UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO

MARCOS DE SOUZA

O COMPORTAMENTO DE TERMOS DA CIÊNCIA DA INFORMAÇÃO POR MEIO DA MODELAGEM DE TÓPICOS

Belo Horizonte – MG

2020

MARCOS DE SOUZA

O COMPORTAMENTO DE TERMOS DA CIÊNCIA DA INFORMAÇÃO POR MEIO DA MODELAGEM DE TÓPICOS

Versão corrigida

Tese apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação da Universidade Federal de Minas Gerais para obtenção do grau de Doutor, área de concentração Ciência da Informação.

Linha de Pesquisa: Gestão & Tecnologia da Informação e Comunicação (GETIC)

Orientador: Renato Rocha Souza

Belo Horizonte – MG

2020

S729c

Souza, Marcos de.

O comportamento de termos da ciência da informação por meio da modelagem de tópicos [recurso eletrônico] / Marcos de Souza. - 2020.

1 recurso online (404 f. : il., gráf., color.) : pdf.

Orientador: Renato Rocha Souza

Tese (Doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

Referências: f. 256-266.

Apêndices: f. 267-404.

Exigências do sistema: Adobe Acrobat Reader.

1. Ciência da informação – Teses. 2. Modelagem de dados – Teses. 3. Mineração de dados (Computação) – Teses. I. Título. II. Souza, Renato Rocha. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 02:004

Ficha catalográfica: Rosimeire Silva Campos de Lima CRB:6/3145

Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

FOLHA DE APROVAÇÃO

O COMPORTAMENTO DE TERMOS DA CIÊNCIA DA INFORMAÇÃO POR MEIO DA MODELAGEM DE TÓPICOS

MARCOS DE SOUZA

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia.

Aprovada em 30 de setembro de 2020, pela banca constituída pelos membros:

Prof(a). Renato Rocha Souza (Orientador)
FGV/RJ [por videoconferência]

Prof(a). Agnaldo Lopes Martins
Openrobotics Educacional [por videoconferência]

Prof(a). Daniela Lucas da Silva Lemos
UFES [por videoconferência]

Prof(a). Flavio Codeco Coelho
EMAp/FGV [por videoconferência]

Prof(a). Luiz Claudio Gomes Maia
FUMEC [por videoconferência]

Prof(a). Elisângela Cristina Aganette
ECI/UFMG [por videoconferência]

Prof(a). Renata Maria Abrantes Baracho Porto
Escola de Arquitetura/UFMG [por videoconferência]

Belo Horizonte, 30 de setembro de 2020.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

ATA DA DEFESA DE TESE DO ALUNO
MARCOS DE SOUZA

Realizou-se, no dia 30 de setembro de 2020, às 14:00 horas, Videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *O COMPORTAMENTO DE TERMOS DA CIÊNCIA DA INFORMAÇÃO POR MEIO DA MODELAGEM DE TÓPICOS*, apresentada por MARCOS DE SOUZA, por videoconferência, número de registro 2016712303, graduado no curso de SISTEMAS DE INFORMAÇÃO, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Renato Rocha Souza - FGV/RJ [por videoconferência] (Orientador), Prof(a). Agnaldo Lopes Martins - Openrobotics Educacional [por videoconferência], Prof(a). Daniela Lucas da Silva Lemos - UFES [por videoconferência], Prof(a). Flavio Codeco Coelho - EMAp/FGV [por videoconferência], Prof(a). Luiz Claudio Gomes Maia - FUMEC [por videoconferência], Prof(a). Elisângela Cristina Aganette - ECI/UFMG [por videoconferência], Prof(a). Renata Maria Abrantes Baracho Porto - Escola de Arquitetura/UFMG [por videoconferência].

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 30 de setembro de 2020.

Prof(a). Renato Rocha Souza

Prof(a). Agnaldo Lopes Martins

Prof(a). Daniela Lucas da Silva Lemos

Prof(a). Flavio Codeco Coelho

Prof(a). Luiz Claudio Gomes Maia

Prof(a). Elisângela Cristina Aganette

Prof(a). Renata Maria Abrantes Baracho Porto

DEDICATÓRIA

À Maria Olimpia de Souza.

AGRADECIMENTOS

À família Souza, tios - Natalino, Sérgio e Cristiane; primos - Bruno, Camila e Carolina; e afilhados - Júlia, Maria Eduarda, Matheus, Ana Clara e Maria Luíza, alicerces na minha vida pessoal.

À família Gama Chaves, padrinhos - Raimundo e Ana Maria, pelo abrigo e momentos ímpares durante minha passagem por Belo Horizonte. As “irmãs” Andrea e Juliana que sempre torceram por esta conquista, bem como suas respectivas famílias.

À família Lourenço Izo, mais que amigos - Antonio, Viviane, Isabela e Isabel, pelo reencontro, abrigo e capital intelectual acompanhado do Deus Baco em ótimas noites de discussões durante minha passagem pelo Rio de Janeiro.

Ao orientador professor Dr. Renato Rocha Souza e membros da banca examinadora pelas contribuições junto a pesquisa.

Aos coordenadores que atuaram frente ao Programa de Pós-Graduação em Gestão e Organização do Conhecimento (PPG-GOC) da Universidade Federal de Minas Gerais durante minha passagem pelo curso, Dr. Maurício Barcellos Almeida, Dr. Ricardo Rodrigues Barbosa e Dra. Célia da Consolação Dias.

Às secretárias do PPG-GOC, Gisele Reis e Gildenara da Costa Gomes, por toda atenção prestada durante esses anos e principalmente pelas diversas orientações com relação aos procedimentos burocráticos e prazos.

Às bibliotecárias, pesquisadoras, doutoras e parceiras de pesquisas, Fernanda Gomes Almeida e Josiana Florêncio Vieira Régis de Almeida.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais ao qual financiou a pesquisa e a todos os participantes da validação dos resultados pesquisa, sendo professores, bibliotecários e consultores atuantes no mercado de trabalho.

Aos colegas e amizades construídas intra e extraclasse: Adriana Lemos, Aline Azevedo, Ana Junqueira, Amarildo Magalhães, Belkiz Costa, Celsiane Araújo, Cristiano Silva, Danielle Rioga, Eduardo Felipe, Edna Angelo, Elaine Diamantino, Élida Pieri, Fernanda Matos, Gislene Silva, Gustavo Lages, Gracielle Mendonça, Guilherme Porto, Guilherme Rodrigues, Jorge Santa Anna, Junio Lopes, Leila Anastácio, Letícia Peixoto, Livia Marangon, Lúcia Helena Magalhães, Lucinéia Maia, Marcello Rodrigues, Maria Queiroz, Michela Rezende, Mirna Azevedo, Mônica Elisque, Rafael Dias, Silvana Sousa, Wilimar Ruas, Webert Montsho.

À Fernanda Gomes Almeida, ao qual compartilhamos momentos mais que especiais.

A todos que acompanharam e torceram por esta conquista, o meu muito obrigado.

Gratidão!

EPÍGRAFE

[...] Quem bebeu água da fonte
Não vai se esquecer
O que eu sou, eu sou em par
Não cheguei, não cheguei sozinho

Trago no sonho e no sangue
Motivos para lutar
Ladeiras do divino
E becos de fome

Quem cruzou aquela ponte
Não vai se perder
O que eu sou, eu sou em par
Não cheguei, não cheguei sozinho [...]

Castanho
Posada/Lenine

RESUMO

O crescimento da pesquisa, ciência e tecnologia na perspectiva acadêmica tem contribuído para a produção de uma quantidade elevada de informações científicas produzidas em diversos formatos e tipos de documentos da comunicação científica. Levando em consideração a quantidade, variedade e complexidade de informações produzidas, tem sido cada vez mais necessário o uso de tecnologias e métodos para elaboração e produção de registros de informação, além da necessidade de produzir informações sobre informações. A Modelagem de Tópicos, constituída de métodos estatísticos/probabilísticos e recursos tecnológicos, utiliza modelos de algoritmos de aprendizagem que possibilita identificar padrões, organizar coleções, resumir conteúdos, extrair tópicos mais frequentes, identificar relações entre assuntos e mudanças realizadas ao longo do tempo em *corpora* de documentos. Partindo desse princípio, questiona-se: de que forma tem se apresentado, na segunda década do século XXI, os temas da produção científica brasileira na área da Ciência da Informação quando se comparado às áreas e disciplinas já estabelecidas na literatura por pesquisadores como núcleo da área? O objetivo geral buscou verificar a proximidade e o distanciamento entre os temas extraídos dos *corpora* de dados constituídos por documentos científicos com as áreas e disciplinas da Ciência da Informação estabelecidas na literatura. Dentre os objetivos específicos constam identificar, analisar e discutir o comportamento diacrônico dos termos extraídos dos *corpora* de dados, bem com suas respectivas relações, além de analisar e discutir os modelos de treinamento de extração de tópicos, selecionar os resultados significativos e validar junto à comunidade científica brasileira da Ciência da Informação. Justifica-se a importância desta pesquisa uma vez que a comparação entre estudos – mesmo que utilizando de metodologias e intervalos de tempo diferentes na composição de documentos – permite apresentar, por meio do mapeamento científico, novos resultados e prospectar diferentes cenários e perspectivas para a ciência estudada. Para a pesquisa empírica foram realizadas as etapas de coleta de dados e formação dos *corpora* de dados; preparação e pré-processamento referente à limpeza, manipulação, combinação e normalização dos dados; transformação dos dados referentes às operações matemáticas e estatísticas aplicadas; modelagem e processamento, ao qual conecta os dados tratados aos modelos *Latent Semantic Indexing* e *Latent Dirichlet Allocation*; apresentação dos resultados por meio de sínteses textuais e gráficos interativos e estatísticos; validação dos resultados junto a pesquisadores da área estudada; e documentação gerada a partir dos resultados empíricos com o referencial teórico. Dentre os principais resultados constam: o comportamento parcialmente diferente entre o mapeamento científico das disciplinas do núcleo da Ciência da Informação encontrado na literatura com os resultados empíricos desta pesquisa; o comportamento diacrônico e surgimento de termos em pesquisas na área da Ciência da Informação, como *fake news*, *big data* e *machine learning*; a proximidade e o distanciamento entre disciplinas como Sistemas de Informação e Comunicação Científica Eletrônica; os melhores resultados na modelagem de tópicos realizada por meio do modelo *Latent Dirichlet Allocation*, levando em consideração o equilíbrio entre os pesos dos resultados e um maior número de bigramas e trigramas que contribuem para a uma melhor interpretação dos dados, realizada pelo indexador e validada pela comunidade científica.

Palavras-chave: Modelagem de tópicos; Alocação de Dirichlet Latente; Proximidade e distanciamento; Comportamento diacrônico.

ABSTRACT

The growth of research, science and technology from an academic perspective has contributed to the production of a large amount of scientific information produced in various formats and types of scientific communication documents. Considering the amount, variety and complexity of information produced, it has been increasingly necessary to use technologies and methods for the elaboration and production of information records, in addition to the need to produce information about information. The Topic Modeling consisting of statistical / probabilistic methods and technological resources uses models of learning algorithms that make it possible to identify patterns, organize collections, summarize content, extract more frequent topics, identify relationships between issues and changes made over time in corpora of documents. Based on this principle, the question is: in what way has the themes of Brazilian scientific production in the area of Information Science been presented in the second decade of the XXI century when comparing the areas and disciplines already established in the literature by researchers as the core of the area? The general objective was to verify the proximity and the distance between the themes extracted from the data corps constituted by scientific documents and the areas and disciplines of Information Science established in the literature. Among the specific objectives were to identify, analyze and discuss the diachronic behavior of the terms extracted from the data corpora, as well as their respective relationships, and to analyze and discuss the training models for topic extraction, to select the significant results and to validate them with the Brazilian scientific community of Information Science. The importance of this research is justified since the comparison between studies, even if using different methodologies and time intervals in the composition of documents, allows presenting, through scientific mapping, new results and prospecting different scenarios and perspectives for the studied science. For the empirical research were carried out the steps data collection and formation of data corpora, preparation and pre-processing referring to cleaning, manipulation, combination and normalization of data, transformation of the data referring to mathematical operations and applied statistics, modeling and processing to which connects the data treated with the Latent Semantic Indexing models, and Latent Dirichlet Allocation, presentation of the results through textual synthesis and interactive graphics and statistics, validation of the results with researchers in the studied area and documentation generated from the empirical results with the theoretical reference. Among the main results are the partially different behavior between the scientific mapping of the disciplines of the Information Science core found in the literature with the empirical results of this research; diachronic behavior and emergence of terms in research in the area of Information Science such as fake news, big data and machine learning; Proximity and distance between disciplines such as Information Systems and Electronic Scientific Communication; Better results in the modeling of topics using the Latent Dirichlet Allocation model taking into account the balance between the weights of the results and a greater number of bigrams and trigrams that contribute to a better interpretation of the data carried out by the indexer and validated by the scientific community.

Keywords: Topic Modeling; Latent Dirichlet Allocation; Proximity and Distance; Diachronic behavior.

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABRACD – Associação Brasileira de Ciência de Dados

AGFORV – Associação dos grupos de folias de reis de Valença

ANCIB – Associação Nacional de Pesquisa em Ciência da Informação

BDJUR – Banco de Dados Jurídico

BDTD – Biblioteca Digital Brasileira de Teses e Dissertações

BENANCIB – Base de Encontro Nacional de Pesquisa em Ciência da Informação

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CBISSN – Centro Brasileiro do ISSN

CE – Campos Especializados

CIE – Ciência da Informação Geral

CP – Conhecimento Prático

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

Comblin – Centro de Pesquisa e Documentação José Comblin

CYC-ES – Conselho Técnico-Científico da Educação Superior

DAP – Download Accelerator Plus

DGA – Dados Governamentais Abertos

ECM – *Enterprise Content Management* – Gerenciamento de Conteúdo Corporativo

ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação

FA – *Factor Analysis* – Análise Fatorial

FAPERJ – Fundação Carlos Chagas de Amparo à Pesquisa do Rio de Janeiro

FCRB - Fundação Casa de Rui Barbosa

FGV – Fundação Getúlio Vargas

FRBR – *Functional Requirements for Bibliographic* Records – Requisitos Funcionais para Registros Bibliográficos

GAD – Gestão Arquivística de Documentos

GTs – Grupos de Trabalhos

IBBD – Instituto Brasileiro e Bibliografia e Documentação

IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia

IDE – *Integrated Development Environment* – Ambiente de desenvolvimento integrado

IGC – Instituto de Geociências

IR – *Information Retrieval* - Recuperação da Informação

ISBN – *International Standard Book Number* – Número de Livro Padrão Internacional

ISSN – *International Standard Serial Number* – Número de Série Padrão Internacional

HMM – *Hidden Markov Model* – Modelo oculto de Markov

LDA – *Latent Dirichlet Allocation* – Alocação de Dirichlet Latente

LDE – Livro Digital Eletrônico

LSA – *Latent Semantic Analysis* – Análise Semântica Latente

LSI – *Latent Semantic Indexing* – Indexação Semântica Latente

MAST - Museu de Astronomia e Ciências Afins

MDS – *Multidimensional Scale* – Escala Multidimensional

OJS – *Open Journal Systems*

OWL – *Ontology Web Language*

PCA – *Principal Components Analysis* – Análise de Componentes Principais

PLN – Processamento de Linguagem Natural

pLSA – *Probabilistic Latent Semantic Analysis* – Análise Semântica Latente Probabilística

pLSI – *Probabilistic Latent Semantic Indexing* – Indexação Semântica Latente Probabilística

PMLL – Plano Municipal do Livro, Leitura

PMLLLB – Plano Municipal do Livro, Leitura, Literatura e Biblioteca

PRESERVE – Programa de Preservação do Patrimônio Histórico do Ministério dos Transportes

PUC – Pontifica Universidade Católica

RE – *Regular Expression* – Expressão Regular

RDF – *Resource Description and Access* – Descrição de recursos e acesso

RDA – *Resource Description Framework*

Reuni – Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais

RID – Recursos de Informação Digitais

SINASC – Sistema de Informações sobre Nascidos Vivos

SIS – Sistemas de Informação em Saúde

SKOS – *Simple Knowledge Organization System* – Sistema Simples de Organização do Conhecimento

SVD – *Singular Value Decomposition* – Decomposição de Valor Singular

TICs – Tecnoloigas de Informação e da Comunicação

UDESC – Universidade do Estado de Santa Catarina

UEL – Universidade Estadual de Londrina

UFAM – Universidade Federal do Amazonas

UFBA – Universidade Federal da Bahia

UFC – Universidade Federal do Ceará

UFCa – Universidade Federal do Cariri

UFF – Universidade Federal Fluminense

UFMG – Universidade Federal de Minas Gerais

UFPB – Universidade Federal da Paraíba

UFPE – Universidade Federal de Pernambuco

UFRJ – Universidade Federal do Rio de Janeiro

UFRN – Universidade Federal do Rio Grande do Norte

UFRGS – Universidade Federal do Rio Grande do Sul

UFS – Universidade Federal de Sergipe

UFSC – Universidade Federal de Santa Catarina

UFSCAR – Universidade Federal de São Carlos

UFSM – Universidade Federal de Santa Maria

UNB – Universidade de Brasília

UNICAMP – Universidade Estadual de Campinas

UNESP – Universidade Estadual Paulista

UNIRIO – Universidade Federal do Estado do Rio de Janeiro

USP – Universidade de São Paulo

URL – Uniform Resource Locator

WIDat - Workshop de Informação, Dados e Tecnologia

SUMÁRIO

1. CONSIDERAÇÕES INICIAIS

O crescimento da pesquisa, ciência e tecnologia na perspectiva acadêmica tem contribuído para um grande volume de informações científicas produzidas nos mais diversos formatos e tipos de documentos, como teses e dissertações, artigos científicos, resumos expandidos, patentes e livros. Essas informações estão armazenadas em formato digital, em repositórios, como bibliotecas digitais, anais de eventos, periódicos e servidores de dados.

Alguns dos fatores que contribuem para o crescente volume da produção de pesquisas científicas e publicações são: a) critérios para credenciamento e reconhecimento de cursos de graduação e pós-graduação estabelecidos pelo Ministério da Educação (MEC); b) Qualis[1] ou fator de impacto[2] para classificação dos periódicos científicos; e c) índice de relevância de um pesquisador aferido por meio do índice-h[3]. Esses fatores formam uma tríade constituída por: órgão regulamentador; ambientes de suporte, armazenamento e disseminação da informação; e capital intelectual dos pesquisadores e seus pares.

A Ciência da Informação, enquanto área interdisciplinar e que se relaciona com outras ciências – como Administração, Arquivologia, Biblioteconomia, Ciência da Computação, Ciência Política, Ciências Contábeis, Comunicação, Direito, Economia, Educação, Epistemologia, Estatística, Ética, Filosofia, Filosofia da Ciência, História da Ciência, Linguística, Matemática, Museologia, Psicologia e Sociologia da Ciência –, tem produzido uma quantidade representativa de diferentes tipos de pesquisas científicas. Essas, por sua vez, perpassam pelos processos da comunicação científica para serem validadas até chegar à fase de publicação.

Diferentes pesquisas no âmbito da Ciência da Informação abordam estudos que envolvem diversas disciplinas. Dentre elas, podem-se destacar: bases de dados; bibliometria; bibliotecas digitais/virtuais; comunicação científica

---

[1] O Qualis, Qualis-Periódicos ou Qualis/CAPES é um sistema brasileiro de avaliação de periódicos, mantido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf/.

[2] Fator de impacto é o método bibliométrico para avaliar a importância de periódicos científicos em suas respectivas áreas.

[3] Índice-h é uma proposta para quantificar a produtividade e o impacto de cientistas com base nos seus artigos mais citados.

eletrônica; economia da informação, formação e aspectos profissionais; gestão da informação; inteligência competitiva e gestão do conhecimento; mineração de dados; necessidades de informação; políticas de informação; processamento automático da linguagem; representação da informação; sistemas de informação; sistemas de recuperação da informação; tecnologia da informação; e teoria da ciência da informação.

A partir da diversidade de áreas correlatas, que envolvem pesquisas em Ciência da Informação e disciplinas que compõem o núcleo da área, pode-se considerar a necessidade de realizar o mapeamento do conhecimento científico da área, de forma a identificar os principais assuntos discutidos na segunda década do século XXI. Considerando a qualidade dos resultados, faz-se necessário um estudo com um quantitativo de dados representativo e de múltiplas fontes de informações que possibilitam abarcar um maior número de pesquisas e assuntos da Ciência da Informação.

Tal mapeamento, se realizado apenas utilizando-se de esforços cognitivos e técnicos de pesquisadores sem o uso das tecnologias, poderia se tornar uma tarefa humanamente impossível. Alinhado a esse cenário, as bases de dados acabam por gerar dificuldades na organização e no desenvolvimento de análise qualitativa desses documentos. Questões como custos elevados e recursos necessários para análise de dados ressaltam a importância da automatização de sistemas para trabalhos com grandes volumes de dados, visando a mapear o conhecimento científico da área.

Levando em consideração a necessidade do mapeamento científico na área da Ciência da Informação mediante sua interdisciplinaridade e o quantitativo de disciplinas que compõem a área, **questiona-se**: de que forma tem se apresentado, na segunda década do século XXI, os temas da produção científica brasileira na área da Ciência da Informação, comparando-se às áreas e disciplinas já estabelecidas na literatura?

Levando em consideração a quantidade, variedade e complexidade de informações produzidas, não só no meio acadêmico, mas de maneira geral, tem sido cada vez mais necessário o uso de tecnologias e métodos para elaboração e produção de registros de informação. Consequentemente, a necessidade de disponibilizar as informações aos interessados constitui-se uma das principais barreiras na identificação e acesso às informações. Dessa forma, a necessidade

de produzir informações sobre informações tem-se desenvolvido com o objetivo de documentar os registros existentes nas áreas do conhecimento, bem como suas características e localização.

Pressupõe-se que, realizar manualmente quaisquer tipos de estudos em coleções de documentos científicos sem fazer uso de aportes tecnológicos – seja, por exemplo, em uma análise de assunto, análise de conteúdo ou mesmo por meio da modelagem de tópicos que busca descobrir e analisar os temas e suas relações por meio de anotações – possa ser uma tarefa morosa para o pesquisador no que diz respeito ao fator tempo de pesquisa. Dependendo também do tamanho do *corpus* de dados a ser estudado, torna-se humanamente impossível, levando-se em consideração o nível de dificuldade das atividades a serem executadas. Além disso, um estudo dessa característica e amplitude pode se tornar factível ao erro ao ser reproduzido por outros pesquisadores, uma vez que, além do conhecimento específico de cada pesquisador, existe o fator subjetividade na interpretação dos dados entre os pares, podendo interferir no fator qualidade da pesquisa.

Considera-se que a Modelagem de Tópicos constituída de métodos estatísticos/probabilísticos e recursos tecnológicos possa obter melhores resultados por meio de algoritmos de treinamentos que possibilitem identificar determinados padrões, organizar coleções, resumir conteúdo, extrair tópicos mais frequentes, identificar relações entre os assuntos e mudanças realizadas ao longo do tempo contidos em grandes *corpora* de dados eletrônicos, trazendo resultados de forma mais precisa e rápida quando se comparado ao trabalho de análise realizado de forma manual pelo profissional indexador. Apesar da rapidez e precisão proporcionadas pela modelagem de tópicos, sugere-se a validação dos resultados alcançados por especialistas da área.

Esta tese apresenta técnicas integradas para a identificação automática de áreas de pesquisa contidas em *corpora* de dados e sua posterior representação por meio de rótulos para facilitar a compreensão do seu conteúdo. Os algoritmos de *Machine Learning Latent Semantic Indexing* (LSI) - Indexação Semântica Latente e *Latent Dirichlet Allocation* (LDA) - Alocação de Dirichlet Latente utilizados na Modelagem de Tópicos representam um conjunto de palavras importantes contidas em documentos ao qual ocorrem em combinação

com outras palavras de linguagem natural extraídas automaticamente de documentos não marcados e não supervisionados.

Acredita-se como **hipótese** que, dentre os modelos de algoritmos de *Machine Learning* aplicados nos *corpora* de dados, o modelo LDA, mesmo sendo considerado uma técnica mais simples, possa alcançar resultados mais assertivos em relação ao modelo LSI, por decompor um *corpus* em seus temas constituintes e reduzir os efeitos adversos gerados pela sinonímia e polissemia por meio da identificação de associações estatísticas entre os termos. O modelo LSI possui uma variação do método de recuperação vetorial no qual as dependências entre os termos dos documentos de um determinado *corpus* possuem relevância em sua representação. Isso porque é simultaneamente explorada na recuperação por meio de suas inter-relações entre termos e documentos.

A pesquisa tem como **objetivo geral** analisar a proximidade e o distanciamento entre os temas abordados nos *corpora* de dados constituídos por trabalhos científicos, tais como teses e dissertações dos programas brasileiros de pós-graduação em Ciência da Informação, além de artigos científicos e resumos expandidos do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) com as áreas e disciplinas da Ciência da Informação estabelecidas na literatura por Pinheiro (2006) e que também aparecem em perspectivas diferentes em pesquisas de Zins e Santos (2015).

Compõem os **objetivos específicos** desta pesquisa:

- Identificar, analisar e discutir o comportamento diacrônico dos termos extraídos dos *corpora* de dados ao longo do período analisado, bem como suas respectivas relações; e
- Analisar e discutir os resultados dos modelos de treinamento de extração de tópicos, selecionar os resultados significativos e validar junto à comunidade científica brasileira da Ciência da Informação.

Dentre os **diferenciais** da pesquisa está o baixo volume de estudos do tipo teses e dissertações de programas brasileiros de pós-graduação modalidade *stricto sensu* que abordaram em seu título ou assuntos sobre Modelagem de Tópicos. Foram encontradas duas dissertações na Biblioteca

Digital Brasileira de Teses e Dissertações[4] (BDTD) e nove dissertações no Catálogo de Teses e Dissertações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES[5]), sendo uma delas armazenada na Plataforma Sucupira[6]. Torna-se importante destacar os seguintes pontos: i) nenhuma dessas pesquisas foram produzidas por cursos da área da Ciência da Informação; e ii) nenhuma das pesquisas realiza um estudo comparativo entre o que já se encontra na literatura com os resultados encontrados na pesquisa empírica. Faz necessário ressaltar que, mesmo existindo uma lacuna temporal entre a pesquisa de Pinheiro (2006) e esta tese, não foram desenvolvidas durante tal intervalo outras pesquisas que apresentam características similares e que possam contribuir para a consolidação das disciclipnas que copõem o núcleo das áreas da Ciência da Informação no Brasil.

Os *corpora* de dados da pesquisa empírica foram constituídos por documentos elaborados por pesquisadores brasileiros e que representam a comunidade científica da área da Ciência da Informação do país. Essa composição foi realizada justamente para comparar os resultados empíricos com as áreas e disciplinas do núcleo da Ciência da Informação encontradas na literatura, de forma que fosse possível identificar padrões ou alteração na curva de tópicos e assuntos produzidos da área.

Dessa forma, **justifica-se** a importância desta pesquisa, uma vez que a comparação entre estudos – mesmo que utilizando de metodologias e intervalos de tempo diferentes – permite apresentar, por meio do mapeamento científico, novos resultados e prospectar diferentes cenários e perspectivas para a ciência estudada, tais como contribuições práticas, metodológicas e científicas de ensino, pesquisa e extensão.

O **escopo da tese** é constituído entre os itens de introdução, referenciais teórico e empírico, resultados e discussão e considerações finais, sendo estruturada em 6 capítulos:

- Capítulo 1 – apresenta a introdução, composta de contextualização da pesquisa, buscando nortear o leitor sobre a temática do estudo; seguido

[4] Biblioteca Digital Brasileira de Teses e Dissertações. Disponível em: http://bdtd.ibict.br/vufind/. Acesso em: 20/03/2019.
[5] Catálogo de Teses e Dissertações da CAPES. Disponível em: https://catalogodeteses.capes.gov.br/catalogo-teses/#!/. Acesso em: 20/03/2019.
[6] Plataforma Sucupira. https://sucupira.capes.gov.br/sucupira/. Acesso em: 20/03/2019.

da problemática da pesquisa, ponto central a ser respondido; pressupostos e hipóteses a serem respectivamente alcançadas e comprovadas; objetivos geral e específicos a serem atingidos; e justificativa que remete à importância e contribuição da pesquisa junto à comunidade cientifica;

- Capítulo 2 – apresenta o referencial teórico, onde se destacam os conceitos e características sobre os assuntos que formam a espinha dorsal da pesquisa. Na primeira seção do capítulo, são apresentados os conceitos e características da Ciência da Informação, desde sua gênese até a interdisciplinaridade, referindo-se, assim, à área de estudos a ser discutida. Na segunda seção do capítulo, são apresentados os Processos da Comunicação Científica que perpassam pela produção intelectual realizada por pesquisadores e estudiosos, além das etapas de desenvolvimento, validação e disseminação da pesquisa científica. Essa seção diz respeito aos tipos de documentos que serão utilizados na pesquisa empírica deste estudo. A terceira e última seção do capítulo apresenta os conceitos sobre Análise de Assunto, uma área da organização e tratamento da informação, no campo da Ciência da Informação. A análise de assunto se refere à etapa de tradução dos conceitos extraídos dos documentos contidos nos *corpora* de dados. Neste trabalho, a análise de assunto é utilizada pelo profissional especialista no domínio da linguagem estudada para validação dos resultados.
- Capítulo 3 – apresenta o referencial empírico, sendo os conceitos necessários e fundamentais para o desenvolvimento prático da pesquisa. Na primeira seção do capítulo, são abordados os temas Linguística Computacional, Linguística de *Corpus* e *Corpus,* que dão aporte à formação dos *corpora* de dados abordados nesta pesquisa, sendo constituídos por documentos científicos da área da Ciência da Informação (referencial teórico). A seção ainda aborda os conceitos de Processamento de Linguagem Natural, que trata computacionalmente os diversos aspectos da comunicação humana e tem como objetivo realizar análise e extração de significados de conteúdo – nesse caso, dos *corpora* de dados. A segunda seção do capítulo aborda sobre os conceitos de

Modelagem de Tópicos, buscando organizar e resumir, por meio de algoritmos de *Machine Learning*, conteúdos de arquivos eletrônicos que constituem grandes volumes de dados e informações;

- Capítulo 4 – refere-se aos procedimentos metodológicos realizados para alcançar os resultados deste estudo. Nesse capítulo são apresentados: a) a classificação da pesquisa quanto ao método, natureza, objetivos, abordagem ao problema e procedimentos técnicos; e b) as etapas da fase empírica, como interação com o mundo interno, preparação e pré-processamento, transformação, modelagem, apresentação e validação dos resultados e documentação;
- Capítulo 5 – apresenta os resultados e discussão alcançadas a partir dos resultados do referencial teórico e da pesquisa empírica. Fazem parte desse capítulo a análise diacrônica dos termos extraídos dos *corpora* de dados, a proximidade e o distanciamento entre os termos, as etapas da modelagem de tópicos e resultados gráficos e textuais, além da validação dos tópicos extraídos das coleções de documentos junto à comunidade científica da Ciência da Informação;
- Capítulo 6 – são apresentadas as considerações finais da pesquisa, levando em consideração o referencial teórico, empírico e os resultados alcançados. Fazem parte da seção a resposta do problema central da pesquisa, o alcance e comprovação do pressuposto e da hipótese, e se os objetivos propostos foram atingidos. Por fim, como forma de contribuição e continuidade da pesquisa cientifica, são apresentadas as perspectivas para pesquisas futuras. Após esse capítulo, são apresentadas as referências, seguidas dos apêndices da pesquisa.

2. REFERENCIAL TEÓRICO

Este capítulo apresenta os conceitos e características sobre os temas que norteiam este estudo e são fundamentais para realizar a pesquisa empírica. Na primeira seção são apresentados a gênese da Ciência da Informação, bem como os paradigmas, conceitos, disciplinas, interdisciplinaridade e histórico da Ciência da Informação no Brasil. Os conceitos são apoiados em autores como Borko (1968), Saracevic (1996), Le Coadic (2004), Oliveira (2005) e Russo (2010). Após essa etapa conceitual, são apresentados dois estudos, sendo um deles de Pinheiro (1997; 2006) e o outro de Zins e Santos (2015), que definem o núcleo da Ciência da Informação. Ambos são utilizados para realizar a discussão entre os resultados teóricos e empíricos da pesquisa.

Ainda neste capítulo, são apresentados os processos da comunicação científica, tais como a produção, o armazenamento, a disseminação e a divulgação de tipos de documento científico, seja por meio da comunicação formal ou informal. Apresenta-se também os meios de divulgação de trabalhos científicos que acompanharam a evolução tecnológica a partir da criação da rede mundial de computadores, destacando os encontros científicos, periódicos científicos, periódicos científicos eletrônicos, bases de dados, anais de eventos, teses e dissertações. A comunicação científica está intrinsecamente associada aos *corpora* de dados desta pesquisa, utilizados na fase empírica, sendo constituídos por teses, dissertações, artigos e resumos científicos apresentados no capítulo 4. A seção apoia-se em autores como Ziman (1979), Lara e Conti (2003), Campello (2007) e Muller (2007).

Por fim, a última seção do capítulo apresenta os conceitos sobre Análise assunto. Trata-se de um tema específico dentro da Ciência da Informação, que se refere aos processos de interpretação de documentos por meio de técnicas de seleção, compreensão e hierarquização de informações, realizada por um especialista no domínio da linguagem e que tem como finalidade a extração de conceitos. Especificamente para esta tese, o tema Análise de assunto respalda-se em autores como Cesarino e Pinto (1980), Fujita (2003), Lancaster (2003) e Dias e Naves (2007).

2.1. Ciência da Informação

A Ciência da Informação é um campo científico ainda em construção, com divergências em sua gênese e conceituação, principalmente no que diz respeito às suas tecnologias. Autores como Rayward (1994) e Barreto (2008) entendem que os traços básicos da área da Ciência da Informação surgiram antes da explosão informacional e que caracterizou o Período Pós-Segunda Guerra Mundial. Outros autores associam a origem da área da Ciência da Informação à busca por soluções para os problemas informacionais surgidos no período Pós-Guerra (NHACUONGUE; FERNEDA, 2015).

Contudo, na literatura científica é possivel encontrar diversos autores que enfatizam o marco inicial da Ciência da Informação a partir da explosão informacional, caracterizada pelo grande número de relatórios que documentavam a Segunda Guerra Mundial e pelas conferências propostas por Vannevar Bush para debater soluções tecnológicas (NHACUONGUE; FERNEDA, 2015).

Assim, a Ciência da Informação teve seu início na década de 1950, aliada ao uso dos computadores, tanto no desenvolvimento quanto na disseminação de seu uso pós Segunda Guerra Mundial (CAPURRO; HJORLAND, 2007). De acordo com Wersig (1993), sua origem está atrelada à inovação tecnológica, entretanto, sua fundamentação se encontra na explosão informacional pós-guerra. Diferente das ciências clássicas, em que a gênese está na busca pelo entendimento completo do funcionamento do mundo, a Ciência da Informação surgiu da necessidade de desenvolver estratégias para solucionar problemas causados por ciências e tecnologias, de forma que seja possível impor mudanças significativas de conhecimento, como a visão técnico-sistema para visão usuário/humano (WERSIG, 1993).

Dessa forma, considera-se a gênese da Ciência da Informação, em uma perspectiva histórica, entendida como uma inovação tecnológica que sempre motivou o homem, desde a fase das pinturas rupestres, passando pela invenção da escrita e chegando até à era da internet, destacando, dessa forma, o fluxo da informação e sua distribuição ampliada. Mesmo que de maneira indireta, torna-se possível relacionar a busca de soluções para a produção, organização e disseminação da informação (BARRETO, 2008).

A Ciência da Informação traz na sua origem um movimento acelerado das tecnologias da informação e comunicação, que, por sua vez, buscam diferentes soluções tecnológicas, levando em consideração os meios de abordagens, natureza, manifestações e efeitos da informação e conhecimento para garantir o fluxo e o uso da comunicação. É destacado nesse processo o ciclo constituído por produção, organização, armazenamento, representação, disseminação, recuperação, acesso e uso (NHACUONGUE; FERNEDA, 2015).

Três pontos fundamentais são destacados no surgimento da Ciência da Informação (PINHEIRO, 2002), sendo eles: 1 – a explosão da informação demandada durante a Segunda Guerra Mundial mediante o avanço científico e tecnológico, além dos periódicos científicos (RUSSO, 2010); 2 – a necessidade de registro, transmissão de informação e conhecimento da pesquisa em desenvolvimento (MIRANDA, 2002); e 3 – o surgimento de novas tecnologias, com destaque para o computador no processo de informações bibliográficas na década de 1970 (BARRETO, 2007).

Conforme as demais ciências que surgiram no movimento pós Segunda Guerra, a Ciência da Informação apresenta características próprias mediante as necessidades de reunir, organizar e deixar acessível a tríade de conhecimento produzido pelo mundo – conhecimento cultural, conhecimento científico e conhecimento tecnológio (OLIVEIRA, 2011). Dessa maneira, a Ciência da Informação se desenvolveu ao longo do tempo mediante os problemas informacionais, apresentando sua relevância junto à sociedade (SARACEVIC, 1996).

Com a expansão informacional e influenciado pela preocupação de registro de transmissão de informação e conhecimento, em 1945, foi publicado o artigo intitulado *As we may think*, de Vannevar Bush, criador do Memex. O estudo apresentou a importância da preservação e armazenamento de documentos científicos de forma que pudessem ser disponibilizados para pesquisas futuras (QUEIROZ; MOURA, 2015).

Influenciada pelas teorias de Norbert Wiener, Claude Shannon e Warren Weaver, e a de Bertalanffy, a Ciência da Informação foi se moldando ao longo da sua trajetória inicial. Norbert Wiener contribuiu com a sua Teoria da Informação, publicada em 1947, na obra intitulada *Cybernetics or control and communication in the animal* and *the machine*. A Teoria Matemática da

Comunicação ou a Teoria da Informação de Claude Shannon e Warren Weaver inspiraram a Ciência da Informação, apresentando conceitos de entropia, redundância e ruído, fundamentais para os sistemas de recuperação de informação. A influência de Bertalanffy junto à Ciência da Informação está explicitada na sua Teoria Geral dos Sistemas, publicada em 1956, que aborda os conceitos de sistemas de informação e redes (OLIVEIRA, 2011).

A Ciência da Informação, a partir do problema da informação, surgiu no século XX com a documentação, que buscava solucionar o entrave conhecido como “dilúvio da literatura”. São destacados a partir deste viés o foco na recuperação da informação juntamente com complexos da tecnologia que se tornaria ciência. Influenciada pelo advento da tecnologia. A área da Ciência da Informação surge na mudança do papel do conhecimento para indivíduos, organizações e culturas. Exemplo dessa mudança, mediante os meios tecnológicos, está na forma como o conhecimento passou a ser disseminado pela multiplicidade de usuários por meio da comunicação impressa (WERSIG, 1993).

No Brasil, a Documentação precedeu a Ciência da Informação entre as décadas de 1900 e 1921 por meio de Peregrino da Silva, diretor da Biblioteca Nacional (CASTRO, 2000). A Ciência da Informação foi introduzida no Brasil na década de 1970 com o curso *stricto sensu* na modalidade de mestrado em Ciência da Informação pelo substituído Instituto Brasileiro e Bibliografia e Documentação (IBBD) e atual Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Destaca-se, em 1972, a criação do periódico Ciência da Informação (PINHEIRO, 1997; RUSSO, 2010).

Tanto os cursos de pós-graduação quanto os periódicos científicos desempenham um papel de suma importância para a pesquisa científica, principalmente no que diz respeito à promoção da comunicação e da discussão de seus resultados nas sociedades científicas (OLIVEIRA, 1998). Entretanto, um dos desafios a ser reforçado está nos cursos da graduação, que acabam por formar profissionais instrumentalistas, mas com pouca ênfase à pesquisa. Com isso, os cursos de pós-graduação na área da Ciência da Informação têm recebido discentes de diferentes áreas do conhecimento, porém, com pouca transformação da área (SOUZA; ALMEIDA; BARACHO, 2015).

Com o passar dos séculos e mediante os avanços científicos e tecnológicos, ocorrem alterações de paradigmas. No século XXI, com o advento da era digital para difusão do conhecimento, a Ciência da Informação está cada vez mais em evidência e a informação, na era tecnológica, cada vez mais disponível para os usuários da rede mundial de computadores (QUEIROZ; MOURA, 2015). A Ciência da Informação acompanhou os novos paradigmas que surgiram a partir da evolução da sociedade e ao longo do tempo (LE COADIC, 2004).

Oliveira (2011) defende que o paradigma da Ciência da Informação ocorre dentro de um sistema de comunicação humana, sendo constituído por um grupo de ideias associadas ao processo que engloba o movimento da informação. Le Coadic (2004) destaca: a) paradigma do trabalho coletivo, substituindo o paradigma do trabalho individual; b) paradigma do fluxo, substituindo o paradigma do acervo; c) paradigma do uso voltado para o usuário, contrapondo o paradigma voltado para o bibliotecário, documentalista e museólogo; e d) paradigma do elétron, no lugar do paradigma do papel.

Já Capurro (2003) destaca: a) paradigma físico receptor; b) paradigma cognitivo, cujo conhecimento é insuficiente para atender às necessidades de um determinado usuário. Assim, o usuário busca informação de forma que seja possível transformá-lo ou não; e c) paradigma social, em que existe uma associação referente à transmissão de uma mensagem de o emissor e o usuário de informação, podendo o seu conhecimento sofrer influências dos meios sociais e materiais ao qual está inserido.

A indústria da informação surge entre as décadas de 1950 e 1960. A Ciência da Informação teve seu início nos Estados Unidos entre 1961 e 1962 a partir dos estudos a respeito da recuperação da informação. Soma-se a isso a evolução dos cartões perfurados para os CD-ROM e, posteriormente, aos acessos on-line; dos sistemas não interativos para os interativos; dos textos escritos para formatos multimídia; das bases de documentos para as bases de conhecimento; e da recuperação de citações para textos completos (SARACEVIC, 1996).

Harold Borko escreveu o conceito de Ciência da Informação a partir das ideias de Taylor, realizadas em duas reuniões com diversos profissionais, dentre eles, bibliotecários e docentes. As reuniões foram realizadas em outubro de 1961

e abril de 1962, no *Georgia Institute of Technology*, no *Natural Science Foudantion* – Estados Unidos (BARRETO, 2007).

O conceito de Ciência da Informação estabelecido por Borko é definido como:

> Ciência da Informação é aquela disciplina que investiga as propriedades e o comportamento informacional, as forças que governam os fluxos de informação, e os significados do processamento da informação, para uma acessibilidade e usabilidade ótima. Ela está preocupada com o corpo de conhecimentos relacionados à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação, e utilização da informação. Isto inclui a investigação da representação da informação em ambos os sistemas, naturais e artificiais, o uso de códigos para a transmissão eficiente da mensagem, e o estudo do processamento de informações e de técnicas aplicadas aos computadores e seus sistemas de programação (BORKO, 1968, p. 3).

Shera e Cleveland corroboram Borko em relação à investigação e ao comportamento da informação, tanto para as forças que regem o fluxo informacional quanto para os meios de processamento da informação para uma melhor acessibilidade e usabilidade de qualidade. Entretanto, acabam por acrescentar os processos de informação constituídos pela sua geração, disseminação, organização, armazenamento, recuperação e uso (SHERA; CLEVELAND, 1977).

Tal conceito também é enriquecido por Capurro e Hjorland (2007), que destacam os processos da informação, tais como geração, coleta, transformação, interpretação, armazenamento, recuperação, disseminação e uso para o domínio particular das tecnologias modernas da área. Oliveira (2011) corrobora os autores, apresentando o fluxo composto pelos processos de produção, coleta, organização, armazenamento, recuperação, interpretação e transmissão.

Além disso, enfatizam que a Ciência da Informação, enquanto disciplina, procura estudar um corpo de conhecimento científico, tecnológico e de sistemas (CAPURRO; HJORLAND, 2007). Trata-se de uma ciência que está relacionada às tecnologias da informação e voltada para o estudo científico do comportamento humano em busca de informação e maneiras de processá-la. Além dessas características, são destacadas questões como armazenamento, organização e manipulação de dados realizadas por meio de computadores (SOUZA, 2007).

Souza, Almeida e Baracho (2015) também corroboram a definição de Borko (1968) sobre o termo Ciência da Informação, entretanto, identificam um viés mais amplo, relacionado aos processos de representação do conhecimento e de registros associados à informação.

A Ciência da Informação refere-se a uma área de estudos voltada para questões da produção científica e possui amplo interesse nas tecnologias informacionais no qual: “[...] investiga dentro das estruturas e propriedades (e não um conteúdo específico) da informação científica, tanto quanto as regularidades do trabalho de informação científica, suas teorias, história, metodologia, e organização” (MIKHAILOV; GILYAREVSKIJ, 1970, p. 14); ao uso das necessidades da informação mediante a “[...] prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual” (SARACEVIC, 1996, p.47); e “[...] à regularidade do fenômeno relativo à sua dispersão e uso, obsolescência, epidemiologia de sua propagação e outros aspectos detectados no processo de manipulação e analise da literatura” (MIRANDA, 2002, p.10).

A Ciência da Informação é uma área de estudos que busca resolver problemáticas da comunicação do conhecimento e suas respectivas formas de registros entre a raça humana. Essa premissa se aplica tanto na área científica quanto na área de prática profissional, que enfoca o contexto social, institucional ou individual, dentre as necessidades da informação. As tecnologias informacionais possuem um papel fundamental para esse fazer, contribuindo para uma estrutura que possibilita compreender o passado, o presente e o futuro da Ciência da Informação, tais como as questões problemas (SARACEVIC, 1996). Como característica geral, a Ciência da Informação permite o processo de gerar, transmitir e fazer uso da informação por meio da observação e do estudo em grupos sociais, destacando assim a mais própria das Artes enquanto a informação é criada (PINHEIRO, 1997).

O conceito de Ciência da Informação não aparenta uma uniformidade na literatura, apresentando diferentes significados em diferentes domínios do conhecimento. Isso implica diferentes campos que, por sua vez, seguem diferentes abordagens e tradições. Exemplos estão nas abordagens objetivas *versus* cognitivas e nas tradições bibliotecária *versus* documental *versus*

computacional (ZINS, 2007). Entretanto, as diversas definições sobre Ciência da Informação estabelecidas pelos autores podem ser consideradas sinônimas entre si. A variação ocorre mediante as diferentes perspectivas teóricas das escolas de Biblioteconomia e Ciência da Informação no mundo, destacando os diferentes aspectos e tendências construídas ao longo da história da área (HJORLAND, 2000 apud BICALHO, 2009).

Fazem parte das características gerais da Ciência da Informação: a) a interdisciplinaridade – que se apresenta nas relações entre as ciências da computação e inteligência artificial, além dos trabalhos teóricos e experimentos; b) o uso das tecnologias da informação – referente à transformação da sociedade moderna em sociedade da informação de forma que a tecnologia tem possibilitado mudança na quantidade e qualidade da informação e da comunicação; e c) a evolução da sociedade da informação – associado ao papel do desenvolvimento econômico e social da Ciência da Informação (SARACEVIC, 1996).

A estrutura da Ciência da Informação está baseada entre os vínculos da pesquisa empírica e da pesquisa profissional. Ambas buscam solucionar problemas que essa área se dispõe a resolver. A Ciência da Informação pode se relacionar de forma interdisciplinar, multidisciplinar ou transdisciplinar com áreas como a Biblioteconomia, Ciência da Computação, Documentação e Inteligência Artificial mediante o problema de excesso de informação, que se torna cada vez mais complexo a partir dos adventos tecnológicos, sobretudo a Web (NHACUONGUE; FERNEDA, 2015).

As disciplinas da Ciência da Informação reconhecem a existência de uma conexão com áreas tradicionais como a “Psicologia (Psicologia da informação), Sociologia (Sociologia da informação), Economia (Economia da informação), Ciência política (Política da informação) e tecnologia (Tecnologia da informação)” (WERSIG; NEVELING, 1975, p. 36). Tais áreas do conhecimento, sob a denominação das “Ciências da Informação” (no plural), contemplam a própria Ciência da Informação, Biblioteconomia, Arquivologia, Museologia, Comunicação e Educação, que, por sua vez, são vinculadas à Teoria Geral da Informação e aplicáveis em quaisquer áreas do conhecimento. Contemplam áreas e teorias que se relacionam com a Ciência da Informação como: Cibernética, Ciência da Ciência, Ciência da Computação, Direito, Filosofia,

Linguística, Matemática, Semiótica, Teoria da comunicação e Teoria dos sistemas (WERSIG; NEVELING, 1975).

A área da Ciência da Informação estuda, no campo humano, o fluxo da comunicação entre os seus respectivos autores, o que envolve os meios de registros responsáveis por transportar informação e conhecimento (ODDONE, 1998). No campo da computação, as áreas de busca, recuperação, qualidade e uso da informação são tratadas pela área da Ciência da Informação (SARACEVIC, 1996). Trata-se de uma disciplina com origem na documentação, recuperação da informação e bibliografia a qual relaciona-se à interdisciplinaridade com as áreas da Biblioteconomia, Ciência da Computação por meio da área de inteligência artificial e aspecto cognitivos por meio da Comunicação (SOUZA, 2007).

A Ciência da Informação, entre outras ciências interdisciplinares, como a Ciência da Computação, a Comunicação Social e a Ecologia, nasceu no centro da revolução científica e técnica, que se encaminhou à Segunda Guerra Mundial. As disciplinas de Documentação e Recuperação da Informação foram fundamentais para a sua origem e desenvolvimento. A primeira apresentou novas conceituações e a segunda viabilizou o surgimento de sistemas automatizados de recuperação da informação (OLIVEIRA, 2011). No estudo da documentação e recuperação da informação, a ciência da informação no campo interdisciplinar inclui tópicos como classificação, linguagem, linguística, ciência do comportamento, transferência e está relacionada à comunicação e ao comportamento (HARMON, 1971).

A Ciência da Informação é derivada e está relacionada a diversas áreas do conhecimento, tais como artes gráficas, biblioteconomia, comunicações, gestão, linguística, lógica, matemática, psicologia, tecnologias computacionais dentre outras disciplinas. Por isso, caracteriza-se como ciência interdisciplinar. Além disso, possui componentes da ciência dura, que investiga sem a necessidade de aplicação, e a ciência aplicada, que possibilita o desenvolvimento de produtos e serviços (BORKO, 1968).

Embora o fato de Borko não ter experimentado os avanços das telecomunicações e microeletrônica após a invenção dos microcomputadores ao conceituar Ciência da Informação, isso não o impediu de prospectar nove campos de pesquisa da área, tais como análise linguística, tradução, demanda

da informação, análise e projeto de sistemas, padrões de reconhecimento de imagens e voz, sistemas especialistas, linguagens documentárias, produção e reprodução de documentos (SOUZA, 2007).

Chain Zins (2007) corrobora Souza ao caracterizar a Ciência da Informação como uma abordagem científica e interdisciplinar, relativa aos fenômenos da informação e suas aplicações tecnológicas utilizadas para transformar informação em mensagem de conteúdo significativo. São levados em consideração os contextos histórico, cultural e social, além de conceitos, princípios, métodos, teorias e leis relativas aos fenômenos para tal transformação (ZINS, 2007).

A interdisciplinaridade na área da Ciência da Informação tem sido discutida por diversos autores ao longo da história. Borko (1968, p. 3) destaca a "matemática, lógica, linguística, psicologia, tecnologia da computação, pesquisa operacional, artes gráficas, comunicação, biblioteconomia e administração, entre outros". Para Shera e Cleveland (1977, p. 265), “[...] o campo é derivada de ou relacionada à matemática, lógica, linguística, psicologia, tecnologia computacional, pesquisa operacional, artes gráficas, comunicações, biblioteconomia, administração e algumas outras áreas”. Segundo Foskett (1973, p. 164) refere-se à “[...] antiga arte da biblioteconomia, a nova arte da computação, as artes dos novos meios de comunicação e ciências como a psicologia e a linguística”. Wersig e Neveling (1975, p. 30) acreditam ser “parte da matemática, lógica, filosofia da ciência, gramática transformacional e teoria matemática da comunicação”. Para Brookes (1980, p. 128), é “[...] uma mistura peculiar de linguística, comunicação, estatística e metodologia da pesquisa, junto com algumas técnicas da biblioteconomia, como indexação e classificação”. Le Coadic (1996, p. 22) acredita ter relação com “[...] a psicologia, a linguística, a informática, a matemática, a lógica, a estatística, a sociologia, a economia, o direito, a filosofia, a política e as telecomunicações”. Para Saracevic (1996, p. 48) a Ciência da Informação está ligada à “[...] biblioteconomia, ciência da computação, ciência cognitiva (incluindo inteligência artificial - IA) e comunicação”. Oliveira (2011, p. 20) relaciona à “biblioteconomia, ciência da computação, comunicação social, administração, linguística, psicologia, lógica, matemática, filosofia/epistemologia”.

A Ciência da Informação enquanto característica interdisciplinar colabora com disciplinas como direito, economia, eletrônica, filosofia, informática, linguística, lógica, matemática, política, psicologia, sociologia e telecomunicações (LE COADIC, 2004). A interdisciplinaridade também é explorada na Ciência da Informação em outras quatro áreas, sendo elas: a) biblioteconomia – compartilham seu papel social e preocupa-se com os problemas de utilização em questões como registros gráficos; b) ciência da computação – o uso da computação por meio de algoritmos para tratar informações enquanto a ciência da informação foca seus esforços na natureza da informação e sua comunicação realizada pelos seus usuários; c) ciência cognitiva – representada pela inteligência artificial e com contribuições em sistemas de informação, seja sistemas e interfaces inteligentes, bases de conhecimento, hipertextos e questões relacionadas à interação homem-máquina; e d) comunicação – estudo em conjunto do interesse da informação como fenômeno e da comunicação como processo (SARACEVIC, 1996).

A pesquisadora Lena Vania Ribeiro Pinheiro apresentou em sua tese, intitulada “A Ciência da Informação entre sombra e luz: domínio epistemológico e campo interdisciplinar”, uma estrutura classificatória que apresenta as disciplinas científicas, pilares da Ciência da Informação, distribuídas em cinco categorias. Trata-se do resultado de uma pesquisa empírica e teórica que categorizou as disciplinas integrantes da Ciência da Informação com base nas características e modalidades do conhecimento, sendo disciplinas estruturais, disciplinas de representação ou instrumentais, disciplinas gerenciais e disciplinas tecnológicas (PINHEIRO, 1997). As categorias são detalhadas com as disciplinas conforme Pinheiro (2002 apud BRAMBILLA; STUMPF, 2007):

- Disciplinas estruturais: reunindo pesquisas históricas, teóricas e epistemológicas sobre conceitos, metodologias, princípios, leis e interdisciplinaridade da Ciência da Informação, Bibliometria ou Informetria ou, ainda, Cientometria, Comunicação Científica e Tecnológica;
- Disciplinas de representação ou instrumentais: abrangendo pesquisas sobre os processos de descrição e análise (catalogação, classificação e indexação) para o sistema de recuperação da informação, instrumentos como linguagens documentárias, vocabulários controlados e tesauros, normas e padrões nacionais e internacionais, incluindo as específicas para Web, como o Dublin core etc.;
- Disciplinas gerenciais: pesquisas voltadas ao planejamento e administração de bibliotecas especializadas, centros, serviços, redes e sistemas de informação, base de dados, organização e processamento

> da informação, gestão da informação, economia da informação, inteligência competitiva e gestão do conhecimento da empresa, sistemas gerenciais de informação, em abordagem mais de aspectos administrativos e recursos (humanos, bibliográficos, materiais, financeiros);
>
> - Disciplinas tecnológicas: abordando a implantação, operação e avaliação de redes e sistemas, redes e serviços de informação especializados, inclusive DSI e busca retrospectiva, com ênfase na abordagem dos aspectos tecnológicos, de produção e acesso a bases de dados, arquitetura da informação, serviços e produtos de informação na Web, bibliotecas digitais e virtuais, OPACs, arquivos abertos, mecanismos de buscas; e - Disciplinas socioculturais ou de transferência da informação: política de informação, necessidades, acesso e uso da informação ou antigos estudos de usuários, informação em Arte e Cultura, divulgação científica etc.

O professor Chaim Zins, da University of Haifa, de Israel, desenvolveu uma pesquisa denominada "*knowledge map of information science: issues, principles, implications"*, que teve como objetivos: 1) clarificar as diferentes concepções de Ciência da Informação; 2) desenvolver um mapa do conhecimento amplo, sistemático e cientificamente válido do domínio do conhecimento da Ciência da Informação; e 3) fundamentar esse mapa em sólidas bases teórica (ZINS, 2005).

Com o resultado da pesquisa de Zins, a pesquisadora Lena Vania Ribeiro Pinheiro ampliou o número de disciplinas de cada categoria, conforme apontado no Quadro 01. As disciplinas estruturais foram agrupadas em Fundamentos da Ciência da Informação. Já as disciplinas instrumentais foram representadas por Organização e Processamento da Informação. As disciplinas gerenciais foram agrupadas pelas Tecnologias da Informação com o nome de Gestão da Informação. Por fim, a disciplina sociocultural fora associada a Transparências da Informação. A autora justifica as alterações conceituais, terminológicas e disciplinares mediante os avanços e mutações que a Ciência da Informação sofreu ao longo do tempo, com destaque para sua forma de reflexão e o amadurecimento intelectual (PINHEIRO, 2006).

Quadro 01 – Disciplinas e subdisciplinas da Ciência da Informação

Disciplinas em Ciência da Informação	Subdisciplinas
Fundamentos da Ciência da Informação	• Bibliometria / Informetria / Cientometria / Webmetria; • Formação profissional; • Epistemologia da Ciência da Informação; • Estudos interdisciplinares (relações epistemológicas com a Ciência da Computação, Comunicação Social, Museologia, Biblioteconomia, Arquivística, Arte etc.); • História da Ciência da Informação; • Metodologias da Ciência da Informação; • Teoria da Informação;
Organização e processamento da informação	• Arquitetura de informação; • Organização do conhecimento / Representação da informação; o Catalogação; o Classificação; o Indexação; o Metadados; o Tesauros; o Vocabulários controlados; • Ontologia; • Processamento automático de linguagem;
Gestão da Informação	• Disseminação da informação (produtos e serviços de informação); • Economia da informação; • Gestão de qualidade de informação; • Gestão do conhecimento; • Inteligência competitiva; • Marketing de informação;
Tecnologias da Informação	• Automação de bibliotecas; • Bases de dados; • Bibliotecas virtuais e digitais; • Comunicação mediada por computador - Internet/Web; • Mineração de dados; • Preservação e segurança digital; • Redes e sistemas de informação; • Sistemas de recuperação da informação;
Transferência de Informação	• Competência informacional (“*information literacy*”); • Comunicação científica; • Divulgação científica; • Educação à distância; • Estudos de necessidades e usos de informação; • Estudos de usuários; • Ética na informação; • Inclusão digital; • Políticas de informação;
Aplicações de informação	• Informação científica; • Informação tecnológica; • Informação industrial; • Informação em Arte; • Informação em bibliotecas; • Informação em arquivos; • Informação em museus.

Fonte: (PINHEIRO, 2006).

Em seu desenvolvimento epistemológico, a Ciência da Informação prioriza a descrição e a explicação de fenômenos por meio de três vertentes: 1) conceituações básicas de termos da área; 2) descrição de sua estrutura, campo e ação; e 3) estratégias metodológicas. Os interesses aplicados estão relacionados à utilização dos resultados da pesquisa nos contextos práticos, como previsão de acesso a recursos de informação e planejamento (SAVOLAINEN, 1992).

Pinheiro (2006) apresenta o território epistemológico da Ciência da Informação, constituído por estudos teóricos e pesquisas empíricas, tanto na literatura nacional como internacional, ao qual traça, por meio de seus resultados, um núcleo constituído por 17 subdisciplinas e áreas interdisciplinares, conforme apresentado no Quadro 02:

Quadro 02 – Subdisciplinas da Ciência da Informação e suas áreas interdisciplinares

Subáreas	Áreas interdisciplinares
01 - Sistemas de informação	Administração; Ciências Contábeis;
02 - Tecnologia da informação	Ciência da Computação;
03 - Sistemas de recuperação da informação	Biblioteconomia; Ciência da Computação; Linguística;
04 - Políticas de informação	Administração; Ciência Política; Direito;
05 - Necessidades de informação	Arquivologia; Biblioteconomia; Museologia; Psicologia;
06 - Representação da informação	Arquivologia; Biblioteconomia; Filosofia; Linguística; Museologia;
07 - Teoria da ciência da informação	Epistemologia; Filosofia; Filosofia da Ciência; Matemática;
08 - Formação e aspectos profissionais	Educação; Ética; Direito;
09 - Gestão da informação	Administração; Economia; Estatística;
10 - Bases de dados	Ciência da Computação;
11 - Processamento automático da linguagem	Biblioteconomia; Ciência da computação; Linguística;
12 - Economia da informação	Administração; Economia;
13 – Bibliometria	Estatística; História da Ciência; Matemática; Sociologia da Ciência;
14 - Inteligência competitiva e gestão do conhecimento	Administração; Economia;
15 - Mineração de dados	Ciência da computação;
16 - Comunicação científica eletrônica	Ciência da computação; Comunicação; História da Ciência; Sociologia da Ciência;
17 - Bibliotecas digitais/virtuais	Biblioteconomia; Ciência da Computação; Comunicação.

Fonte: (PINHEIRO, 2006).

Pinheiro desenvolveu a pesquisa empírica de sua tese e, quase 10 anos depois, repensou os procedimentos metodológicos e aplicou a pesquisa novamente. Os procedimentos adotados para análise do *corpus* de dados foram

divididos entre leitura dos títulos, resumos e de conteúdos de artigos. Nessa análise, destacam-se os títulos dos tópicos da estrutura de artigos do tipo de revisão, utilizados para definição dos temas e tendo como fonte pesquisas da Revisão Anual da Ciência e Tecnologia da Informação. No primeiro momento, foram analisados 307 documentos publicados entre o período de 1966-1995, alcançando uma frequência de 99,93% (PINHEIRO, 1997). No segundo momento, foram analisados 81 documentos publicados entre 1996-2004 e com frequência de 100,01% (PINHEIRO, 2006). A pesquisa resultou em um total de 17 temas/assuntos e disciplinas – alteração de terminologia –, organizadas por ordem decrescente de frequência, conforme apresentado no Quadro 03:

Quadro 03 – Temas/assuntos e disciplinas por ordem de frequência

1966-1995			1996-2004		
Temas / Assuntos	**Nº de artigos**	**Frequência**	**Disciplinas**	**Nº de artigos**	**Frequência**
Sistemas de informação	43	14	Sistemas de recuperação da informação	15	18,51
Tecnologia da informação	28	9,12	Representação da informação	9	11,11
Disseminação da informação	27	8,79	Tecnologia da informação	8	9,87
Políticas de informação	23	7,49	Sistemas de informação	6	7,40
Necessidades e usos de informação	22	7,16	Bibliometria	6	7,40
Sistemas de recuperação da informação	20	6,51	Inteligência competitiva e Gestão do conhecimento	5	6,17
Computadores e programas	19	6,18	Mineração de dados	5	6,17
Representação da informação	16	5,21	Política de informação	5	6,17
Automação de bibliotecas	15	4,89	Teoria da Ciência da Informação	5	6,17
Redes de informação	14	4,56	Comunicação científica eletrônica	3	3,70
Formação e aspectos profissionais	14	4,56	Necessidades e usos da informação	3	3,70
Bases de dados	13	4,23	Administração de informação	2	2,50
Organização e processamento da informação	13	4,23	Bibliotecas digitais	2	2,50
Administração da informação	12	3,90	Economia da informação	2	2,50

Teoria da Ciência da informação	11	3,58	Formação e aspectos profissionais	2	2,50
Processamento automático de linguagem	9	2,93	Processamento automático de linguagem	2	2,50
Economia da informação	8	2,60	Bases de dados	1	1,23

Fonte: Adaptado de (PINHEIRO, 2006).

É possível observar uma distribuição de frequência equilibrada no primeiro estudo, realizado com artigos publicados entre 1966-1995, exceto para os temas/assuntos de Sistemas de Informação e Tecnologia da Informação. Tal equilíbrio também ocorre com as disciplinas de cunho tecnológico e de caráter social, representadas respectivamente por Sistemas de Informação, Tecnologias da Informação, Sistemas de Recuperação da Informação, Computadores e Programas e Disseminação da Informação, Políticas de Informação, Necessidade e Usos de Informação. A pesquisa empírica, realizada em 2004, apresentou resultados comparativos e evolutivos com relação aos temas/assuntos, denominadas disciplinas (PINHEIRO, 2006).

No segundo estudo são destacadas a Recuperação e a Representação da Informação. Elas passam a ser a questão central e estão estreitamente relacionadas. Se comparado com o estudo anterior, destacam-se dois pontos: 1) disciplinas que não fazem mais parte do quadro 1966-1995 – Disseminação da Informação, Computadores e Programas, Automação de Bibliotecas, Redes de Informação e Organização e Processamento da Informação. Tratam-se de disciplinas amplas, que foram fragmentadas; 2) novos temas incluídos na pesquisa 1996-2004 – Na situação inversa estão a Bibliometria, Inteligência Competitiva e Gestão do Conhecimento, Mineração de dados, Comunicação Científica Eletrônica e Biblioteca Digitais. Torna-se importante ressaltar que Bibliometria e Comunicação Científica Eletrônica apareceram durante a primeira pesquisa, entretanto, com baixa frequência (PINHEIRO, 2006).

Com base nos estudos realizados por Pinheiro – sua tese, defendida em 1997, intitulada “A Ciência da Informação entre sombra e luz: domínio epistemológico e campo interdisciplinar” e um artigo publicado em 2006, intitulado “Ciência da Informação: desdobramentos disciplinares, interdisciplinaridade e transdisciplinaridade” –, foi possível estabelecer um melhor entendimento epistemológico da Ciência da Informação. Os resultados,

organizados por frequência, puderam estabelecer o núcleo das disciplinas consolidadas da área, conforme apresentado no Quadro 04.

Quadro 04 – Disciplinas do núcleo da Ciência da Informação

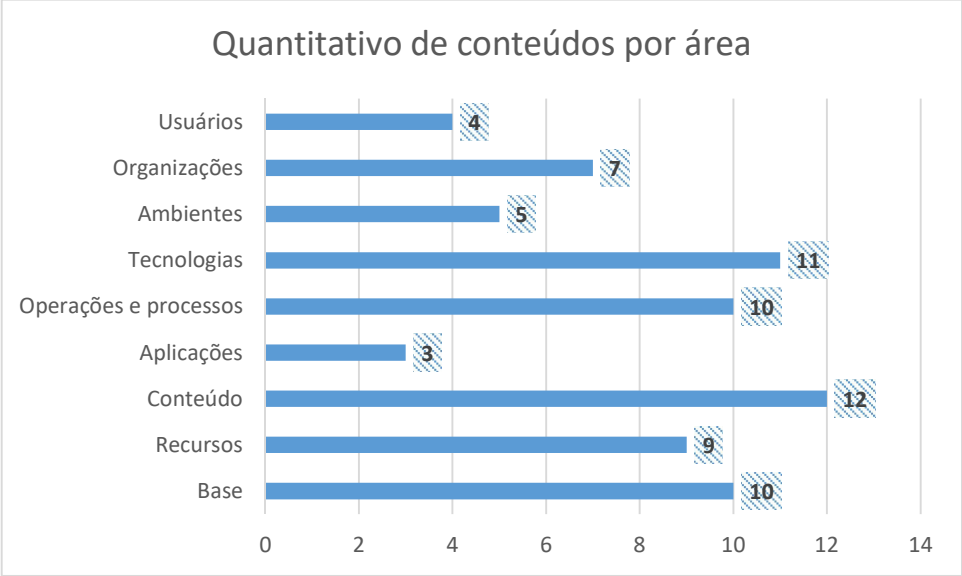
Disciplinas	Frequência
01. Sistemas de informação	49
02. Tecnologia da informação	36
03. Sistemas de recuperação da informação	35
04. Políticas de informação	28
05. Necessidades e usos de informação	25
06. Representação da informação	25
07. Teoria da Ciência da Informação	16
08. Formação e aspectos profissionais	16
09. Gestão da informação	14
10. Bases de dados	14
11. Processamento automático da linguagem	11
12. Economia da informação	10
13. Bibliometria	6
14. Inteligência competitiva e Gestão do conhecimento	5
15. Mineração de dados	5
16. Comunicação científica eletrônica	3
17. Bibliotecas digitais/virtuais	2

Fonte: (PINHEIRO, 2006).

Outra pesquisa foi realizada por Zins e Santos (2015), com o objetivo de melhorar a formação acadêmica dos cursos de Biblioteconomia e Ciência da Informação no Brasil. O estudo intitulado "*Brazilian model of library and information studies in the bachelor's level"* molda os conteúdos dos programas acadêmicos por meio de uma estruturação do raciocínio curricular. Foram utilizadas duas metodologias, sendo Delphi Crítico quem permitiu realizar uma série de críticas reflexivas e de discussões aprofundadas por pares estruturados e moderados por especialistas, além da avaliação assistemática baseada na teoria fundamentada para a formação das categorias de conteúdo. O estudo resultou em um modelo composto por dois modelos auxiliares: um processo sistemático de desenvolvimento em quatro etapas e um plano estruturado de 288 categorias de conteúdo (ZINS; SANTOS, 2015).

O Gráfico 01 apresenta as áreas destacadas nesse estudo e o quantitativo de conteúdos propostos para serem abordados em seu respectivo estágio durante a formação do acadêmico, englobando, assim, as subáreas Ciência da Informação Geral (CIG), Campos Especializados (CE) e Conhecimento Prático (CP).

Gráfico 01 – Áreas de estudos em Ciência da Informação



Fonte: Elaborado pelo autor. Adaptado de (ZINS; SANTOS, 2015).

O Quadro 05 apresenta as áreas, subáreas - Ciência da Informação Geral (CIG), Campos Especializados (CE) e Conhecimento Prático (CP) com os conteúdos relacionados à proposta de conteúdos para os cursos de pós-graduação em Ciência da Informação:

Quadro 05 – Áreas e subáreas do campo da Ciência da Informação

Ciência da Informação	Educação Básica	Conteúdos
Base	CIG	Filosofia da ciência da informação, história da ciência da informação, metodologia de pesquisa, epistemologia, ética, estatística
	CE	Teoria dos estudos arquivísticos, teoria das bibliotecas, teoria dos museus
	CP	Metodologia de Pesquisa
Recursos	CIG	Qualidade da informação, serviços da informação
	CE	Recursos arquivísticos, recursos de biblioteca, recursos do museu, gerenciamento de registros, padrões
	CP	Serviços de informação, padrões internacionais (recursos)
Conteúdo	CIG	Organização do conhecimento, representação do conhecimento, metadados, ontologias, taxonomias
	CE	Informações comerciais, informação educacional, informação jurídica, informação médica, informação científica, informação social
	CP	Desenvolvimento de coleção
Aplicações	CIG	Busca de informação, redes sociais
	CE	Publicação eletrônica
Operações e processos	CIG	Pesquisa de informação, representação do conhecimento, visualização de informação, informatização
	CE	Digitalização, preservação digital, avaliação arquivística
	CP	Pesquisa de informação, tratamento descritivo, classificação
Tecnologias	CIG	Tecnologias da informação, sistemas de informação, tecnologias de rede
	CE	Curadoria digital, arquitetura de informação, bibliotecas digitais, repositórios digitais, sistema de informação, serviços web

	CP	Internet, programação
Ambientes	CIG	Ética da informação, política da informação
	CE	Economia da informação, política da informação
	CP	Ética da informação
Organizações	CIG	Gestão do conhecimento, economia da informação
	CE	Gestão do conhecimento, organizações privadas (empresas), organizações públicas (ONG ou GO), aspecto organizacional
	CP	Gestão do conhecimento
Usuários	CIG	Estudos de usuários, psicologia social
	CE / CP	Estudos de usuários

Fonte: Adaptado de (ZINS; SANTOS, 2015).

O estudo realizado por Zins e Santos (2015) contribui para o desenvolvimento e avaliação dos cursos de Biblioteconomia e para os programas de pós-graduação em Ciência da Informação do Brasil, ao fornecer um modelo estruturado para a seleção de conteúdo a ser abordado nos respectivos cursos.

Uma gama de pesquisadores brasileiros tem contribuído para o desenvolvimento da área da Ciência da Informação. A Figura 01 apresenta a centralidade de grau dos autores mais representativos do país e que possuem publicações em revistas científicas internacionais e nacionais de relevância indexadas em base de dados como a *Web of Science*. Refere-se à centralidade de grau a medida correspondente ao número de enlaces que possui um nó com os demais. O estudo, realizado entre 1994 e 2013, destaca a produção brasileira da Ciência da Informação na *Web of Science* (PINTO; MATIAS; GONZÁLEZ, 2016).

Figura 01 – Autores brasileiros mais representativos da Ciência da Informação

Fonte: (PINTO; MATIAS; GONZÁLEZ, 2016).

Ainda de acordo com Pinto, Matias e González (2016), foi possível localizar um total de 1.334 autores em 742 artigos científicos. Com relação aos pesquisadores brasileiros, tornou-se possível identificar uma produção científica representativa. O Quadro 06 destaca os nomes dos pesquisadores, o quantitativo de publicações e o vínculo com a Instituição de Ensino Superior (IES).

Quadro 06 – Pesquisadores brasileiros com publicações na *Web of Science*

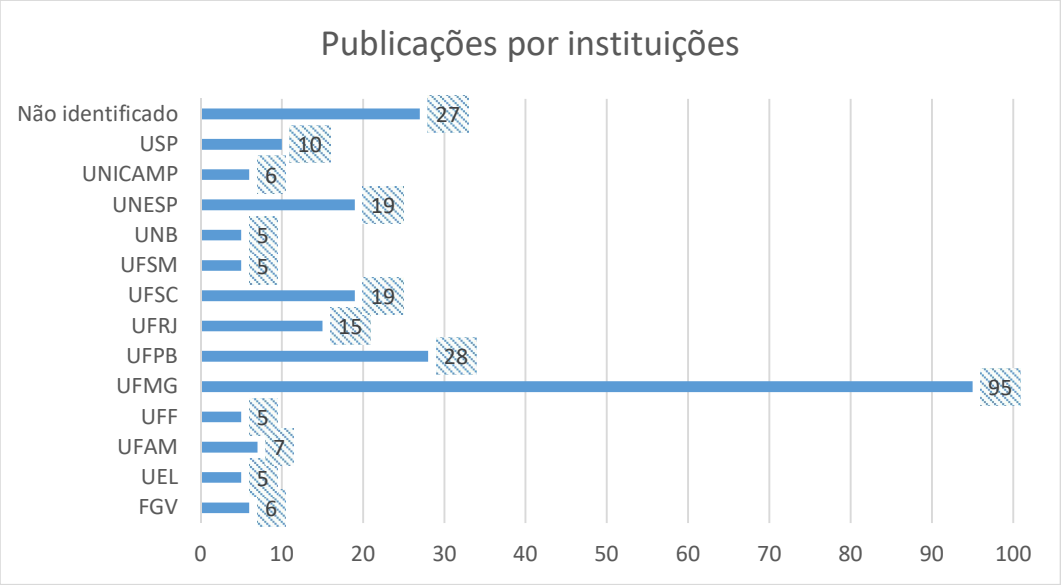
Pesquisadores	Publicações	IES
Marcos André Gonçalves	15	UFMG
Jacqueline Leta	15	UFRJ
Alberto Henrique Frade Laender	13	UFMG
Leilah Santiago Bufrem	10	UFPB
Beatriz Valadares Cendón	9	UFMG
Mariângela Spotti Lopes Fujita	8	UNESP
Rosângela Schwarz Rodrigues	8	UFSC
Ricardo Rodrigues Barbosa	7	UFMG
Altigran Soares da Silva	7	UFAM
E. S. de Moura	7	Não identificado
Isa Maria Freire	7	UFPB
Berthier Ribeiro de Araujo Neto	7	UFMG
Maurício Barcellos Almeida	6	UFMG
Mirian de Albuquerque Aquino	6	UFPB
Carlos Alberto Ávila Araújo	6	UFMG
Miriam Figueiredo Vieira da Cunha	6	UFSC
José Augusto Chaves Guimarães	6	UNESP
Luiz Antonio Joia	6	FGV
Marlene Oliveira	6	UFMG
Renato Rocha Souza	6	UFMG
Jacques Wainer	6	UNICAMP
Aline Elis Arboit	5	UNESP
Ligia Maria Arruda Café	5	UFSC
L. D. Costa	5	Não identificado
Murilo Bastos da Cunha	5	UNB
E. L. da Silva	5	Não identificado
Fabrício José Nascimento da Silveira	5	UFMG
F. D. de Souza	5	Não identificado
Daniel Flores	5	UFSM
Gercina Ângela. Borém De Oliveira Lima	5	UFMG
Carlos Henrique Marcondes de Almeida	5	UFF
Marcelo Seido Nagano	5	USP
J. C. R. Pereira	5	Não identificado
Francisca Arruda Ramalho	5	UFPB
Nicolau Reinhard	5	USP
Nadia Ameno Ribeiro	5	UFMG
Maria Inés Tomáel	5	UEL
Nívio. Ziviani	5	UFMG

Fonte: Adaptado de (PINTO; MATIAS; GONZÁLEZ, 2016).

Relacionando pesquisadores com publicações na revista *Web of Science* e suas respectivas IES, destaca-se o quantitativo de publicações por

universidades, sendo: Universidade Federal de Minas Gerais, com 95 publicações e 13 pesquisadores; Universidade Federal da Paraíba, com 28 publicações e 3 pesquisadores; Universidade Federal de Santa Catarina e Universidade Estadual Paulista, ambas com 19 publicações e 3 pesquisadores, conforme apresentado no Gráfico 02.

Gráfico 02 – Quantitativo de publicações por instituições



Fonte: O autor. Elaborado a partir de (PINTO; MATIAS; GONZÁLEZ, 2016).

A pesquisa sobre o estudo referente à produção científica brasileira da Ciência da Informação na *Web of Science,* ao qual apresenta a relação dos pesquisadores mais citados, é baseada no índice-h, que se refere a um indicador proposto por Jorge E. Hirsch, em 2005, que busca avaliar a produção científica individual dos pesquisadores com base na quantidade relativa de trabalhos. O índice-h é definido da seguinte maneira: “[...] um cientista tem índice h se h de seus Np artigos tiver ao menos h citações cada, e os outros artigos (Np - h) tiver ≤ h citações cada" (HIRSCH, 2005, p. 16.569). Dessa forma, para se definir o índice-h de um determinado pesquisador, faz-se necessário um maior número de “h” de artigos científicos com ao menos o mesmo número “h” de citação para cada publicação.

Um dos indicadores mais simples e utilizados é a contagem de publicações de um determinado pesquisador, entretanto, isso não reflete o impacto do pesquisador junto à comunidade científica. O grande volume de produção científica não está associado ao impacto, visibilidade ou reconhecimento pelos pares (OLIVEIRA; GRACIO, 2011). Faz-se necessário

ressaltar que não existe nenhum método quantitativo com abordagem qualitativa para mensurar a produção científica. Isso porque não há como medir a utilização da obra por pesquisadores da área ou a inserção em bibliografias de disciplinas específicas de forma automática para que as agências de fomento possam utilizar como parâmetro de avaliação (FONSECA, 2015).

Assim como houve alterações nos resultados das pesquisas realizadas por Pinheiro (1997; 2006) e como a publicação de Zins e Santos (2015) já possui cinco anos, as três pesquisas utilizadas como núcleo para temas/assuntos da Ciência da Informação para esta tese, acredita-se que novos termos possam surgir como resultado, mediante o avanço tecnológico e a necessidade de produção, organização, armazenamento, representação, disseminação, recuperação, acesso e o uso da informação, alguns deles como *Big Data*, *Machine Learning*, *Web* Semântica, Computação nas Nuvens, Gestão de Dados e E-science.

A seção seguinte apresenta os processos da comunicação científica, o que envolve a produção intelectual realizada por pesquisadores e estudiosos, tais como as diferentes atividades de comunicação, aos tipos de documentos científicos, à validação por pares, aos tipos de publicação e à disseminação de pesquisas científicas.

2.2. Processos da comunicação científica

Todo o trabalho intelectual realizado por pesquisadores e estudiosos depende de um sistema de comunicação que permita aos cientistas a comunicação dos resultados de suas próprias pesquisas e a obtenção de informações de resultados alcançados por outros pesquisadores. Esse sistema, que envolve diferentes atividades de comunicação entre os pesquisadores, é constituído por dois grupos de comunicação, sendo eles a comunicação formal e a comunicação informal. A comunicação formal utiliza os canais formais de comunicação para a divulgação do conhecimento científico. Normalmente são publicações com divulgação mais amplas, tais como livros e periódicos científicos. Os canais formais permitem um acesso mais amplo, em que as informações são geralmente mais trabalhadas e de fácil coleta e armazenamento. Já a comunicação informal caracteriza-se por canais de caráter

mais pessoal ou de pesquisas ainda em andamento, podendo ser apresentados em congressos ou eventos afins (MUELLER, 2007a).

Uma pesquisa pode produzir uma série de publicações construídas durante e após o seu término. Com isso, podem ser divulgadas em um conjunto de publicações, denominada literatura científica. Existe uma variação entre formatos para disseminação da informação, tais como relatórios, palestras, trabalhos apresentados em congressos, artigos de periódicos e livros que podem ser divulgados em formato impresso ou meio eletrônico, possuir audiência como estudantes ou público em geral ou com a função de informar, registrar autoria, ou indicar a localização dos documentos, por exemplo. Faz-se necessário ressaltar que nesse processo o autor deve expor sua pesquisa ao julgamento por seus pares de forma a almejar consenso, que confere a confiabilidade para validação de uma pesquisa científica (ZIMAN, 1979).

Disseminar informação possibilita tornar público a produção do conhecimento. Sobre o conteúdo disponibilizado pelo produtor ou centro difusor, não existem garantia quanto aos usuários atingidos, à aplicação efetiva das informações e ao sucesso das operações de divulgação. Fatores como conceito de informação e usuários acabam por envolver problemas de delimitação de público e linguagem (LARA; CONTI, 2003). A informação pode fluir por muitos canais de comunicação e por diferentes tipos de documentos que seguem os seus modelos particulares, dependendo do estágio da pesquisa. Nos canais informais as informações podem não possuir uma fácil recuperação da informação, pois nem sempre serão armazenadas em ambientes virtuais. Exemplos estão nos relatórios de pesquisa, textos apresentados em seminários, atas de reuniões ou mesmo anais de eventos. Já os canais formais possibilitam um acesso amplo de forma que as informações possam ser facilmente coletadas e armazenadas (MUELLER, 2007a).

De acordo com Santos (2011), o processo de produção, armazenamento, disseminação e divulgação de um determinado tipo de produção científica, seja nacional ou estrangeira, perpassam por padronizações estéticas de normatizações como Associação Brasileira de Normas Técnicas (ABNT), Association Française de Normalization (AFOR), Associação Mercosul de Normalização (AMN), American Society for Testing and Materials (ASTM), British Standards Institution (BSI), Deutsches Institut für Normung (DIN), Modern

Language Association (MLA), Instituto de Engenheiro Eletricistas e Eletrônicos (IEEE), International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), Japanese Industrial Standards (JIS), National Fire Protection Association (NFPA), entre outras, são adotadas de acordo com cada país. Tais produções padronizadas e normatizadas são utilizadas pelas bases de dados automatizadas e impressas. As normatizações adotadas pelas bases de dados seguem procedimentos e técnicas de responsabilidades das áreas da Biblioteconomia e da Ciência da Informação (SANTOS, 2011).

O acesso livre ao conhecimento científico pode ser considerado um dos movimentos mais interessantes e importantes da comunicação científica. Tal movimento representa um grande desafio decorrente da expansão e de mudanças provocadas no sistema tradicional erradicado da comunicação do conhecimento científico (MUELLER, 2006).

Fenômeno comum em todas as áreas do conhecimento, a explosão bibliográfica possui como característica o número crescente e uma rapidez no qual os documentos científicos são produzidos. Esse processo ficou evidenciado com o desenvolvimento das tecnologias eletrônicas, que expandiu as possibilidades de comunicação ao eliminar as barreiras geográficas. Um dos modelos mais conhecidos de fluxo da informação científica foi desenvolvido na década de 1970 pelos autores norte-americanos Garvey e Griffith, posteriormente adaptados para as diversas áreas do conhecimento. Nesse modelo são representadas diversas atividades que são cumpridas e seus respectivos documentos gerados pelo pesquisador, perpassando, assim, desde o início da pesquisa a relatórios preliminares, ao término da pesquisa, à apresentação em seminários, à apresentação de relatórios em congressos, a publicações em anais de eventos e periódicos científicos, remetendo a revisões bibliográficas anuais até serem citados na literatura e, consequentemente, utilizados em textos didáticos, manuais e enciclopédias (MUELLER, 2007a).

A constante evolução das mudanças científicas e tecnológicas, atrelada à morosidade das publicações dos periódicos científicos e livros, reflete muitas vezes em resultados que acabam sendo ultrapassados/desatualizados quando divulgados nos canais formais de comunicação. Com isso, outras formas de socialização de resultados são utilizadas entre os pesquisadores, tais como

correios eletrônicos e fóruns de discussões que possibilitam uma maior agilidade no processo de disseminação e comunicação da informação científica. Revistas destinadas exclusivamente à divulgação de resultados parciais de pesquisas e relatórios técnicos de relatos de pesquisa têm contribuído para agilizar o processo da comunicação científica de maneira que os pesquisadores tenham conhecimento antes mesmo da divulgação formal da pesquisa. Outros canais de comunicação – como boletins, revistas, jornais ou listas para divulgação da produção científica – são utilizados por universidades, institutos e centros de pesquisas, sendo, em algumas situações, fornecidas apenas informações fragmentadas (CAMPELLO, 2007b).

Os documentos produzidos ao longo do processo de pesquisa podem ser classificados como: primários – produzidos com interferência direta do autor como teses, dissertações, artigos científicos, patentes; secundários – apresentam informações filtradas e organizadas de acordo com a sua finalidade. Exemplos estão em monografia, livros-textos enciclopédias, anuários; e terciários – possuem a função de guiar o usuário para as fontes primárias e secundárias (MUELLER, 2007a).

A comunidade científica é constituída por muitos grupos sociais que formam a sociedade contemporânea. Dessa forma, existem interesses financeiros das editoras que controlam o mercado de periódicos, interesses por instituições de ensino superior e de pesquisa que buscam prestígio e financiamento, interesses de cunho pessoal dos pesquisadores e interesses nacionais, políticos e econômicos que buscam o desenvolvimento e prestígio nacional (MUELLER, 2006).

Mesmo que as tecnologias estejam acelerando o processo de comunicação científica, seja por meio das teleconferências ou listas de discussões, os eventos científicos presenciais acabam por apresentar possibilidades diferencias como contato pessoal e um maior número de pesquisadores participantes reunidos em um único lugar, o que permite uma troca de informações com maior intensidade. Os números elevados de eventos científicos presenciais que ocorrem em todas as áreas do conhecimento acabam por concretizar que se trata de uma forma de comunicação no qual agrada aos pesquisadores que podem expor e discutir suas pesquisas e serem avaliados por pares de uma forma mais ampla (CAMPELLO, 2007a).

Existem diferentes tipos de encontros científicos e sua denominação ocorre de acordo com a sua abrangência e seus objetivos. Dentre eles, destacam-se: i) Congresso – trata-se de um evento com grandes proporções e se caracteriza pelo âmbito nacional ou internacional, com duração média de uma semana e com participantes da comunidade científica. Incluem palestras, mesas redondas, painéis, conferências realizadas por convidados de destaque da área; ii) Colóquio, encontro, fórum, jornada, reunião, seminário e simpósio – ocorrem em proporções menores ao congresso, tanto quanto à duração, quanto ao número de participantes. Trata de assuntos mais especializados (CAMPELLO, 2007a; SEVERINO, 2011).

Destacam-se como forma de produção e disseminação científica os periódicos científicos, os periódicos científicos eletrônicos, as bases de dados, os anais de eventos, as teses e dissertações. O periódico científico surgiu no século XVII, na Europa, época marcada por mudanças, como o surgimento da ciência moderna em que as evidências baseadas na observação e na experiência empírica passaram a ser aceitas como o principal método de pesquisa no meio científico. Até a ciência moderna, os cientistas se comunicavam pessoalmente ou por meio de cartas. A divulgação formal era realizada em livros e tratados. O periódico científico surge como um meio de comunicação mais amplo em relação à comunicação oral e as cartas, sendo mais ágeis do que a divulgação nos livros e tratados. Por isso, acaba por permitir uma troca de ideias mais rápida entre os cientistas e os interessados nos resultados de pesquisa (MUELLER, 2007b).

Os periódicos são fontes de informação indispensáveis para orientação e pesquisa bibliográfica em todas as áreas, consolidando-se como um dos mais eficientes meios de registro e divulgação de pesquisas intelectuais e originais (CUNHA, 2001). Os periódicos científicos possuem quatro funções, sendo elas: i) comunicação formal dos resultados da pesquisa para a comunidade científica ou interessados; ii) preservação e organização do conhecimento, garantindo a possiblidade de acesso às pesquisas registradas ao longo do tempo; iii) estabelecimento da propriedade intelectual, em que o autor registra formalmente a sua autoria e propriedade de descoberta científica; e iv) manutenção do padrão da qualidade da ciência que se refere à aprovação da pesquisa por especialistas e aprovação da comunidade científica (MUELLER, 2007b).

Os periódicos científicos também possuem problemas inerentes, como a demora para publicação de artigos, custos altos de aquisição e manutenção de coleções, e dificuldades para o pesquisador saber o que do seu interesse está sendo publicado, e ao seu acesso, pois nem sempre sua biblioteca assina o periódico ou conseguem renovar assinaturas de revistas, eletrônicas ou não, no qual o pesquisador necessita devido ao custo das coleções. Outra dificuldade pelos pesquisadores está em conseguir publicações nos títulos principais de sua área ou no reconhecimento internacional da comunidade científica (MUELLER, 2006) (MUELLER, 2007b).

Também conhecido como publicação periódica, publicação seriada, revista técnica ou revista científica, o periódico possui um número serial único que o identifica e evita ambiguidades ou problemas com títulos homônimos. O *International Standard Serial Number* (ISSN) possui as seguintes características: periodicidade na publicação dos fascículos; publicação em partes sucessivas (ano, volume, número); continuidade indefinida de publicação; grande variedade de assuntos e autores (CUNHA, 2001).

De acordo com o Instituto Brasileiro de Informação em Ciência e Tecnologia – (IBICT[7]), a rede ISSN é coordenada pelo Centro Internacional do ISSN com sede em Paris, na França. Trata-se de uma organização intergovernamental, criada em 1971, que constitui de uma fonte completa de informações de publicações seriadas. Ainda segundo o IBICT, as atividades no Centro Nacional da Rede ISSN iniciaram-se em 1975. Já em 1980, firmou-se como Centro Brasileiro do ISSN (CBISSN), por meio de acordo firmado entre o ISSN e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), tornando-se responsável pela gestão do ISSN junto aos editores e usuários em geral. Entre as distribuições e os benefícios do uso do ISSN destacam-se uma maior rapidez, produtividade, qualidade, precisão da identificação e controle relacionado às publicações seriadas.

As hierarquias que ocorrem entre indivíduos de uma comunidade científica também acontecem nos diversos veículos da comunicação científica. Entre elas estão os periódicos, livros e trabalhos de congressos. Têm destaque e prestígio junto à comunidade científica os periódicos indexados em bases de

[7] Instituto Brasileiro de Informação em Ciência e Tecnologia. Disponível em: http://cbissn.ibict.br/index.php/centro-internacional-do-issn/. Acesso em: 20/04/2019.

dados com editores e avaliadores reconhecidos na sociedade científica (MUELLER, 2006),

O termo indexação na área da Ciência da Informação representa a inclusão de um periódico em uma base de dados. Trata-se de um processo e representação de uma determinada temática de pesquisa associada ao conteúdo em um formato de um documento redigido pelo descritor. Ressalta-se que os processos de inclusão de um periódico em quaisquer bases de dados perpassam por uma análise e posterior apresentação do resultado de aprovação ou negação por parte do comitê avaliador (SAMPAIO; SABADINI, 2009).

De acordo com a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES[8]), o sistema de classificação de periódicos utilizado no Brasil para produção intelectual científica de pós-graduação denomina-se Qualis. O sistema de classificação foi idealizado para atender a uma demanda específica de avaliação da produção dos cursos de pós-graduação.

Ainda de acordo com a CAPES, a classificação dos periódicos cadastrados na base de dados é realizada anualmente pelas áreas de avaliação por meio de informações fornecidas pelos programas de pós-graduação junto à Plataforma Sucupira e possui um indicativo de qualidade com conceitos de A1, sendo a maior nota, perpassando por A2; B1; B2; B3; B4; B5; até C para a nota de entrada. Essa classificação é realizada pelos comitês de consultores de cada área de atuação e seguem os critérios aprovados pelo Conselho Técnico-Científico da Educação Superior (CTC-ES), que disponibilizam uma lista de classificação dos periódicos vinculados aos programas de pós-graduação. Os periódicos que possuem mais de uma área de atuação podem receber diversas classificações, sendo respectivamente atribuído um conceito a cada área. Severino destaca o papel dos periódicos e das revistas científicas:

> [...] é fundamentalmente a comunicação dos resultados dos trabalhos de pesquisa à comunidade científica e à própria sociedade como um todo. Elas promovem normas de qualidade na condução da ciência e na sua comunicação. Consolidam critérios para a avaliação da qualidade da ciência e da produtividade dos indivíduos e instituições. Consolidam áreas e subáreas de conhecimento. Garantem a memória da ciência. Representam o mais importante meio de disseminação do conhecimento em escala. São instrumentos de grande importância na

[8] Fundação Coordenação de Aperfeiçoamento de Pessoa de Nível Superior. Disponível em: https://www.capes.gov.br/avaliacao/instrumentos-de-apoio/qualis-periodicos-e-classificacao-de-producao-intelectual/. Acesso em: 20/04/2019.

> constituição e institucionalização de novas disciplinas e disposições específicas (2000, p.198).

A internet tem potencializado uma grande variedade de base de dados nas diversas áreas do conhecimento. O acesso a essas bases, realizada por usuários junto a *web*, tem agilizado consideravelmente o acesso às informações de forma democrática e garantindo a acessibilidade imediata, diferente dos documentos impressos (SANTOS, 2011).

Com a evolução tecnológica por meio da disponibilidade de serviços *online* na rede de computadores a partir de 1994 e a busca por alternativas inovadores, os periódicos científicos eletrônicos surgiram como solução para oferecer maior rapidez na comunicação, flexibilidade de acesso, versatilidade, disponibilidade imediata, custo baixo se comparado aos periódicos impressos, não existindo barreiras de acesso e se tornando universal (MUELLER, 2006; 2007b). Dessa forma, o periódico científico deixa de ser, desde sua publicação, um veículo noticioso, transformando-se em um veículo de divulgação do conhecimento a partir das atividades de pesquisa (MIRANDA; PEREIRA, 1996).

O periódico eletrônico surgiu como alternativa frente às deficiências apresentadas pelo periódico científico tradicional ao possibilitar maior rapidez na comunicação, largo alcance, baixo custo relativo, disponibilidade imediata e facilidade de acesso (MUELLER, 2007b). Os periódicos científicos eletrônicos caracterizam-se por publicações que contemplam controle de qualidade de trabalhos publicados e aceitos em nível internacional, que tenham a pretensão de continuidade, que disponibilizem textos completos por meio de acesso *online* e que possam existir também em CD-ROM, não sendo necessária uma versão impressa (OLIVEIRA, 2008). Além disso, são considerados um meio de comunicação versátil e rápido, já que ultrapassam as barreiras geográficas e permitem agilizar o processo de divulgação da pesquisa logo após a sua conclusão, bem como diversos meios para a recuperação da informação (MUELLER, 2007b).

No que diz respeito ao acesso aberto às informações científicas, Björk (2004) aponta quatro canais importantes: i) periódicos científicos eletrônicos, que surgiram no início da década de 1990, realizam avaliação prévia por pelos pares e são acessíveis sem pagamento; ii) servidores de *e-prints*, utilizados como repositórios para áreas específicas; iii) repositórios para assuntos específicos:

repositórios institucionais de universidades específicas que possuem como objetivo disponibilizar textos apresentados em eventos, canais paralelos aos periódicos. Nesse modelo, os próprios autores podem disponibilizar seus textos; e iv) autoarquivamento em páginas pessoais de autores existentes desde o início da rede mundial de computadores e utilizadas pelos pesquisadores para divulgação de sua produção pessoal.

Existem também os documentos gerados por meio de encontros científicos publicados no formato de anais, reunindo um conjunto de trabalhos apresentados em eventos. Também costumam fazer parte dos anais as palestras e conferências que fizeram parte de um determinado evento. Os anais podem ser resumos dos trabalhos ou pesquisas na íntegra, além disso, podem ser publicados antes ou após o evento (CAMPELLO, 2007a).

A partir de 2009, a agência nacional do *International Standard Book Number* (ISBN) passou a considerar também as produções científicas publicadas no formato de anais de eventos realizados por meio de congressos, encontros, simpósios, seminários, entre outros. O ISBN permite uma periodicidade regular ou irregular, como por exemplo, publicações anuais, bianuais ou de acordo com a importância e o período de cada periódico. No caso de eventos contínuos ou uma revista científica que possui periodicidade, utiliza-se como referencial único o ISSN (SANTOS, 2011).

Bases de dados são consideradas coleções de dados bibliográficos que incluem referências bibliográficas e resumos ou textuais, textos completos de artigos de periódicos, jornais ou outras modalidades de documentos e que servem como suporte a um sistema de recuperação da informação (CUNHA, 2001). Base de dados representa o modo como os dados são armazenados nos computadores, e pode ser definida como uma coleção geral e integrada de dados, gerenciadas para atender as necessidades dos usuários. As bases de dados são formadas por arquivos, constituídos por coleções de registros similares, que se relacionam entre si, e por registros, constituídos por informações que dizem respeito a um documento ou item contido na base de dados (ROWLEY, 2002).

As bases de dados podem ser classificadas como de referências ou de fontes. As bases de referências encaminham o usuário a outras fontes como um documento, pessoas ou o texto completo de um documento, podendo ser

subdivididas em: 1 - bases de dados bibliográficos, que inclui citações ou referências bibliográficas e resumos dos artigos; 2 - bases de dados catalográficos, que refletem o acervo de uma biblioteca ou rede de bibliotecas; e 3 - bases de dados referenciais, que referenciam informações ou dados, como nomes e endereços de instituições. Nas bases de dados de fontes que contêm dados originais, os usuários encontram disponíveis os dados que precisam, sem precisar buscá-los em outras fontes. Podem ser agrupadas em: bases de dados numéricos; de texto integral; textuais e numéricos; e multimídia (ROWLEY, 2002).

As teses e dissertações são documentos que apresentam a pesquisa em formato original, a partir de pesquisas desenvolvidas nos cursos de programas de pós-graduação. No Brasil, as dissertações são originadas dos cursos de mestrado, em que o discente deve passar pelo curso formal e ainda elaborar uma dissertação sobre seu trabalho de pesquisa, demonstrando capacidade de sistematização, além do domínio do tema e da metodologia. Já as teses são originárias dos cursos de doutorado. Nele o discente desenvolve uma tese no qual envolva etapas metodológicas como revisão bibliográfica adequada, sistematização das informações, planejamento e a realização de uma pesquisa considerada original (CAMPELLO, 2007c).

Tanto a dissertação como a tese destinam-se ao resultado final do trabalho de pesquisa do discente. É elaborado sobre um tema igualmente único e delimitado por meio da comunicação, análise e resultados, que, após a exposição pública diante de uma banca examinadora responsável pela aprovação, confere-se o grau de mestre ou doutor (CUNHA, 2001; SEVERINO, 2011).

Documentos pertinentes à comunicação científica ganharam ênfase diante do movimento *Open Access*. Refere-se à disponibilização aberta da informação pertinente à divulgação de resultados de pesquisas científicas em espaços abertos avaliados por pares, tais como documentos de conferência, capítulos de livros, artigos científicos, teses, dissertações e monografias. A ideia do movimento é que os autores e leitores tenham acessos aos documentos científicos sem restrições de acesso, seja eletrônico ou por cópias impressas para quaisquer finalidades, diferentes das grandes bases de dados que impõem

pagamento para ter acesso. A Declaração de Budapeste define o Acesso Aberto como:

> "Acesso aberto" à literatura científica revisada por pares significa a disponibilidade livre na Internet, permitindo a qualquer usuário ler, fazer download, copiar, distribuir, imprimir, pesquisar ou referenciar o texto integral desses artigos, recolhe-los para indexação, introduzi-los como dados em software, ou usá-los para outro qualquer fim legal, sem barreiras financeiras, legais ou técnicas que não sejam inseparáveis ao próprio acesso a uma conexão à Internet. As únicas restrições de reprodução ou distribuição e o único papel para o direito autoral neste domínio é dar aos autores o controle sobre a integridade do seu trabalho e o direito de ser devidamente reconhecido e citado (BOAI, 2012).

Acredita-se que o movimento *Open Access* tenha contribuído positivamente na formação dos *corpora* de dados dessa tese, considerando que as bibliotecas digitais brasileiras de teses e dissertações começaram a ser implementadas, na maioria das universidades, a partir de 2012. Com relação aos anais de eventos, também abordados neste estudo, pôde-se perceber maior eficiência e eficácia na disponibilização dos documentos na *web* se comparado às BDTD, talvez decorrente pelos diferentes processos, burocratização ou mesmo pelo quantitativo de colaboradores disponíveis para execução das tarefas.

2.3. Análise de assunto

Constantes problemas de armazenamento de dados e recuperação da informação têm-se apresentado com a crescente produção de documentos e a necessidade de armazenamento de informações iniciada a partir do fenômeno conhecido como "explosão bibliográfica", na década de 1940. Com essa produção de documentos, surgem assuntos complexos em todas as áreas do conhecimento, inclusive nas interdisciplinares, que podem tornar um trabalho árduo para o profissional que desenvolve atividades com essas informações. Isso porque esse profissional necessita dominar técnicas de organização da informação de forma a deixá-las acessíveis aos usuários (NAVES, 1996).

A atividade realizada por um profissional da informação de identificar e descrever um documento de acordo com o assunto relacionado é denominada de indexação (NAVES, 1996). A indexação, enquanto ato de construir índice, é uma prática utilizada desde as bibliotecas da antiguidade no tratamento de documentos. Entretanto, a partir da necessidade da ordenação desses índices,

foi necessária uma alteração no processo mecânico da construção de índices, introduzindo, assim, a etapa de análise de conteúdo dos documentos (FUJITA, 2003).

A indexação enquanto processo de análise documentária ganhou enfoque a partir do aumento na produção dos documentos da comunicação científica, sendo necessária em centros de documentação especializados a criação de mecanismos de controle bibliográfico (ABNT, 1992; FUJITA, 2003). Nesse processo, a análise de assunto é considerada uma das etapas mais importantes realizada pelo indexador, responsável por extrair os conceitos do documento por meio de análise e traduzir para os termos de instrumentos de indexação, tais como lista de cabeçalhos, esquemas de classificação ou tesauros (NAVES, 1996).

A NBR 12676/1992 conceitua indexação como “ato de identificar e descrever o conteúdo de um documento com termos representativos dos seus assuntos e que constituem uma linguagem de indexação” (ABNT, 1992, p. 2). Além disso, instrui ao profissional indexador a utilizar uma abordagem sistemática para identificar os conceitos essenciais na descrição do assunto, tais como: a indagação sobre qual é o assunto principal do documento; a definição do assunto em termos de teorias, hipóteses, dentre outros campos que constituem um documento; a ação, operação ou o processo do assunto; a definição do agente da ação, operação e processo; a constatação de métodos, técnicas e instrumentos especiais; a consideração dos aspectos no contexto de um local ou ambiente especial; a identificação de variáveis dependentes ou independentes; e a consideração do assunto em relação a um ponto de vista (ABNT, 1992).

Entende-se como processos de indexação as etapas de compreensão total do texto, identificação e seleção de conceitos válidos para a indexação e a etapa de tradução, referente à representação dos conceitos por termos a ser utilizada em uma linguagem de indexação (FUJITA, 2003). Pode-se considerar também o estágio de sumarização presente na fase analítica, isso porque o estágio analítico pode ser dividido entre análise e sumarização, englobando a operação de síntese. A análise documentária compreende que a leitura do documento pode ser dividida nos estágios de compreensão do documento,

identificação dos conceitos que representam o conteúdo e a seleção dos conceitos válidos para a recuperação da informação (VICKERY, 1980).

Define-se análise documentária como um conjunto de procedimentos realizados para representar conteúdos de documentos com características distintas de forma que facilite na recuperação da informação. Trata-se de uma operação semântica que realiza a passagem de um documento original para um tipo de representação (DIAS; NAVES, 2007). O processo de extrair conceitos que representem a essência de um determinado documento é conhecido como análise de assuntos, entretanto, é possível encontrar na literatura expressões como análise de informação, análise temática, análise conceitual, análise de conteúdo e análise documentária (FUJITA, 2003; DIAS; NAVES, 2007).

A evolução de técnicas de tratamento da informação permitiu a expansão da análise documentária com o tratamento temático que comporta a geração de resumos e a indexação, possibilitando, de forma estratégica, o estreitamento entre o processo e a finalidade de indexação por meio de uma combinação de tratamento do conteúdo de documentos e a recuperação da informação pelo usuário (FUJITA, 2003).

Para Cesarino e Pinto (1980, p. 32) “A análise de assunto é a operação-base para todo o procedimento de recuperação de informações”. O princípio de indexação, publicado pela *Information System for Science and Technology* (1981, p. 84), é definido como "[...] o ato de descrever ou identificar um documento em termos de seu conteúdo”, podendo considerar dois pontos distintos, em que o primeiro está em realizar a descrição e identificação de documentos com auxílio de representações dos conceitos contidos num documento e o segundo possibilita busca e acesso à informação.

O tratamento da informação tem apresentado grande relevância no desenvolvimento de bibliotecas e em sistemas de recuperação da informação, abordando em sua natureza as atividades de tratamento descritivo, responsável pelas informações dos dados físicos extraídos do documento, e tratamento temático, que busca descrever o conteúdo do documento. A análise de assunto é o primeiro passo realizado no tratamento temático, sendo considerada como a etapa intelectual que visa a excelência no trabalho realizada pelo indexador (DIAS; NAVES, 2007).

Em diferentes situações, o profissional da informação pode ser requisitado para realizar a análise de assunto, podendo citar, por exemplo, a inserção de documentos num determinado sistema de informação, realizando, assim, a análise e seleção do conteúdo informativo e o entendimento do objetivo do sistema e da necessidade do usuário (CESARINO; PINTO, 1980; SOUSA; FUJITA, 2014). Outro cenário pode ocorrer num processo inverso. Nele o profissional da informação recebe um pedido de extração de informação e avalia a necessidade de informação do usuário com os conceitos existentes na solicitação, traduzindo para uma linguagem sistêmica. Ambas as situações possuem o objetivo de identificar a necessidade informacional do usuário (CESARINO; PINTO, 1980).

O termo análise de assunto envolve conhecimento nos documentos a serem indexados, bem como a determinação das características de sua relevância (LANGRIDGE, 1977). Além de ser considerada a fase inicial no processo de indexação, decidirá quais os principais tópicos de assunto de um documento serão utilizados. São considerados quatro estágios nesse processo, sendo eles: i) análise de assunto do documento; ii) expressão do conteúdo do assunto nas palavras dos indexadores, realizada por meio de linguagem natural; iii) tradução para um vocabulário de indexação; e iv) expressão do assunto em termos do índice (CHU; O'BRIEN, 1993). O termo análise de assunto também é considerado por autores, como a etapa de tradução dos conceitos extraídos dos documentos analisados para um vocabulário controlado ou mesmo para um processo de indexação em um todo (FUJITA, 2003).

Com a constante evolução das tecnologias, pesquisadores da área de tratamento da informação questionam sobre a necessidade de estudar os processos de análise de assunto, pois acreditam que se trata de um tema que se caminha para a obsolescência. Entretanto, faz-se necessário destacar que o capital intelectual do indexador durante o processo de identificação de assunto de um documento pode ser considerado o limite entre o homem e a máquina, pois possui características como abstração, percepção, compreensão e interpretação capazes de ser realizadas pelo ser humano. Remete-se aos computadores a tarefa de indexação automática de grandes *corpora* de documentos sem as expectativas de qualidade e precisão nos resultados, bem como no trabalho de recuperação das informações. Ao profissional indexador

cabe realizar o tratamento em coleções pequenas e especializadas, em que se pode adotar grande especificidade e esperar uma alta precisão na busca de informações (DIAS; NAVES, 2007).

Sob o ponto de vista do indexador, a análise de assunto é iniciada por meio da leitura do texto e, para isso, faz-se necessário que o profissional tenha conhecimento de estruturas de textos para que se realize uma leitura com fins específicos. A etapa seguinte está na extração dos conceitos que possam representar o conteúdo temático do texto, levando em consideração a qualidade da representação em que os termos são definidos em linguagem natural. Em sequência está a tradução para uma linguagem de indexação em que os termos passam a ser chamados de cabeçalhos ou descritores de assunto, enunciados, termos de indexação ou mesmo palavras-chave (DIAS; NAVES, 2007).

Sob o ponto de vista dos sistemas de informação, a indexação de análise de assuntos também é um processo de suma importância, porque condiciona os resultados numa estratégia de busca. O reflexo de desempenho dos resultados alcançados numa recuperação da informação, realizada por índices, está relacionada diretamente à qualidade da indexação de documentos. Considera-se na recuperação da informação a pertinência entre a questão de busca, cuja a indexação proporcionou a identificação de conceitos ao conteúdo, correlacionando com precisão o assunto pesquisado em índices (FUJITA, 2003).

A análise de assunto é uma atividade que envolve complexidades como os problemas de terminologias, intrinsecamente associadas à influência direta do profissional indexador. Mesmo se tratando de uma tarefa em que o profissional deve determinar de forma precisa o conteúdo do documento, pode ocorrer subjetividade por meio de seus próprios valores ao realizar a interpretação do conteúdo do documento (NAVES, 1996; 2000).

Após a identificação dos conceitos, o indexador passa a abordá-los de forma mais lógica, buscando selecionar os conceitos que melhor representam tais conteúdos contidos no documento. A partir da leitura do documento o indexador realiza um processo interativo de comunicação entre leitor, texto e contexto que podem estar sujeitos a diferentes condições. Entretanto, o indexador enquanto leitor, passa a ser a variável mais influente na interação de análise de assunto, porque necessita realizar a compreensão da leitura do documento mediante os seus aspectos cognitivos (FUJITA, 2003).

Trata-se de uma atividade com características subjetivas do indexador, levando em consideração o nível de conhecimento prévio pertinente ao documento analisado, bem como sua formação e experiência, além de fatores cognitivos, linguísticos e lógicos. A compreensão de textos e a identificação dos conceitos perpassam por processos intelectuais complexos que envolvem a memória e estruturas cognitivas do indexador. Considera-se uma atividade intelectual em uma representação condessada, associada à identificação das palavras e à compreensão do léxico utilizado (DIAS; NAVES, 2007).

O fato é que não se trata de uma tarefa fácil, mas caberá ao indexador analisar o documento e tomar a decisão de identificar o assunto que melhor represente o texto. São dois os principais instrumentos necessários para o tratamento da informação: o tratamento descritivo e o tratamento temático. O primeiro se desenvolve em aspectos mais objetivos, capazes de identificar num documento elementos como autor, título, editora e elementos similares, além de utilizar códigos de catalogação e formatos de metadados. Já o segundo tratamento possui subjetividade, caracterizando o documento do ponto de vista do conteúdo, além de ser representado pelas linguagens de indexação (DIAS; NAVES, 2007).

Ao realizar a análise de assunto, faz-se necessário fazer uma leitura que possibilite a extração dos conceitos e posterior tradução em termos para indexação. Uma questão indagada por Foskett (1973) está em: “Como podemos determinar o assunto de um documento de modo a especificá-lo?”. O ideal seria que o indexador realizasse a leitura completa do documento, entretanto, esse processo não apresenta garantias de entendimento do conteúdo, podendo interferir na qualidade do resultado e levar tempo para realização da tarefa. O autor sugere encurtar caminhos por meio de leituras em comentário em orelha, prefácio, sumário, introdução e resumo. Além disso, ressalta que muitas vezes o título de um trabalho é escolhido para chamar atenção e não para indicar o assunto (FOSKETT, 1973).

Essa técnica de leitura tornou-se uma estratégia clássica para análise de assunto e denomina-se leitura técnica. Ela consiste numa leitura direcionada para partes específicas do documento, possibilitando encontrar informações importantes para identificação do assunto contido no documento. As seções de título, subtítulo, prefácio, apresentação, sumário, títulos dos capítulos, resumo,

introdução, frases e parágrafos que abrem os capítulos, conclusões, ilustrações, gráficos, tabelas, legendas e referências são indicadas em manuais de catalogação e indexação. Recomenda-se “um misto de ler e ‘passar os olhos’ pelo texto” (LANCASTER, 2004, p. 24).

Na área da Ciência da Informação, identificar e acessar informações são tarefas essenciais, a fim de identificar maior complexidade por se tratar de uma etapa preliminar para o acesso efetivo à informação. Além disso, faz-se necessário destacar a compreensão mediante os aspectos cognitivos do profissional, passando a ser considerado o acesso à informação uma questão de interesse básico da área. Fator determinante para a seleção de assunto ou informação está nas políticas de indexação ao qual o profissional está inserido, cabendo à instituição, de acordo com os seus objetivos institucionais, decidir, com base em seu perfil do usuário, se o assunto extraído do documento será entregue com maior especificidade ou com características generalistas (DIAS; NAVES, 2007).

Diferentes concepções de análise de assunto são adotadas em sistemas de informação e podem afetar o desempenho do indexador enquanto leitor, uma vez que os objetivos institucionais estão intrinsecamente associados aos processos. A concepção simplista considera os assuntos como entidades objetivas absolutas, derivando-se de uma abstração linguística ou métodos estatísticos de indexação, além de uma indexação totalmente automatizada. A concepção orientada para o conteúdo é facilmente percebida pelo indexador humano. Isso envolve a abstração indireta do documento por meio da identificação de tópicos de assuntos inseridos em uma estrutura superficial do documento, o que ultrapassa os limites da estrutura superficial léxica e gramatical. Na concepção orientada por demanda, o indexador analisa o documento ao deixar as informações visíveis para o usuário, de acordo com as suas necessidades. Com isso, desconsidera representar ou resumir informações de forma explícita ou implícita, funcionando como instrumentos de medição e transmissão do conhecimento para o usuário interessado (ALBRECHTSEN, 1993).

Ambas as concepções são importantes para a análise de assunto, sendo as concepções “orientada para o conteúdo” e “orientada por demanda” complementares. Entretanto, a concepção orientada por demanda pode ser

considerada um etapa posterior à análise de assunto, uma vez que os objetivos são traduzir os conceitos extraídos do documento para termos de indexação, buscando atender aos interesses dos usuários (NAVES, 2000).

Um mecanismo composto de perguntas e respostas generalistas constituem o processo de indexação na etapa de identificação de assuntos realizados em documentos na metodologia de Tálamo (1987). Questões como Quem? (ser), O quê? (tema), Como? (modo), Onde? (lugar) e Quando? (tempo) permitem identificar a estrutura temática que contempla o objetivo principal do texto (TÁLAMO, 1987). A composição desse conjunto de perguntas e respostas formulará o tema do documento analisado. O tema possui uma estrutura temática que pode ser constituída por conceitos, categorias ou facetas, permitindo a sua identificação por meio da análise de assunto (FUJITA, 2003).

O sistema de recuperação da informação busca otimizar o acesso ao conteúdo de quaisquer tipos de informações, independentemente de sua localização. No caso das bibliotecas, as informações estão organizadas num sistema de catálogo e, nesse caso, o termo otimizar remete à facilidade de acesso à informação (DIAS; NAVES, 2007). O tratamento da informação é um dos subsistemas responsáveis por descrever os aspectos físicos do documento e podem ser representados como:

- Catalogação e classificação – terminologia utilizada em bibliotecas tradicionais para descrever o tipo de trabalho. A catalogação cria representações dos documentos por meio de fichas catalográficas que descrevem os aspectos físicos, objetivos do documento quanto aos aspectos de conteúdo. A classificação busca identificar o conteúdo do documento, porém, com o objetivo de determinar um lugar para ele numa coleção organizada por assuntos e fazer uso de um sistema que represente os assuntos por meio de classificação bibliográfica (DIAS; NAVES, 2007);
- Indexação – serviço designado para organizar informações de documentos, tendo como principais produtos os índices e *abstracts*, que poderão estar disponíveis em formato impresso ou digital (DIAS; NAVES, 2007);
- Metadados – utilizado em bibliotecas digitais, tem o objetivo de designar a descrição física de recursos eletrônicos, bem como suas respectivas

normalizações. Trata-se de uma função semelhante a de catalogação nas bibliotecas digitais (DIAS; NAVES, 2007). Também são utilizados em cadastros tradicionais para organização da informação em arquivos, bibliotecas e museus. Metadados mantém conexões evolutivas com metodologias de tratamento de informações para representação de características de objetos digitais como *Resource Description and Access* (RDA), *Resource Description Framework* (RDF), RDF *Schema* e *Ontology Web Language* (OWL) (LEMOS; SOUZA. 2018).

- Ontologias – termo utilizado no contexto digital para designar o trabalho de organização de recursos e conteúdos eletrônicos, além da recuperação dessas informações, sendo considerado semelhante ao trabalho de classificação, indexação e catalogação de assunto (DIAS; NAVES, 2007). Trata-se de um assuno abordado em diversas áreas do conhecimento como filosifia, ciência da computação e informação e em domínios como, biologia, direito, engenharia, geografia e medicina (ALMEIDA, 2013).

A análise de assunto é a etapa mais importante para o desenvolvimento da atividade do indexador. Tem como objetivo identificar e selecionar os conceitos que melhor representam a essência de um determinado documento. A identificação dos conceitos é uma tarefa árdua e envolve complexidade, justamente por exigir do indexador o uso de metodologias e abordagens sistematizadas que devem ser adequadas para cada situação, a fim de obter uma recuperação da informação de qualidade (FUJITA, 2003).

Cabe ao profissional indexador conhecer os processos, técnicas, métodos e processos relativos ao tratamento da informação, tais como: descrição física e temática dos documentos, seja numa biblioteca tradicional ou em um sistema de recuperação de informação; desenvolvimento de instrumentos como códigos, linguagens, normas ou padrões a serem utilizados na descrição; e implantação de estruturas físicas ou bases de dados para armazenamentos de documentos como fichas ou registros eletrônicos (DIAS; NAVES, 2007).

Aboutness é o termo originário da língua inglesa utilizado por Fairthorne (1969) que, em uma tradução não literal, pode significar em português: “do que se trata o texto”. O termo não apresenta uma consolidação do seu conceito ao ser traduzido para a língua portuguesa. Pesquisadores se divergem entre

temacidade, substantivo próximo ao termo temático e atinência empregado na tradução do livro “Indexação e resumos: teoria e prática” de Lancaster e utilizado por Dias e Naves (2007). A relevância da temacidade se refere à problemática envolvida na identificação do tema em um determinado documento. Temacidade enquanto objetivo principal permite realizar a análise de assunto ao identificar temas ou assuntos por meio de análise conceitual, que possui como característica a identificação e seleção de conceitos (FUJITA, 2003).

O termo *aboutness* passou a ser utilizado na literatura como sinônimo de assunto (TODD, 1992) e ser pesquisado em substituição ao conceito de *subject* (ALBRECHTSEN, 1993). Begthol (1986) fez distinção entre *aboutness* – entender o conteúdo permanente no documento – e *meanings* – o significado compreendido pelo usuário. Lancaster (2003), por sua vez, não buscou entrar numa discussão filosófica sobre o significado de antinência por não apresentar um direcionamento pragmático para a definição do termo enquanto melhoria no processo de indexação, entretanto, discute sobre *aboutness* em obras textuais, obras de arte e o termo *of-ness.* Para o autor, atinência é um tema próximo ao conceito de relevância “[...] a relação entre um documento e uma necessidade de informação ou entre um documento e um enunciado de necessidade de informação (uma consulta)” (LANCASTER, 2003, p. 14).

Lancaster (2003) questiona a necessidade de compreensão teórica do termo atinência para realizar uma indexação e assuntos. Fujita (2003) corrobora ao considerar que, para se entender o conteúdo de um documento de forma que seja compreendido e representado, é necessário entender o processo cognitivo do profissional especialista em análise de assunto, levando em consideração a capacidade de processar, compreender, sintetizar e representar a informação.

O capítulo seguinte apresenta o referencial empírico da pesquisa. Ele se inicia a partir dos conceitos e características de linguística computacional, linguística de *corpus* até o processamento de linguagem natural. A segunda seção do capítulo destaca a modelagem de tópicos e os modelos *Latent Semantic Indexing* e *Latent Dirichlet Allocation*, utilizados por meio de algoritmos de *Machine Learning* para identificação de tópicos, termos, pesos e frequências de termos extraídos dos *corpora* de dados.

3. REFERENCIAL EMPÍRICO

A primeira seção inicia com a apresentação de conceitos fundamentais responsáveis por nortear a metodologia da pesquisa. Assuntos como linguística computacional, linguista de *corpus* e *corpus* envolvidos no processamento de linguagem natural são discutidos por meio de contribuições de autores como Othero (2006), Sardinha (2000; 2004), Vieira e Lima (2001; 2003), Granger (1998), Pustejovsky e Stubbs (2012), Grus (2016) e Marquesone (2016).

A segunda seção apresenta os conceitos sobre Modelagem de Tópicos e os padrões *Latent Semantic Indexing* e *Latent Dirichlet Allocation*, utilizados no desenvolvimento prático por meio de algoritmos de *Machine Learning* aplicados aos *corpora* de dados para o alcance dos resultados desta pesquisa. Posteriormente, são apresentados e debatidos no capítulo de resultados e discussão com o núcleo de disciplinas da Ciência da Informação, exposto no referencial teórico. Os temas abordados no capítulo são embasados em autores como Berry, Dumais e O'brien (1995), Hofmann (1999), Ayodele (2010), Blei (2012), Aggarwal e Zhai (2012), Steyvers e Griffiths (2015), Santos (2015) e Kaszubowski (2016).

3.1. Linguística computacional

Refere-se à área de estudos, teórica e aplicada, que atua na investigação do tratamento computacional da linguagem e das linguagens naturais. Destaca-se na linguística as áreas como a Sintaxe, Semântica, Análise do Discurso, Fonética e Fonologia, utilizadas para compreender e produzir as linguagens naturais e dominar o conhecimento linguístico envolvido no domínio de uma linguagem natural (OTHERO, 2006). Trata-se de uma área do conhecimento que investiga as relações entre a linguística e a informática de forma que seja possível reconhecer e produzir informações em sistemas computacionais apresentadas por meio da linguagem natural (VIEIRA; LIMA, 2001).

A linguística computacional divide-se em duas subáreas, sendo elas: 1 - Linguística de *Corpus*; e 2 – Processamento de Linguagem Natural (PLN) (OTHERO, 2006). A primeira realiza estudos a partir de *corpora* de dados eletrônicos que contemplam amostras de linguagem natural. Buscam estudar fenômenos linguísticos e sua ocorrência em grandes volumes de dados de um

determinado idioma, podendo haver uma variedade, modalidade ou dialeto da língua. Além disso, trata-se de uma área de estudos que nem sempre possui como objetivo desenvolver sistemas computacionais (SARDINHA 2000b; 2004; OTHERO, 2006). A segunda se refere ao estudo da linguagem voltado para a construção de sistemas especialistas capazes de interpretar e/ou gerar informações fornecidas em linguagem natural. Para esse processo, vários subsistemas são necessários para poder diferenciar os aspectos da língua natural, bem como sons, sentenças, palavras, discursos nos níveis estruturais. Tradutores e reconhecimento automático de voz, *parsers*, *chatterbots*, geradores automáticos de resumos entre outras aplicações são exemplos de PLN (SARDINHA, 2000b; OTHERO, 2006).

3.1.1. Linguística de *corpus*

A linguística de *corpus* busca coletar de forma criteriosa e explorar um conjunto de dados linguísticos textuais no qual constituem um *corpus* de dados de forma que possam ser utilizadas para pesquisa de língua ou variações linguísticas. As amostras de um *corpus* podem ser formadas por diferentes tipos e fontes, como por exemplo, *corpora* de linguagem falada ou linguagem escrita, que contemplam textos de jornais em diversos formatos. Realiza-se a exploração da linguagem por meio de evidências empíricas extraídas por meio de aplicações computacionais (SARDINHA, 2000a; 2000b).

A linguística de *corpus* desenvolve suas atividades por meio de textos reais que ocorrem naturalmente em uma determinada linguagem natural. Normalmente, estão fora do contexto do analista onde são oferecidos, geralmente, apenas o co-texto (HUNSTON, 2002). *Corpora* de documentos podem possuir informações ou classificações complementares como período de produção dos textos, autores ou até participantes em interações que possibilita ao analista recuperar, mesmo que parcialmente, contextos situacional ou cultural ao qual os textos foram desenvolvidos. Entretanto, tal possibilidade não tem se apresentado como um problema teórico relevante para a área estudada, uma vez que o foco de estudo tem-se voltado mais para identificação de padrões do que para usos particulares da língua em determinadas situações (OLIVEIRA, 2009).

Pesquisas realizadas na área de linguística de *corpus* envolvem os processos de coleta, compilação e organização de repositórios que contenham trechos da linguagem escrita literária, falada, por textos de jornais e até por falas de crianças em desenvolvimento linguístico geradas de forma natural e espontânea (OTHERO, 2006; VIEIRA; LIMA, 2001). Todo esse processo tornou-se viável a partir da década de 1960, com o suporte dos computadores (VIEIRA; LIMA, 2001). Após o surgimento dos computadores, a linguística de *corpus* tem basicamente realizado pesquisas a partir de *corpora* de dados eletrônicos que contenham diferente tipos de amostras em diferentes tipos de fontes (OTHERO, 2006) ,possibilitando armazenar, acessar e analisar grandes volumes de dados linguísticos (SVARTVIK, 1996).

A *corpora* surgiu bem antes da era do computador e, em seu sentido original, *corpus* é corpo e refere-se a um conjunto de documentos. Já na Grécia antiga, Alexandre o Grande definiu o *Corpus* Helenístico e na antiguidade e idade média foram produzidos *corpora* de citações da bíblia. Durante o século XX, existiram muitas pesquisas relacionadas à descrição da linguagem por meio de *corpora*. Com o surgimento dos computadores, houve uma mudança de paradigma e um condicionamento das tecnologias com a linguística de *corpus*, que, por sua vez, permitiu realizar o armazenamento e exploração de *corpora* (SARDINHA, 2000a). Duas diferenças são destacadas entre a era antes e pós tecnológica: na primeira, os dados que eram coletados, mantidos e analisados manualmente passaram a ser utilizados por meio de computadores; na segunda, o foco – que em geral estava no ensino de línguas – passou a ser a literatura e a descrição da linguagem (GRANGER, 1998).

O *corpus* não computadorizado *Survey of English Usage* (SEU), construído a partir de 1953, foi projetado para ter o tamanho de 1 milhão de palavras com um número fixo de 200 textos que contemplava a quantidade igual de 5000 palavras para cada texto, organizados em fichas de papel contendo uma palavra do *corpus* inserida em 17 linhas de texto. Esse *corpus* acabou por servir como base de referencial para outros *corpora,* como o *Brown Corpus*. Cada ficha recebeu uma categoria gramatical com base nas palavras analisadas gramaticalmente. Por meio da análise gramatical das palavras, cada ficha recebeu uma categoria gramatical. Isso possibilitou a base para o desenvolvimento dos etiquetadores computadorizados e que realizam

automaticamente a identificação dos traços gramaticais. O processo de transformação realizado do *corpus* não computadorizado para o meio eletrônico ocorreu em 1989, entretanto, a parte falada, conhecida como *London-Lund Corpus*, foi computadorizada anteriormente (SARDINHA, 2000a; 2000b).

Três momentos são destacados nos estudos da linguística baseada em *corpus*: i) com o surgimento dos computadores *mainframes* nas universidades na década de 1960, houve um melhor aproveitamento para pesquisas realizadas em linguagem; ii) com a entrada dos computadores junto às universidades, houve um maior volume de pesquisas relacionado em Processamento da Linguagem Natural (PLN), juntamente com a criação e manutenção de *corpora* em maior número e escala; e iii) com a popularização dos microcomputadores na década de 1980, os *corpora* e as ferramentas de processamento ficaram em evidência, o que fortaleceu as pesquisas baseadas em linguística de *corpus*. Essa área está intimamente relacionada ao *corpus* eletrônico. No Quadro 07 são apresentados os principais *corpora* compilados e ainda em compilação (SARDINHA, 2000a; 2000b).

Quadro 07 – Histórico da linguística de *Corpus*

Corpus	Lançamento / Referência na literatura	Palavras	Composição
Brown Corpus (Brown University Standard Corpus of Present-Day American English)	1964	1 milhão	Inglês americano, escrito
American Heritage Intermediate Corpus – AHI	1971	5 milhões	Inglês americano, escrito
Lancaster-Oslo-Bergen – LOB	1978	1 milhão	Inglês americano, escrito
London-Lund Corpus – LLC	1980	500 mil	Inglês americano, escrito
Birmingham Corpus (Birmingham University International Language Database)	1987	20 milhões	Inglês britânico
Kolhapur Corpus (of Indian English)	1988	1 milhão	Inglês indiano, escrito
Tools for Syntactic Corpus Analysis – TOSCA	1988	1,5 milhão	Inglês britânico, escrito
Survey of English Usage – SEU	1989	1 milhão	Inglês britânico, escrito e falado
Child Language Data Exchange – CHILDES	1990	20 milhões	Inglês infantil, falado
Nijmengen Corpus	1991	132 mil	Inglês britânico, escrito e falado
Longman-Lancaster English Language Corpus – LLELC	1991 (previsão)	50 milhões (previsão)	Inglês de vários tipos, escrito e falado
Map Task Corpus	1991	147 mil	Inglês escocês, falado

Corpus os Spoken American English	1991	2 milhões	Inglês americano, falado
Longman Corpus of Learner's English – LCLE	1992	10 milhões	Inglês escrito por estrangeiros
Lancaster/IBM Spoken English Corpus – SEC	1992	53 mil	Inglês britânico, falado
Wellington Corpus (of Written New Zealand English)	1993	1 milhão	Inglês neozelandês, escrito
Polytechnic of Wales Corpus – POW	1993	65 mil	Inglês infantil, falado
Wellington Corpus of Spoken New Zealand English	1995	1 milhão	Inglês neozelandês, falado
British National Corpus – BNC	1995	100 milhões	Inglês britânico, escrito e falado
International Corpus of Learner English – ICLE	1997	200 mil (cada variedade nacional)	Inglês escrito por estrangeiros
Bank of English	1997	320 milhões	Inglês britânico

Fonte: (SARDINHA, 2000a, 2000b).

Destacam-se junto ao Quadro 07 os *corpora* considerados marco de referência histórica. O *corpus Brown*, coletado e compilado por Henry Kucera e Winthrop Nelson Francis, da Universidade de Brown, em 1961, foi o pioneiro de uma ampla gama de *corpus* do inglês americano contemporâneo e *British National Corpus*, sendo o primeiro a conter o quantitativo de 100 milhões de palavras e o único disponível para compra dentro da comunidade europeia, contendo textos de uma ampla gama de gêneros, domínios e mídias. *O corpus Brown* foi lançado em uma época no qual a coleta de registros linguísticos não era bem vista, uma vez que se gastava muito tempo e recurso para executar os procedimentos (SARDINHA, 2000a; 2000b; PUSTEJOVSKY; STUBBS, 2012).

Outro *corpus* destacado é o *Birmingham* que, posteriormente, se tornaria *Bank of English*: o primeiro a ultrapassar o quantitativo de 1 milhão de palavras iniciados por *Brown*. Além disso, está em constante crescimento, com acesso restrito aos pesquisadores ligados ao COBUILD e à editora Collins (SARDINHA, 2000a; 2000b).

No Brasil a linguística de *corpus* tem se desenvolvido em centros voltados ao processamento de linguagem natural, lexicografia e linguística computacional (SARDINHA, 1999). Os espaços acadêmicos têm consolidado parcerias entre universidades e empresas privadas com interesses em aplicações baseadas em *corpora* de dados para informatização de bases de dados, processamento automático de textos, reconhecimento de voz e gerenciamento de informações (SARDINHA, 2000a). Estudos na área ganharam força com a publicação, em

2004, do primeiro livro brasileiro, intitulado “Linguística de *corpus*”, que aborda características e metodologias para análise de *corpora* (SARDINHA, 2004). A fim de buscar maior compreensão da linguística de *corpus*, pesquisadores e gramáticos interessados procuram de forma sistemática, por meio de contribuições teóricas, novos conhecimentos da linguagem para a descrição do português do Brasil (AZEREDO, 2008).

Teses e dissertações com estudos voltados para o português do Brasil têm surgido após a inserção da disciplina linguística de *corpus* em programas de pós-graduação *stricto sensu*. Os *corpora* de dados podem ser classificados como: i) gerais, que buscam representar a língua de forma ampla e servir de base para diferentes pesquisas e incluem a variedade de registros, assuntos e atores; e ii) especializados, coletados para objetivos específicos de pesquisas e constituídos por documentos de gêneros específicos (OLIVEIRA, 2009).

A expansão de estudos na área da linguística de *corpus* pode apresentar problemas por designar uma empreitada coletiva, já que volumes de pesquisas independentes apresentam, de maneira não sistematicamente organizadas, diferentes aspectos das línguas (KENNEDY, 1998). Considera-se a formação de um *corpus* representativo de conhecimentos gramaticais em diferentes línguas quando organizadas as pesquisas independentes (OLIVEIRA. 2009).

O *corpus* é um artefato produzido para fins específicos de pesquisa. Formado por uma coleção de textos produzidos em um ambiente natural de comunicação. Seu conteúdo é definido por fenômenos linguísticos e representa uma variedade de linguagem que devem ser legíveis por máquinas (PUSTEJOVSKY; STUBBS, 2012). São considerados como autênticos os textos produzidos pelo homem em ambientes naturais e que não foram criados com o propósito de comporem um *corpus*, desconsiderando a produção de textos gerados por *software* (SARDINHA, 2000a).

Nem toda coleção de documentos constituída por dados linguísticos naturais e legíveis por computador pode ser considerava um *corpus*. Outras terminologias podem se diferenciar como: i) arquivo: depósito de texto sem organização prévia; ii) biblioteca eletrônica: coleção de documentos com critérios de seleção; iii) *corpus*: parte de uma biblioteca eletrônica construída a partir de um desenho formalizado e com objetivos específicos; e iv) sub*corpus*: parte de um *corpus* que pode ser fixa ou flexível durante a análise (SARDINHA, 2000a).

O *corpus* é uma amostra de uma determinada população ao qual não se conhece a sua real dimensão. Dessa forma, acaba por representar uma fração limitada da língua. Esta característica não desclassifica um *corpus* de dados, já que representa um potencial de significados (HALLDAY, 1994; SARDINHA, 2004). Um *corpus* enquanto fragmento incompleto da língua, por não possuir uma medida da proporção de uso de textos, seja escrito ou falado, pode ser considerado um sistema global ou parte de um sistema, já que permite refletir possibilidades de ocorrências de usos linguísticos (OLIVEIRA; DIAS, 2009).

Acessar, analisar ou contrastar dados em *corpora* estão entre as características gerais em estudos de *corpus* de dados, dependendo da abordagem metodológicas de pesquisa utilizada. De acordo com os objetivos e escopo da pesquisa, pode-se desenvolver estudos em *corpus* por análise multidimensional, prosódia semântica, fraseologia, colocações e cálculo de frequência de termos (BIBER, 1988; CONRAD; BIBER, 2001).

Oliveira (2009) considera as seguintes características baseadas em estudos em *corpora* de dados: i) investigação de uma linguagem natural; ii) coleções de textos selecionados por meio de critérios; iii) análise automática ou interativa por meio de computadores; iv) análises qualitativa e quantitativa; v) análise de textos longos e diversificados; vi) uso do mesmo *corpus* para alcançar novos resultados; vii) apresentação de novas contribuições para linguísticas teóricos e aplicados; e viii) maior precisão e credibilidade das análises quantitativas (OLIVEIRA, 2009).

Os pesquisadores especializados em reconhecimento da fala começaram a utilizar os *corpora* de dados a partir da década de 1980 para modelar os algoritmos de treinamento e fenômenos de linguagem. Na ocasião, criaram um modelo de linguagem que copilou os dados da linguagem falada e funcionava para reconhecimento de um vocabulário limitado de palavras. Esse modelo foi construído por meio dos modelos N-*grams* e *Hidden Markov Model* (HMM) - Modelo oculto de Markov. A década seguinte foi marcada pelo aumento representativo no volume de dados na tradução automática, correspondendo a um aumento da modelagem de linguagem estatística para a tradução. Com a evolução dos *hardwares* de computadores, foi possível coletar e analisar *corpora* de dados com números representativos de fragmentos de linguagem se comparado com os anteriores. Dessa forma, os modelos de linguagem

estatísticas têm apresentado resultados cada vez mais precisos quando aplicados em diferentes situações de linguagem natural (PUSTEJOVSKY; STUBBS, 2012).

Os N-gramas, encontrados em *corpora* de dados, refere-se a um conjunto de itens sequenciais como palavras, letras e fonemas no qual torna-se possível realizar previsões de uma sequência ou aprender o uso de uma linguagem ao se examinar com que frequência determinados itens ocorrem juntos. São frequentemente utilizadas em diferentes situações, como por exemplo, em sites que oferecem sugestões de novas pesquisas quando existente algum erro de digitação realizado pelos usuários ou usada em desambiguação da linguagem falada por um usuário em um sistema, podendo ajudar a reconhecer o problema e encontrar a palavra pretendida por meio de um modelo *N-grama* (PUSTEJOVSKY; STUBBS, 2012).

Quando utilizado no contexto de PLN, encontram-se muitas aplicações em processamento de sinais ou processamento da fala, sendo o HMM considerada uma poderosa ferramenta de estatística para modelar um amplo intervalo de dados de séries temporais, sobretudo quando aplicada em problemas como marcação de partes da fala e frase-substantivo. Torna-se necessário destacar que os HMMs também obtiveram sucesso quando aplicados em PLN de baixo nível, como extração de informação de documentos, por exemplo (BLUNSOM, 2004).

3.1.2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) estuda os problemas de geração e compreensão automática da linguagem humana natural, por meio de sons, palavras, sentenças e discursos que possam ser interpretados computacionalmente através de algoritmos e que levam em consideração formatos e referências, estruturas e significados, contextos e usos por níveis que possam ser fonético, fonológico, sintático, semântico e pragmático (GONZALEZ; LIMA, 2003).

O PLN vem sendo estudado desde o advento dos computadores e em seus primeiros sistemas em que se discutia o interesse em *softwares* que fossem capazes de interpretar os objetivos de seus usuários por meio de sua linguagem. Um dos maiores teóricos da computação, Allan Turing, enfatizava a inteligência

dos computadores com a capacidade de lidar com a linguagem natural dos humanos. Turing enxergou, por meio da máquina de Von Neumann, idealizada para aplicações numéricas, que os computadores eram considerados recursos com capacidades inteligentes para apoiarem atividades como habilidades de compreender e produzir linguagem natural (VIEIRA; LIMA, 2001).

Ainda sobre os sistemas que utilizam PLN, trata-se de técnicas computacionais que envolvem linguagens. Trata-se de uma área ampla com técnicas que podem variar das mais simples até as mais complexas. Estão entre alguns exemplos de PLN: nuvens de palvras, modelos de n-gramas, gramáticas, amostragem de Gibbs, centralidade, centralidade de intermediações, centralidade de vetor próprio, gráficos direcionados *PageRank* e modelagem de tópicos (GRUS, 2016). Destacam-se a preocupação com a avaliação dos sistemas desenvolvidos e a construção de sistemas com capacidade de processamento de linguagem em larga escala. Com os avanços tecnológicos e do PLN foi possível realizar, por exemplo, auxílios aos editores de textos durante a edição de documentos eletrônicos, fornecendo correções gramaticais e de ortografias. Pelo fato de ser um trabalho contínuo e que busca melhorias, o sistema é capaz de promover mais satisfação do que frustração ao usuário. Outro exemplo está nos sistemas ditados, baseado no desenvolvimento de interfaces da fala, em que a aplicação é capaz de interpretar os comandos ou mesmo digitar um texto falado pelo usuário. Além disso, os sintetizadores de fala são capazes de realizar leituras de textos escritos (VIEIRA; LIMA, 2001).

O PLN trata computacionalmente os diversos aspectos da comunicação humana e tem como objetivo realizar análise e extração de significados de conteúdo, seja por meio de sons, palavras, sentenças e discursos – em que as entradas de dados são no formato textual por meio da análise da linguagem escrita que buscam entender o significado de cada palavra e suas estruturas dentro de um contexto – ou por voz – como por exemplo, a partir da análise de uma série de vídeos em um determinado idioma de forma que possa gerar automaticamente a legenda do áudio em outro idioma, considerando formatos e referências, estruturas e significados, contextos e usos (GONZALEZ; LIMA, 2003; MARQUESONE, 2016). Trata-se de um campo da engenharia e da ciência da computação que se desenvolveu a partir do estudo linguístico computacional na área da inteligência artificial e seu objetivo está em estreitar a comunicação

humana com máquinas e outros dispositivos tecnológicos por meio da linguagem natural (PUSTEJOVSKY; STUBBS, 2012).

A área de PLN realiza estudos da linguagem para a construção de sistemas computacionais específicos capazes de interpretar e gerar informações em linguagem natural (OTHERO, 2006). Para isso, fazem-se necessário vários subsistemas utilizados para interpretar os diferentes aspectos da língua, seja falada ou escrita, como sons, palavras e sentenças, além dos discursos nos níveis estruturais (VIEIRA, 2002). São exemplos de aplicações e serviços que utilizam PLN: reconhecimento de voz, tradução de línguas, detecção de plágio, classificação de um texto em categorias, extração de dados e informações a partir de textos, geradores automáticos de resumos, sistemas de atendimento de perguntas, tradução automática, reconhecimento da fala e classificação de documentos (OTHERO, 2006; PUSTEJOVSKY; STUBBS, 2012; (MARQUESONE, 2016).

Destacam-se as áreas de desenvolvimento da linguística computacional e, consequentemente, no PLN, que estão em constante desenvolvimento e atualizações (VIEIRA; LIMA, 2001):

- **Reconhecimento e sintetizadores**: sistemas de reconhecimento da fala são utilizados por meio de uma interface de comando de voz, seja para comandar um editor de texto por meio de ditado, em que o sistema faz a transcrição da fala por meio da linguagem natural para o texto para navegar na internet por comando de voz ou para serviços automatizados de informação por telefone. Sistemas sintetizadores de fala são utilizados para realizar a leitura de textos digitais escritos, interfaces adaptadas para deficientes visuais ou em serviços automatizados de informação por telefone;
- **Corretores ortográficos e gramaticais**: os corretores ortográficos e gramaticais dos editores de textos verificam a origem do vocabulário de cada palavra digitada e posteriormente suas construções gramaticais, tais como regras de concordância, léxico, gramática, pontuação, prefixos, acentuação. Além disso, comtemplam o módulo mecânico, especializado em erros de fácil detecção, como início da frase com letra maiúscula, ausência de pontuação no final da sentença, palavras e pontuações repetidas, espaçamentos

inadequados entre palavras e símbolos, capitalização inadequada, uso não balanceado de parênteses, aspas e chaves;

- **Tradutores automáticos**: os sistemas de tradução automáticos de idiomas possuem distribuições do tipo proprietário e *open source*, entretanto, ainda são considerados preliminares, pois não realizam uma tradução refinada, com imperfeições nos resultados. Algumas metodologias são destacas, dentre elas: i) sistemas diretos - buscam palavras correspondentes diretas; ii) sistemas de transferência - efetuam a análise sintática da frase da linguagem natural de origem e constroem a representação sintática da linguagem natural desejada; e iii) sistemas interlinguais - desenvolvem a representação intermediária entre a linguagem natural de origem e a desejada;
- **Geradores de textos e resumos**: os geradores de textos recebem os elementos de conteúdo e os objetivos de comunicação para produção de um texto linguisticamente correto. Faz-se necessário um processo de planejamento envolvendo a geração do discurso para determinar o que será dito e de que forma será dito. Já os geradores de resumos devem proporcionar o maior número de informações em um mínimo de espaço, envolvendo o estudo da linguagem para que a informação seja concisa. São úteis para o processo de busca pela informação e podem ser utilizados para serem categorizados por uma pessoa de acordo com a relevância de um documento. Os geradores de textos e resumos, associados às técnicas estatísticas, possibilitam a identificação dos modos como as palavras são utilizadas para análise em *corpora* de dados;
- **Interfaces em linguagem natural**: sistemas baseados na linguagem escrita ou falada são utilizados em sistemas de perguntas e respostas de um domínio de aplicação específico e limitado. A interação pode ser delimitada por palavras-chave em um sistema de processamento de linguagem que serve como interface de conexão entre a linguagem natural e a base de dados;
- **Recuperação da informação**: sistemas baseados na Recuperação da Informação são definidos como um conjunto de técnicas constituídas por indexação, busca, filtragem, organização, tratamento

de línguas e múltiplas mídias no qual possuem como objetivos encontrar documentos relevantes de acordo com a necessidade de informação. Duas abordagens distintas são destacadas: i) busca por metadados, referente ao cabeçalho ou palavras-chave que descrevem o conteúdo de um documento e podem ser adicionados de forma anual ou automaticamente; e ii) busca por conteúdo que, de maneira geral, atinge melhores resultados. Faz-se necessário destacar que as duas abordagens não remetem à compreensão automática de textos – área de baixa efetividade em domínios irrestritos –, entretanto, são baseadas em técnicas estatísticas que aferem a similaridade de textos e da consulta;

- **Extração de Informação**: trata-se de uma área de interesse da pesquisa computacional e que se refere à leitura, coleta e armazenamento de um texto não estruturado em um banco de dados por meio sistema de extração de informação. A extração da informação tem motivado os pesquisadores da linguística computacional a migrarem de sistemas de pequena escala e dados artificiais para sistemas de larga escala e dados linguísticos reais, a fim de realizar uma comparação na execução de tarefas entre o desempenho de sistemas de extração da informação com a performance humana. Esses sistemas são responsáveis por analisar e transformar a forma como as informações contidas em um conjunto de documentos são apresentadas;
- **Avaliação de sistemas de processamento de linguagem natural**: as avaliações de sistemas de PLN são frequentemente discutidas. Dentre elas destacam-se os sistemas de recuperação e de extração da informação. Os sistemas de recuperação são constantemente avaliados em termos de alcance – afere o número de documentos relevantes encontrados por meio de uma consulta entre o conjunto total de documentos –, e precisão – mede o número de documentos que realmente são relevantes entre os documentos indicados pelo sistema. Com relação aos sistemas de extração de informação, diferentes aplicações utilizam critérios próprios, como linguístico, operacional ou econômico para se realizar as avaliações. Uma técnica

importante para o processo de avaliação de sistemas está nos *corpora* anotados, em que se constitui em realizar uma avaliação de acordo com o *corpus* anotado nos níveis morfológico, sintático ou semântico, minimizando informações linguísticas como subjetividade. Algumas medidas podem ser consideradas quando houver diferentes sujeitos realizando anotações em um *corpus*, como um sistema pode ser avaliado com uma anotação derivada de várias anotações ou aferir o desempenho por meio do grau de concordância de conceitos entre sistema e anotação manual.

- **Processamento de linguagem natural e sistemas multi-agentes**: refere-se a uma abordagem computacional alternativa em que, no PLN, remete a uma organização em sociedade por agentes. A abordagem multi-agentes estuda os diferentes fenômenos linguísticos como ambiguidade, anáforas e elipses.

O PLN busca realizar a comunicação entre a linguagem humana e o computador, entretanto, nem sempre todos os níveis de entendimentos são compatíveis. Apresentam-se questões relativas e de importância aos diferentes níveis de estudos do domínio da linguística relacionados à engenharia da computação ou ao formalismo computacional envolvido por trás deles para o desenvolvimento de programas de PLN (VIEIRA; LIMA, 2001; GONZALEZ; LIMA, 2003; OTHERO, 2006; PUSTEJOVSKY; STUBBS, 2012).

- **Fonético e fonológico**: áreas da linguística que estudam os sons das palavras produzidas pelas línguas humanas. A fonética refere-se aos estudos dos fones e sons concretizados pela fala. Além disso, suas áreas de interesse estão na acústica, articulação e fisiologia da produção dos sons pela fala. A área de estudo se divide em subáreas como a fonética articuladora, que envolve questões dos mais de 100 músculos no controle direto e contínuo para a produção das ondas sonoras da fala e a fonética acústica, que estuda as propriedades físicas das ondas sonoras da fala. Já a fonologia busca compreender os fonemas e o sistema fonológico subjacente de uma língua. Trata-se do estudo das regras e princípios relacionados à organização, estrutura e distribuição dos sistemas de sons por meio de uma linguagem natural. Destacam-se por meio do desenvolvimento dessas

duas áreas da linguística junto ao PLN os sistemas de reconhecimento de fala e a síntese de fala e sistemas de diálogos em língua falada;

- **Morfológico**: estudo da construção e classificação em categorias das palavras a partir de uma unidade significativa primitiva. Refere-se ao tratamento específico do conhecimento sobre as estruturas. Exemplos estão na palavra “árvore”, que não pode ser quebrada em unidades menores, entretanto, essa ocorrência acontece com as palavras “árvores” e “arvorezinhas”. As unidades constituintes das palavras são denominadas morfemas, como no caso de “impossível” ou “sobremesa”, em que as constituintes podem ser independentes ou dependentes. Na classificação das palavras são utilizadas as categorias como substantivos, verbos, adjetivos, preposições e advérbios;
- **Sintático e semântico**: o sintático é o estudo das palavras, das construções estruturais das frases e das constituições das sentenças por meio de regras, condições, princípios subjacentes à organização estrutural dos elementos envolvidos. O semântico é o estudo do relacionamento das palavras e preposições e de como são combinadas para formar os significados das sentenças de maneiras mais independentes. Os estudos em Sintaxe Generativa e Funcional e em Semântica Cognitiva, Computacional e Formal podem apresentar bons resultados em sistemas que lidam com geração automática de sentenças. O estudo do significado está centralizado no significado da palavra por meio da semântica lexical, que considera as propriedades de cada uma das palavras ou por meio do valor verdade de uma preposição. Nesse caso, acontece por meio da semântica lógica, que, com base na teoria dos conjuntos, estuda o significado a partir das especificações do domínio de conhecimento, podendo explorar a lógica temporal e/ou a lógica intencional;
- **Pragmático**: uso das frases e sentenças em diferentes contextos de forma que afetam seu significado. Refere-se ao estudo das relações dos significados com o contexto da enunciação e procura integrar fatores como contexto e falantes. Refere-se ao acordo mútuo entre os falantes em uma conversação ou por meio da teoria dos atos de fala,

> que remetem a uma nova maneira de compreensão da linguagem natural. O domínio pragmático é uma área de estudos da Linguística e também da Sociolinguística, Antropologia, Filosofia, Psicologia, Psicolinguística e Ciências da Computação.

O PLN perpassa por um ciclo de desenvolvimento em que os recursos tecnológicos utilizados para realizar a codificação de um determinado fenômeno linguístico devem ser capazes de capturar o comportamento desejado por meio dos algoritmos que estão em treinamento. As descrições linguísticas são extraídas a partir da modelagem teórica do fenômeno e formam a base dos valores de anotação da linguagem de especificações que, por sua vez, podem ser construídas em diferentes modelos como *Syntatic Bracketing*, *Syntatic Tree Structure*, *Ontology* ou mesmo, de forma mais simples, por meio do *Question Answering*, geralmente, com perguntas do tipo *Who*, *What*, *Where*, *When?* de uma sentença para identificar os papéis semânticos. As anotações são utilizadas como recursos para o ciclo de desenvolvimento no qual irá treinar e testar o algoritmo. Por fim, com base em uma análise de avaliação, o modelo é revisado e colocado em testes para novos treinamento (PUSTEJOVSKY; STUBBS, 2012). Outras formas técnicas de anotações também são utilizadas, como anotação gramatical, sintática parcial e de discurso (VIEIRA; LIMA, 2001).

Entretanto, faz-se necessário que os dados sejam preparados de maneira que os algoritmos possam encontrar facilmente os padrões e as inferências. Uma forma de realizar esse tratamento está na inserção de metadados relevantes aos *corpora* de dados. Qualquer *tag* de metadados recebe o nome de anotação sobre a entrada quando utilizado para marcar elementos de um conjunto de dados. Para que os algoritmos aprendam em eficiência e eficácia, é preciso realizar uma anotação precisa e relevante para a tarefa que a máquina está sendo solicita a executar (PUSTEJOVSKY; STUBBS, 2012).

3.2. Modelagem de tópicos

A modelagem de tópicos possibilita organizar e resumir, por meio de algoritmos que utilizam métodos estatísticos, conteúdos de arquivos eletrônicos que constituem grandes volumes de dados e informações, chamados de *corpus* de dados. Sua relevância está no fato de que, em escalas elevadas, torna-se humanamente impossível descobrir e analisar os temas e suas relações a partir

de anotações manuais. A modelagem de tópicos permite, por meio de algoritmos de *Machine Learning* utilizados em estruturas não supervisionadas, identificar determinados padrões, relações e mudanças de termos realizados ao longo de um determinado período de tempo estabelecido em *corpus* estudado por meio de suas estruturas latentes e de um conjunto de informações pressupostas (BLEI, 2012).

Considerada um método estatístico para descobrir temas em *corpora* de dados, a modelagem de tópicos também é vista como uma clusterização *Fuzzy* ou análise de agrupamento. Trata-se de uma técnica de mineração de dados multivariados realizada por meio de métodos numéricos extraídos unicamente das informações contidas no conjunto de dados, com o objetivo de agrupar de forma automática os dados em grupos independentes (DE OLIVEIRA; PEDRYCZ, 2007; NOLASCO; OLIVEIRA, 2016b).

As estruturas latentes relacionam-se a conjuntos temáticos formados por tópicos que, consequentemente, reúnem palavras semanticamente próximas e determinantes para a escolha de palavras que formarão um documento de acordo com a escolha do modelo idealizado (BLEI, 2012; KASZUBOWSKI, 2016). A modelagem probabilística de tópicos é utilizada para solucionar problemas de agrupamento e organização dos dados em formato textual. Seu objetivo está na descoberta de tópicos e anotação de *corpus* de dados por meio da classificação temática, a fim de analisar as palavras contidas nos textos originais de forma que possa descobrir os temas contidos na coleção de documentos (BLEI, 2012).

Na modelagem de tópicos, os algoritmos de aprendizagem de máquina utilizam uma estrutura não supervisionada. Dessa forma, os dados que não possuem rótulos históricos e os resultados não são conhecidos previamente (AYODELE, 2010; PUSTEJOVSKY; STUBBS, 2012). Por conta disso, a clusterização é considerada uma técnica primitiva quando comparada ao conceito de classificação que, por sua vez, desenvolve-se por meio de aprendizado supervisionado. Na classificação, são utilizados exemplos para que algoritmos de treinamentos possam alocar os dados em cada classe (NOLASCO; OLIVEIRA, 2016b).

A clusterização se baseia em uma estrutura que se utilizada de dois parâmetros básicos, sendo N para um número de casos da base de dados, como

o número de documentos, e K o número de grupos como os temas existentes (DE OLIVEIRA; PEDRYCZ, 2007).

Tais algoritmos perpassam por um processo de treinamento necessário para realizar previsões. As previsões são executadas ao longo do processo até que se atinja um nível satisfatório de precisão dos dados de treinamento. O objetivo está em explorar os dados e apresentar alguma estrutura dentro deles. Com isso, o computador aprenderá a realizar determinados procedimentos de aprendizagem sem ser informado como. Uma das abordagens está em ensinar o agente sem apresentar categorizações explícitas, mas utilizando um sistema de recompensas para indicar o êxito (AYODELE, 2010; PUSTEJOVSKY; STUBBS, 2012).

Esses métodos probabilísticos para modelagem de tópicos utilizam conceitos das áreas da ciência da computação e da estatística para entender o conteúdo de documentos e são utilizadas para estruturar grandes coleções de documentos (BLEI, 2012; KASZUBOWSKI, 2016). Com isso, um dos grandes desafios da Ciência da Informação está em desenvolver interfaces inteligentes para uma melhor interação homem-máquina, de forma que possam apoiar os usuários na busca e recuperação da informação. A partir do uso de elementos ergonômicos, como a visualização da informação por meio da computação gráfica, têm-se encontrado resultados positivos no acesso e interpretação dos dados ao qual o processo em questão está na inteligência da máquina. Buscando uma interação de maneira mais natural entre os computadores e os seres humanos, faz-se necessário assegurar um potencial de ambivalência e imprecisão das solicitações dos usuários de forma que seja possível interpretar a diferenciação entre o que o usuário pode dizer ou fazer, além de poder compreender o que realmente se pretende (HOFMANN, 1999b).

Quando se fala em realizar pesquisas por meio da internet, existem duas maneiras principais que se destacam: pesquisa e *links*. Ao digitar uma palavra-chave em um sistema de buscas, por exemplo, como resultado apresenta-se ao usuário a possibilidade de interação por meio de uma lista de resultados associadas, respectivamente através de *hiperlinks*, ao qual existe um conjunto de documentos relacionados ao termo pesquisado (BLEI, 2012).

Esse típico cenário de interação homem-máquina no processo de recuperação da informação ocorre por meio de consultas a partir da utilização

de linguagem natural. Nesse caso, é formulada uma solicitação, fornecendo *N* número de palavras-chave ou quaisquer textos no formato livre. Posteriormente, o sistema retorna às informações desejáveis ou dados em que possua representações acessíveis. Métodos de recuperação baseados em estratégias de correspondência de palavras simples são utilizados para determinar o grau de relevância de um documento associado à consulta. Entretanto, a correspondência literal de termos possui desvantagens em questões como ambiguidade de palavras e imprecisão devido às diferenças individuais no uso da palavra (HOFMANN, 1999b).

Trata-se de uma maneira eficiente para o processo de interação homem-computador no qual destina-se a recuperação da informação em arquivos on-line. Entretanto, outras possibiliades podem ser discutidas, como por exemplo, explorar os documentos com base nos temas que os executam ao invés de palavra-chave. Dessa forma, pode-se aplicar *zoom in* e *zoon out* para encontrar temas específicos ou mais amplos, além de acompanhar a análise temporal de determinado tema ou compreender como os temas se conectam uns aos outros (BLEI, 2012):

> [...] considere usar temas para explorar a história completa do New York Times. Em um nível amplo, alguns dos temas podem corresponder às seções do jornal - política externa, assuntos nacionais, esportes. Poderíamos ampliar um tema de interesse, como a política externa, para revelar vários aspectos dele - a política externa chinesa, o conflito no Oriente Médio, o relacionamento dos EUA com a Rússia. Poderíamos então navegar no tempo para revelar como esses temas específicos mudaram, acompanhando, por exemplo, as mudanças no conflito no Oriente Médio nos últimos 50 anos. E, em toda essa exploração, seríamos apontados para os artigos originais relevantes aos temas. A estrutura temática seria um novo tipo de janela para explorar e digerir a coleção (p. 77, tradução nossa).

Embora a interação com arquivos eletrônicos não ocorra dessa maneira e o ser humano não possua capacidade de ler e estudar o quantitativo de textos já publicados de uma determinada área, bem como textos que surgem a todo momento para poderem fornecer uma experiência de navegação como a apresentada por Blei, pesquisadores da área de *Machine Learning* desenvolveram um conjunto de algoritmos que busca descobrir e anotar informações temáticas de grandes arquivos de documentos por meio da modelagem probabilística de tópicos. Tais algoritmos de modelagem de tópicos utilizam-se de métodos estatísticos, responsáveis por analisar as palavras dos textos originais para descobrirem os termos que os executam, como os temas

se conectam e como eles podem sofrer alterações durante determinado período (BLEI, 2012). Além disso, os algoritmos não necessitam de nenhuma anotação ou rotulagem de documentos, sendo considerados em *Machine Learning* como abordagem não supervisionada (ZHU et al., 2012). Dessa forma os tópicos emergem dos documentos originais, permitindo organizar, gerenciar, anotar e resumir documentos eletrônicos em larga escala, o que seria humanamente impossível (BLEI, 2012; ZHU et al., 2012).

Enquanto os documentos de uma coleção são observados, toda estrutura de tópicos, tais como os tópicos, as distribuições de tópicos por documentos e as atribuições por palavras por documentos permanecem ocultas durante o processo generativo. Os algoritmos de modelagem de tópicos normalmente se dividem em duas categorias, sendo elas baseadas em:

- **Amostragem**: possui como característica a coleta de amostra do posterior para aproximá-lo com uma distribuição empírica. O algoritmo de amostragem de Gibbs[9] é um dos mais utilizados na modelagem de tópicos para construção de uma cadeia Markov que, por sua vez, é definida nas variáveis dos tópicos ocultos de um determinado *corpus* de dados. O algoritmo possibilita executar a cadeia por um grande período de tempo, coletar amostras da distribuição limite e posteriormente aproximar a distribuição com as amostras coletadas (BLEI, 2012). Trata-se de uma técnica para gerar amostras de distribuições multidimencionais quando apenas se conhecem algumas das distribuições condicionais (GRUS, 2016);
- **Variacionais**: postulam uma família parametrizada em uma estrutura oculta para localizar o membro da família mais próximo do posterior. Isso permite transformar o problema de inferência em um problema de otimização de forma que possa ter impacto prático na modelagem probabilística. Ambas as categorias, por meio dos algoritmos ,realizam uma pesquisa sobre uma estrutura de tópicos. Destaca-se que uma coleção de documentos por meio das variáveis aleatórias observadas

[9] Amostragem de Gibbs é um algoritmo para gerar uma sequência de amostras da distribuição conjunta de probabilidades de duas ou mais variáveis aleatórias.

no modelo são mantidos fixos e funcionam como guia de localização (BLEI, 2012).

Para a descoberta da estrutura semântica latente, também conhecida como tópicos de uma coleção de documentos, podem ser utilizadas diferentes abordagens dos modelos de extração de tópicos. Durante o processo de modelagem de tópicos em um conjunto de documentos, a estrutura latente acaba por ser desconhecida e somente é possível ter acesso às variáveis observadas, sendo os termos em cada documento, quando geradas a partir dela (ZHU et al., 2012).

Os modelos de extração de tópicos probabilísticos utilizam-se de premissas no qual os documentos são um conjunto de tópicos misturados. Com isso, um tópico é constituído por uma distribuição probabilística de palavras. Essa formação do modelo de tópico probabilístico é composta por um modelo generativo probabilístico e quando aplicada aos documentos que formam determinado *corpus*, por sua vez, acaba por especificar um procedimento probabilístico pelo qual os documentos podem ser gerados. Nesse processo de gerar um determinado documento, uma nova distribuição sobre os tópicos é estabelecida. Consequentemente, três processos ocorrem: i) cada palavra será inserida num documento; ii) um tópico é escolhido aleatoriamente com base na distribuição definida anteriormente; e iii) uma palavra tópica é selecionada (STEYVERS; GRIFFITHS, 2007).

A possibilidade de inverter o processo pode ser gerada com técnicas estatísticas que inferem o conjunto de tópicos utilizados para gerar a coleção de documentos (STEYVERS; GRIFFITHS, 2007). A construção de um documento por meio da modelagem de extração de tópicos é realizada após a definição de um número quantitativo de assuntos ou tópicos a serem abordados. Tais tópicos são responsáveis por determinar os termos que serão utilizados nesse documento. Com a existência de um modelo generativo no qual o documento emerge e os parâmetros utilizados para construção dos documentos são desconhecidos, pode-se estimar o quantitativo de termos a partir dos documentos e dos termos por meio das variáveis observadas (SANTOS, 2015).

O autor ainda reforça que o processo de extração de tópicos relaciona-se em como encontrar a melhor estimativa possível dos parâmetros responsáveis por originarem os documentos constituintes de um *corpus* de forma que possam

assumir o modelo generativo. Obtém-se como resultado uma representação denominada documento-tópico, responsável por determinar um peso de cada tópico para cada documento e uma representação termo-tópico, que se relaciona ao modelo generativo escolhido e que pode representar: i) uma probabilidade da ocorrência do termo quando um determinado tópico ocorre em um documento; ii) frequência esperada desse termo; ou iii) peso estimado matematicamente não traduzido para um significado em um contexto linguístico (SANTOS, 2015).

Nos modelos não probabilísticos, um "tópico" pode ser entendido como um grupo de termos com pesos indicando a importância ou significância desses termos para algum assunto. Consequentemente, descobrir tópicos com modelos não-probabilísticos equivale a agrupar termos em conjuntos significativos. Trata-se de um modelo no qual utiliza-se a decomposição de matrizes (CHENG et al., 2013).

Com o crescente volume de informações oriundas de diferentes fontes de dados *online,* tem-se apresentado inúteis as tentativas de organizar, manipular ou pesquisar de maneira eficiente as informações, como apresentado em obras da década de 1990. Desde então, técnicas de recuperação da informação foram fundamentais para abordar de forma eficiente os problemas específicos. Destaca-se como técnica insuficiente a correspondência lexical simples que produz resultados imprecisos e de baixa relevância por causa da sinonímia e diferentes maneiras de expressar determinado conceito, além da polissemia com múltiplos significados para palavras comuns em um conjunto de informações a serem pesquisadas (BERRY; DUMAIS; O'BRIEN, 1995; DEERWESTER *et al*., 1990; BLEI, 2012).

Obtém-se como resultado da modelagem de tópicos aplicado em um determinado conjunto de dados uma combinação de tópicos, em que cada tópico é representado por um conjunto de termos e seus respectivos pesos. Dessa maneira, cada tópico resultante da modelagem de tópicos possui um termo de maior relevância com base na sua respectiva probabilidade (HOFMANN, 1999b; BLEI; NG; JORDAN, 2003; BLEI; CARIN; DUNSON, 2010).

Os tópicos emergentes sugerem uma representatividade do(s) tema(s) e assunto(s) contido(s) na coleção de documentos, entretanto, esses resultados, quando fora do contexto, podem gerar dificuldades de interpretação por pessoas que não tenham conhecimento sobre o domínio da linguagem. Dessa forma, faz-

se necessário uma maior interpretação semântica dos tópicos para uma melhor identificação dos temas (NOLASCO; OLIVEIRA, 2016b).

Um exemplo está nos termos “foguete” e “espaço”, que podem ser associados com uma maior probabilidade ao tópico “viagens espaciais”, diferente do tópico “genética” que, por sua vez, contempla os termos “gene” e “DNA”. Termos generalistas como “planeta”, “isso” e “desde” podem aparecer de forma equivalente em ambos os tópicos. Dessa forma, cada tópico extraído poderia ser representado por um conjunto dos termos mais comuns, juntamente com os documentos que melhor representam o tópico (NOLASCO; OLIVEIRA, 2016b).

Os resultados dos modelos de extração de tópicos não apresentam de forma automatizada possíveis nomes para os temas ou assuntos em que os termos estejam relacionados, sendo necessária a colaboração de um especialista no domínio da linguagem para definir a suposição dos nomes para os tópicos (SOUZA; JÚNIOR; SOUZA, 2019; SOUZA; SOUZA, 2019).

Uma maneira de auxiliar no processo da interpretação semântica dos tópicos está na rotulagem, que possibilita explorar a coleção ou o(s) tópico(s) com o conjunto de termos por meio do título do documento mais representativo para aquele tópico. Essa técnica reduz a necessidade de conhecimento especializado do domínio da linguagem para se realizar a interpretação dos dados (MANNING; RAGHAVAN; SCHÜTZE, 2009). Técnicas como o uso de uma matriz de termos-documentos (PAPADIMITRIOU et al., 1998) e uso de estrutura do documento e a relevância dos termos para cada seção (NOLASCO; OLIVEIRA, 2016b) são utilizadas para a criação de rótulos mais informativos.

Um dos primeiros algoritmos mais representativos de modelagem de tópicos utilizados para organização de documentos textuais foi descrito por Papadimitriou *et al*. (1998). O marco de desenvolvimento da técnica é conhecida como *Latent Semantic Analysis* (LSA) - Análise de Semântica Latente (LANDAUER; DUMAIS, 1997). Ela se destaca na área da pesquisa em recuperação de informações por retornar documentos correspondentes ao realizar uma busca por palavras-chave, categorizar documentos e generalizar resultados por meio de documentos equivalentes disponibilizados em diversas línguas (CHANG *et al*., 2009). Posteriormente, diversos outros modelos foram criados para otimizar a extração de tópicos em corpora de dados. O *Latent*

Dirichlet Allocation (LDA) - Alocação Latente de Dirichlet é um dos modelos generativos mais populares e serviu como base para criação de outros modelos probabilísticos, além de ter tido como base os modelos LSA e *Probabilistic Latent Semantic Indexing* (pLSI) - Indexação Semântica Latente Probabilística (STEYVERS; GRIFFITHS, 2007; DAVID, 2012).

3.2.1. Latent Semantic Indexing

O modelo *Latent Semantic Analysis* (LSA) - Análise Semântica Latente foi o marco inicial para desenvolvimento de modelos de extração de tópicos aplicada para *Information Retrieval* (IR) - Recuperação da Informação de documentos e termos, antecedendo à extração automática de assunto num espaço semântico latente (LANDAUER; DUMAIS, 1997; HOFMANN, 1999b; CHANG *et al*., 2009; AGGARWAL; ZHAI, 2012). Em IR, o LSA recebe também o nome *Latent Semantic Indexing* (LSI) - Indexação Semântica Latente é utilizado para recuperar e categorizar documentos e generalizar resultados.

Trata-se de um conjunto de procedimentos automatizados que busca medir, por meio quantitativo, a semelhança de significado entre duas palavras ou grupos de palavras. O modelo LSI é uma técnica utilizada para reduzir o tamanho dos descritores contidos em um determinado *corpus* de dados e amplamente utilizada na área da recuperação da informação baseada em modelo vetorial introduzido (DEERWESTER *et al.*, 1990; BERRY; DUMAIS; O'BRIEN, 1995) a partir de 1988 (DEERWESTER *et al.*, 1988) e obtendo bons resultados, apesar de sua simplicidade (GRAESSER *et al*., 2000).

A recuperação dos documentos ocorre por meio de buscas por palavras-chave enquanto a categorização é realizada por especialistas de acordo com a área de domínio. Além disso, o modelo generaliza os resultados por meio de documentos próximos em diversas línguas (CHANG *et al*., 2009). Dessa forma, assume-se que nos documentos contenham alguma estrutura subjacente ou latente no padrão de uso das palavras, utilizando, assim, técnicas estatísticas para estimar a estrutura latente do conteúdo semântico dos documentos na coleção (DEERWESTER *et al.*, 1990; DUMAIS, 1995). Para representar e recuperar informações, ao invés de utilizar palavras em nível de superfície, utiliza-se uma descrição dos termos, documentos e consultas de usuários com base na estrutura subjacente semântica latente. (DUMAIS, 1995).

A semelhança entre documentos ou documentos e consultas podem possuir uma maior confiabilidade em sua representação quando reduzida no espaço latente do que em sua representação original. Além disso, os documentos que compartilham de termos coocorrentes terão representação semelhantes no espaço latente. Dessa forma, o modelo apresenta uma redução de ruídos e potencializa a detecção de sinônimos de palavras que se referem ao mesmo tópico (HOFMANN, 1999b).

O modelo LSI, utilizado de um método de indexação automática que projeta intencionalmente os documentos de alta dimensão e seus termos com base em suas frequências em um espaço de baixa dimensionalidade, representa o conceito semântico no documento. Com isso, o modelo LSI permite realizar a análise de documentos puramente baseada em termos e de maneira conceitual, quando projetados os documentos num espaço semântico (DEERWESTER *et al.*, 1990; AGGARWAL; ZHAI, 2012). As dependências entre os termos dos documentos de um determinado *corpus* possuem relevância em sua representação e é simultaneamente explorada na recuperação por meio de suas inter-relações entre termos e documentos. Destaca-se como vantagem na utilização do método referente a representação LSI que uma consulta pode possuir similaridade aos documentos, mesmos quando não compartilham de palavras (DUMAIS, 1995).

O LSI utilizou-se do ferramental da álgebra linear para decompor um *corpus* nos seus temas constituintes e reduzir os efeitos adversos gerados pela sinonímia e polissemia por meio da identificação de associações estatísticas entre os termos. Dessa forma, aplica-se mais especificamente à decomposição *Singular Value Decomposition* (SVD) - Decomposição de Valor Singular em uma matriz com que realiza a contagem de frequência dos termos contidos nos documentos de todo o *corpus* ou de apenas fragmentos desses documentos (HOFMANN, 1999a; CHANG *et al.*, 2009; AGGARWAL; ZHAI, 2012). O princípio dessa técnica é que o espaço original formado pelos termos W é rotacionado de maneira que: i) o primeiro eixo aponte para a direção de maior variância dos documentos; e ii) o segundo eixo aponte para a direção de segunda maior variância e assim sucessivamente (AGGARWAL; ZHAI, 2012).

O modelo LSI refere-se a palavras utilizadas em um mesmo contexto e que tendem a ter significados semelhantes. Desenvolve-se utilizando um

conjunto de procedimentos estatísticos e automatizados que possibilita aferir, de forma quantitativa, a semelhança de significados entre duas palavras ou um grupo de palavras utilizadas no mesmo contexto, o que possibilita a extração de tal conteúdo conceitual de um determinado *corpus* e estabelece a associação entre os termos (WITTER; BERRY, 1998).

O LSI permite realizar associações desconhecidas entre as palavras de forma que possam ser induzidas a partir de uma grande análise, ao identificar como as palavras contidas em documentos ocorrem em combinação com as outras palavras por meio da língua natural. Além disso, o modelo LSI também pode ser utilizado para determinar a similaridade entre palavras ou documentos do *corpus* com documentos externos (MARTIN; BERRY, 2011).

O LSI utiliza a SVD de uma matriz de termos por documentos responsável por identificar um subespaço linear que apresenta uma maior variação no espaço de características. O SVD está relacionado à análise fatorial para modelar as relações associativas e a decomposição de autovalores (DUMAIS, 1995; WITTER; BERRY, 1998; HOFMANN, 1999b). Os vetores singulares produzem um espaço-k e os vetores singulares correspondentes são utilizados para codificar os termos e os documentos em um espaço vetorial de dimensão *k* juntamente com uma consulta de usuário. Dessa forma, por meio do modelo LSI, termos e/ou documentos de importância podem ser recuperados e correspondidos até mesmo quando não houver palavras em comum com os documentos relevantes (WITTER; BERRY, 1998).

Em uma matriz de documentos de termos acontece a decomposição em um conjunto de *k,* onde a matriz original pode se aproximar de uma cominação linear por meio dos fatores ortogonais. A representação de documentos e consultas – que seriam realizadas por um conjunto de palavras independentes – são representados como valores contínuos em cada uma das dimensões *k* de indexação ortogonais. Com isso, as palavras não serão independentes, uma vez que os números de fatores e dimensões são menores em relação ao número de termos exclusivos. Dessa forma, se dois ou mais termos foram utilizados em contextos semelhantes em documentos, acabaram por ter vetores semelhantes na representação LSI de dimensão reduzida. Faz-se necessário destacar que a SVD pode capturar melhor essa estrutura do que realizar uma simples correlação de termos ou documento-documento e *clusters* (DUMAIS, 1995).

A SVD é uma forma geral de análise fatorial no qual condensa uma grande matriz de dados *word-by-context* para uma consideravelmente menor. Mesmo assim, ainda contém informações de relevância sobre os dados. Essa técnica consiste em quatro etapas: i) construir uma matriz de documentos a partir de um *corpus* ou *corpora* de documentos; ii) realizar a decomposição SVD da matriz; iii) escolher *n* componentes principais; e iv) utilizar uma métrica de semelhança como cosseno para encontrar o documento mais semelhante (SCARPA, 2017).

A *Principal Components Analysis* (PCA) - Análise de Componentes Principais está relacionada ao conceito de SVD e trata-se de um método de redução de dimensionalidade que permite encontrar uma projeção dos pontos alocados em um subespaço de dimensão *k*. Dessa forma, acaba por preservar dois pontos fundamentais: características genéticas e *clustering* dos pontos originais (SCARPA, 2017). Trata-se de um dos métodos de extração de características que trabalha com projeção linear por meio de aprendizagem não supervisionada. A PCA assemelha-se a *Factor Analysis* (FA) – Análise Fatorial e a *Multidimensional Scale* (MDS) - Escala Multidimensional (ALPAYDIN, 2010).

3.2.2. Latent Dirichlet Allocation

Um dos modelos probabilísticos generativos mais utilizados é o *Latent Dirichlet Allocation* (LDA) - Alocação de Dirichlet Latente. O modelo utiliza uma abordagem bayesiana e parte do princípio de que os documentos contidos em um determinado *corpus* sejam representados como misturas aleatórias de tópicos latentes. Posteriormente, cada tópico passa a ser caracterizado por uma distribuição de palavras que compreendem a cada um dos documentos (BLEI, 2012).

No que diz respeito à modelagem probabilística de tópicos, as estruturas-base referentes ao modelo LDA foram baseadas por meio do trabalho seminal dos modelos *Latent Semantic Analysis* (LSA) - Análise Semântica Latente e *Probabilistic Latent Semantic Indexing* (pLSI) - Indexação Semântica Latente Probabilística que, consequentemente, destacaram-se pela criação de outros modelos probabilísticos de tópicos. O LDA refere-se a uma evolução do modelo LSA com o uso de fórmulas probabilísticas (STEYVERS; GRIFFITHS, 2007; BLEI; LAFFERTY, 2007; 2009) e uma extensão do modelo *Probabilistic Latent Semantic Analysis* (pLSA) - Análise Semântica Latente Probabilística, permitindo

assim propor um modelo generativo probabilístico com base no vocabulário fixo de termos. Dessa forma, os tópicos são definidos por meio de uma distribuição de probabilidades (SANTOS, 2015). Trata-se de uma técnica que possui aplicações na recuperação e filtragem da informação, Processamento de Linguagem Natural (PLN) e *Machine Learning* a partir de texto e suas áreas de relacionamento, levando em consideração a técnica estatística para análise de modo e coocorrência de dados (HOFMANN, 1999a).

A modelagem de tópicos induz a relação entre tópicos e documentos de um ou mais *corpus*. Dessa forma, o LDA é uma técnica de modelagem de tópicos considerada mais simples e utilizada para extrair tópicos de dados textuais (BLEI; NG; JORDAN, 2003; BLEI, 2012; GRUS, 2016). Com isso, os modelos de tópicos aprendem tópicos representados por um conjunto de palavras importantes extraídas automaticamente de documentos não marcados e de maneira não supervisionada. Os algoritmos não possuem informações sobre os assuntos e os documentos não são rotulados por palavras-chave ou tópicos, entretanto, faz-se necessário medidas de coerência para diferenciar os tópicos bons dos ruins, uma vez que pode existir ou não a garantia da interpretação dos dados (BLEI; NG; JORDAN, 2003; BLEI, 2012).

Trata-se de um modelo estatístico descrito pelo seu processo generativo de indexação semântica probabilística – aleatório imaginário, que quando aplicado coleções de documentos, resulta em distribuições tópicas de probabilidade multinominal interpretáveis sobre os termos gerados pelo agrupamento flexível de palavras (BLEI, 2012; BLEI; NG; JORDAN, 2003). O modelo LDA é um modelo bayesiano hierárquico que possui três níveis: i) cada item de uma coleção é moldado como uma mistura finita sobre um conjunto subjacente de tópicos; ii) cada tópico é modelado como uma mistura infinita sobre um conjunto subjacente de probabilidade de tópico; e iii) as probabilidades dos tópicos fornecem uma representação explícita de um documento (BLEI; NG; JORDAN, 2003; NOLASCO; OLIVEIRA, 2016b).

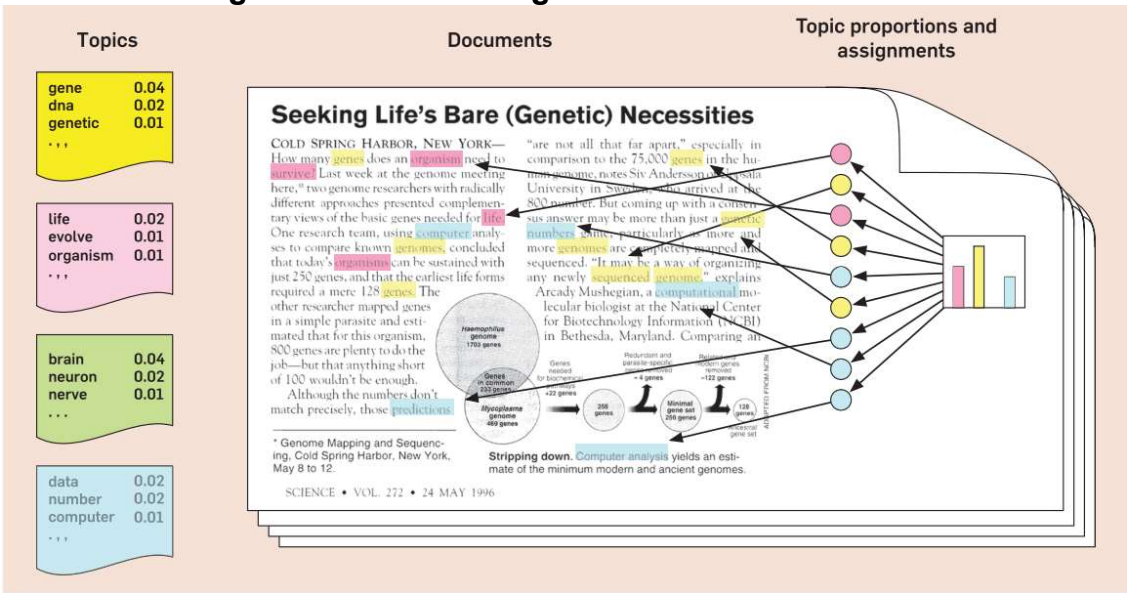
O modelo LDA presume que existe um número fixo *K* de tópicos e uma variável aleatória responsável por atribuir a cada tópico uma probabilidade de distribuição associada às palavras. Essa distribuição pode ser pensada como a probabilidade de ver a palavra *w* para o tópico *K*. Existe também outra distribuição aleatória: para cada documento é atribuída a probabilidade de

distribuição do tópico, podendo ser considerado como uma mistura de tópicos no documento *d.* Dessa forma, as palavras são geradas inicialmente pela escolha aleatória de um tópico, sendo da distribuição do tópico dos documentos para depois ser gerada a palavra referente à distribuição das palavras dos tópicos (BLEI, 2012; GRUS, 2016).

O modelo define um tópico por meio da distribuição de probabilidade sobre o vocabulário fixo, antes mesmo dos documentos. Em um tópico de "genética" por exemplo, será constituído por palavras relacionadas aos termos que possuam maior probabilidade de ocorrência. Probabilidade de baixa ocorrência ou zero poderá ocorrer em tópicos que se relacionam com quaisquer outros assuntos diferente de "genética". Todos os tópicos contêm distribuição de palavras com probabilidades sobre o vocabulário fixo (NOLASCO; OLIVEIRA, 2016b).

A Figura 02 ilustra o processo generativo do modelo LDA de Blei (2012). Inicialmente, assume-se um certo número de tópicos constituídos por uma distribuição de palavras com os respectivos percentuais de representatividade para todo o *corpus* de dados, representado à esquerda da ilustração. Os documentos são gerados a partir da escolha da distribuição sobre os tópicos representados pelo histograma à direita. Posteriormente, para cada palavra há uma atribuição de tópico, representada pelas moedas coloridas. Por fim, a escolha da palavra que será associada ao tópico correspondente, representada pelos retângulos coloridos (BLEI, 2012).

Figura 02 – Processo generativo do modelo LDA

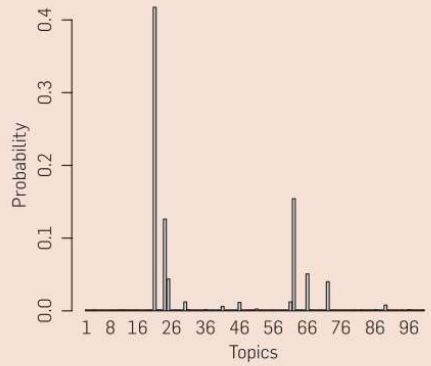


Fonte: (BLEI, 2012).

Ao aplicar o modelo estatístico LDA em um conjunto de documentos, os tópicos são interpretáveis como temas na coleção e as representações do documento remetem aos temas de cada documento. Destacam-se três pontos importantes: variáveis aleatórias ocultas codificam a estrutura temática; os tópicos aprendidos resumem a coleção e as representações dos documentos; e os *corpora* em grupos sobrepostos são organizados pela representação do documento (CHANEY; BLEI, 2012). Cada documento contido em um *corpus* possui sua destruição própria de tópicos. Dessa forma, cada documento pode conter vários tópicos e cada um deles contém a sua proporção de relevância. Tal distribuição de tópicos para cada documento está relacionada à distribuição multivariada de *Direchlet* (SANTOS, 2015).

A Figura 03 ilustra o modelo LDA, configurado para identificar 100 tópicos em um *corpus* de dados formado por 17.000 artigos científicos da revista científica Science. À esquerda da imagem são representadas as proporções dos tópicos inferidos e, à direita, as proporções dos tópicos mais frequentes com as 15 palavras de maior frequência para cada tópico (BLEI, 2012).

Figura 03 – Tópicos extraídos do modelo LDA



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Fonte: (BLEI, 2012).

O modelo LDA é descrito por meio da notação:

1. Dado os tópicos $\theta_{1:N}$, onde cada θ_n vocabulário V.
2. As proporções dos tópicos para o d-ésimo documento são ρ_d, onde ρ_{dn} é a proporção do tópico n no documento d.

3. As atribuições de tópicos para o d-ésimo documento são z_d, onde $z_{d,i}$ é a atribuição do tópico para a i-ésima palavra no documento d.
4. Por fim, as palavras observadas para o documento d são w_d, onde $w_{d,i}$ é a i-ésima palavra no documento d, a qual é um elemento do vocabulário V.

O processo generativo em LDA corresponde à distribuição conjunta das variáveis observadas e ocultas representada pela expressão:

$$p(\theta_{1:N}, \rho_{1:D}, z_{1:D}, w_{1:D}) =$$

$$\prod_{j=1}^{N}$$

$$\prod_{j=1}^{N} p(\theta_d) \prod_{d=1}^{D} p(\rho_d) \left(\prod_{i=1}^{l} p(z_{d,i}|\rho_d) p(w_{d,i}|\theta_{1:N}, z_{d:i}) \right)$$

Os modelos de tópicos são considerados de suma importância para a exploração de dados onde os tópicos acabam por apresentar um resumo do *corpus* de dados. A análise de um modelo de tópico pode revelar conexões e concorrências entre documentos, o que seria impossível ou não estariam óbvias quando realizadas de forma manual. Importa ressaltar que os modelos de extração de tópicos não são definitivos, sendo necessário realizar ajustes do modelo ao corpus. Dessa forma, é preciso utilizar outros métodos que evidenciem os assuntos contidos na coleção de documentos (BLEI; LAFFERTY, 2009).

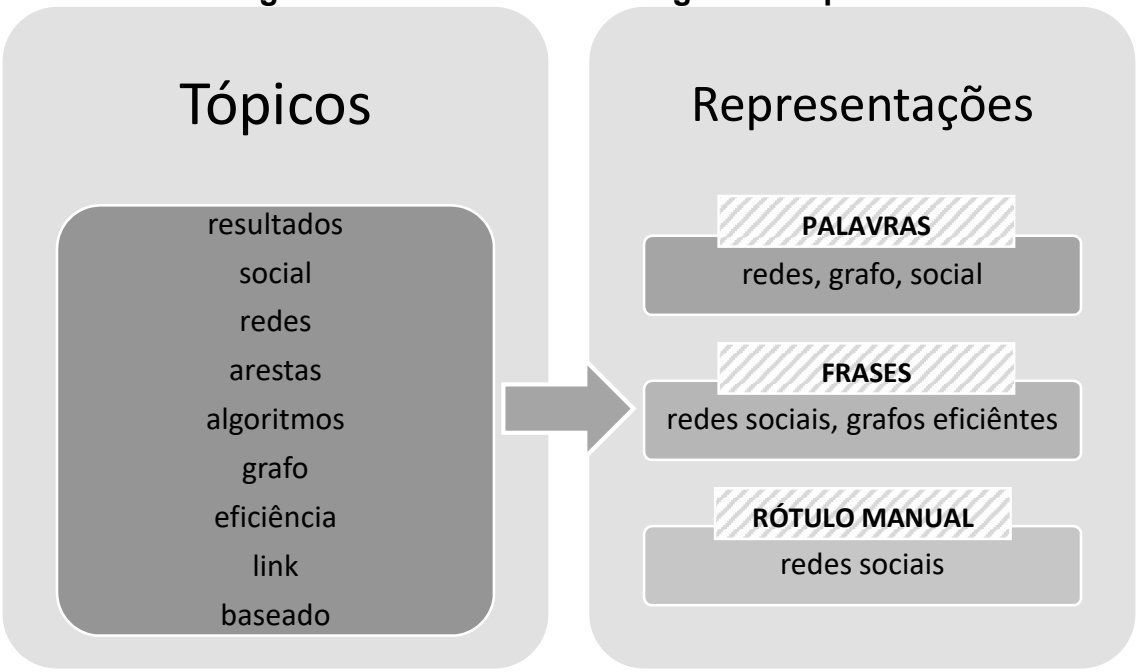
3.2.3. Rotulagem

A rotulagem de tópicos é uma técnica que diminui a dependência de especialistas da linguagem de domínio do *corpus* analisado, exibindo, junto ao resultado, tópicos semanticamente mais coerentes. Após conhecimento sobre o quantitativo de áreas contidas em um determinado *corpus*, torna-se possível realizar o agrupamento entre áreas e, consequentemente, gerar uma representação visual de tópicos da coleção. A rotulagem apresenta temas e assuntos extraídos de *corpora* de dados para que os analistas, pesquisadores ou usuários possam explorar nas temáticas de coleções informações que

desejam em bases desconhecidas ou numa base totalmente nova (NOLASCO; OLIVEIRA, 2016b).

Uma das técnicas mais utilizadas e encontradas na literatura é a geração de uma lista para que o usuário possa entender o assunto da área descrita (PAPADIMITRIOU *et al*., 1998; BLEI, 2012). Outra técnica utilizada e de preferência por classificadores humanos está em descrever os termos emergentes, expressando o conceito relacionado, seja por palavras, pequenas frases ou até mesmo sentenças (CHANG *et al*., 2009) conforme apresentadas na Figura 04.

Figura 04 – Técnica de rotulagem de tópicos



Fonte: (NOLASCO; OLIVEIRA, 2016b).

O uso da lista elaborada a partir do agrupamento pode auxiliar os classificadores humanos durante o processo de identificação dos assuntos, entretanto, faz-se necessário que haja um determinado conhecimento da área analisada com o domínio do *corpus* de dados durante o processo de criação dos rótulos. Uma pessoa que não tenha conhecimento da temática do domínio, pode encontrar dificuldades ao interpretar a lista e identificar os conceitos, o principal assunto ou mesmo associar as palavras de maneira que seja possível constituir termos com significados que representem o assunto analisado (NOLASCO; OLIVEIRA, 2016b).

A modelagem de tópicos, quando utilizada para agrupamento, em sua maioria, acaba por usar a distribuição de termos extraídos como própria

representação (HOFMANN, 1999b; BLEI, 2012). Uma opção está em deixar que o processo de rotulagem seja realizado por um especialista do domínio da linguagem, capaz de criar os rótulos manualmente de acordo com a lista de documentos contidos na coleção de documentos (CHANG *et al*., 2009). A técnica de rotulagem consiste no agrupamento de termos semanticamente próximos e coerentes de forma que seja possível gerar as representações de tópicos da coleção (SOUZA; SOUZA, 2019). Trata-se de uma opção confiável, uma vez que o especialista possui conhecimento para interpretar os dados e transmitir de forma correta os rótulos, entretanto, interpretações particulares podem ocorrer, dependendo do *background* do especialista. Além disso, ambientes com grandes volumes de informações e dados correlacionados podem ser custosos e levar muito tempo (NOLASCO; OLIVEIRA, 2016b).

Outras técnicas de rotulagem também são utilizadas, como a abordagem semi-supervisionada e a abordagem de aprendizado ativo. Na primeira, os rótulos criados são genéricos e posteriormente refinados com a ajuda de um especialista (LAU *et al*., 2011; RAMAGE; MANNING; DUMAIS, 2011). Já na segunda abordagem, o sistema extrai de forma simples os termos de maior relevância e o especialista dá retorno ao sistema, informando-o se o rótulo está bom ou precisa ser melhorado (YANG *et al*., 2014).

Dessa forma, as abordagens não-supervisionadas acabam por se tornar mais rápidas pela falta do conhecimento especializado. Já as tradicionais, apresentam melhores resultados, entretanto, utilizam mais recursos pessoais e de tempo. Por fim, as abordagens semi-supervisionadas buscam otimizar os processos das abordagens tradicionais, reduzindo o montante de trabalho introdutório do especialista da linguagem de domínio antes do trabalho manual. O grande desafio na geração de rótulos após a realização da modelagem de tópicos está na criação automática de sua representação, auxiliando na interpretação para os conhecedores do domínio de linguagem ou auxiliando na definição dos temas para usuários não especialistas no conteúdo do *corpus* (NOLASCO; OLIVEIRA, 2016b).

3.2.4. Desafios para a modelagem de tópicos

Por se tratar de uma área de pesquisa ativa, a modelagem de tópicos apresenta uma série de pressupostos apontados como melhorias para o

desenvolvimento de novos modelos de tópicos e a inserção de metadados no *corpus* de dados como forma de enriquecer a modelagem de documentos (NOLASCO; OLIVEIRA, 2016a).

O primeiro pressuposto está relacionado ao conceito de *bag-of-words* – saco de palavras, referindo-se que a ordem das palavras contidas nos documentos de um *corpus* de dados não possui relevância. Esse pressuposto pode ser relevante desde que o único objetivo da aplicação seja unicamente relevar a estrutura semântica contida no corpus (NOLASCO; OLIVEIRA, 2016a). Uma forma de considerar a ordem das palavras está no uso de bigramas ao invés de unigramas (WALLACH, 2006) ou N-gramas por meio de bigramas (WANG; MCCALLUM; WEI, 2007).

O segundo pressuposto traz que a ordem dos documentos contidos em um *corpus* de dados não é relevante para a modelagem de tópicos. Esse pressuposto pode ser relevante desde que o objetivo não seja analisar alterações dos tópicos ao longo dos anos ou séculos (NOLASCO; OLIVEIRA, 2016a). Pesquisas surgiram no campo da modelagem de tópicos dinâmica a fim de modelar e identificar temas contidos em coleções ao longo do tempo (CALTABIANO; VERZI; SCAPPUZZO, 1989).

O terceiro pressuposto diz respeito ao número de tópicos ser conhecido e fixo, sendo necessário definir, antes da execução do modelo, o quantitativo de tópicos a serem extraídos (BLEI, 2012). Para esse pressuposto a modelagem de tópicos hierárquica não necessita de um número de tópicos como entrada, assim como a clusterização não necessita de um número inicial de *clusters* (PUJARA; SKOMOROCH, 2012; TEH *et al.*, 2006). O uso da prática de identificar o quantitativo de tópicos de um *corpus* por suposição, tentativa, erro e acerto passou a ser solucionada com o uso da técnica de análise de estabilidade que verifica qual o número de tópicos estáveis que melhor representa a coleção (GREENE; O'CALLAGHAND; CUNNINGHAM, 2014).

O quarto pressuposto impossibilita modelar a correlação entre os termos ao considerar que os tópicos são independentes (BLEI, 2012). Esse pressuposto expande a pesquisa na área para a modelagem de tópicos correlacionais que busca extrair ou identificar as correlações entre os tópicos, levando em consideração que a modelagem de tópicos é considerada uma clusterização fuzzy (BLEI; LAFFERTY, 2007).

Os últimos pressupostos buscam um melhor desempenho e praticidade em grandes fluxos de dados de maneira que possa ser atualizado constantemente, sendo capaz de identificar o número de tópicos ideal para o *corpus* de dados, sendo eles: i) identificar a evolução dos tópicos ao longo do tempo; ii) correlacionar os tópicos de pesquisa multidisciplinares; e iii) aplicar no âmbito acadêmico com grande volume de dados (NOLASCO, 2016).

Levando em consideração aspectos de ordem prática para execução e atualização dos modelos, uma vez realizada a modelagem de tópicos e caso haja necessidade de quaisquer alterações nos documentos, faz-se necessário executar o algoritmo novamente para que se alcancem novos resultados. Outro ponto a ser considerado está no tamanho do vocabulário, que, quanto maior, exigirá mais processamento do equipamento para execução do algoritmo. Técnicas que removam termos insignificantes ou frequência nominal possibilitam otimizar o desempenho. A última consideração está relacionada à escolha do número de tópicos a serem extraídos de um *corpus* de dados. Uma vez que o usuário não tenha conhecimento do domínio e tamanho do *corpus*, poderá gerar saídas com números pequenos ou grandes de tópicos, o que resultará, respectivamente, em fusões de termos indesejáveis ou granularidade muito fina (BERRY; KOGAN, 2010).

O capítulo seguinte apresenta a metodologia de pesquisa utilizada para construção desta tese. Destaca os procedimentos técnicos realizados para durante a fase exploratória, preparação, pré-processamento, transformação, extração dos tópicos, validação e apresentação dos dados empíricos.

4. METODOLOGIA DA PESQUISA

Distinguir metodologia de métodos é premissa fundamental para descrição deste capítulo, uma vez que todos os procedimentos e técnicas para o desenvolvimento desta pesquisa, tanto para fase exploratória quanto para os dados empíricos, são apresentados de forma detalhada e servirão como norteamento para solicitações, questionamentos ou críticas a serem realizadas pela comunidade científica.

Metodologia se refere aos procedimentos e regras utilizadas por determinado método e possui como intenção a discussão sobre a problemática da pesquisa por meio do debate teórico, buscando criticamente maneiras de se fazer ciência. Método é definido como o caminho percorrido para se alcançar os objetivos, o êxito ideal para a produção do saber e constituído por meio de etapas para se alcançar determinado fim ou objetivo, levando em consideração a realidade empírica. A metodologia é constituída por regras direcionadoras para os métodos e ambos os conceitos são complementares (DEMO, 1995; CARVALHO, 2010; RICHARDSON, 2010; MICHEL, 2015).

O capítulo em questão está dividido em duas seções. Na primeira, são apresentados os aspectos gerais da pesquisa, tais como sua classificação, fase exploratória e empírica, procedimentos para análise e tratamento dos dados e as etapas da pesquisa. Na segunda seção são detalhadas a descrição dos processos metodológicos e etapas realizadas ao longo do percurso para obtenção dos resultados.

O comportamento científico é constituído de uma forma reacional, planejada e intencional, permitindo ao ser humano refletir e criticar suas ações, além de modificar o seu comportamento frente a uma sociedade ou a si mesmo. Esses processos ou etapas possibilitam que o indivíduo, enquanto membro de uma comunidade, possa identificar, elaborar e aplicar ações que atinjam seus objetivos e solucionem problemas de forma que modifique o ambiente no qual está imerso ou até mesmo o mundo (MICHEL, 2015).

A classificação da pesquisa tem surgido de forma com frequência nas diversas modalidades da pesquisa científica. Trata-se de uma característica da racionalidade humana para melhor organização e entendimento dos fatos, permitindo ao pesquisador reconhecer e aplicar, dentre um sistema de

classificação, melhor solução para o desenvolvimento da sua problemática de pesquisa (GIL, 2010).

4.1. Aspectos gerais

A pesquisa se classifica, de acordo com Creswell (2010), Gil (2010), Michel (2015) e Severino (2016):

- Quanto ao método, a pesquisa se classifica como ***indutiva***, caminhando do registro dos fatos particulares ou específicos para chegar a uma conclusão ampliada ou uma premissa geral. Parte-se da observação dos fatos ou fenômenos no qual se deseja conhecer. Posteriormente, compará-los e entender as relações existentes entre eles e, por fim, proceder à generalização com base nas relações verificadas;
- Quanto à finalidade/natureza, a pesquisa se classifica como ***aplicada***, uma vez que pode contribuir para a aplicação do conhecimento científico, buscando solucionar um problema específico e sugerir novas questões a serem investigadas. A pesquisa não visa somente a gerar um novo conhecimento, mas aplicá-la e intervir no mundo real;
- Quanto aos objetivos, a pesquisa se classifica como ***exploratória***, pois, além de buscar maior familiaridade com o problema, tem como objetivo identificar melhor um fato ou fenômeno no âmbito da pesquisa e, consequentemente, tornar o problema explícito. Além disso, permite utilizar um conjunto de procedimentos técnicos;
- Quanto à abordagem do problema, a pesquisa se classifica como método ***misto*** ou ***quali-quantitativo***, empregando a combinação ou associação das abordagens qualitativa e quantitativa, perpassando entre a interpretação dos fenômenos e a atribuição de significados por meio de dados quantificáveis, utilizando técnicas e recursos para classificá-las e analisá-las;
- Quanto aos procedimentos técnicos, a pesquisa se classifica como: ***bibliográfica***, a fim de construir o referencial teórico com base em materiais científicos já publicados, como artigos, livros, dissertações e teses; ***experimental***, a fim de determinar o objeto de estudo em sua concretude de fonte, sendo capaz de selecionar as variáveis que

possam influenciá-lo, consequentemente, definir as formas de controle e de observação dos efeitos em que as variáveis possam interferir sobre o objeto; e ***comparativo***, no qual procede a investigação entre as diferenças e similaridades contidas nos fatos e fenômenos nas bases de dados investigadas, levando em consideração as particularidades e relações entre elas.

4.2. Fase empírica

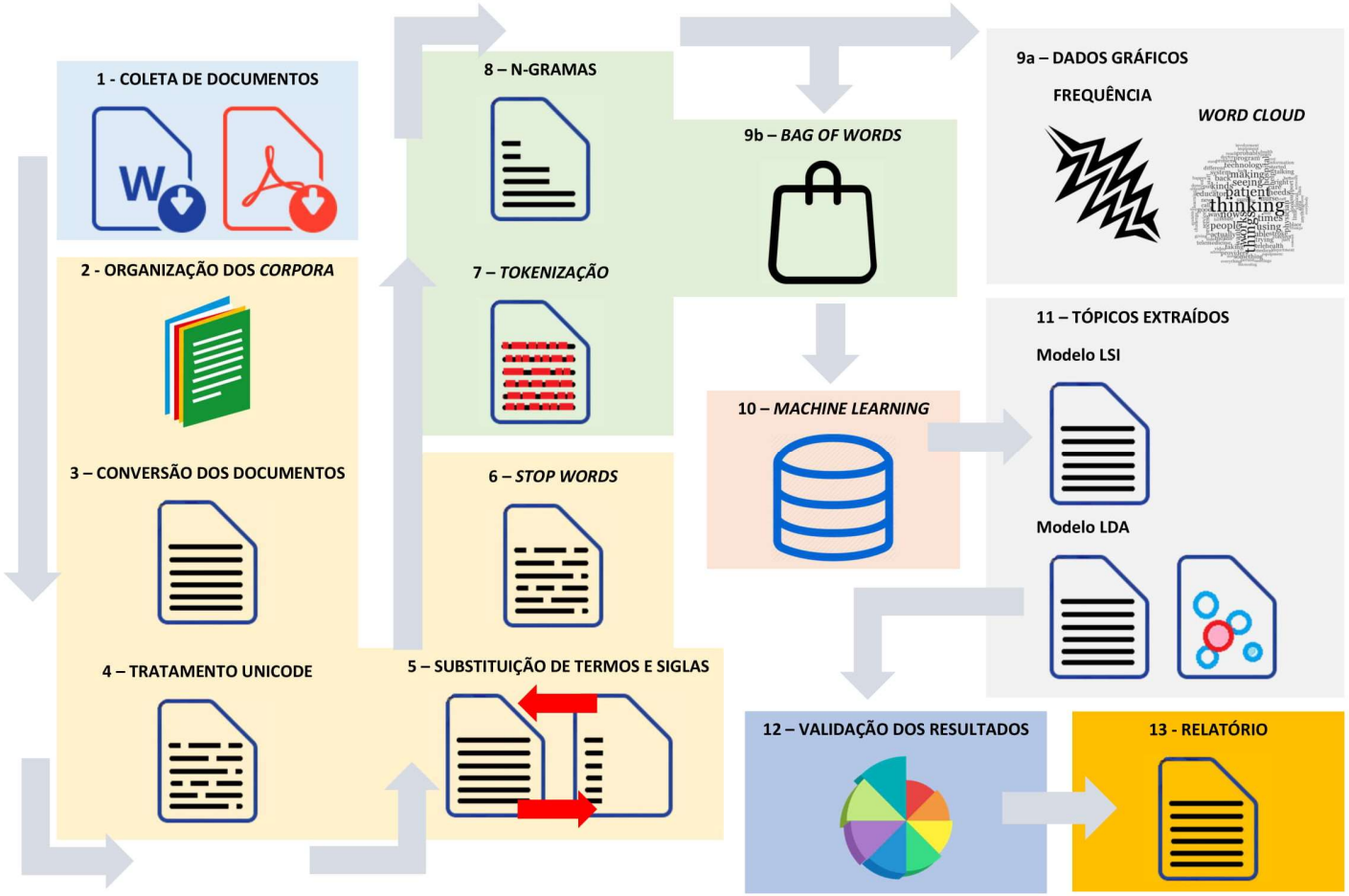
Trata-se da fase no qual o controle é consolidado nas fases de observação e a experimentação por meio do papel do pesquisador com a ciência na experiência e da evidência, experiência sensorial e a formação de ideias. O empirismo lógico é constituído pelos princípios do Empirismo e Logicismo, respectivamente, o conceito ou ideia terá significância na medida em que houver fundamento na experiência e que exista uma exata formulação da linguagem lógica para que de fato um sistema ou enunciado possa ter validade (CARVALHO, 2006).

A fase empírica da pesquisa foi constituída por meio das etapas adaptadas de Mckinney (2018) e ilustrada na Figura 05:

- **Interação com o mundo externo**: tange à coleta e constituição de *corpora* de dados, destacando os pontos de leitura e escrita a partir de uma variedade de formatos de arquivos e repositórios de dados;
- **Preparação e pré-processamento**: referente à organização, limpeza, manipulação, combinação, normalização, tratamento dos dados para realização da análise descritiva;
- **Transformação**: diz respeito às operações matemáticas e estatísticas aplicadas em grupos de conjuntos de dados a fim de obter resultados significativos;
- **Modelagem e processamento**: refere-se à conexão dos dados já tratados a modelos estatísticos, algoritmos de aprendizagem de máquina e outras ferramentas de processamento;
- **Apresentação**: diz respeito à criação de visualizações gráficas interativas ou estáticas, ou sínteses textuais;
- **Validação**: refere-se à validação dos resultados junto à comunidade científica do domínio da linguagem estudada; e

- **Documentação**: refere-se à construção da redação do relatório da pesquisa e disponibilização das codificações.

Figura 05 – Fluxo da modelagem de tópicos da pesquisa[10]



Fonte: Elaborado pelo autor.

[10] Fluxo de processamento de modelagem de tópicos adaptada de Mckinney (2018): 1 – Interação com o mundo externo; 2, 3, 4, 5, 6 – Preparação e pré-processamento dos dados; 7, 8, 9b – Transformação dos dados; 10 – Modelagem e processamento; 9b, 11 – Apresentação dos resultados; 12 – Validação dos resultados junto a comunidade científica; 13 – Elaboração da documentação e relatórios.

4.2.1. Descrição dos corpora de dados

Para a fase empírica da pesquisa, foram utilizados dois *corpora* de dados constituídos por documentos científicos extraídos da internet, sendo o primeiro *corpora* constituído por teses e dissertações dos programas brasileiros de pós-graduação na modalidade *stricto sensu* em Ciência da Informação entre o período de 2012 e 2017. O segundo *corpora* foi constituído por artigos completos e resumos expandidos publicados nos anais do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) entre 2012 e 2018.

4.2.1.1. Universo dos *corpora*

O primeiro *corpora* estabelece como universo os cursos de pós-graduação brasileiros em Ciência da Informação encontrados na Sucupira[11]. A Plataforma possui registrado o quantitativo de 24 universidades, 27 programas e 40 cursos em funcionamento entre mestrado acadêmico, mestrado profissional, doutorado acadêmico e doutorado profissional. Utilizou-se como amostragem para a coleta de dados os programas de pós-graduação da área de estudo vinculados à Associação Nacional de Pesquisa em Ciência da Informação (ANCIB)[12], sociedade civil sem fins lucrativos fundada em junho de 1989 com empenho de alguns cursos de pós-graduação em Ciência da Informação nos quais foram coletados os dados. A ANCIB tem como finalidade acompanhar e estimular as atividades de ensino de pós-graduação e de pesquisa da área - ENANCIB.

Fazem parte da ANCIB[13] o quantitativo de 27 instituições, sendo distribuídos em 6 diferentes cursos na modalidade acadêmicos e 7 cursos na modalidade profissionais, ambos na modalidade *stricto sensu*, conforme apresentado no Quadro 08:

---

[11] Plataforma Sucupira. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/quantitativos/quantitativoles.jsf?areaAvaliacao=31&areaConhecimento=60700009/. Acesso em: 13/03/2019.

[12] ANCIB - Associação Nacional de Pesquisa em Ciência da Informação. Disponível em: http://gcancib.eci.ufmg.br/. Acesso em: 13/03/2019.

[13] ANCIB - Associação Nacional de Pesquisa em Ciência da Informação – Programas de pós-graduação. Disponível em: https://www.ancib.org.br/menu-lateral/revistas-da-ancib/. Acesso em: 06/02/2020.

Quadro 08 – Universidades e programas na área da Ciência da Informação - ANCIB

Programas de Pós-Graduação Acadêmicos	
Universidade	**Programa**
Instituto Brasileiro de Informação em Ciência e Tecnologia/Universidade Federal do Rio de Janeiro (IBICT/UFRJ)	• Ciência da Informação (Mestrado/Doutorado)
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT)	• Informação e Comunicação em Saúde (Mestrado/Doutorado)
Universidade de Brasília (UNB)	• Ciência da Informação (Mestrado/Doutorado)
Universidade de São Paulo (USP)	• Ciência da Informação (Mestrado/Doutorado) • Museologia (Mestrado)
Universidade Estadual de Londrina (UEL)	• Ciência da Informação (Mestrado)
Universidade Estadual Paulista (UNESP)	• Ciência da Informação (Mestrado/Doutorado)
Universidade Federal da Bahia (UFBA)	• Ciência da Informação (Mestrado/Doutorado) • Museologia (Mestrado)
Universidade Federal da Paraíba (UFPB)	• Ciência da Informação (Mestrado/Doutorado)
Universidade Federal de Minas Gerais (UFMG)	• Ciência da Informação (Mestrado/Doutorado) • Gestão e Organização do Conhecimento (Mestrado/Doutorado)
Universidade Federal de Pernambuco (UFPE)	• Ciência da Informação (Mestrado/Doutorado)
Universidade Federal de Santa Catarina (UFSC)	• Ciência da Informação (Mestrado/Doutorado)
Universidade Federal de São Carlos (UFSCAR)	• Ciência da Informação (Mestrado)
Universidade Federal do Ceará (UFC)	• Ciência da Informação (Mestrado)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)	• Memória Social (Mestrado/Doutorado) • Museologia e Patrimônio (Mestrado/Doutorado)
Universidade Federal do Rio Grande do Sul (UFRGS)	• Comunicação e Informação (Mestrado/Doutorado)
Universidade Federal Fluminense (UFF)	• Ciência da Informação (Mestrado/Doutorado)
Programas de Pós-Graduação Profissionais	
Fundação Casa de Rui Barbosa (FCRB)	• Memória e Acervos (Mestrado)
Fundação Universidade Federal do Piauí (UFPI)	• Artes, Patrimônio e Museologia (Mestrado)

Museu de Astronomia e Ciências Afins (MAST)	• Preservação de Acervos de Ciência e Tecnologia (Mestrado)
Universidade do Estado de Santa Catarina (UDESC)	• Gestão da Informação (Mestrado)
Universidade Federal de Sergipe (UFS)	• Gestão da Informação e do Conhecimento (Mestrado)
Universidade Federal do Cariri (UFCa)	• Biblioteconomia (Mestrado)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)	• Biblioteconomia (Mestrado)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)	• Gestão de Documentos e Arquivos (Mestrado)
Universidade Federal do Rio Grande do Norte (UFRN)	• Gestão da Informação e do Conhecimento (Mestrado)

Fonte: ANCIB[14].

O segundo *corpora* estabelece como universo a produção do conhecimento científico registrada por meio de anais do ENANCIB[15]. Trata-se do principal evento da área de Ciência de Informação do Brasil voltado para troca de experiências acadêmico-cientifica entre pesquisadores, discutindo e refletindo ao longo dos anos diversos temas, perspectivas e tendências da área.

Tomou-se como base o projeto de pesquisa denominado “Questões em Rede”. Ele disponibiliza trabalhos e palestras apresentados ao longo das edições do ENANCIB, conforme apresentado no histórico de realização do evento, no Quadro 09. Na Base de Encontro Nacional de Pesquisa em Ciência da Informação (BENANCIB) é possível realizar navegação por comunidades e coleções, data do documento, autores, títulos, palavras-chave ou assunto. Faz-se necessário destacar que a BENANCIB[16] foi criada pelo grupo de pesquisa Informação, Discurso e Memória da Universidade Federal Fluminense (UFF), possuindo cadastro no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e parceria com a ANCIB. A BENANCIB é viabilizada por financiamento da Fundação Carlos Chagas de Amparo à Pesquisa do Rio de

[14] ANCIB – Lista de programas de Pós-Graduação na área da Ciência da Informação. Disponível em: https://web.archive.org/web/20190103110139/https://www.ancib.org.br/menu-lateral/revistas-da-ancib/. Acesso em 06/02/2020..

[15] ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação – sobre. Disponível em: http://www.enancib2019.ufsc.br/sobre/. Acesso em 06/02/2020.

[16] BENANCIB – Base de Encontro Nacional de Pesquisa em Ciência da Informação. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/2. Acesso em: 06/02/2020.

Janeiro (FAPERJ) e apoio técnico da Superintendência de Tecnologia da Informação da própria universidade.

Quadro 09 – Histórico de realização dos ENANCIB's

ENANCIB	Ano	Instituição	Cidade / Estado
I	1994	UFMG	Belo Horizonte – MG
II	1995	PUC[17]	Valinhos – SP
III	1997	IBICT[18]	Rio de Janeiro – RJ
IV	2000	UnB	Brasília – DF
V	2003	UFMG	Belo Horizonte - MG
VI	2005	UFSC	Florianópolis – SC
VII	2006	UNESP	Marília – SP
VIII	2007	UFBA	Salvador – BA
IX	2008	USP	São Paulo – SP
X	2009	UFPB	João Pessoa – PB
XI	2010	IBICT	Rio de Janeiro - RJ
XII	2011	UnB	Brasília – DF
XIII	2012	Fiocruz	Rio de Janeiro - RJ
XIV	2013	UFSC	Florianópolis – SC
XV	2014	UFMG	Belo Horizonte - MG
XVI	2015	UFPB	João Pessoa – PB
XVII	2016	UFBA	Salvador – BA
XVIII	2017	UNESP	Marília – SP
XIV	2018	UEL	Londrina – PR
XX	2019	UFSC	Florianópolis – SC

Fonte: Adaptado da BENANCIB.

O ENANCIB é promovido anualmente pela ANCIB desde 1994, desempenhando atividades como conferências proferidas por pesquisadores estrangeiros convidados, reuniões dos grupos temáticos e dos coordenadores dos programas de pós-graduação *stricto sensu* da área da Ciência da Informação e afins, reunião de estudantes de pós-graduação, entrega do Prêmio ANCIB às melhores dissertações e teses, lançamentos de livros, fóruns e debates por meio dos grupos de trabalhos, entre outras atividades.

As pesquisas cientificas realizadas por docentes, discentes e pesquisadores da área da Ciência da Informação e afins submetidas e aprovadas por meio de comitês científicos junto ao ENANCIB são apresentadas durante o evento por meio dos grupos de trabalhos e suas respectivas temáticas, conforme Quadro 10:

[17] Pontifica Universidade Católica de Campinas.
[18] Instituto Brasileiro de Informação em Ciência e Tecnologia.

Quadro 10 – Grupos de trabalhos do ENANCIB

Grupos de Trabalho	Nome do Grupo
GT 1	- Estudos Históricos e Epistemológicos da Ciência da Informação
GT 2	- Organização e Representação do Conhecimento
GT 3	- Mediação, Circulação e Apropriação da Informação
GT 4	- Gestão da Informação e do Conhecimento nas Organizações
GT 5	- Política e Economia da Informação
GT 6	- Informação, Educação e Trabalho
GT 7	- Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação
GT 8	- Informação e Tecnologia
GT 9	- Museu, Patrimônio e Informação
GT 10	- Informação e Memória
GT 11	- Informação & Saúde

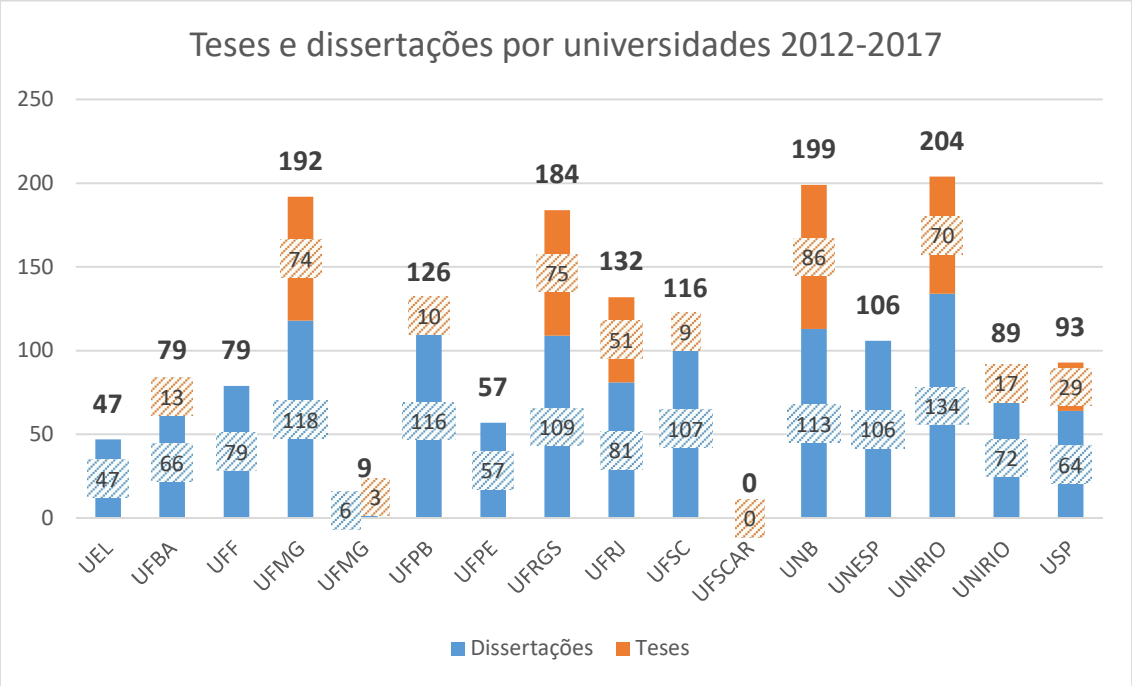
Fonte: ANCIB.

4.2.1.2. Amostragem dos *corpora*

A amostragem dos dados do primeiro *corpora*, constituído por documentos científicos do tipo tese e dissertação, foi coletada entre 27 de abril e 01 de maio de 2018 em um total de 14 universidades e 16 cursos, tendo como critérios os programas de pós-graduação da área de Ciência da Informação associados a ANCIB até a data da coleta dos dados.

Foram coletados para a formação dos *corpora* o quantitativo de 437 teses e 1275 dissertações publicados entre o período de 2012 e 2017 junto às bibliotecas digitais de teses e dissertações das respectivas instituições de ensino superior, totalizando 1712 documentos. O Gráfico 03 apresenta o quantitativo por universidades de teses e dissertações coletadas para a formação desses *corpora* de dados.

Gráfico 03 – Quantitativo de teses e dissertações defendidas por universidades



Fonte: Elaborado pelo autor.

Torna-se necessário ressaltar que o curso *stricto sensu* de mestrado acadêmico em Ciência da Informação da UFSCAR, até a data da coleta de dados, não possuía dissertações defendidas, uma vez que se trata de um curso novo, com primeira turma ingressante no segundo semestre de 2016. A UEL, por meio do curso de pós-graduação em Ciência da Informação, e a UFMG, com o curso de Gestão e Organização do Conhecimento, tiveram suas primeiras dissertações defendidas, respectivamente, em 2014 e 2016. Por isso, justifica-se o baixo quantitativo de trabalhos de conclusão de curso. Outros cursos também foram criados após a coleta de dados, tais como mestrado acadêmico em Ciência da Informação pelas Universidades Federais de Alagoas, do Espírito Santo e do Pará.

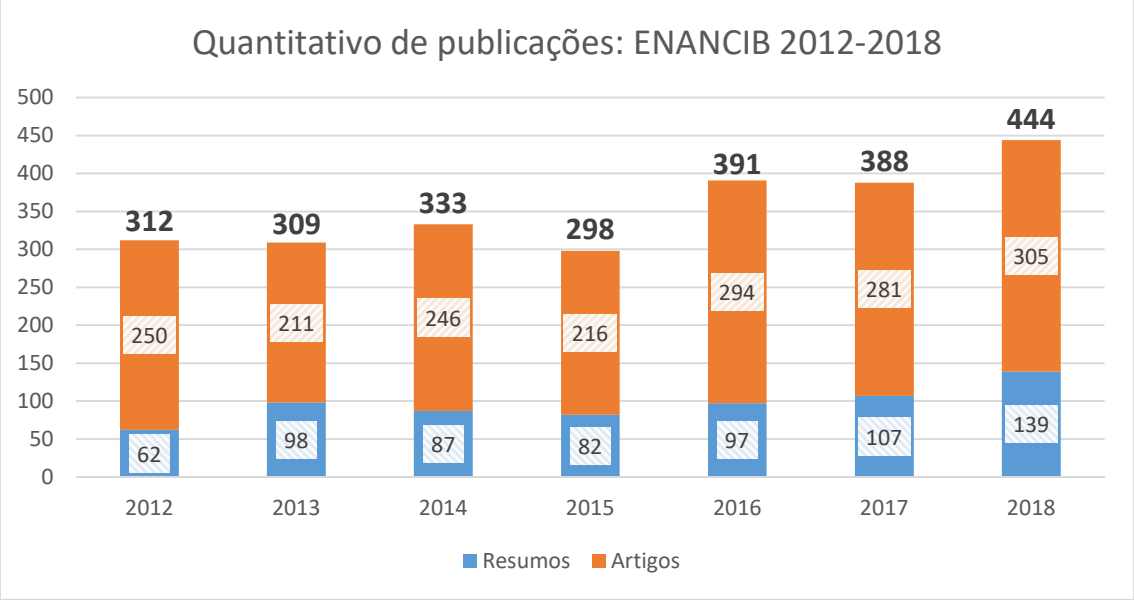
Referente ao quantitativo de teses e dissertações da amostragem selecionada, destaca-se a UNIRIO, com o Programa de Pós-graduação em Memória Social, com 204 trabalhos, sendo 70 teses e 134 dissertações; seguido da UnB, com 199 trabalhos, sendo 86 teses e 113 dissertações; UFMG, por meio do Programa de Ciência da Informação, com 192 trabalhos, sendo 74 teses e 119 dissertações; UFRGS, com 184 trabalhos, sendo 75 teses e 109 dissertações durante o período de coleta dos documentos. Os cursos da UEL,

UFF, UFPE e UNESP possuem, até a data da coleta, somente a modalidade de mestrado acadêmico.

O segundo *corpora* de documentos foi constituído por artigos completos (apresentações orais) e resumos expandidos (pôsteres), referente aos anais do ENANCIB, publicados entre 2012 e 2017. A coleta dos dados foi realizada entre 27 de abril e 01 de maio de 2018. Em 2019, foram acrescidas as publicações dos anais referentes ao XIV ENANCIB. Dessa forma, o *corpora* passou a ser constituído por documentos publicados entre os anos de 2012 a 2018.

Foram coletados para a formação do *corpus* o quantitativo de 1.803 artigos completos e 672 resumos expandidos, publicados entre 2012 e 2018 do evento junto a BENANCIB e também nas páginas dos eventos anuais do ENANCIBs, totalizando 2.475 documentos. O quantitativo de trabalhos científicos por modalidade e totalizadores publicados pelos anos da amostragem podem ser observados no Gráfico 04:

Gráfico 04 – Quantitativo anual de artigos completos e resumos expandidos publicados por anais do ENANCIB 2012-2018



Fonte: Elaborado pelo autor.

Com base nos números dos trabalhos apresentados nas modalidades de artigos completos e resumos expandidos publicados nos anais do ENANCIB entre 2012 e 2018, podem-se destacar alguns dados estatísticos do *corpora* a fim de apresentar um maior detalhamento sobre o cenário estudado, sendo eles as médias anuais de 248 artigos científicos, 92 resumos expandidos e de 322,5 trabalhos de ambas as modalidades.

Os anos em que tiveram uma menor produção científica aprovada para o ENANCIB na modalidade de artigos completos, resumos expandidos e em ambas as modalidades juntas foram, respectivamente: 2013, com 211 artigos completos; 2012, com 63 resumos expandidos; e 2015, com 298 produções científicas. Já o ano com a maior produção científica anual nas modalidades de artigos completos, resumos expandidos e ambas as modalidades foi 2018, com 305 artigos completos, 139 resumos expandidos e 444 trabalhos em ambas as modalidades.

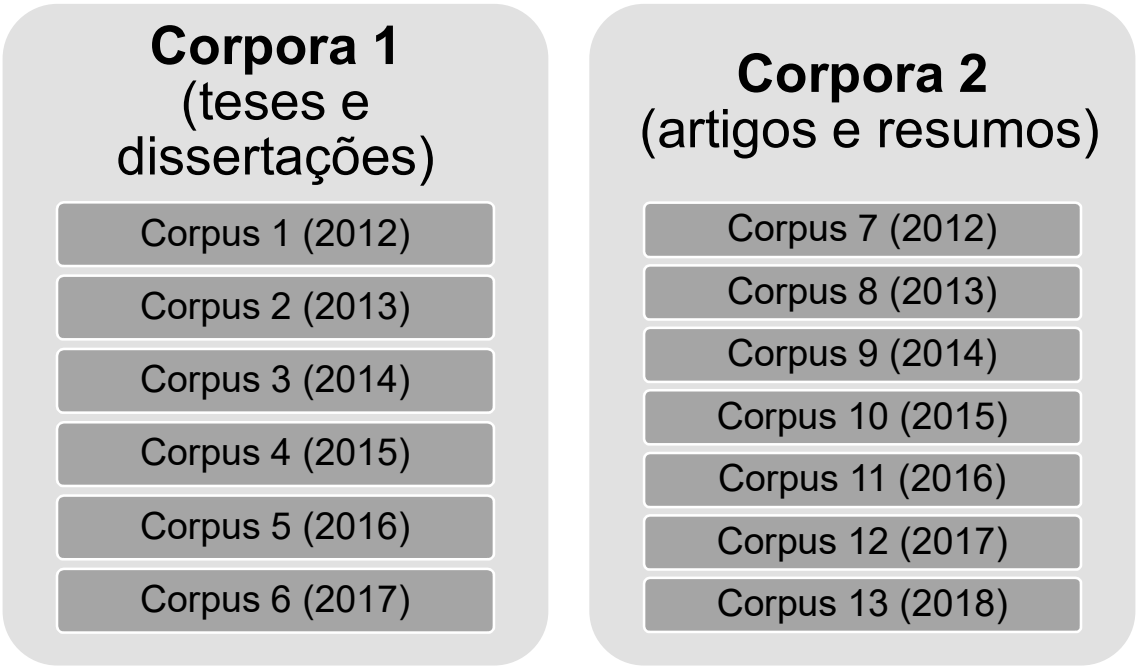
Também são destacados nesse *corpora* de dados os Grupos de Trabalho (GT) com menor e maior produção científica publicadas nos anais do ENANCIB durante o intervalo da amostragem. Para os artigos completos, o GT 11 publicou o menor quantitativo, com 8 trabalhos em 2016 e 2017. Já a maior produção científica dessa modalidade aconteceu no GT 02, com 50 trabalhos publicados nos anais em 2016. Para a modalidade de resumos expandidos, o GT 02 publicou somente um (1) trabalho apresentado em 2012, assim como o GT 09, em 2014, 2015 e 2017. A maior publicação dos resumos expandidos ficou para o GT 03, que publicou 27 resumos expandidos em 2017.

Foi necessário estabelecer o ponto de recorte da amostragem mediante a quantidade de documentos, tamanho dos *corpora* associados a questões de desempenho referente ao processamento dos modelos de extração de tópicos. Cada *corpora* de dados foi organizado em *corpus* – analisados separadamente, constituídos com os respectivos documentos do canal formal da comunicação científica organizados por ano de publicação, conforme apresentado na Figura 06.

Todos os documentos, organizados por *corpus* e utilizados na pesquisa empírica estão disponibilizados no Figshare[19] - repositório online de acesso aberto, sob a licença Creative Commons CC BY 4.0.

[19] Figshare – Corpora utilizado na pesquisa empírica. Disponível em: https://figshare.com/account/collections/5351453. Acesso em: 18/11/2019.

Figura 06 – Representação dos *corpora* de dados



Fonte: Elaborado pelo autor.

A próximo seção apresenta a descrição dos processos metodológicos e técnicos executados na fase empírica da pesquisa, destacando as etapas para a execução da modelagem de tópicos, realizada por meio dos modelos de extração de tópicos *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA).

4.3. Descrição dos procedimentos empíricos

Correspondem aos procedimentos empíricos que se iniciam desde a coleta de dados, descrição e utilização dos ambientes de produção e linguagem e bibliotecas, preparação dos documentos, pré-processamentos e transformação dos dados, execução dos modelos LSI e LDA até a visualização dos resultados.

4.3.1. Coleta de dados

A coleta dos dados para formação dos *corpora* foi realizada por meio de acesso direto aos documentos disponibilizados nas bases de teses e dissertações das universidades e das páginas dos respectivos programas de pós-graduação, Base de Encontro Nacional de Pesquisa em Ciência da Informação (BENANCIB) e nas páginas das edições do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB).

Foi utilizada a extensão do navegador Chrome chamada *Copy ALL URL*[20], que realiza cópia tabulada das *Uniform Resource Locator* (URL) abertas no navegador para a área de transferência. Em conjunto com essa extensão, foi utilizado o *software* gerenciador de *downloads* chamado *Download Accelerator Plus*[21] (DAP), que permitiu gerar os arquivos dos documentos e organizá-los em diretórios, formando dois *corpora* de dados.

Tanto a extensão de navegador utilizado para copiar as URLs quanto o *software* gerenciador de *downloads* estão disponíveis gratuitamente na internet para utilização. Faz-se necessário ressaltar que não foi desenvolvido nenhum código para extração dos documentos para compor os *corpora* de dados. O que de fato poderia ser interessante para a comunidade científica, esbarrou no entrave da falta de padronização entre as bases de dados, levando em consideração, por exemplo, a variedade de estruturas, diferentes linguagens e a falta de interoperabilidade entre os ambientes/*frameworks* de armazenamento de documentos eletrônicos.

Entre os programas de pós-graduação dos cursos utilizados na amostragem, pôde-se perceber uma variação entre os sistemas próprios com diferentes características utilizadas pelas universidades. Já entre os anais do ENANCIB, a maioria das edições disponibilizam os documentos por meio do sistema de publicação e gerenciamento de periódicos *Open Journal Systems* (OJS), entretanto, há edições que disponibilizam anais do evento no formato PDF único, gerando mais esforços na etapa de preparação dos documentos.

Mediante o cenário apresentado, seria necessário acrescer de esforços e tempo para desenvolver diferentes sistemas de extração de documentos e com pouca utilização para esta tese. Além disso, o acesso manual para cópia das URLs e *downloads* dos documentos permitiu identificar diversos *hiperlinks* quebrados, que foram informados aos programas responsáveis e posteriormente acertados.

4.3.2. Ambiente de desenvolvimento

[20] Copy ALL URL. Disponível em: https://chrome.google.com/webstore/detail/copy-all-urls/djdmadneanknadilpjiknlnanaolmbfk?hl=pt-BR/. Acesso em: 18/11/2019.
[21] Download Accelerator Plus. Disponível em: https://download-accelerator-plus.br.uptodown.com/. Acesso em: 18/11/2019.

O *Integrated Development Environment* (IDE) – Ambiente de desenvolvimento integrado escolhido para o desenvolvimento da modelagem de tópicos foi o *framework* Anaconda – *Jupyter Notebook*[22], a linguagem de programação Python 3.6[23] e as bibliotecas *Pdfminer*[24], *Gensim*[25], *NLTK*[26], *Numpy*[27], *Matplotlib*[28], *Plotly*[29] e *pyLDAvis*[30], além de bibliotecas básicas, mas fundamentais para o desenvolvimento do código.

A escolha do *framework* se deu por apresentar um ambiente de desenvolvimento iterativo, interface amigável, possibilidade para se conectar a mais de 40 tipos de linguagens de programação e por ser construída sobre algumas bibliotecas *open source*. Já a opção pela linguagem de programação *Python*, além de ser ideal para Ciência de Dados, tem se destacado na comunidade de programadores, tornando-se a terceira[31] linguagem de programação mais utilizada do mundo.

Todos os códigos foram desenvolvidos utilizando o notebook, tendo como sua configuração o processador Core i7 - 2630QM 2.00GHz e memória de 8GB, entretanto, o equipamento foi insuficiente ao executar a modelagem de tópicos, além de apresentar uma lentidão nos processos. Posteriormente, utilizou-se o servidor virtual da Fundação Getúlio Vargas (FGV) com a configuração de 24 processadores virtuais i7 2.7GHz e memória de no máximo de 64GB de RAM, porém, possui um quantitativo representativo de processos, sendo executados simultaneamente, o que resultou em resultados inferiores à primeira tentativa utilizando o notebook. Por fim, foi utilizado outro notebook contendo a configuração de processador Intel® Core (TM) i7-8750H CPU @ 2.20GHz e memória de 16GB, que apresentou resultados de processamento, em alguns casos, superior a 1000% quando comparado ao primeiro notebook. Mesmo apresentando resultados superiores, o equipamento não foi capaz de executar

[22] Jupyter Notebook. Disponível em: https://jupyter.org/. Acesso em: 18/11/2019.
[23] Python 3.6. Disponível em: https://www.python.org/downloads/release/python-360/. Acesso em: 18/11/2019.
[24] PFDMiner. Disponível em: https://pypi.org/project/pdfminer/. Acesso em: 18/11/2019.
[25] Gensim. Disponível em: https://pypi.org/project/gensim/. Acesso em: 18/11/2019.
[26] NLTK. Disponível em: https://pypi.org/project/nltk/. Acesso em: 18/11/2019.
[27] Numpy. Disponível em: https://pypi.org/project/numpy/. Acesso em: 18/11/2019.
[28] Matplotlib. Disponível em: https://pypi.org/project/matplotlib/. Acesso em: 18/11/2019.
[29] Plotly. Disponível em: https://pypi.org/project/plotly/. Acesso em: 18/11/2019.
[30] pyLDAvis. Disponível em: https://pypi.org/project/pyLDAvis2/. Acesso em: 18/11/2019.
[31] TIOBE – The software quality company. Disponível em: https://www.tiobe.com/tiobe-index/. Acesso em: 18/11/2019.

grandes *corpora* de dados, sendo necessário adequar a organização dos documentos de forma que fosse executado em blocos.

Tendo como base o equipamento utilizado para processamento, a metodologia e detalhes como tamanho dos *corpora* de dados, foi possível apresentar, dentre os resultados, o fator tempo de processamento, que serve apenas como ponto norteador para pesquisadores e/ou desenvolvedores.

4.3.3. Preparação e pré-processamento

A próxima etapa após a coleta de dados refere-se à organização dos diretórios para o recebimento dos documentos coletados. Foram criados dois diretórios, sendo um para receber os arquivos das teses e dissertações e o outro para receber os arquivos dos artigos completos e resumos expandidos, de acordo com a amostragem estabelecida. Os diretórios foram organizados respectivamente com o quantitativo de seis e sete subdiretórios, sendo cada um deles destinado para receber os documentos referentes ao ano pertinente. Foram organizados um total de 4.187 arquivos no formato PDF com mais de 9GB de tamanho de dados. Entende-se cada subdiretório como um *corpus* e os dois conjuntos de seis ou sete subdiretórios como *corpora*.

A próxima etapa está na conversão dos arquivos contidos nos *corpora* de maneira que os dados possam ser interpretados e processados pelo computador. Para isso, foram desenvolvidos dois códigos. O primeiro código[32] lista os arquivos encontrados no diretório e subdiretório apontado e já organizado. O código gera um arquivo TXT com o caminho de todos os arquivos encontrados. Já o segundo código[33] realiza a leitura do arquivo gerado pelo primeiro código, localizando todos os arquivos no formato PDF do *corpus* e realizando a conversão dos textos para um único arquivo no formato TXT, um padrão interpretável pelo computador. Os caracteres do documento TXT foram padronizados para caixa baixa. Para essa etapa foi utilizada a biblioteca chamada Pdfminer que realiza a extração dos textos na íntegra contidos nos arquivos PDF. Cada texto convertido e que compõe cada um dos *corpus* de

[32] Código para listar arquivos em diretórios e subdiretórios. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/01_listar_arquivos.ipynb/.
[33] Código para leitura, extração e junção de texto em documentos PDF para um único documento TXT. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/02_leitura_arquivos.ipynb/.

documentos fica alocado em uma das linhas no documento TXT separados por colchetes. Durante a conversão, o programa apresenta algumas informações como os nomes dos documentos, o número de páginas de cada documento do *corpus* e o total de documentos convertidos. Os 4.187 arquivos no formato PDF foram convertidos para 13 arquivos no formato TXT com tamanho de 656MB.

Após a conversão dos documentos, cada arquivo que constitui os *corpora* de dados foi analisado com o objetivo de identificar e excluir caracteres Unicode indesejáveis gerados durante o processo de conversão dos documentos PDF para TXT. Alguns exemplos de caracteres Unicode excluídos dos *corpora* são “\x0c”, “uf0b7”, “x0c6011”, “\uf0fc”, “\x0c5701”, “\x0c7”, gerados ao converter imagens contidas nos documentos ou arquivos criptografados.

Com os *corpora* de dados limpos, foi desenvolvido o terceiro algoritmo,[34] que contempla as fases de pré-processamento, transformação, modelagem e processamento e apresentação dos resultados.

Após a importação das bibliotecas, foi utilizada a função *Regular Expression* (RE) – Expressão Regular por meio da função “re.sub” para padronizar uma lista de termos da linguagem de domínio utilizados encontradas nos *corpora* de dados. Essa função permite substituir termos em siglas, siglas em termos ou termos em termos. Tal prática possibilita potencializar o peso dos termos com o mesmo significado. Por exemplo, “eci” e “escola de ciência da informação” possuem o mesmo significado e, quando convertidas todas as siglas para termos ou todos os termos para siglas, os pesos passam a não ser divididos entre eles. Uma boa prática está na conversão de termos logos para siglas como “fundação brasileira à pesquisa do estudo de minas gerais” para “fapemig”. O apêndice M apresenta uma lista contendo 153 termos que foram substituidos por outros termos do domínio da linguagem. Dessa forma, consegue-se obter melhores resultados para os N-gramas bem como calibração dos pesos dos termos calculados pelos próprios modelos. Uma boa prática também está na unificação de termos com o mesmo significado, como apresentados em diferentes tipos de linguagem natural e que podem ser apresentadas em seções diferentes de um documento, como *abstract* e o corpo do texto (SOUZA; SOUZA, 2019).

[34] Código da modelagem de tópicos. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2012.ipynb/.

A próxima etapa foi setar[35] as *stop words* por meio da biblioteca utilizando a lista padrão da biblioteca NLTK. A biblioteca possui uma lista padrão de palavras de parada. O uso dessa técnica permite realizar uma limpeza das palavras irrelevantes como “as”, “e”, “os”, “de”, “para”, “com”, “sem”, “foi”, “aquela”. Mesmo sendo *corpora* constituídos por documentos redigidos em língua portuguesa, foram configurados também as *stop words* nos idiomas inglês e espanhol por apresentarem textos como resumos ou termos encontrados nesses dois idiomas. Além disso, foi adicionada uma lista contendo 130 novas palavras de parada que potencializam o processamento e também melhora os resultados nos N-gramas e, consequentemente, para os especialistas realizarem a análise de assunto com base nos diferentes tipos de resultados gerados a partir da Modelagem de Tópicos, vide apêndice N.

As técnicas de lematização – que busca encontrar a palavra em seu lema – e *Stemming* – que busca fatiar o início ou o fim das palavras com a intenção de remover os afixos – não foram utilizadas no processo de limpeza e tratamento dos dados. No primeiro caso, ainda não existe uma biblioteca para a linguagem natural utilizada nos *corpora* de dados e palavras com significados diferentes como “biblioteca”, “bibliotecário” e “biblioteconomia” seriam convertidos para “biblio” somando todos as frequênias. Já o segundo caso, por se tratar de *corpora* com grandes volumes de informações, acaba por exigir uma demanda de tempo para análise de forma que palavras com significados diferentes não sejam unificadas. Além disso, o *Stemming* pode não apresentar bons *tokens*.

Posteriormente, o *corpus* é carregado e lido após cumpridas as etapas de limpeza, manipulação e combinação. O carregamento do *corpus* é realizado por meio de apontamento do arquivo no diretório e a leitura é realizada para a memória do computador, a partir da substituição da lista de termos setada anteriormente. Após a leitura do arquivo, o algoritmo informa o quantitativo de documentos contidos no *corpus*.

Ainda na fase de preparação, foram setadas as funções de criação N-gramas (WANG; MCCALLUM; WEI, 2007). Posteriormente, foi realizada a normalização e limpeza do *corpus* por meio da*s* *stop words* e uso das expressões regulares. Anterior à *tokenização*, todas as *stop words* setadas e

[35] Neologismo derivado do verbo inglês "To Set" que significa "Definir".

adicionais foram excluídas do *corpus*, assim como todos os caracteres especiais e numéricos, por meio da função RE “re.match”. Além disso, foram criados os unigramas, bigramas e trigramas do conteúdo do *corpus*, utilizados na fase de transformação para obtenção de novos resultados. Ao final é apresentado o quantitativo de cada N-grama.

4.3.4. Transformação

No que diz respeito às operações matemáticas e estatísticas aplicadas em grupos de conjuntos de dados e a fim de obter resultados significativos, a primeira etapa da fase de transformação foi a criação da frequência dos termos contidos no *corpus*. A *tokenização* dos textos foi realizada por meio da biblioteca NLKT utilizando o *word_tokenize*. Dessa forma, os textos do *corpus* foram divididos em palavras, símbolos e em outros elementos convertidos em uma sequência de palavras separadas por meio do tratamento de pontuação e espaçamento. Quando a linguagem escrita é armazenada em um documento de computador, passa a ser representada por sequência de *strings* de caracteres. Esses caracteres podem ser alfanuméricos, caracteres especiais que representam os espaços, tabulações ou mesmo indicativo de uma nova linha, entre outras possibilidades.

A função N-gramas definida na etapa anterior e executada nessa fase, cria, tanto na memória do computador quanto em arquivo externo no formato CSV, uma lista de frequência contendo os 1.000 primeiros termos por categorias de N-gramas, sendo unigramas, bigramas, trigramas e geral (N-gramas). Esses dados possibilitam criar análises de comportamento dos termos quando se comparado a um determinado espaço de tempo. Paralelo a esse algoritmo e com os resultados dos N-gramas, foi desenvolvido um algoritmo[36] que apresenta no formato de gráfico dinâmico a evolução de uma seleção de termos realizados por frequência e relevância junto ao domínio de linguagem dos dois *corpora*. Posteriormente, foram gerados para cada um dos 13 *corpus*, dois gráficos, com um deles contendo os 50 termos mais frequentes e o outro no formato de nuvem de palavras, contendo os 250 termos do *corpus*.

[36] Gráfico de frequência de termos. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/04_grafico_freq_teses_dissertacoes.ipynb/.

Outra etapa de suma importância dessa fase está na criação do dicionário que utilizou a biblioteca Gensim para definir um *corpus* de *bag of words* – saco de palavras. Nesse processo, os textos contidos no *corpus* são convertidos em uma representação matricial, desconsiderando a estrutura gramatical e até mesmo a ordenação delas, mas mantendo sua multiplicidade. Cada palavra recebe um identificador único, representado no documento como uma lista de palavras e apresenta o número de ocorrências de cada palavra contida no *corpus*. Após essa etapa, os dados estão prontos para serem conectados aos modelos.

4.3.5. Modelagem e processamento

Nessa etapa são apresentados os procedimentos para realização da modelagem de tópicos a partir da utilização dos modelos de extração de tópicos LSI e LDA, bem como das definições e configurações dos quantitativos de tópicos generativos, treinados nos dois *corpora* constituídos de *corpus*, organizados e analisados separadamente por cada ano de referência. Cada modelo apresenta um conjunto de nove resultados para cada *corpus*/ano, contendo um número de tópicos, termos e pesos que foram configurados para extrair um determinado quantitativo de tópicos, sendo eles: 1 – 10 tópicos; 2 – 14 tópicos; 3 – 18 tópicos; 4 – 22 tópicos; 5 – 26 tópicos; 6 – 30 tópicos; 7 – 34 tópicos; 8 – 38 tópicos; e 9 – 42 tópicos. Dessa forma, cada *corpus* treinado gerou o quantitativo de 18 modelos de resultados entre LSI e LDA por *corpus*/ano, conforme apresentado nos Quadros 12 e 13.

Quadro 12 – Quantitativo de análises do *corpora* teses e dissertações

Corpora	**TESES E DISSERTAÇÕES**											
Corpus	2012 (1)		2013 (2)		2014 (3)		2015 (4)		2016 (5)		2017 (6)	
Modelos	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA
Resultados por modelos	9	9	9	9	9	9	9	9	9	9	9	9
Resultados por *corpus*	18		18		18		18		18		18	
Resultados por *corpora*	**108**											

Fonte: Elaborado pelo autor.

O *corpora* de documentos do canal formal da comunicação científica contendo teses e dissertações resultou em 108 modelos, analisadas com o objetivo de encontrar a melhor representação de cada corpus/ano.

Quadro 13 – Quantitativo de análises do *corpora* artigos completos e resumos expandidos

Corpora	**ARTIGOS E RESUMOS**													
Corpus	2012 (7)		2013 (8)		2014 (9)		2015 (10)		2016 (11)		2017 (12)		2018 (13)	
Modelos	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA	LSI	LDA
Resultados por modelos	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Resultados por *corpus*	18		18		18		18		18		18		18	
Resultados por *corpora*	**126**													

Fonte: Elaborado pelo autor.

Já o *corpora* de documentos contendo artigos científicos e resumos expandidos resultou um total de 126 modelos, em que também se buscou analisar o quantitativo de tópicos que melhor representasse o *corpus*/ano.

4.3.5.1. Execução do modelo LSI

A execução do modelo LSI utiliza a técnica de SVD para elaboração da matriz de documentos. Por meio da biblioteca *Gensim*, foi utilizado um algoritmo randômico e interativo que perpassou pelas seguintes etapas: i) elaborar uma matriz de documentos a partir dos documentos que constituíram os *corpora* de dados; ii) realizar a decomposição da matriz obtida; iii) escolher *n* componentes principais; iv) utilizar uma métrica de semelhança para identificar os documentos mais semelhantes. Foi utilizada a configuração padrão do modelo LSI para realizar a modelagem de tópicos, bem como a definição do quantitativo de tópicos a serem extraídos com o parâmetro *num_topics* = K.

4.3.5.2. Execução do modelo LDA

A técnica de execução do modelo LDA utiliza a redução de dimensionalidade. Também busca preservar as características mais importantes do *corpus* de dados e descartar o que não é importante. O modelo LDA utilizou de métodos mais sofisticados em termos de matemática e de estatística com vários algoritmos de processamento de linguagem natural. Além da quantidade de tópicos a serem extraídos, foram configurados os seguintes parâmetros: *num_topics* = K, referente ao número de tópicos a serem extraídos; em *chunksize* = 1.000, que se refere ao número de documentos a serem usados em cada bloco de treinamento; *passes* = 40, referente ao número de passagens de treinamento pelos documentos; e *iterations* = 600, referente ao número máximo de iterações no *corpus* ao inferir a distribuição de tópico de um *corpus*.

Ambos os modelos foram configurados os *idx* = 10, referente ao número de palavras por tópico. As palavras são representadas por N-gramas, podendo

ser unigrama, bigrama ou trigrama. Para o *corpora* contendo os *corpus* de dados com os documentos de artigos completos e resumos expandidos foi utilizada a biblioteca *pyLDAvis*, configurada com o quantitativo de 12 tópicos para gerar uma visualização dinâmica dos resultados. A biblioteca permite gerar uma visão global dos tópicos plotados em formato de círculos em um plano bidimensional, sendo os os centros determinados pelo cálculo da distância entre tópicos. Posteriormente é usado o dimensionamento multidimensional para projetar a distância intertópica em das dimensões (CHUANG; MANNING; HEER, 2012).

4.3.6. Apresentação dos resultados

Tanto para as etapas que antecedem a conexão com os modelos de extração de dados LSI e LDA, quanto para a pós-execução dos modelos, apresentam resultados que possibilitam diversas discussões acerca da temática abordada. Dentre os resultados, são destacados para cada *corpus* analisado: i) lista no formato de texto dos N-gramas mais frequentes; ii) grafíco de frequência dos 50 N-gramas mais frequentes; iii) nuvem de palavras dos 250 N-gramas mais frequentes; iv) tópicos textuais dos modelos LSI e LDA e para cada *corpus* de artigos científicos e resumos expandidos, utilizou-se a biblioteca pyLDAvis para gerar visualizações dinâmicas dos tópicos; e, para a etapa que antecede a modelagem de tópicos; e v) gráfico dinâmico que apresenta o comportamento dos principais termos obtidos por meio da frequência dos *corpora* de dados quando se comparado a um período de tempo. Para seleção dos dados, foi utilizada a técnica de observação comportamental e importância dos termos frente ao domínio de linguagem.

Todos os algoritmos desenvolvidos, bem como os resultados de todos os *corpora* de dados, estão disponibilizados para a comunidade científica por meio do *Github*[37]. Outra apresentação e discussão de resultados está na validação dos resultados do melhor modelo de extração de tópicos junto à comunidade científica da área da Ciência da Informação.

4.3.7. Validação dos resultados

[37] Github. Plataforma de desenvolvimento colaborativo para hospedar, revisar códigos, gerenciar projetos e criar software de maneira colaborativa. Algoritmos desenvolvidos e utilizados na tese e resultados disponíveis em: https://github.com/marcosdesouza82/topic-model-tese/.

Com base nos resultados dos modelos de extração de tópicos, foram selecionados 30 conjuntos de tópicos constituídos por termos e pesos dos *corpora* de dados, sendo 15 para o *corpora* de teses e dissertações entre o *corpus* de 2012 a 2017 e 15 para o *corpora* de artigos completos e resumos expandidos entre o *corpus* de 2012 a 2018, apenas do modelo que apresentou os melhores resultados. Utilizou-se como critério de qualidade para seleção dos tópicos a coesão entre os termos, bem como pesos com representatividade ao domínio de linguagem. São considerados tópicos fortes aqueles que possuem um conjunto de termos coesos que apontam para 1 único documento.

Foi criado um formulário para validação dos resultados junto à comunidade científica da área de Ciência da Informação representada por professores, pesquisadores, bibliotecários e profissionais. O questionário[38] foi constituído por três seções: i) identificação do perfil respondente; ii) conjuntos de tópicos com os seus respectivos termos e pesos extraídos do *corpora* teses e dissertações; e iii) conjuntos de tópicos também com os seus respectivos termos e pesos extraídos do *corpora* de artigos científicos e resumos expandidos. As seções 2 e 3 possuem 11 opções de rótulos para associação a cada conjunto de termos. Os rótulos foram extraídos dos nomes dos grupos de trabalhos do ENANCIB, principal evento da área da Ciência da Informação do Brasil. Além disso, consta, para cada conjunto de termos, um *link* de um documento para consulta que melhor representa determinado tópico. O *link* foi disponibilizado por meio do encurtador de URL *Bitly*[39], que permite ter acesso ao número de vezes que o documento foi consultado.

O formulário foi construído por meio do ambiente virtual *Google Forms*[40] e o convite foi realizado por e-mail aos para professores dos programas de pós-graduação utilizados na amostragem da pesquisa, bem como para os diretores e coordenadores das bibliotecas universitárias das referidas instituições com o adendo para socializarem o convite aos demais bibliotecários da instituição. Os e-mails foram extraídos das páginas dos programas de pós-graduação e das

[38] Questionário de validação da pesquisa. Disponível em: https://forms.gle/KX45xXYs5Wh5bjLRA
[39] Bitly. Encurtador de URL, links personalizados e gerenciamento de links. Disponível em: https://bitly.com/.
[40] Google Forms. Aplicativo de administração de pesquisas que permite coletar informações de usuários por meio de uma pesquisa ou questionário personalizado. Disponível em: https://www.google.com/forms/about/.

bibliotecas das universidades que constituíram a amostragem da pesquisa, sendo convidados 333 professores e 12 bibliotecas. O formulário ficou disponibilizado entre 30 de março e 15 de abril de 2020, alcançando a participação de 88 respondetes.

O formulário foi colocado em pré-teste entre 23 e 27 de março de 2020 com o objetivo de receber sugestões de melhorias antes de ser enviado para que o público específico realizasse a validação dos dados. Os critérios utilizados para seleção dos avaliadores foram baseados em suas respectivas formações e atuações junto à comunidade científica. As considerações realizadas pelos avaliadores, bem como os respectivos perfis acadêmicos, estão disponibilizadas junto ao apêndice A.

4.3.8. Documentação.

Elaboração da documentação final da pesquisa de tese realizada por meio dos resultados alcançados nas fases da pesquisa empírica em consonância com o referêncial teórico.

5. RESULTADOS E DISCUSSÃO

O capítulo apresenta os resultados alcançados e uma discussão frente aos referenciais teórico e empírico, sendo dividido em 3 seções. A primeira seção destaca os resultados e a discussão referente aos termos extraídos dos *corpora* de dados, a fim de realizar uma análise diacrônica a partir da etapa de pré-processamento. A segunda seção aborda a proximidade e o distanciamento entre as disciplinas do núcleo da Ciência da Informação encontrados na literatura com os resultados extraídos dos *corpora* de dados. Por fim, a última seção apresenta as etapas de transformação, modelagem e processamento e apresentação dos resultados dos *corpus* de dados, além do processo de validação dos resultados junto à comunidade científica da área da Ciência da Informação a partir dos resultados do melhor modelo de extração.

5.1. Comportamento diacrônico dos termos dos *corpora* de dados

Os dois *corpora* de dados constituídos por documentos do canal formal da comunicação científica – o primeiro por teses e dissertações e o segundo por artigos completos e resumos expandidos – apresentam uma lista de termos em que é possível analisar, por meio de gráficos dinâmicos, seus respectivos comportamentos ao longo dos anos analisados. Constam, dentre os resultados apresentados, termos generalistas – que podem se compor em diversos outros significados – e termos especialistas – que apontam um único significado dentro domínio da linguagem analisada. O primeiro *corpora* de dados possui um conjunto de seis *corpus* analisados separadamente e extraídos os termos e suas respectivas frequências, sendo cada *corpus* organizado por documentos de um determinado ano. O segundo *corpora* apresenta as mesmas características, entretanto, possui sete *corpus* de dados. Os gráficos dinâmicos foram construídos com base na lista de N-gramas extraídos do *corpus* de dados pelo algoritmo, constando cada lista de mil termos mais frequentes de cada ano.

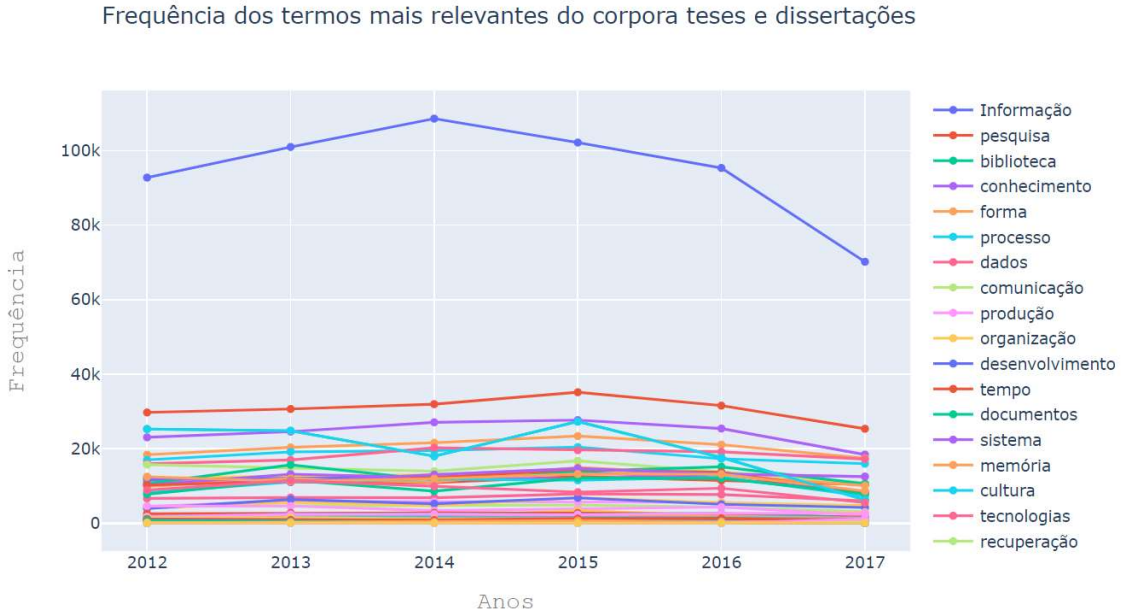
Foram extraídos do primeiro *corpora* de dados um total de 80 termos e suas respectivas frequências dos documentos do tipo teses e dissertações

publicadas entre 2012 e 2017, representados por meio de gráfico dinâmico[41]. Esse gráfico permite realizar diferentes combinações e análises entre os termos e períodos. Para o segundo *corpora* de dados foram extraídos um total de 84 termos e suas respectivas frequências dos documentos do tipo artigos completos e resumos expandidos publicadas entre 2012 e 2018, representados por meio de gráfico dinâmico[42], que também permite realizar diferentes combinações e análises entre os termos e períodos.

5.1.1. Termos mais frequentes dos *corpora* teses e dissertações

O Gráfico 05 apresenta uma visão geral do comportamento dos termos selecionados por meio do número de frequência e/ou relevância do termo junto ao domínio de linguagem.

Gráfico 05 – Frequência geral de comportamento de termos: teses e dissertações



Fonte: Elaborado pelo autor.

O Gráfico 06 apresenta o comportamento dos termos generalistas que constituem a tríade “dado”, “informação” e “conhecimento” extraídos do primeiro

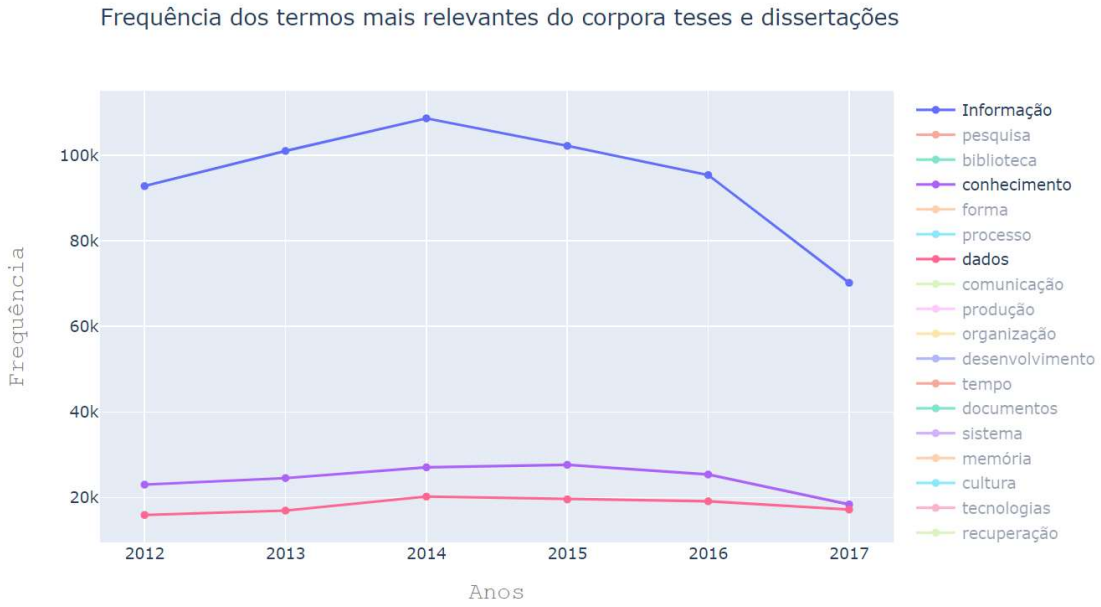
[41] Gráfico dinâmico da frequência dos termos do *corpora* de dados de teses e dissertações. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/04_grafico_freq_teses_dissertacoes.html/.

[42] Gráfico dinâmico da frequência dos termos do *corpora* de dados de artigos completos e resumos expandidos. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/04_grafico_freq_artigos_resumos.html/.

corpora de dados. O termo “dado” (representado pela linha lilás e localizado na parte inferior do gráfico) apresentou frequência de 15.928 em 2012. Em 2013, o termo apresentou um aumento de 6% referente ao ano anterior, atingindo a frequência de 16.953. Em 2014 é possível perceber um aumento de 19%, em que a frequência resultou em 20.225. Já em 2015 ocorreu a primeira queda do termo, em -3%, resultando em uma frequência de 19.571. A queda de frequência se manteve nos próximos dois anos, sendo 2016 com frequência de 19.137 ou -2% e 2017 com frequência de 17.191 ou -10% referente ao ano anterior. O termo “informação” (representado pela cor azul e localizado ao topo do gráfico) apresentou, em 2012, uma frequência de 92.776 e um aumento nos dois anos seguintes, sendo 2013 com frequência de 100.997 ou 9% e 2014 com frequência de 108.609 ou 8% referente ao ano anterior.

Já nos anos seguintes houve uma queda na frequência dos termos, sendo 2015 com frequência de 102.184 ou -6%, 2016 com frequência de 95.368 ou -7% e 2017 com frequência de 70.181 ou -26%. Já o termo “conhecimento” (representado pela cor roxa e localizado ao centro do gráfico) apresentou frequência de 23.025 em 2012 e um aumento no quantitativo de termos nos próximos três anos, sendo 2013 com frequência 24.558 de ou 7%, 2014 com frequência de 27.061 ou 10% e 2015 com frequência de 27.656 ou 2%. Já os anos seguintes apresentam queda de -8% em 2015, obtendo frequência de 25.394 e -28% em 2017 com frequência de 18.395.

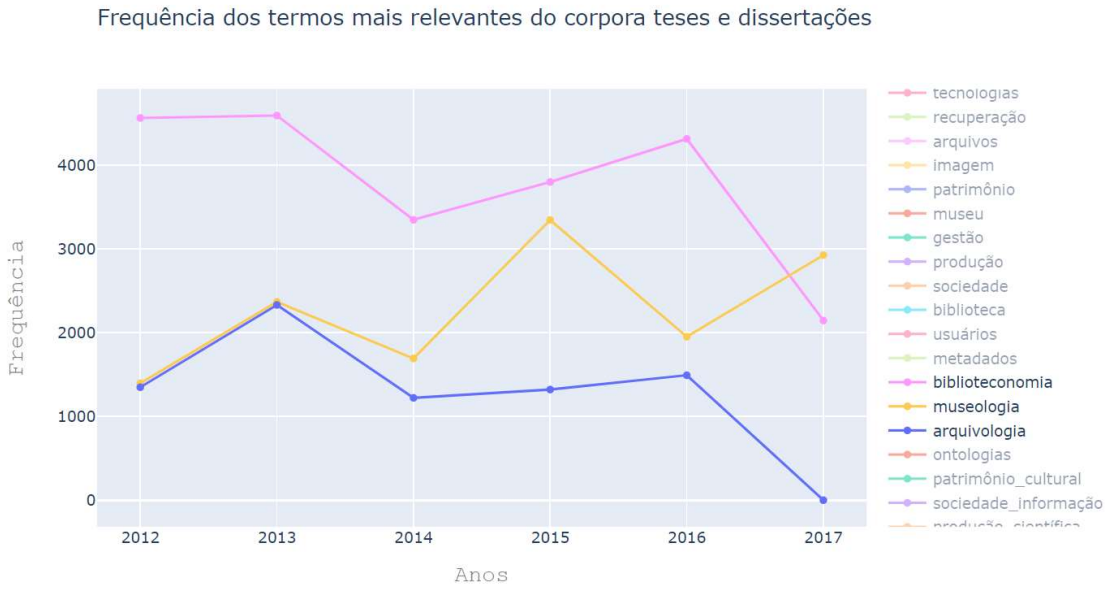
Gráfico 06 – Frequência de termos: teses e dissertações



Fonte: Elaborado pelo autor.

As quedas no percentual de frequência entre 2015 e 2016 não estão associados aos números de documentos contidos nesses *curpus* de dados, uma vez que esse anos são equivalentes ou superiores ao quantitativo de documentos dos *corpus* de 2012, 2013 e 2014. Pode ser considerado um indicativo referente à queda desse percentual o aumento de pesquisas contendo termos específicos ou mesmo sendo de cunho prático e menos teórico. Já com relação a 2017, a queda no número de frequência se justifica pelo período de coleta de dados, em que nem todas as teses e dissertações estavam publicadas em seus respectivos repositórios. Outra tríade a ser destacada está no Gráfico 07, com os termos “arquivologia”, “biblioteconomia” e “museologia”.

Gráfico 07 – Frequência de termos: teses e dissertações



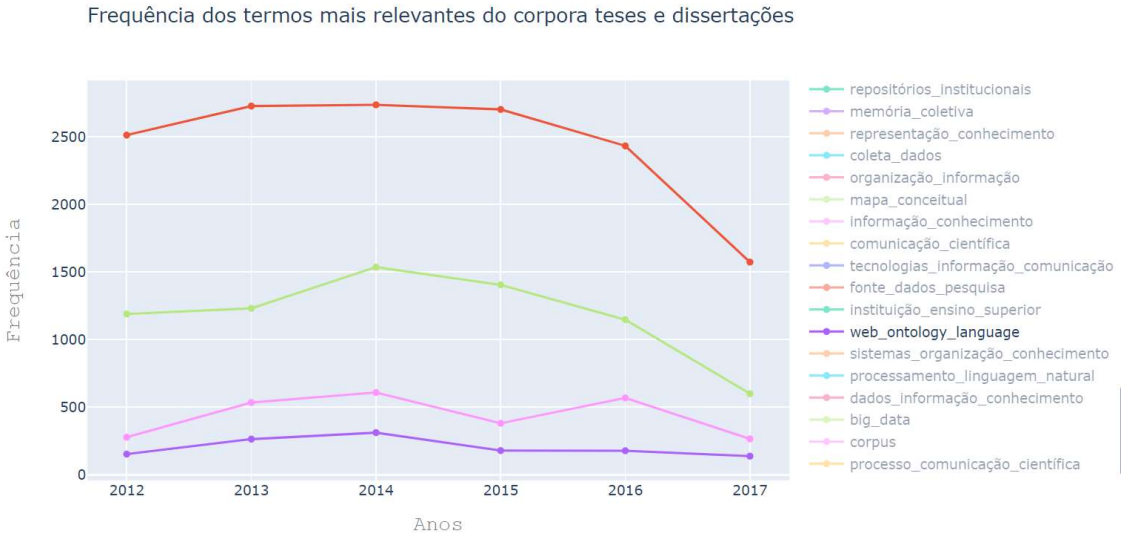
Fonte: Elaborado pelo autor.

É possível observar no gráfico de comportamento uma proximidade entre os termos “arquivologia” (representado pela linha azul) e “museologia” (representado pela linha amarela), especificamente em 2012 e 2013, que apresentam uma diferença de frequência de 3% e 2%, respectivamente. Ao contrário dessa proximidade, é possivel observar um distanciamento entre os termos em 2017. Nesse ano, alcança-se uma diferença superior a 100% entre os termos, uma vez que para o termo museologia apresenta frequência de 2.927, enquanto o termo arquivologia aparece somente após o milésimo termo mais frequente do *corpus* de dados, representado com o valor igual a 0.

Outra proximidade aparece em 2015 entre os termos museologia e biblioteconomia, representados pela cor lilás, no topo do gráfico. A diferença entre os termos é de uma frequência igual a 454 ou 14%. Também é possível observar uma inversão de posição entre esses mesmos termos em 2017, em que biblioteconomia apresenta uma queda de -50% referente ao ano anterior, com uma frequência de 2.136, enquanto o termo museologia tem um aumento de 50%, também referente ao ano anterior, obtendo uma frequência de 2.927.

O Gráfico 08 apresenta o comportamento dos termos “recuperação_informação” (representado pela cor vermelha, no topo do gráfico), “arquitetura_informação” (representado pela cor verde, no centro dos resultados), “web_semântica” (representado pela cor rosa) e “web_ontology_language” (representado pela cor lilás).

Gráfico 08 – Frequência de termos: teses e dissertações



Fonte: Elaborado pelo autor.

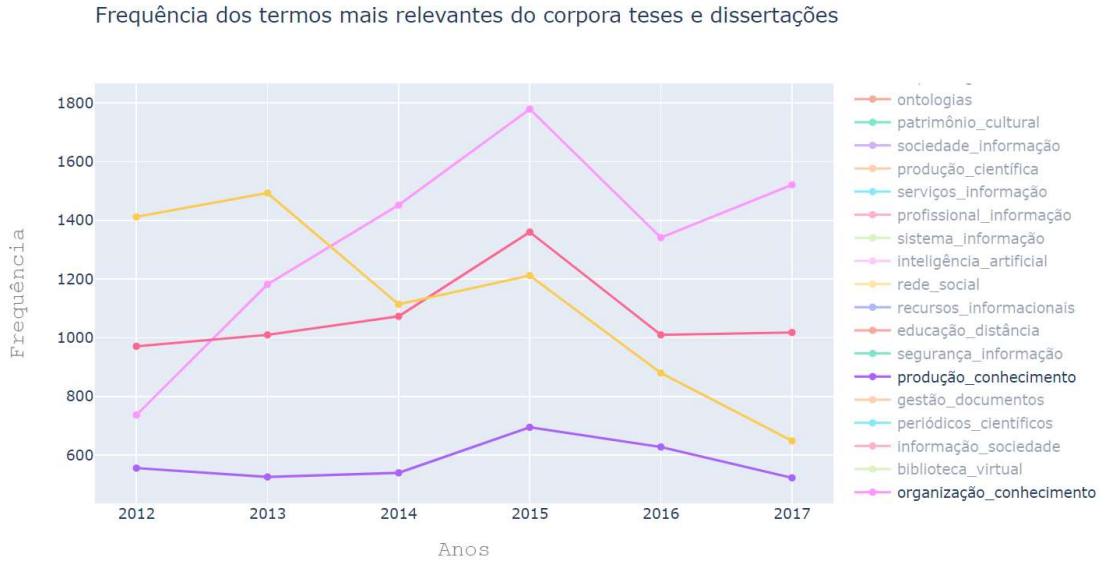
O termo “recuperaçao_informação” apresenta um equilíbrio em sua frequência entre os anos analisados, sendo possível observar uma frequência crescente entre 2012 e 2014 e um valor decrescente nos anos seguintes. Em 2014 o termo resultou em seu maior número de frequência, sendo encontrado 2.737 vezes no *corpus* de dados, enquanto no em 2016 obteve frequência de 2.433, representando uma queda de -10%. Com um comportamento similar ao termo “recuperação_informação”, o termo “arquitetura_informação” resultou em sua maior frequência em 2014, atingindo 1.536 repetições, enquanto em 2016 obteve a sua segunda menor frequência, com 1.147 repetições encontradas no *corpus* de dados, resultado em uma diferença de 25%.

Os termos “web_semântica” e “web_ontology_language”, localizados na parte inferior do gráfico, apresentam comportamentos similares entre 2012 e 2015 e diferentes entre 2016 e 2017. O termo “web_semântica” apresenta frequência de 277 em 2012 e um aumento nos anos seguintes de 93%, com frequência de 534 para 2013, 14% e frequência de 608 para 2014. Em 2015 o termo apresenta frequência de 380, equivalente a uma queda de -38%. O termo “web_ontology_language” apresenta frequência de 152 em 2012 e também um aumento nos anos seguintes, sendo frequência de 262 ou 73% referente a 2013 e 311 ou 18% em 2014. Já em 2015, o termo apresenta uma redução de -43%, obtendo frequência de 178. A diferença no comportamento entre os dois termos ocorre em 2016: o termo “web_semântica” se destaca com mais evidência enquanto o termo “web_ontology_language” se mantém na mesma linha de

frequência em relação ao ano anterior, atingindo uma diferença de 221% entre os termos.

Já o Gráfico 09 apresenta o comportamento de frequência dos termos “organização_conhecimento” representado pela linha rosa, “organização_informação” representado pela cor lilás, “comunicação_científica” representado pela linha amarela e “produção_conhecimento” representado pela linha roxa”.

Gráfico 09 – Frequência de termos: teses e dissertações



Fonte: Elaborado pelo autor.

O termo “organização_conhecimento” apresentou uma frequência crescente entre 2012 a 2015, iniciando com frequência de 737 e atingindo em seu maior volume com o quantitativo de 1.779, representando um aumento de 141%. Em 2016 o termo apresentou queda (-25%), resultando em uma frequência de 1.341. Já em 2017, mesmo havendo um volume menor de documentos no *corpus* de dados, o termo esteve em evidência e apresentou um valor crescente de 13% referente ao ano anterior, atingindo a frequência de 1.521. O termo “organização_conhecimento” apresentou frequência abaixo que o termo “organização_informação” somente em 2012. Além disso, é possível observar um comportamento similar entre esses dois termos, porém, em escalas diferentes. O termo “organização_conhecimento” iniciou 32%, atrás do termo “organização_informação”, e terminou 49% a frente do mesmo termo.

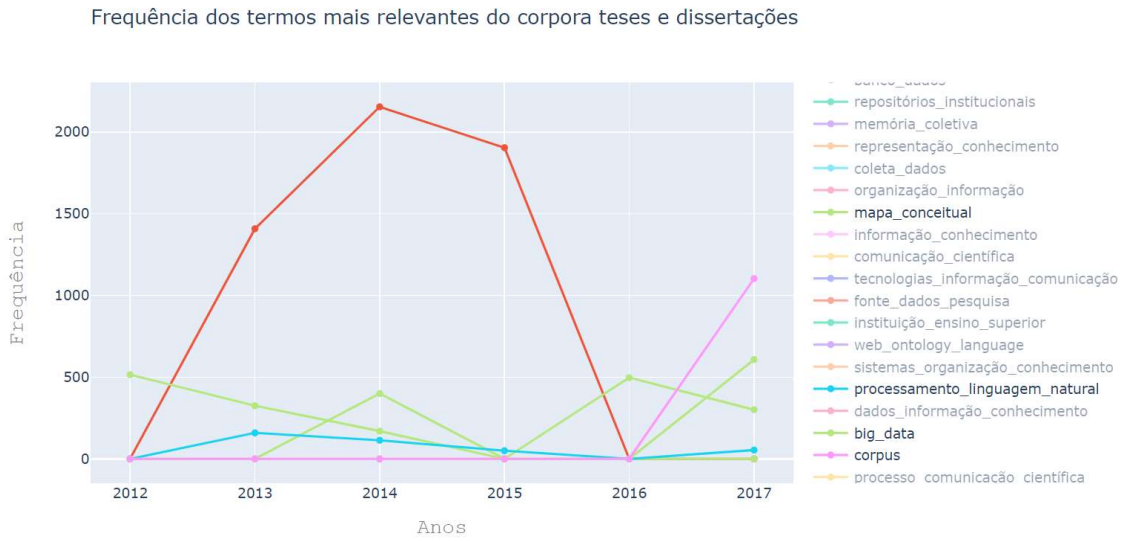
Dentre os termos apresentados no gráfico, o termo “comunicação_científica” apresentou maior frequência quando se comparado

apenas com 2012, resultando em um quantitativo 1.412 contra 971 de “organização_informação”, 737 de “organização_conhecimento” e 556 de “produção_conhecimento”. Em 2013 o termo apresentou um aumento de 6%, atingindo frequência de 1.493. No ano seguinte, o termo apresentou queda de -25%, atingindo frequência de 1.114. Já 2015 apresentou um aumento de 9%, obtendo frequência de 1.212. Em 2016 e 2017, o termo apresentou quedas de -27% e -26%, respectivamente, alcançando, no acumulado do primeiro ao último ano analisado, uma queda de -52%.

O termo “produção_conhecimento” apresentou frequências próximas entre os intervalos analisados, sendo o comportamento representado em alguns momentos com aumento e em outros com queda de frequência. Em 2012 o termo apresentou frequência de 556 e em 2017 frequência de 523, representando uma queda de -6% ao longo dos anos. Dentre os quatro termos analisados no gráfico, foi possível perceber que o termo “organização_conhecimento” ficou em evidência entre as pesquisas científicas realizadas, bem como a redução no quantitativo de pesquisas relacionadas à “comunicação_científica”.

Já o Gráfico 10 apresenta os termos que surgiram ou caíram em desuso ao longo do período analisado, de acordo com a frequência obtida. Constam os termos “ontologias” (destacado na cor vermelha), “biblioteca_virtual” (representado pela cor verde), com o maior valor em 2012, “mapa_conceitual” (também representado pela cor verde), com destaque de maior valor em 2014, “processamento_linguagem_natural” (representado pela linha azul), “*big_data*” (representado pela cor verde), obtendo maior destaque em 2017 e “*corpus*” (representado pela cor rosa).

Gráfico 10 – Frequência de termos: teses e dissertações



Fonte: Elaborado pelo autor.

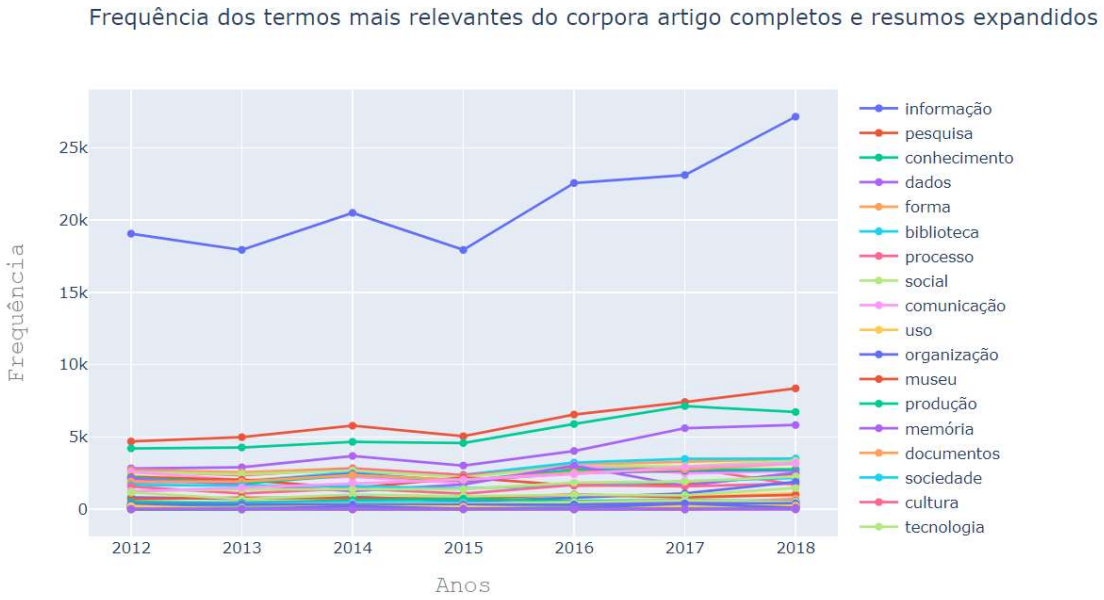
O termo “ontologias” apresentou frequência entre 2013 e 2015, sendo o primeiro ano com uma frequência de 1.408 e em 2014 com um aumento de 53% e frequência de 2.154. Já em 2015 o termo apresentou queda de -12%, totalizando uma frequência de 1.904. O termo “biblioteca_virtual” apresentou frequência de termos nos três primeiros anos analisados, sendo 512 para 2012, 325 para 2013 e 170 para 2014. É possível observar uma queda de -67% na frequência do termo entre 2012 a 2014. O termo “processamento_linguagem_natural” apresenta ausência de frequência entre 2012 e 2016. Em 2013 o termo apresenta frequência de 159 e em 2014 uma queda de -28%, resultando em uma frequência de 114. 2017 apresenta a menor frequência do termo, com 54 repetições no *corpus* de dados.

Os termos “*big_data*” e “*corpus*” ficaram em evidência em 2017, com frequências, respectivamente, de 608 e 649 extraídas do *corpus* de dados. Ambos os termos não apresentaram frequências ou apresentaram frequências após o milésimo termo nos *corpus* de dados anteriores, entretanto, faz-se necessário destacar que o *corpus* de dados referente a 2017 possui um menor quantitativo de documentos do canal formal da comunicação científica do tipo teses e dissertações quando se comparado aos *corpus* dos anos anteriores.

5.1.2. Termos mais frequentes dos *corpora* artigos completos e resumos expandidos

O Gráfico 11 apresenta uma visão geral do comportamento dos termos, selecionados por meio do número de frequência e/ou relevância junto ao domínio de linguagem.

Gráfico 11 – Frequência geral de comportamento de termos: artigos e resumos expandidos



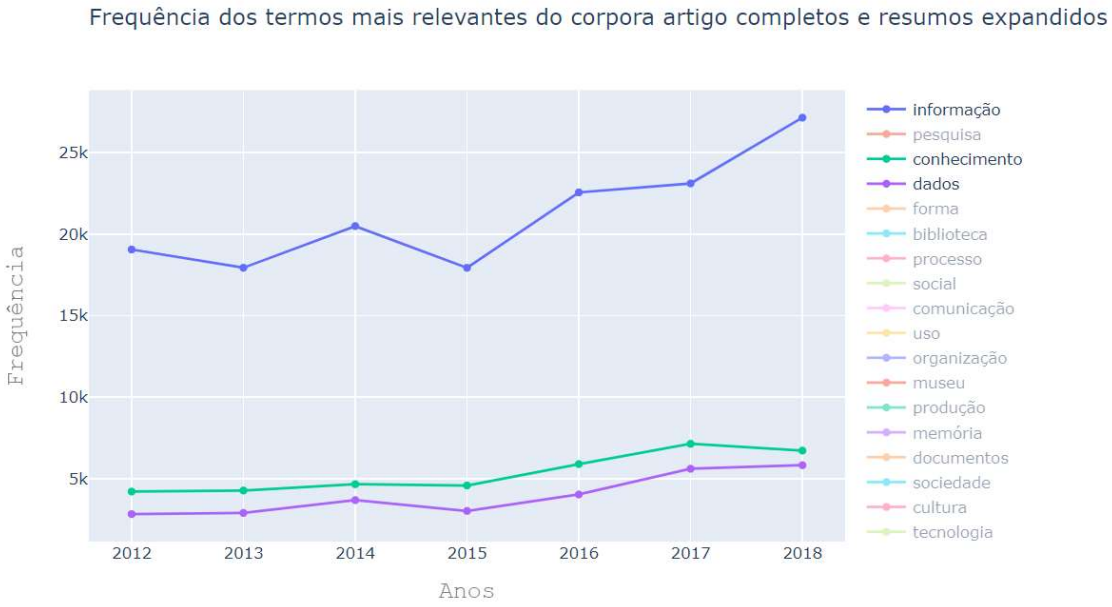
Fonte: Elaborado pelo autor.

O Gráfico 12 apresenta o comportamento dos termos generalistas “dado”, “informação” e “conhecimento”, que constituem uma das tríades da Ciência da Informação. O termo “informação” (representado pela linha azul) apresentou um comportamento inconstante entre 2012 a 2016, com alternância entre valores altos e baixos de frequência. A partir de 2016 o termo apresentou um comportamento de forma crescente até 2018, sendo 2016 com frequência de 22.559 e no ano seguinte com um aumento de 2% e frequência de 23.111. Já em 2018 apresentou um aumento de 17% referente ao ano anterior, resultando em uma frequência de 27.146.

Já o termo “conhecimento” (representado pela cor verde) apresenta em 2012 uma frequência de 4.207 e um valor crescente até 2014, atingindo frequência de 4.657 ou aumento de 11%. Em 2015 o termo apresentou uma queda na frequência de -2%, resultando no valor de 4.580. Os anos seguintes foram representados por um aumento de frequência do termo, atingindo, em

2018, uma frequência de 27.146 ou 42% quando se comparado a 2012. O termo “dado” (representado pela cor roxa) apresentou comportamento similar ao do termo “conhecimento”, entretanto, o valor acumulado ao longo do período analisado atingiu 106% mediante o crescimento de termos especialistas evidenciados na pesquisa científica, como “base_dados”, ”coleta_dados”, “fonte_dados_pesquisa” e “ciclo_vida_dados”.

Gráfico 12 – Frequência de termos: artigos e resumos expandidos



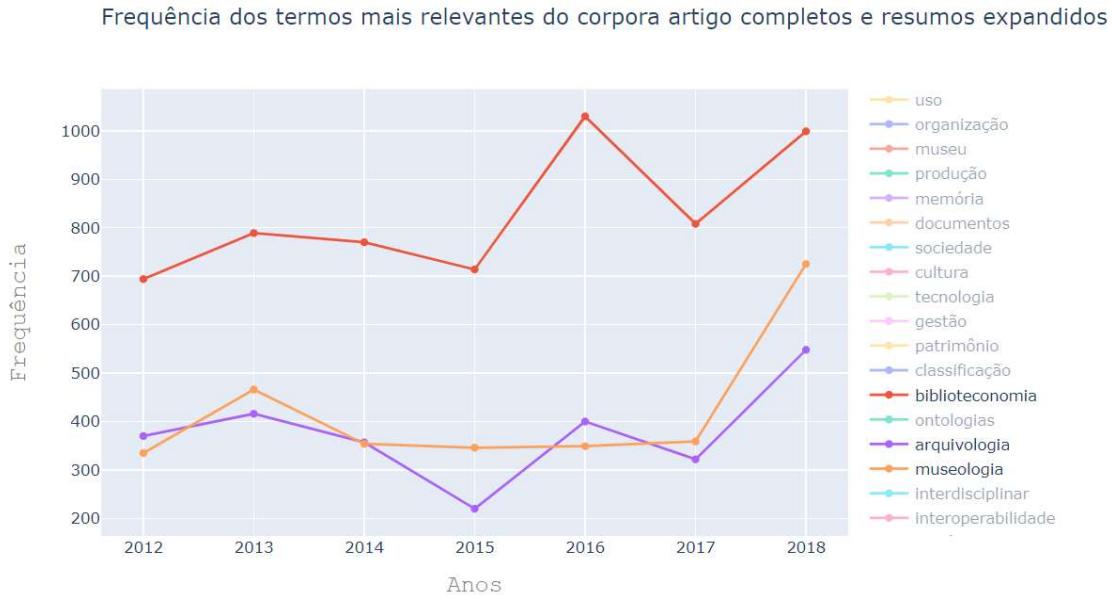
Fonte: Elaborado pelo autor.

Quando comparado o comportamento da tríade “dado”, “informação” e “conhecimento” entre os *corpora* de dados de teses e dissertações e artigos completos e resumos expandidos, torna-se possível perceber que os termos “dado” e “conhecimento” apresentam uma proximidade durante o período analisado. Já o termo “informação” possui um distanciamento na frequência de termos dos dois outros termos em ambos os *corpora* de dados. Uma diferença evidenciada está no termo “informação”, que apresenta caminhos inversos entre os *corpora* de dados. Enquanto o *corpora* contendo os documentos científicos do tipo artigos completos e resumos expandidos apresenta um aumento de 42% entre os anos analisados, o *corpora* de dados contendo os documentos do tipo teses e dissertações apresenta uma queda de -24%, também durante o período analisado.

O Gráfico 13 apresenta o comportamento dos termos “arquivologia” (representado pela cor roxa), “biblioteconomia” (representado pela cor vermelha)

e “museologia” (representado pela cor marrom), sendo a tríade dos cursos de graduação em Ciência da Informação. É possível perceber uma alternância de comportamento entre os termos “arquivologia” e “museologia” durante o período de 2012 e 2017 e um crescimento superior de 102% do termo “museologia”, contra 70% do termo “arquivologia” em 2018. Já o termo “biblioteconomia” possui resultados mais evidenciados e distantes dos termos “arquivologia” e “museologia”. Quando comparado a maior frequência de cada termo, pode-se identificar em 2016 o termo “biblioteconomia” com frequência de 1.030, enquanto “arquivologia” e “museologia” apresentam maior frequência em 2018, respectivamente 548 e 725, equivalentes a uma diferença de 88% e 42% para o primeiro termo.

Gráfico 13 – Frequência de termos: artigos e resumos expandidos



Fonte: Elaborado pelo autor.

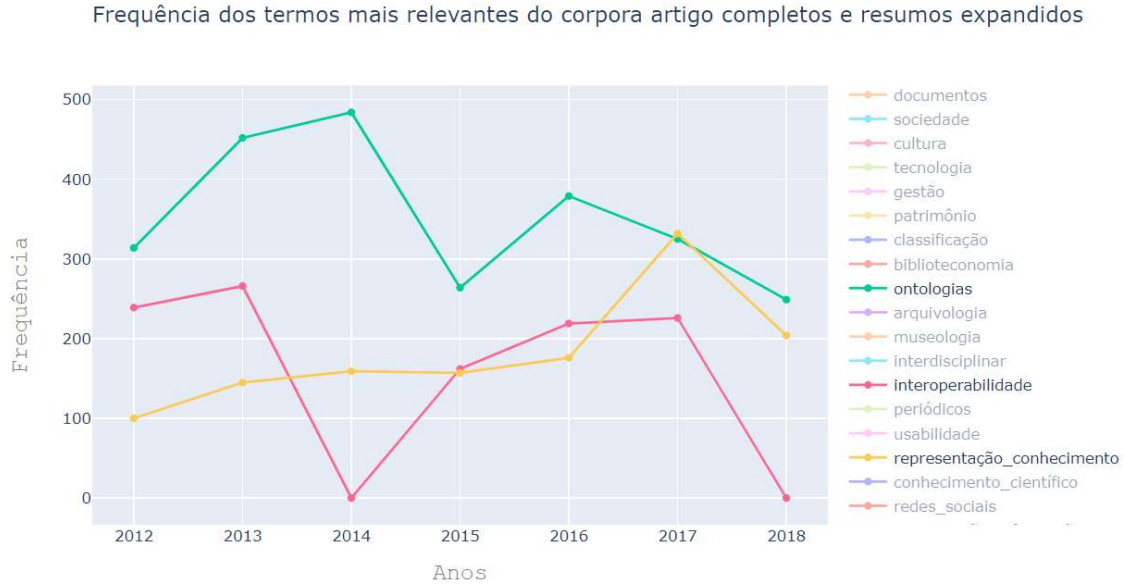
Quando comparado o comportamento dos resultados da tríade “arquivologia”, “biblioteconomia” e “museologia” entre os dois *corpora* de dados, sendo o primeiro constituído por documentos de teses e dissertações e o segundo por artigos completos e resumos expandidos, é possível afirmar que o termo “arquivologia” possui um menor quantitativo de pesquisas em ambos os *corpora* de dados, entretanto, enquanto no primeiro *corpora* apresenta um aumento de 11% até 2016, o segundo *corpora* de dados apresenta um crescimento superior de 48% ao longo dos anos analisados. Já o termo “museologia” teve o seu maior número de frequência no primeiro *corpora* de

dados durante 2015, enquanto no segundo *corpora* o seu melhor resultado foi em 2018. O termo “biblioteconomia” apresentou comportamento inverso entre os *corpora* de dados, enquanto o *corpora* de teses e dissertações apresentou queda na frequência de termos de -53% durante o intervalo analisado, no *corpora* constituído por artigos completos e resumos expandidos apresentou alta de 44%.

Faz necessário enfatizar que os temos “arquivologia”, “biblioteconomia” e “muselogia” não representam unicamente os seus respectivos cursos, podendo por exemplo representar a origem da pesquisa ou o departamento ao qual o pesquisador está alocado.

O Gráfico 14 apresenta o comportamento dos termos “ontologias” (destacado na cor verde), “interoperabilidade” (destacado pela cor rosa) e “representação_conhecimento” (destacado na cor amarela). O termo “ontologias” apresentou frequência de 314 em 2012 e dois aumentos consecutivos, sendo 2013 com 44% e frequência de 452 e 2014 com 7% e frequência de 484. Em 2015 o termo apresentou queda de -45% quando comparado ao ano anterior, atingindo a frequência de 264. Em 2016 o termo apresentou alta de 44%, atingindo a frequência de 379 e, nos anos seguintes, duas quedas, sendo 2017 de -14% e 2018 de -23%, resultando em uma frequência de 249 – o menor número extraído ao longo dos anos analisados.

Gráfico 14 – Frequência de termos: artigos e resumos expandidos

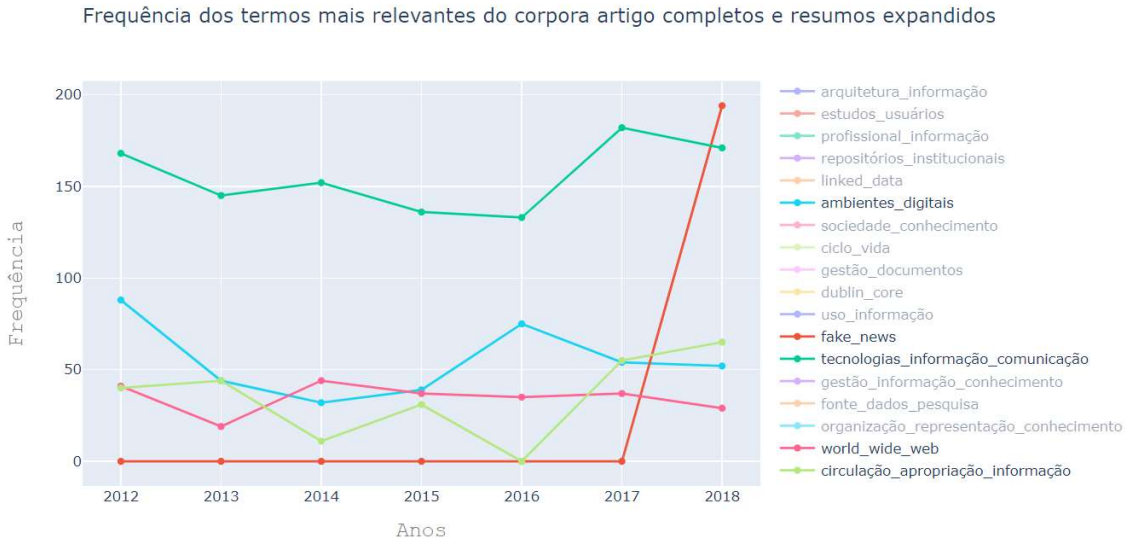


Fonte: Elaborado pelo autor.

O termo “interoperabilidade” apresentou um comportamento próximo ao do termo “ontologias”, entretanto, quando representada a queda de frequência, pode-se encontrar valores igual a zero pelo fato do termo não aparecer entre o milésimo termo na lista de unigramas. Entre 2012, com frequência de 239, a 2017, com frequência de 226, é possível identificar uma queda de -5%. Já o termo “representação_conhecimento” apresentou um comportamento crescente e de forma constante entre 2012 e 2016. O termo resultou na frequência de 176 em 2016 e um aumento de 89% em 2017, atingindo a maior frequência entre todos os termos apresentados no gráfico, com valor de 332. No ano seguinte, o termo apresentou uma queda de 39%, resultando em uma frequência de 204.

Já o Gráfico 15 apresenta o comportamento dos termos “tecnologia_informação_comunicação” (representado pela linha da cor verde escuro), “ambientes_digitais” (representado pela cor azul), “*word_wide_web*” (cor rosa), “circulação_apropriação_informação” (cor verde claro) e “*fake_news*” (vermelha). Entre o intervalo analisado, o termo “tecnologia_informação_comunicação” apresentou uma variação em seu comportamento com alguns momentos, obtendo baixas e altas frequências, entretanto, quando analisado somente o ano de 2012 e 2018, percebe-se um crescimento de 2% ou frequência igual a 3.

Gráfico 15 – Frequência de termos: artigos e resumos expandidos



Fonte: Elaborado pelo autor.

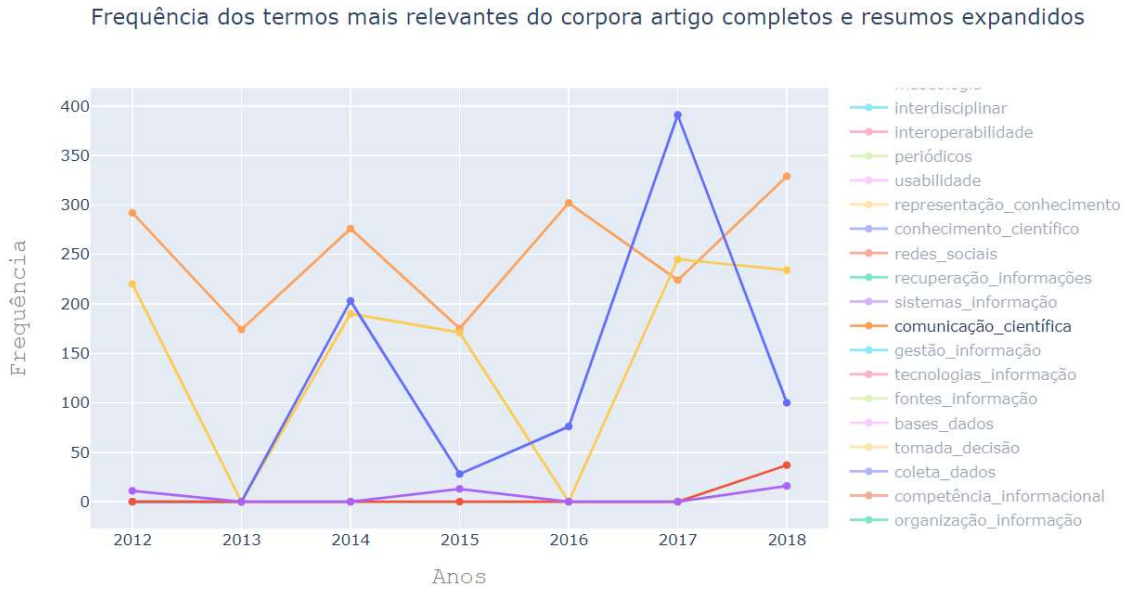
Já os termos “ambientes_digitais”, “*world_wide_web*” e “circulação_apropriação_informação” apresentam comportamentos variados com diversos altos e baixos ao longo do período analisado e apresentando

diversas trocas de posições de ranqueamento entre os termos. O termo “ambientes_digitais” iniciou em 2012 com frequência de 88 –maior frequência entre os três termos – e encerrou em 2018, com frequência de 52, representando uma queda de -41% e ocupando a segunda posição de ranqueamento entre os termos. Já o termo “*world_wide_web*” apresentou frequência de 41 em 2012 e 29 em 2018, também apresentando uma queda ao longo do período, equivalente a -29%, fazendo com que o termo encerrasse na terceira posição de ranqueamento. Já o termo “circulação_apropriação_informação” iniciou com a frequência de 40 em 2012 – menor frequência entre os três termos – e encerrou 2018 com frequência de 65, sendo a maior frequência entre os termos, apresentando um aumento de 63% e ocupando a posição mais alta entre os termos.

O termo “*fake_news*” ganhou notoriedade em 2018, apresentando a maior frequência de todos os termos, com 194 repetições extraídas do *corpus* de dados. Nos anos anteriores, o termo apresentou frequência igual a 0 pelo fato de não aparecer entre as listagens de unigramas do *corpus* de dados que contemplam mil termos mais frequentes.

O Gráfico 16 apresenta o comportamento de termos que constituem parcialmente o referencial teórico e empírico da pesquisa ou estejam relacionados com os assuntos abordados, quais sejam “comunicação_científica” (representado pela cor marrom), “*corpus*” (cor amarela), “*big_data*” (cor azul), “processamento_linguagem_natural” (cor roxa) e “*machine_learning*” (cor vermelha). Termos como “análise_assunto” e “modelagem_tópicos” não apresentaram frequências entre o milésimo termo extraído dos *corpus* de dados, o que pode se caracterizar como áreas pouco exploradas dentro do domínio da linguagem por pesquisadores que participam do programas brasileiros de pós-graduação em Ciência da Informação ou do ENANCIB.

Gráfico 16 – Frequência de termos: artigos e resumos expandidos



Fonte: Elaborado pelo autor.

O termo “comunicação_científica” apresentou evidencialidade em seu comportamento quando se comparado aos demais termos apresentados no gráfico durante quase todo o período analisado. Como resultado, teve, em 2012, uma frequência de 292 e, em 2018, uma frequência de 329, equivalente a um aumento de 12% ao longo do período. O termo “comunicação_científica” somente obteve frequência abaixo de outro em 2017, quando o termo “*big_data*” ultrapassou o quantitativo de frequência, resultando em 391 contra 224, representando uma diferença de 75% entre os termos.

Já o termo “*big_data*” não apresentou frequência ou apresentou frequência após o milésimo bigrama extraídos do *corpus* de dados entre 2012 e 2013. Em 2014 o termo obteve frequência de 203 e, em 2015, uma queda de 86%, resultando em uma frequência de 28. No ano seguinte, o termo apresentou um resultado crescente com frequência de 76 ou 141% e, em 2017, alcançou o maior resultado quando se comparado aos demais termos, com frequência de 391 ou 414% de crescimento referente ao ano anterior. Em 2018 o termo apresentou queda nos resultados de 74%, apresentando a frequência igual a 100 extraídos do *corpus* de dados.

O termo “processamento_linguagem_natural” apresentou baixa frequência entre os anos analisados, inclusive em quatro dos sete anos com frequência após o milésimo termo ou não existentes entre os *corpus* de dados.

O termo apresentou frequência de 11 em 2012, 13 em 2015 e 16 em 2018. Já o termo “*machine_learning*” apresentou frequência de 37 somente em 2018.

5.1.3. Considerações sobre os comportamentos dos termos dos *corpora* de dados

Por meio dos dois gráficos dinâmicos, sendo o primeiro constituído por termos extraídos do *corpora* de dados de teses e dissertações e o segundo do *corpora* de dados de artigos completos e resumos expandidos, disponibilizados por meio da GitHub, pode-se gerar inúmeras possibilidades e combinações entre termos e assuntos para realização de análises de comportamento ao longo do período analisado. Os valores dos termos foram extraídos com base em suas respectivas frequências obtidas em cada um dos 13 *corpus* de dados analisados separadamente.

Faz-se necessário ressaltar que as coleções possuem características diferentes, como tamanho e tipos de documentos do canal formal da comunicação científica, o que permite justificar a diferença de frequência entre os mesmos termos, porém, em *corpus* diferentes, como por exemplo, o termo “informação” que, em 2012, apresenta frequência de 92.726 no primeiro *corpora* de dados e 19.056 no segundo *corpora* de dados, representando uma diferença de 387%. Quando se comparado ao mesmo *corpora* de dados, permite-se analisar o comportamento anual dos termos. Para a construção dos gráficos, os termos foram selecionados de acordo com a sua frequência e/ou relevância ao domínio da linguagem estudada.

Dentre os cenários apresentados, foi possível observar os seguintes comportamentos durante o período analisado: i) termos estáveis, termos que mantêm uma fluidez constante entre o ponto inicial e final, apresentando pequenas alternâncias para baixo ou para cima durante o intervalo; ii) termos instáveis, aqueles que apresentam comportamentos irregulares com alternâncias de valores altos e baixos durante o intervalo analisado e em determinados momentos, apresentando valores igual a 0; iii) termos em desuso, termos que apresentam quedas constantes nos valores de suas respectivas frequências até um ponto que desaparecem do gráfico e não retornam; iv) termos em ascensão, termos que aparecem em determinado ponto no gráfico e

apresentam valores crescentes nos anos posteriores; e v) combinação entre os comportamentos.

Uma característica distinta apresentada entre os resultados dos *corpora* de dados está nos termos estáveis contidos na coleção de documentos constituídos por teses e dissertações e na aparição de termos em ascensão destacados na coleção constituída por documentos do tipo artigos completos e resumos expandidos. Enquanto o primeiro *corpora* de dados apresenta termos com frequências durante todo o intervalo analisado como "sociedade_informação", "busca_informação", "serviços_informação", "segurança_informação", "arquitetura_informação", "gestão_informação", "tecnologia_informação, "organização_informação" e "representação_conhecimento", o segundo *corpora* de dados apresenta termos que se destacam somente no último intervalo analisado, tais como "*machine_learning*", "*fake_news*", "processamento_linguagem_natural", "processo_comunicação_científica".

Pode-se considerar os termos que apresentam valores elevados de frequência ao longo de todo período analisado como termos já estabilizados junto ao domínio da linguagem e/ou utilizados como referencial teórico para novas pesquisas cientificas, mesmo que em outras áreas do conhecimento. Já os termos que surgem em um determinado período de tempo analisado podem ser considerados termos em ascensão junto ao domínio da linguagem, podendo apresentar uma área do conhecimento ainda a ser estudada ou pouco explorada.

A diferença entre os tipos de documentos do canal formal da comunicação científica que constituem os *corpora* de dados, bem como o processo disseminação das pesquisas realizada por meio dos bancos de dados teses e dissertações das universidades federais e dos anais do ENANCIB, podem contribuir para essa diferença entre os termos estáveis e o surgimento de novos termos no domínio da linguagem. Enquanto uma pesquisa de curso de pós-graduação *stricto sensu* na modalidade de mestrado ou doutorado pode levar de dois a quatro anos para ser concluída, sem prazo de prorrogação, as pesquisas aprovadas para o ENANCIB são publicadas anualmente, levando menos tempo para serem realizadas quando se comparado a pesquisas de conclusão de curso.

O surgimento de novos programas de pós-graduação em Ciência da Informação no Brasil, bem como os critérios de avaliação de cursos utilizados pelo Ministério da Educação, contribuem para um maior quantitativo de artigos completos ou resumos expandidos publicados em periódicos científicos ou anais de eventos quando comparado à diferença do quantitativo de produção de teses ou dissertações publicadas a cada ano.

Eventos como o ENANCIB permitem discutir e refletir a cada ano sobre a produção do conhecimento da área da Ciência da Informação e áreas interfaceadas por meio de diálogos realizados entre pesquisadores e compartilhamento de pesquisas que refletem ao estado da arte e o avanço do conhecimento, incluindo perspectivas e tendências da Ciência da Informação.

Outros fatores que contribuem para o surgimento de novos termos, destacado com maior frequência no segundo *corpora* de dados, estão no fato do ENANCIB publicar um maior volume e variedade de pesquisas aplicadas ao mercado profissional, que apresentam de forma prática, dinâmica e em um curto período, publicações contendo soluções que buscam atender necessidades de uma demanda específica e mercadológica. Além disso, o evento permite a participação de pesquisadores que já encerraram vínculos entre discentes e docentes em programas de pós-graduação, não sendo necessário produzir pesquisas alinhadas às temáticas ou linhas de pesquisas dos professores orientadores e seus respectivos programas de pós-graduação, o que ocasiona em uma expansão de temas.

Os comportamentos dos termos instáveis também podem ocorrer por abordarem assuntos pontuais e que buscam solucionar determinada demanda em um período específico. Já os termos em desuso podem ocorrer por uma saturação de pesquisas relacionadas a determinado tema. Essas situações também podem ocorrer mediante a temáticas estabelecidas anualmente por meio da comissão organizadora do ENANCIB.

Sobre os rótulos apresentados nos gráficos, pode-se destacar:

- Dado, informação e conhecimento, sendo considerados unigramas generalistas dentro do domínio da linguagem dos *corpora* de dados, podendo possuir diversas composições de significados quando constituídos de outros termos, formando bigramas ou trigramas. Dessa forma, justifica-se o número elevado de frequência entre os termos. Cabe

ressaltar que os termos, enquanto conceito, são abrangentes estudados na Ciência da Informação, bem como cursos e disciplinas que correlacionam com os termos. Um padrão identificado entre os *corpora* de dados está na maior frequência para os termos Dado e Conhecimento num mesmo ano, sendo 2014 para teses e dissertações e 2018 para artigos completos e resumos expandidos. Já o termo Informação apresentou comportamento diferente de maior frequência entre os *corpora* de dados, sendo 2014 para teses e dissertações e 2018 para artigos completos e resumos expandidos;

- Arquivologia, Biblioteconomia e Museologia enquanto cursos de graduação extraídos da base de dados do e-MEC[43], constam para Arquivologia 18 cursos, sendo 16 ativos, dois não iniciados e primeiro curso iniciado em 1911 pela Universidade Estadual do Rio de Janeiro; Biblioteconomia, com 60 cursos, sendo 48 ativos, três em extinção, dois extintos, sete não iniciados e o primeiro curso iniciado em 1945 pela Pontifícia Universidade Católica de Campinas; Museologia, com 16 cursos, sendo três ativos, dois em extinção e um (1) extinto e não iniciado e primeiro curso iniciado em 1931 pela Universidade Federal do Rio de Janeiro.

  A partir de 2008 houve um aumento no quantitativo de cursos superiores mediante o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (Reuni[44]) cuja meta estipulada estava em dobrar o número de alunos dos cursos de graduação nos próximos 10 anos e permitir o ingresso de 680 mil novos estudantes. Com isso, os cursos de Biblioteconomia, Arquivologia e Museologia foram criados e iniciados em outras IES, respectivamente, 15, seis e oito cursos.

  A criação dos cursos pode representar um aumento no quantitativo de publicações científicas, seja por meio de projetos de pesquisa e extensão na graduação ou publicações de pesquisas científicas em diferentes

[43] Sistema Eletrônico do Ministério da Educação - base de dados oficial dos cursos e Instituições de Educação Superior. Consulta textual por nome do curso utilizando os termos “Arquivologia”, “Biblioteconomia”, “Museologia”. Disponível em: https://emec.mec.gov.br/. Acesso em: 05/06/2020.
Disponível em: https://emec.mec.gov.br/. Acesso em: 04/06/2020.

[44] Reuni – Disponível em: http://portal.mec.gov.br/reuni-sp-93318841/. Acesso em: 07/06/2020.

canais formais da comunicação científica, envolvendo docentes e discentes de 29 novos cursos surgidos após 2008. Com base nos *corpora* de dados analisados, as maiores frequências dos termos Arquivologia e Museologia aconteceram para teses e dissertações, respectivamente entre 2013 e 2015, enquanto para artigos completos e resumos expandidos ocorreu em 2018, com frequência de 548 para Arquivologia e 725 para Museologia, edição do evento ao qual teve como tema: “Sujeito informacional e as perspectivas atuais na Ciência da Informação”.

Acredita-se que a variação de frequência de termos entre os anos pode ocorrer por fatores que envolvem a tríade ensino, pesquisa e extensão, dentre eles: i) criação de cursos de graduação e pós-graduação; ii) atendimento dos critérios de avaliação de curso do MEC, como por exemplo, o critério de publicações científicas dos discentes e docentes; e iii) desenvolvimento das áreas enquanto ciência;

- Publicado nos anais do IX ENANCIB, em 2008, o artigo intitulado “Organização da Informação ou Organização do Conhecimento?”, das autoras Brascher e Café (2008), destaca que para uma comunicação científica eficiente, faz-se necessário a diferenciação entre os conceitos de Organização da Informação e Organização do Conhecimento utilizados em pesquisas da Ciência da Informação. Pode-se supor que a variação dos termos ao longo dos anos, ao qual foram extraídas as frequências, esteja relacionado a outros conceitos correlacionados, como Representação da Informação, Representação do Conhecimento e Arquitetura da Informação, além de recursos tecnológicos utilizados como instrumentos para organização, tratamento, recuperação da informação e do conhecimento;
- Quanto ao comportamento do termo Ontologia, pode-se perceber a ausência do termo no *corpora* de dados constituído por teses e dissertações em 2012, 2016 e 2017, enquanto no *corpora* constituído por artigos completos e resumos expandidos o termo aparece em todos os anos. O termo Ontologia – com frequência analisada neste estudo – possui característica generalista que permite diversas composições de significados dentro do domínio da linguagem estudada ou aplicações da área. Justifica-se ainda essa variação por se tratar de uma área

interdisciplinar que envolve estudos em domínios como a Linguística, Ciência da Computação e Ciência da Informação, além da própria da Filosofia e Lógica (ALMEIDA; OLIVEIRA; COELHO, 2010).

Há de se considerar dois pontos, sendo o primeiro a utilização da frequência relavita ao invés da frequência absoluta para a análise dos resultados, uma vez que se diferem o tamanho dos *corpus* de dados. O segundo ponto diz respeito a observação dos dados da cauda longa como indicativo para mudança da área, bem como a desconsideração de termos com frequências elevadas e que se repetem em todo os *corpora*.

5.2. Considerações sobre a proximidade e o distanciamento entre os termos

Os estudos de Pinheiro (1997; 2006) apresentam frequência das disciplinas do núcleo da Ciência da Informação. De acordo com a autora, o núcleo é constituído por 17 disciplinas, considerando consolidadas as disciplinas que possuem maior frequência, e de tendências as disciplinas com baixa frequência referente à época ao qual o estudo foi realizado – 1966-1995 com 307 trabalhos e 1996-2004, com 81 trabalhos analisados. Utilizando dos mesmos termos da pesquisa de Pinheiro, foi extraída a frequência dos dois *corpora* de dados desta pesquisa, sendo o primeiro constituído por 1.620 teses e dissertações entre os dados de 2012 e 2017 e o segundo por 2.448 artigos científicos e resumos expandidos entre 2012 e 2018.

Na primeira parte do Quadro 14 são apresentadas as disciplinas e suas respectivas frequências, conforme pesquisa realizada de Pinheiro (1997; 2006). Na segunda parte do quadro são apresentadas as mesmas disciplinas, porém em ordens diferentes, já que se trata de outros resultados extraídos de documentos diferentes.

Quadro 14 – Frequência das disciplinas do núcleo da Ciência da Informação

Disciplinas do núcleo da Ciência da Informação (PINHEIRO, 2006).		
Nº	**Disciplinas**	**Frequência**
01	Sistemas de informação	49
02	Tecnologia da informação	36
03	Sistemas de recuperação da informação	35
04	Políticas de informação	28
05	Necessidades e usos de informação	25

06	Representação da informação	25
07	Teoria da Ciência da Informação	16
08	Formação e aspectos profissionais	16
09	Gestão da informação	14
10	Bases de dados	14
11	Processamento automático da linguagem	11
12	Economia da informação	10
13	Bibliometria	6
14	Inteligência competitiva e Gestão do conhecimento	5
15	Mineração de dados	5
16	Comunicação científica eletrônica	3
17	Bibliotecas digitais/virtuais	2
Atualização de frequência extraída dos *corpora* de dados desta pesquisa referente as disciplinas do núcleo da Ciência da Informação de acordo com Pinheiro (2006).		
Nº	**Disciplinas**	**Frequência**
01	Gestão da informação - ^ 8	13.415
02	Sistemas de informação - v1	9.657
03	Bibliotecas digitais/virtuais - ^ 14	8.791
04	Bases de dados - ^ 6	8.784
05	Gestão do conhecimento - ^ 9	8.360
06	Tecnologia da informação - v4	6.234
07	Representação da informação - v1	5.048
08	Economia da informação - ^ 8	4.425
09	Inteligência competitiva - ^ 8	2.835
10	Políticas de informação - v6	2.751
11	Sistemas de recuperação da informação - v8	1.528
12	Bibliometria	Não ranqueado
	Comunicação científica eletrônica	Não ranqueado
	Formação e aspectos profissionais	Não ranqueado
	Mineração de dados	Não ranqueado
	Necessidades e usos de informação	Não ranqueado
	Processamento automático da linguagem	Não ranqueado
	Teoria da Ciência da Informação	Não ranqueado

Fonte: Elaborado pelo autor a partir de Pinheiro (2006).

Torna-se possível observar que todos os resultados sofreram alterações em suas ordens de relevância mediante os valores de suas respectivas frequências. A disciplina de Gestão da Informação subiu oito posições no *ranking* e passou a ser a disciplina com maior frequência, alcançando o somatório entre os *corpora* de dados de 13.415 repetições. O termo Sistema de Informação ocupou a segunda posição, enquanto o termo Representação da Informação ocupou a sétima posição, resultando nas menores alterações de ordem de relevância. Com isso, cada uma das disciplinas caiu uma posição em comparação ao estudo de Pinheiro (2006).

Outros termos que apresentaram queda no ranqueamento foram: Tecnologia da Informação, com queda de quatro posições e na sexta posição no *ranking*; Políticas de Informação, com queda de seis posições e na décima posição no *ranking*; e Sistema de Recuperação de Informação, com queda de

oito posições e na décima primeira posição no *ranking*. Entretanto, se fosse analisado somente o bigrama “recuperação_informação” sem o termo “sistema”, ocuparia a sexta posição no *ranking*, com frequência de 6.729.

Já os termos Bibliotecas Digitais, Bases de Dados, Gestão do Conhecimento, Economia da Informação e Inteligência Competitiva subiram posições no ranqueamento de acordo com a frequência. O termo Bibliotecas Digitais passou a ocupar a terceira posição, subindo 14 posições no *ranking*. Já o termo Bases de Dados alcançou a quarta posição e subiu seis posições no *ranking*. O termo Gestão do Conhecimento alcançou a quinta posição, subindo nove posições no *ranking*. O termo Economia da Informação alcançou a oitava posição e subiu oito posições no *ranking*, enquanto o termo Inteligência Competitiva alcançou a nona posição e subiu oito posições no *ranking*. Nesse cenário a frequência dos termos Inteligência Competitiva e Gestão do Conhecimento foi analisada separadamente; diferentemente de Pinheiro (2006), que supôs não ser o procedimento mais correto mantê-las unificadas por serem disciplinas distintas.

Os termos Bibliometria, Comunicação Científica Eletrônica, Formação e Aspectos Profissionais, Mineração de Dados (*Data Mining*), Necessidades e Usos de Informação, Processamento Automático da Linguagem e Teoria da Ciência da Informação não apresentaram ou apresentaram frequência após o milésimo termo de cada tipo de N-grama. Dessa forma, todas as disciplinas foram alocadas na última posição do *ranking*.

Faz-se necessário destacar que os termos Comunicação Científica Eletrônica e Necessidades e Uso de Informação possuem termos próximos com frequências como: “canais_comunicação_científica,52”, “comunicação_científica,8521”, “produção_comunicação_científica,203”, “processo_comunicação_científica,515”, “sistema_comunicação_científica,445”, “busca_uso_informação,2069”, “compartilhamento_uso_informação,60”, “disseminação_uso_informação,436”, “mediação_uso_informação,287”, “organização_uso_informação,738”, “produção_uso_informação,53”, e “recuperação_uso_informação,56”. O termo “comunicação_científica” sem o termo “eletrônica” ocuparia a quinta posição do *ranking*, com frequência de 8.532 e o termo “uso_informação”, se analisado sem o termo “necessidade”, estaria no top do *ranking,* com frequência de 14.192.

Pinheiro (2006) destaca que os resultados de sua pesquisa são reflexos da Sociedade da Informação influenciada a partir da implantação da internet em uma explosão da informação em proporções maiores. Utilizando-se de metodologias distintas, *corpora* de dados maiores e uma diferença de no mínimo 17 anos e no máximo 52 entre as amostragens de Pinheiro e desta pesquisa, pode-se destacar dentre a diferença dos resultados alcançados, no que diz respeito às frequências, que disciplinas como Sistemas de Informação continuam em voga, abordando assuntos ou instrumentos que podem envolver diversas subáreas da Ciência da Informação para diversos fins.

Contribuindo com Pinheiro (2006), no qual destaca os reflexos da globalização e o uso das Tecnologias de Informação e da Comunicação (TICs) no período de sua pesquisa, pode-se perceber, quase 15 anos depois, que cada vez mais a sociedade está imersa no uso das TICs, inclusive, no meio acadêmico, sendo utilizadas por instituições de ensino superior, discentes e docentes como forma de dinamizar os processos de comunicação científica.

Os termos Bibliotecas Digitais/Virtuais, Bases de Dados e Gestão do Conhecimento apresentaram mudanças ao longo do intervalo analisado. Além de melhores posições de ranqueamento, houve uma aproximação entre essas disciplinas. Uma contribuição para essa alteração pode estar na legislação 12.527/2011, mais conhecida como Lei de Acesso à Informação[45] no qual contribuiu para que as instituições de ensino superior brasileiras elaborassem sistemas de informação que registrassem teses e dissertações de programas de pós-graduação em meio eletrônico, colocando em evidência tais disciplinas da área e possibilitando aos discentes, por meio de metodologias de ensino e aprendizagem utilizadas pelos docentes de suas respectivas universidades, bem como o desenvolvimento de pesquisas, desenvolvessem competências para esse grupo de disciplinas.

Embora a disciplina Tecnologia da Informação tenha caído quatro posições no ranqueamento, torna-se possível observar que todas as disciplinas que estão à sua frente utilizam ou podem utilizar de alguma forma diversas tecnologias da informação que buscam atender os mais variados segmentos. Em uma sociedade cada vez mais imersa nas tecnologias, cabe uma reflexão

[45] Lei de Acesso a Informação. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm/. Acesso em: 15/04/2020.

que termos podem cair em desuso rapidamente enquanto novos termos podem surgir, de acordo com as diversas situações. Já o termo Representação da Informação, que também pode fazer uso das tecnologias da informação, apresentou queda de uma ranqueamento. Entretanto, vem apresentando, ao longo dos anos, um aumento em pesquisas na área de ontologias, inclusive com eventos próprios como Ontobras[46].

Um exemplo de termo que caiu em desuso está na disciplina Comunicação Científica Eletrônica. De acordo com Pinheiro (2006), abordava pesquisas em torno de recursos eletrônicos de comunicação e de informação enquanto na pesquisa com o *corpora* de dados de teses e dissertações e artigos completos e resumos expandidos da Ciência da Informação o termo sequer aparece entre o milésimo N-grama. Em contrapartida, o bigrama "comunicação_científica", enquanto disciplina que aborda os recursos eletrônicos, são discutidos amplamente na comunidade científica de pesquisadores da área da Ciência da Informação.

A disciplina Bibliometria, considerada por Pinheiro (2006, p. 18) como disciplina consolidada, "[...] estudadas desde a instauração da Ciência da Informação e que se mantêm como questões centrais da área, até hoje" não apresentou comportamento similar ao desta pesquisa. Dentro do cenário destacado na metodologia da pesquisa para o quantitativo de mil unigramas extraídos em cada *corpus* de dados, o termo não apareceu na lista de nenhum dos 13 *corpus*, podendo ser classificada como disciplina de tendência, de acordo com a autora ou termo em desuso, o que não significa que não esteja sendo realizadas pesquisas na área de medição da ciência, como bibliometria, cienciometria, webometria e altmetria.

Embora as disciplinas de Mineração de Dados e Processamento Automático de Linguagem possam ter sido vistas como áreas promissoras na Ciência da Informação – considerando o avanço tecnológico por Pinheiro (2006) –, o estudo atual mostrou-se ao contrário no que diz respeito às atividades relacionadas a pesquisas científicas, pois as disciplinas sequer foram ranqueadas.

[46] Ontobras - Seminário de Pesquisa em Ontologias no Brasil.

Entretanto, percebe-se um movimento crescente mercadológico de prestação de serviços, eventos e pesquisas na área de Ciência de Dados, como por exemplo, a Associação Brasileira de Ciência de Dados[47], Big Data Brasil Day[48], Workshop de Informação, Dados e Tecnologia – WIDaT[49] e o curso ofertado na modalidade de superior tecnológico ou bacharelado de 12 instituições de ensino superior no Brasil, com turmas iniciadas a partir de 2019 e 13 outras instituições com autorização aguardando fechamento de turmas para iniciarem suas atividades acadêmicas, de acordo com o Ministério da Educação[50].

Acredita-se que a curva de crescimento de publicações, com temas pertinentes à Ciência de Dados, possa aumentar a partir de 2025, já que a duração dos cursos de graduação dessa área dura entre dois anos e meio e quatro anos, acrescido do tempo de uma especialização *lato* ou *stricto sensu* ao qual, geralmente, exigem publicações científicas. Faz-se necessário destacar que tais temas foram explorados dentro do cenário da Ciência da Informação, entretanto, a Ciência de Dados está alocada, geralmente, dentro das áreas da Ciência da Computação ou Matemática Aplicada.

Com relação aos estudos de Zins e Santos (2015), que utilizam a metodologia *Delphi* e não sendo possível realizar uma análise comparativa por meio de frequência com os resultados desta pesquisa, pode-se perceber por meio das áreas e subáreas, como Ciência da Informação Geral, Campos Especializados e Conhecimento Prático dos autores, um alinhamento referente ao conteúdo das disciplinas e subdisciplinas da Ciência da Informação de Pinheiro (2006).

5.3. Resultados e discussão sobre a modelagem de tópicos

Durante a fase de preparação e pré-processamento, especificamente no que diz respeito à conversão dos documentos do formato PDF para TXT, foi

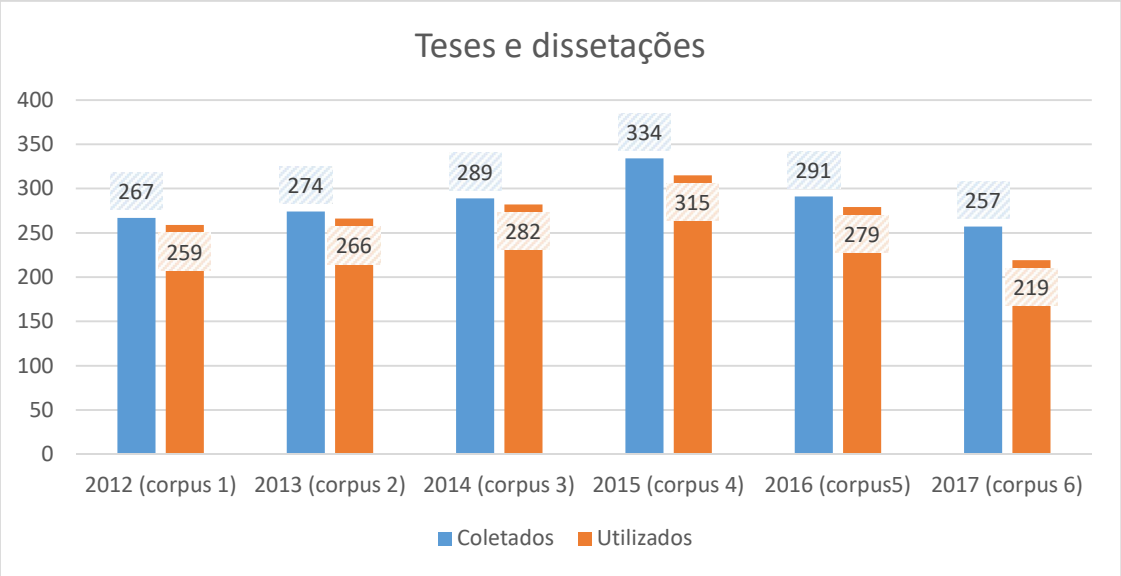
[47] ABRACD – Associação Brasileira de Ciência de Dados. Disponível em: https://abracd.org/. Acesso em: 02/02.2020.
[48] Bid Data Brasil Day. Disponível em: https://bigdatabrasilday.com.br/. Acesso em: 02/02/2020.
[49] Workshop de Informação, Dados e Tecnologia – WIDat. Disponível em http://widat2019.fci.unb.br/. Acesso em: 02/02/2020.
[50] Ministério da Educação – e-MEC. Consulta textual por nome do curso utilizando o termo “Ciência de Dados”. Disponível em: https://emec.mec.gov.br/. Acesso em: 05/06/2020.

necessário verificar todos os *corpora* de dados mediante a erros de Unicode gerados durante processo. Um erro comum diante das limitações da biblioteca está, por exemplo, na conversão de uma imagem contida em um documento para texto ou mesmo documentos inteiros no formato PDF, que acabam sendo convertidos como imagem. Foram eliminados dos documentos mais de 60 tipos de códigos Unicode que interferiam diretamente em resultados como frequência, nuvem de palavras e até mesmo nos resultados alcançados por meio da modelagem de tópicos. Em alguns casos, foi necessária a exclusão de documentos inteiros, seja tese, dissertação, artigos completos ou resumos expandidos.

Dos 4.187 documentos no formato PDF coletados e com tamanho superior a 9GB de dados utilizados para a composição dos *corpora*, sendo 1.712 teses e dissertações e 2.475 artigos completos e resumos expandidos, ambos documentos do canal de comunicação formal da comunicação científica, foram convertidos para o formato TXT já organizados em dois corpora de dados e/ou 13 *corpus* analisados individualmente. Após a conversão e limpeza dos dados, os *corpora* totalizaram 4.068 documentos, representando uma redução de 3% dos dados coletados. Além disso, o tamanho dos arquivos também foi reduzido para um total de 656MG.

Práticas como conversão de documentos de texto para documentos em imagem ou grupos de documentos como anais de eventos em um único arquivo contribuem para uma baixa qualidade na conversão de arquivos em PDF para TXT, entretanto, o percentual de documentos excluídos dos *corpora* apresentou resultados com qualidades superiores quando comparados aos testes com os *corpora* de dados contendo problemas de Unicode. Os Gráficos 17 e 18 apresentam a quantidade de documentos coletados e utilizados por *corpus*/ano.

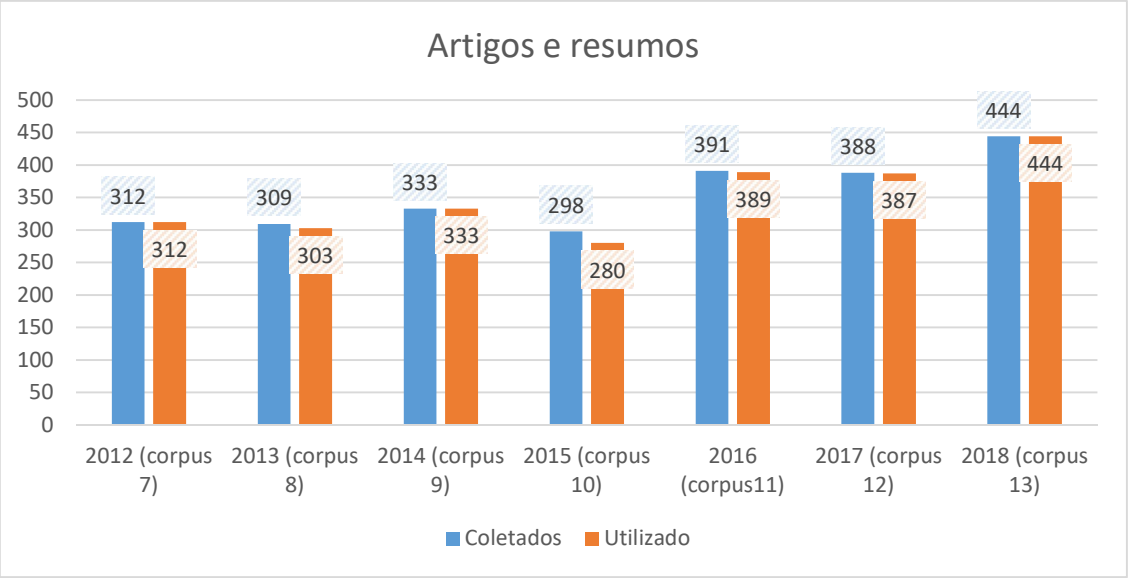
Gráfico 17 – Dados coletados e utilizados do *corpora* de teses e dissertações



Fonte: Elaborado pelo autor.

Pelo fato de as teses e dissertações não possuírem um valor estipulado com relação ao número de páginas, tanto mínimo quanto máximo, diferente dos artigos completos e resumos expandidos do ENANCIB e também por se tratar de pesquisas realizadas, geralmente, entre dois e quatro anos, se tornam documentos grandes, com 100, 200 ou mais laudas, contendo textos, quadros, tabelas, gráficos e imagens. Com isso, a possibilidade de maior descarte de documentos acontece no *corpora* de dados de teses e dissertações, justamente por possuírem um quantitativo maior de informações. Para esse *corpora* foram descartados 92 documentos, sendo: oito documentos para o *corpus* 1; oito documentos para o *corpus* 2; sete documentos para o *corpus* 3; 19 documentos para o *corpus* 4; 12 documentos para o *corpus* 5; e 38 para o *corpus* 6.

Gráfico 18 – Dados coletados e utilizados do *corpora* de artigos completos e resumos expandidos

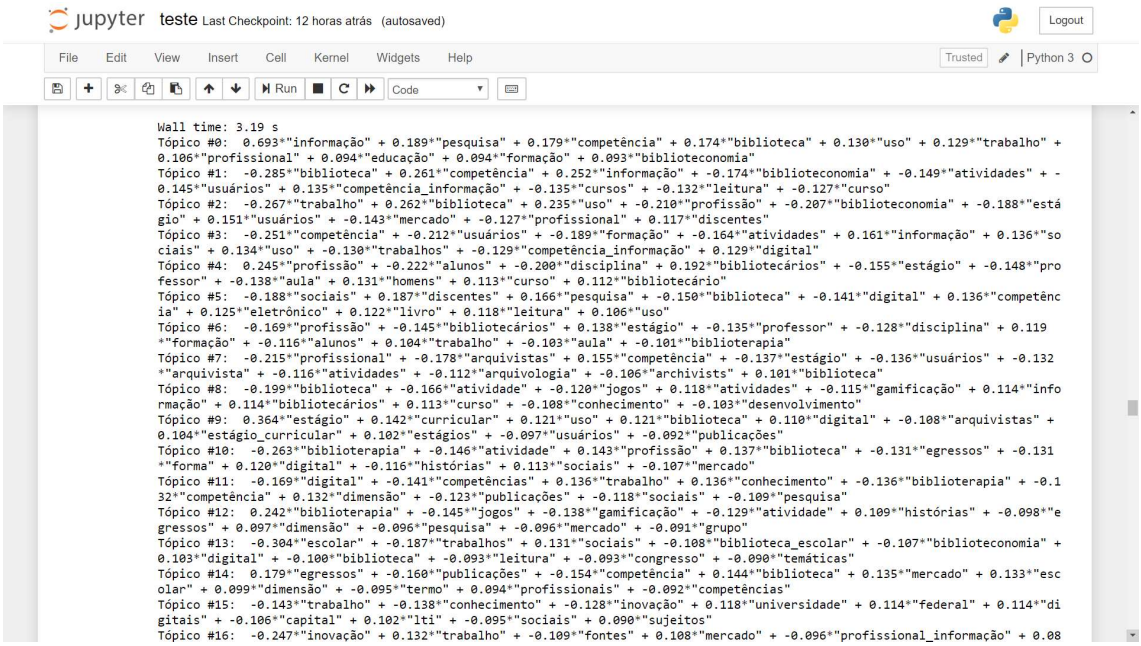


Fonte: Elaborado pelo autor.

Já o segundo *corpora* de dados, constituído por artigos completos e resumos expandidos do ENANCIB, possui o quantitativo maior de documentos, entretanto, o quantitativo de laudas é inferior ao primeiro corpora, uma vez que são estabelecidas por meio das diretrizes para autores o quantitativo mínimo e máximo de laudas por documento/modalidade, sendo 15 a 20 laudas para artigos completos e sete a oito para resumos expandidos. Dessa forma, foram descartados um total de 27 arquivos, sendo seis documentos para o *corpus* 8; 18 documentos para o *corpus* 10; dois documentos para o *corpus* 11; um documento para o *corpus* 12; e os *corpus* 7, 9 e 13 não houveram descartes de documentos.

Ainda na etapa de preparação e pré-processamento, a função re.sub permitiu realizar uma melhor calibração dos pesos entre os termos contidos nos tópicos e extraídos das coleções por meio dos modelos *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA). A partir dessa função, uma lista de termos extraídos do domínio da linguagem e que possuem o mesmo significado encontrados nos *corpora* foram unificados. Exemplo dessa prática é apontado no *corpus* 11, em que o termo no idioma da língua inglesa "*archivists*" foi substituído por “arquivistas”, conforme resultados apresentados nas Figuras 07 e 08:

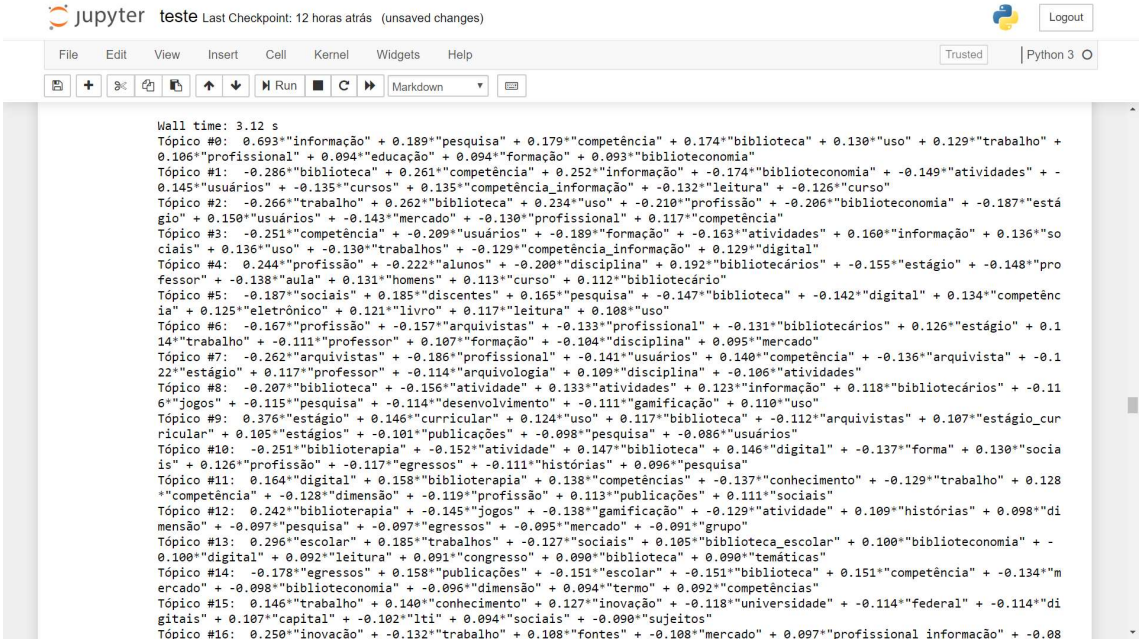
Figura 07 – Exemplo de resultado sem calibração de termos



Fonte: Elaborado pelo autor.

No tópico 07 são apresentados os termos e pesos: -0.215*"profissional" + -0.178*"arquivistas" + 0.155*"competência" + -0.137*"estágio" + -0.136*"usuários" + -0.132*"arquivista" + -0.116*"atividades" + -0.112*"arquivologia" + -0.106*"archivists" + 0.101*"biblioteca". Torna-se possível observar que o tópico, apesar de apresentar termos com pesos expressivos, acaba por se caracterizar fraco por conter termos com o mesmo significado, dividindo os pesos como 0.178*"arquivistas" e -0.106*"archivists".

Figura 08 – Exemplo de resultado com calibração de termos



Fonte: Elaborado pelo autor.

Já com a função re.sub aplicada, o resultado se apresentou mais consolidado, conforme destacado no Tópico 07 da Figura 08, sendo: -0.262*"arquivistas" + -0.186*"profissional" + -0.141*"usuários" + 0.140*"competência" + -0.136*"arquivista" + -0.122*"estágio" + 0.117*"professor" + -0.114*"arquivologia" + 0.109*"disciplina" + -0.106*"atividades". É possível observar que a calibração dos pesos permitiu que o termo “arquivistas” se tornasse o de maior relevância do tópico. Além disso, a união desses termos possibilitou o surgimento de um termo e a substituição de outro, sendo 0.117*"professor" e 0.109*"disciplina". Com essas alterações, o tópico apresenta características que podem ser melhor interpretadas pelo especialista.

A função re.sub pode ser abastecida com mais termos, podendo prospectar melhorias nos resultados de maneira mais eficiente, entretanto, faz-se necessário conhecimento por parte do especialista no domínio da linguagem dos documentos e termos a serem extraídos, como por exemplo, os termos 0.262*"arquivistas" e 0.136*"arquivista", que podem possuir o mesmo significado, conforme apresentado no Tópico 07. O uso dos resultados dos N-gramas pode auxiliar para um melhor uso dessa função, uma vez que os termos podem conter significados diferentes, bem como “informação_arquivística”, “descrição_arquivística”, “instituições_arquivísticas” e “arquivistas_museólogos” encontrados no formato de bigrama nesse mesmo *corpus*. Torna-se importante

ressaltar que os pesos dos termos são resultados dos treinamentos dos modelos de extração de tópicos.

A partir dessa etapa, os dados de cada *corpus* contendo os documentos do canal formal da comunicação cientifica foram executados separadamente, perpassando pelas etapas de eliminação das *stop words*, *tokenização* e criação dos N-gramas do tipo unigrama, bigrama e trigrama. Após identificar os N-gramas, foi calculada a frequência dos termos, que diz respeito a quantas vezes aquele determinado termo aparece em cada *corpus* de dados. Na frequência de termos foram exportadas quatro listas contendo os mil termos mais frequentes de unigrama, bigrama, trigrama e uma lista contemplando todos os tipos de N-gramas, também por ordem de frequência para cada *corpus* de dados.

Foram extraídos um total de 121.198.708 de N-gramas do *corpora* 1, sendo 41.635.208 unigramas, 39.782.560 bigramas e 39.780.940 trigramas enquanto no *corpora* 2 foram extraídos um total de 19.914.348 N-grama, sendo 6.640.564 unigramas, 6.638.116 bigramas e 6.635.668. Por meio da diferença de N-gramas, pode-se ter a noção de diferença de tamanho entre os conteúdos dos *corpora* de dados.

Justifica-se o quantitativo de mil termos para cada lista de N-gramas mediante a qualidade dos resultados extraídos associados ao tamanho dos *corpora* de dados, podendo apresentar um maior quantitativo de N-gramas bons do que ruins. Exemplo é o trigrama “doc_doc_doc”. Ele não apresenta relevância e obtém frequência igual a 20, ocupando ranqueamento entre o milésimo termo em um dos *corpus* de dados, enquanto “arquivologia_biblioteconomia_museologia”, com representatividade ao domínio da linguagem, apresenta frequência igual a 112 e ocupa a posição de ranqueamento 207 entre mil termos. Dessa forma, faz-se necessário ao especialista realizar a análise de assunto dos termos com relação ao domínio da linguagem para seleção dos dados com maior relevância, de acordo com a finalidade.

Uma discussão acerca das *stop words* diz respeito à eliminação de todas as palavras de parada contidas em um *corpus* de dados. Isso, além de ser um trabalho exaustivo de testes de execução e análise para cada *corpus* estudado, não apresenta resultados de grande relevância, uma vez que, além da biblioteca possuir um número relevante de palavras de parada, foram adicionadas a lista

outras *stop words* que contribuem para melhores resultados. Algumas palavras de parada são encontradas entre os N-gramas, entretanto, podem ser facilmente descartadas por especialistas ao realizar uma filtragem dos principais termos, desconsiderando o número elevado de frequência e considerando a representatividade do termo ao domínio da linguagem. Essa análise pode ser realizada quando os dados são utilizados para outros objetivos, como a criação de um gráfico de comportamento de termos ou criação de nuvem de palavras.

A seguir são apresentados os resultados e discussão alcançados por meio das etapas de transformação, modelagem, processamento e apresentação dos resultados do *corpus* 1 e do *corpus* 7, sendo os primeiros *corpus* de cada *corpora* de dados. Além dos resultados diferentes, o *corpus* 7 apresenta uma visualização dinâmica de tópicos extraídos. Os *corpus* 2, 3, 4, 5 e 6, constituídos por documentos do tipo tese e dissertações e que fazem parte do primeiro *corpora* de dados, estão disponibilizados sequencialmente nos anexos B, C, D, E e F. Já os *corpus* 8, 9, 10, 11, 12 e 13, constituídos por documentos do tipo artigos completos e resumos expandidos, estão disponibilizados respectivamente nos anexos G, H, I, J, K e L desta tese.

5.3.1. *Corpus* 1: teses e dissertações 2012

O *corpus* 1 possui o quantitativo de 259 documentos defendidos em 2012 e um tamanho de 89.356kb. Com a transformação dos dados, o *corpus* resultou em um quantitativo de 6.139.088 unigramas, 6.138.829 bigramas e 6.138.570 trigramas. O Quadro 15 apresenta a lista de frequência dos 50 primeiros termos mais frequentes, separados por tipos de N-gramas extraídos no *corpus*.

Quadro 15 – Lista de N-gramas por ordem de frequência do *corpus* 1

Unigramas
informação,92726; pesquisa,29708; biblioteca,25254; conhecimento,23025; forma,18363; trabalho,17004; processo,16969; social,16838; dados,15928; comunicação,15696; relação,14001; uso,13594; sociais,12557; sociedade,12458; produção,12164; organização,12003; desenvolvimento,11666; tempo,10830; meio,10753; brasil,10687; documentos,10584; sistema,10228; museu,10217; estudo,9909; pessoas,9885; anos,9877; memória,9813; paulo,9726; história,9252; universidade,9175; educação,9038; usuários,8853; nacional,8790; científica,8700; fonte,8638; atividades,8573; digital,8456; contexto,8340; termo,8289; estudos,8143; cultura,8142; espaço,8139; grande,8034; busca,7785; gestão,7776; gente,7681; grupo,7601; vida,7573; rede,7497; serviços,7415.
Bigramas
redes_sociais,3087; universidade_federal,3084; recuperação_informação,2513; fontes_informação,2404; produção_científica,1951 ensino_superior,1888;

informação_comunicação,1798; tecnologias_informação,1741; informação_conhecimento,1711; belo_horizonte,1682; uso_informação,1616; dissertação_mestrado,1604; biblioteca_universitárias,1573; ponto_vista,1565; gestão_informação,1469; muitas_vezes,1463; sociedade_informação,1462; coleta_dados,1444; santa_catarina,1418; comunicação_científica,1412; informação_tecnologia,1408; inclusão_digital,1371; biblioteca_escolar,1342; zero_hora,1333; dados_pesquisa,1258; informação_science,1249; sistemas_informação,1193; arquitetura_informação,1189; porto_alegre,1180; conhecimento_científico,1177; competência_informacional,1168; biblioteca_universitária,1165; informação_informação,1157; informação_brasília,1136; fonte_dados,1118; busca_informação,1104; tecnologia_informação,1062; necessidades_informação,1058; novas_tecnologias,1055; bases_dados,996; informação_científica,986; organização_informação,971; biblioteca_digitais,969; tendo_vista,963; gestão_conhecimento,954; meios_comunicação,909; base_dados,878; memória_social,874; biblioteca_digital,838; teses_dissertações,825.
Trigramas
tecnologias_informação_comunicação,1082; fonte_dados_pesquisa,963; instituição_ensino_superior,668; federal_santa_catarina,634; universidade_federal_santa,609; dissertação_mestrado_informação,543; informação_belo_horizonte,466; informação_universidade_federal,455; fonte_elaborado_autora,393; gestão_informação_conhecimento,367; universidade_federal_grande,364; universidade_federal_paraíba,360; instituições_ensino_superior,360; instituições_arquivísticas_nacionais,354; anos_anos_anos,341; perspectivas_informação_belo,318; federal_grande_sul,307; international_organization_standardization,301; american_society_informação,293; society_informação_science,289; federal_minas_gerais,284; portal_periódicos_capes,284; museu_astronomia_afins,272; universidade_federal_minas,269; general_public_licence,266; arquivo_teste_gerado,257; teste_gerado_versão,257; universidade_federal_bahia,253; novas_tecnologias_informação,251; classificação_decimal_universal,242; total_total_total,239; encontro_nacional_pesquisa,231; american_library_association,230; brasília_briquet_lemos,229; journal_american_society,229; fase_fase_fase,226; documentação_informação_histórica,225; nacional_pesquisa_informação,223; informação_histórica_regional,220; patrimônio_histórico_artístico,219; extensible_markup_language,214; informação_sociedade_estudos,213; centro_humanas_sociais,212; segunda_guerra_mundial,207; informação_science_technology,207; requisito_parcial_obtenção,205; datagramazero_revista_informação,205; ensino_pesquisa_extensão,205; fundação_oswaldo_cruz,203; world_wide_web,201.

Fonte: Elaborado pelo autor.

Entre a lista dos mil N-gramas, os primeiros bigramas com maior frequência são: “redes_sociais”, que aparece na posição 270 com frequência de 3.087; seguido de “universidade_federal”, que ocupa a posição 273 e frequência de 3.084; “recuperação_informação” na posição 365 e frequência de 2.513; “fontes_informação”, na posição 382 e frequência de 2.404; e “produção_científica”, na posição 501 com frequência de 1951.

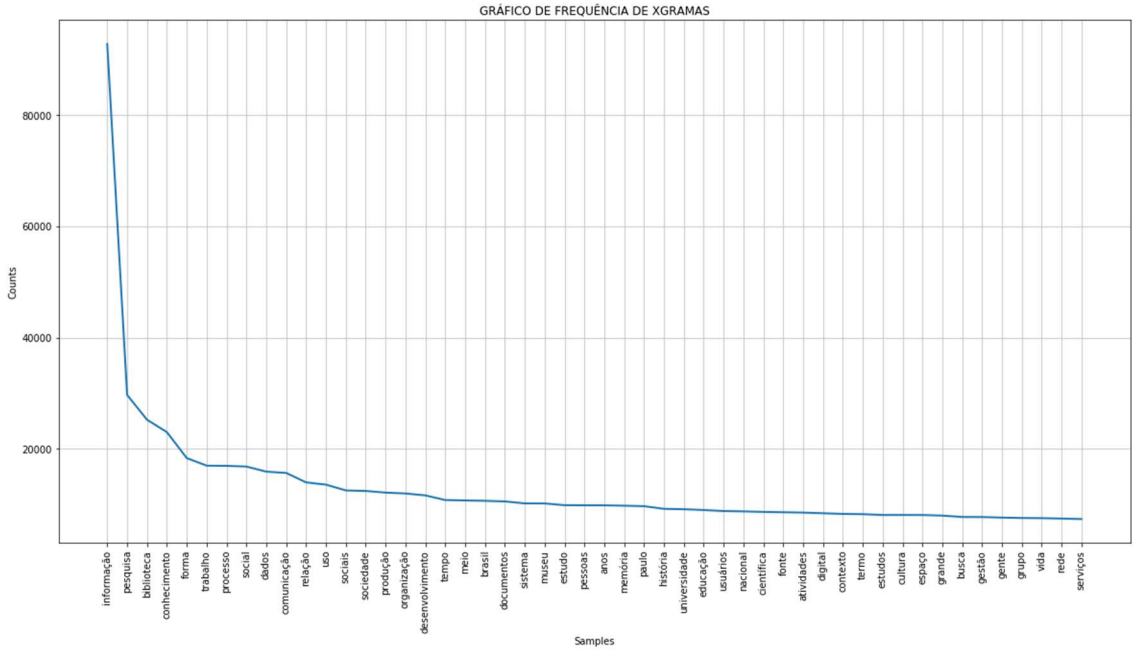
Nesse *corpus*, todos os trigramas aparecem após o milésimo termo da lista geral de N-gramas, estando entre os mais frequentes:

“tecnologias_informação_comunicação”, com frequência de 1082; “fonte_dados_pesquisa”, com frequência de 963; “instituição_ensino_superior”, com frequência de 668; “federal_santa_catarina”, com frequência de 634; e “universidade_federal_santa”, com frequência de 609.

Percebe-se que os unigramas possuem maior frequência em relação aos bigramas e trigramas. O unigrama “informação”, com maior frequência, alcança 92.726 repetições, sendo sua frequência 2.904% maior que o primeiro bigrama “redes_sociais” com a frequência de 3087 repetições. A diferença de percentual alcança 8.470% quando comparado o primeiro unigrama “informação” com o primeiro trigrama “tecnologias_informação_comunicação”, com 1.082 repetições.

O Gráfico 19 apresenta os 50 termos mais frequentes extraídos do *corpus* 1 de dados. No gráfico, é possível perceber que todos os termos apresentados são do tipo unigrama. Além disso, existe uma diferença considerável entre o primeiro termo para com os demais. Quando comparado, por exemplo, o primeiro termo “informação” com frequência de 92.726 e o segundo termo “pesquisa” com frequência de 29.708, encontra-se uma diferença de 212% entre os termos.

Gráfico 19 – Termos mais frequentes do *corpus* 1



Fonte: Elaborado pelo autor.

O termo “informação”, extraído do *corpus* e representado por meio de unigrama com maior frequência pode ser de difícil interpretação quando analisado sozinho. Além de ser um termo central de estudos da Ciência da

Informação, conforme destacado por Borko (1968, p. 3), “[...] é aquela disciplina que investiga as propriedades e o comportamento informacional, as forças que governam os fluxos de informação, e os significados do processamento da informação [...]”. O termo pode ser melhor interpretado quando analisado por meio de bigramas e trigramas, por poder apresentar um contexto de emprego do termo, como apresentado os “termo,frequência”: recuperação_informação,2513; fontes_informação,2404; informação_comunicação,1798; tecnologias_informação,1741; informação_conhecimento,1711; uso_informação,1616; gestão_informação,1469; sociedade_informação,1462; sistemas_informação,1193; arquitetura_informação,1189; gestão_informação_conhecimento,367; novas_tecnologias_informação,251; documentação_informação_histórica,225; informação_histórica_regional,220; informação_sociedade_estudos,213; e sistemas_recuperação_informação,198.

É possível identificar na relação de N-gramas termos que apresentam características de baixa qualidade, entretanto, possuem frequências elevadas que podem aparecer entre os resultados da modelagem de tópicos. Esses resultados podem dificultar a análise e interpretação dos dados realizada pelo especialista no domínio da linguagem, como por exemplo, os “termo,frequência”: informação_belo_horizonte,466; informação_universidade_federal,455; belo_horizonte,1682; universidade_federal_minas,269; federal_santa_catarina,634; e paulo,9726 ,que podem ocupar um dos 10 termos contidos em cada tópico.

Termos únicos necessitam de uma exploração dos dados, uma vez que podem representar conteúdos diferentes. Alguns exemplos de “termo,frequência” neste *corpus* são: processo,16969; social,16838; dados,15928; e comunicação,15696. O unigrama “processo” pode representar outros campos de estudo, como por exemplo, “gestão_processo”, “processo_comunicação” e/ou “processo_produção” enquanto “social” pode abordar “rede_social”, “construção_social” e/ou “vida_social” e “comunicação” pode estar associado a “canais_comunicação”, “comunicação_científica” e/ou “tecnologia_comunicação”.

A Figura 09 apresenta uma nuvem de palavras contendo os 250 N-gramas mais frequentes do *corpus* 1, não sendo realizados quaisquer tratamentos qualitativos de termos para construção da imagem.

Figura 09 – Nuvem de palavras do *corpus* 1

Fonte: Elaborado pelo autor.

A imagem apresenta termos fortes e fracos quando se comparada ao domínio de linguagem e realizada a análise de assunto. Pode-se destacar, por exemplo, os termos “usuário”, “discurso” e “forma” como termos fortes, mesmo sendo necessário consultar a lista de bigramas e trigramas para identificar as vertentes dos termos. Já os termos fracos podem ser exemplificados por meio dos unigramas “meio”, “maneira” e “grande” que não possuem representatividade junto ao domínio de linguagem.

Após a etapa de transformação, os dados foram conectados aos modelos de extração de tópicos e processados por meio de *Machine Learning* para gerar os resultados. O Quadro 16 apresenta os resultados obtidos através do modelo LSI em um conjunto de 30 tópicos com os respectivos termos e pesos. O processo foi executado no *corpus* 1 em 7 minutos e 51 segundos. Os demais resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do *GitHub*[51].

Quadro 16 – Tópicos extraídos do *corpus* 1 usando o modelo LSI

Tópico 0: 0.668*"informação" + 0.195*"pesquisa" + 0.191*"biblioteca" + 0.151*"conhecimento" + 0.100*"comunicação" + 0.099*"dados" + 0.097*"processo" + 0.097*"forma" + 0.096*"trabalho" + 0.093*"social";
Tópico 1: 0.554*"informação" + -0.474*"biblioteca" + -0.158*"gente" + -0.132*"museu" + -0.103*"social" + -0.097*"espaço" + -0.092*"educação" + -0.088*"memória" + -0.086*"trabalho" + -0.083*"história";

[51] Algoritmo de modelagem de tópicos. *Corpus* 1: teses e dissertações 2012. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2012.ipynb/.

Tópico 2: 0.648*"biblioteca" + -0.183*"museu" + 0.144*"usuários" + -0.137*"memória" + -0.114*"história" + 0.110*"informação" + -0.090*"forma" + -0.090*"trabalho" + -0.082*"cultura" + -0.075*"produção";
Tópico 3: 0.276*"gente" + 0.228*"informação" + -0.221*"pesquisa" + -0.210*"científica" + -0.193*"documentos" + -0.162*"museu" + 0.141*"zero" + -0.139*"produção" + -0.134*"conhecimento" + 0.124*"hora";
Tópico 4: -0.444*"museu" + 0.235*"zero" + -0.199*"informação" + 0.182*"hora" + 0.179*"zero_hora" + 0.155*"jornal" + 0.144*"comunicação" + 0.130*"produção" + 0.127*"redes" + 0.122*"web";
Tópico 5: -0.270*"pesquisa" + -0.247*"conhecimento" + -0.231*"trabalho" + 0.166*"digital" + 0.159*"usuários" + 0.158*"museu" + 0.153*"biblioteca" + 0.151*"web" + -0.150*"inovação" + -0.139*"escola";
Tópico 6: 0.279*"documentos" + -0.260*"museu" + -0.239*"redes" + -0.202*"sociais" + -0.191*"rede" + 0.187*"arquivo" + -0.178*"comunicação" + 0.177*"arquivos" + 0.170*"documento" + -0.163*"social";
Tópico 7: 0.216*"digital" + 0.206*"museu" + -0.200*"redes" + -0.179*"rede" + -0.178*"sociais" + 0.160*"zero" + 0.142*"pesquisa" + -0.141*"linguagem" + -0.141*"social" + 0.141*"gente";
Tópico 8: 0.326*"museu" + -0.216*"arquivo" + -0.202*"documentos" + 0.180*"conhecimento" + -0.170*"trabalho" + -0.169*"arquivos" + 0.150*"pesquisa" + -0.136*"social" + 0.134*"zero" + 0.125*"científica";
Tópico 9: -0.291*"digital" + 0.286*"conhecimento" + 0.228*"inovação" + 0.214*"organização" + 0.183*"museu" + 0.140*"gestão" + -0.131*"pesquisa" + 0.125*"processo" + -0.112*"inclusão" + 0.111*"empresa";
Tópico 10: 0.393*"digital" + -0.191*"arquivo" + 0.181*"conhecimento" + -0.157*"arquivos" + 0.153*"inovação" + -0.141*"documentos" + 0.134*"digitais" + -0.132*"comunicação" + 0.122*"organização" + -0.121*"zero";
Tópico 11: -0.251*"digital" + -0.222*"redes" + -0.205*"rede" + -0.200*"museu" + -0.164*"gente" + -0.129*"escolar" + -0.125*"escola" + -0.117*"redes_sociais" + 0.116*"inovação" + -0.115*"documentos";
Tópico 12: 0.333*"portal" + -0.204*"zero" + 0.202*"usuários" + 0.194*"uso" + -0.156*"zero_hora" + -0.151*"hora" + -0.125*"educação" + 0.118*"dados" + 0.118*"web" + -0.113*"digital";
Tópico 13: 0.300*"gente" + 0.256*"memória" + -0.229*"trabalho" + 0.155*"conhecimento" + 0.136*"história" + 0.115*"organização" + -0.115*"linguagem" + 0.113*"tortura" + -0.111*"leitura" + -0.108*"museu";
Tópico 14: -0.292*"trabalho" + 0.191*"inovação" + 0.154*"memória" + 0.136*"escolar" + 0.129*"comunicação" + 0.127*"conhecimento" + -0.121*"biblioteca_universitárias" + -0.121*"indexação" + -0.120*"universitárias" + -0.115*"universidade";
Tópico 15: 0.350*"trabalho" + -0.231*"cartas" + 0.185*"portal" + -0.165*"carta" + -0.144*"dados" + -0.123*"leitura" + -0.123*"escolar" + 0.116*"tempo" + 0.114*"uso" + -0.100*"jornal";
Tópico 16: -0.326*"conhecimento" + -0.252*"web" + 0.173*"trabalho" + -0.170*"portal" + 0.156*"dados" + 0.150*"produção" + -0.136*"sociedade" + 0.132*"rede" + -0.120*"social" + 0.120*"biblioteca";
Tópico 17: -0.272*"trabalho" + 0.204*"redes" + -0.194*"comunicação" + 0.182*"rede" + 0.152*"linguagem" + 0.150*"gente" + 0.142*"portal" + -0.127*"dispositivos" + 0.122*"retórica" + -0.118*"produção";
Tópico 18: 0.231*"cartas" + 0.163*"carta" + 0.156*"conhecimento" + 0.139*"trabalho" + -0.138*"memória" + 0.133*"museu" + -0.128*"indexação" + 0.125*"artigos" + 0.111*"digital" + -0.100*"educação";

Tópico 19: -0.274*"livros" + 0.211*"pesquisa" + -0.173*"livro" + -0.166*"comunicação" + -0.149*"leitura" + -0.147*"dados" + 0.143*"cartas" + -0.139*"políticas" + 0.123*"história" + 0.119*"portal";
Tópico 20: 0.309*"portal" + -0.207*"web" + 0.188*"preservação" + -0.187*"pesquisa" + 0.170*"uso" + 0.126*"memória" + 0.124*"falta" + -0.114*"avaliação" + -0.104*"inclusão" + 0.101*"documentos";
Tópico 21: -0.187*"portal" + 0.173*"pesquisa" + -0.171*"governo" + 0.165*"comunicação" + 0.148*"cartas" + -0.136*"nacional" + -0.121*"escolar" + -0.116*"brasil" + -0.115*"biblioteca" + 0.113*"gente";
Tópico 22: -0.267*"web" + 0.156*"inclusão" + 0.148*"social" + -0.137*"instituições" + -0.132*"escolar" + -0.118*"escola" + 0.115*"sociais" + 0.113*"cultura" + -0.112*"segurança" + 0.111*"indexação";
Tópico 23: 0.201*"livros" + 0.157*"profissional" + 0.157*"memória" + 0.138*"cultura" + 0.126*"web" + 0.121*"leitura" + 0.119*"inovação" + -0.118*"governo" + -0.115*"pesquisa" + 0.115*"serviços";
Tópico 24: 0.209*"profissional" + -0.165*"trabalho" + -0.151*"livros" + 0.146*"dados" + 0.141*"capaz" + 0.139*"bibliotecário" + 0.135*"competências" + 0.124*"bibliotecários" + 0.124*"usuário" + 0.119*"serviço";
Tópico 25: 0.344*"artigos" + 0.288*"neural" + 0.182*"referências" + 0.159*"networks" + 0.149*"network" + 0.147*"neural_networks" + -0.140*"trabalho" + -0.136*"livros" + -0.134*"leitura" + 0.128*"neural_network";
Tópico 26: 0.228*"direito" + 0.191*"livros" + 0.174*"competência" + 0.166*"leitura" + 0.161*"trabalho" + 0.156*"conhecimento" + 0.154*"jurídica" + 0.149*"segurança" + 0.132*"livro" + 0.123*"informacional";
Tópico 27: 0.208*"história" + -0.196*"arte" + 0.172*"museu" + -0.146*"cultura" + -0.120*"patrimônio" + -0.119*"cultural" + 0.114*"dados" + -0.113*"instituições" + -0.109*"projeto" + -0.107*"cidade";
Tópico 28: -0.242*"segurança" + 0.189*"inovação" + -0.178*"políticas" + 0.167*"fontes" + -0.150*"segurança_informação" + -0.125*"instituição" + 0.118*"comunicação" + -0.101*"arte" + -0.101*"avaliação" + 0.099*"direito";
Tópico 29: -0.239*"conhecimento" + 0.204*"inovação" + 0.154*"linguagem" + -0.143*"arte" + 0.134*"artigos" + 0.126*"inclusão" + 0.125*"indexação" + 0.122*"neural" + -0.119*"científico" + 0.114*"retórica".

Fonte: Elaborado pelo autor.

Nos resultados obtidos por meio do modelo LSI foi possível perceber tópicos que contemplam termos e pesos com valores fortes quando se comparado aos demais tópicos. Dentre eles estão os termos “informação”, nos tópicos 0 e 1, e “biblioteca”, no tópico 2, que apresentam valores acima de 0.500. Os termos com os maiores valores são os que melhor representam aquele determinado tópico. É possível perceber que todos os termos de todos os tópicos possuem algum tipo de valor atribuído em seus respectivos pesos, sendo positivo ou negativo. Mesmo que haja tópicos contendo termos iguais, como por exemplo, “digital” para o tópico 8, 9, 10 e 11, existem outros termos nos tópicos que possibilitam ao especialista tomar a decisão de unificar os tópicos ou excluir os tópicos com menor valor de relevância ou menor significância frente ao domínio de linguagem.

É possível identificar também termos-chave como “cartas” nos tópicos 15, 18, 19 e 21, “espaço” no tópico 1 e “neural” nos tópicos 25 e 26, que servem como norte ao especialista no domínio da linguagem para utilizar, por meio de técnicas de análise de assunto, a identificação do melhor tema para os conjuntos, já que os demais termos dos respectivos tópicos apresentam características generalistas e podem se compor para diversos termos com significados diferentes, como por exemplo, o unigrama “dados” que pode se compor em “gestão_dados” ou “banco_dados”. Os termos-chave são ideais para que o especialista explore documentos externos como o próprio *corpus* de dados, caso tenha acesso, para identificar os trabalhos que abordem determinado termo e, consequentemente, obter mais informações por meio do documento de forma que facilite a criação da suposição de um nome para o tópico. Além disso, as listas de N-gramas também podem auxiliar o profissional indexador caso tenha dúvidas na interpretação dos resultados.

O Quadro 17 apresenta os resultados extraídos do *corpus* 1 utilizando o modelo LDA, executado em 42 minutos e 43 segundos. Os demais resultados com 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[52].

Quadro 17 – Tópicos extraídos do *corpus* 1 usando o modelo LDA

Tópico 0: 0.002*"segurança" + 0.002*"segurança_informação" + 0.001*"políticas_instituição" + 0.001*"instituição_pesquisada" + 0.000*"mayara" + 0.000*"pesquisada" + 0.000*"políticas_instituição_pesquisada" + 0.000*"critério" + 0.000*"petruso" + 0.000*"mayara_petruso";
Tópico 1: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"conhecimento" + 0.000*"biblioteca" + 0.000*"processo" + 0.000*"trabalho" + 0.000*"dados" + 0.000*"forma" + 0.000*"uso" + 0.000*"profissional";
Tópico 2: 0.002*"docentes" + 0.002*"repositórios" + 0.001*"produção_científica" + 0.001*"repositórios_institucionais" + 0.001*"repositório" + 0.001*"qualis" + 0.001*"científica" + 0.001*"coletâneas" + 0.001*"chile" + 0.001*"teses";
Tópico 3: 0.002*"recife" + 0.001*"pernambuco" + 0.001*"monumentos" + 0.001*"anníbal" + 0.001*"inspetoria" + 0.000*"pernambuco_recife" + 0.000*"província" + 0.000*"fernandes" + 0.000*"província_recife" + 0.000*"anníbal_fernandes";
Tópico 4: 0.001*"unirio" + 0.001*"astronomia" + 0.001*"museu_astronomia" + 0.001*"astronomia_afins" + 0.001*"museu_astronomia_afins" + 0.001*"edifício" + 0.001*"museologia" + 0.001*"afins" + 0.000*"mast" + 0.000*"itu";
Tópico 5: 0.001*"sepetiba" + 0.000*"ambiental" + 0.000*"praia" + 0.000*"inea" + 0.000*"reabilitação" + 0.000*"praia_sepetiba" + 0.000*"contato" + 0.000*"rio-águas" + 0.000*"saneando" + 0.000*"saneando_sepetiba";

[52] Algoritmo de modelagem de tópicos. *Corpus* 1: teses e dissertações 2012. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2012.ipynb/.

Tópico 6: 0.000*"informação" + 0.000*"conhecimento" + 0.000*"pesquisa" + 0.000*"social" + 0.000*"forma" + 0.000*"sociedade" + 0.000*"relação" + 0.000*"museu" + 0.000*"processo" + 0.000*"trabalho";
Tópico 7: 0.000*"informação" + 0.000*"museu" + 0.000*"pesquisa" + 0.000*"rede" + 0.000*"comunicação" + 0.000*"dados" + 0.000*"relação" + 0.000*"social" + 0.000*"sociais" + 0.000*"forma";
Tópico 8: 0.002*"artigos" + 0.002*"neural" + 0.001*"networks" + 0.001*"neural_networks" + 0.001*"referências" + 0.001*"network" + 0.001*"neural_network" + 0.001*"rna" + 0.000*"categorização" + 0.000*"recurrent";
Tópico 9: 0.001*"informa" + 0.001*"coordenadoria" + 0.001*"assembléia" + 0.001*"legislativa" + 0.001*"diretoria" + 0.001*"organiza" + 0.001*"assembléia_legislativa" + 0.001*"indexação" + 0.001*"epc" + 0.001*"desenho";
Tópico 10: 0.001*"libraries" + 0.001*"digital_libraries" + 0.001*"library" + 0.001*"digital_library" + 0.001*"biblioteca_digitais" + 0.001*"digital" + 0.001*"usabilidade" + 0.001*"evaluation" + 0.001*"bvs" + 0.000*"usability";
Tópico 11: 0.001*"ceará" + 0.001*"museu_ceará" + 0.001*"general_public" + 0.001*"general_public_licence" + 0.001*"public_licence" + 0.001*"licence" + 0.001*"general" + 0.001*"public" + 0.001*"gpl" + 0.000*"gabinete";
Tópico 12: 0.002*"escola" + 0.002*"gente" + 0.002*"presos" + 0.002*"trabalho" + 0.002*"nietzsche" + 0.001*"professor" + 0.001*"professores" + 0.001*"trabalhador" + 0.001*"homem" + 0.001*"alunos";
Tópico 13: 0.000*"informação" + 0.000*"biblioteca" + 0.000*"pesquisa" + 0.000*"conhecimento" + 0.000*"trabalho" + 0.000*"forma" + 0.000*"processo" + 0.000*"dados" + 0.000*"social" + 0.000*"relação";
Tópico 14: 0.019*"biblioteca" + 0.004*"escolar" + 0.002*"usuários" + 0.002*"biblioteca_escolar" + 0.002*"bibliotecário" + 0.002*"bibliotecários" + 0.002*"ensino" + 0.002*"biblioteca_universitária" + 0.002*"redes_sociais" + 0.001*"universitária";
Tópico 15: 0.002*"otlet" + 0.001*"retórica" + 0.001*"lgbt" + 0.001*"linguagem" + 0.001*"artigo" + 0.001*"ndihr" + 0.001*"documentação" + 0.001*"bioética" + 0.001*"resumo" + 0.001*"organização";
Tópico 16: 0.003*"jornalismo" + 0.002*"jornal" + 0.002*"cinema" + 0.002*"caderno" + 0.002*"imagens" + 0.001*"violência" + 0.001*"leitor" + 0.001*"audiovisual" + 0.001*"imagem" + 0.001*"mulheres";
Tópico 17: 0.001*"mbe" + 0.000*"oswaldo_cruz" + 0.000*"oswaldo" + 0.000*"fundação_oswaldo_cruz" + 0.000*"fundação_oswaldo" + 0.000*"medicina_baseada" + 0.000*"ccda-fiocruz" + 0.000*"medicina_baseada_evidências" + 0.000*"baseada_evidências" + 0.000*"gestão_documentos";
Tópico 18: 0.000*"informação" + 0.000*"forma" + 0.000*"pesquisa" + 0.000*"comunicação" + 0.000*"relação" + 0.000*"social" + 0.000*"biblioteca" + 0.000*"documentos" + 0.000*"sociais" + 0.000*"educação";
Tópico 19: 0.001*"indexação" + 0.001*"preservação" + 0.001*"ontologias" + 0.001*"alinhamento" + 0.001*"assunto" + 0.001*"jogos" + 0.001*"ontologia" + 0.001*"digital" + 0.001*"falta" + 0.001*"multinucleate";
Tópico 20: 0.003*"zero" + 0.003*"zero_hora" + 0.002*"hora" + 0.001*"jornal" + 0.001*"hora.com" + 0.001*"zero_hora.com" + 0.001*"impresso" + 0.001*"ipad" + 0.001*"jornal_impresso" + 0.001*"manchetes";
Tópico 21: 0.001*"mineração" + 0.001*"mineração_dados" + 0.001*"ontologias" + 0.001*"difusa" + 0.000*"fuzzy" + 0.000*"vaguidade" + 0.000*"swrlb" + 0.000*"difusos" + 0.000*"difusas" + 0.000*"ontologia";
Tópico 22: 0.001*"gonçalo" + 0.001*"josé" + 0.001*"olympio" + 0.001*"josé_olympio" + 0.001*"telenovela" + 0.001*"ibge" + 0.001*"lobato" + 0.001*"família" + 0.001*"morros" + 0.001*"centro_cultural";

Tópico 23: 0.001*"seer" + 0.001*"navegação" + 0.000*"qtd" + 0.000*"processo_editorial" + 0.000*"fluxo_processo" + 0.000*"usabilidade" + 0.000*"fluxo_processo_editorial" + 0.000*"comunidades" + 0.000*"preenchidas_responderam" + 0.000*"fichas_preenchidas_responderam";
Tópico 24: 0.001*"portuária" + 0.001*"ceu" + 0.001*"quilombolas" + 0.001*"região" + 0.001*"zona_portuária" + 0.001*"zona" + 0.001*"ipn" + 0.001*"cidade" + 0.001*"biblioteca_comunitária" + 0.001*"comunitária";
Tópico 25: 0.013*"informação" + 0.004*"pesquisa" + 0.003*"conhecimento" + 0.003*"forma" + 0.002*"processo" + 0.002*"social" + 0.002*"trabalho" + 0.002*"dados" + 0.002*"comunicação" + 0.002*"biblioteca";
Tópico 26: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"biblioteca" + 0.000*"digital" + 0.000*"conhecimento" + 0.000*"museu" + 0.000*"forma" + 0.000*"gente" + 0.000*"trabalho" + 0.000*"documentos";
Tópico 27: 0.001*"sociologia" + 0.000*"sal" + 0.000*"bolsistas" + 0.000*"veronese" + 0.000*"sinfonia" + 0.000*"sal_prata"" + 0.000*"prata"" + 0.000*"reputação" + 0.000*"cena" + 0.000*"egressos";
Tópico 28: 0.002*"cartas" + 0.002*"carta" + 0.002*"maya" + 0.001*"castro_maya" + 0.001*"castro" + 0.001*"san" + 0.001*"jornal" + 0.001*"total_total" + 0.001*"publicada" + 0.001*"monde";
Tópico 29: 0.004*"arquivos" + 0.004*"arquivística" + 0.004*"documentos" + 0.003*"documental" + 0.003*"arquivologia" + 0.002*"identificação" + 0.002*"arquivo" + 0.001*"diplomática" + 0.001*"documentais" + 0.001*"arquivístico".

Fonte: Elaborado pelo autor.

Já com os resultados obtidos por meio do modelo LDA foi possível perceber que os termos apresentam pesos menores quando se comparado ao modelo LSI, além de muitos termos possuírem pesos igual a 0. Dessa forma, os pesos podem dificultar a interpretação do especialista ao realizar a suposição do nome de um determinado tópico, entretanto, constam, entre os resultados extraídos, um maior volume de termos N-gramas do tipo bigrama e trigrama, o que contribui para uma interpretação mais assertiva ao realizar a suposição do nome do tópico, já que esses termos possuem características que indicam termos especialistas, diferente dos termos generalistas.

É possível perceber uma redução no quantitativo de termos generalistas e um conjunto maior de termos especialistas, como por exemplo, no tópico 0, com os bigramas “segurança_informação”, “políticas_institucionais” e “mayara_petruso” acabam por representar assuntos mais especialistas do que os termos “segurança” ou “mayara”. Sozinhos esses termos podem apresentar diversas composições de assuntos dentro do domínio de linguagem. Caso o especialista tenha acesso aos documentos utilizados para a modelagem de tópicos, poderá obter mais informações como área de concentração ou linha de pesquisa de um determinado trabalho, tornando o resultado mais assertivo.

O tópico 18 pode ser considerado fraco por apresentar todos os termos genéricos, tais como “informação” ou “comunicação”, que podem conter composições para outros significados dentro do domínio da linguagem, por exemplo, “ciência_informação”, “informação_tecnológica”, “comunicação_científica” ou “processo_informação”. Além disso, todos os pesos possuem valor igual a zero, não sendo possível identificar uma ordem de relevância entre os termos.

É possível perceber dentre os resultados que a maioria dos tópicos possuem termos especialistas. O tópico 21, por exemplo, contempla em sua extração os resultados “mineração_dados”, “ontologias”, “difusas” e “fuzzy”, podendo ao especialista supor, por meio da análise de assuntos, que o tópico está relacionado à área de Organização e Representação do Conhecimento. Outro exemplo está no tópico 24, que apresenta termos específicos como “região_portuária”, “quilombolas”, “comunitária” e “ipn (Instituto de Pesquisas e Memória Pretos Novos)”, podendo supor que o assunto do tópico esteja relacionado, por exemplo, à Memória e Patrimônio ou Memória, Cultura e Patrimônio ou Memória e Aspectos Sociais.

Torna-se importante destacar, entre os resultados obtidos por meio do modelo LDA, uma menor quantidade de tópicos correlacionados. Além disso, é possível perceber uma menor quantidade de termos que apresentam o mesmo significado, o que possibilita espaço para alocação para novos termos especialistas que possuem características fortes e contribuem para realização da análise de assunto.

5.3.2. *Corpus* 7: artigos completos e resumos expandidos 2012

O *corpus* 7 é formado por 312 documentos e possui o tamanho equivalente a 12.328kb. Os dados desse *corpus* possuem 870.421 unigramas, 870.109 bigramas e 869.797 trigramas. O Quadro 18 apresenta uma lista contendo os 50 termos mais frequentes organizados por tipos de N-gramas.

Quadro 18 – Lista de N-gramas por ordem de frequência do *corpus* 7

Unigramas
informação,19056; pesquisa,4683; conhecimento,4207; dados,2824; forma,2713; processo,2648; social,2588; comunicação,2567; uso,2292; biblioteca,2256; museu,2228; produção,2176; trabalho,2164; memória,2106; documentos,1899; organização,1872;

brasil,1847; relação,1842; estudo,1715; desenvolvimento,1703; sociais,1698; sociedade,1677; científica,1663; estudos,1659; paulo,1635; sistema,1609; meio,1580; cultura,1565; campo,1562; termo,1544; busca,1519; contexto,1510; digital,1435; usuários,1408; informacional,1380; diferentes,1345; tempo,1320; autores,1270; construção,1267; cultural,1260; relações,1255; tecnologia,1251; gestão,1226; resultados,1213; rede,1199; nacional,1193; história,1180; processos,1176; periódicos,1154; objeto,1152.
Bigramas
informação_science,563; recuperação_informação,497; informação_conhecimento,474; informação_tecnologia,458; universidade_federal,445; produção_científica,370; informação_comunicação,369; redes_sociais,334; uso_informação,333; belo_horizonte,327; sistemas_informação,297; comunicação_científica,292; gestão_informação,277; comunicação_oral,265; tecnologias_informação,265; oswaldo_cruz,263; ponto_vista,257; informação_informação,251; competência_informacional,246; fontes_informação,237; bases_dados,230; tomada_decisão,221; dissertação_mestrado,216; informação_brasília,211; coleta_dados,201; comportamento_informacional,199; muitas_vezes,188; periódicos_científicos,186; organização_informação,183; gestão_conhecimento,182; ensino_superior,180; patrimônio_cultural,179; sociedade_informação,174; base_dados,174; inclusão_digital,174; informação_científica,174; organização_conhecimento,173; xiii_xiii,172; novas_tecnologias,165; biblioteca_digitais,162; pesquisa_informação,161; arquitetura_informação,161; busca_informação,160; teses_dissertações,152; objeto_estudo,151; minas_gerais,147; direito_informação,144; tendo_vista,142; memória_social,140; universidade_estadual,139.
Trigramas
tecnologias_informação_comunicação,168; fundação_oswaldo_cruz,138; gestão_informação_conhecimento,114; american_society_informação,114; society_informação_science,114; journal_american_society,99; universidade_federal_paraíba,96; fonte_dados_pesquisa,87; informação_science_technology,83; dissertação_mestrado_informação,81; international_organization_standardization,81; encontro_nacional_pesquisa,79; nacional_pesquisa_informação,78; resource_description_framework,72; ambientes_informacionais_digitais,70; associação_arquivistas_brasileiros,69; informação_belo_horizonte,65; modalidade_apresentação_comunicação,64; apresentação_comunicação_oral,64; library_informação_science,62; informação_universidade_federal,60; informação_tecnologia_was,60; portal_periódicos_capes,58; organização_representação_conhecimento,57; pesquisa_cultural_listado,57; cultural_listado_edital,57; brasília_briquet_lemos,55; informação_sociedade_estudos,53; busca_uso_informação,48; federal_minas_gerais,46; universidade_federal_minas,45 patrimônio_histórico_artístico,44; instituições_ensino_superior,43; universidade_federal_santa,42; world_wide_web,41; circulação_apropriação_informação,40; machine_readable_cataloging,40; novas_tecnologias_informação,40; xiii_xiii_informação,40; perspectivas_informação_belo,39; exame_nacional_desempenho,39; históricos_epistemológicos_informação,38; sob_ponto_vista,38; sistema_recuperação_informação,38; mediação_circulação_apropriação,37; produção_comunicação_informação,37; paulo_martins_fontes,37; datagramazero_revista_informação,37; zaphiris_kurniawan_ghiawadwala,37; sistemas_recuperação_informação,36.

Fonte: Elaborado pelo autor.

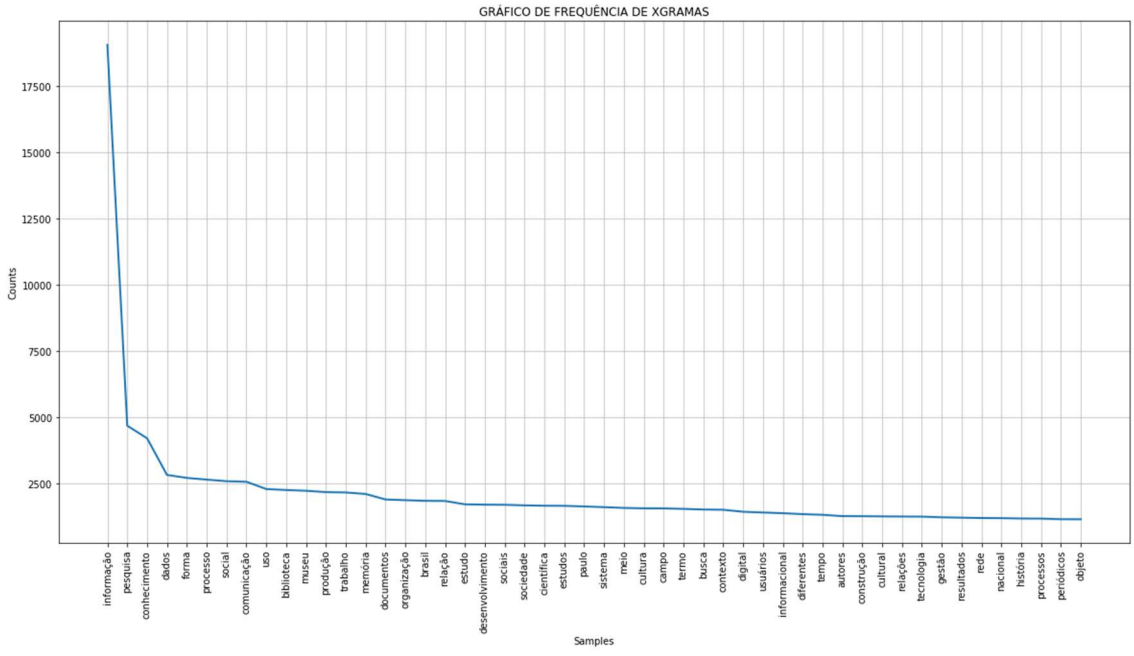
Numa lista geral contendo os mil N-gramas mais frequentes extraídos do *corpus* de dados estão os bigramas: “informação_science”, com frequência de 563 na posição 179; “recuperação_informação”, com frequência de 497 e

posição 222; “informação_conhecimento, com frequência de 474 na posição 241; “informação_tecnologia”, com frequência de 458 na posição 261; e “universidade_federal”, com frequência de 445 e posição 270. Já os trigramas melhores ranqueados são: “tecnologias_informação_comunicação”, com frequência de 168 na posição 936; “fundação_oswaldo_cruz”, com frequência de 138; “gestão_informação_conhecimento”, “american_society_informação” e “society_informação_science”, todos com frequência de 114 e ranqueados após o milésimo termo.

Os termos do tipo unigramas apresentam maior frequência quando se comparados aos outros tipos de N-gramas. Ao comparar, por exemplo, o primeiro unigrama representado pelo termo “informação” como frequência de 19.059 ao primeiro bigrama “informação_science” e frequência de 563, encontra-se uma diferença de 3.285%. Esse percentual aumenta para 11.245% quando comparado ao primeiro unigrama “informação” com o primeiro trigrama, representado pelo termo “tecnologias_informação_comunicação”, com frequência de 168.

O Gráfico 20 apresenta os 50 N-gramas mais frequentes extraídos do *corpus* 7. Os resultados apresentam somente termos do tipo unigramas, sendo os primeiros: “informação”, com frequência de 19.056; “pesquisa”, com frequência de 4.683; “conhecimento”, com frequência de 4.207; “dados”, com frequência de 2.824; e “forma”, com frequência de 2.713.

Gráfico 20 – Termos mais frequentes do *corpus* 7

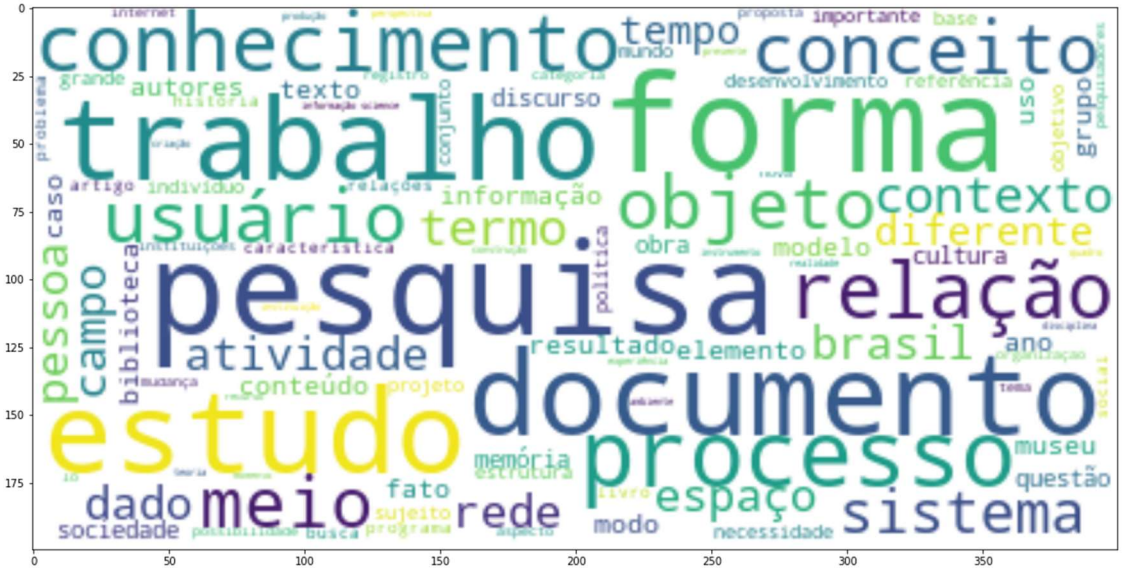


Fonte: Elaborado pelo autor.

É possível perceber no gráfico um crescimento constante entre o 50º e o 2º termo, respectivamente, “objeto” com frequência de 1.152 e “pesquisa” com frequência de 4.683. Entretanto, há um salto no quantitativo de frequência dos termos quando se comparado ao primeiro unigrama, representado pelo termo “informação” e com frequência de 19.056. No gráfico também são encontrados termos fortes como “memória”, “documentos”, “organização”, “informacional’ e “processo”. Eles podem ser explorados junto as listas de bigramas e trigramas a fim de encontrar resultados mais sólidos, uma vez que os termos podem sofrer composições para outros significados. Também é possível encontrar termos fracos como “autores”, “resultados”, “meio”, “diferentes” e “contexto”, que podem ser descartados das análises por não contribuírem ao domínio da linguagem.

A Figura 10 ilustra uma nuvem de palavras constituída pelos 250 termos mais frequentes contidos no *corpus* de dados. Diferente do primeiro *corpora*, constituídos pelos seis primeiros *corpus* de dados, o *corpus* em questão apresenta, embora seja um número reduzido, um maior quantitativo de bigramas. Para a construção da figura foram utilizados dados quantitativos, não sendo realizados quaisquer tipos de análise qualitativa.

Figura 10 – Nuvem de palavras do *corpus* 7



Fonte: Elaborado pelo autor.

Como resultado, a imagem apresenta termos fortes e fracos quando se comparado ao domínio da linguagem ou realizada a análise de assunto apoiada nos bigramas e trigramas extraídos do *corpus* de dados. São exemplos de

tópicos fortes os unigramas “museu”, “dado” e “conhecimento” que, mesmo com características generalistas, podem remeter a outros significados embasados nos bigramas e trigramas como termos de relevância para o domínio de linguagem. Dentre os termos com características fracas estão unigramas “tempo”, “modo” e “perspectiva”, que não apresentam conceitos relacionados ao domínio de linguagem. Pode-se destacar neste *corpus* um maior quantitativo de termos fortes que, comparado ao *corpora* 1, é constituído por seis *corpus* de dados.

Com os dados conectados aos modelos de treinamento, foi possível realizar a extração dos tópicos, termos e pesos do *corpus* de dados. O Quadro 19 apresenta uma lista contendo 30 tópicos extraídos por meio do modelo LSI, levando o tempo de 32.2 segundos para execução. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[53].

Quadro 19 – Tópicos extraídos do *corpus* 7 usando o modelo LSI

Tópico 0: 0.780*"informação" + 0.157*"conhecimento" + 0.150*"pesquisa" + 0.092*"dados" + 0.090*"processo" + 0.090*"forma" + 0.085*"social" + 0.085*"comunicação" + 0.082*"uso" + 0.072*"trabalho";
Tópico 1: 0.576*"museu" + -0.402*"informação" + 0.240*"memória" + 0.151*"cultural" + 0.128*"pesquisa" + 0.115*"patrimônio" + 0.112*"cultura" + 0.092*"brasil" + 0.091*"social" + 0.090*"história";
Tópico 2: -0.577*"museu" + -0.310*"informação" + 0.187*"biblioteca" + 0.170*"pesquisa" + 0.163*"científica" + 0.157*"dados" + 0.153*"periódicos" + 0.119*"produção" + -0.117*"virtual" + 0.116*"artigos";
Tópico 3: 0.384*"memória" + -0.323*"museu" + -0.236*"dados" + 0.187*"social" + -0.164*"periódicos" + -0.159*"pesquisa" + 0.156*"cultura" + 0.154*"conhecimento" + -0.140*"científica" + -0.135*"artigos";
Tópico 4: -0.396*"biblioteca" + 0.245*"científica" + 0.208*"produção" + -0.197*"usuários" + 0.173*"periódicos" + 0.162*"artigos" + -0.151*"usuário" + -0.138*"digital" + 0.136*"pesquisadores" + 0.116*"brasil";
Tópico 5: -0.517*"conhecimento" + 0.297*"memória" + 0.271*"biblioteca" + -0.176*"gestão" + -0.159*"linguagem" + -0.141*"organização" + -0.138*"processo" + 0.132*"informação" + -0.116*"museu" + 0.096*"pesquisa";
Tópico 6: 0.295*"documentos" + -0.266*"biblioteca" + 0.249*"dados" + -0.206*"conhecimento" + 0.193*"memória" + 0.157*"documento" + -0.144*"social" + -0.137*"rede" + -0.134*"informacional" + -0.129*"educação";
Tópico 7: -0.281*"biblioteca" + 0.244*"rede" + 0.211*"digital" + 0.199*"dados" + -0.183*"conhecimento" + 0.181*"redes" + 0.177*"comunicação" + 0.172*"web" + -0.159*"documentos" + 0.154*"sociais";
Tópico 8: 0.357*"memória" + 0.285*"conhecimento" + 0.277*"dados" + -0.200*"documentos" + 0.172*"gestão" + 0.149*"cultural" + -0.142*"documento" + -0.131*"linguagem" + 0.129*"cultura" + 0.125*"pesquisa";

[53] Algoritmo de modelagem de tópicos. *Corpus* 7: artigos completos e resumos expandidos 2012. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2012.ipynb/.

Tópico 9: -0.304*"memória" + -0.161*"periódicos" + 0.159*"brasil" + 0.155*"trabalho" + 0.153*"dados" + 0.150*"documentos" + 0.150*"política" + 0.146*"gestão" + -0.143*"linguagem" + 0.142*"defesa";
Tópico 10: 0.433*"digital" + -0.339*"dados" + -0.193*"biblioteca" + 0.184*"digitais" + 0.174*"preservação" + 0.173*"documentos" + 0.136*"memória" + -0.131*"data" + -0.111*"web" + -0.106*"linked";
Tópico 11: 0.336*"cultural" + 0.321*"cultura" + -0.269*"memória" + -0.164*"trabalho" + 0.161*"culturais" + 0.150*"pesquisa" + -0.144*"rede" + -0.126*"documentos" + -0.124*"redes" + -0.120*"dados";
Tópico 12: -0.365*"biblioteca" + -0.215*"defesa" + 0.202*"educação" + -0.190*"conhecimento" + 0.186*"pesquisa" + 0.176*"informacional" + -0.175*"direito" + -0.171*"política" + 0.139*"competência" + -0.124*"gestão";
Tópico 13: -0.269*"defesa" + 0.222*"rede" + -0.181*"linguagem" + -0.164*"educação" + 0.163*"redes" + 0.147*"cultura" + -0.147*"memória" + -0.142*"brasil" + -0.139*"conhecimento" + -0.133*"política";
Tópico 14: 0.292*"uso" + 0.249*"periódicos" + -0.200*"digital" + 0.197*"defesa" + 0.166*"portal" + -0.152*"científica" + 0.139*"usuários" + -0.133*"repositórios" + -0.118*"livros" + -0.114*"metadados";
Tópico 15: -0.219*"campo" + 0.190*"direito" + 0.183*"informacional" + 0.164*"obra" + 0.164*"registros" + 0.162*"bibliográficos" + 0.159*"periódicos" + -0.137*"usuários" + -0.130*"avaliação" + -0.128*"usabilidade";
Tópico 16: 0.264*"direito" + -0.232*"defesa" + -0.205*"registros" + -0.196*"bibliográficos" + -0.187*"obra" + -0.156*"pesquisa" + -0.145*"registros_bibliográficos" + 0.145*"lei" + -0.129*"rede" + 0.118*"cultura";
Tópico 17: 0.293*"texto" + 0.258*"textos" + -0.181*"documentos" + 0.180*"moraes" + 0.158*"percurso" + -0.136*"campo" + -0.134*"documento" + -0.126*"informacional" + 0.126*"ficção" + -0.120*"cultura";
Tópico 18: 0.228*"pesquisa" + 0.216*"direito" + -0.211*"informacional" + 0.159*"rede" + 0.157*"conhecimento" + -0.150*"produção" + 0.146*"linguagem" + -0.138*"contexto" + 0.132*"comunicação" + -0.123*"comportamento";
Tópico 19: -0.234*"trabalho" + 0.230*"rede" + -0.181*"processo" + 0.173*"documentos" + -0.173*"direito" + 0.129*"redes" + 0.125*"sociais" + 0.121*"biblioteca" + 0.113*"documental" + 0.110*"cultura";
Tópico 20: -0.211*"trabalho" + 0.203*"livros" + 0.178*"maya" + 0.178*"castro" + 0.175*"castro_maya" + 0.173*"brasil" + 0.161*"arte" + 0.139*"organização" + 0.115*"coleção" + -0.110*"cultural";
Tópico 21: -0.175*"processo" + 0.172*"obra" + 0.166*"campo" + 0.158*"bibliográficos" + 0.150*"conhecimento" + 0.149*"registros" + 0.129*"direito" + -0.127*"trabalho" + 0.121*"usuários" + 0.118*"registros_bibliográficos";
Tópico 22: -0.195*"cultura" + -0.166*"organização" + -0.161*"classificação" + -0.154*"sistema" + 0.142*"digital" + 0.139*"estudos" + -0.129*"comunicação" + 0.126*"linguagem" + -0.110*"metadados" + 0.098*"processo";
Tópico 23: 0.202*"uso" + 0.183*"livros" + 0.150*"trabalho" + -0.148*"avaliação" + 0.133*"portal" + -0.126*"atores" + -0.125*"memória" + -0.122*"educação" + -0.117*"direito" + -0.112*"usabilidade";
Tópico 24: -0.323*"trabalho" + 0.181*"comunicação" + 0.171*"pesquisa" + 0.140*"linguagem" + -0.136*"digital" + -0.126*"periódicos" + 0.119*"livros" + -0.118*"preservação" + 0.113*"científica" + -0.111*"classificação";
Tópico 25: -0.263*"linguagem" + -0.154*"arquivistas" + 0.126*"classificação" + -0.125*"brasileiros" + -0.116*"organização" + -0.110*"arquivistas_brasileiros" + -0.105*"social" + -0.103*"associação_arquivistas_brasileiros" + -0.103*"associação_arquivistas" + -0.101*"metadados";

Tópico 26: 0.236*"trabalho" + -0.176*"arte" + -0.128*"avaliação" + -0.126*"preservação" + 0.120*"arquivistas" + -0.108*"produção" + -0.106*"oiticica" + -0.102*"uso" + 0.101*"brasileiros" + 0.101*"direito";
Tópico 27: 0.188*"usuários" + 0.156*"maya" + 0.153*"castro_maya" + 0.151*"castro" + -0.147*"digital" + -0.121*"campo" + 0.116*"documentos" + 0.114*"coleção" + 0.111*"arte" + 0.111*"web";
Tópico 28: -0.280*"cruz" + -0.265*"oswaldo" + -0.263*"oswaldo_cruz" + -0.195*"univ" + -0.189*"comunicação" + 0.129*"informacional" + 0.115*"direito" + -0.111*"fundação_oswaldo_cruz" + -0.111*"fundação_oswaldo" + -0.110*"fundação";
Tópico 29: -0.159*"livros" + 0.158*"informacional" + 0.141*"cruz" + 0.138*"oswaldo" + 0.136*"oswaldo_cruz" + 0.133*"classificação" + 0.129*"univ" + -0.129*"patrimônio" + 0.118*"pesquisa" + -0.110*"mediação".

Fonte: Elaborado pelo autor.

É possível observar, com base nos resultados obtidos por meio do modelo LSI, a existência de termos fortes com pesos acima de 0.500, como “informação”, no tópico 0, e “museu”, nos tópicos 1 e 2. Os termos contidos nos tópicos são apresentados sem ordem de relevância.

O tópico 0 pode ser caracterizado como um tópico generalista por apresentar termos gerais ao domínio da linguagem. Esses termos, quando aprofundados, podem possuir diversas composições, como por exemplo, “informação”, que pode estar associado a termos mais significativos como “ciência_informação”, gestão_informação” ou “informação_conhecimento”. Cabe ao especialista no domínio da linguagem decidir, por exemplo, pela exclusão de tópicos generalistas ou criar uma categoria generalista para tópicos com tais características.

A identificação de termos específicos nos resultados contribui para uma interpretação mais assertiva pelo especialista. O tópico 10, por exemplo, apresenta termos como “digital”, “linked” e “data” e pode ser interpretado pelo especialista como um tópico que aborda assunto sobre Informação e Tecnologia. O tópico pode receber outro rótulo ao se realizar também a interpretação do termo “preservação”, podendo supor por exemplo que se trata de um tópico relacionado a Organização e Representação do Conhecimento.

O tópico 23 também apresenta características fortes em seus termos como “uso”, “portal”, “usabilidade” e “direito”. O especialista pode supor que o tópico aborde assuntos relacionados à Informação e Tecnologia ou Política e Economia da Informação. É possível identificar tópicos com termos específicos, como “castro_maya”, no tópico 27, e “fundação_oswaldo_cruz”, no tópico 28. Essa identificação possibilita ao especialista explorar o *corpus* de dados – caso

tenha acesso – antes mesmo de realizar a interpretação dos tópicos, podendo realizar a suposição de um rótulo de maneira mais assertiva para o tópico. Esses termos, embora não possuam relevância para domínio da linguagem, apresentam características específicas para busca por palavra-chave dentro de um *corpus* de dados, pois, além de serem termos gatilhos, podem não apresentar grandes volumes de documentos no *corpus* de dados.

O Quadro 20 apresenta um conjunto formado por 30 tópicos e seus respectivos termos e pesos extraídos do *corpus* de dados por meio do modelo LDA. O processo de treinamento foi executado em 7 minutos e 18 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[54].

Quadro 20 – Tópicos extraídos do *corpus* 7 usando o modelo LDA

Tópico 0: 0.000*"informação" + 0.000*"biblioteca" + 0.000*"memória" + 0.000*"social" + 0.000*"paulo" + 0.000*"pesquisa" + 0.000*"cultural" + 0.000*"brasil" + 0.000*"cultura" + 0.000*"belo";
Tópico 1: 0.007*"informação" + 0.002*"conhecimento" + 0.002*"biblioteca" + 0.002*"pesquisa" + 0.001*"linguagem" + 0.001*"uso" + 0.001*"social" + 0.001*"comunicação" + 0.001*"sistema" + 0.001*"forma";
Tópico 2: 0.002*"cultural" + 0.002*"pesquisa" + 0.001*"pesquisa_cultural" + 0.001*"edital" + 0.001*"listado" + 0.001*"pernambuco" + 0.001*"cultural_listado_edital" + 0.001*"listado_edital" + 0.001*"pesquisa_cultural_listado" + 0.001*"cultural_listado";
Tópico 3: 0.002*"periódicos" + 0.002*"artigos" + 0.002*"produção" + 0.002*"científica" + 0.002*"informação" + 0.001*"trabalho" + 0.001*"pesquisa" + 0.001*"citação" + 0.001*"indicadores" + 0.001*"estudos";
Tópico 4: 0.001*"sinasc" + 0.000*"sis" + 0.000*"cultura_informacional" + 0.000*"cultura_organizacional" + 0.000*"registro" + 0.000*"preenchimento" + 0.000*"estrutura_organizacional" + 0.000*"nova_estrutura" + 0.000*"nova_estrutura_organizacional" + 0.000*"bloco";
Tópico 5: 0.000*"informação" + 0.000*"memória" + 0.000*"conhecimento" + 0.000*"social" + 0.000*"científica" + 0.000*"cultura" + 0.000*"pesquisa" + 0.000*"produção" + 0.000*"sociedade" + 0.000*"brasil";
Tópico 6: 0.002*"preservação" + 0.001*"digital" + 0.001*"preservação_digital" + 0.001*"liber" + 0.001*"digitais" + 0.000*"patente" + 0.000*"malária" + 0.000*"diretrizes" + 0.000*"acervos" + 0.000*"objetos";
Tópico 7: 0.001*"negros" + 0.001*"planeta" + 0.001*"origem" + 0.001*"macacos" + 0.001*"mito" + 0.001*"planeta_macacos" + 0.000*"memória_iconográfica" + 0.000*"iconográfica" + 0.000*"jornada_estrelas" + 0.000*"spock";
Tópico 8: 0.002*"informação" + 0.002*"memória" + 0.001*"pesquisa" + 0.001*"documentos" + 0.001*"nacional" + 0.001*"museu" + 0.001*"preservação" + 0.001*"brasil" + 0.001*"oiticica" + 0.001*"periódicos";

[54] Algoritmo de modelagem de tópicos. *Corpus* 7: artigos completos e resumos expandidos 2012. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2012.ipynb/.

Tópico 9: 0.002*"livros" + 0.001*"ex-votos" + 0.001*"scielo" + 0.001*"cartas" + 0.001*"editoras" + 0.001*"scielo_livros" + 0.001*"brasil" + 0.001*"projeto" + 0.001*"publicação" + 0.001*"memória";
Tópico 10: 0.001*"biblioteca" + 0.001*"livros" + 0.001*"tradução" + 0.001*"livro" + 0.000*"mulheres" + 0.000*"fforde" + 0.000*"literatura" + 0.000*"thursday" + 0.000*"dewey" + 0.000*"bibliotecária";
Tópico 11: 0.013*"informação" + 0.003*"conhecimento" + 0.002*"pesquisa" + 0.002*"processo" + 0.002*"forma" + 0.002*"dados" + 0.001*"trabalho" + 0.001*"organização" + 0.001*"social" + 0.001*"relação";
Tópico 12: 0.002*"informação" + 0.002*"comunicação" + 0.001*"processo" + 0.001*"nacional" + 0.001*"sociedade" + 0.001*"fiscal" + 0.001*"industrial" + 0.001*"produção" + 0.001*"paulo" + 0.001*"patrimônio";
Tópico 13: 0.001*"informação" + 0.001*"conhecimento" + 0.001*"metadados" + 0.001*"dados" + 0.001*"documentos" + 0.001*"social" + 0.001*"rede" + 0.001*"verdade" + 0.001*"sociais" + 0.001*"campo";
Tópico 14: 0.000*"informação" + 0.000*"trabalho" + 0.000*"pesquisa" + 0.000*"forma" + 0.000*"museu" + 0.000*"documentos" + 0.000*"conhecimento" + 0.000*"produção" + 0.000*"processo" + 0.000*"campo";
Tópico 15: 0.001*"arquivos" + 0.001*"documentos" + 0.000*"sensíveis" + 0.000*"justiça" + 0.000*"regime" + 0.000*"repressão" + 0.000*"verdade" + 0.000*"thiesen" + 0.000*"memória" + 0.000*"história";
Tópico 16: 0.002*"trabalho" + 0.001*"autores" + 0.001*"grupos" + 0.001*"informação" + 0.001*"poder" + 0.001*"grupos_trabalho" + 0.001*"rede" + 0.001*"bourdieu" + 0.001*"braman" + 0.001*"redes";
Tópico 17: 0.001*"representação" + 0.001*"internet" + 0.001*"conhecimento" + 0.001*"documentos" + 0.001*"informação" + 0.001*"imagens" + 0.001*"descrição" + 0.001*"ontologia" + 0.001*"arquivística" + 0.001*"ontologias";
Tópico 18: 0.009*"informação" + 0.002*"dados" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.002*"museu" + 0.001*"social" + 0.001*"uso" + 0.001*"contexto" + 0.001*"brasil" + 0.001*"forma";
Tópico 19: 0.004*"informação" + 0.001*"pesquisa" + 0.001*"culturais" + 0.001*"cultural" + 0.001*"campo" + 0.001*"memória" + 0.001*"patrimônio" + 0.001*"categorização" + 0.001*"cultura" + 0.001*"artigos";
Tópico 20: 0.001*"brasil" + 0.001*"registro" + 0.001*"marcas" + 0.001*"informação" + 0.001*"positivismo" + 0.001*"história" + 0.001*"pernambuco" + 0.000*"comte" + 0.000*"exposição" + 0.000*"nacional";
Tópico 21: 0.006*"informação" + 0.002*"direito" + 0.002*"uso" + 0.002*"pesquisa" + 0.001*"lei" + 0.001*"biblioteca" + 0.001*"direito_informação" + 0.001*"portal" + 0.001*"periódicos" + 0.001*"dados";
Tópico 22: 0.003*"biblioteca" + 0.002*"informação" + 0.002*"pesquisa" + 0.001*"avaliação" + 0.001*"produção" + 0.001*"digitais" + 0.001*"usabilidade" + 0.001*"arquitetura" + 0.001*"biblioteca_digitais" + 0.001*"dados";
Tópico 23: 0.004*"informação" + 0.001*"obra" + 0.001*"pesquisa" + 0.001*"cruz" + 0.001*"oswaldo" + 0.001*"oswaldo_cruz" + 0.001*"autores" + 0.001*"artigos" + 0.001*"comunicação" + 0.001*"relações";
Tópico 24: 0.010*"informação" + 0.003*"pesquisa" + 0.002*"memória" + 0.002*"museu" + 0.002*"conhecimento" + 0.001*"usuários" + 0.001*"forma" + 0.001*"documentos" + 0.001*"dados" + 0.001*"comunicação";
Tópico 25: 0.002*"informação" + 0.002*"cultura" + 0.002*"cultural" + 0.001*"culturais" + 0.001*"pesquisa" + 0.001*"museu" + 0.001*"acervo" + 0.001*"políticas" + 0.001*"pesquisas" + 0.001*"política";

Tópico 26: 0.001*"fotografia" + 0.001*"belo" + 0.001*"belo_horizonte" + 0.001*"horizonte" + 0.001*"informação" + 0.001*"museu" + 0.001*"cultura" + 0.001*"crav" + 0.001*"pesquisa" + 0.001*"cultural";
Tópico 27: 0.003*"memória" + 0.002*"informação" + 0.002*"biblioteca" + 0.001*"cultura" + 0.001*"paulo" + 0.001*"social" + 0.001*"patrimônio" + 0.001*"cultural" + 0.001*"instrumentos" + 0.001*"conhecimento";
Tópico 28: 0.005*"museu" + 0.002*"informação" + 0.001*"comunicação" + 0.001*"exposição" + 0.001*"museologia" + 0.001*"exposições" + 0.001*"objetos" + 0.001*"objeto" + 0.001*"castro" + 0.001*"maya";
Tópico 29: 0.001*"documento" + 0.001*"informação" + 0.001*"brazil" + 0.001*"produção" + 0.001*"citação" + 0.001*"vital" + 0.001*"objetos" + 0.001*"tempo" + 0.001*"linguagem" + 0.001*"trabalho".

Fonte: Elaborado pelo autor.

É possível observar que os resultados obtidos por meio do modelo LDA apresentam características diferentes quando se comparados aos resultados do modelo LSI. Essas diferenças contribuem para que o especialista realize uma interpretação dos dados com menor esforço cognitivo e de maneira mais assertiva. Dentre elas, constam um maior número de N-gramas dos tipos bigramas e trigramas entre os termos dos tópicos e uma menor quantidade de termos generalistas, que podem se compor em diversos significados. Entretanto, torna-se possível observar uma fragilidade entre os resultados com baixos valores dos pesos atribuídos aos termos. Em muitos casos, esses valores são iguais a 0, o que dificulta a interpretação da ordem de relevância dos termos.

É possível encontrar, entre os resultados, tópicos que contemplam termos com características fracas, como por exemplo, “paulo”, “brasil” e “belo”, extraídos no tópico 0. Além disso, todos os demais termos do tópico possuem peso com valor igual a 0, enfatizando a fragilidade do tópico. Uma solução para realizar um melhor ajuste do modelo está na adição dos termos na fase de pré-processamento, que diz respeito à limpeza dos dados.

Também é possível encontrar tópicos que apresentam termos especialistas, a fim de reduzir o esforço cognitivo por parte do especialista ao realizar a interpretação dos dados por meio de análise de assuntos e, posteriormente, definir a suposição de um rótulo para o tópico. Dentre eles, destacam-se o tópico 3, que apresenta os termos “periódicos”, “artigos”, “produção”, “científica”, “citação” e “indicadores”. Eles podem remeter, por exemplo, ao tema Comunicação Científica ou Produção e Comunicação da Informação.

O tópico 7 também pode ser considerado especialista por conter termos com características fortes, bem como “planeta_macacos”, “jornada_estrelas”, “spock” e “mito”, que permitem ao especialista, por exemplo, investigar documentos externos como o próprio *corpus* de dados e identificar de maneira mais assertiva a área ou assunto do tópico por meio do título, resumo ou GT de trabalho, o que nesse caso, remete ao rótulo relacionado à Informação e Memória.

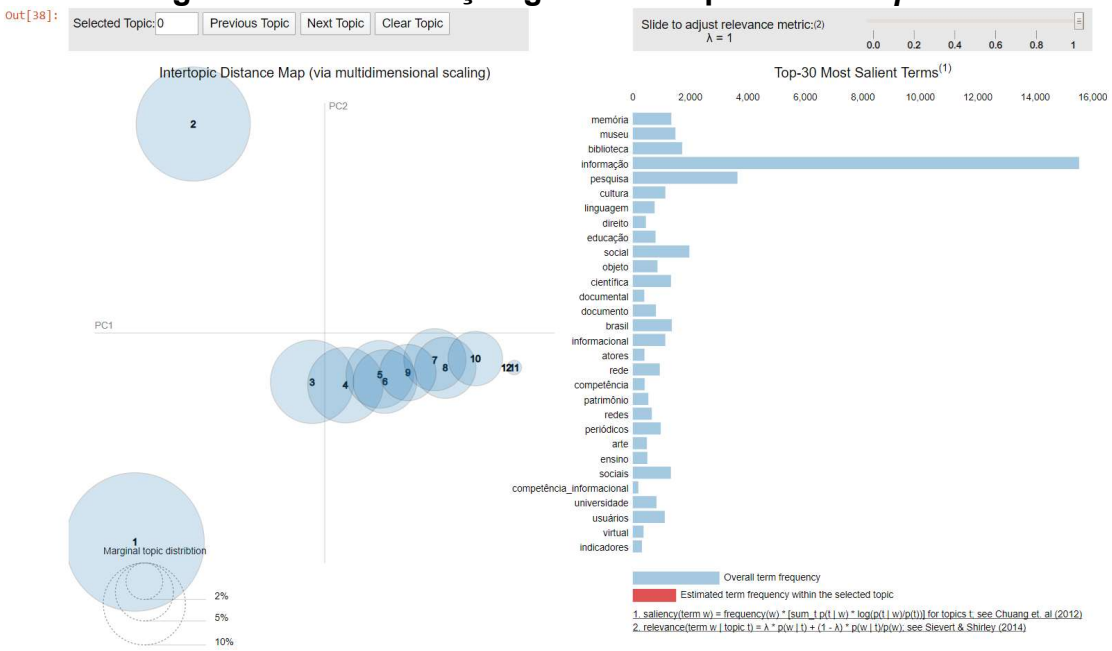
Também é possível encontrar entre os resultados tópicos que abordem mais de uma área ou assunto, como por exemplo, os apresentados no tópico 4. Ele possui dois termos específicos com pesos: “sinasc (Sistema de Informações sobre Nascidos Vivos)” e “sis (Sistemas de Informação em Saúde)”, relacionados à área de Informação e Saúde, e os termos “cultura_informacional”, “cultura_organizacaional”, “estrutura_organizacional” e “nova_estrutura_organizacional”, com pesos igual a zero e que podem remeter ao especialista ao rótulo de Gestão da Informação e do Conhecimento.

O tópico 6 também apresenta as mesmas características que o tópico 4. Contém termos especialistas como “preservação”, “digital” e “liber”, que permitem explorar o *corpus* de dados e identificar o documento que melhor representa tais termos, para, posteriormente, identificar a área ou assunto do tópico. Nesse caso, pode supor que se trata de um rótulo relacionado à Informação e Tecnologia. Além disso, constam entre os resultados termos especialistas com peso igual a zero e de áreas diferentes, bem como “patente” e “malária”, remetendo, por meio de exploração do *corpus* de dados, à área de Informação e Saúde.

Outra maneira que contribui para o desenvolvimento do trabalho do especialista na definição da suposição dos nomes dos tópicos extraídos do *corpus* de dados está na visualização dinâmica dos tópicos, construída a partir dos resultados do modelo LDA e utilizando a biblioteca pyLDAvis. A Figura 11 apresenta, em seu lado esquerdo, uma visão global do modelo de tópicos em que estão plotados os tópicos em formato de círculos no plano bidimensional, sendo os centros utilizados cálculos para identificar a distância entre os tópicos. Posteriormente, é usado o dimensionamento multidimensional para projetar as distâncias intertópicas em duas dimensões. À direita da imagem é apresentado o gráfico de barras, que representa os termos individuais, úteis para realizar a

interpretação do tópico selecionado. O *download* do arquivo para a visualização dinâmica dos tópicos no formato HTML pode ser realizado através do GitHub[55].

Figura 11 – Visualização geral dos tópicos do *corpus* 7

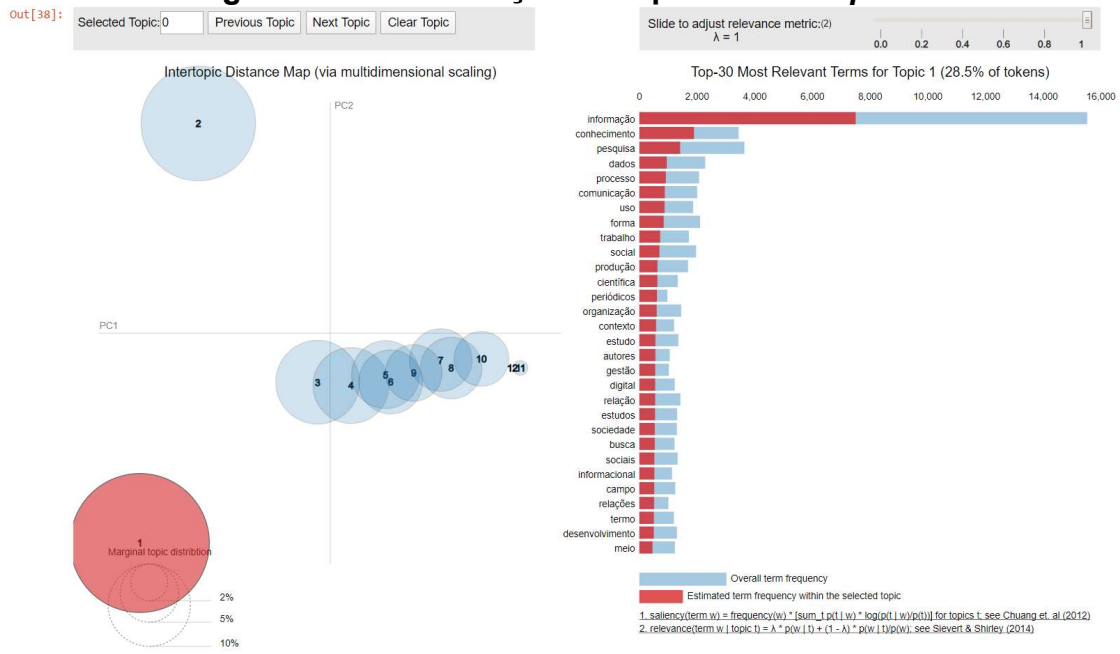


Fonte: Elaborado pelo autor.

A Figura 12 apresenta, à direita da imagem, os 30 termos mais relevantes do tópico 1 representado à esquerda da imagem. Os termos representam 28.5% dos *tokens* quando ajustados à métrica de relevância para valor igual a 1.0. Esses termos são representados nas cores azul para a frequência geral dos termos e vermelho para frequência estimada do termo no tópico selecionado.

[55] Visualização dinâmica dos tópicos. *Corpus* 7: artigos completos e resumos expandidos 2012. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2012_gts.html/.

Figura 12 – Visualização do tópico 1 do *corpus* 7



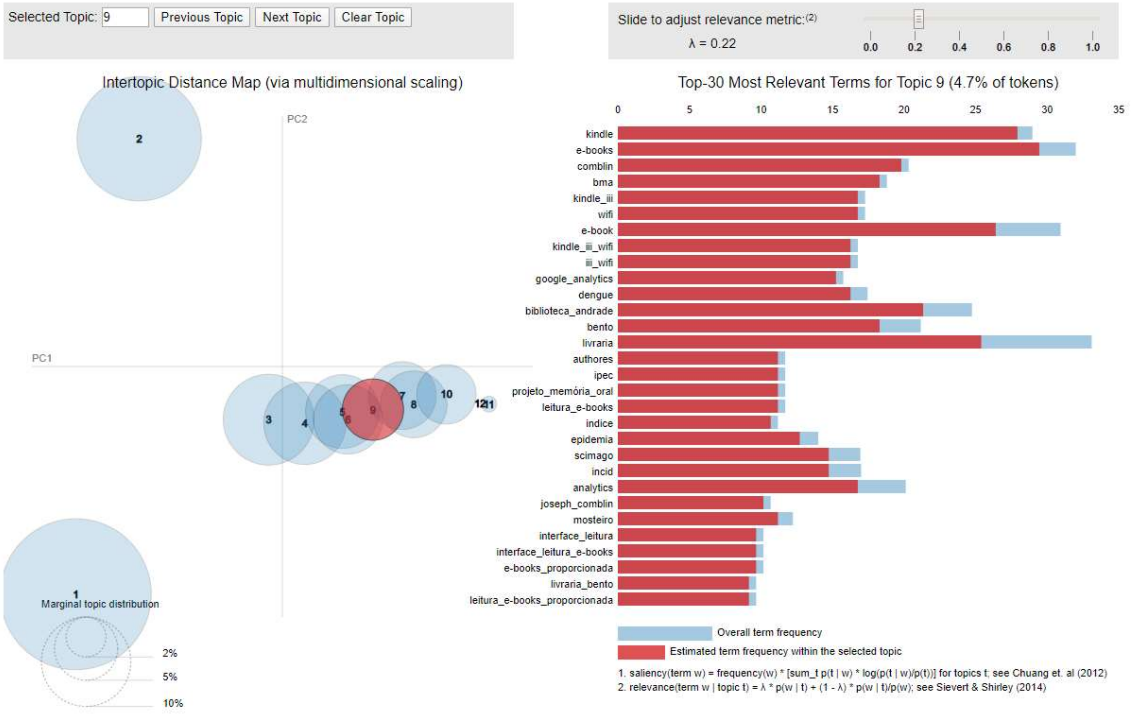
Fonte: Elaborado pelo autor.

O tópico 1 possui termos fortes, entretanto, acabam por se caracterizar como generalistas a área do *corpus* estudada, como por exemplo, “informação”, “conhecimento”, “pesquisa”, “dados’ e “processo”, assuntos abordados em diversas outras áreas do domínio da linguagem. Quando os valores da métrica são reduzidos para 0.2 é possível identificar termos mais específicos, tendo maior representatividade em relação aos termos generalistas e auxiliando na identificação de um possível rótulo para o tópico, como por exemplo, os bigramas “gestão_conhecimento”, “preservação_digital”, “informação_conhecimento”, “periódicos_eletronicos” e “informação_orgânica”. Por meio desses ajustes, o especialista pode visualizar diversas opções que auxiliem na definição da suposição dos tópicos.

Os tópicos mais ao centro do plano bidimensional possuem inter-relação entre os termos com maior frequência contidos em cada tópico, tais como “informação”, “pesquisa”, “biblioteca”, “conhecimento” e “forma” quando ajustada a métrica para 1.0. Ao analisar o tópico 9 com a métrica no valor 0.2, os termos apresentam características mais fortes, tais como “kindle”, “kindle_wifi” ,“e-books”, “Comblin” (Centro de Pesquisa e Documentação José Comblin) , “bma” (Biblioteca Mário Andrade), “leitura_e-books”, “interface_ebooks”, “leitura_interface_e-books”, conforme apresentado na Figura 13. Dessa forma, pode-se supor que o tópico esteja relacionado, por exemplo, à Informação e

Tecnologia, enquadrando-se em Aplicações de Informação em disciplinas e subdisciplinas do campo da Ciência da Informação e Sistemas de Informação e Tecnologia da Informação em temas/assuntos e disciplinas, conforme estudos de Pinheiro (2006).

Figura 13 – Visualização do tópico 9 do *corpus* 7



Fonte: Elaborado pelo autor.

Outro ponto a ser destacado está na proximidade entre os tópicos 3 e 10. Há uma correlação entre os principais termos contidos entre os tópicos, exigindo ao especialista uma navegação variada por meio de ajustes de métricas de relevância entre os tópicos.

5.3.3. Considerações sobre os resultados gerados a partir dos modelos LSI e LDA

Os resultados textuais obtidos por meio dos modelos LSI e LDA em ambos os *corpora* de dados apresentam duas características distintas. A primeira refere-se aos pesos contidos em cada um dos termos de cada tópico e a segunda refere-se às próprias características dos termos, podendo ser fortes ou fracos mediante a sua especificidade ou generalidade junto ao domínio da linguagem.

É possível observar entre os resultados, utilizando como exemplo o tópico 0 do *corpus* 1, que os pesos dos termos do modelo LSI são mais representativos em relação aos pesos do modelo LDA por apresentar valores distintos em todos os termos. Isso facilita para que o especialista identifique os assuntos do tópico e sugira o nome de um rótulo representativo de acordo com o domínio de linguagem.

- Modelo LSI: 0.668*"informação" + 0.195*"pesquisa" + 0.191*"biblioteca" + 0.151*"conhecimento" + 0.100*"comunicação" + 0.099*"dados" + 0.097*"processo" + 0.097*"forma" + 0.096*"trabalho" + 0.093*"social";
- Modelo LDA: 0.002*"segurança" + 0.002*"segurança_informação" + 0.001*"políticas_instituição" + 0.001*"instituição_pesquisada" + 0.000*"mayara" + 0.000*"pesquisada" + 0.000*"políticas_instituição_pesquisada" + 0.000*"critério" + 0.000*"petruso" + 0.000*"mayara_petruso".

No modelo LSI é possível observar que o termo “informação” apresenta maior peso entre os demais termos, seguindo uma ordem de relevância de termos como “pesquisa”, “biblioteca” até o termo “social”, diferente do modelo LDA, que apresenta pesos com valores baixos e próximos uns dos outros entre os quatro primeiros termos e peso igual a zero para os demais termos do tópico. Com os pesos com valor igual a zero, os resultados acabam por não apresentar uma ordem de relevância, dificultando a identificação dos assuntos por parte do especialista. Entretanto, os termos extraídos com base no modelo LDA possuem mais N-gramas do tipo bigrama e trigrama em relação ao modelo LSI, possibilitando ao especialista uma maior gama de opções para realizar a análise de assuntos e, posteriormente, realizar a suposição do nome do tópico, indo ao encontro de Wallach (2006) e Wang, McCallum e Wei (2207) sobre o uso de bigramas ou N-gramas na modelagem de tópicos para resultados com maior significância.

Embora existam práticas para identificar a quantidade de tópicos ideal de um determinado *corpus* de dados, como por exemplo, a técnica de análise de estabilidade, que verifica a quantidade de tópicos estáveis que melhor representa a coleção, optou-se por utilizar a técnica de suposição por meio da tentativa de erro. Isso gerou nove conjuntos de resultados, sendo eles 10, 14,

18, 22, 26 ,30, 34, 38 e 42 tópicos de cada *corpus* de dados. Por meio dos resultados foi possível identificar a quantidade ideal de tópicos para cada coleção de documentos, uma vez que, dependendo do tamanho do *corpus* de dados, uma quantidade baixa de tópicos pode apresentar resultados com mais de um assunto e uma quantidade elevada de tópicos pode apresentar termos fracos que não estejam relacionados ao domínio da linguagem.

O modelo LSI utiliza da *Singular Value Decomposition* (SVD) - Decomposição de Valor Singular à matriz de documentos, enquanto o LDA possui uma robustez em termos matemáticos e estatísticos. Isso possibilita, por exemplo, inferir a partir de documentos, tópicos relevantes que sumarizam os textos contidos no *corpus*. Dessa forma, o modelo LDA acaba por apresentar melhores resultados que o modelo LSI. Um exemplo está em um maior quantitativo de termos do tipo bigramas que podem ser considerados especialistas ou chave, contribuindo assim para uma melhor interpretação dos dados por parte do especialista ao realizar a análise de assunto, mesmo que não tenha sido desenvolvido nenhum tipo de algoritmo de comparação de performance e resultados dos modelos.

Entre os resultados, considera que o modelo LSI utiliza uma técnica mais simples, porém com bons resultados (GRAESSER et al., 2000) enquanto o modelo LDA utiliza uma técnica mais moderna, baseada em métodos de estatística Bayesiana (BLEI; NG; JORDAN, 2003).

Com relação à visualização dinâmica dos tópicos, realizada por meio da biblioteca pyLDAvis – através dos resultados extraídos do modelo LDA do segundo *corpora* de dados constituído de artigos completos e resumos expandidos –, torna-se possível ao especialista do domínio de linguagem realizar interpretações dos tópicos de maneira mais assertiva, com recursos de interações e distância multidimensional entre os tópicos. Além disso, é possível analisar a frequência geral e a estimada dos termos de cada tópico, ajustar métricas de relevância e distribuição de tópico condicional dado prazo, e examinar um grande número de relacionamentos termo-tópico de maneira compacta.

Um tópico LDA utiliza uma distribuição com milhares de termos contidos em um vocabulário de um determinado *corpus*, entretanto, para analisar os resultados, são utilizados um quantitativo entre três e 30 termos mais prováveis

encontrados em cada tópico. O problema encontrado para a interpretação dos dados está nos termos comuns que aparecem no topo das listas de vários tópicos, dificultando a diferenciação dos significados dos tópicos (SIEVERT; SHIRLEY, 2014).

Dependendo do conteúdo de um *corpus* de dados, a configuração das métricas com menor valor pode apresentar um conjunto de termos mais especialistas enquanto os valores elevados podem apresentar uma lista de termos generalistas, dificultando a interpretação dos termos contidos no *corpus*. Para identificar a qualidade dos termos, faz-se necessário que o especialista realize um passeio semântico por meio das configurações de métricas, uma vez que os termos são alterados de acordo com a configuração utilizada.

Pode-se perceber uma maior interação na interpretação dos dados de um *corpus* de dados e menor esforço cognitivo ao realizar a análise de assunto por parte do especialista quando comparados os resultados textuais extraídos por meio dos modelos LSI e LDA aos resultados dinâmicos exibidos através de gráficos e construídos por meio da biblioteca pyLDAvis. Entretanto, faz-se necessário ressaltar que a biblioteca exige processamento e memória para sua execução. Com a configuração do equipamento utilizado para processamento da visualização dinâmica, foi possível gerar os resultados apenas com os *corpus* de dados inferiores a 18.000kb e uma quantidade abaixo de 18 tópicos, o que permitiu o processamento do *corpus* 7 ao 13º com 12 tópicos cada. A visualização dinâmica gerada a partir de 13 tópicos apresentou tópicos e termos fracos entre seus resultados, enquanto a visualização dinâmica até 12 tópicos abordou mais de um assunto por tópico. Isso exigiu do especialista a utilização de outros parâmetros de relevância para definição da suposição do nome para um determinado tópico.

Pode-se criar uma classificação quanto às características, tanto do termo quanto para o peso de cada assunto extraído por meio da modelagem de tópicos, sendo: i) generalista - pode apresentar diversos significados por meio da composição de termos; ii) especialista - apresenta significado único e claro referente ao termo; e iii) chave - ideal para explorar documentos externos como o *corpus* de dados para obtenção de mais informações para os tópicos; e a) forte - com pesos representativos que permitem identificar a ordem de relevância dos termos nos tópicos; e b) fraco - termos com pesos de baixo valor de um tópico

ou valores iguais a zero, apresentando dificuldade de identificar a ordem de relevância dos tópicos.

A combinação da identificação de termos especialistas contidos nos tópicos com a exploração de documentos externos como o *corpus* de dados permite ao especialista realizar a interpretação dos dados e criar a suposição do nome ou área do tópico de maneira mais assertiva, uma vez que os documentos contidos nos *corpus* de dados podem conter informações de relevância como título, resumos e palavras-chaves.

5.3.4. Tópicos relevantes dos *corpora*

Considerando que o modelo LDA apresentou melhores resultados em relação ao modelo LSI, foram selecionados os conjuntos de tópicos, termos e pesos dos dois *corpora* de dados. Utilizou-se como critério para essa seleção a qualidade dos termos e pesos extraídos de cada coleção de documentos. Um exemplo de qualidade está na seleção de tópicos fortes, com termos que representam uma única área do conhecimento ou que contemplam áreas correlacionadas, descartando tópicos que não possuem conjunto de termos coesos e de difícil interpretação da linguagem natural pelo especialista do domínio da linguagem. O Quadro 21 apresenta um conjunto de tópicos relevantes – selecionado do primeiro *corpora* de dados – que contém documentos do tipo teses e dissertações extraídos dos *corpus* entre 2012 e 2017.

Quadro 21 – Tópicos mais relevantes do *corpora* de dados teses e dissertações

Número	*Corpus* / Tópico	Termos e pesos
1	*Corpus* 1 / tópico 2	0.002*"docentes" + 0.002*"repositórios" + 0.001*"produção_científica" + 0.001*"repositórios_institucionais" + 0.001*"repositório" + 0.001*"qualis" + 0.001*"científica" + 0.001*"coletâneas" + 0.001*"chile" + 0.001*"teses";
2	*Corpus* 1 / tópico 4	0.001*"unirio" + 0.001*"astronomia" + 0.001*"museu_astronomia" + 0.001*"astronomia_afins" + 0.001*"museu_astronomia_afins" + 0.001*"edifício" + 0.001*"museologia" + 0.001*"afins" + 0.000*"mast (Museu de Astronomia e Ciências Afins)" + 0.000*"itu";
3	*Corpus* 1 / tópico 21	0.001*"mineração" + 0.001*"mineração_dados" + 0.001*"ontologias" + 0.001*"difusa" + 0.000*"fuzzy" + 0.000*"vaguidade" + 0.000*"swrlb" + 0.000*"difusos" + 0.000*"difusas" + 0.000*"ontologia";
4	*Corpus* 2 / tópico 7	0.004*"zoo" + 0.003*"park" + 0.001*"zoológico" + 0.001*"filmes" + 0.001*"aves" + 0.001*"museu" + 0.001*"animais" + 0.001*"aquarium" + 0.001*"doação" + 0.001*"jardim";

5	*Corpus* 2 / tópico 27	0.002*"inteligência" + 0.001*"metadados" + 0.001*"competitiva" + 0.001*"inteligência_competitiva" + 0.001*"rdf" + 0.001*"concordo" + 0.001*"language" + 0.001*"semântica" + 0.001*"modelagem" + 0.001*"registros";
6	*Corpus* 3 / tópico 0	0.001*"perfis" + 0.001*"fakes" + 0.001*"celebridades" + 0.001*"twitter" + 0.000*"tweet" + 0.000*"mussumalive" + 0.000*"perfis_fakes" + 0.000*"ato_linguagem" + 0.000*"tweet_dia" + 0.000*"nairbello";
7	*Corpus* 3 / tópico 17	0.001*"vagas" + 0.001*"reservadas" + 0.001*"vagas_reservadas" + 0.001*"candidatos" + 0.001*"cotas" + 0.000*"percentual" + 0.000*"requisitos" + 0.000*"inscritos" + 0.000*"matriculados" + 0.000*"renda";
8	*Corpus* 4 / tópico 0	0.003*"universidades" + 0.002*"ranking" + 0.001*"top" + 0.001*"usp" + 0.001*"classificadas" + 0.001*"rankings" + 0.001*"posição" + 0.001*"universidades_brasileiras" + 0.001*"brasileiras" + 0.001*"unicamp";
9	*Corpus* 4 / tópico 15	0.003*"jongo" + 0.001*"roda" + 0.001*"lapa" + 0.001*"moda" + 0.001*"campo_grande" + 0.001*"jongo_lapa" + 0.000*"serrinha" + 0.000*"tambor" + 0.000*"nscg" + 0.000*"jongueira";
10	*Corpus* 4 / tópico 19	0.002*"multimídia" + 0.002*"ontologias" + 0.002*"domínio" + 0.002*"mpeg7" + 0.002*"ontologia" + 0.002*"escopo" + 0.002*"domínio_escopo" + 0.002*"conteúdo" + 0.001*"metadados" + 0.001*"mpeg-7";
11	*Corpus* 5 / tópico 2	0.001*"centro_memória" + 0.000*"centros_memória" + 0.000*"cemef (Centro de Memória da Educação Física)" + 0.000*"memória_documentação" + 0.000*"centros_memória_documentação" + 0.000*"escola_educação" + 0.000*"linhales" + 0.000*"garimpando" + 0.000*"projeto_garimpando" + 0.000*"garimpando_memórias";
12	*Corpus* 5 / tópico 11	0.001*"acervo" + 0.001*"bibliófilos" + 0.001*"fase" + 0.001*"imagens" + 0.001*"livros" + 0.001*"movimento" + 0.001*"imagens_movimento" + 0.001*"bce (Biblioteca Central)" + 0.001*"cem" + 0.001*"cem_bibliófilos";
13	*Corpus* 5 / tópico 29	0.001*"maioridade" + 0.001*"penal" + 0.001*"texto" + 0.001*"maioridade_penal" + 0.001*"pec" + 0.001*"redução" + 0.001*"redução_maioridade" + 0.000*"folha" + 0.000*"dia_publicação_dia" + 0.000*"publicação_dia_semana";
14	*Corpus* 6 / tópico 0	0.001*"editores" + 0.001*"livro_digital" + 0.001*"editorial" + 0.001*"revistas" + 0.000*"conduta" + 0.000*"mayer" + 0.000*"editoras" + 0.000*"multimídia" + 0.000*"maria_anjos" + 0.000*"livros_digitais";
15	*Corpus* 6 / tópico 15	0.002*"taxonomia" + 0.001*"auditoria" + 0.001*"contábeis" + 0.001*"riscos" + 0.001*"procedimentos" + 0.001*"auditor" + 0.001*"distorções" + 0.001*"risco" + 0.001*"demonstrações" + 0.000*"inspeção".

Fonte: Elaborado pelo autor.

O Quadro 22 apresenta os tópicos relevantes do segundo *corpora* de dados, que contém os documentos do tipo artigos completos e resumos expandidos extraídos dos *corpus* entre 2012 e 2018.

Quadro 22 – Tópicos mais relevantes do *corpora* de dados artigos completos e resumos expandidos

Número	*Corpus* / Tópico	Termos e pesos
1	*Corpus* 7 / Tópico 3	0.002*"periódicos" + 0.002*"artigos" + 0.002*"produção" + 0.002*"científica" + 0.002*"informação" + 0.001*"trabalho" +

		0.001*"pesquisa" + 0.001*"citação" + 0.001*"indicadores" + 0.001*"estudos";
2	*Corpus* 7 / Tópico 8	0.002*"informação" + 0.002*"memória" + 0.001*"pesquisa" + 0.001*"documentos" + 0.001*"nacional" + 0.001*"museu" + 0.001*"preservação" + 0.001*"brasil" + 0.001*"oiticica" + 0.001*"periódicos";
3	*Corpus* 8 / Tópico 1	0.002*"livro" + 0.001*"eletrônico" + 0.001*"livro_eletrônico" + 0.001*"informação" + 0.001*"documentos" + 0.001*"programa" + 0.001*"ufes" + 0.001*"papel" + 0.001*"suporte" + 0.001*"e-book";
4	*Corpus* 8 / Tópico 5	0.001*"dados" + 0.001*"publicação" + 0.001*"publicações" + 0.001*"recursos" + 0.000*"publicações_ampliadas" + 0.000*"ampliadas" + 0.000*"modelo" + 0.000*"digitais" + 0.000*"publicação_ampliada" + 0.000*"ampliada";
5	*Corpus* 9 / Tópico 2	0.001*"indexação" + 0.001*"facetada" + 0.001*"analistas" + 0.001*"usabilidade" + 0.001*"classificação" + 0.000*"etiquetagem" + 0.000*"taxonomia" + 0.000*"taxonomia_facetada" + 0.000*"usuário" + 0.000*"faceted";
6	*Corpus* 9 / Tópico 3	0.001*"cinema" + 0.001*"canais" + 0.001*"televisão" + 0.001*"audiovisual" + 0.001*"lei" + 0.001*"brasil" + 0.001*"globo" + 0.001*"filmes" + 0.000*"brasileiro" + 0.000*"programação";
7	*Corpus* 10 / Tópico 14	0.001*"gestão" + 0.001*"conhecimento" + 0.001*"modelo" + 0.001*"documentos" + 0.001*"empresa" + 0.001*"organização" + 0.001*"gestão_conhecimento" + 0.001*"gad" + 0.001*"motivacional" + 0.001*"modelagem";
8	*Corpus* 10 / Tópico 15	0.002*"ontologia" + 0.001*"ontologias" + 0.001*"domínio" + 0.001*"hemonto" + 0.001*"sangue" + 0.001*"etapa" + 0.001*"ontoforinfoscience" + 0.001*"ontology" + 0.001*"digitais" + 0.001*"construção";
9	*Corpus* 10 / Tópico 24	0.003*"serviços" + 0.002*"marketing" + 0.002*"informativos" + 0.002*"serviços_informativos" + 0.001*"orientação" + 0.001*"gestão" + 0.001*"prestadora" + 0.001*"lde" + 0.001*"unidade" + 0.001*"prestadora_serviços";
10	*Corpus* 11 / Tópico 5	0.003*"dados" + 0.001*"reuso" + 0.001*"conteúdo" + 0.001*"data" + 0.001*"reuso_dados" + 0.001*"arquétipos" + 0.001*"sistemas" + 0.001*"openehr" + 0.001*"sentenças" + 0.000*"metadados;
11	*Corpus* 11 / Tópico 12	0.000*"documentos" + 0.000*"fotografia" + 0.000*"audiovisuais" + 0.000*"documentos_audiovisuais" + 0.000*"sonoros" + 0.000*"iconográficos" + 0.000*"ctdais" + 0.000*"iconográficos_sonoros" + 0.000*"audiovisuais_iconográficos" + 0.000*"documentos_audiovisuais_iconográficos";
12	*Corpus* 12 / Tópico 17	0.000*"feministas" + 0.000*"feminismo" + 0.000*"estudos_feministas" + 0.000*"periódico" + 0.000*"estudos" + 0.000*"mulheres" + 0.000*"feminista" + 0.000*"revista_estudos_feministas" + 0.000*"revista_estudos" + 0.000*"produtivos";
13	*Corpus* 12 / Tópico 19	0.004*"conhecimento" + 0.003*"informação" + 0.001*"gestão" + 0.001*"organização" + 0.001*"conceitos" + 0.001*"processo" + 0.001*"competências" + 0.001*"meio" + 0.001*"seleção" + 0.001*"trabalho";
14	*Corpus* 13 / Tópico 3	0.001*"periódicos" + 0.001*"zona" + 0.001*"bradford" + 0.001*"artigos" + 0.001*"bahia" + 0.001*"consumo" + 0.001*"documentos" + 0.001*"igc (Instituto de Geociências)" + 0.001*"docentes" + 0.000*"produção";
15	*Corpus* 13 / Tópico 18	0.002*"conhecimento" + 0.002*"museu" + 0.001*"biblioteca" + 0.001*"gestão" + 0.001*"inovação" + 0.001*"pesquisa" + 0.001*"ferroviário" + 0.001*"políticas" + 0.001*"uso" + 0.001*"ferramentas".

Fonte: Elaborado pelo autor.

Faz-se necessário ressaltar a existência de outros tópicos com características similares as dos critérios utilizados para a seleção dos tópicos relevantes. Entretanto, optou-se por utilizar como amostragem o quantitativo de 15 tópicos para cada *corpora* de dados para realização da validação dos resultados.

5.3.5. Validação dos resultados

Com a seleção dos conjuntos de tópicos, termos e pesos mais relevantes dos dois *corpora* de dados extraídos por meio do modelo LDA, foi construído um formulário contendo 30 questões. 15 perguntas delas continham tópicos dos resultados mais representativos do *corpora* de dados constituído por teses e dissertações e 15 perguntas contendo os resultados dos tópicos mais relevantes do *corpora* de dados de artigos completos e resumos. O questionário foi enviado para profissionais da Ciência da Informação que atuam em universidades ou no mercado de trabalho.

Para cada questão, foi apresentado um quantitativo de 11 rótulos para que os respondentes, especialistas no domínio da linguagem, selecionassem a opção que melhor representasse o conjunto de termos e pesos. Os rótulos foram extraídos dos nomes dos Grupos de Trabalhos (GTs) do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), sendo eles: “Estudos Históricos e Epistemológicos da Ciência da Informação”, “Gestão da Informação e do Conhecimento”, “Informação & Saúde”, “Informação e Memória”, “Informação e Tecnologia”, “Informação, Educação e Trabalho”, “Mediação, Circulação e Apropriação da Informação”, “Museu, Patrimônio e Informação”, “Organização e Representação do Conhecimento”, “Política e Economia da Informação”, “Produção e Comunicação da Informação em Ciência” e “Tecnologia & Inovação”. Além disso, foi adicionada a opção “Outros” para que fosse utilizado pelo respondente para definir rótulos que representassem o tópico, caso necessário.

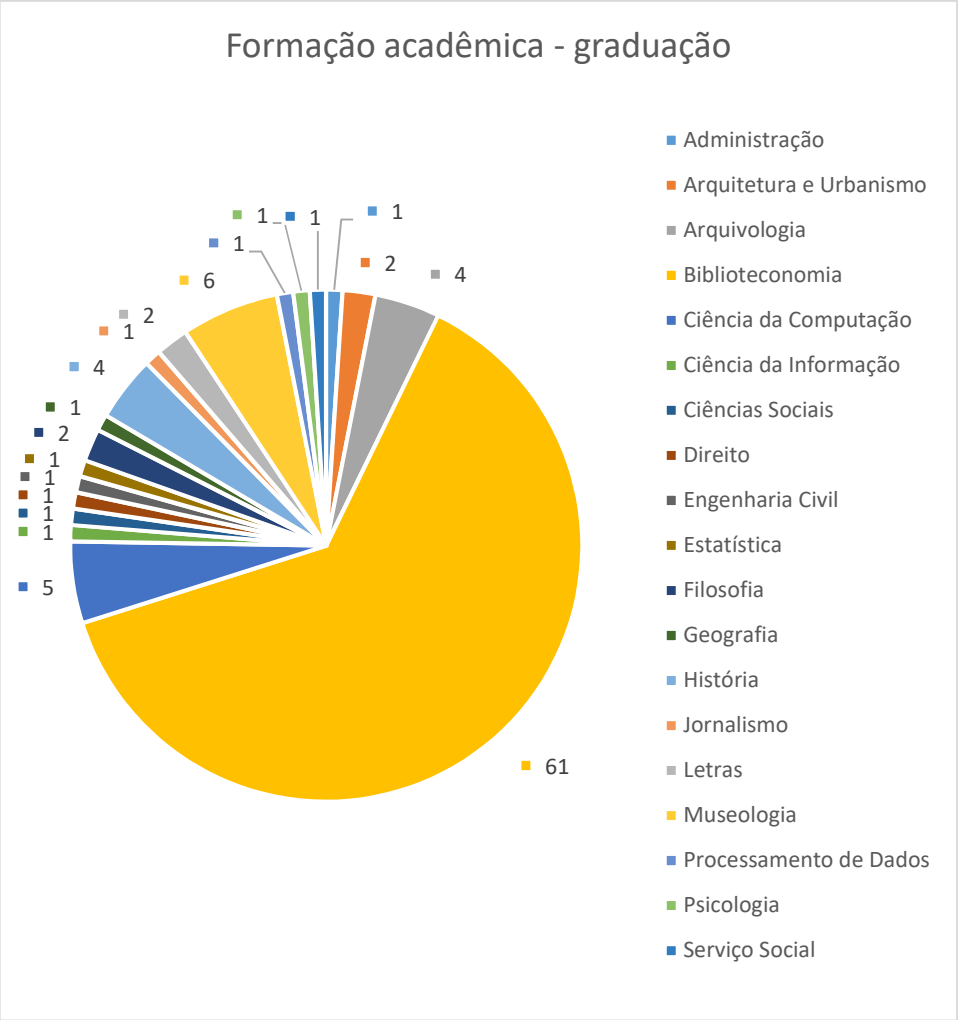
O questionário obteve o quantitativo de 88 respondentes, sendo 70 professores dos programas de pós-graduação da área de Ciência da Informação das universidades utilizadas na amostragem da pesquisa, 17 bibliotecários das universidades: Universidade Estadual de Londrina, Universidade Federal de Minas Gerais, Universidade Federal de Pernambuco, Universidade Federal do

Triângulo Mineiro, Universidade Estadual Paulista e Universidade Federal do Estado do Rio de Janeiro e 1 consultor que atua no mercado de trabalho. Dentre os professores, 7 deles exercem outras atividades, sendo 1 museólogos e gestor público, 1 cientista, 1 consultor palestrante, 1 pesquisador e 3 profissionais da tecnologia da informação. Também constam entre os respondentes 1 arquivista, 1 bibliotecário e 1 museólogo, que desenvolvem atividades na docência do ensino superior como segunda opção de atividade profissional.

Do quantitativo de respostas obtidas durante o período de coleta, foi excluído durante a fase de análise dos resultados 1 respondente com o perfil docente, que escreveu em todas as 30 perguntas a seguinte resposta: “Não entendi a pergunta”. Dessa forma, todos os dados foram tabulados e gerados a partir de 87 respondentes.

O Gráfico 21 apresenta a formação acadêmica - nível graduação dos respondentes que participaram da validação dos resultados alcançados por meio da modelagem de tópicos. Faz-se necessário ressaltar a existência de respondentes com mais de uma graduação, totalizando 95 formações acadêmicas.

Gráfico 21 – Formação acadêmica - graduação



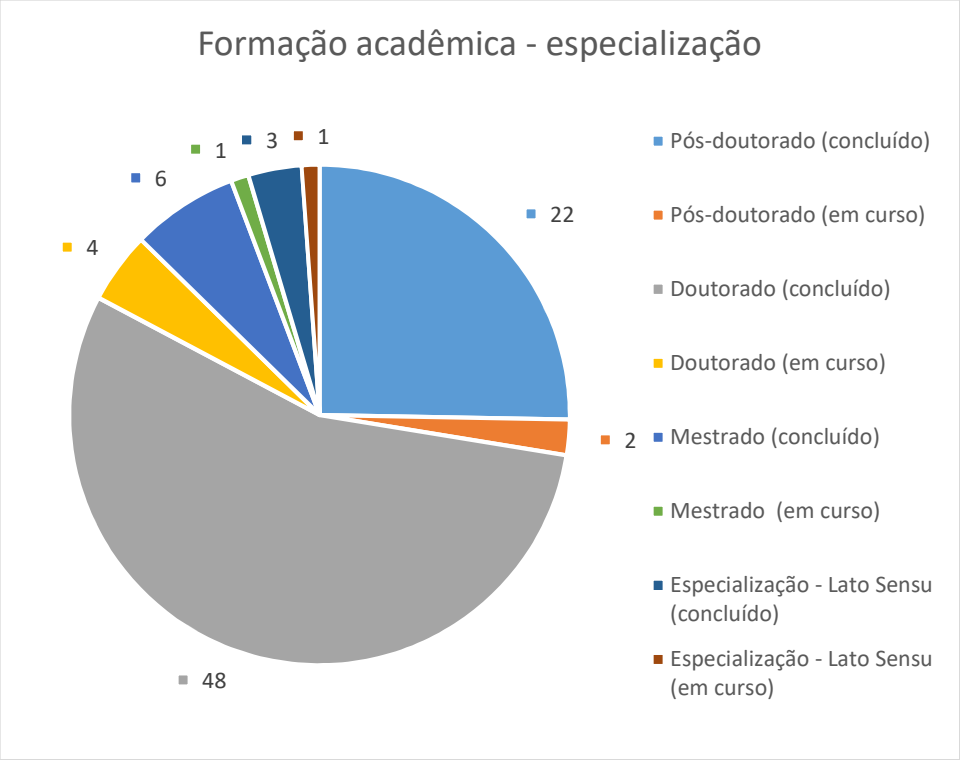
Fonte: Elaborado pelo autor.

Estão dentre as formações em nível de graduação: 1 ou 1,0% formado em Administração; 2 ou 2,1% em Arquitetura e Urbanismo; 4 ou 4,1% em Arquivologia; 61 ou 62,9% em Biblioteconomia; 5 ou 5,2% em Ciência da Computação; 1 ou 1,0% em Ciência da Informação; 1 ou 1,0% em Ciências Sociais; 1 ou 1,0% em Direito; 1 ou 1,0% em Engenharia Civil; 1 ou 1,0% em Estatística; 2 ou 2,1% em Filosofia; 1 ou 1,0% em Geografia; 4 ou 4,1% em História; 1 ou 1,0% em Jornalismo; 2 ou 2,1% em Letras; 6 ou 6,2% em Museologia; 1 ou 1,0% em Processamento de Dados; 1 ou 1,0% em Psicologia; e 1 ou 1,0% em Serviço Social.

Dos 61 respondentes com formação em Biblioteconomia, 17 ou 28% atuam profissionalmente como bibliotecários, 43 ou 70% como professores universitários e 1 ou 2% como consultor. Já todos os 26 respondentes com demais formações atuam como professores universitários em instituições de ensino superior utilizadas na amostragem.

O Gráfico 22 apresenta a formação acadêmica – nível especialização dos respondentes que participaram da validação dos resultados obtidos por meioda modelagem de tópicos. Sabe-se que pós-doutorado não é considerado título, entretanto, a opção foi disponibilizada aos respondentes como forma de identificar vivências e experiências enquanto discente para a sua formação profissional.

Gráfico 22 – Formação acadêmica - especialização



Fonte: Elaborado pelo autor.

Dentre os respondentes, 22 ou 25,3% possuem pós-doutorado concluído; 2 ou 2,3% com pós-doutorado em andamento; 48 ou 55,2% com título de doutorado concluído; 4 ou 4,6% com doutorado em curso; 6 ou 6,9% com título de mestrado concluído; 1 ou 1,1% com mestrado em andamento; 3 ou 3,4% com especialização *lato sensu* concluído; e 1 ou 1,1% com especialização *lato sensu* em andamento. Todos os respondentes com pós-doutorado concluído ou em curso atuam como professores universitários. Entre os doutores, 40 ou 85,1% atuam como professores universitários; 5 ou 10,6% atuam como bibliotecários; e 2 ou 4,3% possuem outras atuações no mercado. Já todos os doutorandos atuam como bibliotecários. Dentre os profissionais que possuem título de mestrado, 4 ou 66,7% são bibliotecários; e 2 ou 33.3% atuam como professores universitários. Já o mestrando atua como bibliotecário, bem como os respondentes com especialização *lato sensu* concluída ou em andamento.

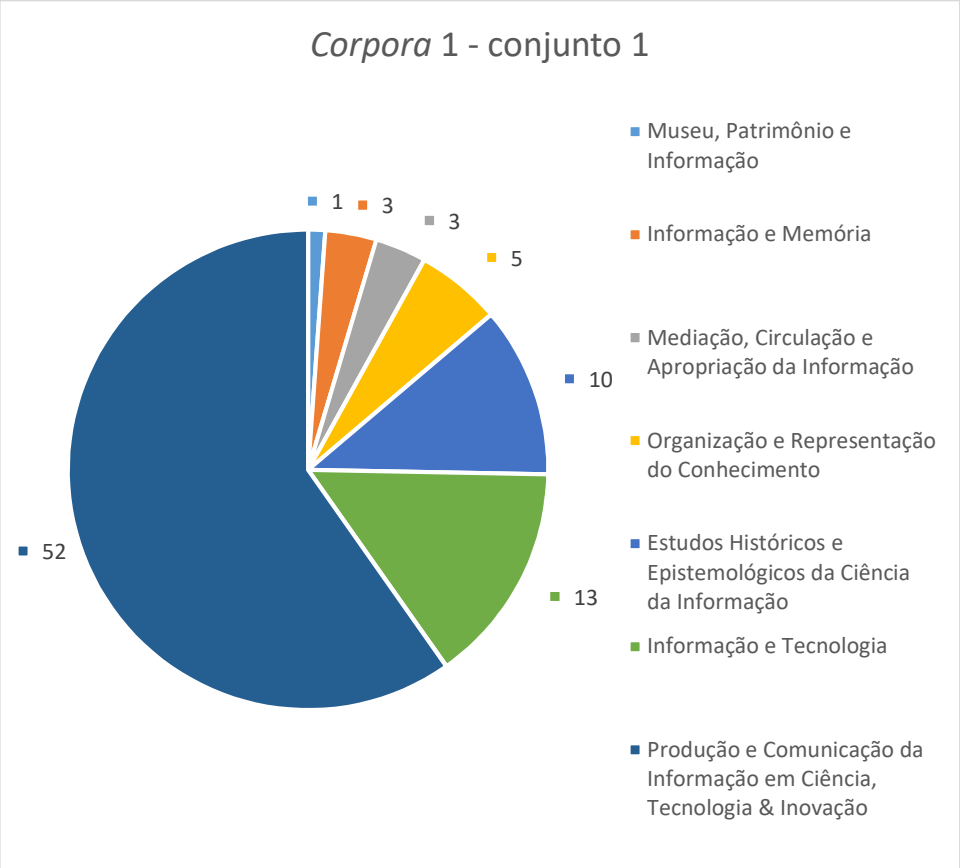
5.3.5.1. Validação dos resultados – *corpora* 1

A seguir são apresentados os resultados referentes à segunda seção do questionário, que diz respeito à validação dos termos extraídos por meio da

modelagem de tópicos realizada no *corpora* 1, constituído por documentos do tipo teses e dissertações.

O Gráfico 23 apresenta os resultados sobre qual rótulo melhor representa o primeiro conjunto de termos: [0.002*"docentes" + 0.002*"repositórios" + 0.001*"produção_científica" + 0.001*"repositórios_institucionais" + 0.001*"repositório" + 0.001*"qualis" + 0.001*"científica" + 0.001*"coletâneas" + 0.001*"chile" + 0.001*"teses"].

Gráfico 23 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 1



Fonte: Elaborado pelo autor.

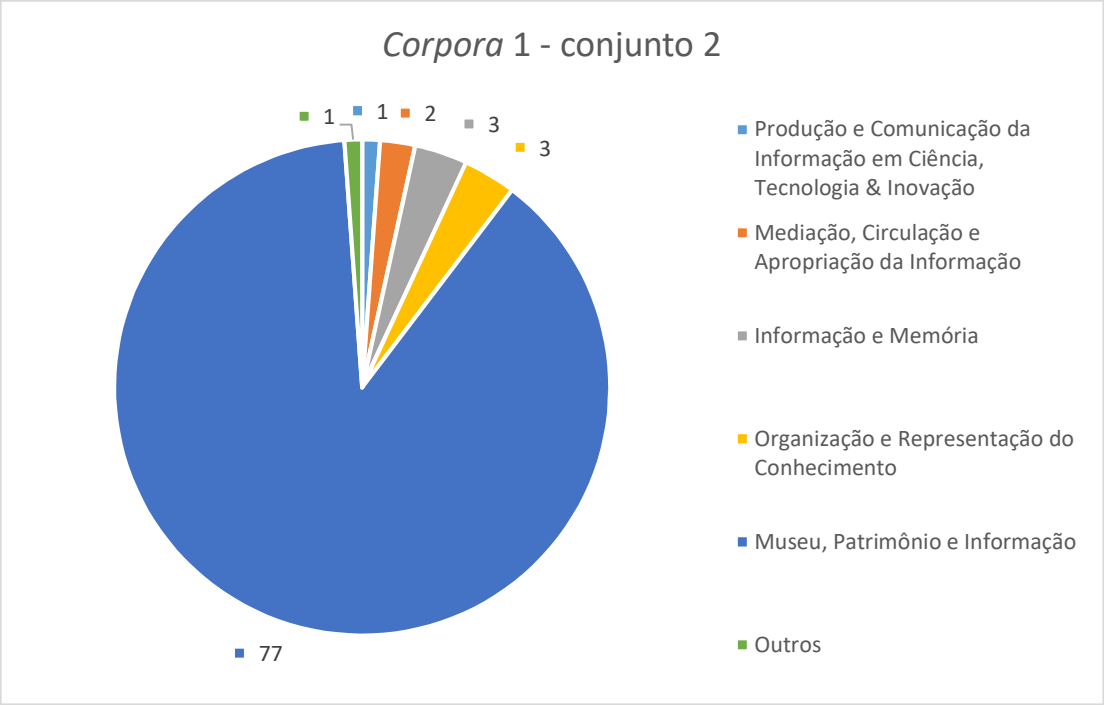
Dentre as respostas, 1 ou 1% escolheu o rótulo Museu, Patrimônio e Informação; 3 ou 3,4% escolheram o rótulo Informação e Memória; 3 ou 3,4% para Mediação, Circulação e Apropriação da Informação; 5 ou 5,7% para Organização e Representação do Conhecimento; 10 ou 11,5% para Estudos Históricos e Epistemológicos da Ciência da Informação; 13 ou 14,9% para Informação e Tecnologia; e 52 ou 59,8% Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação.

É possível observar que quase 60% dos respondentes optaram pela opção Produção e Comunicação da Informação em Ciência, Tecnologia &

Inovação que, na opinião dos participantes, melhor representa o conjunto de termos apresentados. Dos respondentes, 47 ou 54% consultaram o documento do tipo tese, intitulado “BIBLIOTECAS DIGITAIS: modelo metodológico para avaliação de usabilidade”, disponibilizado no *link* externo que faz parte da linha de pesquisa Organização e Uso da Informação e possui como palavras-chave os termos Bibliotecas digitais, Avaliação de Bibliotecas Digitais, Usabilidade de Bibliotecas Digitais e Metodologias para Avaliação de Bibliotecas Digitais.

O Gráfico 24 apresenta os resultados sobre qual rótulo melhor representa o segundo conjunto de termos do *corpora* teses e dissertações: [0.001*"unirio" + 0.001*"astronomia" + 0.001*"museu_astronomia" + 0.001*"astronomia_afins" + 0.001*"museu_astronomia_afins" + 0.001*"edifício" + 0.001*"museologia" + 0.001*"afins" + 0.000*"mast (Museu de Astronomia e Ciências Afins)" + 0.000*"itu"].

Gráfico 24 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 2



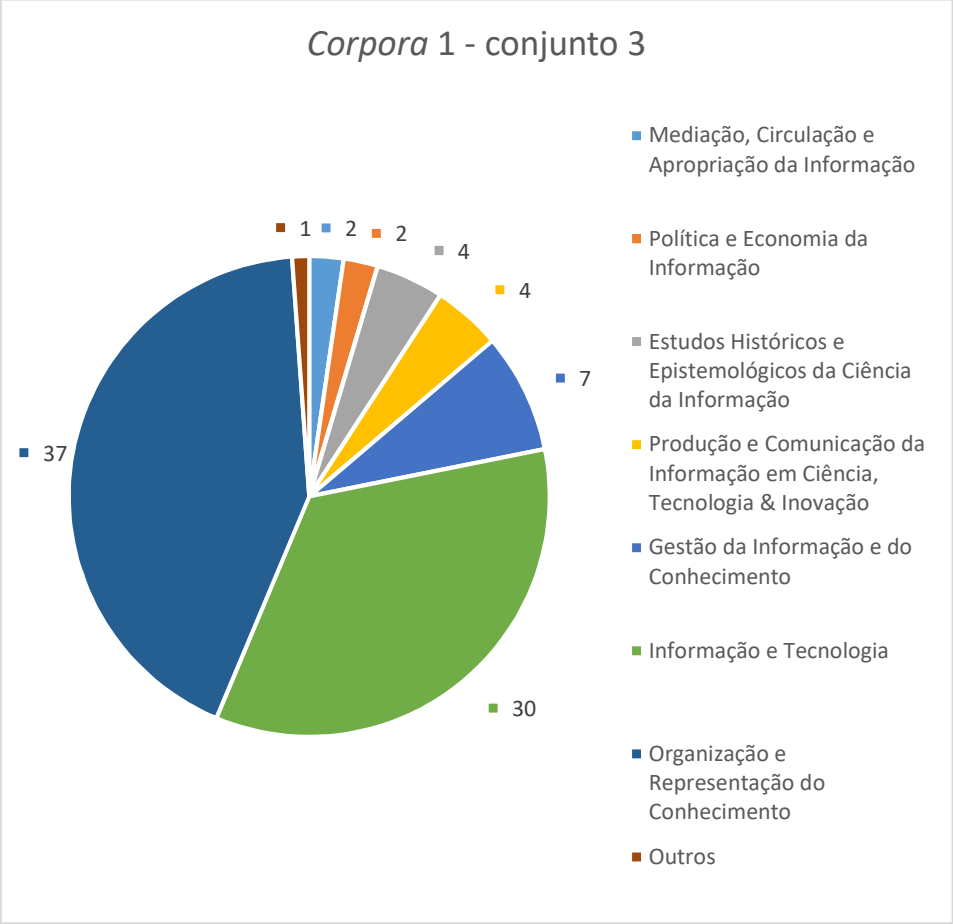
Fonte: Elaborado pelo autor.

Dentre as respostas, 1 ou 1,1% selecionou a opção Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 2 ou 2,3% para Mediação, Circulação e Apropriação da Informação; 3 ou 3,4% para Informação e Memória; 3 ou 3,4% para Organização e Representação do Conhecimento; 77 ou 88,5% para Museu, Patrimônio e Informação; e 1 ou 1,1% para opção Outros, sendo “não acho que seja CI - Arquitetura: uso de edifícios históricos”.

É possível observar uma coesão entre os respondentes com relação à seleção do rótulo com maior representatividade, sendo Museu, Patrimônio e Informação. Dos respondentes, 14 ou 16,1% realizaram a consulta ao documento que melhor representa o conjunto de termos disponibilizados no *link* externo junto ao questionário. O documento do tipo dissertação, intitulado “VIVÊNCIAS NO MUSEU: a arquitetura e os caminhos da museografia no Museu de Astronomia e Ciências Afins”, apresenta as palavras-chave Arquitetura, Museologia, Museu, Museografia e Patrimônio. Pode-se considerar um conjunto de termos fortes extraídos por meio da modelagem de tópicos, uma vez que um percentual elevado de respondentes, que selecionou uma única opção de resposta, e um baixo percentual de respondentes, que realizou a consulta ao documento através do *link* externo.

O Gráfico 25 apresenta os resultados sobre qual rótulo melhor representa o terceiro conjunto de termos do *corpora* teses e dissertações: [0.001*"mineração" + 0.001*"mineração_dados" + 0.001*"ontologias" + 0.001*"difusa" + 0.000*"fuzzy" + 0.000*"vaguidade" + 0.000*"swrlb" + 0.000*"difusos" + 0.000*"difusas" + 0.000*"ontologia"].

Gráfico 25 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 3



Fonte: Elaborado pelo autor.

Dentre as respostas, constam 2 ou 2,3% para Mediação, Circulação e Apropriação da Informação; 2 ou 2,3% para Política e Economia da Informação; 4 ou 4,6% para Estudos Históricos e Epistemológicos da Ciência da Informação; 4 ou 4,6% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 7 ou 8,0% para Gestão da Informação e do Conhecimento; 30 ou 34,5% para Informação e Tecnologia; 37 ou 42,5% para Organização e Representação do Conhecimento; e 1 ou 1,1% marcou a opção Outros, sendo "Recuperação da Informação e Representação do Conhecimento".

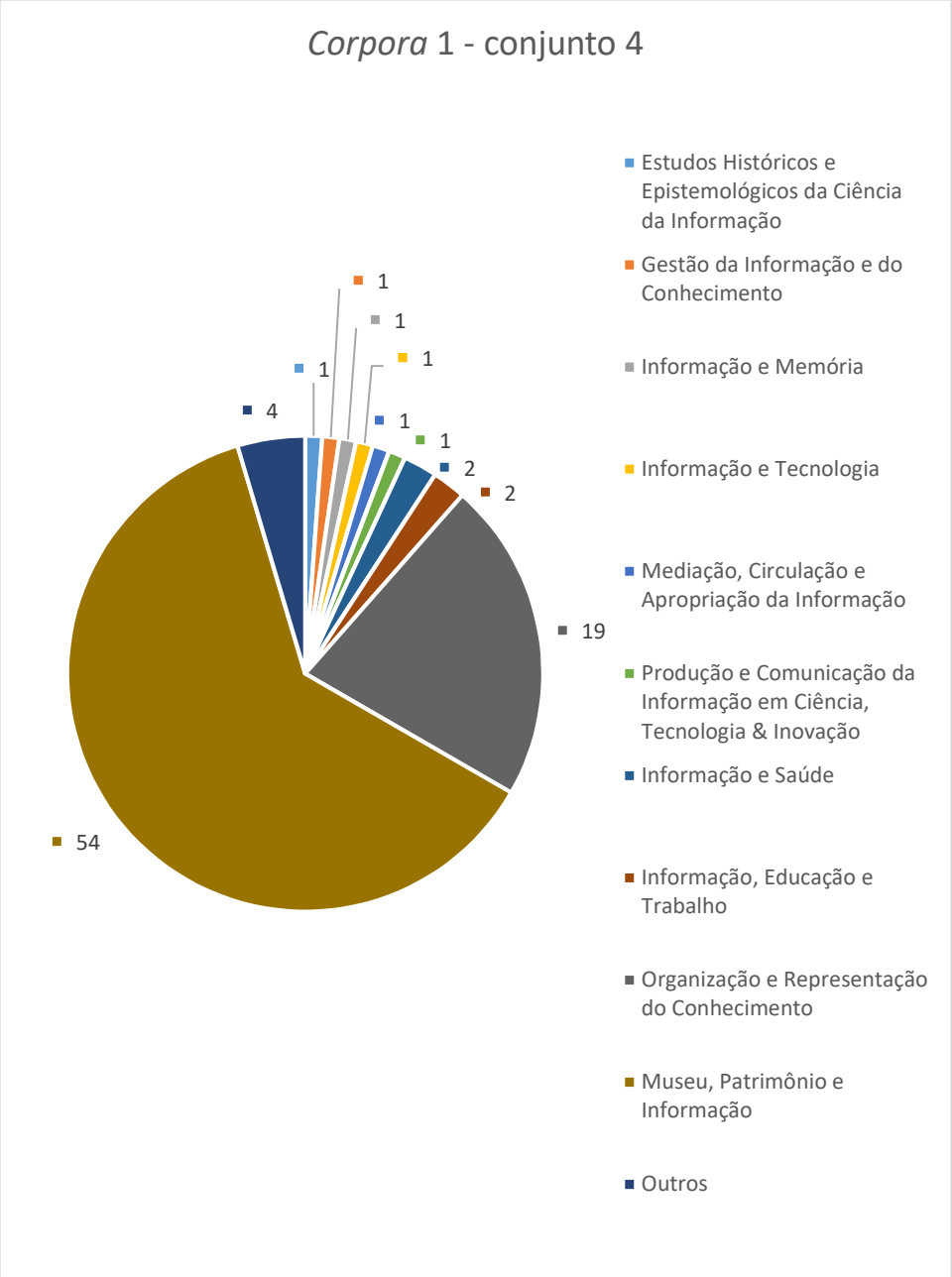
Os rótulos Informação e Tecnologia e Organização e Representação do Conhecimento apresentam os percentuais mais elevados dentre as classificações realizadas pelos respondentes, resultando numa diferença entre os rótulos de 7 ou 8%. Dos respondentes, 11 ou 12,6% realizaram a consulta ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo. O documento do tipo tese, intitulado "ONTOLOGIAS DIFUSAS NO SUPORTE À MINERAÇÃO DE DADOS: aplicações na Secretaria de

Finanças da Prefeitura Municipal de Belo Horizonte", faz parte da área de concentração Produção, Organização e Utilização da Informação, linha de pesquisa Gestão da Informação e do Conhecimento e possui as palavras-chave Ciência da Informação, Representação do Conhecimento (Teoria da Informação), Ontologias (Recuperação da Informação) e Mineração de dados (Computação).

Além do conjunto de termos extraído por meio da modelagem de tópicos e das informações obtidas junto ao documento que melhor representa o conjunto de termos, mesmo obtendo um número baixo de acessos mediante o quantitativo de respondentes, pode-se supor, por exemplo, que o tópico esteja relacionado aos rótulos com maior representatividade selecionada pelos respondentes.

O Gráfico 26 apresenta os resultados sobre qual rótulo melhor representa o quarto conjunto de termos do *corpora* teses e dissertações: [0.004*"zoo" + 0.003*"park" + 0.001*"zoológico" + 0.001*"filmes" + 0.001*"aves" + 0.001*"museu" + 0.001*"animais" + 0.001*"aquarium" + 0.001*"doação" + 0.001*"jardim"].

Gráfico 26 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 4



Fonte: Elaborado pelo autor.

Dentre os resultados, 1 ou 1,1% selecionou a opção Estudos Históricos e Epistemológicos da Ciência da Informação; 1 ou 1,1% para Gestão da Informação e do Conhecimento; 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Informação e Tecnologia; 1 ou 1,1% para Mediação, Circulação e Apropriação da Informação; 1 ou 1,1% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 2 ou 2,3% selecionaram o rótulo Informação e Saúde; 2 ou 2,3% para Informação, Educação e Trabalho; 19 ou 21,8% para Organização e Representação do Conhecimento; 54 ou 62,1% para

Museu, Patrimônio e Informação; e 4 ou 4,6% para opção Outros, sendo “Catalogação?”, “Não acho que seja CI - Museologia - novos ambientes”, “Zoologia” e “Zoológico”.

O rótulo Museu, Patrimônio e Informação foi a opção que apresentou um maior quantitativo de respondentes, alcançando mais de 60%. Outro rótulo a ser levado em consideração – ainda que em proporções menores e com uma diferença de 40,2% para o rótulo mais representativo – é o rótulo de Organização e Representação do Conhecimento. Dos respondentes, 15 ou 17,2% realizaram a consulta ao documento que melhor representa o conjunto de termos através do *link* externo disponibilizado junto ao formulário. O documento do tipo dissertação possui o título “JARDIM ZOOLÓGICO: Desafios para a aplicação do conceito de museu aos espaços de exposição de organismos vivos” e as palavras-chave Fundação Jardim Zoológico da Cidade do Rio de Janeiro, Museologia, Museus, Patrimônio Natural, Jardins Zoológicos e Coleções Vivas – Exposições.

Faz-se necessário destacar que somente uma das opções dos rótulos apresentada aos respondentes não foi selecionada, entretanto, o número elevado de rótulos selecionados entre as respostas não é indicativo de que o conjunto de termos extraídos por meio da modelagem de tópicos seja fraco. É possível notar entre os resultados que 14 respondentes selecionaram nove dos rótulos apresentados, enquanto o restante se dividiram em apenas 2 rótulos.

O Gráfico 27 apresenta os resultados sobre qual rótulo melhor representa o quinto conjunto de termos do *corpora* teses e dissertações: [0.002*"inteligência" + 0.001*"metadados" + 0.001*"competitiva" + 0.001*"inteligência_competitiva" + 0.001*"rdf" + 0.001*"concordo" + 0.001*"language" + 0.001*"semântica" + 0.001*"modelagem" + 0.001*"registros"].

Gráfico 27 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 5

Fonte: Elaborado pelo autor.

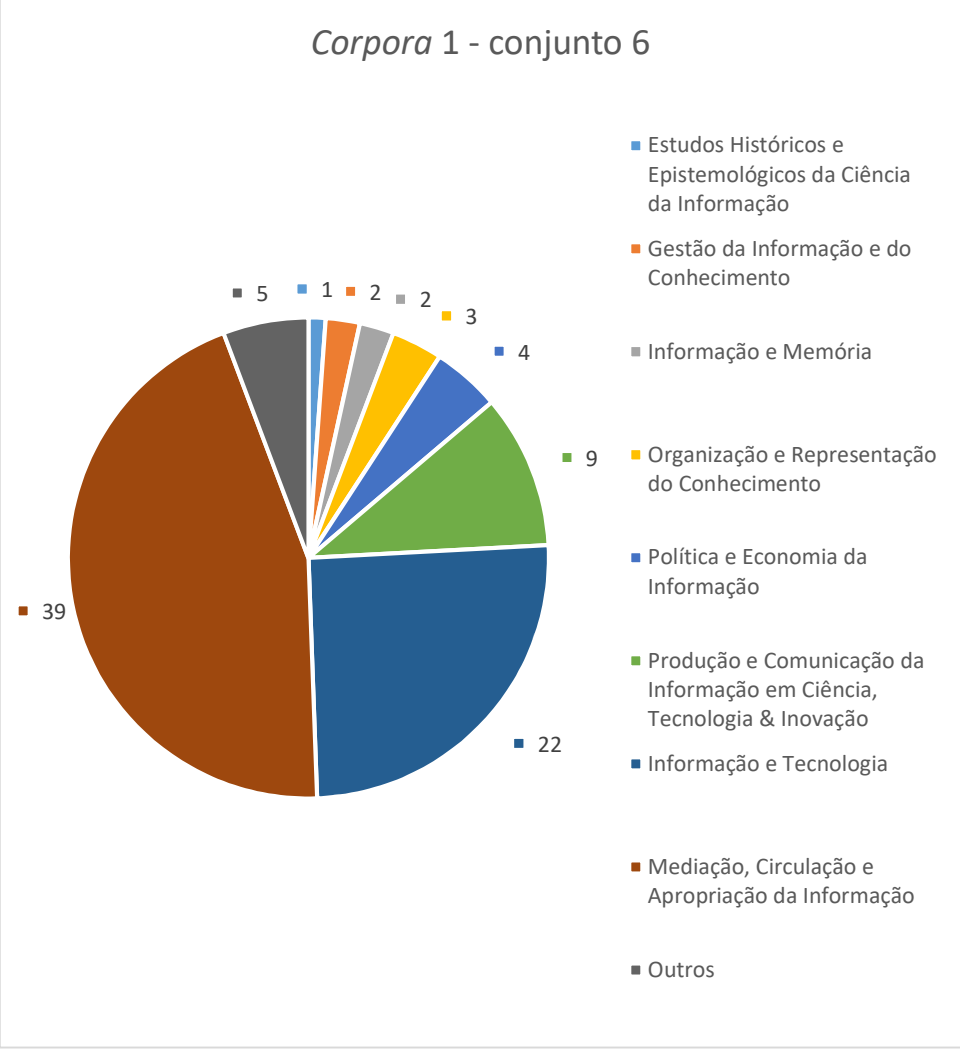
Estão entre os rótulos registrados 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Informação e Saúde; 1 ou 1,1% para Política e Economia da Informação; 3 ou 3,4% para Estudos Históricos e Epistemológicos da Ciência da Informação; 7 ou 8,0%, para Informação e Tecnologia; 29 ou 33,3% para Gestão da Informação e do Conhecimento; e 45 ou 51,7% para Organização e Representação do Conhecimento.

Dois dos rótulos apresentaram números representativos entre os respondentes, sendo um deles Organização e Representação do Conhecimento com mais de 50% dos participantes. Dentre todos os respondentes, 9 ou 10,3% realizaram a consulta ao documento que melhor representa o conjunto de termos disponibilizado junto ao formulário. O documento do tipo tese, intitulado "ORGANIZAÇÃO DO CONHECIMENTO EM BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES: uma abordagem baseada na classificação facetada e taxonomias dinâmicas", faz parte da linha de pesquisa Organização e Uso da Informação e possui as palavras-chave Ciência da Informação, Organização da

informação, Sistemas de Recuperação da Informação, Bibliotecas Digitais e Classificação Facetada, estando de acordo com as classificações realizadas pelos respondentes.

O Gráfico 28 apresenta os resultados sobre qual rótulo melhor representa o sexto conjunto de termos do *corpora* teses e dissertações: [0.001*"perfis" + 0.001*"fakes" + 0.001*"celebridades" + 0.001*"twitter" + 0.000*"tweet" + 0.000*"mussumalive" + 0.000*"perfis_fakes" + 0.000*"ato_linguagem" + 0.000*"tweet_dia" + 0.000*"nairbello"].

Gráfico 28 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 6



Fonte: Elaborado pelo autor.

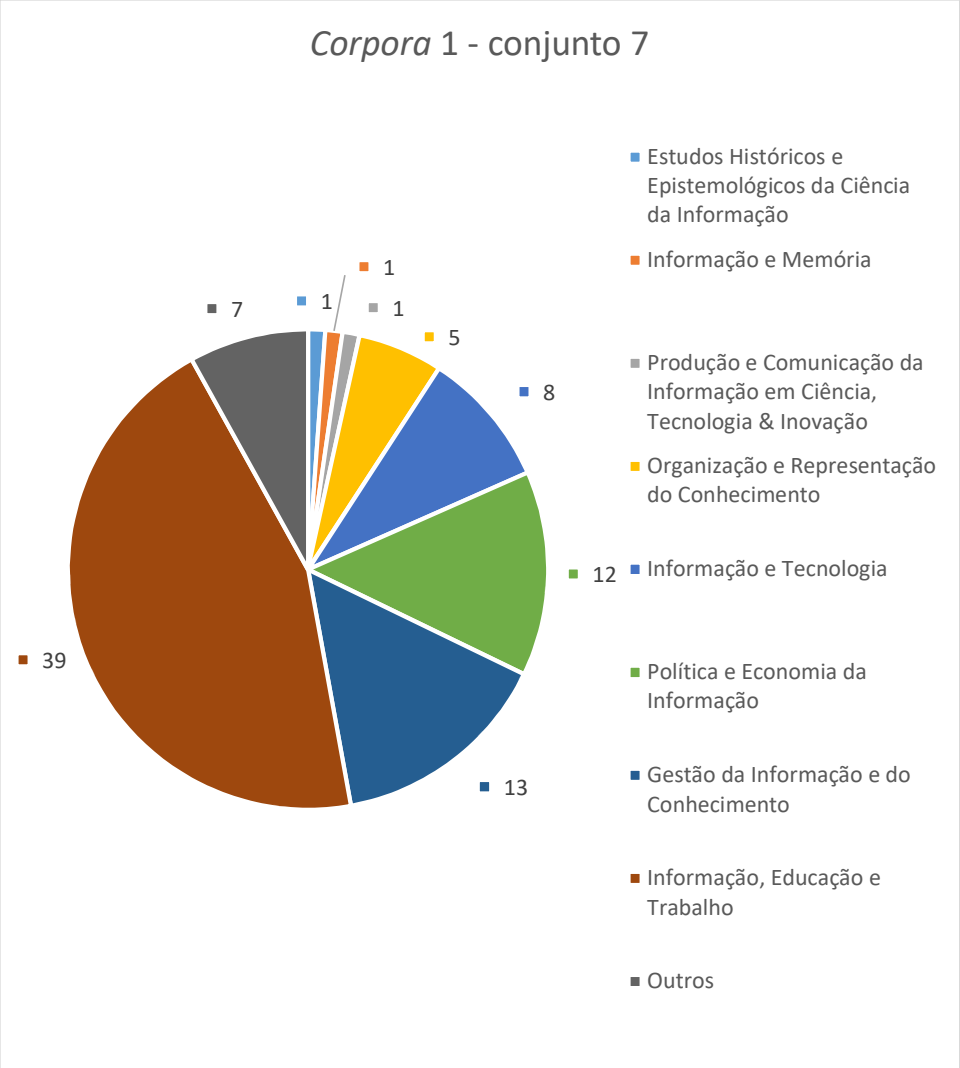
Dentre os rótulos registrados pelos respondentes estão 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 2 ou 2,3% para Gestão da Informação e do Conhecimento; 2 ou 2,3% para Informação e Memória; 3 ou 3,4% para Organização e Representação do Conhecimento; 4 ou

4,6% para Política e Economia da Informação; 9 ou 10,3% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 22 ou 25,3% para Informação e Tecnologia; 39 ou 44,8% para Mediação, Circulação e Apropriação da Informação; e 5 ou 5,7% para Outros, sendo “Cultura e Sociedade”, “Informação e Sociedade”, “Letramento Informacional”, “Não sei dizer” e “Redes Sociais”.

Mediação, Circulação e Apropriação da Informação foi o rótulo com maior número de escolhas realizadas pelos respondentes, seguido de Informação e Tecnologia. Embora os valores dos rótulos não sejam absolutos, é possível perceber – ao explorar o documento disponibilizado através do *link* externo junto ao questionário – que os resultados extraídos por meio da modelagem de tópicos se alinham ao rótulo mais votado, o que permite afirmar que se trata de um tópico forte. O documento que melhor representa o conjunto de termos é do tipo dissertação, intitulado “FAKES E CELEBRIDADES NO TWITTER: contratos de comunicação nos perfis @nairbello, @hebecamargo e @MussumAlive”. A pesquisa busca discutir as especificidades situacionais (finalidade, identidade, dispositivo) e discursivas (legitimidade, credibilidade e captação) nas quais se inscrevem os tweets produzidos pelos perfis e que foi consultado por 11 ou 12,6% dos respondentes.

O Gráfico 29 apresenta os resultados sobre qual rótulo melhor representa o sétimo conjunto de termos do *corpora* teses e dissertações: [0.001*"vagas" + 0.001*"reservadas" + 0.001*"vagas_reservadas" + 0.001*"candidatos" + 0.001*"cotas" + 0.000*"percentual" + 0.000*"requisitos" + 0.000*"inscritos" + 0.000*"matriculados" + 0.000*"renda"].

Gráfico 29 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 7



Fonte: Elaborado pelo autor.

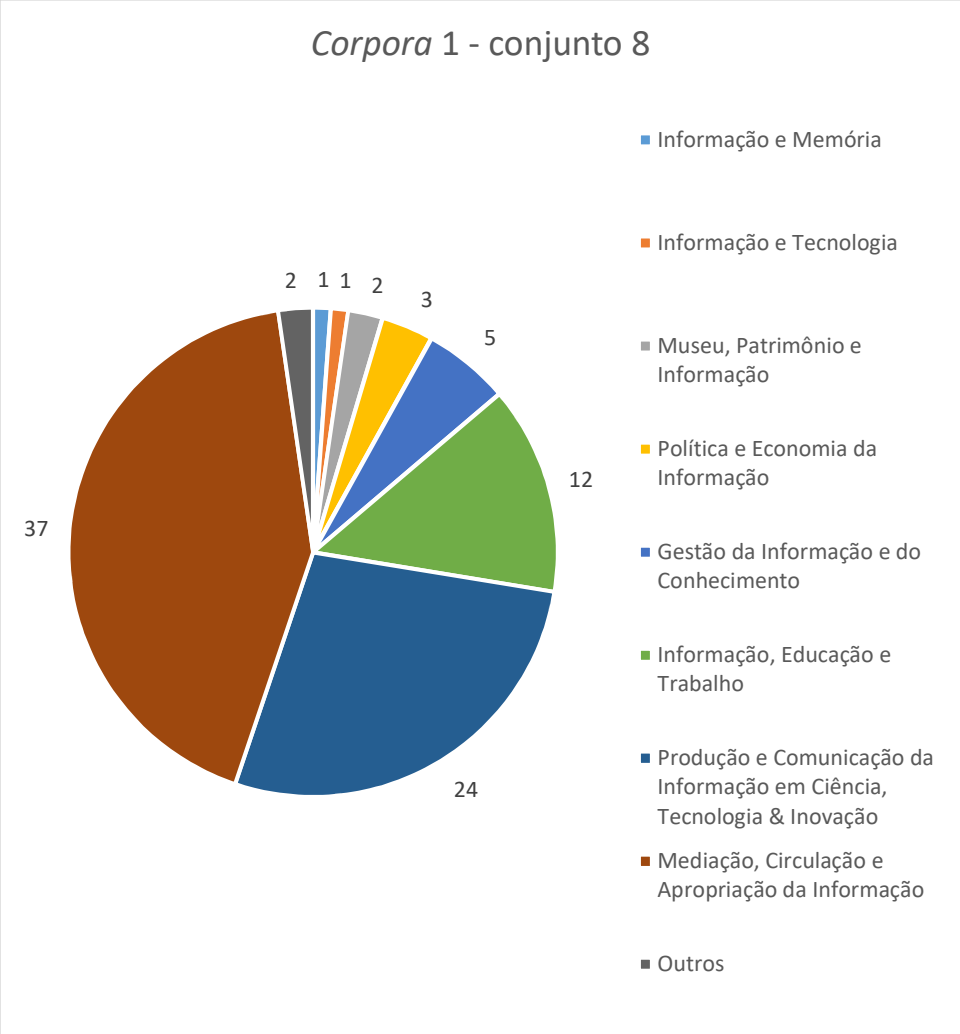
Dentre os rótulos selecionados pelos respondentes estão 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 5 ou 5,7% para Organização e Representação do Conhecimento; 8 ou 9,2% para Informação e Tecnologia; 12 ou 13,8% para Política e Economia da Informação; 13 ou 14,9% para Gestão da Informação e do Conhecimento; 39 ou 44,8% para Informação, Educação e Trabalho; e 7 ou 8,0% para a opção Outros, sendo “?”, “??”, “Informação e Sociedade”, “outra que não Ciência da Informação”, “Políticas Afirmativas de Educação”, “Políticas de Acesso ao Ensino Superior”, “Política Pública e Recuperação da Informação”. Os respondentes que não conseguiram identificar

um possível rótulo atuam como docentes do curso superior de biblioteconomia, entretanto, não possuem formação inicial na área.

O conjunto de termos pode ser representado por diversas áreas do conhecimento quando pensado na atividade fim, entretanto, torna-se possível perceber que o rótulo Informação, Educação e Trabalho, com maior número de suposições, possui maior proximidade para categorização. Dentre os respondentes, 11 ou 12,6% consultaram o *link* externo contido no formulário para o documento que melhor representa o conjunto de termos. O documento do tipo dissertação, intitulado “LEI DE COTAS: um estudo da reserva de vagas em uma instituição federal de ensino através da descoberta de conhecimento em bases de dados”, da linha de pesquisa Gestão da Informação e do Conhecimento, possui as palavras-chave Ciência da Informação, Gestão do Conhecimento, Universidades e Faculdades Públicas – Brasil, Políticas Públicas, Mineração de Dados e Processo Decisório. As opções citadas no campo “Outros”, que envolvem termos sobre Políticas e Recuperação da Informação, podem ser agregadas a outros rótulos já existentes.

O Gráfico 30 apresenta os resultados sobre qual rótulo melhor representa o oitavo conjunto de termos do *corpora* teses e dissertações: [0.003*"universidades" + 0.002*"ranking" + 0.001*"top" + 0.001*"usp" + 0.001*"classificadas" + 0.001*"rankings" + 0.001*"posição" + 0.001*"universidades_brasileiras" + 0.001*"brasileiras" + 0.001*"unicamp"].

Gráfico 30 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 8



Fonte: Elaborado pelo autor.

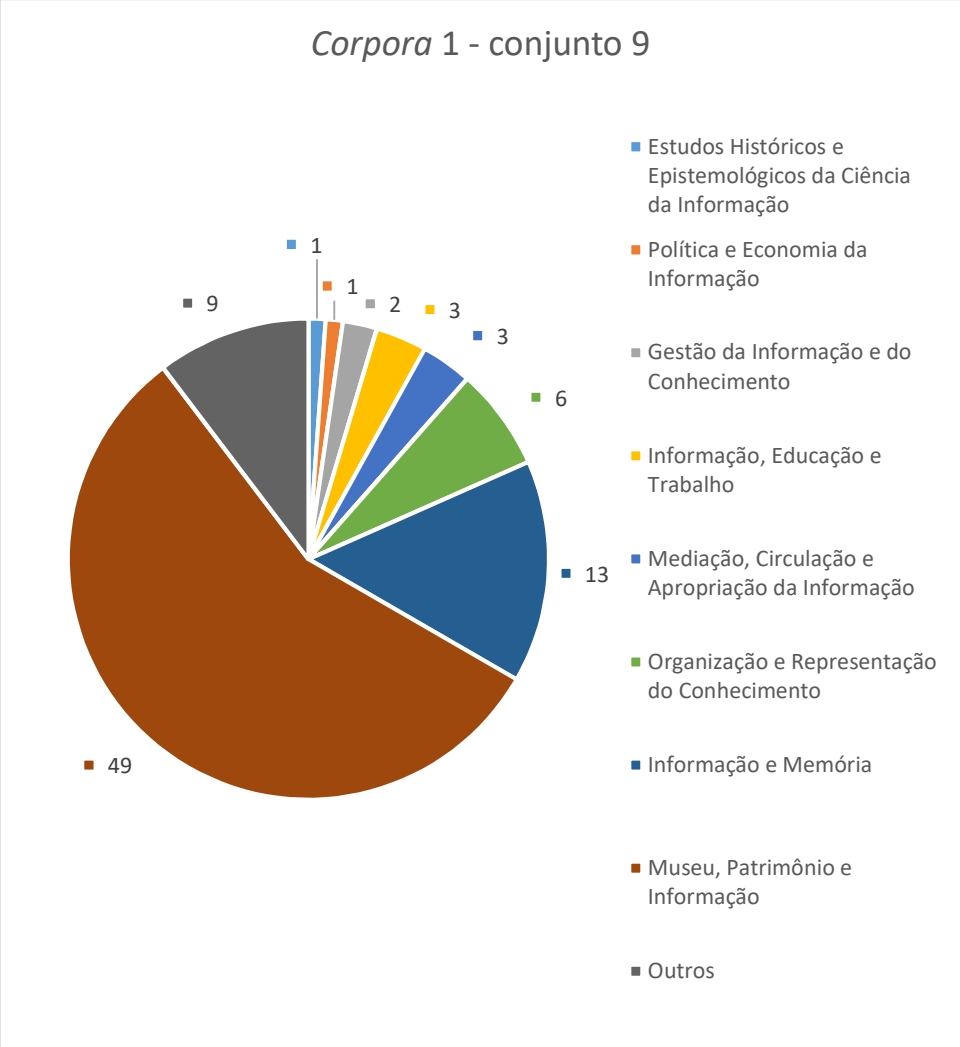
Constam entre números, percentuais e rótulos selecionados pelos respondentes 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Informação e Tecnologia; 2 ou 2,3% Museu, Patrimônio e Informação; 3 ou 3,4% Política e Economia da Informação; 5 ou 5,7% Gestão da Informação e do Conhecimento; 12 ou 13,8% Informação, Educação e Trabalho; 24 ou 27,6% Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 37 ou 42,5% Mediação, Circulação e Apropriação da Informação; e 2 ou 2,3% para a opção "Outros", sendo "Educação Superior" e "O arquivo da tese não abriu".

Embora dois dos rótulos tenham recebido os maiores números de suposições, sendo uma diferença de 13 votos ou 14,9%, é possível perceber, por meio das informações disponibilizadas no documento através do *link* externo junto ao conjunto de termos, que o rótulo Mediação, Circulação e Apropriação da Informação, com o maior número de votantes, possui maior aderência ao

conjunto de termos extraído por meio da modelagem de tópicos. O documento do tipo tese, intitulado “AVALIAÇÃO DA PÓS-GRADUAÇÃO BRASILEIRA: análise dos quesitos utilizados pela CAPES e das críticas da comunidade acadêmica”, pertence à área de concentração de Cultura e Informação e possui as palavras-chave Avaliação de Pós-Graduação do Brasil, Sistema Nacional de Pós-Graduação, CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e Critérios e Instrumentos de Avaliação de Cursos de Pós-Graduação.

O Gráfico 31 apresenta os resultados sobre qual rótulo melhor representa o nono conjunto de termos do *corpora* teses e dissertações: [0.003*"jongo" + 0.001*"roda" + 0.001*"lapa" + 0.001*"moda" + 0.001*"campo_grande" + 0.001*"jongo_lapa" + 0.000*"serrinha" + 0.000*"tambor" + 0.000*"nscg" + 0.000*"jongueira"].

Gráfico 31 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 9



Fonte: Elaborado pelo autor.

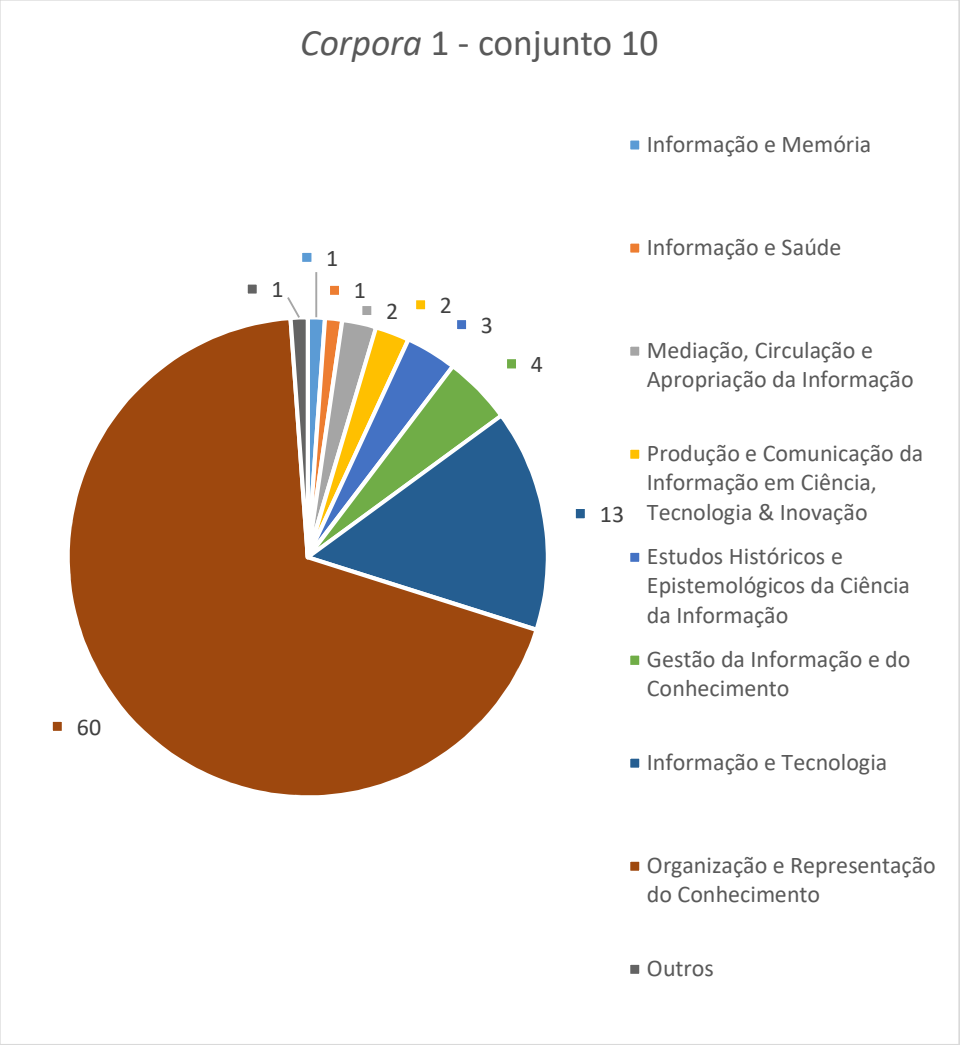
Dentre os rótulos selecionados pelos respondentes estão 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 1 ou 1,1% para Política e Economia da Informação; 2 ou 2,3% para Gestão da Informação e do Conhecimento; 3 ou 3,4% para Informação, Educação e Trabalho; 3 ou 3,4% para Mediação, Circulação e Apropriação da Informação; 6 ou 6,9% para Organização e Representação do Conhecimento; 13 ou 14,9% para Informação e Memória; 49 ou 56,3% para Museu, Patrimônio e Informação; e 9 ou 10,3% para a opção “Outros”, sendo “?”, “Cultura afro-brasileira”, “Cultura Popular”, “Informação e Cultura”, “Informação e Sociedade”, “não acho que seja CI - Museologia - representações emergentes”, “Não identifico”, “Outra que não Ciência da Informação” e “Patrimônio Cultural”.

O rótulo Museu, Patrimônio e Informação, representado por 56,3% dos respondentes, possui aderência aos termos extraídos por meio da modelagem de tópicos. O documento disponibilizado através do *link* junto ao conjunto de termos no formulário foi utilizado por 5 ou 5,7% dos respondentes. O documento do tipo dissertação é intitulado “O VALOR DO NEGRO: o processo de musealização do Museu do Ceará” e possui as palavras-chave Museu do Ceará, Museus, Museologia, Patrimônio cultural e Negros – Brasil.

Torna-se possível observar, junto a opção “Outros” deste conjunto de rótulos, que os respondentes sugeriram que o grupo de termos fosse da área de Museologia. Isso ocorreu porque os programas de Pós-Graduação de Memória Social e Museologia e Patrimônio da UNIRIO constituem os *corpora* de dados desta pesquisa, sendo informado aos respondentes por meio do questionário de validação da modelagem de tópicos que os documentos constituintes dos *corpora* de dados são dos programas de pós-graduação que fazem parte da ANCIB, não explicitando os nomes dos cursos utilizados na amostragem.

O Gráfico 32 apresenta os resultados sobre qual rótulo melhor representa o décimo conjunto de termos do *corpora* teses e dissertações: [0.002*"multimídia" + 0.002*"ontologias" + 0.002*"domínio" + 0.002*"mpeg7" + 0.002*"ontologia" + 0.002*"escopo" + 0.002*"domínio_escopo" + 0.002*"conteúdo" + 0.001*"metadados" + 0.001*"mpeg-7"].

Gráfico 32 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 10



Fonte: Elaborado pelo autor.

Dentre os rótulos selecionados pelos respondentes estão 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Informação e Saúde; 2 ou 2,3% para Mediação, Circulação e Apropriação da Informação; 2 ou 2,3% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 3 ou 3,4% para Estudos Históricos e Epistemológicos da Ciência da Informação; 4 ou 4,6% para Gestão da Informação e do Conhecimento; 13 ou 14,9% para Informação e Tecnologia; 60 ou 69,0% para Organização e Representação do Conhecimento; e 1, ou 1,1% para “Outros”, sendo “O arquivo não abriu”.

Dentre os resultados alcançados, o rótulo Organização e Representação do Conhecimento foi a opção selecionada por 69% dos respondentes. Dos respondentes, 7 ou 8,0% utilizaram da consulta ao documento externo que melhor representa o conjunto de termos. O número elevado de respondentes que escolheram uma única opção e o baixo número de acesso ao documento

auxiliar pode caracterizar o conjunto de termos como um tópico forte. O documento disponibilizado para consulta do tipo tese é intitulado “ONTOLOGIAS PARA REPRESENTAÇÃO DE DOCUMENTOS MULTIMÍDIA: análise e modelagem”, da linha de pesquisa em Organização e Uso da Informação. Tem como palavras-chave: Representação do Conhecimento, Anotação Semântica, Ontologias, Ontologias Multimídia, Padrões de Metadados e Web Semântica.

Alguns dos rótulos selecionados pelos respondentes podem ser descartados por não apresentarem aderência, tanto para os termos extraídos por meio da modelagem de tópicos quanto para o documento que melhor representa o conjunto de termos disponibilizados para consulta, como por exemplo, os rótulos Informação e Memória e Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação.

O Gráfico 33 apresenta os resultados sobre qual rótulo melhor representa o 11º conjunto de termos do *corpora* teses e dissertações: [0.001*"centro_memória" + 0.000*"centros_memória" + 0.000*"cemef (Centro de Memória da Educação Física)" + 0.000*"memória_documentação" + 0.000*"centros_memória_documentação" + 0.000*"escola_educação" + 0.000*"linhales" + 0.000*"garimpando" + 0.000*"projeto_garimpando" + 0.000*"garimpando_memórias"].

Gráfico 33 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 11

Fonte: Elaborado pelo autor.

Constam dentre os rótulos selecionados pelos respondentes 1 ou 1,1% para Informação e Saúde; 1 ou 1,1% para Informação, Educação e Trabalho; 1 ou 1,1% para Política e Economia da Informação; 3 ou 3,4% para Organização e Representação do Conhecimento; 7 ou 8,0% para Gestão da Informação e do Conhecimento; 8 ou 9,2% para Museu, Patrimônio e Informação; 64 ou 73,6% para Informação e Memória; e 2 ou 2,3% para a opção “Outros”, sendo “Arquivologia” e “Não acho que seja CI - Arquivologia – Memória”.

O rótulo Informação e Memória foi selecionado por mais de 70% dos participantes. Assim como nos exemplos anteriores, alguns dos rótulos também poderiam ser descartados dos resultados, pois, além de não possuírem aderência ao assunto, sequer podem causar dúvidas ao profissional especialista ao realizar a análise de assunto dos termos extraídos por meio da modelagem de tópicos, como por exemplo, os rótulos Informação e Saúde, Informação, Educação e Trabalho ou Política e Economia da Informação.

O documento disponibilizado para consulta através de *link* externo junto ao questionário foi utilizado por 8 ou 9,2% dos respondentes. O documento do tipo dissertação, intitulado “CENTROS DE MEMÓRIA E DOCUMENTAÇÃO DA UNIVERSIDADE FEDERAL DE MINAS GERAIS: perfis institucionais e políticas de acervo”, pertence à linha de pesquisa Informação, Cultura e Sociedade e possui as palavras-chave Ciência da Informação, Centros de Documentação – Organização, Centros de Memória e Arquivos.

Tanto o número elevado de respostas que marcaram o rótulo Informação e Memória quanto o baixo número de acesso ao documento externo que melhor representa o conjunto de termos permite caracterizar como um tópico forte.

O Gráfico 34 apresenta os resultados sobre qual rótulo melhor representa o 12º conjunto de termos do *corpora* teses e dissertações: [0.001*"acervo" + 0.001*"bibliófilos" + 0.001*"fase" + 0.001*"imagens" + 0.001*"livros" + 0.001*"movimento" + 0.001*"imagens_movimento" + 0.001*"bce (Biblioteca Central)" + 0.001*"cem" + 0.001*"cem_bibliófilos"].

Gráfico 34 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 12

Fonte: Elaborado pelo autor.

Dentre os rótulos selecionados pelos respondentes estão 2 ou 2,3% para Informação e Tecnologia; 2 ou 2,3% para Museu, Patrimônio e Informação; 9 ou 10,3% para Mediação, Circulação e Apropriação da Informação, 10 ou 11,5% para Gestão da Informação e do Conhecimento; 12 ou 13,8% para Estudos Históricos e Epistemológicos da Ciência da Informação; 13 ou 14,9% para Organização e Representação do Conhecimento; 36 ou 41,4% para Informação e Memória; e 3 ou 3,4% para a opção "Outros", sendo "Informação Cultural". "Não acho que seja CI - Biblioteconomia - livros raros" e "Materialidade da Informação, Coleções e Instituições".

O documento disponibilizado junto ao questionário que melhor representa o conjunto de tópicos foi consultado por 14 ou 16,1% dos respondentes. O documento do tipo dissertação é intitulado "O estudo da coleção de livros da Sociedade dos Cem Bibliófilos do Brasil da Biblioteca Central da Universidade de Brasília". Pertence à área de concentração de Gestão da Informação, linha

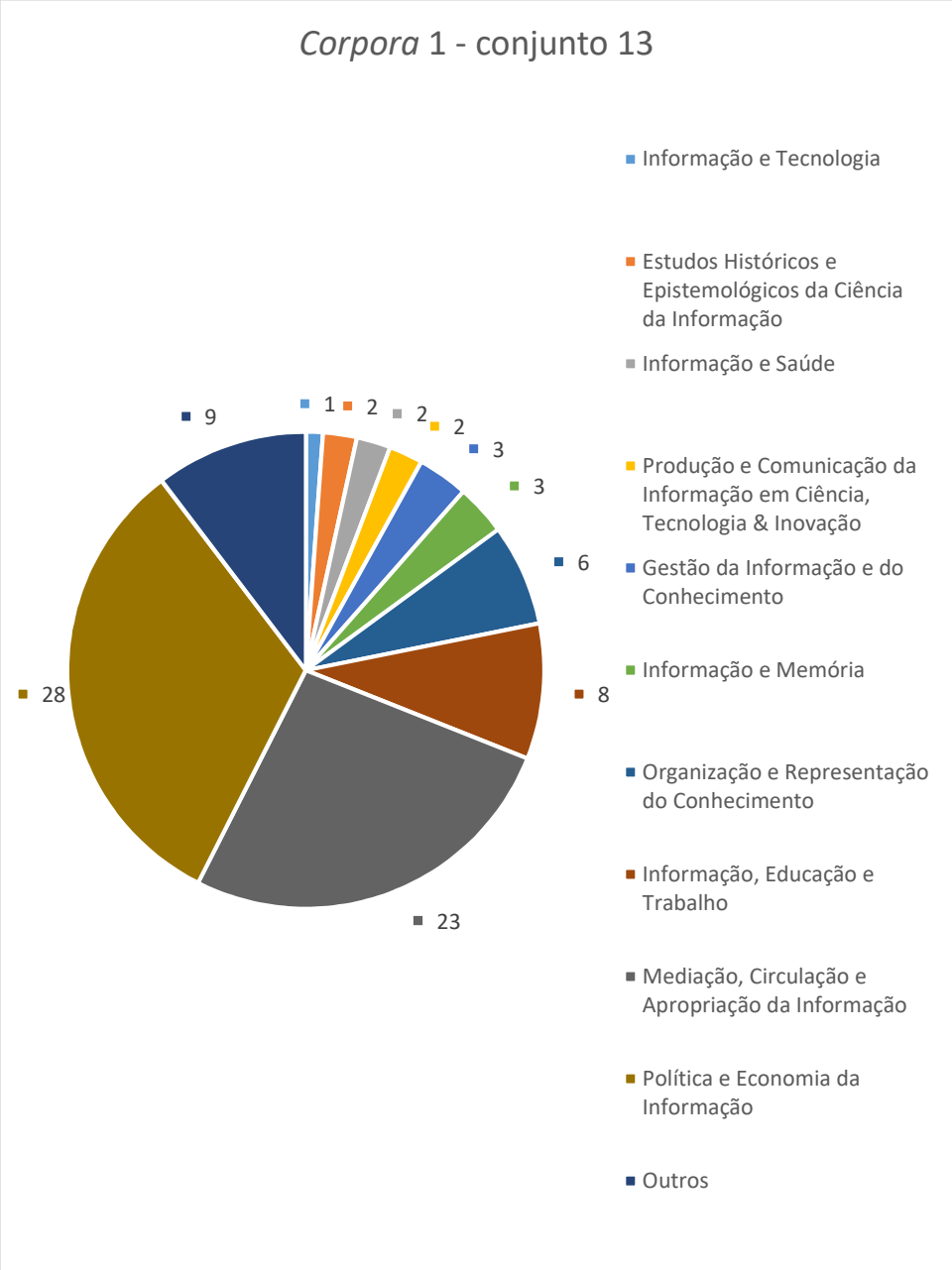
de pesquisa em Organização da Informação e palavras-chave Sociedade dos Cem Bibliófilos do Brasil, Livro de Arte, Livro de Bibliófilos e Obras Raras.

O rótulo Informação e Memória foi selecionado por 41,4% dos respondentes e, quando consultado o documento externo, é possível identificar uma coesão entre os resultados. Posteriormente, é possível identificar um equilíbrio formado por quatro rótulos que possuem números próximos, sendo Mediação, Circulação e Apropriação da Informação, Gestão da Informação e do Conhecimento, Estudos Históricos e Epistemológicos da Ciência da Informação e Organização e Representação do Conhecimento. Entretanto, ao analisar o documento externo é possível identificar que o conjunto de termos extraído por meio da modelagem de tópicos de fato não está relacionado a esses rótulos.

Embora os termos apresentados no resultado fossem associados pela maioria dos respondentes ao rótulo Informação e Memória – um grupo constituído por quatro termos, que somados alcançam 44 ou 50,6% dos respondentes –, acabam por apontar para caminhos incoerentes em relação às opções marcadas. Com isso, algumas suposições podem ser refletidas: i) os termos não estão claros ou não oferecem indicativos para os rótulos; ii) o formulário de validação está extenso ou exaustivo, não havendo tempo ou prática para consultar documentos externos; e iii) com base no conhecimento ou subjetividade o respondente ter convicção na escolha da suposição correta do rótulo. Entretanto, a primeira opção pode ser descartada, uma vez que a maioria dos respondentes marcou o rótulo que melhor representa o conjunto de termos.

O Gráfico 35 apresenta os resultados sobre qual rótulo melhor representa o 13º conjunto de termos do *corpora* teses e dissertações: [0.001*"maioridade" + 0.001*"penal" + 0.001*"texto" + 0.001*"maioridade_penal" + 0.001*"pec" + 0.001*"redução" + 0.001*"redução_maioridade" + 0.000*"folha" + 0.000*"dia_publicação_dia" + 0.000*"publicação_dia_semana"].

Gráfico 35 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 13



Fonte: Elaborado pelo autor.

Dentre a quantidade, percentual e rótulos selecionados pelos respondentes estão 1 ou 1,1% para Informação e Tecnologia; 2 ou 2,3% para Estudos Históricos e Epistemológicos da Ciência da Informação; 2 ou 2,3% para Informação e Saúde; 2 ou 2,3% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 3 ou 3,4% para Gestão da Informação e do Conhecimento; 3 ou 3,4% para Informação e Memória; 6 ou 6,9% para Organização e Representação do Conhecimento; 8 ou 9,2% para Informação; Educação e Trabalho; 23 ou 26,4% para Mediação, Circulação e Apropriação da Informação; 28 ou 32,2% para Política e Economia da Informação; e 9 ou 10,3%

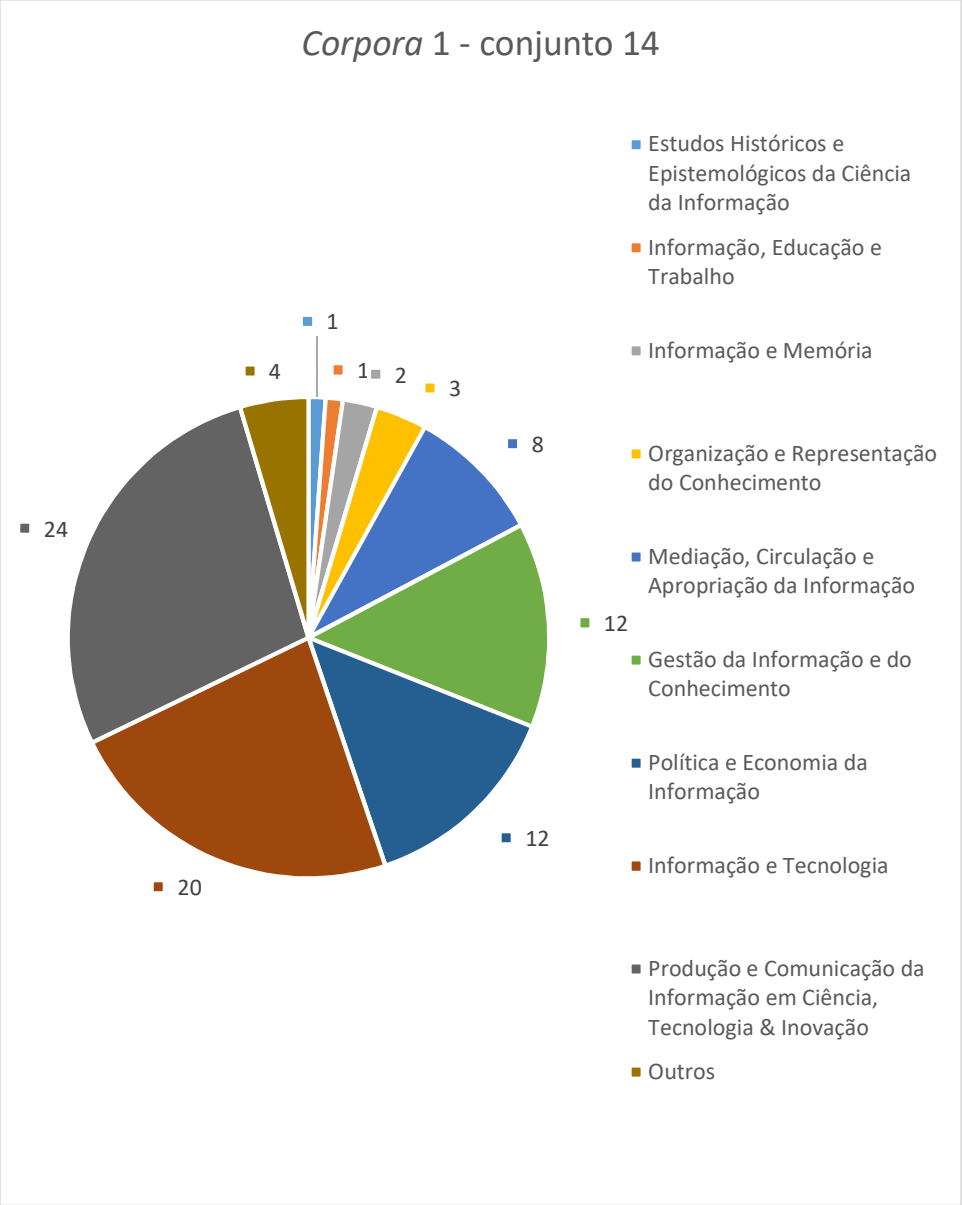
para a opção “Outros”, sendo “?”, “??”, “Comunicação, Sociedade e Política”, “Direito”, “Direito Penal”, “Informação e Sociedade”, “Jornalismo”, “Não identifico” e “Outra que não Ciência da Informação”. Constam entre os perfis dos respondentes que não identificaram a suposição do tema do tópico três docentes, sendo dois deles com formação em nível de graduação fora da área da Ciência da Informação e um com formação na área dura da pesquisa.

O documento que melhor representa o conjunto de termos disponibilizado através do formulário foi acessado por 15 ou 17,2% dos respondentes. O documento do tipo dissertação, o intitulado “A REDUÇÃO DA IDADE PENAL NO JORNALISMO DE REFERÊNCIA BRASILEIRO: uma análise dos sentidos sobre segurança pública”, da área de concentração de Informação, Ciência e Sociedade, possui as palavras-chave Jornalismo, Discurso, Segurança Pública, Redução da Maioridade Penal.

É possível perceber uma aproximação entre os dois rótulos mais votados, sendo Mediação, Circulação e Apropriação da Informação com 26,4% e Política e Economia da Informação com 32,2%. Embora o rótulo que tenha recebido o maior número de suposições pelos respondentes seja mais coerente com os termos, faz-se necessário ressaltar a necessidade da prática do especialista para realizar o acesso às informações contidas no documento que melhor representa o conjunto de termos, já que é possível identificar mais informações que possibilitam realizar uma inferência da suposição do rótulo de maneira mais assertiva. Justifica-se o volume de respondentes que optaram pela opção “Outros”, uma vez que os termos do tópico, bem como o documento externo utilizado para consulta, apontam para uma pesquisa do programa de pós-graduação em Comunicação e Informação. Faz-se necessário ressaltar que a amostragem da pesquisa foi coletada com base nos programas de pós-graduação da ANCIB, não sendo divulgado junto às instruções do formulário de validação dos dados os nomes dos cursos.

O Gráfico 36 apresenta os resultados sobre qual rótulo melhor representa o 14º conjunto de termos do *corpora* teses e dissertações: [0.001*"editores" + 0.001*"livro_digital" + 0.001*"editorial" + 0.001*"revistas" + 0.000*"conduta" + 0.000*"mayer" + 0.000*"editoras" + 0.000*"multimídia" + 0.000*"maria_anjos" + 0.000*"livros_digitais"].

Gráfico 36 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 14



Fonte: Elaborado pelo autor.

Constam dentre os resultados alcançados na identificação da suposição do tópico os números, percentuais e rótulos 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 1 ou 1,1% para Informação, Educação e Trabalho; 2 ou 2,3% para Informação e Memória; 3 ou 3,4% para Organização e Representação do Conhecimento; 8 ou 9,2% para Mediação, Circulação e Apropriação da Informação; 12 ou 13,8% para Gestão da Informação e do Conhecimento; 12 ou 13,8% para Política e Economia da Informação, 20 ou 23,0% para Informação e Tecnologia; 24 ou 27,6% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; e 4 ou 4,6% para a opção “Outros”, sendo “Biblioteconomia”, “Materialidade da

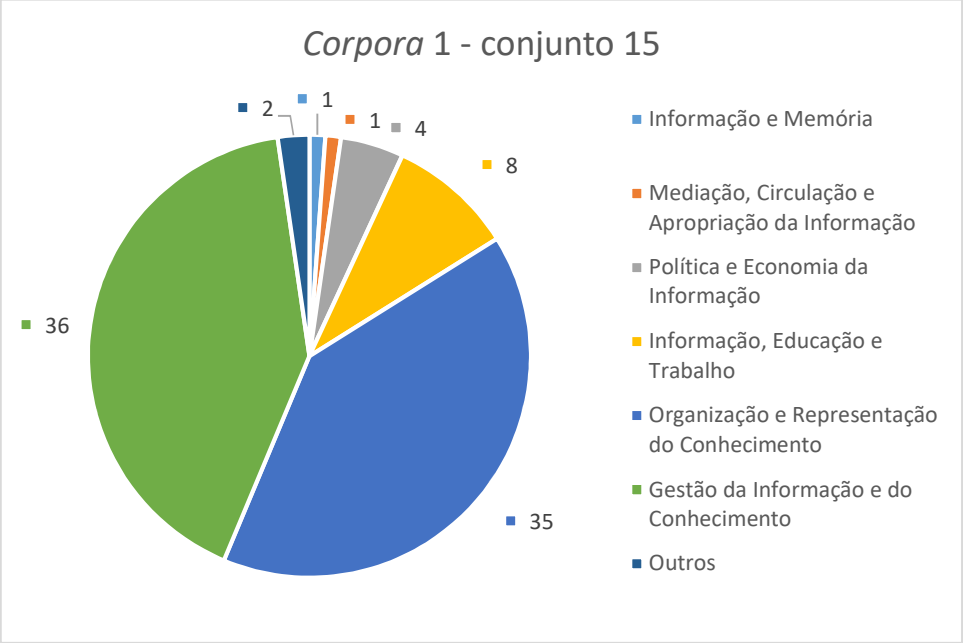
Informação, Coleções e Instituições”, “Não sei” e “Produção, Editoração e Comunicação Científica”.

O gráfico apresenta um equilíbrio entre dois grupos, sendo o primeiro formado por dois rótulos que alcançaram os maiores números e percentuais de suposições selecionadas pelos respondentes e o segundo constituído do terceiro ao quinto rótulo selecionado. Dos respondentes, 10 ou 11,5% acessaram o documento que melhor representa o conjunto de termos disponibilizado através de *link* externo junto ao questionário. O documento do tipo dissertação é intitulado “LIVRO DIGITAL: estudo de cenários do setor editorial nacional”. É parte da linha de pesquisa Ética, Gestão da Informação Estratégica e Políticas de Informação e possui as palavras-chave Gestão da Informação, Mercado Editorial e Cenários Prospectivos.

Mesmo fazendo uso da prática de acesso ao documento que melhor representa o conjunto de termos, torna-se uma tarefa de difícil execução, uma vez que o conjunto de termos pode fazer parte de mais de um rótulo. O gráfico também apresenta resultados que poderiam ser descartados, pois não possuem aderência ao conjunto de termos apresentados ou mesmo ao conteúdo do documento que melhor representa o tópico. Podem ser exemplos desse descarte os rótulos Estudos Históricos e Epistemológicos da Ciência da Informação, Informação, Educação e Trabalho, Informação e Memória e Organização e Representação do Conhecimento,

O Gráfico 37 apresenta os resultados sobre qual rótulo melhor representa 15º conjunto de termos do *corpora* teses e dissertações: [0.002*"taxonomia" + 0.001*"auditoria" + 0.001*"contábeis" + 0.001*"riscos" + 0.001*"procedimentos" + 0.001*"auditor" + 0.001*"distorções" + 0.001*"risco" + 0.001*"demonstrações" + 0.000*"inspeção"].

Gráfico 37 – Validação da modelagem de tópicos: *corpora* 1 – conjunto 15



Fonte: Elaborado pelo autor.

Dentre as suposições dos rótulos apresentados aos respondentes para o conjunto de termos, obteve-se os resultados 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Mediação, Circulação e Apropriação da Informação; 4 ou 4,6% para Política e Economia da Informação; 8 ou 9,2% para Informação, Educação e Trabalho; 35 ou 40,2% para Organização e Representação do Conhecimento; 36 ou 41,4% para Gestão da Informação e do Conhecimento; e 2 ou 2,3% para a opção “Outros”, sendo “Acesso ao conhecimento e informação através da tomada de decisão” e “Contabilidade”.

Dos respondentes, 5 ou 5,7% realizaram o acesso ao documento que melhor representa o conjunto de termos através do *link* externo disponibilizado no questionário. O documento do tipo tese, intitulada “Estudo do emprego da taxonomia como instrumento auxiliar para decisões táticas no processo de auditoria”, pertence à área de concentração Gestão da Informação, linha de pesquisa Organização da Informação e palavras-chave Modelo de Taxonomia, Procedimentos de Auditoria, Gestão de Risco e Redução de Subjetividade.

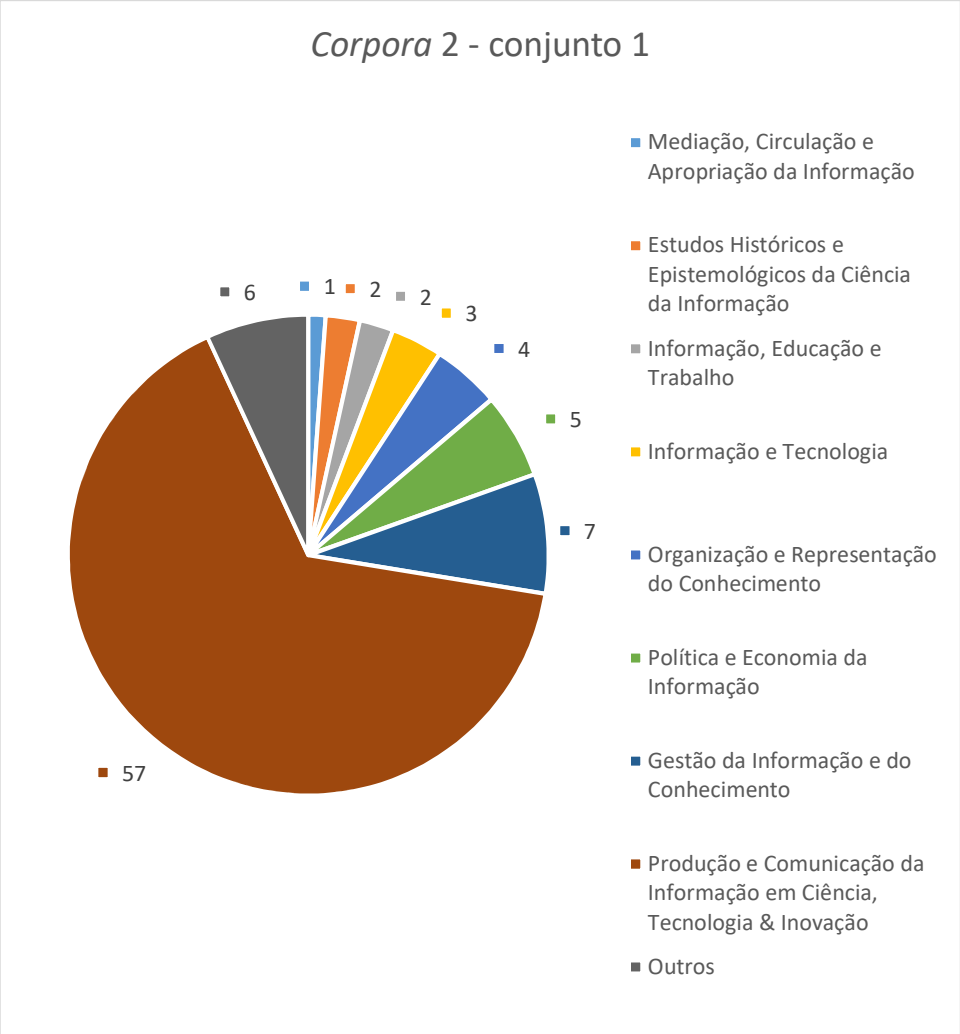
Considerando as respostas obtidas por meio da validação dos dados, bem como as características dos termos extraídos por meio da modelagem de tópicos, é possível obter dois rótulos, sendo Organização e Representação do Conhecimento como opção de 40,2% e Gestão da Informação e do Conhecimento como opção de 41,4% dos respondentes.

5.3.5.2. Validação dos resultados – *corpora* 2

A seguir são apresentados os resultados referentes à terceira seção do questionário. Diz respeito à validação dos termos extraídos por meio da modelagem de tópicos realizada no *corpora* 2, constituído por documentos do tipo artigos completos e resumos expandidos.

O Gráfico 38 apresenta os resultados sobre qual rótulo melhor representa o primeiro conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.002*"periódicos" + 0.002*"artigos" + 0.002*"produção" + 0.002*"científica" + 0.002*"informação" + 0.001*"trabalho" + 0.001*"pesquisa" + 0.001*"citação" + 0.001*"indicadores" + 0.001*"estudos"].

Gráfico 38 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 1



Fonte: Elaborado pelo autor.

Constam entre os rótulos selecionados pelos respondentes 1 ou 1,1% para Mediação, Circulação e Apropriação da Informação; 2 ou 2,3% para

Informação, Educação e Trabalho; 3 ou 3,4% para Informação e Tecnologia; 4 ou 4,6% para Organização e Representação do Conhecimento; 5 ou 5,7% para Política e Economia da Informação; 7 ou 8,0% para Gestão da Informação e do Conhecimento; 57 ou 65,5% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; e 6 ou 6,9% para a opção “Outros”, sendo “Comunicação Científica”, “Comunicação Cientifica, Produção Científica, Indicadores Bibliométricos”, “Estudos Métricos”, “Falha no Arquivo”, “Métricas e Indicadores Bibliométricos” e “Produção, Editoração e Comunicação Científica”.

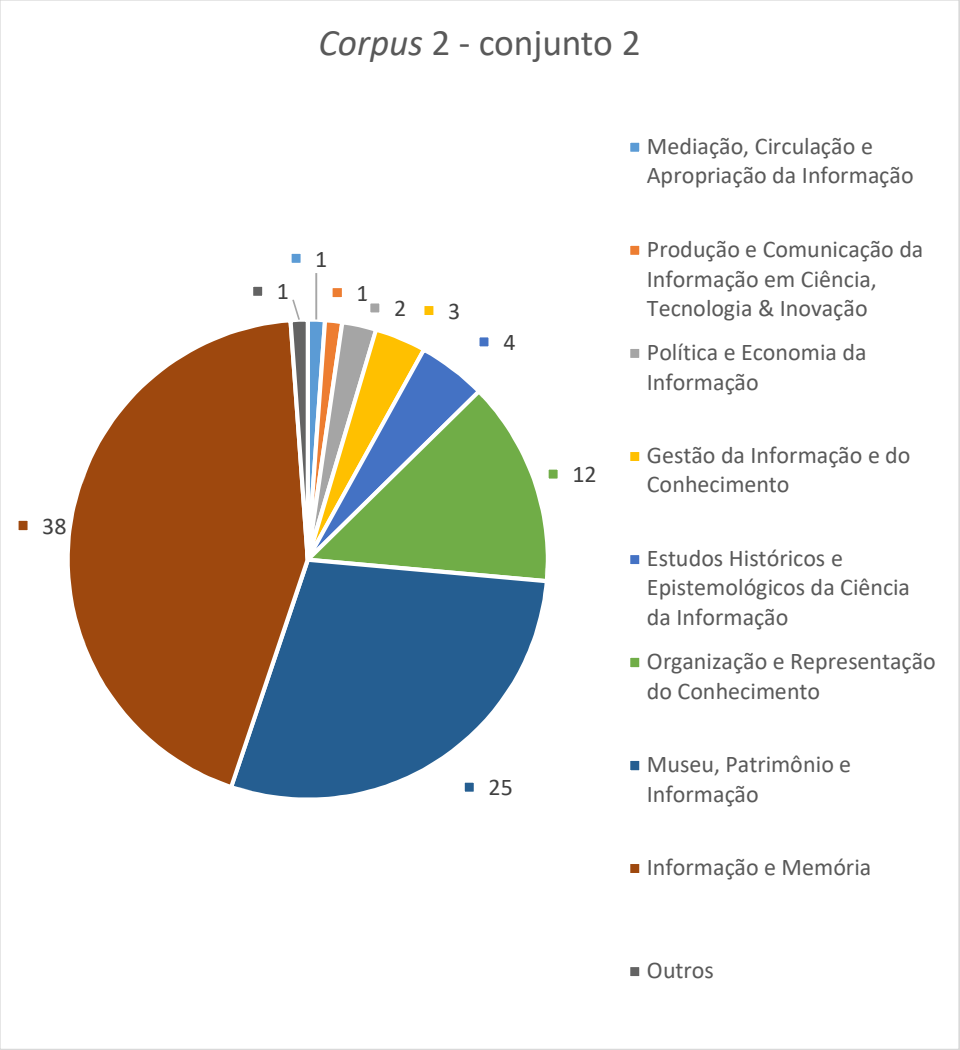
Dos respondentes, 5 ou 5,7% realizaram o acesso ao documento que melhor representa o conjunto de termos disponibilizados através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “ESTUDO DOS INDICADORES NACIONAIS A PARTIR DA BASE BRAPCI: uma análise comparativa com os indicadores das bases internacionais”, possui as palavras-chave Indicadores de Impacto, Indicadores de Citação, Base BRAPCI e Bases Internacionais.

O rótulo Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação foi escolhido por 65,5% dos respondentes, possibilitando inferir que se trata de um conjunto de termos fortes extraídos por meio da modelagem de tópicos, pois, além de um número representativo de respondentes que escolheram esta opção, existe um baixo número de consultas ao documento externo disponibilizado. Por fim, o artigo completo também está alocado ao GT-7 do ENANCIB, que possui o mesmo nome do rótulo.

Os demais 30 ou 34,5% das escolhas realizadas pelos respondentes estão distribuídas entre sete grupos. Dentre os respondentes que optaram pela opção “Outros”, 3 ou 50% supõem que a pesquisa pode estar relacionada a estudos Métricos e 3 ou 50% supõem que pode estar relacionada à Comunicação Científica.

O Gráfico 39 apresenta os resultados sobre qual rótulo melhor representa o segundo conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.002*"informação" + 0.002*"memória" + 0.001*"pesquisa" + 0.001*"documentos" + 0.001*"nacional" + 0.001*"museu" + 0.001*"preservação" + 0.001*"brasil" + 0.001*"oiticica" + 0.001*"periódicos"].

Gráfico 39 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 2



Fonte: Elaborado pelo autor.

Dentre as suposições dos rótulos apresentadas aos respondentes para o conjunto de termos, obtiveram-se os resultados 1 ou 1,1% para Mediação, Circulação e Apropriação da Informação; 1 ou 1,1% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 2 ou 2,3% para Política e Economia da Informação; 3 ou 3,4% Gestão da Informação e do Conhecimento; 4 ou 4,6% Estudos Históricos e Epistemológicos da Ciência da Informação; 12 ou 13,8% Organização e Representação do Conhecimento; 25 ou 28,7% Museu, Patrimônio e Informação; 38 ou 43,7% Informação e Memória; e 1 ou 1,1% para a opção “Outros”, sendo “Indexação e Recuperação da Informação”.

Dentre os respondentes, 1 ou 1,1% realizou o acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado

“FRASEOLOGIA OITICIANA DESVENDA O LABIRINTO: categorias documentais de Hélio Oiticica aplicadas à sua produção artística”, possui as palavras-chave Hélio Oiticica, Informação em Arte, Termos e Conceitos Artísticos, Fraseologia Oiticiana, Indexação e Recuperação da Informação Artística.

Os rótulos Museu, Patrimônio e Informação e Informação e Memória resultaram em 72,4% das suposições realizadas pelos participantes, sendo, respectivamente, 28,7% para o primeiro e 43,7% para o segundo rótulo. É possível destacar uma divergência entre os resultados apresentados e a alocação do artigo que melhor representa o conjunto de tópicos no GT do ENANCIB. O artigo que melhor representa o conjunto de termos está alocado no GT-9 do ENANCIB, enquanto, na análise da maioria dos respondentes, o conjunto de termos se enquadra melhor no GT-10. Faz-se necessário destacar que a análise foi realizada sobre o conjunto de termos, uma vez que o acesso ao documento externo foi realizado por apenas um dos respondentes.

O Gráfico 40 apresenta os resultados sobre qual rótulo melhor representa o terceiro conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.002*"livro" + 0.001*"eletrônico" + 0.001*"livro_eletrônico" + 0.001*"informação" + 0.001*"documentos" + 0.001*"programa" + 0.001*"ufes" + 0.001*"papel" + 0.001*"suporte" + 0.001*"e-book"].

Gráfico 40 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 3

Fonte: Elaborado pelo autor.

Dentre a quantidade, percentual e rótulos selecionados pelos respondentes estão 1 ou 1,1% para Informação e Memória; 4 ou 4,6% para Gestão da Informação e do Conhecimento; 4 ou 4,6% para Política e Economia da Informação; 6 ou 6,9% para Estudos Históricos e Epistemológicos da Ciência da Informação; 7 ou 8,0% para Organização e Representação do Conhecimento; 9 ou 10,3% para Mediação, Circulação e Apropriação da Informação; 12 ou 13,8% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 42 ou 48,3% para Informação e Tecnologia; e 2 ou 2,3% para a opção "Outros", sendo "Design editorial" e "Materialidade da informação, coleções e instituições".

Dentre os respondentes, 10 ou 11,5% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado "EM

BUSCA DE UMA DEFINIÇÃO PARA O LIVRO ELETRÔNICO: o conteúdo informacional e o suporte físico como elementos indissociáveis", possui as palavras-chave E-book, Livro eletrônico e Tecnologia da Informação.

O rótulo Informação e Tecnologia foi a opção de 48,3% dos respondentes, estando de acordo com o GT-8 no qual o artigo completo está alocado. Esse percentual de quase 50% dos respondentes para um único rótulo, bem como uma distribuição uniforme de votos entre outros dois termos e com o baixo número de consulta ao documento externo, pode significar que o conjunto de tópicos extraídos por meio da modelagem de tópicos pode ser considerado forte.

O Gráfico 41 apresenta os resultados sobre qual rótulo melhor representa o quarto conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.001*"dados" + 0.001*"publicação" + 0.001*"publicações" + 0.001*"recursos" + 0.000*"publicações_ampliadas" + 0.000*"ampliadas" + 0.000*"modelo" + 0.000*"digitais" + 0.000*"publicação_ampliada" + 0.000*"ampliada"].

Gráfico 41 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 4

Fonte: Elaborado pelo autor.

Constam, dentre os resultados alcançados na identificação da suposição do tópico, os números, percentuais e rótulos 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 2 ou 2,3% para Mediação, Circulação e Apropriação da Informação; 2 ou 2,3% para Política e Economia da Informação; 3 ou 3,4% para Gestão da Informação e do Conhecimento; 3 ou 3,4% para Organização e Representação do Conhecimento; 22 ou 25,3% para Informação e Tecnologia; 52 ou 59,8% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; e 2 ou 2,3% para a opção "Outros", sendo "Comunicação Científica" e "Produção, Editoração e Comunicação Científica".

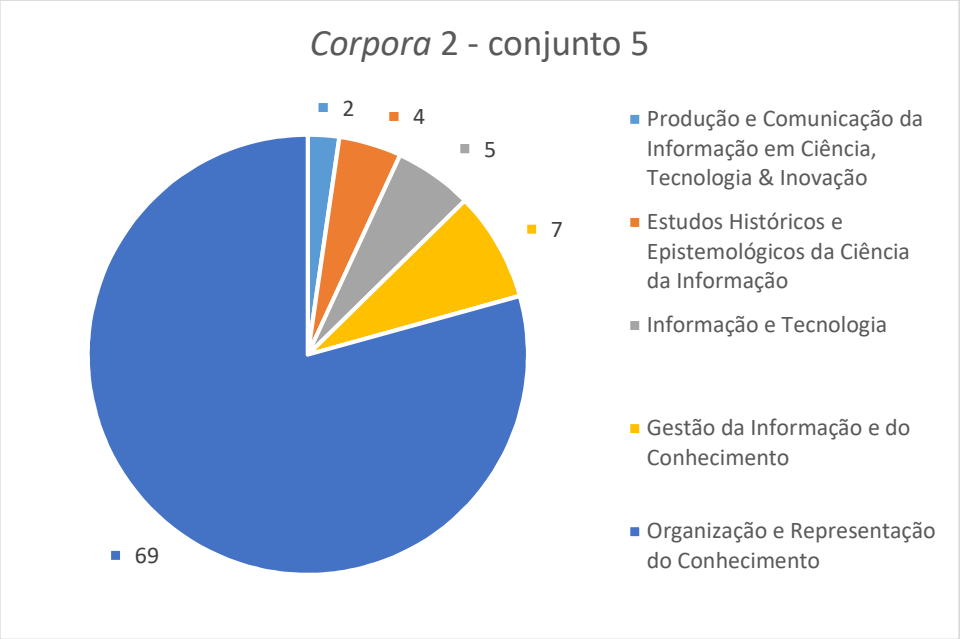
Dentre os respondentes, 5 ou 5,7% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado "PUBLICAÇÕES AMPLIADAS: um novo modelo de publicação acadêmica para o ambiente de e-science", possui as palavras-chave Publicações Ampliadas, Dados Digitais de Pesquisa, Norma oai-ore e E-science.

O rótulo Informação e Tecnologia foi a segunda opção mais selecionada pelos respondentes, atingindo 25,3%, enquanto o rótulo Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação foi a opção que atingiu o maior percentual, sendo 59,8% dos respondentes. Embora o conjunto de termos tenha remetido em quase 60% dos participantes para o GT-7 do ENANCIB, o artigo completo, utilizando de técnicas empregadas na literatura para analisar o documento que melhor representa o conjunto de termos, está enquadra na verdade no segundo rótulo mais votado, sendo no GT-8 do ENANCIB. Dessa forma, pode-se considerar que os rótulos não possuem características fortes ou que a atividade de catalogação foi realizada pelos respondentes levando em consideração os aspectos subjetivos, bem como vivências e experiências de cada respondente.

Também é possível identificar entre os resultados rótulos que podem ser descartados, uma vez que não condizem com os conjuntos de termos, tampouco com o documento disponibilizado para consulta, como por exemplo, os rótulos Estudos Históricos e Epistemológicos da Ciência da Informação, Política e Economia da Informação e Gestão da Informação e do Conhecimento.

O Gráfico 42 apresenta os resultados sobre qual rótulo melhor representa o quinto conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.001*"indexação" + 0.001*"facetada" + 0.001*"analistas" + 0.001*"usabilidade" + 0.001*"classificação" + 0.000*"etiquetagem" + 0.000*"taxonomia" + 0.000*"taxonomia_facetada" + 0.000*"usuário" + 0.000*"faceted"].

Gráfico 42 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 5



Fonte: Elaborado pelo autor.

Constam entre os rótulos selecionados pelos respondentes 2 ou 2,3% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 4 ou 4,6% Estudos Históricos e Epistemológicos da Ciência da Informação; 5 ou 5,7% Informação e Tecnologia; 7 ou 8,0% Gestão da Informação e do Conhecimento; e 69 ou 79,3% Organização e Representação do Conhecimento.

Entre os respondentes, 4 ou 4,6% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “Modelo de colaboração para indexação de recursos web”, possui as palavras-chave Classificação Facetada, Folksonomia, Web 2.0, Colaboração e Indexação.

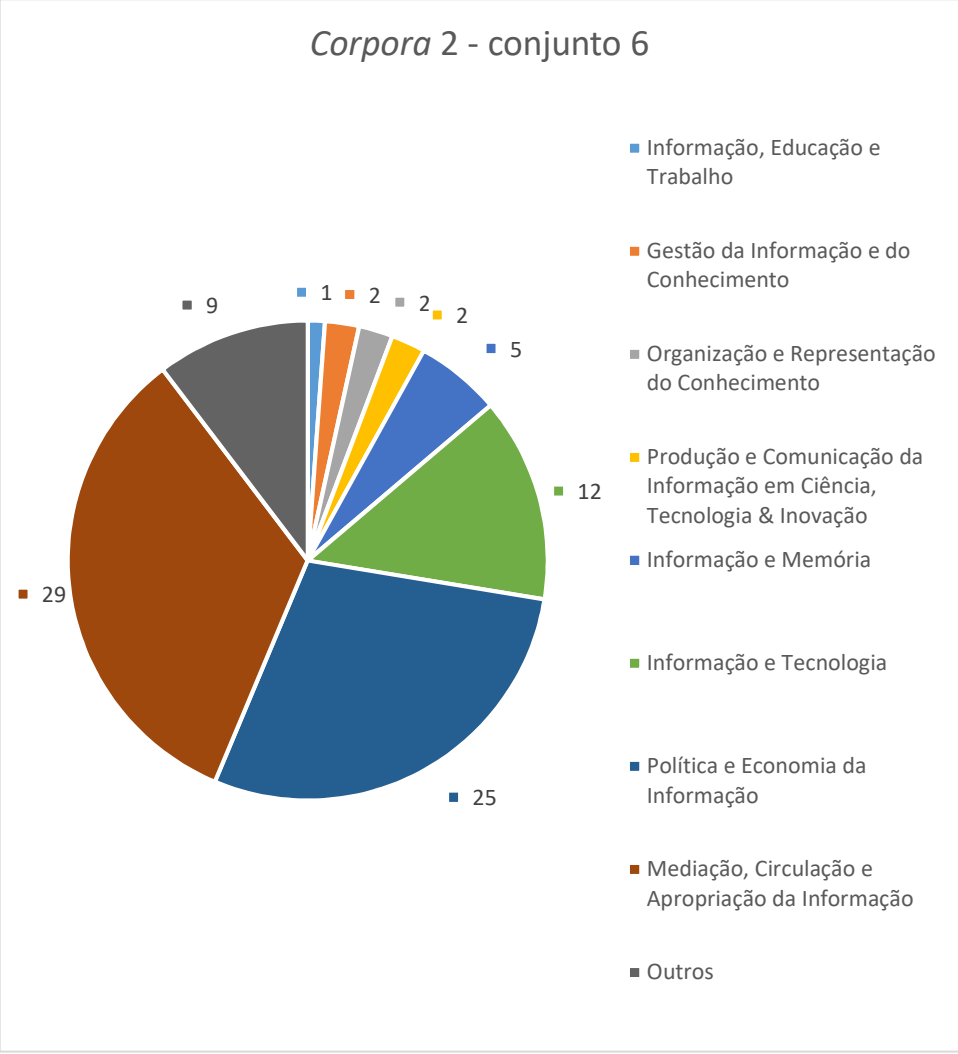
O conjunto de termos extraído por meio de modelagem de tópicos pode ser considerado forte, uma vez que, além do rótulo Organização e Representação do Conhecimento ter sido selecionado por 79,3% dos respondentes e o baixo número de consultas ao documento externo, o artigo

está alocado ao GT-2 do ENANCIB, sendo mesmo do rótulo. Faz-se necessário destacar que nenhum rótulo foi sugerido pelos respondentes.

Dois dos cinco rótulos selecionados pelos respondentes podem ser descartados do gráfico, pois não possuem aderência ao conjunto de termos e ao documento disponibilizado para consulta, sendo eles Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação e Estudos Históricos e Epistemológicos da Ciência da Informação.

O Gráfico 43 apresenta os resultados sobre qual rótulo melhor representa o sexto conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.001*"cinema" + 0.001*"canais" + 0.001*"televisão" + 0.001*"audiovisual" + 0.001*"lei" + 0.001*"brasil" + 0.001*"globo" + 0.001*"filmes" + 0.000*"brasileiro" + 0.000*"programação"].

Gráfico 43 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 6



Fonte: Elaborado pelo autor.

Dentre as suposições dos rótulos apresentados aos respondentes para o conjunto de termos obtiveram-se os resultados de 1 ou 1,1% para Informação, Educação e Trabalho; 2 ou 2,3% para Gestão da Informação e do Conhecimento; 2 ou 2,3% para Organização e Representação do Conhecimento; 2 ou 2,3% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 5 ou 5,7% para Informação e Memória; 12 ou 13,8% para Informação e Tecnologia; 25 ou 28,7% para Política e Economia da Informação; 29 ou 33,3% para Mediação, Circulação e Apropriação da Informação; e 9 ou 10,3% para a opção "Outros", sendo "Audiovisuais", "Comunicação" duas vezes, "Comunicação da Informação", "Comunicação Social", "Direito e Industria entretenimento", "Mídias informacionais", "Não me parece CI - Comunicação" e "outra que não Ciência da Informação".

Entre os respondentes, 13 ou 14,9% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado "As barreiras à integração do audiovisual e a nova lei da tv a cabo", possui as palavras-chave Cinema, Televisão, Comunicação, Políticas Cinematográficas e TV a cabo.

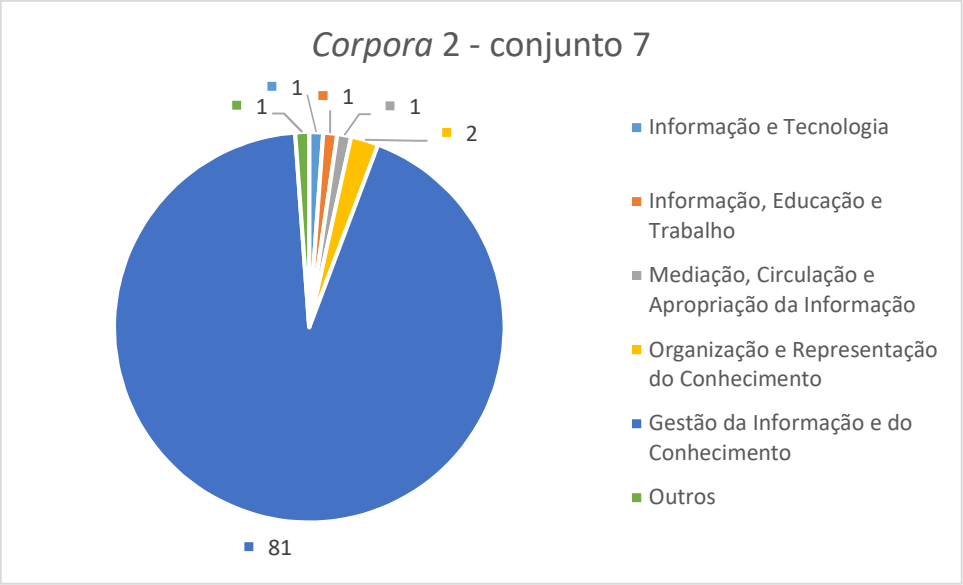
O conjunto de termos possui características específicas que permitem apontar para um único documento que melhor os representa, entretanto, é possível identificar três rótulos em destaque que foram selecionados respondentes, sendo 13,8% para Informação e Tecnologia; 28,7% para Política e Economia da Informação; e 33,3% para Mediação, Circulação e Apropriação da Informação. O número de respondentes que consultaram o documento externo também é representativo quando se comparado aos demais documentos que foram consultados. É possível observar que não existe uma coesão com relação ao rótulo que melhor represente o conjunto de termos. Um exemplo disso está no fato de o artigo científico representado pelo conjunto de termos fazer parte do GT-7 do ENANCIB e ter sido a segunda opção mais votada pelos respondentes.

Outro ponto a ser analisado está no número elevado de novas suposições realizadas pelos respondentes, inclusive, apontando para áreas diferentes da que realmente o artigo está alocado. O tópico constituído pelo conjunto de termos, mesmo apontando para um único documento, possui termos de difícil

classificação que necessitam de um aprofundamento nas informações contidas no documento, entretanto, o tópico não pode ser considerado fraco, pois os termos são coesos e possuem pesos equilibrados.

O Gráfico 44 apresenta os resultados sobre qual rótulo melhor representa o sétimo conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.001*"gestão" + 0.001*"conhecimento" + 0.001*"modelo" + 0.001*"documentos" + 0.001*"empresa" + 0.001*"organização" + 0.001*"gestão_conhecimento" + 0.001*"gad (gestão arquivística de documentos)" + 0.001*"motivacional" + 0.001*"modelagem"].

Gráfico 44 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 7



Fonte: Elaborado pelo autor.

Constam dentre os resultados alcançados na identificação da suposição do tópico os números, percentuais e rótulos 1 ou 1,1% para Informação e Tecnologia; 1 ou 1,1% Informação, Educação e Trabalho; 1 ou 1,1% Mediação, Circulação e Apropriação da Informação; 2 ou 2,3% Organização e Representação do Conhecimento; 81 ou 93,1% Gestão da Informação e do Conhecimento; e 1 ou 1,1% para opção “Outros”, sendo “Arquitetura da Informação”.

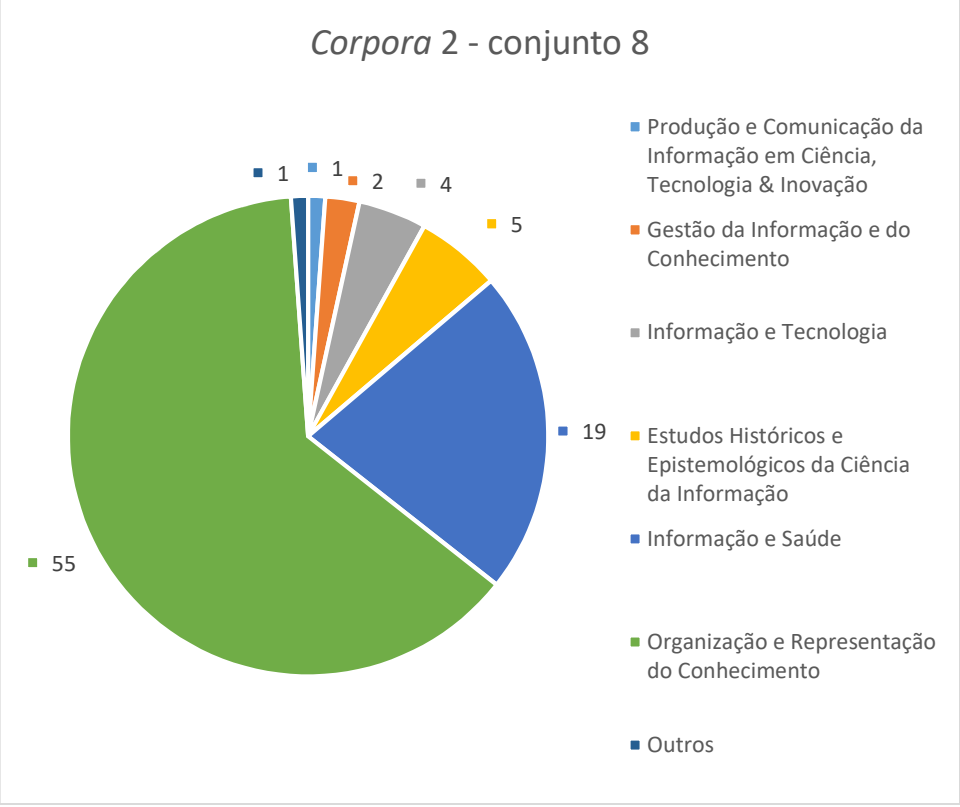
Dentre os respondentes, 7 ou 8,0% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “Quão estratégica pode ser a gestão arquivística de documentos? Aportes da

arquitetura corporativa", possui as palavras-chave Gestão Arquivística de Documentos, GAD, Estratégia, Modelo Motivacional e Arquitetura Corporativa.

O conjunto de termos apresentado possibilitou que 93,1% dos respondentes escolhessem o rótulo Gestão da Informação e do Conhecimento, sendo o maior percentual alcançado para um rótulo dos 30 tópicos apresentados durante o processo de validação da pesquisa, entretanto, apenas 1,1% dos respondentes optou pelo rótulo Informação e Tecnologia, sendo o rótulo igual do GT-8 do ENANCIB, ao qual o trabalho de fato está alocado. Trata-se de um conjunto de termos com características fortes, entretanto, faz-se necessário ao profissional especialista, caso tenha acesso, explorar mais informações sobre documento que melhor representa o tópico.

O Gráfico 45 apresenta os resultados sobre qual rótulo melhor representa o oitavo conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.002*"ontologia" + 0.001*"ontologias" + 0.001*"domínio" + 0.001*"hemonto" + 0.001*"sangue" + 0.001*"etapa" + 0.001*"ontoforinfoscience" + 0.001*"ontology" + 0.001*"digitais" + 0.001*"construção"].

Gráfico 45 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 8



Fonte: Elaborado pelo autor.

Constam entre os rótulos selecionados pelos respondentes 1 ou 1,1% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 2 ou 2,3% para Gestão da Informação e do Conhecimento; 4 ou 4,6% Informação e Tecnologia; 5 ou 5,7% para Estudos Históricos e Epistemológicos da Ciência da Informação; 19 ou 21,8% para Informação e Saúde; 55 ou 63,2% para Organização e Representação do Conhecimento; e 1 ou 1,1% para a opção "Outros", sendo "Não sei informar qual rótulo melhor representa".

Entre os respondentes, 7 ou 8,0% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado "ONTOFORINFOSCIENCE: uma metodologia detalhada para construção de ontologias e sua aplicação no domínio da biomedicina", possui as palavras-chave Ontologias, Desenvolvimento de Ontologias, Ontologias Biomédicas e Hematologia.

O rótulo Organização e Representação do Conhecimento foi escolhido por 62,3% dos respondentes, seguido do rótulo Informação e Saúde, com 21,8%. Os termos que constituem o tópico possuem aderência para ambos os rótulos. O documento disponibilizado para consulta está alocado ao GT-2 do ENANCIB, sendo o mesmo do rótulo com maior representatividade. O fato de o rótulo possuir um percentual elevado quando se comparado aos demais rótulos, o número baixo de consultas realizadas por meio do documento que melhor representa os termos e a coerência na qualidade dos termos e pesos permite inferir que o tópico extraído através da modelagem de tópicos possui características fortes.

Os resultados poderiam obter percentuais mais altos, uma vez que opções de rótulos escolhidas pelos respondentes como Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, Gestão da Informação e do Conhecimento, Estudos Históricos e Epistemológicos da Ciência da Informação e a opção Outros não possui aderência ao conjunto de termos, tampouco quanto ao documento disponibilizado para consulta que melhor representa o tópico.

O Gráfico 46 apresenta os resultados sobre qual rótulo melhor representa o nono conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.003*"serviços" + 0.002*"marketing" + 0.002*"informativos" +

0.002*"serviços_informativos" + 0.001*"orientação" + 0.001*"gestão" + 0.001*"prestadora" + 0.001*"lde" + 0.001*"unidade" + 0.001*"prestadora_serviços"].

Gráfico 46 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 9

Fonte: Elaborado pelo autor.

Dentre os rótulos selecionados pelos respondentes estão 1 ou 1,1% para Informação e Tecnologia; 1 ou 1,1% para Política e Economia da Informação; 1 ou 1,1% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 2 ou 2,3% para Estudos Históricos e Epistemológicos da Ciência da Informação; 6 ou 6,9% para Informação, Educação e Trabalho; 13 ou 14,9% para Mediação, Circulação e Apropriação da Informação; 62 ou 71,3% para Gestão da Informação e do Conhecimento; e 1 ou 1,1% para a opção Outros, sendo "?".

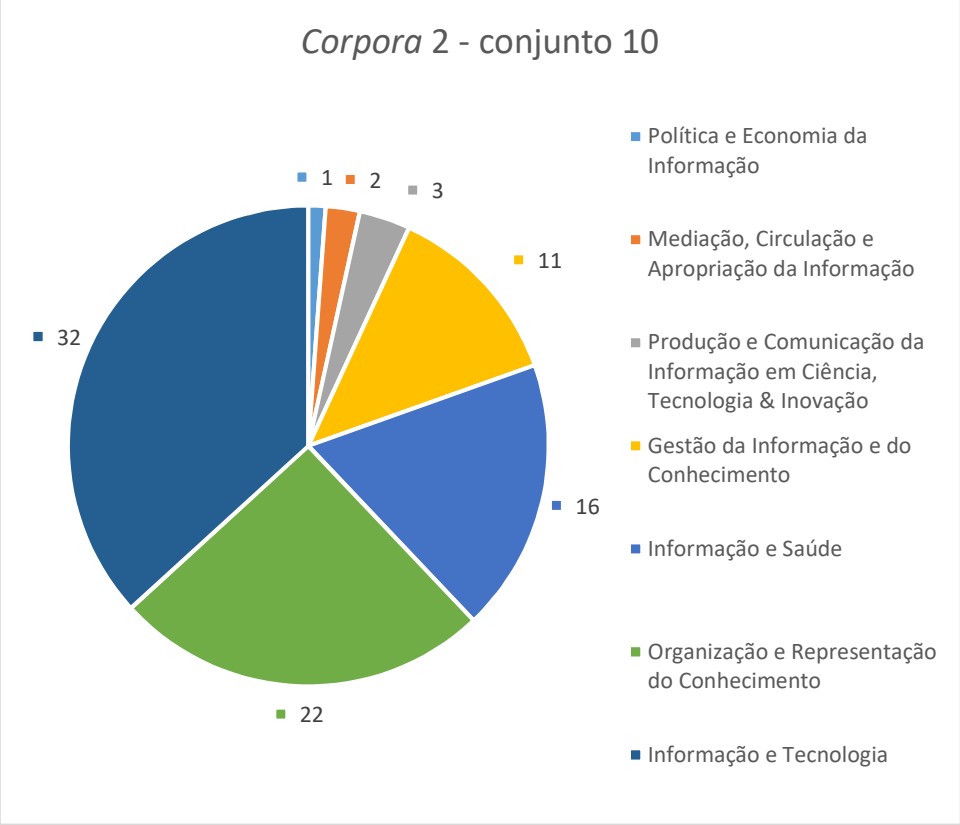
Entre os respondentes, 9 ou 10,3% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado "Fundamentos teóricos da orientação de marketing na gestão de serviços informativos", possui

as palavras-chave Marketing da Informação, Orientação de Marketing, Serviço de Informação, Gestão e Teoria.

O rótulo Gestão da Informação e do Conhecimento foi selecionado por 71,3% dos respondentes, sendo o mesmo do GT-4 do ENANCIB, ao qual o documento que melhor representa os termos está alocado. Tanto a qualidade dos termos e pesos quanto o rótulo mais votado permitem inferir que se trata de um tópico com características fortes.

O Gráfico 47 apresenta os resultados sobre qual rótulo melhor representa o décimo conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.003*"dados" + 0.001*"reuso" + 0.001*"conteúdo" + 0.001*"data" + 0.001*"reuso_dados" + 0.001*"arquétipos" + 0.001*"sistemas" + 0.001*"openehr" + 0.001*"sentenças" + 0.000*"metadados].

Gráfico 47 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 10



Fonte: Elaborado pelo autor.

Constam entre os rótulos selecionados pelos respondentes 1 ou 1,1% para Política e Economia da Informação; 2 ou 2,3% para Mediação, Circulação e Apropriação da Informação; 3 ou 3,4% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 11 ou 12,6% para Gestão da

Informação e do Conhecimento; 16 ou 18,4% para Informação e Saúde; 22 ou 25,3% para Organização e Representação do Conhecimento; 32 e 36,8% para Informação e Tecnologia.

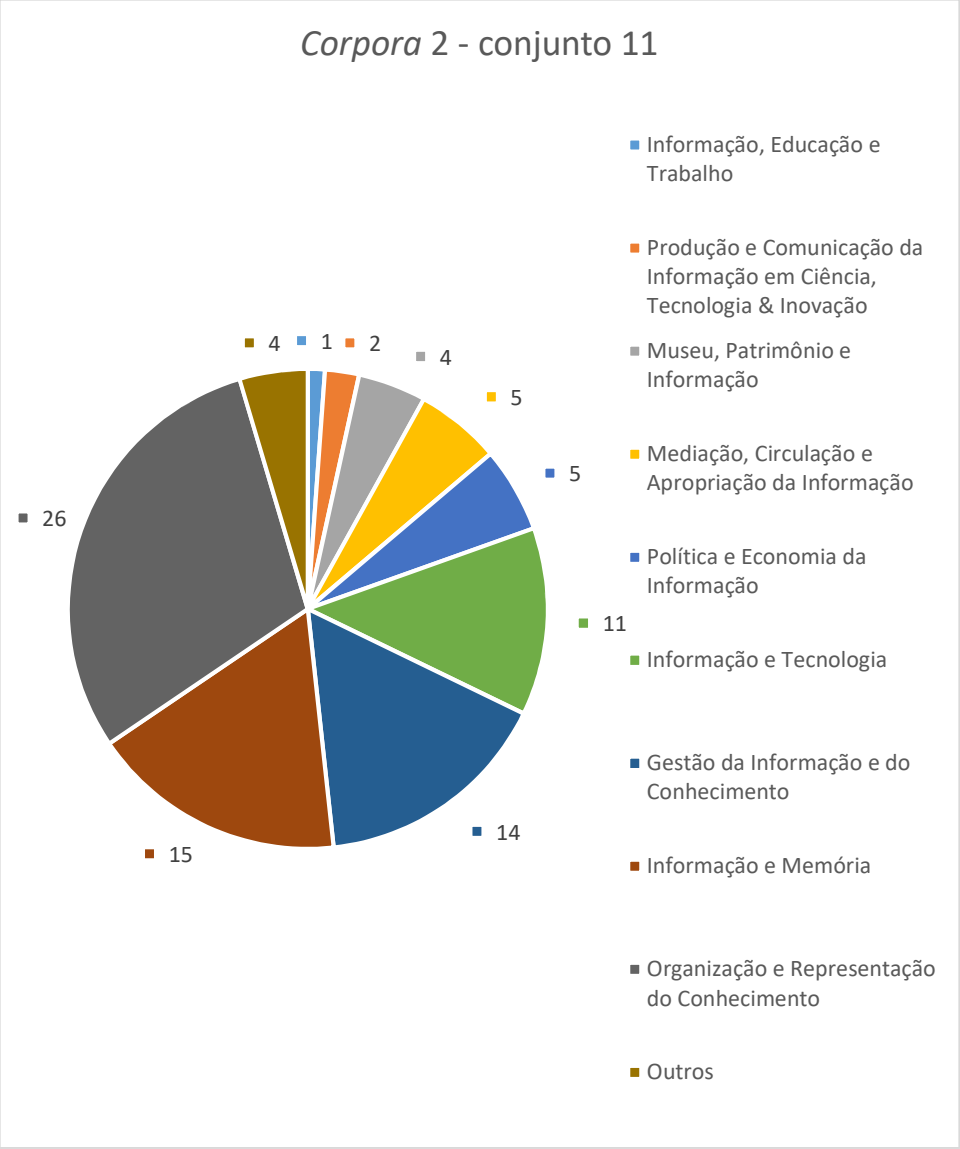
Entre os respondentes, 10 ou 15,5% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “IMPLEMENTANDO O PRONTUÁRIO ELETRÔNICO OPENEHR EM SISTEMAS GESTORES DE CONTEÚDO: similitude entre arquétipos e conteúdos”, possui as palavras-chave Registro Eletrônico de Saúde, Interoperabilidade, Arquétipos OpenEHR e Sistema Gestor de Conteúdos.

O conjunto de termos permite inferir diversos rótulos, como apresentado nos resultados alcançados, justamente por envolver áreas que possam se correlacionar com representação, gestão, saúde e tecnologias. Entretanto, esse tipo de situação pode apresentar dificuldades no processo de classificação realizado pelo profissional especialista.

Nesses casos, técnicas para obter mais informações sobre o documento são indicadas na literatura e essenciais para um resultado mais assertivo. O rótulo com a terceira maior representatividade, sendo Informação e Saúde com 25,3% da escolha realizada pelos participantes da pesquisa, é o mesmo do qual o documento que melhor representa os termos do tópico está alocado no ENANCIB, diferente do rótulo escolhido pela maioria dos participantes. Faz-se necessário destacar que os pesos dos termos estão equilibrados. Mesmo assim, trata-se de um tópico de difícil classificação.

O Gráfico 48 apresenta os resultados sobre qual rótulo melhor representa o 11º conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.000*"documentos" + 0.000*"fotografia" + 0.000*"audiovisuais" + 0.000*"documentos_audiovisuais" + 0.000*"sonoros" + 0.000*"iconográficos" + 0.000*"ctdais" + 0.000*"iconográficos_sonoros" + 0.000*"audiovisuais_iconográficos" + 0.000*"documentos_audiovisuais_iconográficos"].

Gráfico 48 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 11



Fonte: Elaborado pelo autor.

Dentre a quantidade, o percentual e os rótulos selecionados pelos respondentes está 1 ou 1,1% para Informação, Educação e Trabalho; 2 ou 2,3% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 4 ou 4,6% para Museu, Patrimônio e Informação; 5 ou 5,7% para Mediação, Circulação e Apropriação da Informação; 5 ou 5,7% para Política e Economia da Informação; 11 ou 12,6% para Informação e Tecnologia; 14 ou 16,1% para Gestão da Informação e do Conhecimento; 15 ou 17,2% para Informação e Memória; 26 ou 29,9% para Organização e Representação do Conhecimento; e 4 ou 4,6% para a opção “Outros”, sendo “Documentação”, “Documentação Arquivística”, “Estudos Epistemológicos em Arquivologia” e “Materialidade da Informação, Coleções e Constituições”.

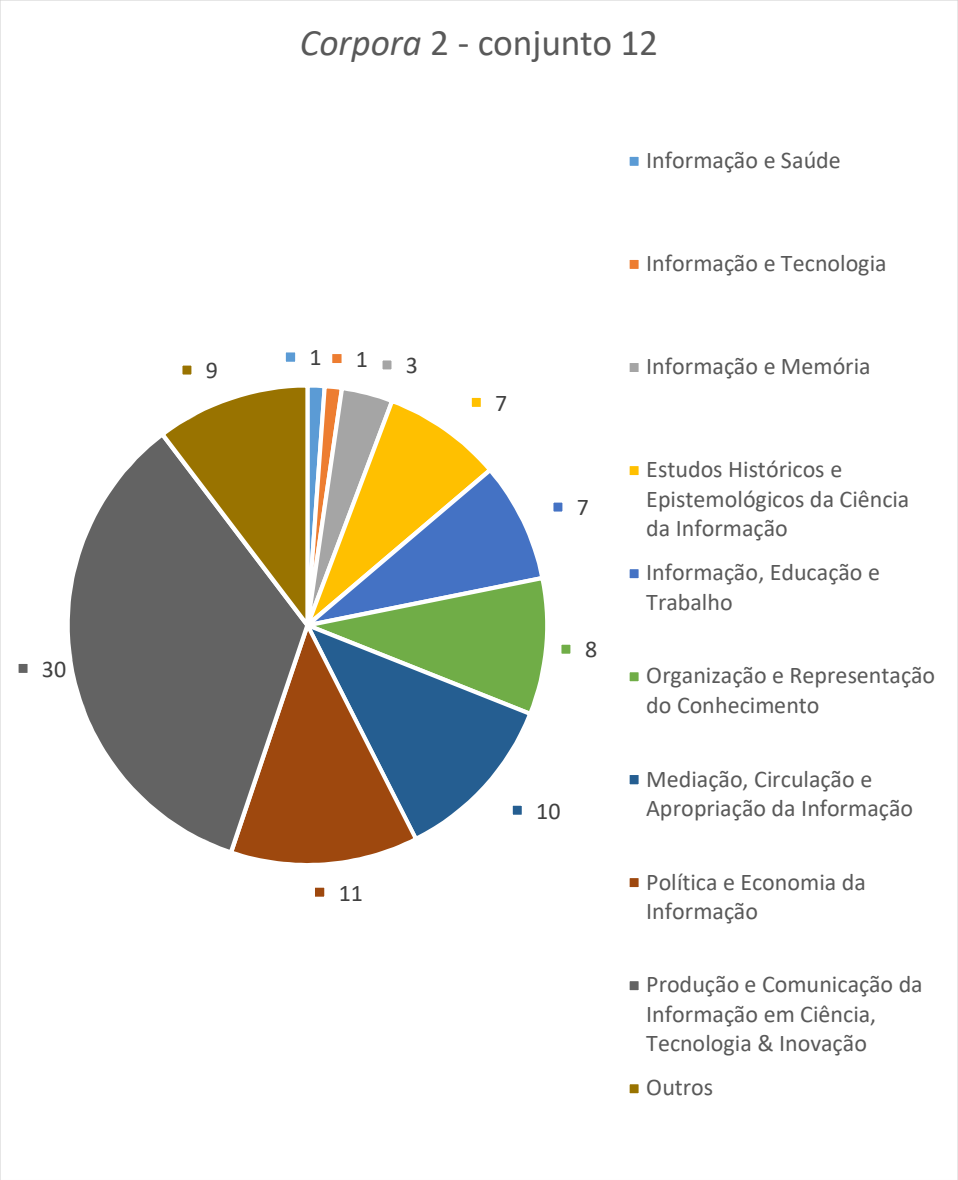
Entre os respondentes, 18 ou 20,7% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “Ações da câmara técnica de documentos audiovisuais, iconográficos e sonoros - CTDAIS para institucionalização de documentos não textuais”, possui as palavras-chave Documentos Imagéticos, Arquivologia, Documento de Arquivo e Ciência da Informação.

O conjunto de termos permite inferir diversas áreas do conhecimento, assim como apresentado nos resultados por meio da variação de escolhas selecionados pelos respondentes. Isso pode correr porque o tópico não possui um termo chave que possa indicar um rótulo de maneira mais assertiva. Além disso, os termos podem ser empregados em pesquisas de quaisquer rótulos apresentados.

Embora exista uma coesão entre os termos e um padrão de normalização nos pesos, o que permite inferir que seja um tópico com caracteristica forte, os resultados da validação do tópico apresentam um equilíbrio entre os rótulos escolhidos, sendo Organização e Representação do Conhecimento a opção de 29,9% dos respondentes. Outro fator que contribui para esse percentual e equilíbrio entre os rótulos está no percentual de 20,7% dos respondentes que consultaram o documento que melhor representa o tópico. Todavia, o rótulo escolhido pela maioria dos respondentes é o mesmo ao qual o artigo completo está alocado junto ao ENANCIB, sendo o GT-2.

O Gráfico 49 apresenta os resultados sobre qual rótulo melhor representa o 12º conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.000*"feministas" + 0.000*"feminismo" + 0.000*"estudos_feministas" + 0.000*"periódico" + 0.000*"estudos" + 0.000*"mulheres" + 0.000*"feminista" + 0.000*"revista_estudos_feministas" + 0.000*"revista_estudos" + 0.000*"produtivos"].

Gráfico 49 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 12



Fonte: Elaborado pelo autor.

Dentre as suposições dos rótulos apresentados aos respondentes para o conjunto de termos, obtiveram-se os resultados 1 ou 1,1% para Informação e Saúde; 1 ou 1,1% para Informação e Tecnologia; 3 ou 3,4% para Informação e Memória; 7 ou 8,0% para Estudos Históricos e Epistemológicos da Ciência da Informação; 7 ou 8,0% para Informação, Educação e Trabalho; 8 ou 9,2% para Organização e Representação do Conhecimento; 10 ou 11,5% para Mediação, Circulação e Apropriação da Informação; 11 ou 12,6% para Política e Economia da Informação; 30 ou 34,5% para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; e 9 ou 10,3% para a opção “Outros”, sendo “?”, “Feminismo”, “Informação e Bibliometria”, “Informação e Gênero”, “Informação e

Sociedade”, “Métricas e Indicadores Bibliométricos”, “Outra que não Ciência da Informação”, “Produção Científica e Bibliometria” e “Produção, Editoração e Comunicação Científica”.

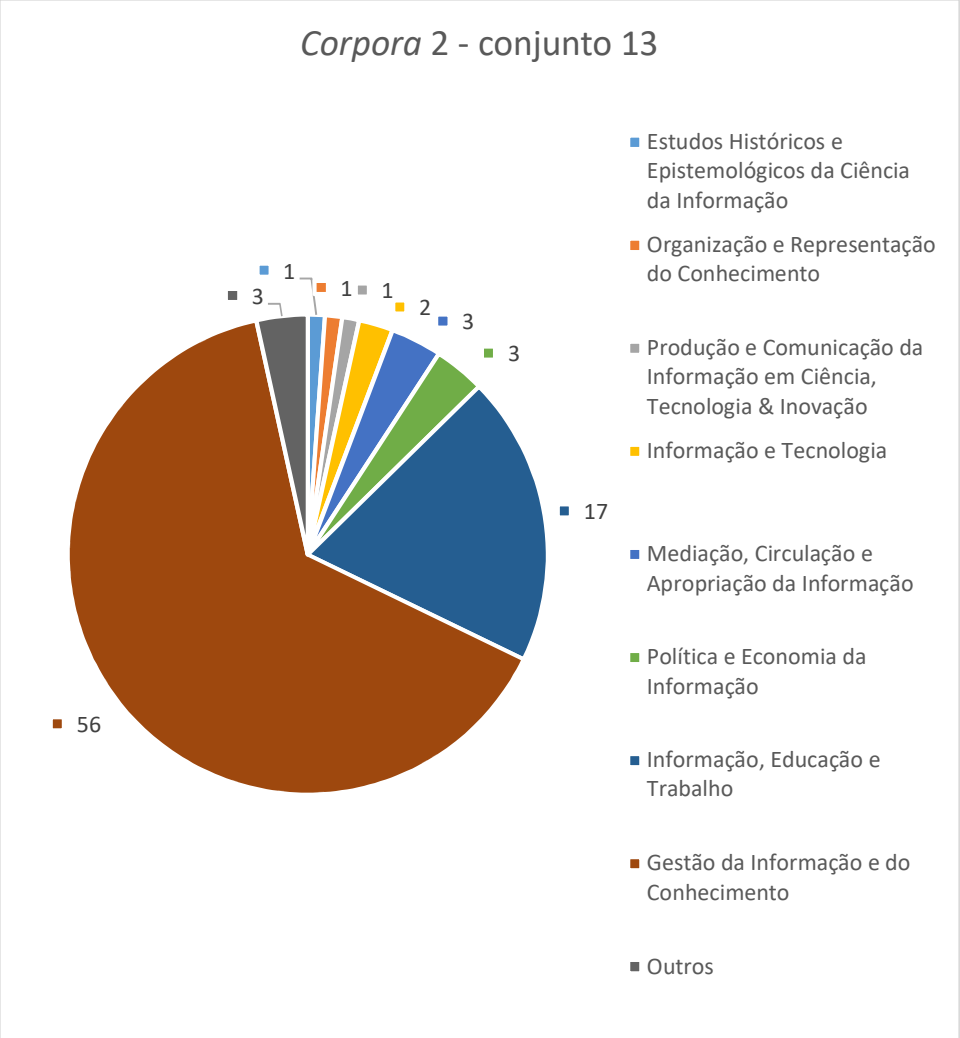
Entre os respondentes, 9 ou 10,3% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “FEMINISMOS E ESTUDOS DE GÊNERO: uma abordagem bibliométrica”, possui as palavras-chave Feminismo, Estudos de Gênero, Produção Científica.

O conjunto de termos também possibilitou uma variação entre os rótulos selecionados pelos respondentes e um número elevado de 10,3% dos respondentes que sugeriram outros rótulos. Isso ocorreu justamente pelo tópico não possuir um termo chave, o que de fato não permite caracterizar como um tópico fraco, já que os termos são coesos e os pesos equilibrados.

O rótulo com maior representatividade foi Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, com 34,5% da escolha dos respondentes, sendo o mesmo do GT-7 do ENANCIB no qual o artigo completo que melhor representa os termos está alocado.

O Gráfico 50 apresenta os resultados sobre qual rótulo melhor representa o 13º conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.004*"conhecimento" + 0.003*"informação" + 0.001*"gestão" + 0.001*"organização" + 0.001*"conceitos" + 0.001*"processo" + 0.001*"competências" + 0.001*"meio" + 0.001*"seleção" + 0.001*"trabalho"].

Gráfico 50 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 13



Fonte: Elaborado pelo autor.

Constam, entre os rótulos selecionados pelos respondentes, 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 1 ou 1,1% para Organização e Representação do Conhecimento; 1 ou 1,1%, para Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; 2 ou 2,3% para Informação e Tecnologia; 3 ou 3,4% para Mediação, Circulação e Apropriação da Informação; 3 ou 3,4% para Política e Economia da Informação; 17 ou 19,5% para Informação, Educação e Trabalho; 56 ou 64,4% para Gestão da Informação e do Conhecimento; e 3 ou 3,4% para a opção “Outros”, sendo “Cultura e Clima Organizacional”, “Gestão da Informação e Trabalho” e “Gestão de Pessoas”.

Dentre os respondentes, 7 ou 8,0% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “Sistemática de

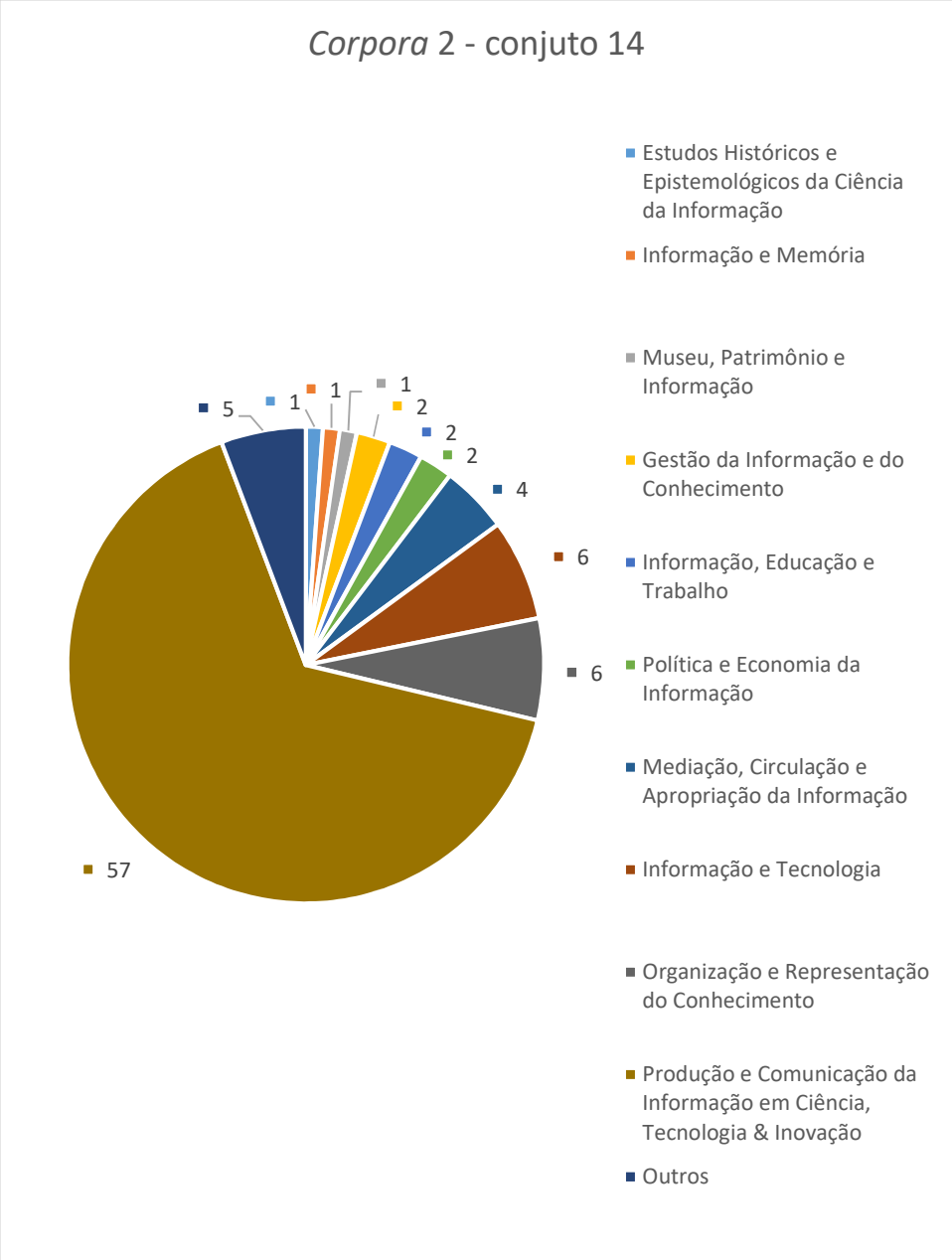
seleção de servidores para as funções de confiança, baseada na abordagem das competências", possui as palavras-chave Seleção de Pessoas por Competências, Função Gratificada e Organização Pública.

O tópico apresentou um conjunto de termos coesos e pesos equilibrados que permitiram que 64,4% dos respondentes optassem pelo rótulo Gestão da Informação e do Conhecimento, sendo o mesmo ao qual o artigo completo que melhor representa o conjunto de termos está publicado no GT-4 do ENANCIB. O tópico pode ser considerado forte por possuir um elevado percentual de respondentes que optaram pelo mesmo rótulo e baixo percentual de respondentes que consultaram o documento disponibilizado no questionário.

O resultado poderia ser mais expressivo, uma vez que as opções citadas no campo "Outros" se enquadram no rótulo Gestão da Informação e do Conhecimento. Além disso, os rótulos Estudos Históricos e Epistemológicos da Ciência da Informação, Organização e Representação do Conhecimento, Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, Informação e Tecnologia, Mediação, Circulação e Apropriação da Informação que somam 9,2% dos respondentes não possuem aderência aos termos e ao documento apresentado.

O Gráfico 51 apresenta os resultados sobre qual rótulo melhor representa o 14º conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.001*"periódicos" + 0.001*"zona" + 0.001*"bradford" + 0.001*"artigos" + 0.001*"bahia" + 0.001*"consumo" + 0.001*"documentos" + 0.001*"igc (Instituto de Geociências)" + 0.001*"docentes" + 0.000*"produção"].

Gráfico 51 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 14



Fonte: Elaborado pelo autor.

Dentre as suposições dos rótulos apresentados aos respondentes para o conjunto de termos, obtiveram-se os resultados 1 ou 1,1% para Estudos Históricos e Epistemológicos da Ciência da Informação; 1 ou 1,1% para Informação e Memória; 1 ou 1,1% para Museu, Patrimônio e Informação; 2 ou 2,3% para Gestão da Informação e do Conhecimento; 2 ou 2,3% para Informação, Educação e Trabalho; 2 ou 2,3% para Política e Economia da Informação; 4 ou 4,6% para Mediação, Circulação e Apropriação da Informação; 6 ou 6,9% para Informação e Tecnologia; 6 ou 6,9% para Organização e Representação do Conhecimento; 57 ou 65,5% para Produção e Comunicação

da Informação em Ciência, Tecnologia & Inovação; e 5 ou 5,7% para a opção “Outros”, sendo “Comunicação Científica, Produção Científica, Indicadores Bibliométricos”, “Estudos Métricos”, “Geociências”, “Produção Científica e Indicadores Bibliométricos” e “Produção, Editoração e Comunicação Científica”.

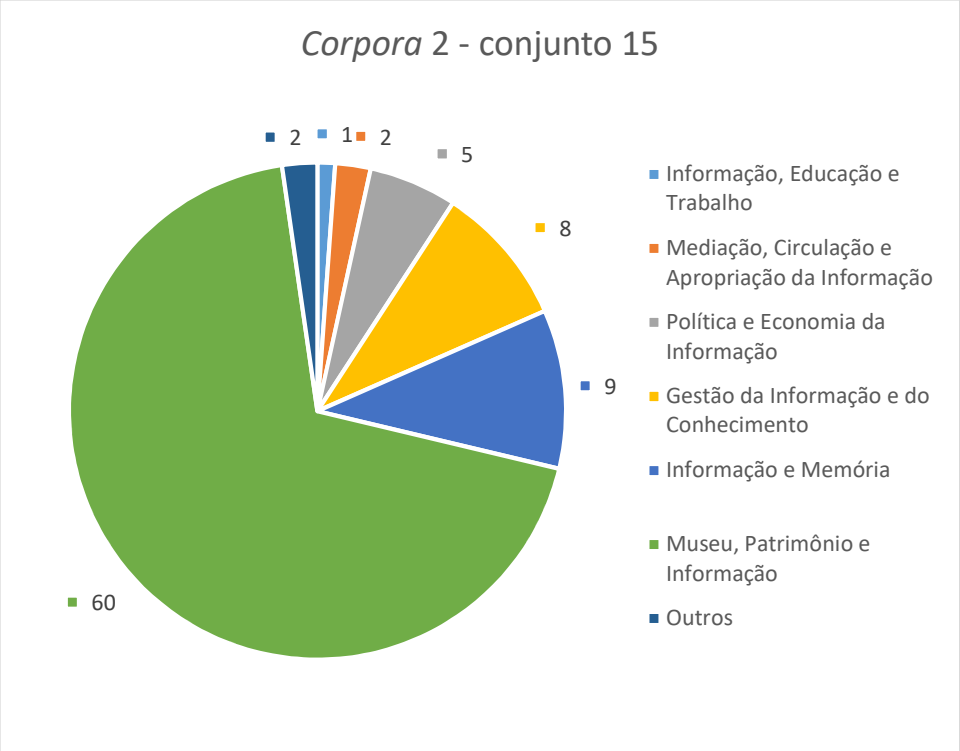
Dentre os respondentes, 8 ou 9,2% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado “PRODUÇÃO E CONSUMO NAS GEOCIÊNCIAS: estudo de dispersão em diferentes níveis de agregação”, possui as palavras-chave Lei de Bradford, Produtividade, Análise de Citação, Níveis de Agregação e Geociências.

O tópico possui termos coesos e pesos equilibrados, possibilitando que 65,5% dos respondentes optassem pelo rótulo Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, sendo o mesmo do GT-7 do ENANCIB ao qual o artigo completo que melhor representa o conjunto de termos está alocado. Outro fator a ser destacado está no baixo percentual dos respondentes que consultaram o artigo completo, podendo inferir que se refere a um tópico forte.

Entre os resultados também constam um percentual de 10,3% de rótulos somados como Estudos Históricos e Epistemológicos da Ciência da Informação, Informação e Memória, Museu, Patrimônio e Informação, Gestão da Informação e do Conhecimento, Informação, Educação e Trabalho, e Política e Economia da Informação que não possuem aderência com o conjunto de termos, tampouco quanto ao documento que melhor representa o tópico.

O Gráfico 52 apresenta os resultados sobre qual rótulo melhor representa o 15º conjunto de termos do *corpora* artigos científicos e resumos expandidos: [0.002*"conhecimento" + 0.002*"museu" + 0.001*"biblioteca" + 0.001*"gestão" + 0.001*"inovação" + 0.001*"pesquisa" + 0.001*"ferroviário" + 0.001*"políticas" + 0.001*"uso" + 0.001*"ferramentas"].

Gráfico 52 – Validação da modelagem de tópicos: *corpora* 2 – conjunto 15



Fonte: Elaborado pelo autor.

Dentre os rótulos selecionados pelos respondentes estão 1 ou 1,1% para Informação, Educação e Trabalho; 2 ou 2,3% para Mediação, Circulação e Apropriação da Informação; 5 ou 5,7% para Política e Economia da Informação; 8 ou 9,2% para Gestão da Informação e do Conhecimento; 9 ou 10,3% para Informação e Memória; 60 ou 69% para Museu, Patrimônio e Informação; e 2 ou 2,3% para a opção "Outros", sendo "Indefinido" e "Não me parece CI - Museologia – Memória".

Entre os respondentes, 6 ou 6,9% realizaram acesso ao documento que melhor representa o conjunto de termos disponibilizado através do *link* externo no questionário. O documento do tipo artigo completo, intitulado "Museus ferroviários e a trajetória da preservação do patrimônio ferroviário no Brasil", possui as palavras-chave Museologia, Museus, Museus Ferroviários, Patrimônio e PRESERVE (Programa de Preservação do Patrimônio Histórico do Ministério dos Transportes).

O Tópico com termos coesos e pesos equilibrados resultou num percentual de 69% dos respondentes que optaram pelo rótulo Museu, Patrimônio e Informação, sendo o mesmo GT-9 ao qual o documento que melhor representa o conjunto de termos está alocado. Tanto o elevado percentual escolhido por um

grupo de respondentes, quanto o baixo percentual de consultas ao documento externos, bem como as características dos termos permite inferir que se trata de um tópico com forte.

O termo 0.001*"gestão" pode ter contribuído para que 9,2% dos respondentes tenham optado pelo rótulo Gestão da Informação e do Conhecimento. Entretanto, o conjunto de termos não aponta indicativos para os rótulos de Informação, Educação e Trabalho, Mediação, Circulação e Apropriação da Informação, Política e Economia da Informação e Informação e Memória que, quando somados, alcançam 19,5%.

5.3.6. Considerações sobre a validação dos resultados

O processo de validação dos resultados obtidos por meio da modelagem de tópicos em *corpora* de dados constituídos por documentos formais da comunicação científica e realizado por profissionais que atuam na área da Ciência da Informação, sendo professores universitários e bibliotecários, possibilitou reflexões acerca da pesquisa em questão.

Dos 15 tópicos que representam o primeiro *corpora* de dados composto por documentos do tipo de teses e dissertações utilizados no processo de validação dos dados, 12 ou 80,0% foram relacionados à suposição dos rótulos de maneira mais assertiva, enquanto 3 ou 20,0% dos tópicos apresentaram incertezas na escolha do rótulo que melhor representa os conjuntos de termos, uma vez que os tópicos podem ser alocados em mais de um rótulo.

Ainda sobre o primeiro *corpora* de dados e o percentual de tópicos com maior volume de assertividade pelo número de respondentes, é possível destacar que 1 ou 6,7% dos tópicos foram selecionados por 80 a 89,9% dos respondentes; 1 ou 6,7% dos tópicos foram selecionados por 70 a 79,9% dos respondentes; 2 ou 13,3% dos tópicos foram selecionados por 60 a 69,9% dos respondentes; 3 ou 20,0% dos tópicos foram selecionados por 50 a 59,9% dos respondentes; 6 ou 40,0% dos tópicos foram selecionados por 40 a 49,9% dos respondentes; 1 ou 6,7% dos tópicos foram selecionados por 30 a 39,9% dos respondentes; e 1 ou 6,7% dos tópicos foram selecionados por 20 a 29,9% dos respondentes. Com esses dados é possível perceber que 11 ou 73,3% dos 15 tópicos apresentaram assertividade abaixo de 70% do volume de respondentes,

enquanto 4 ou 26,7% dos tópicos apresentaram assertividade igual ou acima dos 70% do volume de respondentes.

Já entre os tópicos extraídos do segundo *corpora* de dados constituído por artigos completos e resumos expandidos utilizados no formulário para validação dos resultados apresentaram assertividade nas suposições dos tópicos em 10 ou 66,7%, enquanto 5 ou 33,3% dos tópicos foram classificados em rótulos diferentes aos dos documentos que melhor representavam os tópicos e 7 ou 46,7% dos rótulos poderiam ser classificados em mais de um rótulo mediante a distribuição e equilíbrio do percentual de votação das suposições dos rótulos definidas pelos respondentes.

Sobre o segundo *corpora* de dados e o percentual de tópicos com maior volume de assertividade pelo número de respondentes, é possível destacar que 1 ou 6,7% dos tópicos foram selecionados por 90 a 100,0% dos respondentes; 2 ou 13,3% dos tópicos foram selecionados por 70 a 79,9% dos respondentes; 5 ou 33,3% dos tópicos foram selecionados por 60 a 69,7% dos respondentes; 1 ou 6,7% dos tópicos foram selecionados por 50 a 59,9% dos respondentes; 2 ou 13,3% dos tópicos foram selecionados por 40 a 49,9% dos respondentes; 3 ou 20,0% dos tópicos foram selecionados por 30 a 39,9% dos respondentes; e 1 ou 6,7% dos tópicos foram selecionados por 20 a 29,9% dos respondentes. Dessa forma, apenas 3 ou 20,0% dos tópicos foram selecionados por 70 ou mais por cento dos respondentes, enquanto 12 ou 80% dos tópicos apresentaram suposições de assertividade abaixo de 70%.

O *corpora* de dados constituído por documentos do tipo teses e dissertações apresentou três tópicos que podem ser relacionados em mais de um rótulo enquanto o *corpora* com documentos do tipo artigos completos e resumos expandidos apresentou cinco tópicos que não foram classificados de acordo com os documentos que melhor representavam os conjuntos de termos. Além disso, apresentou uma maior distribuição entre as opções de rótulos selecionados em tópicos, alcançando 7 ou 46,7% dos tópicos com que apresentam rótulos com percentuais próximos.

Constam, entre os resultados, rótulos selecionados com baixos percentuais de representatividade, incondizentes com os conjuntos de termos ou mesmo com os documentos disponibilizados para consulta. Entretanto, esses resultados não interferem nas suposições assertivas realizadas pelos

respondentes, tampouco nos rótulos que apresentam representatividade. Esse tipo de resultado poderia ser descartado, entretanto, levou-se em consideração o fator subjetividade dos respondentes. Acredita-se que esse grupo de resultados tenha surgido mediante o tamanho do questionário, que pode se tornar cansativo ao responder.

Os tópicos selecionados para validação dos resultados apresentam características fortes entre os termos de forma que apontam para somente um documento que melhor representa cada tópico. Entretanto, especificamente no segundo *corpora* de dados, contemplado por documentos publicados nas edições do ENANCIB, foi possível identificar um número superior de tópicos que podem ser associados a mais de um dos rótulos, sendo esses rótulos extraídos dos nomes dos GTs do evento. Esse fenômeno ou incoerência apresentada nos resultados dos *corpora* de dados, constituído por artigos completos e resumos expandidos, pode estar relacionado à interdisciplinaridade da área da Ciência da Informação, bem como na transversalidade dos GTs do ENANCIB que abrangem mais de uma área do conhecimento.

O questionário utilizado para validação dos tópicos apresentou duas possibilidades para que os respondentes utilizassem no processo de identificação e suposição do rótulo que melhor representasse cada conjunto de termos. A primeira foi realizada por meio da interpretação dos termos, levando em consideração os aspectos subjetivos, bem como conhecimento e experiência do especialista na área. Já a segunda possibilidade foi baseada no referencial teórico desta tese, que se baseia em técnicas de análise de assunto, disponibilizando aos respondentes documentos que representassem os conjuntos de termos para serem explorados.

A média de acessos aos documentos que melhor representam os tópicos do primeiro *corpora* de dados foi de 14,9% enquanto no segundo *corpora* de dados foi de 9,1%. No perfil dos participantes, fica claro que 79,3% atuam como docentes em universidades e 19,5% atuam como bibliotecários, o que não significa que exercem a função de indexador. Esse percentual pode justificar o baixo número de consultas realizadas junto aos documentos que melhor representam os conjuntos de termos, uma vez que não se trata de uma atividade cotidiana dos respondentes. Entretanto, esse fato não desclassifica a pesquisa, uma vez que os 62,9% dos respondentes são formados em biblioteconomia e

cursaram disciplinas de catalogação presentes nas matrizes curriculares dos cursos. Além disso, dentre os respondentes, constam professores e pesquisadores atuantes da área da Ciência da Informação.

O baixo percentual de consultas aos documentos do segundo *corpora* de dados permite inferir que a validação dos resultados foi realizada por meio da subjetividade, levando em consideração os aspectos cognitivos, conhecimento e experiências dos respondentes, o que pôde ter resultado numa divergência de 33,3% entre os rótulos mais votados com os rótulos onde realmente os artigos completos estão alocados junto ao ENANCIB. Acredita-se que a prática de consulta das informações por parte dos respondentes aos documentos que melhor representam os tópicos poderia resultar em um percentual maior de assertividade entre termos, documentos e rótulos.

Além disso, em ambos os *corpora* de dados foram identificadas sugestões para novos rótulos que representassem os tópicos, entretanto, muitas das sugestões fazem parte dos rótulos transdisciplinares dos GTs do ENANCIB, de acordo com os ementários[56]. Também constam sugestões que não contemplam no ementário dos GTs, como Comunicação Científica englobando também Produção Científica, citado 18 vezes, Estudos Métricos englobando Biometria, citado 9 vezes, e Materialidade da Informação. Já o termo Informação e Sociedade englobando Informação e Cultura e Informação e Gênero foi citado 8 vezes.

[56] Ementas dos GTs do ENANCIB. Disponível em: http://www.enancib2019.ufsc.br/gts/. Acesso em: 25/04/2020.

6. CONSIDERAÇÕES FINAIS

A pesquisa, norteada a partir do referencial teórico e empírico, possibilitou a execução de algoritmos de *Machine Learning* por meio de técnicas de modelagem de tópicos em *corpora* de dados constituídos por documentos científicos da área da Ciência da Informação. Os diferentes tipos de resultados possibilitaram gerar discussões sobre os vieses acadêmico e técnico.

Do ponto de vista acadêmico, as contribuições estão relacionadas ao comportamento de termos da Ciência da Informação ao longo dos anos analisados, ao distanciamento e à proximidade entre disciplinas do núcleo da área encontrado na literatura com os resultados alcançados na fase empírica da pesquisa e na validação dos resultados da modelagem de tópicos junto à comunidade científica. Do ponto de vista técnico, a pesquisa possibilitou identificar procedimentos como calibração de pesos por meio de *Regular Expression* e uso de bigramas e N-gramas para obter melhores resultados na modelagem de tópicos, bem como o modelo que apresentou melhores resultados.

O **problema da pesquisa** foi respondido ao identificar, por meio de metodologias diferentes entre pesquisas estabelecidas na literatura com as utilizadas nesta tese, bem como a diferenciação de *corpora* de dados analisados, que os documentos formais da comunicação científica da mesma área de atuação produzidos a partir da segunda década do século XXI apresentam comportamentos parcialmente diferentes aos do mapeamento científico das disciplinas e áreas apontadas por Pinheiro (2006) e que também aparecem como áreas e subáreas nos estudos de Zins e Santos (2015).

O objetivo geral da pesquisa foi atingindo ao verificar a proximidade e o distanciamento entre os temas adotados nas áreas e disciplinas da Ciência da Informação e estabelecidas na literatura por Pinheiro (2006) e que também aparecem em pesquisas Zins e Santos (2015) de maneira não ranqueável com os termos extraídos dos *corpora* de dados da pesquisa empírica realizada neste estudo. Mesmo se tratando de metodologias e tamanho diferentes de *corpora* de dados, é possível perceber que, dentre as 17 disciplinas estabelecidas por Pinheiro (2006), 11 ou 64,7% dos termos ainda estão em evidência quando comparados à pesquisa empírica. Faz-se necessário destacar a existência

mínima de 8 anos e máxima de 52 anos entre os documentos utilizados por Pinheiro (2006) e a amostragem desta pesquisa.

Torna-se possível perceber, entre o intervalo mínimo e máximo de anos dos resultados do referencial teórico e da pesquisa empírica, uma mudança no comportamento dos termos referentes às frequências e ao ranqueamento, destacando as disciplinas de Gestão da Informação, Bibliotecas Digitais/Virtuais, Gestão do Conhecimento, Economia da Informação e Inteligência Competitiva, que alcançaram média de crescimento de 9,4 posições ao longo dos anos. Outro fator relevante está na disciplina de Sistemas de Informação, que apresentou queda de uma única posição, mantendo-se como disciplina consolidada – conceito de Pinheiro (2006) – e ocupando a segunda posição no *ranking*.

Também foi possível identificar um distanciamento entre as disciplinas do referencial teórico com a pesquisa empírica. 7 ou 41,2% das disciplinas não apresentaram frequência entre os milésimos N-gramas (unigrama, bigrama, trigrama e lista geral) extraídos dos *corpora* de dados, sendo elas Bibliometria, Comunicação Científica Eletrônica, Formação e Aspectos Profissionais, Mineração de Dados, Necessidades e Usos de Informação, Processamento Automático da Linguagem e Teoria da Ciência da Informação, que passam a ser consideradas disciplinas de tendência da época, conceito também estabelecido por Pinheiro (2006). Registra-se que, se as disciplinas Comunicação Científica Eletrônica e Necessidades e Uso de Informação tivessem sido utilizadas como Comunicação Científica e Uso da Informação, estariam entre as disciplinas consolidadas. Nesse cenário, a disciplina de Uso da Informação manteria a proximidade entre as pesquisas.

O **primeiro objetivo específico** identificou, por meio de frequência extraída dos *corpora* de dados e representada através de gráficos dinâmicos, o comportamento de conjuntos de termos representados por N-gramas. Dentre os resultados, foi possível identificar pontos de comportamentos diferentes nos termos durante o intervalo analisado, sendo eles termos estáveis, termos instáveis, termos em ascensão, termos em descensão e combinações entre os comportamentos.

A diferença entre os tipos de documentos constituintes dos *corpora* de dados pode interferir nos comportamentos citados, uma vez que as pesquisas do tipo teses e dissertações podem levar, normalmente, entre 2 e 4 anos para

serem concluídas, enquanto os artigos completos e resumos expandidos são publicados anualmente e em maior volume, justificado pelos critérios de avaliação de cursos de pós-graduação estabelecidos pelo Ministério da Educação. O *corpora* de dados de teses e dissertações apresentou um maior volume de termos já estabelecidos na área da Ciência da Informação enquanto o *corpora* de dados constituído de artigos completos e resumos expandidos apresentou um maior número de termos em ascensão.

Com a necessidade de produção, organização, armazenamento, representação, disseminação, recuperação, acesso e o uso da informação, principalmente numa era tecnológica, tem surgido na produção científica brasileira – ainda que modestamente quando se comparados a termos estabilizados da Ciência da Informação – pesquisas em ascensão, com temas como *Big Data*, *Machine Learning*, Web Semântica, Computação nas Nuvens, Gestão de Dados e E-science.

Quanto ao **segundo objetivo específico**, o modelo *Latent Semantic Indexing* (LSI) - Indexação Semântica Latente que usa *Singular Value Decomposition* (SVD) - Decomposição de Valor Singular apresentou resultados significativos na modelagem de tópicos obtidos através da uma grande matriz de dados do tipo termo/documento em um processo de reducionalidade para uma matriz menor, mantendo informações sobre os dados e fornecendo associações desconhecidas entre palavras de maneira que possam ser induzidas em uma análise de como essas palavras co-ocorrem com a linguagem natural.

Entretanto, o modelo *Latent Dirichlet Allocation* (LDA) - Alocação de Dirichlet Latente – um modelo probabilístico generativo embasado em fundamentação estatística rigorosa e desenvolvido especificamente para análise em *corpora* de dados – apresentou melhores resultados, tanto para os termos quanto para os pesos de cada tópico de maneira que apresentam características no qual permitem ser classificados por indexadores da área de conhecimento com menor esforço cognitivo. Além disso, o modelo LDA permite realizar uma série de configurações, como número de documentos a serem utilizados por blocos de treinamento, número de passagens de treinamento pelos documentos e número máximo de iterações por *corpus* de dados que, por meio do *Machine Learning,* apresentou os melhores resultados, mesmo não sendo desenvolvido quaisquer códigos que avaliasse o desempenho dos dois modelos.

A classificação criada, referente às características dos termos e pesos, contribui para uma melhor interpretação dos resultados ao serem analisados por um profissional especialista no domínio da linguagem estudada, considerando os termos generalistas, especialistas e chave, além dos pesos fortes e fracos que constituem os tópicos.

Ainda no segundo objetivo específico, foi realizada a validação da seleção dos melhores resultados extraídos por meio do melhor modelo de extração de tópicos, sendo realizada por um grupo constituído por 87 profissionais formado por professores, bibliotecários e consultores atuantes no meio acadêmico e no mercado profissional que representam a comunidade científica brasileira da área da Ciência da Informação. Os resultados da validação entre os tópicos constituídos por termos e pesos e os possíveis rótulos que melhor representam os conjuntos apresentaram médias entre os *corpora* de dados de 76,7% de assertividade, 6,7% para tópicos que podem ser vinculados em dois rótulos e 17,7% que foram vinculados a rótulos diferentes aos que realmente os documentos disponibilizados para consulta representavam.

Os tópicos vinculados a rótulos diferentes aos dos documentos que melhor representam os conjuntos de termos fazem parte dos *corpora* de dados constituído por documentos do tipo artigo completo e resumo expandido extraídos do ENANCIB, bem como os nomes dos rótulos utilizados para validação dos resultados – os mesmos dos GTs do evento. Entretanto, ficou notória a dificuldade por parte dos respondentes ao categorizar os tópicos extraídos de documentos científicos do ENANCIB com os próprios GTs do evento. Constam entre os resultados que apresentaram rótulos diferentes aos dos documentos um equilíbrio nos percentuais que vão de 2 até 5 rótulos.

Independente de práticas de análise de assunto apresentada no referencial teórico ou documentos disponibilizados para consulta – o que poderia reduzir o percentual de 17,7% de tópicos vinculados a rótulos diferentes e aumentar o percentual de assertividade – faz-se necessário refletir sobre a pluralidade dos GTs do ENANCIB enquanto norteadores para as pesquisas na área da Ciência da Informação, uma vez que é possível identificar uma miscelânea de disciplinas em Ciência da Informação, Subdisciplinas – Quadro 1, Subáreas – Quadro 2, por Pinheiro (2006), Áreas e subáreas e conteúdo do

campo da Ciência da Informação – Quadro 5, por Zins e Santos (2015), com os ementários dos GTs.

Pôde-se alcançar o **pressuposto** da pesquisa no que diz respeito à morosidade e incerteza na qualidade dos resultados, levando em consideração o fator subjetividade do profissional ao analisar grandes *corpora* de dados. O uso das tecnologias, a partir de algoritmos estatísticos de extração de tópicos que utilizam métodos de *Machine Learning*, apresentaram dinamismo com relação aos fatores tempo e qualidade dos resultados. O modelo LSI realizou a modelagem de tópicos em 1 hora, 5 minutos e 18 segundos, enquanto o modelo LDA levou 5 horas, 36 minutos e 54 segundos para identificar os tópicos mais representativos de 4.068 documentos distribuídos em 2 *corpora* de dados – ou 13 *corpus* de dados quando se analisados separadamente. Já a qualidade dos tópicos do melhor modelo foi validada por professores, pesquisadores e bibliotecários atuantes na área da Ciência da Informação com média de assertividade entre os *corpora* de dados de 76,7%, levando em consideração a dificuldade de representatividade dos rótulos que são os mesmos dos GTs do ENANCIB utilizado no questionário, bem como o baixo percentual de média de 12% na utilização de técnicas de análise de assuntos em documentos que melhor representam os tópicos.

A **hipótese** da pesquisa foi comprovada ao identificar que o algoritmo de *Machine Learning* LDA utilizado na realização da Modelagem de Tópicos nos *corpora* de dados da amostragem da pesquisa empírica apresentou melhores resultados que o modelo LSI. Considera-se, para isso, a qualidade dos termos e o equilíbrio nos pesos dos conjuntos de tópicos extraídos por meio de modelo LDA. Destaque referente à qualidade dos tópicos está no maior número de variedade de termos N-gramas, sendo unigramas, bigramas e trigramas, que permite ao indexador inferir rótulos de maneira mais assertiva ao analisar um tópico. Isso acontece porque o modelo LDA permite decompor os *corpora* de dados em temas constituintes, reduzindo os efeitos adversos gerados pela sinonímia e polissemia por meio da identificação de associações estatísticas entre os termos.

A **limitação da pesquisa** está relacionada aos procedimentos técnicos, bem como aos equipamentos para processamento dos algoritmos, uma vez que a ideia inicial seria processar dois *corpora* de dados. Entretanto, por insuficiência

de memória em três equipamentos testados, foi necessário fragmentar os *corpora* de dados em 13 *corpus* de dados, o que potencializou em um quantitativo exorbitante de volume de resultados a serem analisados. Foi cogitada a possibilidade de reduzir o intervalor entre os anos aos quais os documentos foram analisados, entretanto, essa decisão interferiria na representatividade de um dos objetivos que apresentou o comportamento dos termos mais relevantes ao longo do período analisado. O fator processamento também interferiu na visualização dinâmica dos tópicos, uma vez que acabou por limitar o número ideal de tópicos a ser gerado mediante ao tamanho do *corpus* de dados.

Quanto às **contribuições da pesquisa** para a área da Ciência da Informação estão o norteamento para áreas de pesquisa em ascensão com viés para o uso das tecnologias identificada através de frequências nos *corpora* de dados, podendo ser utilizadas pelos programas para elaboração de projetos com contribuições práticas, metodológicas e científicas de ensino, pesquisa e extensão.

Também cabe uma reflexão sobre a estruturação dos GTs do ENANCIB, justamente por ser o principal evento da área da Ciência da Informação e apresentar um número representativo de pesquisas publicadas anualmente, entretanto, constam entre os ementários dos GTs uma mistura de conceitos que se divergem das estruturas encontradas na literatura. Por fim, os resultados da modelagem de tópicos apresentaram um alto percentual de coesão com os rótulos apresentados, sendo indicativo para aprofundamento nos estudos relacionados à indexação automática de conteúdo. Num viés tecnológico, pôde-se constatar que o uso de bigramas ou trigramas apresentam melhores resultados que contribuem para uma análise de assuntos mais assertiva, já que termos com essas características são mais representativos e menos generalistas em relação aos unigramas.

Sugere-se, para **pesquisas futuras**, ainda em viés tecnológico, que sejam utilizados modelos diferentes ao *bag-of-words* (saco de palavras), usado nesta pesquisa para comparação de resultados e desempenho. Essa análise comparativa poderia ser realizada utilizando os mesmos *corpora* de dados entre os modelos: i) bag-of-words – refere-se a um conjunto de palavras que desconsidera a gramática e ordem das palavras de um *corpus* de documento,

mas mantendo a multiplicidade; ii) word2vec - produz espaço vetorial de modo que as palavras que compartilham contextos comuns num *corpus* sejam localizadas próximas umas das outras no espaço; e iii) lda2vec - constrói representações de documentos sobre a incorporação de palavras.

Já em um viés acadêmico, sugere-se a comparação por meio da mesma metodologia entre a produção de pesquisas nacionais em Ciência da Informação com o que é produzido no mundo, como por exemplo, utilizando *corpus* de dados constituídos de documentos científicos da International *Society for Knowledge Organization* (ISKO) - Sociedade Internacional para Organização do Conhecimento ou realizando o mapeamento científico dos cursos de uma instituição de ensino superior. Além disso, os resultados extraídos dos *corpora* de dados podem possibilitar uma atualização do estudo de Pinheiro (2006) e constatação de Zins e Pinheiro (2015) com áreas, subáreas da Ciência da Informação.

Sobre as **considerações do autor**, a modelagem de tópicos apresentou eficiência nos resultados de extração de tópicos, comprovadas por meio do alto percentual de assertividade alcançado na validação dos resultados por profissionais que atuam no domínio da linguagem dos *corpora* de dados estudados. Embora seja uma área de estudos relativamente nova e promissora como instrumentalização na organização de informações, ainda apresenta lacunas a serem estudadas/preenchidas e, consequentemente, implementadas como novas soluções tecnológicas. Enquanto mapeamento científico da área da Ciência da Informação, espera-se para a próxima década um maior quantitativo de termos com viés tecnológico em publicações científicas mediante o surgimento de novos cursos de ensino superior, como Ciência de Dados, que estão intrinsecamente associados à Ciência da Informação.

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. New York: Springer Science & Business Media, 2012.

ALBRECHTSEN, H. Subject analysis and indexing: from automated indexing to domain analysis. **The Indexer**, Liverpool, v. 18, n. 4, p. 219–224, 1993.

ALMEIDA, M. B. Revisiting Ontologies: a necessary clarification. **Journal of the American Society for Information Science and Technology**, New York, v. 64, n. 8, p. 1682-1693, 2013.

ALMEIDA, M. B.; DE OLIVEIRA, V. N. P.; COELHO, K. C. Estudo exploratório sobre ontologias aplicadas a modelos de sistemas de informação: perspectivas de pesquisa em Ciência da Informação. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 15, n. 30, p. 32-56, 2010.

ALPAYDIN, E. **Introduction to machine learning**. 2. ed. Cambridge: The MIT Press, 2010.

AYODELE, T. O. Types of Machine Learning Algorithms. *In*: ZHANG, Y. (org.) **New Advances in Machine Learning**. London: Intechopen, 2010. p. 19-48.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**: Métodos para análise de documentos - determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992.

AZEREDO, J.C. **Gramática Houaiss da língua portuguesa**. 3. ed. São Paulo: PubliFolha, 2008.

BRASCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento?. *In*: Encontro Nacional de Pesquisa em Ciência da Informação, 9., 2008, São Paulo - SP. ***Anais***... Salvador: USP, 2008.

BARRETO, A. A. Uma história da Ciência da Informação. *In*: TOUTAIN, L. M. B. (org.). **Para entender a Ciência da Informação**. Salvador: EDUFBA, 2007. p. 13-34.

BARRETO, A. A. Uma quase história da ciência da informação. **DataGramaZero - Revista de Ciência Da Informação**, Rio de Janeiro, v. 9, n. 2, p. 1–17, 2008.

BERRY, M. W.; DUMAIS, S. T.; O'BRIEN, G. W. Using linear algebra for intelligent information retrieval. **Society for Industrial and Applied Mathematics**, Philadelphia, v. 37, n. 4, p. 573–595, 1995.

BERRY, M. W.; KOGAN. J. **Text Mining Applications and Theory**. West Sussex: John Wiley & Sons, 2010.

BEGHTOL, C. Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. **Journal of Documentation**, London, v. 42, n. 2, p. 84-113, 1986.

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988.

BICALHO, L. M. **As relações interdisciplinares refletidas na literatura brasileira da Ciência da Informação**. 2009. 268 f. Tese (Doutorado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte - MG, 2009.

BJÖRK, B. C. Open access to scientific publications-an analysis of the barriers to change? **Information Research**, Borås, v. 9, n. 2, 2004.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, New York, v. 55, n. 4, p. 77–84, 2012.

BLEI, D.; CARIN, L.; DUNSON, D. Probabilistic topic models. I**EEE Signal Processing Magazine**, New York, v. 27, n. 6, p.55–65, 2010.

BLEI, D. M.; LAFFERTY, J. D. A correlated topic model of science. **The Annals of Applied Statistics**, Cleveland, v. 1, n. 1, p. 17-35, 2007.

BLEI, D. M.; LAFFERTY, J. D. Topic models. *In*: SRIVASTAVA, A. N.; SAHAMI, M. (org.). **Text mining:** Classification, clustering, and applications. Minneapolis: Chapman & Hall/CRC, 2009. p. 71-94.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, Cambridge, v. 3, n. jan, p. 993-1022, 2003.

BLUNSOM, P. Hidden markov models. **Lecture Notes**, Hoboken, v. 15, n. 18-19, p. 1-7, 2004.

BOAI (Budapest Open Access Initiative). **Dez anos da Iniciativa de Budapeste em Acesso Aberto**: a abertura como caminho a seguir. Budapeste – HU, 2012. Disponível em: http://www.budapestopenaccessinitiative.org/boai-10-translations/portuguesebrazilian-translation. Acesso em: 5 nov. 2019.

BORKO, H. Information science: what is it? **American Documentation**, Washington, v.19, n.1, p.3-5, 1968.

BRAMBILLA, S. D. S.; STUMPF, I. R. C. Interfaces da Informação: Tendências Temáticas da Pós-Graduação. *In*: Encontro Nacional de Pesquisa em Ciência da Informação, viii., 2007, Salvador - BA. ***Anais***... Salvador: UFBA, 2007.

BROOKES, B. C. The foundations of Information Science: part I - philosophical aspects. **Journal of Information Science**, New York, v. 2, n. 3-4, p. 125-133, 1980.

CALTABIANO, M.; VERZI, P.; SCAPPUZZO, S. G. Head posture in orthodontics: physiopathology and clinical aspects 2. **Mondo ortodontico**, Milano, v. 14, n. 3, p. 313-324, 1989.

CAMPELLO, B. S. Encontros científicos. *In*: CAMPELLO, B. S.; CENDÓN, B. V.; KRENMER J. M. (org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora UFMG, 2007a, p. 55-71.

CAMPELLO, B. S. Pesquisa em andamento. *In*: CAMPELLO, B. S.; CENDÓN, B. V.; KRENMER J. M. (org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte – MG: Editora UFMG, 2007b, p. 49-54.

CAMPELLO, B. S. Teses e dissertações. *In*: CAMPELLO, B. S.; CENDÓN, B. V.; KRENMER J. M. (org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora UFMG, 2007c, p. 121-128.

CAPURRO, R. Epistemologia e Ciência da Informação. *In*: Encontro Nacional de Pesquisa em Ciência da Informação, v., 2003, Belo Horizonte - MG. ***Anais***... Belo Horizonte: UFMG, 2003.

CAPURRO, R.; Hjorland, B. O conceito de informação. **Perspectivas em Ciência Da Informação**, Belo Horizonte, v. 12, n. 1, p. 148-207, 2007.

CARVALHO, M. C. M. **Construindo o saber**: metodologia científica: fundamentos e técnicas. 22. ed. Campinas: Papirus, 2010.

CASTRO, C. A. **História da biblioteconomia brasileira**: perspectiva histórica. Brasília: Thesaurus, 2000.

CESARINO, M. A. N.; PINTO, M. C. M. F. Análise de assunto. **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 1, p. 32-43, 1980.

CHANEY, A. J. B.; BLEI, D. M. Visualizing topic models. *In*: **Sixth international AAAI conference on weblogs and social media**, p. 419-422, 2012.

CHANG, J. et al. Reading Tea Leaves: How humans interpret topic models. *In*: **Advances in Neural Information Processing Systems**, p. 288-296. 2009.

CHENG, X. et al. Coupled term-term relation analysis for document clustering. *In*: **The 2013 International Joint Conference on Neural Networks**, IEEE, p. 1-8. 2013.

CHU, C. M.; O'BRIEN, A. Subject analysis: the critical first estage in indexing. **Journal of Information Science**, New York, v. 19, n. 6, p. 439-454, 1993.

CHUANG, J.; MANNING, C. D.; HEER, J. Termite: Visualization Techniques for Assessing Textual Topic Models. *In*: **Proceedings of the International Working Conference on Advanced Visual Interfaces**, p. 74-77, 2012.

CONRAD, S.; BIBER, D. **Variation in English:** multi-dimensional studies. New York: Longman, 2001.

CRESWELL, J. W. **Projeto de pesquisa**: Métodos qualitativos, quantitativos e misto. 3. ed. Porto Alegre: Artmed, 2010.

CUNHA, M. B. **Para saber mais**: fontes de informação em ciência e tecnologia. Brasília: Briquet de Lemos, 2001.

OLIVEIRA, J. V; PEDRYCZ, W. (org.). **Advances in fuzzy clustering and its applications**. West Sussex: John Wiley & Sons, 2007. 454p.

DEERWESTER, S. *et al*. Improving information retrieval with latent semantic indexing. *In*: AMERICAN SOCIETY FOR INFORMATION SCIENCE ANNUAL MEETING, 51., 1988, Atlanta. **Proceedings**... Medford: Information Today, 1988. p. 36-40.

DEERWESTER, S. *et al*. Indexing by latent semantic analysis. **Journal of the american society for information science**, New York, v. 41, n. 6, p. 391-407, 1990.

DEMO, P. **Metodologia Científica em Ciências Sociais**. 3. ed. São Paulo: Atlas, 1995.

DIAS, E. W.; NAVES, M. M. L. **Análise de Assunto Teoria e Prática**: Estudos Avançados em Ciência da Informação. Brasília: Thesaurus, 2007.

DUMAIS, S. T. Latent semantic indexing (lsi): Trec-3 report. *In* Proceedings of the Text REtrieval Conference (TREC-3), 1995. p. 219-230.

FAIRTHORNE, R.A. Content analysis, specification, and control. **Annual Review of Information Science and Technology**, v. 4, p. 73-109, 1969.

FONSECA, M. S. **Produtividade e impacto de pesquisadores brasileiros em Ciência da Informação**: análise dos autores do ENANCIB 2013. 2015. 239 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis, 2015.

FOSKETT, A. C. **A abordagem temática da informação**. São Paulo: Polígono, 1973.

FUJITA, M. S. L. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas – SP, v. 1, n. 1, p. 60–90, 2003.

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010.

GIL, A. C. **Métodos e técnicas de Pesquisa Social**. 6. ed. Campos Elísios: Atlas, 2016.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. *In*: XXIII Congresso da Sociedade Brasileira de Computação, 2003. p. 347-395.

GRANGER, S. **Learner english on computer**. London: Longman, 1998.

GREENE, D.; O'CALLAGHAN, D; CUNNINGHAM. How many topics? Stability analysis for Topic models. *In*: CALDERS, T *et al*. (org). **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. Berlin: Springer, 2014. p. 498-513.

GRUS, J. **Data Science do zero**: primeiras regras com Python. Rio de Janeiro: Alta Books, 2016.

HARMON, G. On the evolution of Information Science. **Journal of the American Society for Information Science and Technology**, New York, v.22, n.4, p.235-241, 1971.

HIRSCH, J. E. An index to quantify an individual's scientific research output. *In*: **Proceedings of the National academy of Sciences**, v. 102, n. 46, p. 16569-16572, 2005.

HOFMANN, T. Probabilistic Latent Semantic Analysis. *In*: **Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence**, p. 289-296, 1999a.

HOFMANN, T. Probabilistic Latent Semantic Indexing. *In*: **Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval**, p. 50-57, 1999b.

KASZUBOWSKI, E. **Modelo de tópicos para associações livres**. 2016. 227 f. Tese (Doutorado em Psicologia) – Centro de Filosofia e Ciências Humanas, Universidade Federal de Santa Catarina, Florianópolis, 2016.

KENNEDY, G. **An Introduction to corpus linguistics**. London: Longman, 1998.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos Livros, 2004.

LANDAUER, T. K.; DUMAIS, S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. **Psychological Review**, Washington, v. 104, n. 2, p. 211-240, 1997.

LANGRIDGE, D. **Classificação**: abordagem para estudantes de biblioteconomia. Rio de Janeiro: Interciência, 1997.

LARA, M. L. G.; CONTI, V. L. Disseminação da informação e usuários. **São Paulo em Perspectiva**, São Paulo, v. 17, n. 3-4, p. 26-34, 2003.

LAU, J. H. et al. Automatic labeling of topic models. *In*: **Proceedings of the 49th annual meeting of the association for computational linguistics**, p. 1536-1545, 2011.

LE COADIC, Y.F. **A ciência da informação**. 2. ed. Brasília: Briquet de Lemos Livros, 2004.

LEMOS, D. L. S.; SOUZA, R. R. Organização de recursos bibliográficos e multimídia na web: contribuições interdisciplinares, **Informação & Informação**, Londrina, v. 23, n. 2, p. 98-126, 2018.

MARQUESONE, R. **Big data**: técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016.

MANNING, C.D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2009.

MARTIN, D. I.; BERRY, M. W. Mathematical Foundations Behind Latent Semantic Analysis. In Handbook of latent semantic analysis. *In*: LANDAUER, T.

K. et al. **Handbook of latent semantic analysis**. New York: Routledge, 2011, p. 35–56.

MCKINNEY, W. **Python para análise de dados**: tratamento de dados com pandas, numpy e ipython. São Paulo: Novatec, 2018.

MICHEL, M. H. **Metodologia e pesquisa científica em ciências sociais**: um guia prático para acompanhamento da disciplina e elaboração de trabalhos monográficos. 3. ed. São Paulo: Atlas, 2015.

MIKHAILOV, A. I.; GILYAREVSKIJ, R.G. **An Introductory course on informatics/documentation**. Programme and meeting document. 1970.

MIRANDA, A. A Ciência da Informação e a teoria do conhecimento objetivo: um relacionamento necessário. *In*: AQUINO, M. A. (org). **O campo da Ciência da Informação:** gênese, conexões e especificidades. João Pessoa: Universitária/UFPB, 2002, p. 9-24.

MIRANDA, D. B.; PEREIRA, M. N. F. O periódico científico como veículo de comunicação: uma revisão de literatura. **Ciência da Informação**, Brasília, v. 25, n. 3, p. 375-382, 1996.

MUELLER, S. P. M. A comunicação científica e o movimento de acesso livre ao conhecimento. **Ciência da Informação**, Brasília, v. 35, n. 2, p. 27-38, 2006.

MUELLER, S. P. M. A ciência, o sistema de comunicação e a literatura científica. *In*: CAMPELLO, B. S.; CENDÓN, B. V.; KRENMER J. M. (org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora UFMG, 2007a, p. 21-34.

MUELLER, S. P. M. O periódico científico. *In*: CAMPELLO, B. S.; CENDÓN, B. V.; KRENMER J. M. (org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora UFMG, 2007b, p. 73-95.

NAVES, M. M. L. Análise de assunto: Concepções. **Revista de Biblioteconomia de Brasília**, Brasília, v. 20, n. 2, p. 215-226, 1996.

NAVES, M. M. L. **Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores**. Tese (Tese em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2000. Universidade Federal de Minas Gerais.

NHACUONGUE, J. A.; FERNEDA, E. O campo da ciência da informação: contribuições, desafios e perspectivas. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 20, n. 2, p. 3-18, 2015.

NOLASCO, D. **Identificação automática de áreas de pesquisa em C&T**. 2016. 195 f. Dissertação (Mestrado em Informática) – Departamento de Ciência da Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

NOLASCO, D.; OLIVEIRA, J. Detecting knowledge innovation through automatic topic labeling on scholar data. *In*: **2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE**, p. 358-367, 2016a.

NOLASCO, D.; OLIVEIRA, J. Modelagem de tópicos e criação de rótulos: identificando temas em dados semi-estruturados e não-estruturados. *In*: OGASAWARE, E.; VIEIRA, V. (org.). **Tópicos em gerenciamento de dados e informação**. Salvador: Sociedade Brasileira da Computação, 2016b, p. 87-112.

ODDONE, N. E. **Atividade editorial & Ciência da Informação**: convergência epistemológica. 1998. 266 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Faculdade de Estudos Sociais Aplicados, Universidade de Brasília, Brasília, 1998.

OLIVEIRA, E. B. P. M. Periódicos científicos eletrônicos: definições e histórico. **Informação & Sociedade**, João Pessoa, v. 18, n. 2, p. 69-77, 2008.

OLIVEIRA, E. F. T.; GRACIO, M. C. C. Indicadores bibliométricos em ciência da informação: análise dos pesquisadores mais produtivos no tema estudos métricos na base Scopus. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 16, n. 4, p. 16-28, 2011.

OLIVEIRA, L. P. Linguística de Corpus: teoria, interfaces e aplicações. **Matraga - Estudos Linguísticos e Literários**, Rio de Janeiro, v. 16, n. 24, p. 48–76, 2009.

OLIVEIRA, L. P.; DIAS, M. C. P. Compilação de corpus: representatividade e o CORPOBRAS. **Calidoscópio**. São Leopoldo, v. 7, n. 3, p. 192-198, 2009.

OLIVEIRA, M. **A investigação científica na ciência da informação**: análise da pesquisa financiada pelo CNPq. 1998. 218 f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 1998.

OLIVEIRA, M. Origens e evolução da ciência da informação. *In*: OLIVEIRA, M. (org.). **Ciência da informação e biblioteconomia**: novos conteúdos e espaços de atuação. 2. ed. Belo Horizonte: Editora UFMG, 2011. p. 9-28.

OTHERO, G. Á. Lingüística Computacional: uma breve introdução. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 341–351, 2006.

PAPADIMITRIOU, C. H. et al. Latent semantic indexing: A probabilistic analysis. *In*: **Proceedings of the 1998 17th ACM SIGART-SIGMOD-SIGART Symposium on Principles of Database Systems**, p. 159-168, 1998.

PINHEIRO, L. V. R. **A Ciência da Informação entre sombra e luz: domínio epistemológico e campo interdisciplinar**. 1997. 278 f. Tese (Doutorado em Comunicação) – Centro de Filosofia e Ciências Humanas, Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1997.

PINHEIRO, L. V. R. Gênese da Ciência da Informação ou sinais anunciadores da nova área. *In*: AQUINO, M. A. (org). **O campo da Ciência da Informação**: gênese, conexões e especificidades. João Pessoa: UFPB, 2002. p. 61-86.

PINHEIRO, L. V. R. Ciência da Informação: desdobramentos disciplinares interdisciplinaridade e transdisciplinaridade. *In*: GOMEZ, M. N. G.; ORRICO, E. G, D. (Org.). **Políticas de memória e informação**: reflexos na organização do conhecimento. Natal: EDUFRN, 2006. p. 111-141.

PINTO, A. L.; MATIAS, M.; GONZÁLEZ, J. A.M. Produção brasileira da Ciência da Informação na Web of Science entre 1994 e 2013 e a lista Qualis/Capes da Área. **Ibersid: revista de sistemas de información y documentación**. Saragoça, v. 10, n. 1, p. 51-61, 2016.

PUJARA, J.; SKOMOROCH, P. Large-Scale Hierarchical Topic Models. *In*: **NIPS Workshop on Big Learning**, p. 1–8, 2012.

PUSTEJOVSKY, J.; STUBBS, A. **Natural Language Annotation for Machine Learning**: a guide to corpus-building for applications. Champaign: O'Reilly Media, 2012.

QUEIROZ, D. G. D. C.; MOURA, A. M. M. Ciência da Informação: história, conceito e características. **Em Questão**. Porto Alegre, v. 21, n. 3, p. 26-42, 2015.

RAMAGE, D.; MANNING, C. D.; DUMAIS, S. Partially labeled topic models for interpretable text mining. *In*: **Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 457-465, 2011.

RAYWARD, W. B. Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext. **Journal of the American Society for Information Science**, New York, v. 45, n. 4, p. 235-250, 1994.

RICHARDSON, R. J. **Pesquisa Social**: Métodos e Técnicas. 3. ed. São Paulo: Atlas, 2010.

ROWLEY, J. **A biblioteca eletrônica**. 2. ed. Brasília: Briquet de Lemos, 2002.

RUSSO, M. **Fundamentos de biblioteconomia e Ciência da Informação**. Rio de Janeiro: E-papers, 2010.

SAMPAIO, M. I. C.; SABADINI, A. A. Z. P. Indexação e fator de impacto. *In*: SAMPAIO, M. I. C.; SABADINI, A. A. Z. P.; KOLLER, S. H. (org). **Publicar em Psicologia:** um enfoque para a revista científica. São Paulo: Associação Brasileira de Editores Científicos de Psicologia / Instituto de Psicologia da Universidade de São Paulo, 2009. p. 109-121.

SANTOS, F. F. **Extração de tópicos baseado em agrupamento de regras de associação**. 2015. 129 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo, São Carlos, 2015.

SANTOS, G. C. **Fontes de indexação para periódicos científicos**: um guia para bibliotecários e editores. Campinas: E-color Editora, 2011.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**. Belo Horizonte, v. 1, n. 1, p. 41–62, 1996.

SARDINHA, T. B. Processamento Computacional do Português. **Simpósio, 9o. InPLA, PUCSP**, São Paulo, 1999.

SARDINHA, T. B. Lingüística de corpus: histórico e problemática. **Delta: documentação de estudos em lingüística teórica e aplicada**, São Paulo, v. 16, n. 2, p. 323-367, 2000a.

SARDINHA, T. B. O que é um corpus representativo? **Direct Papers**, São Paulo, v. 44, 2000b.

SARDINHA, T. B. **Lingüística de corpus**. Barueri: Manole LTDA, 2004.

SVARTVIK, J. Corpora are becoming mainstream. *In*: THOMAS, J.; SHORT, M. (org). **Using corpora for language research**. London: Longman, 1996. p. 3-13.

SAVOLAINEN, Reijo. The sense-making theory: An alternative to intermediary-centered approaches. *In*: library and information science. Conceptions of Library and Information Science, Taylor Graham, London, p. 149-64, 1992.

SCARPA, A. D. **Técnicas de Processamento de Linguagem Natural Aplicadas às Ciências Sociais**. 2017. 86 f. Dissertação (Mestrado em Matemática Aplicada) – Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, 2017.

SEVERINO, A. J. **Metodologia do trabalho científico**. 21. ed. São Paulo: Cortez, 2000.

SEVERINO, A. J. **Metodologia do trabalho científico**. 24. ed. São Paulo: Cortez, 2016.

SHERA, J. H.; CLEVELAND, D. B. History and foundations of Information Science. **Annual Review of Information Science and Technology**, New York, v. 12, p. 249-275, 1977.

SIEVERT, C.; SHIRLEY, K. E. LDAvis: A method for visualizing and interpreting topics. *In*: **Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces**, p. 63-70, 2014.

SOUSA, B. P.; FUJITA, M. S. L. Análise de assunto no processo de indexação: um percurso entre teoria e norma. **Informação & Sociedade**, João Pessoa, v. 24, n. 1, p. 19–34, 2014.

SOUZA, M. P. N. Abordagem inter e transdiciplinar em ciência da informação. *In* TOUTAIN, L. M. B. B. (org.). **Para entender a ciência da informação**. Salvador: EDUFBA, 2007, p. 75-90.

SOUZA, M.; JÚNIOR, A. I.; SOUZA, R. R. Modelagem de tópicos: mapeamento científico do gt-8 do enancib. *In*: Encontro Nacional de Pesquisa em Ciência da Informação, 20., 2007, Florianópolis - SC. ***Anais***... Florianópolis: UFSC, 2019.

SOUZA, M.; SOUZA, R. R. Modelagem de tópicos: resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina. **Múltiplos Olhares em Ciência da Informação**, Belo Horizonte, v. 9, n. 2, p. 1–11, 2019.

SOUZA, R. R.; ALMEIDA, M. B.; BARACHO, R. M. A. Ciência da informação em transformação: big data, nuvens, redes sociais e web semântica. **Ciência da Informação**, Brasília, v. 42, n. 2, p. 159–173, 2015.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. *In*: LANDAUER, T. K. et al. **Handbook of latent semantic analysis**. Mahwah: Lawrence Erlbaum Associates, 2007, p. 424-440.

TÁLAMO, M. F. G. M. **Elaboração de resumos**. São Paulo: Escola de Comunicação e Artes, 1987.

TEH, Y. W. *et al*. Hierarchical dirichlet processes. **Journal of the American Statistical Association**, New York, v. 101, n. 476, p. 1566-1581, 2006.

TODD, R. T. Academic indexing: what`s it all about? **The Indexer**, London, v. 18, n. 2, p. 101-104, 1992.

VICKERY, B. C. **Classificação e indexação nas ciências**. Rio de Janeiro: BNG/BRASILART, 1980.

VIEIRA, R. Lingüística computacional: fazendo uso do conhecimento da língua. **Entrelinhas**, São Leopoldo, ano 2, n. 4, p. 20–25, 2002.

VIEIRA, R.; LIMA, V. L. S. Lingüística computacional: princípios e aplicações. *In*: **Congresso da Sociedade Brasileira de Computação, I Jornada de Atualização em Inteligência Artificial**, xxi, 2001. p. 47-86.

WALLACH, H. M. Topic modeling: beyond bag-of-words. *In*: **Proceedings of the 23rd international conference on machine learning**, p. 977-984, 2006.

WANG, X.; MCCALLUM, A.; WEI, X. Topical N-grams: Phrase and topic discovery, with an application to information retrieval. *In*: **Seventh IEEE international conference on data mining (ICDM 2007)**, p. 697-702, 2007.

WERSIG, G.; NEVELING, U. The phenomena of interest to Information Science. **The information scientist**, London, v. 9, n. 4, p. 127-140, 1975.

WERSIG, G. Information science: the study of postmodern knowledge usage. **Information processing & management**, New York, v. 29, n. 2, p. 229-239, 1993.

WITTER, D. I.; BERRY, M. W. Downdating the Latent Semantic Indexing Model for Conceptual Information Retrieval. **The Computer Journal**, Manchester, v. 41, n. 8, p. 589-601, 1998.

WORLD INFORMATION SYSTEM FOR SCIENCE AND TECHNOLOGY. Princípios de indexação. **Revista Da Escola de Biblioteconomia Da UFMG**, Belo Horizonte, v.10, n. 1, p. 83–94, 1981.

YANG, Yi et al. Active learning with constrained topic model. *In*: **Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces**. p. 30-33, 2014.

ZHU, D. et al. Intuitive Topic Discovery by Incorporating Word-Pair's Connection Into LDA. *In*: **2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology**, p. 303-310, 2012.

ZINS, C. **knowledge map of information science: issues, principles, implications**. Jerusalem, 2005.

ZINS, C. Conceptions of informations science. **Journal of the American Society for Information Science and Technology**, New York, v. 58, n. 3, p. 335-350, 2007.

ZINS, C,; Santos, P. L. V. A. C. Brazilian model of library and information studies in the bachelor's level. **Informação & Sociedade**, João Pessoa, v. 25, n. 3, p. 185-203, 2015.

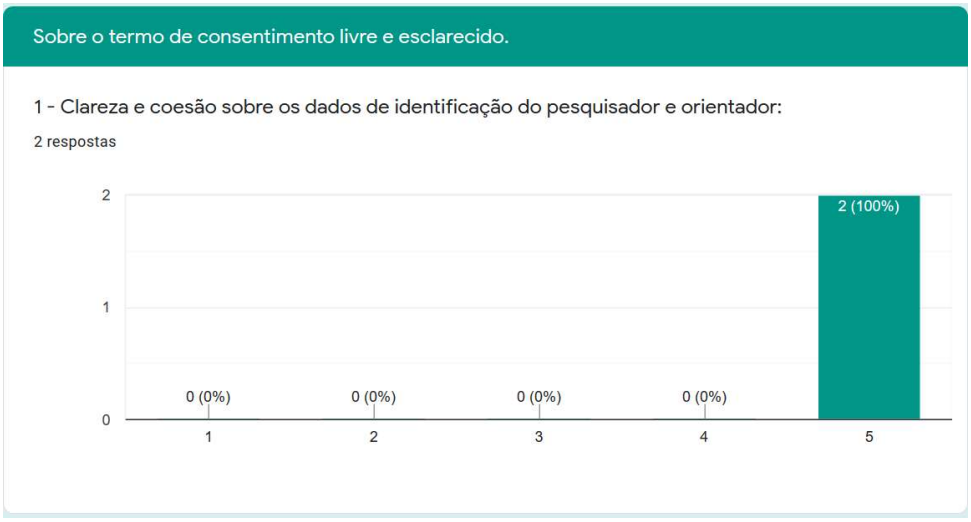
ZIMAN, John. **Conhecimento público**. Belo Horizonte: Itatiaia, 1979.

APÊNDICE A - Avaliação do pré-teste

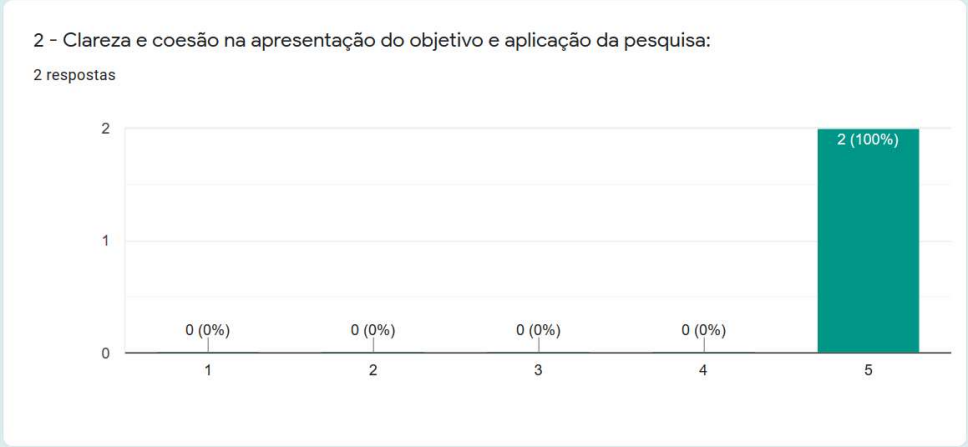
Perfil avaliador 1: Professora universitária com doutorado e mestrado em Cognição e Linguagem pela Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF). Especialista em Planejamento, Implementação e Gestão da Educação a Distância pela Universidade Federal Fluminense (UFF). Licenciada em Letras - Língua Portuguesa e Literaturas de Língua Portuguesa pela Universidade Federal de Viçosa (UFV) e em Pedagogia pelo Centro Universitário Claretiano (Polo Campos dos Goytacazes/RJ).

Perfil avaliador 2: Professor universitário com doutorado em Ciência da Informação (Ciência da Comnicação) pela Universidade de São Paulo (USP), mestrado em *Organization & Management* pela *Central Connecticut State University* (CCSU). Bacharel em Direito pelo Centro Universitário de João Pessoa (UNIPE). Graduado em Ciência da Computação pela Universidade Federal da Paraíba (UFPB).

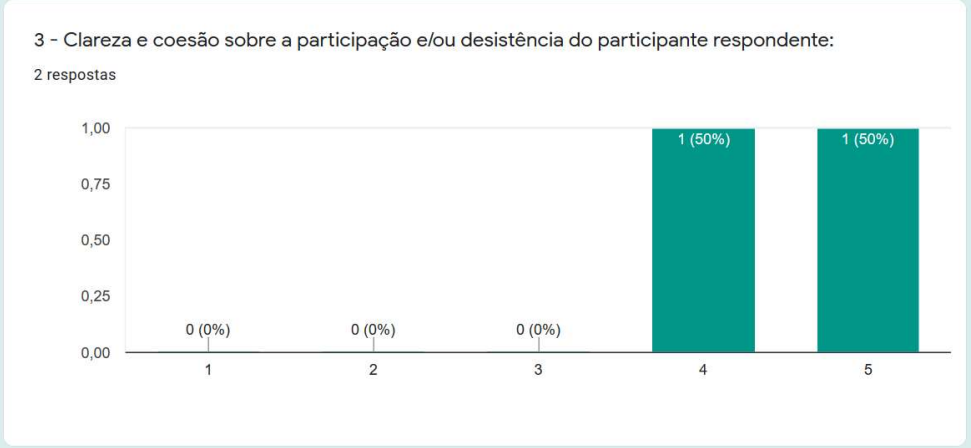
- Sobre claresza e coesão referente ao termo de consentimento livre e esclarecido.



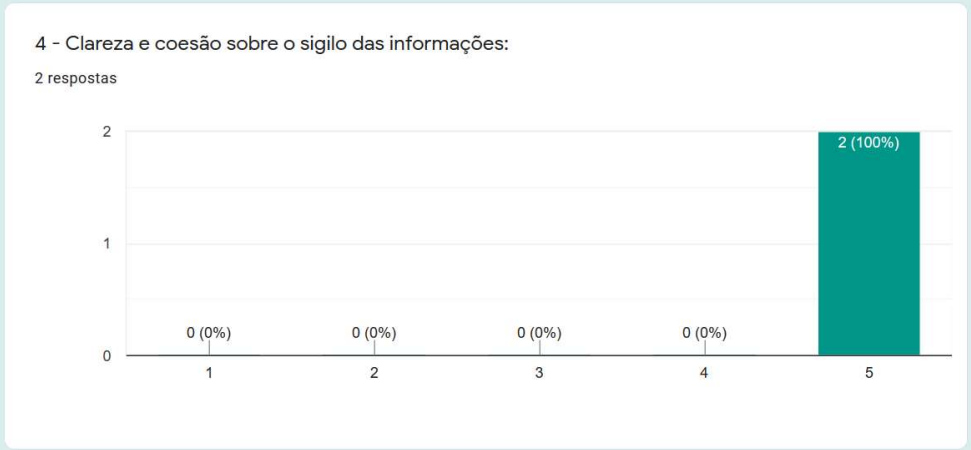
- Sobre clareza e coesão referente a apresentação do objetivo e aplicação da pesquisa.



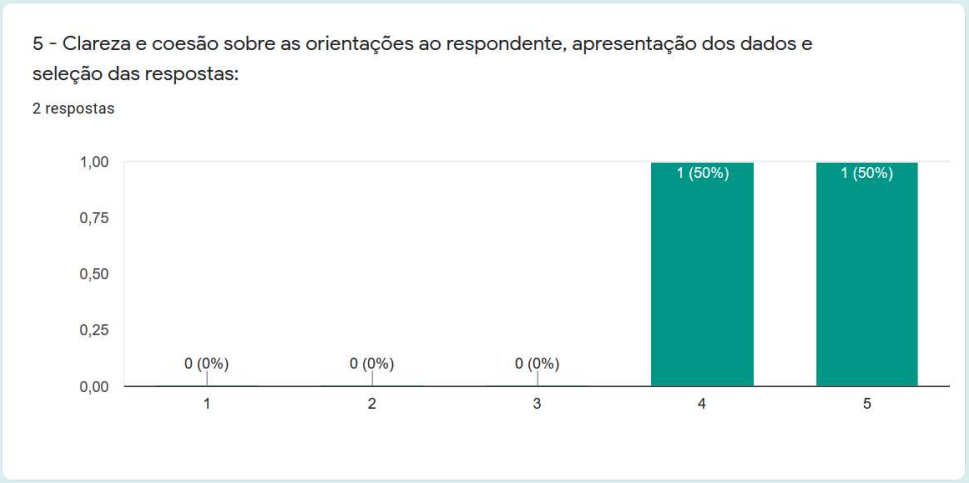
- Sobre clareza e coesão sobre a participação e/ou desistência do participante respondente.



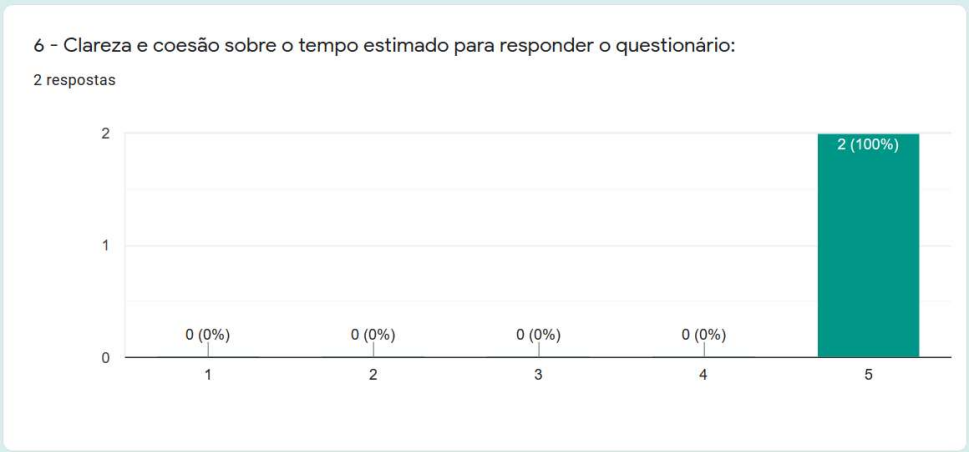
- Sobre clareza e coesão sobre o sigilo das informações.



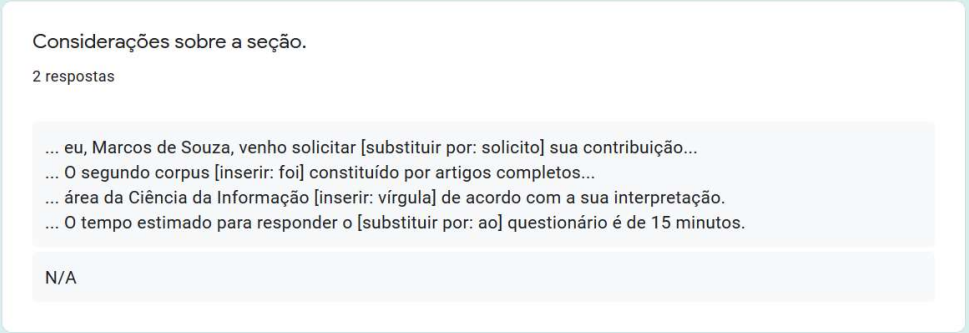
- Sobre clareza e coesão referente as orientações ao respondente, apresentação dos dados e seleção das respostas.



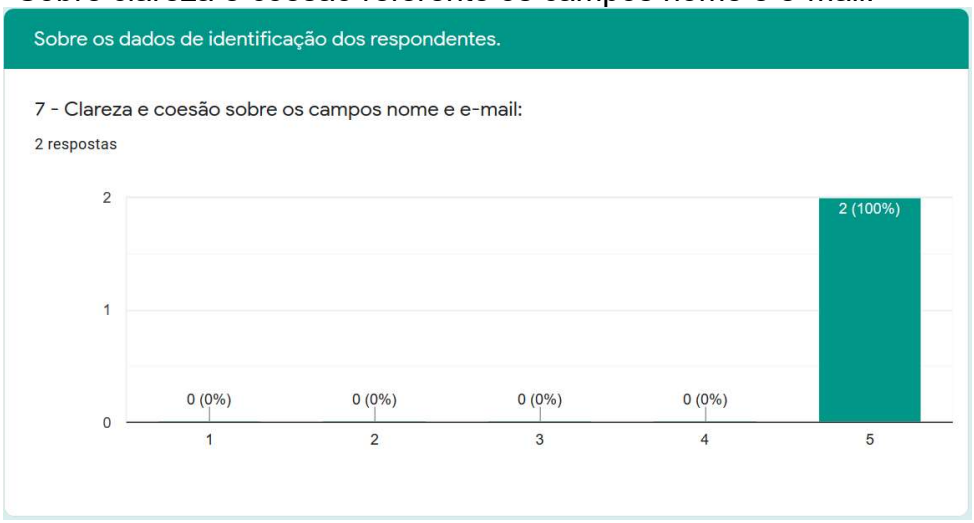
- Sobre clareza e coesão referente ao tempo estimado para responder o questionário.



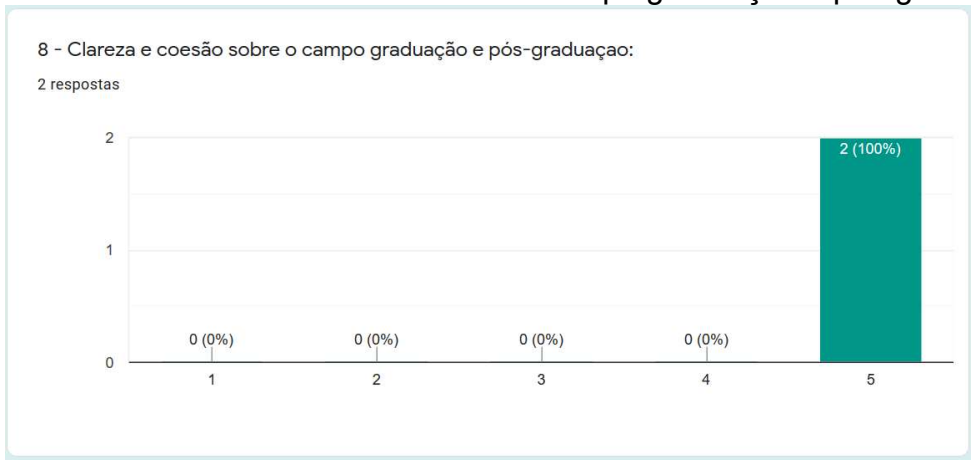
- Sobre as considerações da seção.



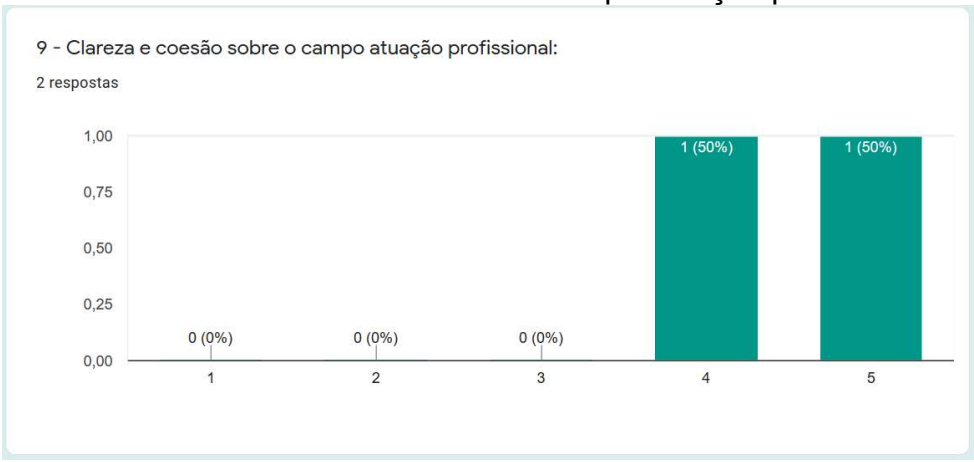
- Sobre clareza e coesão referente os campos nome e e-mail.



- Sobre clareza e coesão referente ao campo graduação e pós-graduaçao.



- Sobre clareza e coesão referente ao campo atuação profissional.



- Sobre as considerações da seção.

Considerações sobre a seção.

2 respostas

Sugiro que no campo Graduação as opções estejam em ordem alfabética.

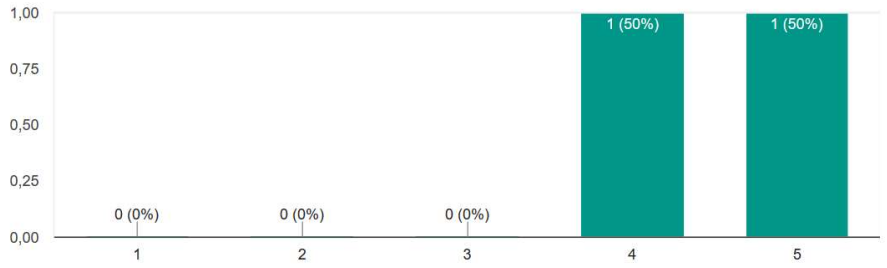
N/A

- Sobre clareza e coesão referente a descrição dos *corpora* de dados.

Sobre a descrição dos corpora de dados, orientações, perguntas e respostas.

10 - Clareza e coesão sobre a descrição do corpora de dados:

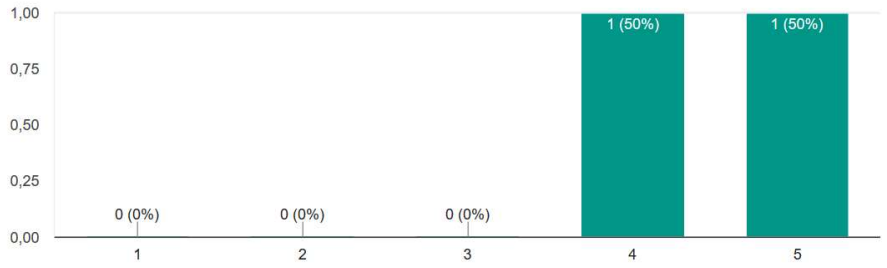
2 respostas



- Sobre clareza a orientação de consultas a *links* externos (apresentado na descrição da seção).

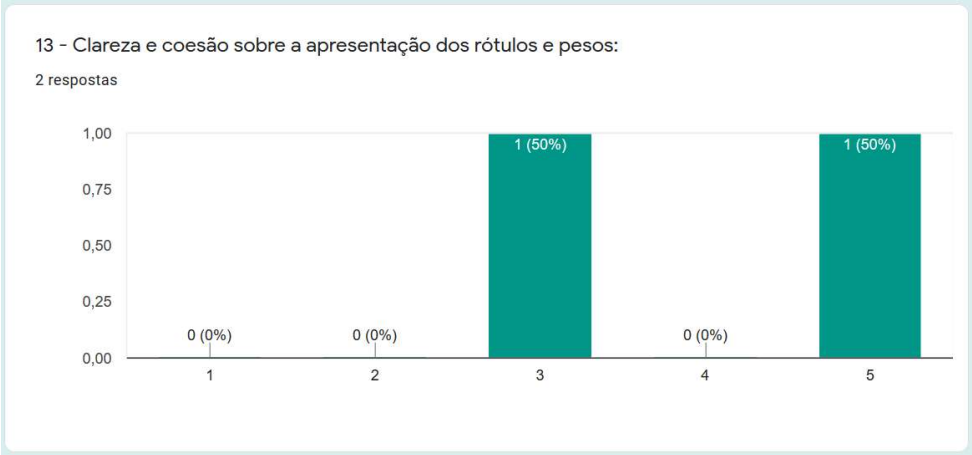
11 - Clareza sobre a orientação de consultas a links externos (apresentado na descrição da seção):

2 respostas

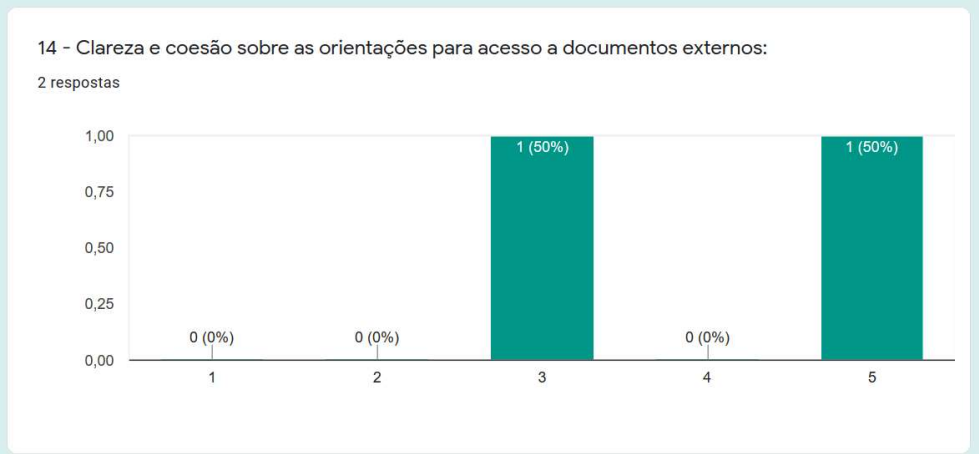


Sobre clareza e coesão referente as perguntas.

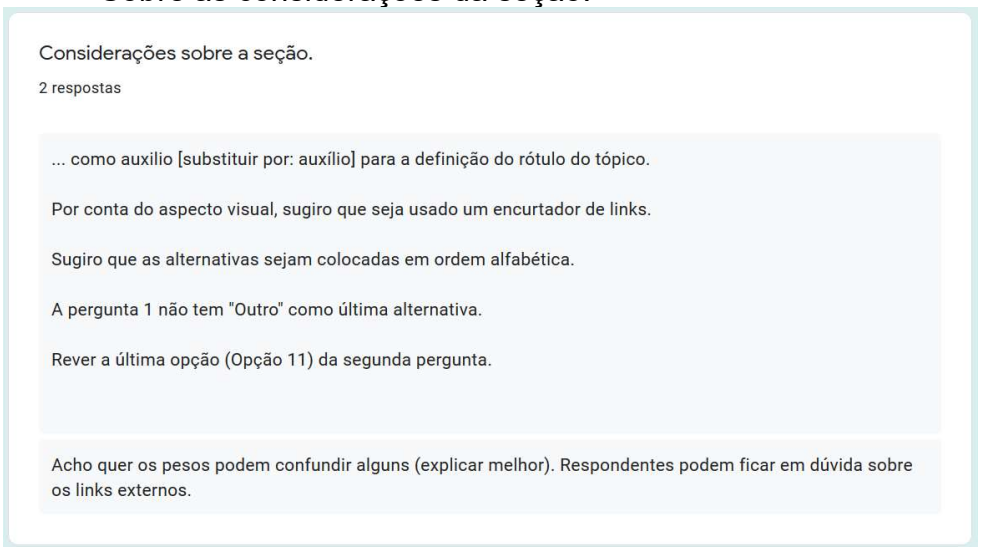
- Sobre clareza e coesão referente a apresentação dos rótulos e pesos.



- Sobre clareza e coesão referente as orientações para acesso a documentos externos.



- Sobre as considerações da seção.



APÊNDICE B - *Corpus* 2: teses e dissertações 2013

O *corpus* 2, constituído por teses e dissertações defendidas no ano de 2013 contempla o quantitativo de 266 documentos com tamanho de 96.053kb. Esse *corpus*, além pertencer ao primeiro *corpora* de dados, possui o quantitativo de 6.638.452 unigramas, 6.63.818 bigramas e 6.637.920 trigramas. O Quadro 23 apresenta uma lista contendo os 50 termos mais frequentes separados por tipo dos N-gramas.

Quadro 23 – Lista de N-gramas por ordem de frequência do *corpus* 2

Unigramas
informação,100977; pesquisa,30636; biblioteca,24794; conhecimento,24558; forma,20329; processo,19113; dados,16953; documentos,15630; trabalho,15304; uso,14881; comunicação,14823; relação,14763; brasil,14538; social,14526; sistema,13146; organização,12641; desenvolvimento,12086; memória,11964; meio,11941; termo,11471; gestão,11470; usuários,11400; sociedade,11196; produção,11114; museu,11106; cultura,11036; tempo,11001; sociais,10963; paulo,10891; estudo,10775; nacional,10359; fonte,10251; universidade,9869; busca,9652; anos,9306; atividades,9009; estudos,8910; pessoas,8892; cultural,8848; contexto,8819; educação,8662; usuário,8560; modelo,8486; política,8422; espaço,8420; diferentes,8268; vida,8233; caso,8195; serviços,8152; científica,8144.
Bigramas
universidade_federal,2895; recuperação_informação,2728; redes_sociais,2542; uso_informação,2385; informação_conhecimento,2329; belo_horizonte,1943; gestão_informação,1825; biblioteca_universitárias,1804; ensino_superior,1794; ponto_vista,1757; coleta_dados,1694; muitas_vezes,1644; fontes_informação,1638; patrimônio_cultural,1613; dissertação_mestrado,1600; sistemas_informação,1579; informação_tecnologia,1572; gestão_documentos,1517; produção_científica,1494; comunicação_científica,1493; informação_science,1491; tecnologias_informação,1406; competência_informacional,1404; informação_informação,1396; tomada_decisão,1305; informação_comunicação,1303; minas_gerais,1236; arquitetura_informação,1231; bases_dados,1226; busca_informação,1204; sociedade_informação,1197; biblioteca_universitária,1185; organização_conhecimento,1182; porto_alegre,1151; dados_pesquisa,1117; base_dados,1106; necessidades_informação,1081; tendo_vista,1071; biblioteca_digitais,1060; organização_informação,1010; comportamento_informacional,993; sistema_informação,981; biblioteca_digital,979; memória_social,969; produtos_serviços,935; gestão_conhecimento,916; informação_brasília,913; fonte_dados,913; educação_distância,907; qualidade_informação,881.
Trigramas
fonte_dados_pesquisa,846; tecnologias_informação_comunicação,762; machine_readable_cataloging,555; international_organization_standardization,510; extensible_markup_language,460; dissertação_mestrado_informação,452; instituição_ensino_superior,441; portal_periódicos_capes,434; instituições_ensino_superior,406; patrimônio_histórico_artístico,380; fonte_elaborado_autora,345; informação_universidade_federal,318; american_society_informação,317 universidade_federal_bahia,313; society_informação_science,312; federal_minas_gerais,312; gestão_informação_conhecimento,311; resource_description_framework,310; informação_belo_horizonte,303; universidade_federal_minas,302;

universidade_federal_grande,301; segunda_guerra_mundial,300; sistemas_organização_conhecimento,284; universidade_federal_paraíba,283; trabalho_conclusão_curso,281; busca_uso_informação,279; american_library_association,279; histórico_artístico_nacional,278; sob_ponto_vista,272; informação_science_technology,268; universidade_federal_santa,266; journal_american_society,264; web_ontology_language,263; informação_conhecimento_sociedade,254; arquivologia_biblioteconomia_museologia,252; brasília_briquet_lemos,248; federal_grande_sul,245; informado_informado_informado,245; paulo_companhia_letras,242; estação_memória_cambury,238; aluno_aluno_aluno,232; federal_santa_catarina,231; conhecimento_sociedade_contemporânea,226; sistema_recuperação_informação,225; poder_executivo_brasil,225; museu_astronomia_afins,224; educação_superior_distância,221; requisito_parcial_obtenção,220; compartilhamento_informação_conhecimento,218; anos_anos_anos,215.

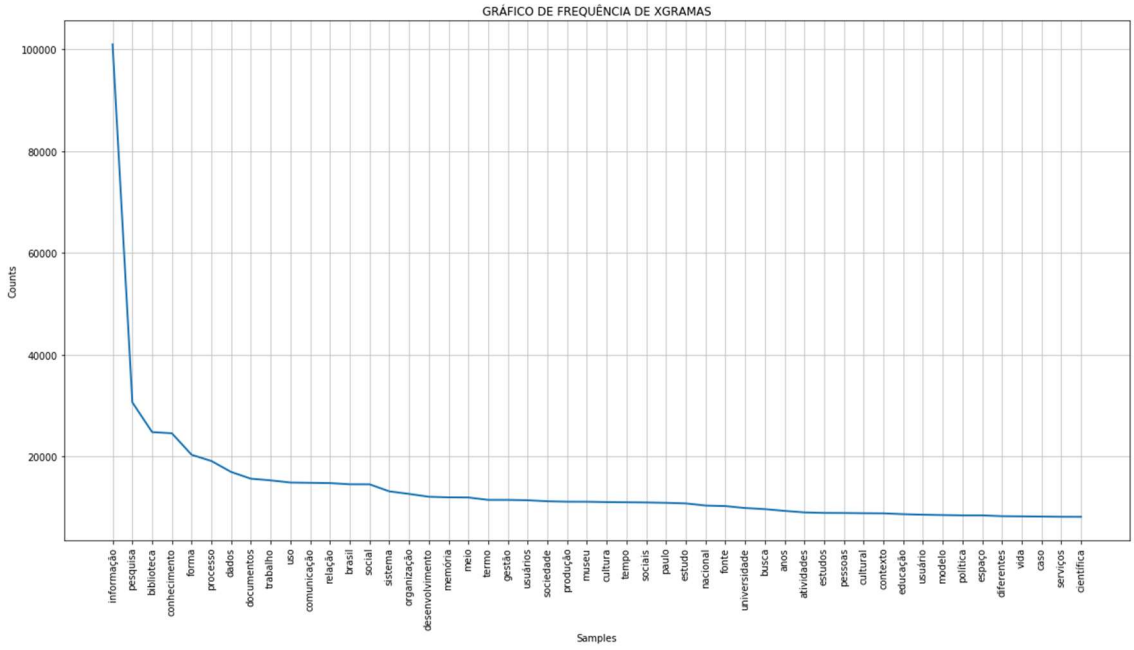
Fonte: Elaborado pelo autor.

Dentre os primeiros bigramas do *corpus* 2 estão os termos “universidade_federal” com frequência de 2.895 e ocupando a posição 323 dentre a lista contendo os mil N-gramas mais frequentes, seguido de “recuperação_informação” com frequência de 2.728 e ocupando a posição 351, “redes_sociais” com frequência de 2.542 na posição 377, “uso_informação” com frequência de 2.385 na posição 410 e “informação_conhecimento” com frequência de 2.329 na posição 425.

Já entre os primeiros trigramas estão “fonte_dados_pesquisa” com frequência de 846, “tecnologias_informação_comunicação” com frequência de 762, “machine_readable_cataloging” com frequência de 555, “international_organization_standardization” com frequência de 510 e “extensible_markup_language” com frequência de 460. Ambos os resultados estão ranqueados em ordem de frequência acima do milésimo termo da lista geral de N-gramas.

O Gráfico 53 apresenta os 50 termos mais frequentes do *corpus* 2 de dados. Nele, é possível observar que todos os termos são do tipo unigrama. A diferença entre frequências do primeiro termo “informação” extraído 100.977 vezes no corpus chega a 230% quando se comparado ao segundo termo “pesquisa” que possui frequência de 30.636. O percentual aumenta para 1140% quando comparado do primeiro ao 50º termo ”científica” com frequência de 8.144.

Gráfico 53 – Termos mais frequentes do *corpus* 2



Fonte: Elaborado pelo autor.

Embora os unigramas alcancem frequências maiores que os bigramas e trigramas, faz-se necessário destacar a importância do especialista no conhecimento da linguagem de domínio ou para a prática de explorar os demais resultados, umas vez que termos únicos como “uso”, “memória” ou “educação” podem ser generalistas, podendo representar outras áreas ou assuntos como “uso_informação”, “busca_uso”, “uso_recursos”, “uso_tecnologias”, “memória_história”, “memória_institucional”, “memória_patrimônio”, “preservação_memória”, “memória_social”, “memória_coletiva”, “secretaria_educação”, “educação_distância”, “educação_superior”, “ministério_educação” e “educação_cultura”.

Também são encontrados unigramas contendo frequências elevadas, mas com características fracas, como por exemplo “paulo” que aparece 10.891 e está associado a assuntos como autores, editora, referências, cidades ou universidades, não contribuindo assim para áreas, subáreas ou disciplinas da Ciência da Informação abordadas neste estudo.

A Figura 14 ilustra uma nuvem de palavras constituída por 250 N-gramas mais frequentes do *corpus* 2. Para essa construção, não foi utilizado nenhum tipo de tratamento qualitativo dos dados de maneira que termos pudessem ser excluídos ou substituídos.

Figura 14 – Nuvem de palavras do *corpus* 2

Fonte: Elaborado pelo autor.

Ainda na imagem, pode-se destacar termos fortes e fracos quando se comparado ao domínio de linguagem e realizado a análise de assunto. Exemplo de termos fortes estão “conceito”, “tempo” e “relação”, que, mesmo tendo que realizar análises através dos bigramas e trigramas para obter melhores resultados, possuem frequência representativa e familiaridade de termos no domínio da linguagem. Destaca-se como exemplos para os termos fracos “sempre”, “texto”, “objeto” e “brasil”, ao qual, mesmo quando analisado os bigramas e trigramas não apresentam valores e significados de relevância junto ao domínio de linguagem.

Após conectado os dados do *corpus* 2 ao modelo de treinamento LSI, foram extraídos conjuntos de resultados contendo tópicos, termos e pesos. O Quadro 24 apresenta os resultados contento a extração de 30 de tópicos que foram processados em 9 minutos e 47 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[57].

Quadro 24 – Tópicos extraídos do *corpus* 2 usando o modelo LSI

Tópico 0: 0.708*"informação" + 0.175*"pesquisa" + 0.168*"biblioteca" + 0.147*"conhecimento" + 0.099*"processo" + 0.096*"dados" + 0.093*"forma" + 0.092*"uso" + 0.086*"comunicação" + 0.082*"gestão";
Tópico 1: -0.533*"informação" + 0.190*"biblioteca" + 0.158*"museu" + 0.135*"brasil" + 0.135*"memória" + 0.116*"cultural" + 0.113*"forma" + 0.111*"documentos" + 0.109*"patrimônio" + 0.106*"cultura";

[57] Algoritmo de modelagem de tópicos. Corpus 2: teses e dissertações 2013. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2013.ipynb/.

Tópico 2: -0.728*"biblioteca" + -0.224*"usuários" + -0.161*"serviços" + -0.151*"usuário" + -0.129*"redes" + 0.125*"documentos" + -0.120*"redes_sociais" + -0.116*"web" + 0.096*"museu" + 0.096*"informação";
Tópico 3: 0.532*"documentos" + 0.245*"arquivo" + 0.232*"arquivos" + 0.228*"gestão" + 0.143*"documento" + 0.116*"arquivística" + 0.113*"documental" + -0.110*"conhecimento" + 0.108*"nacional" + -0.105*"aprendizagem";
Tópico 4: 0.274*"museu" + -0.192*"termo" + 0.180*"biblioteca" + -0.180*"usuário" + 0.164*"informação" + 0.145*"cultura" + 0.137*"cultural" + -0.136*"usuários" + -0.129*"documentos" + -0.128*"dados";
Tópico 5: 0.312*"curso" + 0.231*"universidade" + 0.209*"educação" + 0.206*"pesquisa" + 0.205*"distância" + 0.186*"cursos" + 0.180*"estudantes" + 0.168*"brasil" + 0.153*"ensino" + -0.149*"informação";
Tópico 6: -0.236*"científica" + 0.228*"curso" + 0.228*"museu" + -0.191*"brasil" + -0.191*"universidade" + -0.186*"pesquisa" + -0.147*"comunicação" + 0.143*"distância" + -0.142*"produção" + 0.142*"aprendizagem";
Tópico 7: -0.458*"zoo" + -0.295*"park" + -0.288*"museu" + -0.171*"zoológico" + -0.137*"aves" + 0.128*"curso" + -0.126*"usuário" + -0.117*"usuários" + -0.116*"animais" + -0.104*"aquarium";
Tópico 8: 0.200*"gente" + 0.198*"usuário" + -0.182*"biblioteca" + -0.180*"aprendizagem" + 0.171*"sistema" + 0.168*"usuários" + 0.156*"qualidade" + -0.156*"conhecimento" + -0.135*"competência" + -0.131*"documentos";
Tópico 9: 0.458*"museu" + -0.292*"zoo" + -0.189*"park" + 0.143*"usuário" + 0.134*"usuários" + -0.128*"qualidade" + -0.127*"gestão" + -0.109*"zoológico" + 0.108*"arte" + 0.108*"obras";
Tópico 10: -0.430*"cinema" + -0.238*"cinemas" + -0.214*"salas" + 0.182*"museu" + -0.139*"qualidade" + 0.137*"sociais" + -0.133*"cine" + -0.126*"cidade" + -0.121*"exibição" + -0.111*"filmes";
Tópico 11: -0.230*"brasil" + -0.203*"universidade" + -0.196*"gente" + -0.183*"usuários" + 0.180*"termo" + -0.173*"usuário" + -0.141*"trabalho" + 0.140*"curso" + 0.110*"obra" + 0.107*"distância";
Tópico 12: -0.340*"qualidade" + -0.220*"museu" + -0.195*"gestão" + 0.162*"cinema" + 0.127*"biblioteconomia" + 0.126*"documentos" + 0.124*"gente" + 0.122*"digital" + -0.122*"organização" + 0.117*"zoo";
Tópico 13: -0.259*"redes" + 0.220*"qualidade" + -0.219*"sociais" + -0.211*"redes_sociais" + 0.208*"biblioteca" + -0.200*"web" + 0.155*"brasil" + 0.120*"leitura" + 0.119*"universidade" + -0.119*"parque";
Tópico 14: -0.204*"biblioteconomia" + 0.197*"conhecimento" + -0.179*"brasil" + -0.178*"museu" + 0.173*"cultura" + 0.169*"cultural" + -0.167*"museologia" + -0.159*"sociais" + -0.154*"arquivologia" + -0.153*"cinema";
Tópico 15: 0.217*"museu" + 0.203*"comunicação" + 0.192*"qualidade" + -0.192*"universidade" + -0.175*"conhecimento" + -0.174*"brasil" + -0.171*"biblioteconomia" + 0.151*"científica" + -0.149*"organização" + 0.112*"aborto";
Tópico 16: 0.228*"memória" + -0.203*"trabalho" + -0.164*"museu" + -0.154*"conhecimento" + -0.144*"clube" + 0.143*"qualidade" + -0.137*"futebol" + -0.121*"inteligência" + -0.120*"biblioteca" + -0.118*"cinema";
Tópico 17: -0.412*"trabalho" + -0.200*"termo" + -0.182*"convenção" + -0.175*"crianças" + 0.157*"futebol" + 0.154*"clube" + -0.138*"qualidade" + -0.123*"infantil" + 0.121*"imagem" + -0.116*"parque";
Tópico 18: -0.216*"imagem" + -0.216*"termo" + 0.180*"conhecimento" + 0.170*"memória" + 0.168*"científica" + 0.158*"qualidade" + 0.150*"nietzsche" + 0.141*"obra" + -0.133*"imagens" + 0.124*"biblioteconomia";

Tópico 19: 0.243*"biblioteconomia" + 0.230*"parque" + -0.197*"trabalho" + 0.152*"museologia" + 0.152*"conservação" + 0.147*"arquivologia" + -0.136*"crianças" + -0.134*"museu" + 0.132*"conhecimento" + -0.122*"qualidade";
Tópico 20: -0.257*"termo" + -0.230*"conhecimento" + -0.216*"cultura" + 0.196*"parque" + 0.184*"imagem" + 0.134*"conservação" + -0.128*"cultural" + 0.125*"imagens" + -0.113*"sistemas" + -0.105*"política";
Tópico 21: -0.274*"conhecimento" + 0.190*"cultural" + -0.172*"museu" + 0.158*"inteligência" + 0.148*"modelo" + 0.144*"patrimônio" + 0.142*"cultura" + -0.136*"imagem" + 0.128*"biblioteconomia" + -0.128*"trabalho";
Tópico 22: -0.265*"imagem" + -0.208*"imagens" + 0.170*"memória" + -0.163*"conhecimento" + -0.157*"consumo" + 0.155*"clube" + 0.152*"termo" + 0.136*"obra" + 0.135*"futebol" + -0.122*"biblioteconomia";
Tópico 23: 0.213*"pesquisa" + 0.192*"termo" + 0.166*"leitura" + 0.164*"inteligência" + -0.154*"imagem" + -0.153*"trabalho" + -0.153*"imagens" + -0.146*"comunicação" + -0.143*"obra" + 0.139*"museu";
Tópico 24: 0.255*"memória" + -0.213*"termo" + 0.198*"dantas" + 0.160*"acontecimento" + 0.145*"satiagraha" + -0.126*"cultura" + 0.121*"operação" + -0.116*"aborto" + -0.112*"duelo" + 0.109*"abin";
Tópico 25: 0.244*"memória" + 0.242*"termo" + 0.191*"inteligência" + -0.158*"leitura" + -0.136*"digitais" + 0.135*"trabalho" + 0.135*"imagem" + -0.133*"modelo" + 0.120*"científica" + -0.120*"digital";
Tópico 26: 0.188*"inteligência" + -0.173*"aborto" + -0.152*"memória" + -0.150*"periódicos" + 0.149*"corpo" + 0.147*"obra" + -0.134*"alunos" + -0.129*"aluno" + -0.115*"busca" + -0.108*"patrimônio";
Tópico 27: 0.205*"leitura" + 0.197*"termo" + -0.173*"política" + 0.170*"processo" + -0.145*"aborto" + 0.145*"obra" + -0.139*"trabalho" + 0.138*"comunicação" + -0.135*"nietzsche" + -0.126*"clube";
Tópico 28: 0.215*"aborto" + -0.182*"alunos" + -0.175*"aluno" + -0.157*"cultura" + 0.153*"memória" + -0.141*"busca" + 0.138*"comunicação" + -0.134*"cidade" + -0.134*"duelo" + -0.129*"mcs";
Tópico 29: -0.235*"planejamento" + -0.218*"processo" + 0.214*"leitura" + 0.193*"conhecimento" + -0.144*"estratégico" + -0.141*"aborto" + -0.129*"modelo" + 0.128*"clube" + 0.125*"futebol" + -0.123*"obra".

Fonte: Elaborado pelo autor.

Ë possível perceber na composição dos tópicos extraídos do modelo LSI termos com pesos fortes acima de 0.500 quando se comparado aos demais resultados, bem como “informção” no tópico 0 e “documentos” no tópico 3. Os termos de cada tópico são apresentados por ordem de relevância de acordo com os pesos atribuídos. Embora existam tópicos que contemplam termos generalistas do domínio da linguagem e que possam representar diversos assuntos, cabe ao especialista identificar conjuntos de termos dentro do tópico que possa apontar para uma determinada área ou assunto, como por exemplo no tópico 1 que contempla termos como “informação”, “biblioteca”, “documentos” e “forma”, presente em tatos outros tópicos, entretanto, o tópico também contempla os termos “memória”, “cultural”, “museu” e “patrimônio”, apresentando

um norte para que o especialita possa supor um rótulo que melhor se adéque ao domínio de linguagem.

Tópicos contemplando um grande volume de termos fortes também foram extraídos nesta coleção de documentos, como por exemplo o tópico 7 e 9 que apresentam características de uma área específica no domínio da linguagem, destacando tópicos como “zoológico”, “park”, “museu”, “animais”, “aves”, “usuários”, “gestão” e “qualidade” bem como o tópico 10 com os termos “cinema”, “salas”, “exibição”, “filmes” e “sociais”. É possível perceber também tópicos fracos que contemplam mais de um assunto representados por seus termos, como por exemplo o tópico 29 que apresentam os termos especialistas “aborto” e “futebol”, além dos termos generalistas “processo” e “conhecimento”.

O Quadro 25 apresenta um conjunto de 30 tópicos compostos por termos e pesos extraídos do *corpus* 2 usando do modelo de extração LDA, sendo executado em 48 minutos e 11 segundos. Os demais resultados com 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[58].

Quadro 25 – Tópicos extraídos do *corpus* 2 usando o modelo LDA

Tópico 0: 0.001*"passione" + 0.001*"finaestampa" + 0.001*"telenovela" + 0.001*"insensatocoração" + 0.001*"tuiteiros" + 0.001*"twitter" + 0.001*"tweets" + 0.001*"novela" + 0.000*"xxx" + 0.000*"dec";
Tópico 1: 0.001*"iphaep" + 0.000*"dec._iphaep" + 0.000*"iphaep_edificação" + 0.000*"dec." + 0.000*"dec._iphaep_edificação" + 0.000*"edificação" + 0.000*"pessoa_dec." + 0.000*"pessoa_dec._iphaep" + 0.000*"iphaep_conjunto" + 0.000*"edificação_igreja";
Tópico 2: 0.008*"patrimônio" + 0.006*"museu" + 0.003*"museologia" + 0.003*"cinema" + 0.003*"cultural" + 0.002*"patrimônio_cultural" + 0.002*"cidade" + 0.002*"obras" + 0.002*"cinemas" + 0.001*"preservação";
Tópico 3: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"museu" + 0.000*"forma" + 0.000*"processo" + 0.000*"zoo" + 0.000*"conhecimento" + 0.000*"brasil" + 0.000*"meio" + 0.000*"comunicação";
Tópico 4: 0.003*"aborto" + 0.002*"dantas" + 0.001*"satiagraha" + 0.001*"dilma" + 0.001*"acontecimento" + 0.001*"operação" + 0.001*"stf" + 0.001*"abin" + 0.001*"protógenes" + 0.001*"cpi";
Tópico 5: 0.002*"universidade" + 0.002*"brasil" + 0.001*"university" + 0.001*"usa" + 0.001*"instituto" + 0.001*"artigos" + 0.001*"instituto_pesquisa" + 0.001*"escola_superior" + 0.001*"biológicas" + 0.001*"ufrgs";
Tópico 6: 0.000*"paisagem" + 0.000*"geografia" + 0.000*"ibge" + 0.000*"nacional_geografia" + 0.000*"conselho_nacional_geografia" + 0.000*"unidade_categoria_administrativa" + 0.000*"unidade_categoria" + 0.000*"categoria_administrativa" + 0.000*"agrícolas" + 0.000*"tibor";
Tópico 7: 0.004*"zoo" + 0.003*"park" + 0.001*"zoológico" + 0.001*"filmes" + 0.001*"aves" + 0.001*"museu" + 0.001*"animais" + 0.001*"aquarium" + 0.001*"doação" + 0.001*"jardim"

[58] Algoritmo de modelagem de tópicos. Corpus 2: teses e dissertações 2013. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2013.ipynb/.

Tópico 8: 0.011*"museu" + 0.003*"futebol" + 0.002*"nietzsche" + 0.002*"arte" + 0.002*"clube" + 0.001*"homem" + 0.001*"cidade" + 0.001*"arendt" + 0.001*"mediadores" + 0.001*"mast";
Tópico 9: 0.000*"informação" + 0.000*"museu" + 0.000*"conhecimento" + 0.000*"processo" + 0.000*"pesquisa" + 0.000*"brasil" + 0.000*"sociedade" + 0.000*"dados" + 0.000*"forma" + 0.000*"organização";
Tópico 10: 0.001*"boatos" + 0.001*"referência" + 0.001*"chat" + 0.001*"boato" + 0.001*"bibliotecário" + 0.001*"preço" + 0.001*"serviço_referência" + 0.000*"via_chat" + 0.000*"termo_indexação" + 0.000*"conclusão_curso";
Tópico 11: 0.004*"curso" + 0.004*"distância" + 0.003*"estudantes" + 0.001*"educação_distância" + 0.001*"educação" + 0.001*"cursos" + 0.001*"moodle" + 0.001*"curso_curso" + 0.001*"tutores" + 0.001*"aprendizagem";
Tópico 12: 0.003*"dança" + 0.002*"corpo" + 0.001*"movimento" + 0.001*"mulher" + 0.001*"filme" + 0.001*"cinema" + 0.001*"memória" + 0.001*"cassino" + 0.001*"verdade" + 0.001*"deputados";
Tópico 13: 0.002*"duelo" + 0.002*"mcs" + 0.002*"ombudsman" + 0.001*"duelo_mcs" + 0.001*"jornalismo" + 0.001*"departamento_cultura" + 0.001*"hip-hop" + 0.001*"refugiados" + 0.001*"departamento" + 0.001*"cidade";
Tópico 14: 0.002*"performance" + 0.001*"musical" + 0.001*"ufjf" + 0.001*"mdct" + 0.001*"eejf" + 0.001*"eletricidade" + 0.000*"performance_musical" + 0.000*"deyrolle" + 0.000*"farmácia" + 0.000*"mflma";
Tópico 15: 0.001*"termo" + 0.001*"reuters" + 0.001*"scielo" + 0.001*"candidato" + 0.001*"notícias" + 0.001*"afp" + 0.001*"candidatos" + 0.001*"efe" + 0.001*"eleições" + 0.000*"vassouras";
Tópico 16: 0.002*"cambury" + 0.001*"estação_memória" + 0.001*"memória_cambury" + 0.001*"estação_memória_cambury" + 0.001*"mediação_cultural" + 0.001*"parceiros_muda" + 0.001*"cambury_mediação" + 0.001*"cambury_mediação_cultural" + 0.001*"memória_cambury_mediação" + 0.001*"cultural_parceiros_muda";
Tópico 17: 0.000*"informação" + 0.000*"processo" + 0.000*"dados" + 0.000*"pesquisa" + 0.000*"sistema" + 0.000*"mulher" + 0.000*"forma" + 0.000*"comunicação" + 0.000*"relação" + 0.000*"organização";
Tópico 18: 0.002*"telenovela" + 0.001*"consumo" + 0.001*"imagem" + 0.001*"beleza" + 0.001*"imagens" + 0.001*"garotas" + 0.001*"mulher" + 0.001*"publicidade" + 0.001*"telenovelas" + 0.001*"moda";
Tópico 19: 0.001*"indústria" + 0.001*"resp" + 0.000*"robatto" + 0.000*"usina" + 0.000*"telefone" + 0.000*"agentes_mediadores" + 0.000*"departamentos" + 0.000*"ltda" + 0.000*"segurança_trabalho" + 0.000*"gente";
Tópico 20: 0.001*"acontecimento" + 0.001*"memórias_coletivas" + 0.001*"engenharia" + 0.001*"comportamento_informacional" + 0.001*"personalidade" + 0.000*"coletivas" + 0.000*"memórias" + 0.000*"arte_digital" + 0.000*"atentados" + 0.000*"trending_topics";
Tópico 21: 0.000*"remontagem" + 0.000*"companhia_dança" + 0.000*"dança_cidade" + 0.000*"companhia_dança_cidade" + 0.000*"processo_remontagem" + 0.000*"trabalho_remontagem" + 0.000*"remontagem_obras" + 0.000*"obras_dança" + 0.000*"graciela" + 0.000*"obras_dança_moderna";
Tópico 22: 0.014*"informação" + 0.004*"pesquisa" + 0.003*"conhecimento" + 0.003*"biblioteca" + 0.003*"forma" + 0.003*"processo" + 0.002*"dados" + 0.002*"documentos" + 0.002*"trabalho" + 0.002*"uso";
Tópico 23: 0.002*"digitais" + 0.002*"ontologias" + 0.002*"web" + 0.001*"facetada" + 0.001*"ontologia" + 0.001*"entidades" + 0.001*"semântica" + 0.001*"biblioteca_digitais" + 0.001*"representação" + 0.001*"catalogação";
Tópico 24: 0.003*"parque" + 0.001*"biblioteca" + 0.001*"conservação" + 0.001*"entorno" + 0.001*"parques" + 0.001*"indicadores" + 0.001*"tijuca" + 0.001*"desempenho" + 0.001*"unidades_conservação" + 0.001*"parna";

Tópico 25: 0.002*"aprendizagem" + 0.001*"pergunta" + 0.001*"competência" + 0.001*"informado" + 0.001*"competência_informacional" + 0.001*"informado_informado" + 0.001*"informacional" + 0.001*"fonte_aprendizagem" + 0.001*"atribuíram" + 0.001*"atribuíram_nota";
Tópico 26: 0.001*"enunciado" + 0.001*"expertise" + 0.001*"experts" + 0.001*"resolução_problemas" + 0.001*"novatos" + 0.001*"solucionador" + 0.000*"resolução" + 0.000*"experimento" + 0.000*"escore" + 0.000*"escore_alcançado";
Tópico 27: 0.002*"inteligência" + 0.001*"metadados" + 0.001*"competitiva" + 0.001*"inteligência_competitiva" + 0.001*"rdf" + 0.001*"concordo" + 0.001*"language" + 0.001*"semântica" + 0.001*"modelagem" + 0.001*"registros";
Tópico 28: 0.000*"honneth,2003" + 0.000*"habermas,2007" + 0.000*"ouseja" + 0.000*"alberti" + 0.000*"pereira,2007" + 0.000*"nobrasil" + 0.000*"honneth,2008" + 0.000*"emancipaçãoejustiça" + 0.000*"discriminaçãoracial" + 0.000*"riodejaneiro";
Tópico 29: 0.001*"recife" + 0.000*"partitura" + 0.000*"pernambuco" + 0.000*"bairro" + 0.000*"recife_pernambuco_brasil" + 0.000*"recife_pernambuco" + 0.000*"pernambuco_brasil" + 0.000*"paisagem" + 0.000*"cais" + 0.000*"paisagem_vista".

Fonte: Elaborado pelo autor.

Os resultados extraídos a partir do modelo LDA são caracterizados por apresentar uma maior quantidade de bigramas e trigramas e uma menor quantidade de termos generalistas quando se comparado ao modelo LSI, o que permite ao especialista um menor esforço cognitivo para realizar a definição da suposição dos nomes dos tópicos. Uma característica fraca apresentada nos resultados do modelo está nos baixos valores dos pesos e em alguns casos, todos termos do tópico apresentam valores igual a 0, o que pode dificultar a interpretação dos dados pelo especialista por não apresentar uma ordem de relevância entre os termos.

Embora os resultados possam ser melhores interpretados pelas caracterizas apresentadas, o uso de documentos externos como acesso ao *corpus* de dados pode auxiliar na interpretação dos tópicos pelo especialista de forma mais assertiva. Um exemplo dessa técnica está no tópico 18 que apresenta termos de difícil interpretação para o especialista como “telenovela”, “consumo”, “mulher”, “imagem’ e “publicidade” quando analisados separadamente. Quando acessado ao *corpus* de dados, e utilizando de técnica simples de localização de termos específicos, percebe-se que se trata de uma pesquisa na área de museologia que está relacionada a tevenovela brasileira, podendo assim supor de maneira mais assertiva que o tópico esteja relacionado ao rótulo Informação e Patrimônio Imaterial.

Já o tópico 19 apresenta características fortes em seus termos, entretanto, faz-se necessário realizar a interpretação dos termos por meio de associação, principalmente por não apresentar dente os resultados nenhum

termo generalista como “informação” e “comunicação” pertinentes ao domínio de linguagem. Os termos destaque do tópico são “industría”, “usina”, “agentes_mediadores”, “departamentos” e “segurança_trabalho” que estão associados a pesquisa de Processos de Comunicação e Mediação da Informação, podendo supor por exemplo que o tópico esteja relacionado a Mediação da Informação ou Teoria da Informação.

Também é possível encontrar tópicos fracos entre os resultados, apresentando assim termos desconexos como “memória_coeltiva”, “arte_digiral”, “personalidade”, “acontecimento” e “engenharia” como apresentado no tópico 20. Neste caso o especialista poderá optar pelo descarte do tópico ou criar uma categoria de tópicos não categorizados por exemplo.

APÊNDICE C - *Corpus* 3: teses e dissertações 2014

O *corpus* 3 pertencente ao primeiro *corpora* de dados possui o quantitativo de 282 documentos do tipo teses e dissertações defendidas no ano de 2014. Esses documentos possuem o tamanho de 96.801kb, sendo 6.670.028 unigramas, 6.669.746 bigramas e 6.669.464 trigramas. O Quadro 26 apresenta os 50 termos mais frequentes separados por tipo dos N-gramas.

Quadro 26 – Lista de N-gramas por ordem de frequência do *corpus* 3

Unigramas
informação,108609; pesquisa,31935; conhecimento,27061; forma,21569; dados,20225; processo,19465; biblioteca,17974; social,17068; trabalho,15861; relação,15031; comunicação,13925; brasil,13706; uso,13490; produção,13013; organização,12977; meio,12926; desenvolvimento,12367; sistema,12209; museu,12205; sociedade,11926; sociais,11914; cultura,11858; documentos,11600; busca,11380; termo,11279; memória,10962; estudo,10877; tempo,10844; fonte,10363; nacional,10267; educação,10266; paulo,9981; usuários,9893; anos,9823; pessoas,9728; contexto,9656; construção,9451; universidade,9379; história,9023; estudos,8979; mundo,8864; atividades,8740; caso,8629; gestão,8570; campo,8417; diferentes,8357; base,8196; vida,8163; grande,8146; sistemas,8027.
Bigramas
universidade_federal,3147; recuperação_informação,2737; informação_conhecimento,2401; uso_informação,1958; redes_sociais,1891; fontes_informação,1873; muitas_vezes,1751; produção_científica,1726; informação_tecnologia,1685; base_dados,1682; bases_dados,1661; gestão_informação,1659; ponto_vista,1656; coleta_dados,1626; belo_horizonte,1584; dissertação_mestrado,1572; informação_science,1563; arquitetura_informação,1536; informação_informação,1504; organização_conhecimento,1452; ensino_superior,1418; dados_pesquisa,1414; sistemas_informação,1413; porto_alegre,1373; busca_informação,1347; informação_comunicação,1251; tecnologias_informação,1207; tendo_vista,1158; sociedade_informação,1133; comunicação_científica,1114; organização_informação,1073; necessidades_informação,1062; fonte_dados,1038; direitos_humanos,1025; competência_informacional,1000; patrimônio_cultural,974; educação_distância,970; tomada_decisão,969; minas_gerais,963; competência_informação,954; meio_ambiente,950; estudo_caso,880; biblioteca_escolar,878; objeto_estudo,868; informação_brasília,856; universidade_estadual,846; gestão_conhecimento,823; fonte_elaborado,821; pesquisa_informação,816; sistema_informação,813.
Trigramas
fonte_dados_pesquisa,834; tecnologias_informação_comunicação,701; fonte_elaborado_autora,598; museu_histórico_nacional,524; dissertação_mestrado_informação,477; portal_periódicos_capes,437; universidade_federal_paraíba,420; fotografias_obras_arte,420; domínio_escopo_mpeg7,406; informação_universidade_federal,382; universidade_federal_grande,381; resource_description_framework,375; society_informação_science,342; american_society_informação,341; informação_belo_horizonte,334; instituição_ensino_superior,330; universidade_federal_santa,327; federal_grande_sul,315; sistemas_organização_conhecimento,312; web_ontology_language,311; informação_science_technology,310; busca_uso_informação,302

universidade_federal_bahia,295; federal_santa_catarina,295; federal_minas_gerais,286; encontro_nacional_pesquisa,283; nacional_pesquisa_informação,279; instituições_federais_ensino,269; instituições_ensino_superior,268; sistema_recuperação_informação,267; universidade_federal_minas,263; journal_american_society,261; sistemas_recuperação_informação,261; federais_ensino_superior,256; fundação_oswaldo_cruz,254; ensino_superior_brasil,254; gestão_informação_conhecimento,251; organização_representação_conhecimento,247; trabalho_conclusão_curso,245; paulo_companhia_letras,238; library_informação_science,236; requisito_parcial_obtenção,230; brasília_briquet_lemos,228; extensible_markup_language,223; busca_recuperação_informação,218; produtos_serviços_informação,214; transparência_informação_instituições,214; informação_instituições_federais,212; informação_sociedade_estudos,210; perspectivas_informação_belo,207.

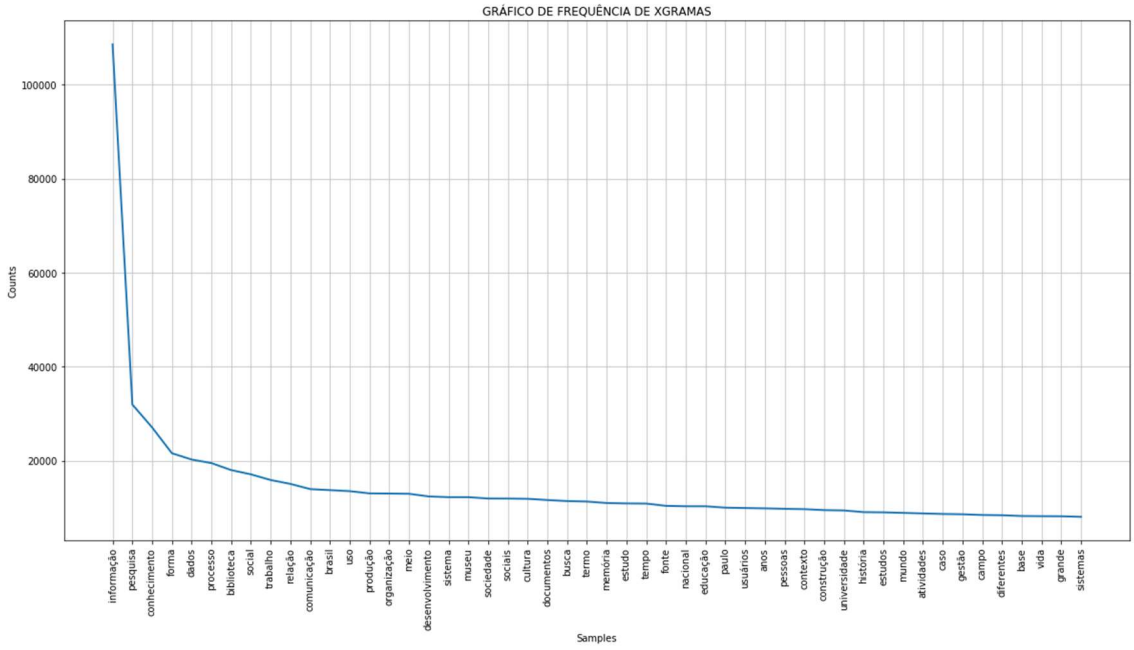
Fonte: Elaborado pelo autor.

Em uma lista de frequência contendo os mil primeiros N-gramas do *corpus* 3, são destacados os 5 bigramas e trigramas com maior número de repetições, sendo eles: “universidade_federal” com frequência de 3.147 e ocupando a posição de 292, “recuperação_informação” com frequência de 2.737 e posição 347, “informação_conhecimento” com frequência de 2.401 e posição de 419, “uso_informação” com frequência de 1.958 e posição 553, “redes_sociais” com frequência de 1.891 e posição 579 para bigramas; e “fonte_dados_pesquisa” com frequência de 834, “tecnologias_informação_comunicação” com frequência de 701, “fonte_elaborado_autora” com frequência de 598, “museu_histórico_nacional” com frequência de 524 e “dissertação_mestrado_informação” com frequência de 477. Todos os trigramas aparecem após ao milésimo termo da lista geral de N-gramas.

Os unigramas possuem maior frequência que os demais tipos de N-gramas. Quando comparado o unigrama com maior frequência com o segundo bigrama melhor ranqueado e o mais representativo do *corpus* 3, sendo eles “informação” com frequência de 108.609 e “recuperação_informação” com 2.737, encontra-se uma diferença de 3.868% entre seus valores. Esse percentual aumenta ainda mais quando comparado o primeiro unigrama com o primeiro trigrama “fonte_dados_pesquisa” com frequência 834, resultando assim numa diferença de 12.923%.

O Gráfico 54 apresenta os 50 termos mais frequentes do *corpus* 3, sendo todos eles caracterizados como unigramas. O termo “informação” possui maior frequência com 108.609, seguido de “pesquisa” com 31.935 e “conhecimento” com 27.061.

Gráfico 54 – Termos mais frequentes do *corpus* 3



Fonte: Elaborado pelo autor.

É possível observar uma evolução na frequência de forma contínua do 50º ao 4º termo e uma quebra desse processo quando analisado os três últimos termos. Além disso, é possível identificar termos fortes como “dados” e “processo” e termos fracos como “grande”, “anos”, “diferentes”, “contexto” e “meio”, sendo necessário assim analisar os bigramas e trigramas do corpus como forma de entender o contexto dos termos.

A Figura 15 apresenta uma nuvem de palavras construída por 250 termos mais frequentes contidos no *corpus* 3. Os resultados apresentados nessa figura não possuem quaisquer tipos de filtragem ou tratamento qualitativo, como por exemplo a exclusão ou substituição de termos fracos quando se comparado ao domínio do corpus através dos N-gramas.

Figura 15 – Nuvem de palavras do *corpus* 3

Fonte: Elaborado pelo autor.

Na imagem é possível identificar termos fortes como “conceito”, “relação” e “usuário”, mesmo sendo necessário realizar um aprofundamento através dos bigramas e trigramas de forma que seja possível encontrar vertentes dos termos com significado ao domínio da linguagem estudado. Destaca-se como exemplos de termos fracos “grande”, “brasil” e “possibilidade” que, mesmo quando analisado os bigramas e trigramas, não apresentam significados pertinentes ao domínio da linguagem.

O processo de treinamento dos modelos resultou em conjuntos de tópicos formados por termos e pesos extraídos do *corpus* de dados. O Quadro 27 apresenta um conjunto de 30 tópicos extraídos através do modelo LSI e foi realizado em 9 minutos e 51 segundos. Os demais resultados do *corpus* 3 contendo os conjuntos treinados e extraídos com 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponível através do GitHub[59].

Quadro 27 – Tópicos extraídos do *corpus* 3 usando o modelo LSI

Tópico 0: 0.772*"informação" + 0.157*"conhecimento" + 0.144*"pesquisa" + 0.103*"forma" + 0.100*"dados" + 0.098*"processo" + 0.086*"social" + 0.075*"organização" + 0.074*"uso" + 0.069*"biblioteca";
Tópico 1: -0.524*"informação" + 0.197*"museu" + 0.143*"memória" + 0.141*"pesquisa" + 0.120*"brasil" + 0.115*"cultura" + 0.111*"dança" + 0.110*"biblioteca" + 0.108*"forma" + 0.106*"trabalho";

[59] Algoritmo de modelagem de tópicos. Corpus 3: teses e dissertações 2014. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2014.ipynb/.

Tópico 2: 0.239*"domínio" + 0.221*"conteúdo" + 0.212*"ontologias" + 0.205*"multimídia" + 0.199*"ontologia" + 0.185*"conhecimento" + 0.163*"dados" + 0.161*"metadados" + 0.159*"mpeg7" + 0.153*"escopo";
Tópico 3: -0.414*"biblioteca" + 0.293*"museu" + -0.238*"pesquisa" + -0.198*"dados" + 0.162*"memória" + 0.152*"cinema" + -0.140*"busca" + 0.120*"informação" + -0.100*"universidade" + -0.099*"periódicos";
Tópico 4: 0.616*"museu" + -0.254*"dança" + -0.154*"social" + 0.142*"acervo" + 0.141*"nacional" + -0.134*"sociais" + 0.133*"patrimônio" + 0.133*"documentos" + 0.113*"objetos" + -0.105*"sujeitos";
Tópico 5: -0.558*"biblioteca" + 0.250*"conhecimento" + 0.156*"científica" + 0.152*"pesquisa" + 0.133*"produção" + -0.131*"leitura" + -0.123*"livro" + 0.118*"dados" + -0.111*"multimídia" + 0.107*"artigos";
Tópico 6: 0.517*"cinema" + -0.267*"conhecimento" + 0.251*"memória" + -0.207*"museu" + -0.161*"biblioteca" + 0.149*"imagens" + -0.130*"dança" + -0.129*"educação" + 0.128*"dados" + -0.126*"cultura";
Tópico 7: 0.715*"dança" + -0.147*"social" + 0.137*"campo" + 0.131*"profissional" + -0.123*"sujeitos" + -0.115*"livro" + 0.110*"arte" + -0.109*"sociais" + 0.103*"cidade" + -0.102*"biblioteca";
Tópico 8: 0.461*"conhecimento" + 0.332*"cinema" + 0.180*"biblioteca" + 0.164*"organização" + -0.149*"sociais" + -0.148*"busca" + -0.146*"museu" + 0.120*"memória" + -0.118*"usuários" + 0.116*"gestão";
Tópico 9: 0.285*"documentos" + -0.267*"museu" + 0.264*"arquivos" + 0.246*"arquivo" + -0.183*"arte" + -0.175*"conhecimento" + -0.173*"busca" + 0.147*"direitos" + 0.132*"nacional" + 0.128*"poder";
Tópico 10: -0.539*"livro" + 0.227*"cinema" + -0.214*"leitura" + -0.177*"livros" + -0.151*"leitor" + 0.149*"sociais" + 0.148*"sujeitos" + -0.137*"skoob" + 0.118*"pesquisa" + 0.111*"social";
Tópico 11: -0.289*"documentos" + -0.204*"arquivo" + -0.191*"busca" + -0.188*"termo" + -0.185*"arquivos" + -0.153*"dança" + 0.122*"nuclear" + 0.118*"produção" + -0.116*"sujeitos" + 0.110*"trabalho";
Tópico 12: 0.268*"busca" + -0.233*"científica" + 0.184*"poder" + -0.182*"produção" + -0.161*"artigos" + 0.145*"sistema" + -0.136*"sujeitos" + 0.128*"regimes" + 0.124*"internet" + 0.113*"regime";
Tópico 13: 0.347*"poder" + 0.229*"regimes" + 0.202*"biblioteca" + 0.197*"regime" + 0.161*"internet" + 0.148*"científica" + 0.122*"power" + 0.120*"that" + -0.114*"gestão" + 0.105*"internacionais";
Tópico 14: -0.236*"busca" + 0.210*"livro" + 0.190*"redes" + -0.169*"maracanã" + 0.168*"rede" + 0.141*"arte" + 0.140*"digital" + 0.127*"cinema" + 0.122*"sites" + -0.121*"construção";
Tópico 15: -0.264*"cidadania" + 0.243*"maracanã" + 0.183*"estádio" + -0.176*"direitos" + 0.172*"brasil" + -0.169*"comunicação" + 0.157*"rede" + 0.132*"redes" + 0.126*"entrevistadora" + -0.114*"digital";
Tópico 16: 0.351*"dados" + 0.216*"cito" + -0.210*"busca" + 0.156*"maracanã" + 0.132*"digital" + 0.130*"relações" + -0.129*"produção" + 0.127*"influencia" + 0.121*"pesquisa" + -0.118*"periódicos";
Tópico 17: -0.431*"arte" + -0.347*"fotografias" + -0.281*"obras" + -0.204*"obras_arte" + -0.173*"fotografias_obras" + -0.171*"fotografias_obras_arte" + 0.136*"museu" + -0.129*"maracanã" + 0.125*"livro" + -0.107*"respondentes";
Tópico 18: 0.343*"nuclear" + 0.287*"emergência" + 0.163*"biblioteca" + 0.142*"plano" + 0.138*"energia" + -0.134*"trabalho" + -0.134*"cultura" + -0.125*"arte" + -0.119*"cidadania" + 0.114*"segurança";

Tópico 19: 0.357*"cultura" + -0.188*"trabalho" + 0.178*"cultural" + 0.140*"organização" + -0.131*"maracanã" + 0.129*"kaxuyana" + 0.117*"nuclear" + -0.117*"curso" + -0.112*"cursos" + -0.112*"social";
Tópico 20: 0.208*"maracanã" + -0.191*"cidadania" + -0.180*"qualidade" + 0.178*"busca" + -0.156*"empresa" + 0.155*"cultura" + 0.149*"estádio" + -0.136*"termo" + -0.123*"comunicação" + 0.115*"informacional";
Tópico 21: -0.274*"cidadania" + 0.210*"cultura" + -0.203*"maracanã" + 0.170*"trabalho" + -0.150*"estádio" + -0.148*"cinema" + -0.147*"direitos" + -0.129*"organização" + 0.126*"kaxuyana" + -0.119*"comunicação";
Tópico 22: -0.199*"direitos" + -0.179*"digital" + 0.143*"pesquisa" + -0.128*"patrimônio" + 0.128*"mulheres" + -0.125*"maracanã" + -0.117*"empresa" + 0.112*"transparência" + -0.110*"trabalho" + -0.110*"conhecimento";
Tópico 23: -0.303*"trabalho" + -0.223*"produção" + 0.209*"transparência" + 0.188*"linguagem" + 0.184*"educação" + 0.169*"cultura" + 0.126*"gestão" + -0.124*"marx" + 0.119*"distância" + 0.117*"verdade";
Tópico 24: -0.250*"transparência" + -0.245*"trabalho" + 0.164*"qualidade" + -0.129*"brasil" + -0.126*"direitos" + 0.123*"documentos" + -0.118*"web" + 0.117*"arquivo" + 0.112*"gestão" + -0.111*"marx";
Tópico 25: -0.188*"conhecimento" + -0.161*"direitos" + 0.157*"produção" + -0.146*"corpo" + -0.143*"mulheres" + -0.139*"memória" + 0.136*"digital" + -0.133*"biblioteca" + -0.130*"dados" + -0.129*"aborto";
Tópico 26: 0.269*"direitos" + 0.189*"qualidade" + 0.183*"humanos" + 0.182*"direitos_humanos" + -0.152*"transparência" + 0.151*"unesco" + -0.150*"cidadania" + 0.121*"arquivos" + -0.120*"gestão" + 0.107*"educação";
Tópico 27: 0.223*"memória" + 0.196*"digital" + 0.185*"corpo" + -0.175*"dados" + -0.144*"cinema" + -0.136*"transparência" + 0.130*"organização" + 0.119*"web" + 0.112*"aborto" + 0.112*"movimento";
Tópico 28: 0.213*"conhecimento" + 0.164*"pessoas" + 0.148*"comunidades" + 0.137*"pesquisa" + 0.130*"digital" + -0.130*"educação" + 0.123*"comunidade" + -0.116*"sistema" + -0.116*"gestão" + -0.111*"cultura";
Tópico 29: -0.292*"kaxuyana" + 0.214*"qualidade" + 0.191*"transparência" + 0.177*"conhecimento" + 0.169*"memória" + -0.154*"aldeia" + -0.128*"tamiriki" + -0.117*"linguagem" + 0.112*"revistas" + -0.108*"organização".

Fonte: Elaborado pelo autor.

Os termos dos tópicos extraídos por meio do modelo LSI possuem pesos representativos. Em alguns casos, existem termos fortes com pesos acima de 0.500 como “informação” no tópico 0, “museu” no tópico 4, “cinema” no tópico 6 e “dança” no tópico 7. Todos os termos seguem uma ordem de relevância onde o maior peso se refere ao termo que melhor representa o determinado tópico.

Os resultados extraídos nesse *corpus* de dados permitem realizar a unificação entre tópicos mediante a termos e pesos similares contidos em mais de um tópico. Exemplos disso estão nos tópicos 24 e 26 que apresentam os termos “transparência”, “qualidade”, “gestão” e “arquivo”, além dos termos que se correlacionam como “direitos” e “direitos_humanos”, “unesco” e “cidadania”.

Também é possível identificar termos com características particulares que possam ter sido utilizados como campo de estudos na área da Ciência da

Informação, passando assim a serem norteadores para que o especialista busque informações em documentos externos e posteriormente apresente uma suposição de nome para o tópico de forma mais assertiva. Exemplo disso está no termo “maracanã” que aparece nos tópicos 14, 15, 16, 17, 20, 21 e 22 junto a outros termos como “kaxuyana”, “tamiriki”, “aldeia”, “cidadania” e “memória”. Esse conjunto de termos pode remeter a estudos na área de Memória e Aspectos Sociais. Entretanto, esse quantitativo de tópicos abordando assuntos correlacionados possibilita identificar na modelagem de tópicos, realizada através de tentativa e erro com relação ao número ideal de tópicos a serem extraídos que existe um número excessivo de tópicos para o *corpus* de dados.

Também existe dentre os resultados aqueles tópicos que contemplam termos coesos, facilitando assim ao especialista na identificação da suposição do assunto sem consultar fontes externas. Exemplo disso está no tópico 10 que apresenta entre os resultados os termos “livro”, “leitura” e “skoob”, podendo supor que o tópico esteja relacionado a Acesso, Uso e Apropriação da Leitura.

O Quadro 28 apresenta os resultados extraídos do *corpus* 3 através do modelo de treinamento LDA. Trata-se de um conjunto de resultados formado por 30 tópicos, termos e pesos que foram treinados durante 54 minutos e 29 segundos. Os demais conjuntos de resultados com 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[60].

Quadro 28 – Tópicos extraídos do *corpus* 3 usando o modelo LDA

Tópico 0: 0.001*"perfis" + 0.001*"fakes" + 0.001*"celebridades" + 0.001*"twitter" + 0.000*"tweet" + 0.000*"mussumalive" + 0.000*"perfis_fakes" + 0.000*"ato_linguagem" + 0.000*"tweet_dia" + 0.000*"nairbello";
Tópico 1: 0.002*"comunidades" + 0.001*"membros" + 0.001*"senai" + 0.001*"arquivos" + 0.001*"aquisição" + 0.001*"arquivo" + 0.001*"educação_profissional" + 0.001*"arquivos_pessoais" + 0.001*"description" + 0.001*"indústrias";
Tópico 2: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"dados" + 0.000*"forma" + 0.000*"busca" + 0.000*"documentos" + 0.000*"memória" + 0.000*"cinema" + 0.000*"relação" + 0.000*"poder";
Tópico 3: 0.002*"cidadania" + 0.002*"kaxuyana" + 0.002*"sujeitos" + 0.002*"turismo" + 0.001*"favela" + 0.001*"sociais" + 0.001*"tamiriki" + 0.001*"orkutização" + 0.001*"mst" + 0.001*"hospitalidade";
Tópico 4: 0.000*"informação" + 0.000*"memória" + 0.000*"brasil" + 0.000*"social" + 0.000*"movimento" + 0.000*"pesquisa" + 0.000*"processo" + 0.000*"tempo" + 0.000*"forma" + 0.000*"política";

[60] Algoritmo de modelagem de tópicos. *Corpus* 3: teses e dissertações 2014. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2014.ipynb/.

Tópico 5: 0.001*"raros" + 0.001*"livros_raros" + 0.001*"furto" + 0.001*"roubo" + 0.000*"roubo_furto" + 0.000*"crime" + 0.000*"evot" + 0.000*"obras_raras" + 0.000*"criminal" + 0.000*"crimes";
Tópico 6: 0.001*"direitos" + 0.001*"direitos_humanos" + 0.001*"jornalismo" + 0.001*"rede" + 0.001*"contábeis" + 0.001*"openehr" + 0.001*"nuclear" + 0.001*"unesco" + 0.001*"marx" + 0.001*"circulação";
Tópico 7: 0.003*"nuclear" + 0.002*"ontologia" + 0.002*"ontologias" + 0.002*"emergência" + 0.001*"domínio" + 0.001*"aquisição" + 0.001*"fukushima" + 0.001*"museu_histórico" + 0.001*"organização_conhecimento" + 0.001*"museu_histórico_nacional";
Tópico 8: 0.004*"museu" + 0.003*"memória" + 0.003*"cultura" + 0.003*"social" + 0.003*"história" + 0.002*"brasil" + 0.002*"forma" + 0.002*"cultural" + 0.002*"nacional" + 0.002*"anos";
Tópico 9: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"social" + 0.000*"forma" + 0.000*"memória" + 0.000*"trabalho" + 0.000*"cinema" + 0.000*"processo" + 0.000*"brasil" + 0.000*"meio";
Tópico 10: 0.002*"fotografias" + 0.002*"arte" + 0.001*"aborto" + 0.001*"obras_arte" + 0.001*"fotografias_obras" + 0.001*"fotografias_obras_arte" + 0.001*"obras" + 0.001*"acervo" + 0.001*"biblioteca_nacional" + 0.001*"federações";
Tópico 11: 0.001*"indexação" + 0.001*"transparência" + 0.001*"transparência_ativa" + 0.001*"indexação_automática" + 0.001*"automática" + 0.001*"obino" + 0.001*"ativa" + 0.001*"padrões_indicadores" + 0.000*"indicadores_transparência" + 0.000*"padrões_indicadores_transparência";
Tópico 12: 0.001*"baixa" + 0.001*"alta" + 0.001*"mpd" + 0.001*"ex-alunos" + 0.001*"baixa_baixa" + 0.001*"alta_baixa" + 0.001*"baixa_alta" + 0.000*"alta_alta" + 0.000*"difusa" + 0.000*"ifmg";
Tópico 13: 0.001*"produção" + 0.001*"regimes" + 0.001*"capes" + 0.001*"ris" + 0.001*"aberto" + 0.001*"dados" + 0.001*"colaboração" + 0.001*"periódicos" + 0.001*"científica" + 0.001*"data";
Tópico 14: 0.000*"biblioteca" + 0.000*"informação" + 0.000*"pesquisa" + 0.000*"escolar" + 0.000*"conhecimento" + 0.000*"social" + 0.000*"trabalho" + 0.000*"escola" + 0.000*"educação" + 0.000*"sociedade";
Tópico 15: 0.000*"informação" + 0.000*"conhecimento" + 0.000*"pesquisa" + 0.000*"documentos" + 0.000*"patrimônio" + 0.000*"dados" + 0.000*"forma" + 0.000*"biblioteca" + 0.000*"cultura" + 0.000*"cultural";
Tópico 16: 0.001*"cristo" + 0.001*"fem" + 0.000*"papai" + 0.000*"papai_noel" + 0.000*"noel" + 0.000*"cristo_redentor" + 0.000*"correios" + 0.000*"redentor" + 0.000*"missivistas" + 0.000*"correios_telégrafos";
Tópico 17: 0.001*"vagas" + 0.001*"reservadas" + 0.001*"vagas_reservadas" + 0.001*"candidatos" + 0.001*"cotas" + 0.000*"percentual" + 0.000*"requisitos" + 0.000*"inscritos" + 0.000*"matriculados" + 0.000*"renda";
Tópico 18: 0.000*"informação" + 0.000*"biblioteca" + 0.000*"pesquisa" + 0.000*"cinema" + 0.000*"sistema" + 0.000*"forma" + 0.000*"busca" + 0.000*"dados" + 0.000*"usuários" + 0.000*"memória";
Tópico 19: 0.002*"multimídia" + 0.002*"ontologias" + 0.002*"domínio" + 0.002*"mpeg7" + 0.002*"ontologia" + 0.002*"escopo" + 0.002*"domínio_escopo" + 0.002*"conteúdo" + 0.001*"metadados" + 0.001*"mpeg-7";
Tópico 20: 0.000*"informação" + 0.000*"brics" + 0.000*"pesquisa" + 0.000*"científica" + 0.000*"comunicação" + 0.000*"trabalho" + 0.000*"dados" + 0.000*"biblioteca" + 0.000*"processo" + 0.000*"brasil";
Tópico 21: 0.000*"informação" + 0.000*"conhecimento" + 0.000*"processo" + 0.000*"forma" + 0.000*"termo" + 0.000*"pesquisa" + 0.000*"social" + 0.000*"uso" + 0.000*"mundo" + 0.000*"organização";

Tópico 22: 0.002*"cito" + 0.001*"influencia" + 0.001*"citação_cito" + 0.001*"voc-ien" + 0.001*"publicação_ampliada" + 0.001*"ampliada" + 0.001*"influencia_cito" + 0.001*"citação" + 0.000*"influencia_voc-ien" + 0.000*"derivação";
Tópico 23: 0.020*"informação" + 0.005*"pesquisa" + 0.005*"conhecimento" + 0.004*"dados" + 0.003*"biblioteca" + 0.003*"forma" + 0.003*"processo" + 0.002*"uso" + 0.002*"organização" + 0.002*"sistema";
Tópico 24: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"dados" + 0.000*"conhecimento" + 0.000*"social" + 0.000*"comunicação" + 0.000*"relação" + 0.000*"processo" + 0.000*"sistema" + 0.000*"base";
Tópico 25: 0.003*"maracanã" + 0.002*"estádio" + 0.002*"entrevistadora" + 0.001*"futebol" + 0.001*"copa" + 0.001*"estádio_maracanã" + 0.000*"copa_mundo" + 0.000*"jayme" + 0.000*"lembro" + 0.000*"flamengo";
Tópico 26: 0.001*"sns" + 0.001*"scopus" + 0.001*"medellín" + 0.001*"tropical" + 0.001*"web_science" + 0.001*"quitandeiras" + 0.001*"medicina_tropical" + 0.001*"roa" + 0.001*"favelas" + 0.000*"medicina";
Tópico 27: 0.001*"arquivologia" + 0.001*"segurança_informação" + 0.001*"segurança" + 0.001*"cpi" + 0.001*"deputado" + 0.001*"transparência" + 0.001*"transparência_informação" + 0.001*"arquivístico" + 0.001*"davis" + 0.001*"ostwald";
Tópico 28: 0.001*"folclore" + 0.000*"celta" + 0.000*"museu_folclore" + 0.000*"celtismo" + 0.000*"celtas" + 0.000*"edison_carneiro" + 0.000*"folclore_edison" + 0.000*"folclore_edison_carneiro" + 0.000*"museu_folclore_edison" + 0.000*"irlanda";
Tópico 29: 0.000*"informação" + 0.000*"conhecimento" + 0.000*"processo" + 0.000*"pesquisa" + 0.000*"forma" + 0.000*"trabalho" + 0.000*"museu" + 0.000*"social" + 0.000*"relação" + 0.000*"meio".

Fonte: Elaborado pelo autor.

Os tópicos extraídos do corpus de dados utilizando o modelo LDA são representados por termos que se caracterizam por possuírem um maior número de N-gramas do tipo bigramas e trigramas e um menor quantitativo de termos generalistas ao domínio da linguagem estudado quando se comparado ao modelo LSI. Em contrapartida, os resultados apresentam poucos termos com pesos representativos e a maioria deles com valor igual a 0, o que dificulta na interpretação dos tópicos por parte do especialista, uma vez que não passa a não existir uma ordem de relevância entre os termos.

O tópico 0 apresenta termos especialistas como “perfis_fakes”, “twitter” e “celebridades”, onde, mesmo não existindo em seus resultados termos generalistas como “informação”, possibilita ao especialista através da análise de assuntos supor por exemplo que o tópico esteja relacionado a Informação e Tecnologia. O tópico 1 apresenta termos como “comunidades” e “membros” que quando analisados sozinhos, podem apresentar composições de significados nos termos, entretanto, a composição do resultado com termos como “arquivos_pessoais”, “educação_profissional” e os termos especialistas “indústrias” e “senai” possibilita ao especialista inferir um rótulo para o tópico ou

utilizar de técnicas de análise de assunto como explorar o *corpus* de dados para apresentar uma suposição mais assertiva. Nesse caso, pode-se supor por exemplo que o tópico esteja relacionado a Gestão da Informação ou Gestão da Informação em Indústrias.

Também é possível encontrar entre os resultados tópicos fracos contendo termos fortes, como por exemplo o tópico 7 que apresenta termos como “ontologias” e “domínio” podendo ser interpretado pelo especialista como um tópico referente a Organização e Representação do Conhecimento, entretanto, também constam no tópico termos especialista como “fukushima”, “nuclear” e “museu_histórico_nacional” que remetem a pesquisa na área de Ciência da Informação ou Controvérsia Científica quando analisado o *corpus* de dados, não existindo assim quaisquer relações entre os resultados.

APÊNDICE D - *Corpus* 4: teses e dissertações 2015

O *corpus* 4 também constitui o primeiro *corpora* de dados contempla o quantitativo de 315 documentos do tipo teses e dissertações defendidas no ano de 2015 e um tamanho de 109.601kb. O *corpus* quando convertidos em N-gramas resultou em 7.561.807 unigramas, 7.561.492 bigramas e 7.561.177 trigramas. O Quadro 29 apresenta uma lista com os 50 termos mais frequentes separados por cada tipo de N-gramas.

Quadro 29 – Lista de N-gramas por ordem de frequência do *corpus* 4

Unigramas
informação,102184; pesquisa,35136; conhecimento,27656; biblioteca,27265; forma,23373; social,20699; processo,20414; dados,19571; relação,17526; trabalho,17494; comunicação,16716; uso,15521; brasil,15508; meio,14963; memória,14941; produção,14668; organização,14625; termo,14473; museu,14393; documentos,13903; desenvolvimento,13627; sociedade,13165; sociais,13111; educação,12581; tempo,12528; pessoas,12319; gestão,12123; sistema,11891; fonte,11616; cultura,11475; paulo,11165; estudo,11010; anos,10944; nacional,10832; universidade,10662; busca,10543; história,10491; contexto,10336; mundo,10158; espaço,9998; grupo,9981; rede,9980; atividades,9711; vida,9637; relações,9456; construção,9410; estudos,9386; grande,9137; caso,9093; científica,9031.
Bigramas
universidade_federal,3697; recuperação_informação,2703; redes_sociais,2631; produção_científica,2355; uso_informação,2329; informação_conhecimento,2252; gestão_informação,2216; muitas_vezes,1957; dados_pesquisa,1957; ponto_vista,1915; belo_horizonte,1859; organização_conhecimento,1779; informação_tecnologia,1696; porto_alegre,1634; coleta_dados,1584; dissertação_mestrado,1571; informação_comunicação,1517; livros_eletrônicos,1470; gestão_conhecimento,1425; minas_gerais,1422; biblioteca_universitárias,1419; arquitetura_informação,1405; patrimônio_cultural,1384; fontes_informação,1368; sistemas_informação,1364; organização_informação,1360; fonte_elaborado,1348; educação_distância,1336; ensino_superior,1335; fonte_dados,1304; bases_dados,1297; informação_science,1283; documentos_arquivísticos,1245; base_dados,1232; tendo_vista,1213; comunicação_científica,1212; tecnologias_informação,1198; sociedade_informação,1191; memória_social,1141; busca_informação,1112; produtos_serviços,1094; informação_informação,1088; espírito_santo,1066; dispositivos_comunicação,1025; competência_informação,1011; tecnologia_informação,973; unidades_informação,972; universidade_estadual,971; teses_dissertações,945; biblioteca_escolar,941.
Trigramas
fonte_dados_pesquisa,1222; documentos_arquivísticos_digitais,832; fonte_elaborado_autora,779; comunicação_web_social,737; tecnologias_informação_comunicação,727; dispositivos_comunicação_web,690; sistemas_organização_conhecimento,661; international_organization_standardization,526; informação_universidade_federal,465; dissertação_mestrado_informação,446; universidade_federal_grande,446; fundação_oswaldo_cruz,431; federal_minas_gerais,421; acute_myeloid_leukemia,409; federal_grande_sul,408; universidade_federal_bahia,404; patrimônio_histórico_artístico,390; gestão_informação_conhecimento,387; instituições_ensino_superior,380; doc_doc_doc,379; universidade_federal_santa,365; universidade_federal_minas,349; informação_belo_horizonte,345;

preservação_documentos_arquivísticos,342;	federal_santa_catarina,331;	
busca_uso_informação,317;	instituto_oswaldo_cruz,317;	histórico_artístico_nacional,311;
american_society_informação,306;	universidade_federal_paraíba,305;	
society_informação_science,304;	instituto_benjamin_constant,300;	sites_redes_sociais,285;
american_library_association,270;	encontro_nacional_pesquisa,267;	
nacional_pesquisa_informação,265;	journal_american_society,263;	
datagramazero_revista_informação,262;	grupos_quadrilhas_juninas,258;	
brasília_briquet_lemos,256;	documento_arquivístico_digital,249;	
membros_equipes_trabalho,247;	requisito_parcial_obtenção,245;	
museu_histórico_nacional,242;	morri_sunga_branca,242;	
informação_science_technology,239;	resource_description_framework,237;	
organização_informação_organização,233;	anos_anos_anos,230;	
informação_organização_conhecimento,226.		

Fonte: Elaborado pelo autor.

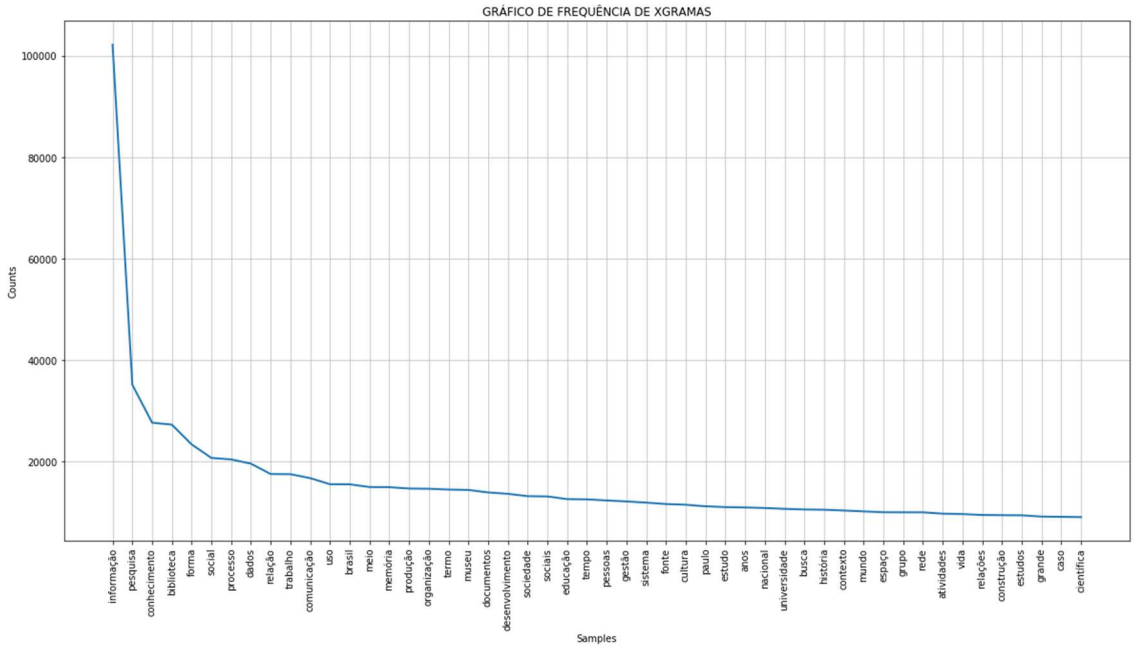
O algoritmo gerou uma lista contendo mil N-gramas mais frequentes de cada tipo, sendo unigramas, bigramas, trigramas e uma lista geral com todos as categorias para observar o comportamento dos termos. O bigrama mais frequente é “universidade_federal” com frequência de 3.697 ocupando a posição 270, seguido de “recuperação_informação” com frequência de 2.703 na posição 410, “redes_sociais” com frequência de 2.631 e posição 426, “produção_científica” com frequência de 2.355 e posição 500 e “uso_informação” com frequência de 2.329 e posição 510. Já entre os primeiros trigramas estão “fonte_dados_pesquisa” com frequência de 1.222, “documentos_arquivísticos_digitais” com frequência de 832, “fonte_elaborado_autora” com frequência de 779, “comunicação_web_social” com frequência de 737 e “tecnologias_informação_comunicação” com frequência de 727, entretanto, todos os trigramas estão acima do milésimo termo da listagem geral de N-gramas.

Os unigramas possuem maior frequência entre os demais tipos de N-gramas. Quando comparado o unigrama “informação” que possui maior frequência, sendo 102.184 repetições, com o segundo bigrama mais frequente e com maior representividade, sendo “recuperação_informação” com 2.703 repetições, encontra-se uma diferença de 3.680%. Esse percentual atinge 9.262% quando comparado o primeiro unigrama com o primeiro trigrama “fonte_dados_pesquisa” que possui frequência de 834.

O Gráfico 55 apresenta os 50 N-gramas mais frequentes extraídos do *corpus* 4. Todos os termos são caracterizados como unigramas, sendo, os mais frequentes, “informação” com frequência de 102.184, seguido de “pesquisa” com

frequência de 35.136, “conhecimento” com frequência de 27.656, “biblioteca” com frequência de 27.265 e “forma” com frequência de 23.373.

Gráfico 55 – Termos mais frequentes do *corpus* 4



Fonte: Elaborado pelo autor.

Torna-se possível observar uma evolução na frequência de forma contínua do 50º ao 6º termo e um salto no quantitativo de frequência entre os cinco primeiros termos. Também nesse gráfico é possível encontrar termos fortes como “organização”, “memória”, “uso” e “gestão” e termos fracos como “contexto”, “grande”, “anos” e “paulo”, entretanto, faz-se necessário analisar os resultados dos bigramas e trigramas para obter inferências mais assertivos, uma vez que esses termos podem possuir diversas composições com diferentes significados.

Com base na frequência dos termos, a Figura 16 apresenta uma nuvem de palavras contendo 250 N-gramas do *corpus* 4. Faz-se necessário ressaltar que não foi realizado nenhum tipo de tratamento qualitativo com relação a seleção, descartes ou substituição dos termos.

Figura 16 – Nuvem de palavras do *corpus* 4.

Fonte: Elaborado pelo autor.

Como resultado, a figura apresenta termos fortes e fracos quando se comparado ao domínio de linguagem e realizado a análise de assunto. Os termos “forma”, “espaço” e “e tempo” são considerados fortes quando realizado análise de assunto através dos bigramas e trigramas, apresentando assim vertentes dos termos condizentes com o domínio da linguagem. Já os termos “ano”, “dia” e “hoje” são considerados termos fracos, pois não estão relacionados do domínio da linguagem estudada ou não apresentam relevância.

A conexão dos dados com os modelos de treinamentos permitiu extrair um conjunto de tópicos formados por termos e pesos do *corpus* 4. O Quadro 30 apresenta os resultados de 40 tópicos obtidos através do modelo LSI em um tempo de 15 minutos e 1 segundo. Os demais resultados contendo os conjuntos de 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[61].

Quadro 30 – Tópicos extraídos do *corpus* 4 usando o modelo LSI

Tópico 0: 0.640*"informação" + 0.176*"biblioteca" + 0.172*"pesquisa" + 0.163*"conhecimento" + 0.107*"forma" + 0.105*"processo" + 0.104*"social" + 0.098*"comunicação" + 0.097*"dados" + 0.094*"trabalho";
Tópico 1: -0.561*"informação" + 0.365*"museu" + 0.206*"biblioteca" + 0.127*"patrimônio" + 0.127*"memória" + 0.107*"cultural" + 0.106*"brasil" + 0.103*"história" + 0.094*"cultura" + 0.093*"nacional";

[61] Algoritmo de modelagem de tópicos. *Corpus* 4: teses e dissertações 2015. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2015.ipynb/.

Tópico 2: -0.809*"biblioteca" + 0.181*"museu" + -0.120*"usuários" + -0.107*"livros" + -0.106*"leitura" + 0.104*"termo" + 0.100*"documentos" + -0.093*"livro" + -0.086*"dispositivos" + -0.071*"dispositivos_comunicação";
Tópico 3: 0.445*"documentos" + -0.369*"museu" + 0.230*"documento" + 0.203*"digitais" + 0.198*"arquivísticos" + 0.188*"documentos_arquivísticos" + -0.188*"informação" + 0.184*"digital" + 0.170*"preservação" + 0.146*"arquivísticos_digitais";
Tópico 4: 0.480*"museu" + 0.269*"documentos" + 0.207*"informação" + -0.181*"pesquisa" + -0.178*"termo" + 0.136*"preservação" + 0.134*"documento" + 0.127*"arquivísticos" + 0.125*"arquivo" + 0.119*"documentos_arquivísticos";
Tópico 5: 0.388*"poder" + -0.351*"museu" + 0.211*"trabalho" + -0.183*"pesquisa" + 0.153*"liderança" + 0.143*"organizacional" + 0.124*"pessoas" + -0.115*"científica" + -0.109*"termo" + -0.105*"universidades";
Tópico 6: 0.423*"termo" + 0.206*"knee" + 0.175*"organização" + 0.173*"conhecimento" + 0.155*"biblioteca" + 0.138*"museu" + 0.135*"ontologia" + 0.134*"definição" + -0.125*"pesquisa" + -0.125*"gente";
Tópico 7: 0.306*"poder" + 0.228*"universidades" + 0.180*"pesquisa" + 0.150*"científica" + 0.149*"liderança" + 0.144*"trabalho" + 0.142*"ranking" + 0.138*"organizacional" + 0.126*"produção" + 0.117*"universidade";
Tópico 8: 0.333*"comunicação" + -0.285*"gente" + 0.234*"dispositivos" + 0.228*"social" + 0.201*"dispositivos_comunicação" + -0.193*"escola" + 0.176*"web" + 0.165*"usuários" + 0.154*"web_social" + 0.147*"comunicação_web";
Tópico 9: -0.355*"gente" + -0.245*"conhecimento" + -0.241*"escola" + -0.197*"gestão" + 0.189*"livros" + 0.167*"informação" + 0.147*"moda" + -0.147*"escolar" + -0.132*"alunos" + -0.127*"educação";
Tópico 10: 0.236*"knee" + 0.214*"termo" + 0.184*"memória" + 0.167*"universidades" + -0.161*"ontologia" + -0.156*"museu" + 0.147*"cultural" + -0.140*"pesquisa" + 0.135*"patrimônio" + -0.122*"moda";
Tópico 11: 0.242*"knee" + -0.217*"ontologia" + -0.210*"conhecimento" + 0.179*"busca" + 0.163*"termo" + 0.152*"rede" + 0.147*"museu" + -0.135*"ontologias" + -0.133*"definição" + 0.124*"dados";
Tópico 12: 0.271*"rede" + 0.242*"pesquisa" + -0.231*"universidades" + 0.189*"redes" + -0.174*"moda" + -0.160*"ranking" + 0.152*"memória" + 0.145*"pesquisadores" + 0.144*"cultural" + 0.129*"científica";
Tópico 13: 0.410*"conhecimento" + 0.296*"moda" + 0.206*"gestão" + 0.162*"organização" + 0.147*"revistas" + 0.142*"revista" + -0.141*"usuários" + -0.122*"ontologia" + -0.115*"web" + -0.102*"pesquisa";
Tópico 14: -0.239*"universidades" + 0.226*"moda" + -0.210*"rede" + -0.174*"ranking" + 0.152*"escola" + -0.149*"redes" + -0.144*"histórias" + 0.141*"gente" + 0.131*"patrimônio" + 0.131*"pesquisa";
Tópico 15: -0.291*"memória" + -0.276*"conhecimento" + -0.205*"livros" + 0.199*"brasil" + 0.163*"rede" + -0.150*"arte" + 0.136*"educação" + -0.109*"eletrônicos" + 0.109*"jornalismo" + -0.105*"livros_eletrônicos";
Tópico 16: -0.270*"rede" + 0.208*"educação" + -0.180*"redes" + -0.178*"cultural" + -0.170*"gente" + -0.142*"moda" + -0.139*"patrimônio" + 0.128*"pesquisa" + -0.118*"publicações" + -0.118*"histórias";
Tópico 17: 0.302*"memória" + 0.246*"moda" + 0.214*"histórias" + 0.206*"educação" + -0.136*"livros" + -0.123*"científica" + 0.117*"profissional" + -0.116*"dados" + 0.113*"formação" + -0.112*"eletrônicos";
Tópico 18: -0.317*"patrimônio" + 0.284*"memória" + -0.238*"cultural" + -0.192*"livros" + -0.134*"eletrônicos" + 0.132*"museu" + -0.124*"livros_eletrônicos" + -0.121*"bens" + -0.115*"educação" + 0.114*"biblioteca";

Tópico 19: -0.297*"museologia" + 0.256*"livros" + 0.175*"eletrônicos" + 0.168*"livros_eletrônicos" + -0.152*"objeto" + 0.141*"memória" + -0.124*"biblioteca" + -0.123*"patrimônio" + -0.111*"filosofia" + -0.106*"sns";
Tópico 20: 0.342*"sns" + 0.186*"rede" + 0.186*"doc" + 0.169*"indexação" + 0.161*"direito" + 0.160*"descritores" + 0.151*"doc_doc" + -0.147*"jongo" + 0.128*"redes" + 0.123*"doc_doc_doc";
Tópico 21: -0.220*"museologia" + -0.205*"memória" + -0.202*"leukemia" + -0.166*"with" + -0.164*"acute" + -0.157*"myeloid" + 0.155*"jongo" + -0.143*"livros" + -0.119*"myeloid_leukemia" + -0.116*"acute_myeloid";
Tópico 22: 0.297*"sns" + 0.239*"jongo" + 0.171*"leukemia" + 0.163*"doc" + -0.141*"memória" + 0.138*"acute" + 0.133*"myeloid" + 0.133*"histórias" + 0.131*"doc_doc" + 0.125*"descritores";
Tópico 23: -0.241*"imagem" + -0.214*"imagens" + -0.174*"leukemia" + 0.173*"jongo" + 0.156*"memória" + -0.141*"acute" + -0.135*"myeloid" + -0.133*"with" + -0.123*"arte" + -0.120*"fotografia";
Tópico 24: 0.310*"histórias" + -0.233*"jongo" + 0.218*"arte" + -0.206*"memória" + -0.137*"pessoas" + 0.126*"jornalismo" + -0.116*"educação" + -0.111*"dados" + 0.108*"narrativa" + -0.104*"roda";
Tópico 25: 0.224*"jongo" + -0.218*"memória" + 0.212*"museologia" + -0.207*"patrimônio" + 0.182*"arquivos" + 0.168*"arquivo" + -0.149*"escola" + 0.113*"documentos" + -0.110*"bens" + 0.098*"lei";
Tópico 26: -0.219*"histórias" + 0.212*"grupos" + -0.167*"jongo" + 0.167*"grupo" + -0.147*"dados" + -0.139*"repositórios" + 0.124*"social" + -0.111*"aberto" + -0.111*"celebridades" + 0.101*"café";
Tópico 27: 0.168*"jongo" + -0.149*"pesquisa" + 0.142*"educação" + -0.141*"memória" + -0.127*"projeto" + 0.115*"científica" + -0.112*"sns" + -0.109*"café" + -0.109*"instituição" + -0.105*"fomento";
Tópico 28: -0.202*"ontologia" + 0.198*"educação" + -0.150*"dados" + 0.133*"conceito" + -0.133*"museologia" + -0.124*"busca" + 0.123*"jongo" + 0.114*"comunicação" + -0.113*"ontologias" + 0.110*"café";
Tópico 29: 0.219*"jornalismo" + 0.195*"economia" + 0.194*"jongo" + -0.186*"arte" + -0.171*"social" + 0.124*"verde" + 0.123*"ambiental" + -0.116*"direito" + 0.113*"ambiente" + -0.104*"obras".

Fonte: Elaborado pelo autor.

É possível identificar através dos resultados obtidos por meio do modelo LSI termos fortes com pesos acima de 0.500 como “informação” no tópico 0 e “informação” e “biblioteca” nós tópicos 1 e 2, entretanto, com valores negativos. Os termos de cada tópico seguem uma ordem de relevância mediante os pesos atribuídos. O tópico 0 apresentar termos generalistas como “processo”, “social” e “forma” referente ao domínio da linguagem, sendo necessário ao especialista utilizar de recursos extremos como lista de bigramas, trigramas ou mesmo acesso ao *corpus* de dados para realizar uma suposição mais assertiva de um nome para o tópico. Uma outra possibilidade a ser analisada pelo especialista está em criar uma categoria geral de assuntos ou mesmo descartar o tópico se necessário.

Os tópicos 3 e 4 constituem de termos similares como “documentos” e “museu” sendo os primeiros tópicos de cada termo, porém, com posições e pesos invertidos. Além disso os tópicos possuem os termos “documentos_arquivísticos”, “preservação” e “informação”, podendo ao especialista por meio de análise de assuntos supor por exemplo que os tópicos estejam relacionados a Gestão da Informação ou Gestão de Documentos.

Já os tópicos 6, 10 e 11 possuem termos em comum como “ontologia”, “museu” e “termo”, além de termos correlacionados como “dado”, “organização” e “conhecimento” e um termo específico, sendo “knee”. Esse conjunto de tópicos permite ao especialista supor por exemplo que o assunto esteja relacionado a área de Organização da Informação e do Conhecimento.

Dentre os resultados também é possível encontrar termos fracos como “doc” ou “doc_doc” que não possue significância para o domínio da linguagem bem como tópicos com termos fortes mas considerados fracos por apresentar uma miscelânea de assuntos como nos tópicos 21, 22 e 23 com os termos “leukemia”, “myeloid” e “acute” que fazem parte do vocabulário Biomédico estudado como Representação da Informação, entretanto constam também termos como “museologia”, “memória”, “jongo”, “arte” e “fotografia“ que permitem supor por exemplo temas como Museu, Patrimônio e Informação ou Informação e Memória. Tópicos com essas características apontam no *corpus* de dados para mais de um documento que melhor representa o conjunto de termos.

O Quadro 31 apresenta os resultados obtidos através de modelo de extração LDA que resultou em um conjunto de 30 tópicos formados por termos e pesos. O treinamento desse modelo foi realizado em 50 minutos e 53 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponíveis através do GitHub[62].

Quadro 31 – Tópicos extraídos do *corpus* 4 usando o modelo LDA

Tópico 0: 0.003*"universidades" + 0.002*"ranking" + 0.001*"top" + 0.001*"usp" + 0.001*"classificadas" + 0.001*"rankings" + 0.001*"posição" + 0.001*"universidades_brasileiras" + 0.001*"brasileiras" + 0.001*"unicamp";
Tópico 1: 0.001*"muf" + 0.001*"graffiti" + 0.001*"favela" + 0.001*"paraty" + 0.001*"fernanda_rodrigues" + 0.001*"casas-tela" + 0.000*"arquivo.mp3" + 0.000*"fernanda" + 0.000*"museu_favela" + 0.000*"entrevistadora_fernanda_rodrigues";

[62] Algoritmo de modelagem de tópicos. *Corpus* 4: teses e dissertações 2015. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2015.ipynb/.

Tópico 2: 0.011*"informação" + 0.004*"pesquisa" + 0.003*"conhecimento" + 0.003*"biblioteca" + 0.002*"forma" + 0.002*"social" + 0.002*"processo" + 0.002*"dados" + 0.002*"relação" + 0.002*"trabalho";
Tópico 3: 0.001*"livros_eletrônicos" + 0.001*"nise" + 0.001*"palavras\" + 0.001*"loucura" + 0.000*"alagoas" + 0.000*"arte_bruta" + 0.000*"bruta" + 0.000*"livro_eletrônico" + 0.000*"imagens_inconsciente" + 0.000*"folclore";
Tópico 4: 0.002*"cinema" + 0.002*"samba" + 0.001*"laura" + 0.001*"imagem" + 0.001*"alvim" + 0.001*"teatro" + 0.001*"benjamin" + 0.001*"laura_alvim" + 0.001*"ferrovia" + 0.001*"filme";
Tópico 5: 0.000*"memória" + 0.000*"informação" + 0.000*"biblioteca" + 0.000*"conhecimento" + 0.000*"social" + 0.000*"pesquisa" + 0.000*"meio" + 0.000*"história" + 0.000*"forma" + 0.000*"tempo";
Tópico 6: 0.002*"celebridades" + 0.001*"blogs" + 0.001*"dou_dado" + 0.001*"lele" + 0.001*"celebridade" + 0.001*"morri" + 0.001*"sunga" + 0.001*"dou" + 0.001*"screenshot" + 0.001*"sunga_branca";
Tópico 7: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"museu" + 0.000*"campo" + 0.000*"conhecimento" + 0.000*"forma" + 0.000*"social" + 0.000*"cultural" + 0.000*"campo_grande" + 0.000*"documentos";
Tópico 8: 0.001*"scopus" + 0.001*"journal" + 0.000*"storytelling" + 0.000*"tablet" + 0.000*"icb/ufmg" + 0.000*"web_science" + 0.000*"departamento" + 0.000*"dados_web" + 0.000*"nota_dados_retirados" + 0.000*"autora_nota_dados";
Tópico 9: 0.002*"dispositivos_comunicação" + 0.002*"dispositivos" + 0.002*"web_social" + 0.002*"otlet" + 0.002*"comunicação_web" + 0.002*"comunicação_web_social" + 0.002*"dispositivos_comunicação_web" + 0.001*"charge" + 0.001*"web" + 0.001*"usuários";
Tópico 10: 0.002*"leukemia" + 0.001*"ambiental" + 0.001*"economia" + 0.001*"acute" + 0.001*"verde" + 0.001*"myeloid" + 0.001*"with" + 0.001*"economia_verde" + 0.001*"myeloid_leukemia" + 0.001*"acute_myeloid";
Tópico 11: 0.001*"quadrilhas" + 0.001*"pdll" + 0.001*"deleuze" + 0.001*"juninas" + 0.001*"quadrilhas_juninas" + 0.001*"grupos_quadrilhas" + 0.001*"livro_leitura" + 0.001*"pulsão" + 0.001*"grupos_quadrilhas_juninas" + 0.001*"pulsões";
Tópico 12: 0.002*"museologia" + 0.001*"documental" + 0.001*"stránský" + 0.001*"indicador" + 0.001*"nietzsche" + 0.001*"marabá" + 0.001*"terminologia" + 0.001*"ibidem" + 0.000*"gonzaga" + 0.000*"percentual";
Tópico 13: 0.011*"museu" + 0.002*"museologia" + 0.001*"jardim" + 0.001*"rodoviário" + 0.001*"escolares" + 0.001*"museu_escolares" + 0.001*"museu_rodoviário" + 0.001*"botânico" + 0.001*"museólogo" + 0.001*"jardins";
Tópico 14: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"conhecimento" + 0.000*"social" + 0.000*"organização" + 0.000*"rede" + 0.000*"comunicação" + 0.000*"memória" + 0.000*"forma" + 0.000*"uso";
Tópico 15: 0.003*"jongo" + 0.001*"roda" + 0.001*"lapa" + 0.001*"moda" + 0.001*"campo_grande" + 0.001*"jongo_lapa" + 0.000*"serrinha" + 0.000*"tambor" + 0.000*"nscg" + 0.000*"jongueira";
Tópico 16: 0.004*"termo" + 0.003*"knee" + 0.001*"arthroplasty" + 0.001*"replacement" + 0.001*"equivalência" + 0.001*"tesauro" + 0.001*"busca" + 0.001*"total" + 0.001*"janeway" + 0.001*"mesh";
Tópico 17: 0.000*"exames" + 0.000*"laboratoriais" + 0.000*"hepatites" + 0.000*"regras_associação" + 0.000*"hepatites_virais" + 0.000*"kdd" + 0.000*"virais" + 0.000*"exames_laboratoriais" + 0.000*"hepatite" + 0.000*"testes_laboratoriais";
Tópico 18: 0.001*"nivalson" + 0.001*"nivalson_miranda" + 0.000*"suellen_barbosa" + 0.000*"suellen" + 0.000*"miranda" + 0.000*"barbosa_galdino" + 0.000*"suellen_barbosa_galdino" + 0.000*"galdino" + 0.000*"bico_pena" + 0.000*"bico";

Tópico 19: 0.001*"livros_eletrônicos" + 0.001*"compra" + 0.001*"eletrônicos" + 0.000*"vendedor" + 0.000*"livros" + 0.000*"vendedores" + 0.000*"pacote" + 0.000*"perpétuo" + 0.000*"compra_livros" + 0.000*"livro_eletrônico";
Tópico 20: 0.002*"sns" + 0.001*"doc" + 0.001*"fae" + 0.001*"doc_doc" + 0.001*"colaboração" + 0.001*"descritores" + 0.001*"doc_doc_doc" + 0.001*"científica" + 0.001*"colaboração_científica" + 0.001*"coautores";
Tópico 21: 0.001*"open" + 0.001*"oer" + 0.001*"quadrinhos" + 0.001*"educational" + 0.001*"resources" + 0.001*"open_educational" + 0.001*"educational_resources" + 0.001*"open_educational_resources" + 0.001*"depósito_produção_científica" + 0.001*"depósito_produção";
Tópico 22: 0.005*"patrimônio" + 0.002*"bens" + 0.002*"patrimônio_cultural" + 0.002*"iphan" + 0.001*"tombamento" + 0.001*"ouro" + 0.001*"monumentos" + 0.001*"ouro_preto" + 0.001*"igreja" + 0.001*"cultural";
Tópico 23: 0.005*"documentos" + 0.002*"arquivísticos" + 0.002*"digital" + 0.002*"documentos_arquivísticos" + 0.002*"digitais" + 0.002*"documento" + 0.002*"preservação" + 0.002*"arquivo" + 0.002*"arquivísticos_digitais" + 0.002*"documentos_arquivísticos_digitais";
Tópico 24: 0.003*"moda" + 0.001*"livros_eletrônicos" + 0.001*"manifestações" + 0.001*"repórter" + 0.001*"revistas" + 0.001*"elle" + 0.001*"vogue" + 0.001*"eletrônicos" + 0.000*"manifestantes" + 0.000*"editorial";
Tópico 25: 0.001*"portela" + 0.000*"samba" + 0.000*"portelense" + 0.000*"desfile" + 0.000*"portelenses" + 0.000*"velha_guarda" + 0.000*"escolas_samba" + 0.000*"agremiação" + 0.000*"quadra" + 0.000*"manter_tradição";
Tópico 26: 0.003*"ontologia" + 0.002*"ontologias" + 0.001*"harry" + 0.001*"domínio" + 0.001*"ontology" + 0.001*"potter" + 0.001*"classes" + 0.001*"blood" + 0.001*"harry_potter" + 0.001*"hemonto";
Tópico 27: 0.001*"institutos_federais" + 0.001*"instituto_federal" + 0.001*"telenovela" + 0.001*"institutos" + 0.000*"salve" + 0.000*"salve_jorge" + 0.000*"merchandising" + 0.000*"professores/pesquisadores" + 0.000*"unodc" + 0.000*"merchandising_social";
Tópico 28: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"conhecimento" + 0.000*"forma" + 0.000*"processo" + 0.000*"biblioteca" + 0.000*"relação" + 0.000*"trabalho" + 0.000*"brasil" + 0.000*"fonte";
Tópico 29: 0.000*"sabático" + 0.000*"guinsburg" + 0.000*"editor_livros" + 0.000*"luiz_schwarcz" + 0.000*"roberto_feith" + 0.000*"paulo_grifo" + 0.000*"editores_livros" + 0.000*"jacó_guinsburg" + 0.000*"charles_cosac" + 0.000*"sergio_machado".

Fonte: Elaborado pelo autor.

É possível perceber que os resultados possuem um maior número de N-gramas do tipo bigramas e trigramas quando se comparado ao modelo LSI, o que contribui para uma interpretação mais assertiva por parte do especialista ao realizar a definição da suposição do nome dos tópicos por meio de análise de assunto. Outro ponto a ser destacado nos resultados está na redução de termos generalistas que possibilitam uma composição de significados, como por exemplo o termo “cultura” que pode se compor para “cultura_organizacinal” ou “cultura_informacional”. Um ponto negativo apresentando nos resultados está no baixo valor dos pesos dos termos e em muitos casos, os valores possuem valor

igual a 0, dificultando assim na interpretação do especialista por não apresentar uma ordem de relevância entre os termos.

O tópico 14 é um exemplo de resultado que possui muitos termos generalistas comuns ao domínio de linguagem como “informação”, “conhecimento”, “comunicação”, “memória”, “pesquisa” e “forma” que podem apresentar diversas composições de significados para outros termos conditos no *corpus* de dados. Dessa forma, cabe ao especialista explorar documentos externos como a lista de bigramas e trigramas geradas a partir dos algoritmos ou mesmo utilizar o *corpus* de dados. Para esses casos, o especialista poderá criar uma classificação geral ou mesmo realizar o descarte dos tópicos se achar necessário.

A existência de termos especialistas permite ao especialista realizar a suposição dos nomes dos tópicos de maneira mais assertiva, ou mesmo nortear para que explore os dados externos de maneira mais objetiva, como por exemplo no tópico 1, que apresenta termos chave como “graffiti”, “favela”, “casas_tela” e “museu_favela”, podendo remeter assim ao especialista supor por exemplo que tópico esteja abordando assuntos relacionados a Memória Social ou Museologia Social. O tópico 0 também possui características iguais ao tópico 1, contendo em seus resultados termos como “universidades_brasileiras”, “ranking”, “classificadas” e “posição”, possibilitando assim ao especialista supor que o tópico esteja relacionado a Métricas Informacionais ou Produção Científica.

É possível encontrar entre os resultados tópicos que abordam mais de um assunto, como por exemplo o tópico 8 que possui os termos “storytelling” e "tablet", podendo remeter ao especialista que o tópico esteja associado a Informação e Tecnologia, entretanto, numa exploração dos dados de maneira mais aprofundada como análise realizada no *corpus* de dados é possível perceber que os termos estão associados também a Jornalismo Digital. Outro conjunto de termos contidos no mesmo tópico está “dados_web”, “web_science”, “scorpus”, “nota_dados_retirados” e “icb/ufmg” que possibilita ao especialista supor que o tópico esteja relacionado a Comunicação Científica e/ou Estudos Bibliométricos. Nesse caso, o especialista poderá supor mais de um rótulo para o tópico ou supor um nome para os termos que possuam maior relevância.

APÊNDICE E - *Corpus* 5: teses e dissertações 2016

O *corpus* 5 que constitui o primeiro *corpora* de dados possui 279 documentos do tipo de teses e dissertações defendidas no ano de 2016 e possuindo o tamanho de 101.948kb. A conversão dos N-gramas resultou em 7.064.656 unigramas, 7.064.377 bigramas e 7.064.098 trigramas. O Quadro 32 apresenta uma lista contendo os 50 termos mais frequentes separados por tipo de N-grama.

Quadro 32 – Lista de N-gramas por ordem de frequência do *corpus* 5

Unigramas
informação,95368; pesquisa,31543; conhecimento,25394; forma,21071; dados,19137; biblioteca,17615; social,17283; processo,17282; documentos,15116; relação,15089; trabalho,15031; brasil,14498; uso,13986; meio,13888; comunicação,13815; museu,13637; sistema,13206; produção,13169; sociedade,13135; memória,12963; cultura,12332; organização,12169; desenvolvimento,12129; gestão,12004; nacional,11913; sociais,11633; termo,11633; fonte,11500; tempo,11420; anos,11031; busca,10455; pessoas,10439; paulo,10313; estudo,9818; contexto,9815; usuários,9361; história,9059; vida,8810; atividades,8755; educação,8712; grande,8671; espaço,8646; cultural,8577; mundo,8480; política,8436; quadro,8407; diferentes,8389; modelo,8380; universidade,8329; campo,8319.
Bigramas
universidade_federal,3136; informação_conhecimento,2632; recuperação_informação,2433; belo_horizonte,2127; uso_informação,2034; competência_informação,2028; redes_sociais,1954; gestão_informação,1900; produção_científica,1884; informação_tecnologia,1868; ponto_vista,1864; dados_pesquisa,1825; muitas_vezes,1755; porto_alegre,1639; gestão_documentos,1519; busca_informação,1448; dissertação_mestrado,1442; fonte_dados,1415; coleta_dados,1399; informação_science,1377; ensino_superior,1361; organização_conhecimento,1341; completos_publicados,1335; informação_comunicação,1318; sistemas_informação,1310; publicados_periódicos,1283; minas_gerais,1262; fontes_informação,1259; fonte_elaborado,1237; artigos_completos,1171; tendo_vista,1156; arquitetura_informação,1147; informação_informação,1143; patrimônio_cultural,1138; tecnologias_informação,1126; biblioteca_nacional,1119; novas_tecnologias,1108; gestão_conhecimento,1073; publicados_anais,1067; tomada_decisão,1053; sociedade_informação,1047; anais_congressos,1033; organização_informação,1010; grande_sul,976; memória_social,948; bases_dados,897; tecnologia_informação,895; comunicação_científica,880; citação_web,878; revisão_pares,859.
Trigramas
fonte_dados_pesquisa,1314; artigos_completos_publicados,1168; completos_publicados_periódicos,1160; publicados_anais_congressos,1007; fonte_elaborado_autora,702; tecnologias_informação_comunicação,665; resumos_publicados_anais,558; gestão_informação_conhecimento,490; universidade_federal_grande,450; instituição_ensino_superior,432; international_organization_standardization,431; busca_uso_informação,416; apresentações_trabalho_conferências,396; trabalho_conferências_palestras,396; sistemas_organização_conhecimento,394; informação_universidade_federal,387; federal_grande_sul,387; dissertação_mestrado_informação,366; public_library_science,338; resource_description_framework,330; federal_minas_gerais,329; classificação_decimal_universal,324; universidade_federal_minas,321;

informação_belo_horizonte,319; society_informação_science,318; encontro_nacional_pesquisa,318; nacional_pesquisa_informação,318; universidade_federal_paraíba,315; american_society_informação,313; classificação_decimal_dewey,312; universidade_federal_bahia,308; universidade_federal_santa,296; expandidos_publicados_anais,291; resumos_expandidos_publicados,290; patrimônio_histórico_artístico,289; cultura_arte_barroca,289; museu_histórico_nacional,284; library_science_biology,281; capítulos_livros_publicados,278; sociedade_cem_bibliófilos,275; journal_american_society,274; american_library_association,273; publicados_periódicos_apresentações,270; periódicos_apresentações_trabalho,270; total_fonte_dados,267; ano_produção_prod,266; instituições_ensino_superior,265; federal_santa_catarina,264; cem_bibliófilos_brasil,264; portal_periódicos_capes,252.

Fonte: Elaborado pelo autor.

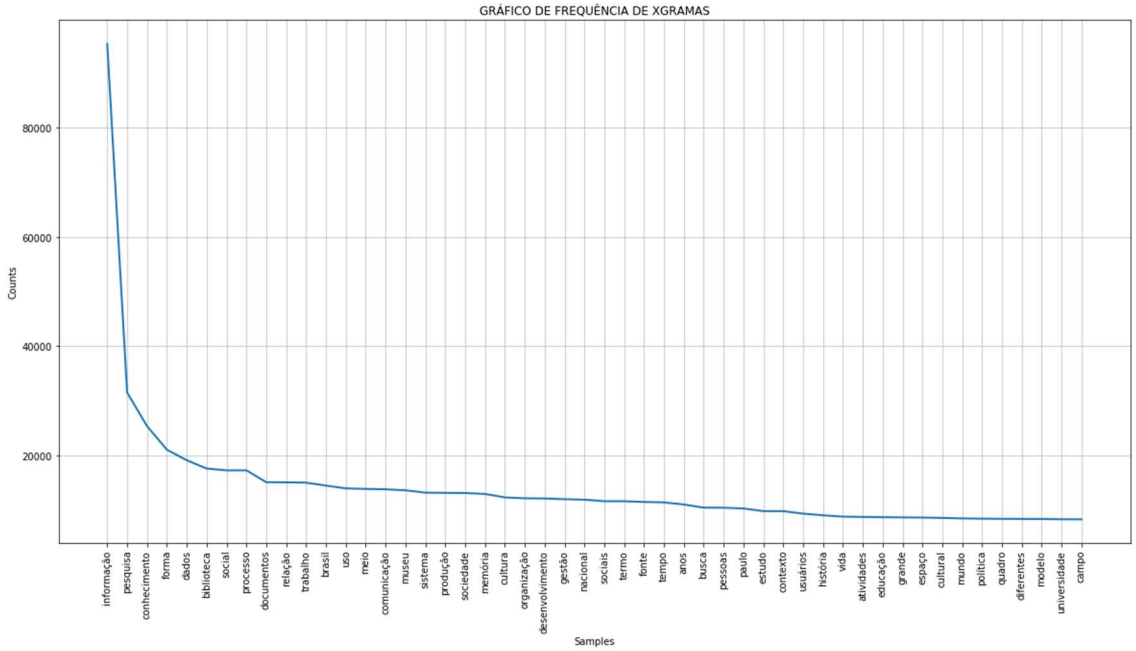
Numa lista contendo os mil N-gramas mais frequentes extraídos do *corpus* de dados, pode-se destacar os bigramas “universidade_federal” com frequência de 3.136 na posição 302, “informação_conhecimento” com frequência de 2.632 na posição 389, “recuperação_informação” com frequência de 2.433 na posição 436, “belo_horizonte” com frequência de 2.127 na posição 528 e “uso_informação” com frequência de 2.034 posição 553 e os trigramas “nacional_pesquisa_informação” com frequência de 198, “encontro_nacional_pesquisa” com frequência de 197, “informação_universidade_federal” com frequência de 189, “modalidade_apresentação_comunicação” com frequência de 166 e “apresentação_comunicação_oralresumo” com frequência de 163, entretanto, nenhum dos trigramas aparecem entre os milésimos termos da lista geral de N-gramas. O trigrama com maior representatividade no domínio de linguagem aparece na sétima posição, sendo “tecnologias_informação_comunicação” com frequência de 133.

Os unigramas possuem maior frequência quando comparado aos demais tipos de N-gramas extraídos do *corpus* de dados. A diferença entre o primeiro unigrama “informação” com frequência de 95.368 para o primeiro bigrama “universidade_federal” com frequência de 3.136 chega a 2.941%. Esse percentual aumenta para 48.006% quando comparado o primeiro unigrama com o primeiro trigrama “nacional_pesquisa_informação” com frequência de 198.

O Gráfico 56 apresenta os 50 termos mais frequentes extraídos do *corpus* 5. Todos os termos desse resultado são do tipo unigramas, estando no topo da lista os termos “informação” com frequência de 95.368, seguido de “pesquisa”

com frequência de 31.543, “conhecimento” com frequência de 25.394, “forma” com frequência de 21.071 e “dados” com frequência de 19.137.

Gráfico 56 – Termos mais frequentes do *corpus* 5



Fonte: Elaborado pelo autor.

Ainda é possível observar no gráfico uma evolução gradual do 50º ao 6º termo onde as frequências partem de 8.319 até 17.615 e um salto de frequências entre os 5 primeiros termos quando se comparado aos demais. Também é possível identificar termos fortes como “processo”, “documentos”, “relação”, “sistema” e “comunicação” que faz-se necessário por parte do especialista no domínio da linguagem explorar os bigramas e trigramas para identificar uma melhor empregabilidade dos termos, uma vez que esses termos possuem composições que possam apontar outras áreas e subáreas. Também constam entre os resultados termos fracos como “anos”, “tempo”, “brasil”, “diferentes” e “grande” podendo estar fora do contexto do *corpus* estudado ou não contribuindo assim para significados de relevância.

A Figura 17 apresenta uma nuvem de palavras contendo os 250 termos mais frequentes extraídos do *corpus* de dados. Os dados são compostos por unigramas, já que os primeiros bigramas e trigramas aparecem somente após o quantitativo de termos utilizando para gerar a figura. Faz-se necessário destacar que a imagem foi construída com dados quantitativos, não sendo realizados quaisquer tratamentos qualitativos.

Figura 17 – Nuvem de palavras do *corpus* 5

Fonte: Elaborado pelo autor.

Ao comparar a imagem com o domínio de linguagem ou realizar a análise de assunto apoiada nos bigramas e trigramas, torna-se possível identificar termos fortes como “espaço”, “usuário” e “conceito”. Mesmo havendo divisão de frequência dos termos quando analisado os bigramas ou trigramas que possuam as mesmas palavras, os termos acabam por apresentar significados de relevância junto ao domínio da linguagem. Também é possível identificar termos fracos como “sempre”, “primeiro” e “cidade” que não remetem significado de valor agregado.

Ao conectar os dados aos modelos de treinamento, foi possível realizar a extração de tópicos do *corpus* de dados, constituídos de termos e pesos. O Quadro 33 apresenta um conjunto de 30 tópicos treinados utilizando o modelo LSI em um processamento que foi realizado em 11 minutos e 49 segundos. Os conjuntos de resultados com 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponível através do GitHub[63].

Quadro 33 – Tópicos extraídos do *corpus* 5 usando o modelo LSI

Tópico 0: 0.664*"informação" + 0.185*"pesquisa" + 0.156*"conhecimento" + 0.116*"dados" + 0.107*"biblioteca" + 0.105*"forma" + 0.098*"processo" + 0.092*"uso" + 0.085*"trabalho" + 0.082*"social";
Tópico 1: -0.395*"publicados" + -0.293*"localizado" + -0.203*"artigos" + -0.202*"completos" + -0.201*"completos_publicados" + -0.189*"periódicos" + 0.184*"informação" + -

[63] Algoritmo de modelagem de tópicos. *Corpus* 5: teses e dissertações 2016. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2016.ipynb/.

0.177*"publicados_periódicos" + -0.177*"artigos_completos" + -0.177*"artigos_completos_publicados";
Tópico 2: 0.517*"informação" + -0.357*"museu" + -0.149*"memória" + -0.135*"cultura" + -0.118*"cultural" + -0.115*"arte" + -0.114*"história" + -0.112*"brasil" + -0.102*"nacional" + -0.097*"anos";
Tópico 3: -0.541*"documentos" + -0.265*"gestão" + -0.260*"museu" + -0.220*"arquivo" + -0.205*"arquivos" + 0.161*"biblioteca" + -0.155*"gestão_documentos" + -0.097*"administração" + -0.086*"informação" + -0.082*"lei";
Tópico 4: 0.649*"museu" + -0.293*"documentos" + 0.283*"informação" + -0.147*"dados" + -0.130*"biblioteca" + -0.123*"pesquisa" + 0.115*"patrimônio" + 0.105*"arte" + -0.098*"gestão" + 0.095*"turismo";
Tópico 5: -0.727*"biblioteca" + -0.158*"livros" + -0.145*"leitura" + -0.135*"livro" + -0.132*"bibliotecário" + -0.113*"biblioteconomia" + -0.111*"educação" + 0.108*"conhecimento" + 0.095*"dados" + -0.094*"usuários";
Tópico 6: 0.386*"dados" + 0.274*"pesquisa" + 0.262*"museu" + 0.171*"científica" + -0.166*"documentos" + -0.141*"memória" + 0.132*"inovação" + 0.125*"pesquisadores" + -0.117*"informação" + -0.109*"jornalismo";
Tópico 7: 0.229*"web" + -0.187*"conhecimento" + -0.180*"gestão" + -0.159*"inovação" + 0.151*"usuários" + 0.147*"livros" + 0.145*"museu" + 0.138*"usuário" + 0.135*"dados" + 0.125*"obra";
Tópico 8: -0.461*"conhecimento" + 0.169*"informação" + 0.164*"jornalismo" + -0.136*"museu" + -0.135*"inovação" + -0.135*"organização" + 0.134*"dados" + -0.120*"web" + 0.119*"brasil" + 0.112*"direito";
Tópico 9: -0.258*"memória" + -0.204*"curso" + 0.178*"jornalismo" + -0.149*"pesquisa" + 0.144*"gestão" + 0.144*"conhecimento" + 0.139*"inovação" + -0.135*"arte" + 0.129*"comunicação" + 0.125*"organização";
Tópico 10: 0.263*"livros" + 0.255*"livro" + 0.245*"conhecimento" + -0.184*"usuários" + 0.168*"coleção" + -0.151*"museu" + -0.148*"uso" + 0.140*"bolso" + -0.121*"gif" + -0.120*"gifs";
Tópico 11: 0.305*"jornalismo" + 0.287*"citação" + 0.258*"web" + -0.249*"dados" + 0.235*"citação_web" + 0.122*"conhecimento" + -0.100*"gestão" + 0.100*"documentos" + 0.099*"museu" + -0.097*"uso";
Tópico 12: 0.257*"dados" + 0.185*"biblioteca" + 0.169*"cultura" + 0.162*"conhecimento" + -0.160*"livros" + 0.159*"web" + -0.125*"livro" + -0.123*"estudantes" + 0.116*"cultural" + -0.114*"educação";
Tópico 13: -0.283*"memória" + 0.267*"cultura" + 0.223*"curso" + 0.140*"ouro" + 0.136*"ouro_preto" + 0.135*"preto" + 0.126*"gif" + 0.125*"gifs" + 0.119*"cultural" + 0.115*"usuários";
Tópico 14: -0.324*"dados" + -0.239*"jornalismo" + 0.160*"jurídica" + 0.153*"direito" + -0.139*"leitura" + -0.136*"educação" + 0.114*"usuário" + 0.112*"bdjur" + -0.103*"leitor" + 0.101*"termo";
Tópico 15: 0.251*"jornalismo" + -0.187*"gif" + -0.184*"gifs" + 0.177*"direito" + 0.161*"avaliação" + 0.156*"jurídica" + 0.152*"curso" + -0.125*"políticas" + 0.120*"busca" + -0.115*"digital";
Tópico 16: -0.237*"jornalismo" + 0.200*"citação" + 0.187*"web" + 0.175*"citação_web" + 0.152*"direito" + 0.132*"jurídica" + 0.131*"dados" + -0.110*"preservação" + -0.107*"revista" + -0.105*"science";
Tópico 17: 0.308*"gif" + 0.305*"gifs" + 0.181*"animados" + 0.176*"gifs_animados" + 0.169*"jornalismo" + -0.153*"sistema" + 0.149*"conhecimento" + -0.145*"jogo" + -0.141*"comunicação" + 0.137*"sujeito";
Tópico 18: -0.389*"memória" + 0.227*"jogo" + -0.154*"preservação" + 0.145*"arte" + -0.132*"conhecimento" + -0.128*"usuários" + 0.118*"dados" + 0.113*"obra" + -0.108*"digital" + 0.096*"curso";

Tópico 19: -0.278*"digital" + -0.200*"jogo" + 0.192*"dados" + 0.185*"usuários" + -0.184*"preservação" + -0.171*"jurídica" + -0.169*"direito" + -0.135*"inovação" + 0.126*"gente" + -0.126*"digitais";
Tópico 20: -0.287*"leitura" + -0.165*"termo" + 0.162*"preservação" + 0.153*"digital" + -0.149*"cultura" + -0.147*"busca" + -0.126*"cultural" + -0.119*"nacional" + -0.115*"texto" + 0.111*"arte";
Tópico 21: -0.199*"memória" + 0.190*"educação" + -0.147*"biblioteca" + 0.146*"social" + -0.145*"jogo" + -0.127*"inovação" + -0.119*"busca" + 0.108*"sociedade" + -0.107*"gente" + -0.104*"preservação";
Tópico 22: 0.433*"inovação" + -0.233*"conhecimento" + -0.221*"jogo" + 0.132*"obra" + -0.118*"gente" + 0.114*"arte" + -0.103*"compartilhamento" + 0.094*"sustentabilidade" + -0.094*"gameplay" + -0.088*"pessoas";
Tópico 23: -0.171*"usuários" + -0.150*"web" + 0.149*"artigos" + 0.143*"científica" + 0.127*"produção" + 0.123*"comunicação" + 0.120*"leitura" + -0.113*"jogo" + 0.112*"busca" + 0.110*"competência";
Tópico 24: -0.183*"sustentabilidade" + 0.183*"gestão" + 0.178*"revista" + -0.171*"termo" + 0.154*"inovação" + 0.142*"revistas" + -0.125*"uso" + 0.110*"arte" + -0.109*"ambiental" + 0.097*"site";
Tópico 25: -0.188*"gente" + 0.188*"memória" + 0.149*"jogo" + -0.131*"documentos" + 0.125*"leitura" + -0.118*"bibliotecário" + -0.115*"funk" + -0.106*"biblioteconomia" + 0.101*"sistema" + -0.098*"digital";
Tópico 26: 0.190*"sustentabilidade" + -0.153*"usuários" + -0.150*"social" + -0.150*"inovação" + -0.140*"cultural" + -0.132*"uso" + -0.130*"e-books" + -0.129*"sociais" + 0.124*"indígenas" + 0.117*"educação";
Tópico 27: -0.427*"inovação" + 0.211*"comunicação" + 0.141*"sustentabilidade" + 0.127*"namitec" + 0.113*"rede" + -0.107*"jogo" + -0.101*"artigos" + -0.100*"brasil" + 0.100*"redes" + -0.097*"indígenas";
Tópico 28: 0.217*"memória" + -0.194*"filme" + 0.186*"sustentabilidade" + -0.170*"conhecimento" + 0.158*"social" + -0.136*"indígenas" + -0.127*"cinema" + -0.119*"digital" + -0.102*"making" + -0.099*"produção";
Tópico 29: -0.247*"inovação" + -0.206*"comunicação" + 0.178*"obra" + 0.167*"cultural" + 0.126*"pedra" + 0.123*"jornalismo" + -0.116*"funk" + 0.116*"sal" + 0.115*"pedra_sal" + -0.114*"tecnologias".

Fonte: Elaborado pelo autor.

Torna-se possível perceber entre os termos formados por termos e pesos extraídos do *corpus* de dados utilizando o modelo LSI a existência de termos fortes como “informação” contidos nos tópicos 0 e 2, “documentos” no tópico 3, “museu” no tópico 4 e “biblioteca” no tópico 5, com valores acima de 0.500. Os termos de cada tópico são ordenados por relevância de acordo com os valores dos pesos. Além disso, é possível observar tópicos inteiros com diferentes características, sendo fracos ou fortes e generalistas ou especialistas quando se comparado ao domínio da linguagem.

O tópico 0 pode ser considerado generalista por apresentar termos gerais pertinentes ao domínio da linguagem, sendo necessário ao especialista realizar pesquisas externas em documentos como lista de bigramas e trigramas ou

acesso ao *corpus* de dados de forma que seja possível criar uma suposição do nome do tópico de maneira mais assertiva ou mesmo descartar o tópico.

Dentre os tópicos fortes podem ser destacados o tópico 1 que apresenta termos coesos como “artigos”, “periódicos” e “artigos_completos_publicados”, podendo ao especialista supor um rótulo que aborde assuntos relacionados a Produção Científica. O tópico 14 também apresenta característica fortes por conter termos chave que podem remeter ao especialista criar a suposição de rótulo como base nos termos “jurídica”, “direito”, “leitura”, “leitor”, “usuário” e “bdjur (Banco de Dados Jurídico)” como por exemplo Informação Jurídica ou Produto de Informação Jurídica.

Além disso, os resultados apresentam um conjunto de tópicos fracos caracterizados por uma aproximação de termos distintos como apresentado nos tópicos 25, 26, 27 28 e 29 que abordam diversos assuntos como “jogo”, “sustentabilidade”, “e-books”, “indígenas”, “funk”, “namitec” e “pedra_sal”. Esse número excessivo de tópicos fracos ocorre mediante o tamanho do *corpus* de dados analisado com o número elevado de tópicos configurados para extração na modelagem de tópicos.

O Quadro 34 apresenta um conjunto com 30 tópicos constituído por termos e pesos extraídos do *corpus* de dados utilizando o modelo LDA. O tempo para execução do modelo foi de 48 minutos e 16 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[64].

Quadro 34 – Tópicos extraídos do *corpus* 5 usando o modelo LDA

Tópico 0: 0.005*"publicados" + 0.004*"localizado" + 0.003*"completos_publicados" + 0.003*"completos" + 0.002*"artigos" + 0.002*"publicados_periódicos" + 0.002*"artigos_completos" + 0.002*"artigos_completos_publicados" + 0.002*"periódicos" + 0.002*"completos_publicados_periódicos";
Tópico 1: 0.000*"informação" + 0.000*"dados" + 0.000*"pesquisa" + 0.000*"cinema" + 0.000*"social" + 0.000*"memória" + 0.000*"processo" + 0.000*"tempo" + 0.000*"brasil" + 0.000*"conhecimento";
Tópico 2: 0.001*"centro_memória" + 0.000*"centros_memória" + 0.000*"cemef" + 0.000*"memória_documentação" + 0.000*"centros_memória_documentação" + 0.000*"escola_educação" + 0.000*"linhales" + 0.000*"garimpando" + 0.000*"projeto_garimpando" + 0.000*"garimpando_memórias";

[64] Algoritmo de modelagem de tópicos. *Corpus* 5: teses e dissertações 2016. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2016.ipynb/.

Tópico 3: 0.001*"pmlllb" + 0.001*"livro_leitura" + 0.000*"plenárias" + 0.000*"saraus" + 0.000*"leitura_literatura" + 0.000*"livro_leitura_literatura" + 0.000*"biblioteca_comunitárias" + 0.000*"pmll" + 0.000*"municipal_livro" + 0.000*"municipal_livro_leitura";
Tópico 4: 0.001*"plano" + 0.001*"planos" + 0.001*"beneficiários" + 0.001*"ans" + 0.001*"operadoras" + 0.001*"suplementar" + 0.001*"beneficiário" + 0.000*"operadora" + 0.000*"site_prefeitura" + 0.000*"then";
Tópico 5: 0.002*"engenharia" + 0.002*"pedra" + 0.002*"pedra_sal" + 0.001*"sal" + 0.001*"deficiência" + 0.001*"peregrino" + 0.001*"biblioteca_nacional" + 0.001*"peregrino_silva" + 0.001*"otlet" + 0.001*"bibliografia";
Tópico 6: 0.000*"museu" + 0.000*"informação" + 0.000*"pesquisa" + 0.000*"dados" + 0.000*"conhecimento" + 0.000*"brasil" + 0.000*"anos" + 0.000*"forma" + 0.000*"comunicação" + 0.000*"bravo";
Tópico 7: 0.002*"indígenas" + 0.001*"ordenação" + 0.001*"linked" + 0.001*"data" + 0.001*"avignon" + 0.001*"festival" + 0.001*"linked_data" + 0.001*"bergman" + 0.001*"livrocensura" + 0.001*"des";
Tópico 8: 0.001*"petróleo" + 0.001*"petrobrás" + 0.001*"indústria_petróleo" + 0.000*"anp" + 0.000*"marco_regulatório" + 0.000*"petróleo_natural" + 0.000*"regulatório" + 0.000*"petróleo_brasil" + 0.000*"indústria_petróleo_brasil" + 0.000*"indústria_petróleo_natural";
Tópico 9: 0.002*"direito" + 0.002*"e-books" + 0.002*"jurídica" + 0.001*"acv" + 0.001*"bdjur" + 0.001*"usuários" + 0.001*"casamento" + 0.001*"stj" + 0.001*"informação_jurídica" + 0.001*"portal";
Tópico 10: 0.003*"web" + 0.002*"citação_web" + 0.002*"citação" + 0.001*"documentos" + 0.001*"mapa" + 0.001*"mapa_conceitual" + 0.001*"taxonomia" + 0.001*"ufpr" + 0.001*"domínio" + 0.001*"ufsc";
Tópico 11: 0.001*"acervo" + 0.001*"bibliófilos" + 0.001*"fase" + 0.001*"imagens" + 0.001*"livros" + 0.001*"movimento" + 0.001*"imagens_movimento" + 0.001*"bce" + 0.001*"cem" + 0.001*"cem_bibliófilos";
Tópico 12: 0.004*"cinema" + 0.002*"filme" + 0.001*"ouro" + 0.001*"preto" + 0.001*"ouro_preto" + 0.001*"filmes" + 0.001*"curso" + 0.001*"arte" + 0.001*"cecab" + 0.001*"tarkovski";
Tópico 13: 0.002*"rossini" + 0.001*"perez" + 0.001*"rossini_perez" + 0.001*"amado" + 0.001*"jorge_amado" + 0.001*"presbitério" + 0.001*"foto" + 0.001*"presbitério_sul" + 0.001*"jorge" + 0.001*"sul_paraíba";
Tópico 14: 0.002*"gif" + 0.002*"gifs" + 0.001*"piva" + 0.001*"animados" + 0.001*"gifs_animados" + 0.001*"musical" + 0.001*"ontologia" + 0.001*"ipad" + 0.001*"animado" + 0.001*"gif_animado";
Tópico 15: 0.001*"segurança_informação" + 0.001*"educação_usuários" + 0.001*"segurança" + 0.001*"ufal" + 0.001*"ufal_ufmg" + 0.000*"bus" + 0.000*"chesf" + 0.000*"moderada" + 0.000*"decg" + 0.000*"progep";
Tópico 16: 0.002*"bolso" + 0.002*"making" + 0.001*"eliza" + 0.001*"filme" + 0.001*"amarildo" + 0.001*"murilo" + 0.001*"poche" + 0.001*"rede_globo" + 0.001*"cinema" + 0.001*"livro_bolso";
Tópico 17: 0.001*"curso_museu" + 0.000*"mhn" + 0.000*"museu_histórico_nacional" + 0.000*"histórico_nacional" + 0.000*"museu_histórico" + 0.000*"objectos" + 0.000*"curso_museu_mhn" + 0.000*"museu_mhn" + 0.000*"barroso" + 0.000*"conservadores_museu";
Tópico 18: 0.000*"folia" + 0.000*"secretário_cultura_turismo" + 0.000*"reis_magos" + 0.000*"folias" + 0.000*"folia_reis" + 0.000*"valença" + 0.000*"herodes" + 0.000*"presidente_agforv" + 0.000*"guarine" + 0.000*"agforv";

Tópico 19: 0.000*"informação" + 0.000*"conhecimento" + 0.000*"pesquisa" + 0.000*"forma" + 0.000*"anos" + 0.000*"meio" + 0.000*"revista" + 0.000*"trabalho" + 0.000*"social" + 0.000*"dados";
Tópico 20: 0.000*"informação" + 0.000*"forma" + 0.000*"dados" + 0.000*"pesquisa" + 0.000*"brasil" + 0.000*"obra" + 0.000*"sistema" + 0.000*"relação" + 0.000*"tempo" + 0.000*"diferentes";
Tópico 21: 0.000*"informação" + 0.000*"termo" + 0.000*"pesquisa" + 0.000*"conhecimento" + 0.000*"pessoas" + 0.000*"forma" + 0.000*"biblioteca" + 0.000*"sistema" + 0.000*"uso" + 0.000*"meio";
Tópico 22: 0.001*"yoga" + 0.000*"yogasūtra" + 0.000*"patañjali" + 0.000*"puruṣa" + 0.000*"libertação" + 0.000*"prakṛti" + 0.000*"buddhi" + 0.000*"aforismo" + 0.000*"yogi" + 0.000*"filliozat";
Tópico 23: 0.011*"informação" + 0.004*"pesquisa" + 0.003*"conhecimento" + 0.003*"forma" + 0.002*"dados" + 0.002*"biblioteca" + 0.002*"social" + 0.002*"processo" + 0.002*"relação" + 0.002*"trabalho";
Tópico 24: 0.006*"documentos" + 0.002*"gestão_documentos" + 0.002*"gestão" + 0.002*"arquivo" + 0.002*"arquivos" + 0.002*"funk" + 0.002*"sustentabilidade" + 0.001*"artigos" + 0.001*"belo_horizonte" + 0.001*"matemática";
Tópico 25: 0.000*"informação" + 0.000*"forma" + 0.000*"conhecimento" + 0.000*"social" + 0.000*"pesquisa" + 0.000*"produção" + 0.000*"desenvolvimento" + 0.000*"inovação" + 0.000*"relação" + 0.000*"dados";
Tópico 26: 0.003*"jornalismo" + 0.002*"filme" + 0.001*"chico" + 0.001*"foto" + 0.001*"vassouras" + 0.001*"aeroporto" + 0.001*"fotografia" + 0.001*"buarque" + 0.001*"chico_buarque" + 0.001*"leitor";
Tópico 27: 0.002*"jogo" + 0.001*"gameplay" + 0.001*"jogador" + 0.001*"kiss" + 0.001*"game" + 0.001*"games" + 0.001*"jogos" + 0.001*"banerj" + 0.001*"jogar" + 0.001*"jogadores";
Tópico 28: 0.000*"informação" + 0.000*"trabalho" + 0.000*"repórter" + 0.000*"pesquisa" + 0.000*"forma" + 0.000*"cultura" + 0.000*"brasil" + 0.000*"publicados" + 0.000*"nacional" + 0.000*"anos";
Tópico 29: 0.001*"maioridade" + 0.001*"penal" + 0.001*"texto" + 0.001*"maioridade_penal" + 0.001*"pec" + 0.001*"redução" + 0.001*"redução_maioridade" + 0.000*"folha" + 0.000*"dia_publicação_dia" + 0.000*"publicação_dia_semana".

Fonte: Elaborado pelo autor.

Entre os resultados obtidos através do modelo LDA é possível identificar que os tópicos possuem maior quantidade de N-gramas do tipo bigramas e trigramas quando se comparado com o modelo LSI. Além disso, a existência de termos generalistas ao domínio da linguagem como “dados” e “memória” que podem apresentar composições em seus significados, tais como “gestão_dados”, “ciclo_vida_dados”, “memória_social” e “memória_coletiva” aparece em quantidades reduzidas também se comparado ao modelo LSI. Essas características contribuem para um menor esforço cognitivo do especialista ao realizar a análise de assunto no domínio da linguagem de forma que possa inferir a suposição do nome de um rótulo de maneira mais assertiva. Entretanto, uma característica fraca apresentada entre os resultados está nos baixos valores dos pesos de cada termo, e em muitos casos, com valores iguais

a 0. Essa característica fraca dificulda por exemplo na ordem de relevância dos termos de cada tópico.

O tópico 3 apresenta termos especialistas como “pmlllb (Plano Municipal do Livro, Leitura, Literatura e Biblioteca)”, “pmll (Plano Municipal do Livro, Leitura)” que quando analisado em documentos externos como o *corpus* de dados, torna-se possível identificar de forma clara a área de estudos, uma vez que poucos documentos contidos no *corpus* de dados trabalham com esses termos. Os demais termos como “livro_leitura_literatura”, “municipal_livro_leitura” e “biblioteca_comunitárias” possuem coesão com a temática, podendo assim ao especialista supor através de técnicas de análise de assunto que o tópico esteja relacionado aos rótulos Política e Cultura, Política do livro e leitura ou Informação, Política e Cultura.

O tópico 18 também apresenta termos específicos como “reis_magos”, “folias”, “secretário_cultura_turismo”, “valença” e “agforv” (Associação dos grupos de folias de reis de Valença) que facilita ao especialista identificar o documento mais relevante através de acesso ao *corpus* de dados para realizar uma interpretação do tópico de forma mais assertiva. Entretanto, mesmo caso o especialista não tenha acesso a documentos externos, é possível identificar uma coesão entre os termos, podendo supor que o tópico esteja relacionado por exemplo a Cultura Popular, Política Pública ou Patrimônio Cultural.

Outro tópico que apresenta a mesma característica é o de número 27, que contempla termos coesos como “jogo”, “jogadores”, “games” e “gameplay”, podendo remeter assim ao especialista e supor que o nome do tópico esteja atrelado a temas como Informação e Tecnologia, Cultura Digital ou Jogos Digirais. O tópico 29 também possui características fortes e termos especialistas, tais como "maioridade", “penal”, “pec”, “redução” além dos bigramas "maioridade_penal" e "redução_maioridade", podendo o especialista supor que o tópico esteja associado a termas como Política, Economia e Informação.

APÊNDICE F - *Corpus* 6: teses e dissertações 2017

O *corpus* 6 diz respeito ao último grupo de documentos que constitui o primeiro *corpora*, apresentando assim o quantitativo de 219 documentos do tipo teses e dissertações defendidas no ano de 2017 e possuindo assim o tamanho de 81.840kb. Justifica-se a quantidade menor de arquivos dessa coleção quando se comparado aos *corpus* 1, 2, 3, 4, e 5 mediante a data de coleta de dados utilizados para amostragem com os processos e prazos estipulados pelos programas de pós-graduação após realização de defesa de tese ou dissertação para entrega da versão final do trabalho, bem como os trâmites para publicação através das bibliotecas digitais de teses e dissertações. Acredita-se que o volume desse *corpus* poderia ser maior caso a coleta de dados fosse realizada no final do ano de 2018.

Os N-gramas do *corpus* 6 resultaram em 5.710.149 unigramas, 5.709.930 bigramas e 5.709.711 trigramas. O Quadro 35 apresenta uma lista contendo os 50 termos mais frequentes de cada tipo de N-grama extraídos no *corpus* de dados.

Quadro 35 – Lista de N-gramas por ordem de frequência do *corpus* 6

Unigramas
informação,70181; pesquisa,25323; conhecimento,18395; forma,17412; dados,17191; processo,15955; social,14610; produção,12488; relação,12482; trabalho,12247; brasil,11543; meio,11488; comunicação,10858; organização,10526; documentos,10524; sociais,10361; tempo,10353; uso,10285; memória,10076; pessoas,9295; termo,9244; paulo,9230; universidade,8781; fonte,8758; sociedade,8611; sistema,8409; anos,8402; estudo,8163; museu,8126; desenvolvimento,7969; espaço,7925; gestão,7867; contexto,7849; vida,7804; estudos,7614; nacional,7602; busca,7566; história,7462; mundo,7311; modo,7276; cultura,7245; caso,7236; campo,7178; construção,7058; diferentes,6835; arte,6743; grupo,6528; relações,6448; processos,6446; biblioteca,6281.
Bigramas
universidade_federal,2826; redes_sociais,1817; informação_tecnologia,1810; porto_alegre,1802; recuperação_informação,1573; organização_conhecimento,1521; produção_científica,1488; uso_informação,1479; muitas_vezes,1453; informação_conhecimento,1362; gestão_informação,1283; ponto_vista,1267; dados_pesquisa,1260; informação_science,1194; universidade_estadual,1135; dissertação_mestrado,1118; coleta_dados,1071; organização_informação,1018; sistemas_informação,1018; fontes_informação,1018; belo_horizonte,1003; informação_comunicação,989; informação_informação,964; memória_social,952; bases_dados,927; tecnologias_informação,917; gestão_conhecimento,883; competência_informação,871; arquivo_nacional,857; fonte_elaborado,844; feed_notícias,831; busca_informação,804; grupos_pesquisa,800; tendo_vista,791; sociais_aplicadas,771; patrimônio_cultural,768; fonte_dados,759; pesquisa_informação,756; universidade_paulo,737; grande_sul,730; meio_ambiente,725; base_dados,724; objeto_estudo,687; comunicação_informação,685; big_data,680; tomada_decisão,660;

tecnologia_informação,650; comunicação_científica,649; deste_trabalho,645; ensino_superior,642.
Trigramas
fonte_dados_pesquisa,603; fonte_elaborado_autora,583; tecnologias_informação_comunicação,546; universidade_federal_grande,451; federal_grande_sul,408; sistemas_organização_conhecimento,371; universidade_federal_santa,335; busca_uso_informação,311; sociais_aplicadas_universidade,298; informação_universidade_federal,277; international_organization_standardization,276; critério_pts_critério,274; pts_critério_pts,273; projeto_ademar_guerra,272; encontro_nacional_pesquisa,267; linguística_letras_artes,266; dissertação_mestrado_informação,265; nacional_pesquisa_informação,264; universidade_federal_paraíba,254; resource_description_framework,240; extensible_markup_language,236; pessoas_deficiência_visual,234; documento_online_tradução,232; humanas_universidade_federal,231; prevenção_combate_corrupção,228; exatas_natureza_universidade,218; museu_arte_moderna,217; patrimônio_histórico_artístico,215; sudeste_humanas_universidade,213; gestão_dados_científicos,212; federal_santa_catarina,210; organização_uso_informação,207; sites_redes_sociais,206; conhecimentos_habilidades_atitudes,205; paulo_companhia_letras,197; katiuscia_neri_cabeça,197; society_informação_science,196; american_society_informação,193; informação_belo_horizonte,193; library_informação_science,191; universidade_federal_bahia,191; requisito_parcial_obtenção,188; programa_comunicação_informação,187; segunda_guerra_mundial,185; informação_science_technology,182; paulo_martins_fontes,182; feira_hippie_ipanema,182; instituição_ensino_superior,180; formação_recursos_humanos,176; letras_artes_universidade,172.

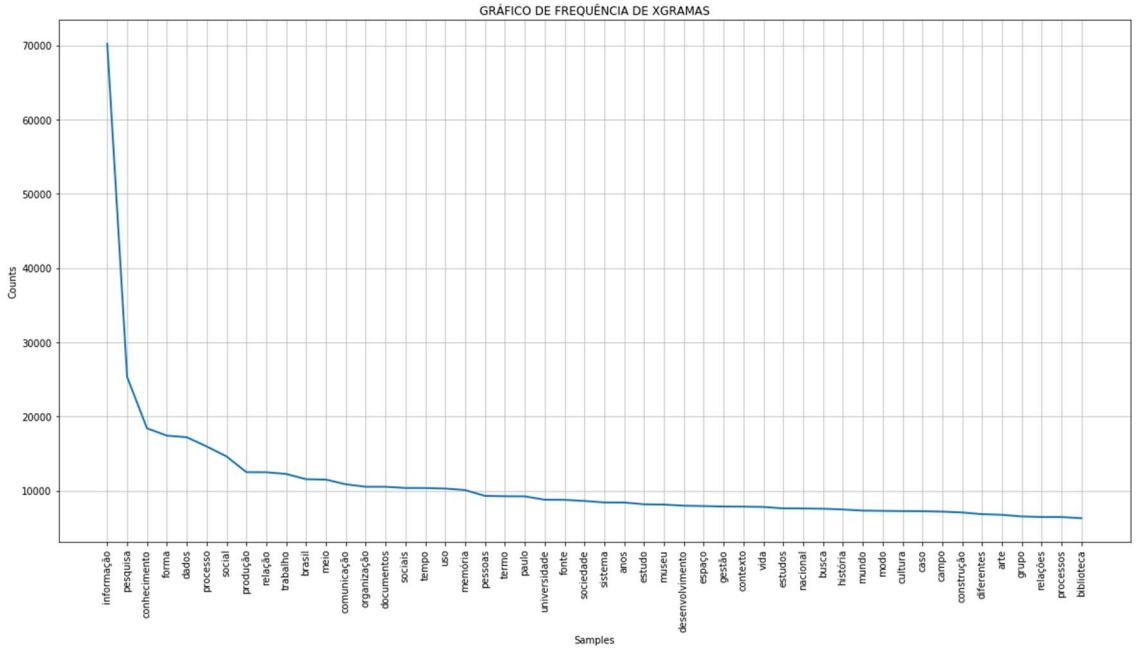
Fonte: Elaborado pelo autor.

Destaca-se numa lista geral contendo os mil N-gramas mais frequentes extraídos do *corpus* 6 os bigramas “universidade_federal” com frequência de 2.826 e ocupando a posição 245, “redes_sociais” com frequência de 1.817 na posição 460, “informação_tecnologia” com frequência de 1.810 e posição 463, “porto_alegre” com frequência de 1.802 na posição 465 e “recuperação_informação” com frequência de 1.573 na posição 553. Todos os trigramas aparecem após ao milésimo termo da lista geral de N-gramas, sendo os mais frequentes “fonte_dados_pesquisa” com frequência de 603, “fonte_elaborado_autora” com frequência de 583, “tecnologias_informação_comunicação” com frequência de 546, “universidade_federal_grande” com frequência de 451 e “federal_grande_sul” com frequência de 408. É possível perceber entre os resultados a existência de termos fortes e fracos que possuem uma maior ou menor representatividade junto ao domínio da linguagem, o que possibilita ao especialistaque decidir por utilizar ou descartar os termos ou mesmo tópicos inteiros durante a análise de assunto.

Os unigramas apresentam maior frequência dentre os tipos de N-gramas extraídos do *corpus* de dados. Se comparado por exemplo o primeiro unigrama representado pelo termo “informação” e frequência de 70.181 com o primeiro bigrama “universidade_federal” e frequência de 2.826, encontra-se um distanciamento entre os termos de 2.383% Esse percentual alcança 11.481% quando comparado o primeiro unigrama com o primeiro trigrama representado pelo termo “fonte_dados_pesquisa” e frequência de 603.

O Gráfico 57 apresenta os 50 N-gramas mais frequentes extraídos do *corpus* de dados. É possível perceber nos resultados que todos os termos são do tipo unigramas. Os termos mais frequentes são “informação” com frequência de 70.181, “pesquisa” com frequência de 25.323, “conhecimento” com frequência de18.395, “forma” com frequência de 17.412 e “dados” com frequência de 17.191.

Gráfico 57 – Termos mais frequentes do *corpus* 6



Fonte: Elaborado pelo autor.

No gráfico ainda é possível perceber uma frequência crescente de forma constante no volume de dados entre o 50º e o 3º termo, sendo eles respectivamente representados por “biblioteca” com frequência de 6.281 e “conhecimento” com frequência de 18.395. Já os dois primeiros termos apresentam um maior volume acumulativo de frequências que destoam dos demais termos, resultando numa diferença de 38% entre o segundo termo “pesquisa” com frequência de 25.323 para o terceiro termo “conhecimento” com

frequência de 18.395 e 282% do primeiro termo “informação” com frequência de 70.181 para o terceiro termo.

Ainda é possível encontrar entre os resultados termos fortes como “busca”, “fonte”, “cultura”, “museu” e “espaço” e termos fracos como “diferentes”, “anos”, “paulo”, “contexto” e ”brasil”. Uma boa prática para o especialista no domínio da linguagem está em explorar os resultados obtidos através dos bigramas e/ou trigramas, podendo encontrar assim opções nos termos que possuem composições e apresentam conceitos diferentes, tais como “mecanismo_busca”, “processo_busca” ou “busca_uso_informação”. Os termos fracos e que não possuem aderência ao domínio da linguagem podem ser descartados pelo especialista.

Mediante o quantitativo de N-gramas, foi gerado uma nuvem de palavras contendo os 250 termos mais frequentes do *corpus* 6 conforme apresentado na Figura 18. Essa ilustração é composta por 249 unigramas e apenas um bigrama, sendo “universidade_federal”. Faz-se necessário ressaltar que para construção dessa figura foi utilizada somente os dados quantitativos, não sendo realizado assim quaisquer tratamentos qualitativos.

Figura 18 – Nuvem de palavras do *corpus* 6

Fonte: Elaborado pelo autor.

Na imagem é possível identificar termos fortes e fracos quando se comparado ao domínio da linguagem. Exemplos de tópicos fortes estão os unigramas “processo”, “relação” e “história”, mesmo remetendo a outros significados quando analisados os bigramas e trigramas, apresentam termos de

relevância pertinentes ao *corpus* de dados. Exemplos de termos fracos encontrados na imagem são “diferente”, “importante” e “palavra” que não apresentam conceitos de relevância relacionado ao domínio de linguagem.

A conexão com os modelos de treinamento permitiu extrair conjuntos de tópicos constituídos por termos e pesos do *corpus* de dados. O Quadro 36 apresenta um conjunto de 26 tópicos extraídos por mdemio de modelagem de tópicos utilizando o modelo LSI que foi processado em 6 minutos e 42 segundos. Os resultados contendo 10, 14, 18, 22, 30, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[65].

Quadro 36 – Tópicos extraídos do *corpus* 6 usando o modelo LSI

Tópico 0: 0.644*"informação" + 0.202*"pesquisa" + 0.150*"conhecimento" + 0.140*"dados" + 0.116*"forma" + 0.113*"processo" + 0.096*"social" + 0.088*"produção" + 0.085*"brasil" + 0.082*"trabalho";
Tópico 1: -0.377*"une" + -0.376*"des" + -0.312*"nous" + -0.304*"est" + -0.206*"qui" + -0.196*"dans" + -0.193*"pour" + -0.175*"patrimoine" + -0.164*"sur" + -0.149*"par";
Tópico 2: 0.594*"informação" + -0.150*"arte" + -0.131*"museu" + -0.113*"memória" + -0.112*"tempo" + -0.110*"forma" + -0.104*"social" + -0.104*"relação" + -0.102*"vida" + -0.098*"espaço";
Tópico 3: 0.361*"revista" + 0.332*"universidade" + 0.228*"pesquisa" + 0.202*"dados" + 0.173*"científica" + -0.152*"informação" + 0.137*"humanas" + 0.134*"sudeste" + 0.127*"federal" + 0.127*"sul";
Tópico 4: -0.420*"dados" + 0.284*"museu" + 0.249*"arte" + 0.203*"revista" + 0.181*"universidade" + 0.173*"informação" + -0.135*"pesquisa" + -0.124*"feed" + 0.116*"obras" + 0.110*"documentação";
Tópico 5: -0.343*"dados" + -0.336*"museu" + -0.253*"arte" + -0.164*"documentos" + -0.147*"pesquisa" + -0.132*"documentação" + -0.126*"obras" + -0.125*"exposição" + 0.125*"informação" + 0.123*"universidade";
Tópico 6: -0.270*"feed" + -0.252*"universidade" + -0.189*"facebook" + -0.189*"notícias" + -0.178*"feed_notícias" + -0.132*"documentos" + -0.131*"tradução" + -0.129*"documento" + -0.128*"humanas" + 0.125*"pesquisa";
Tópico 7: -0.225*"feed" + 0.179*"universidade" + 0.174*"documentos" + -0.162*"maria" + -0.160*"facebook" + -0.157*"ontology" + -0.157*"arte" + -0.153*"museu" + -0.151*"notícias" + -0.148*"feed_notícias";
Tópico 8: 0.273*"documentos" + -0.254*"dados" + -0.193*"arte" + 0.150*"conhecimento" + 0.149*"documento" + 0.143*"maria" + 0.143*"organização" + 0.141*"revista" + -0.136*"universidade" + -0.130*"museu";
Tópico 9: 0.482*"ontology" + 0.241*"conhecimento" + -0.187*"feed" + 0.150*"ontologies" + -0.149*"facebook" + -0.144*"notícias" + 0.142*"figure" + 0.141*"this" + 0.136*"that" + 0.124*"data";
Tópico 10: 0.249*"brasil" + 0.214*"ontology" + -0.175*"arte" + -0.164*"conhecimento" + -0.151*"periódicos" + 0.131*"dados" + 0.129*"patrimônio" + -0.128*"produção" + 0.117*"memória" + 0.109*"jongo";

[65] Algoritmo de modelagem de tópicos. *Corpus* 6: teses e dissertações 2017. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2017.ipynb/.

Tópico 11: 0.302*"dados" + -0.220*"conhecimento" + -0.172*"museologia" + -0.172*"museu" + -0.155*"social" + 0.144*"revista" + 0.134*"arte" + 0.122*"link" + -0.114*"namitec" + 0.114*"digital";
Tópico 12: -0.251*"periódicos" + -0.199*"periódico" + -0.183*"critério" + -0.163*"publicação" + -0.156*"artigos" + -0.152*"fachin" + -0.135*"pts" + -0.135*"critério_pts" + -0.121*"critérios" + -0.121*"editorial";
Tópico 13: 0.298*"museu" + -0.257*"arte" + 0.243*"museologia" + 0.191*"link" + -0.175*"processo" + -0.157*"obras" + 0.151*"reportagem" + 0.122*"história" + -0.121*"ontology" + -0.118*"pessoas";
Tópico 14: -0.306*"dados" + -0.217*"conhecimento" + 0.209*"nacional" + 0.201*"arquivo" + 0.188*"namitec" + 0.151*"arquivos" + 0.150*"documentos" + 0.142*"incts" + -0.136*"museologia" + 0.125*"link";
Tópico 15: -0.216*"memória" + 0.178*"arte" + -0.173*"documentos" + 0.162*"jornalismo" + -0.149*"vida" + 0.142*"contas" + 0.136*"brasil" + -0.128*"patrimônio" + 0.117*"reportagem" + 0.112*"obras";
Tópico 16: 0.175*"contas" + -0.174*"conhecimento" + -0.172*"comunicação" + -0.165*"digital" + -0.163*"organização" + -0.137*"documentos" + -0.129*"link" + -0.124*"consumo" + -0.118*"produção" + 0.116*"informação";
Tópico 17: -0.202*"jongo" + -0.182*"conhecimento" + 0.179*"biblioteca" + 0.166*"estudantes" + 0.136*"pesquisa" + -0.135*"memória" + 0.134*"digital" + 0.134*"educação" + -0.131*"organização" + -0.129*"patrimônio";
Tópico 18: -0.272*"brasil" + 0.257*"jongo" + -0.188*"conhecimento" + 0.169*"patrimônio" + 0.152*"processo" + 0.147*"gente" + -0.124*"negro" + -0.120*"negra" + 0.115*"cultural" + 0.115*"link";
Tópico 19: -0.172*"digital" + -0.158*"link" + -0.149*"conhecimento" + -0.142*"reportagem" + -0.137*"memória" + 0.130*"participantes" + -0.129*"biblioteca" + 0.120*"arquivo" + 0.119*"estudantes" + -0.118*"profissional";
Tópico 20: -0.287*"estudantes" + -0.242*"conhecimento" + 0.195*"memória" + -0.174*"jongo" + 0.172*"social" + -0.149*"gente" + -0.137*"organização" + -0.135*"biblioteca" + -0.115*"anos" + -0.107*"pessoas";
Tópico 21: 0.216*"pessoas" + 0.208*"vida" + -0.173*"livro" + 0.152*"participantes" + 0.147*"cegueira" + 0.135*"conhecimento" + -0.128*"biblioteca" + 0.123*"tam" + 0.121*"museu" + 0.120*"narrativa";
Tópico 22: 0.209*"memória" + 0.168*"negro" + -0.163*"repórter" + -0.162*"documentos" + 0.128*"social" + 0.123*"indígenas" + 0.122*"fonte" + 0.122*"contas" + -0.121*"off" + -0.121*"comunicação";
Tópico 23: 0.154*"processo" + 0.145*"trabalho" + -0.136*"publicidade" + -0.123*"arquivo" + -0.123*"livro" + 0.121*"gestão" + -0.121*"comunicação" + 0.117*"ambiente" + -0.116*"contas" + -0.115*"pesquisa";
Tópico 24: -0.313*"estudantes" + -0.199*"brasil" + 0.155*"biblioteca" + 0.135*"sistema" + -0.130*"processo" + -0.127*"estrangeiros" + -0.114*"estudantes_estrangeiros" + 0.113*"repórter" + 0.113*"organização" + 0.111*"conhecimento";
Tópico 25: -0.327*"digital" + -0.173*"preservação" + -0.169*"competências" + -0.129*"digitais" + 0.121*"educação" + 0.116*"processo" + 0.115*"dados" + 0.111*"social" + 0.109*"arquivo" + -0.108*"pessoas".

Fonte: Elaborado pelo autor.

Dentre o grupo de termos extraídos do corpus de dados utilizando o modelo LSI é possível identificar tópicos generalista que apresentam termos gerais ao domínio de linguagem, sendo necessário ao especlista analisar documentos externos como lista de bigramas, trigramas ou mesmo acesso ao

corpus de dados para identificar o assunto e criar uma suposição de rótulos mais assertiva para o tópico. Para este tipo de situação, pode ser realizado o descarte de tópicos com características fracas como forma de não atribuir informações irrelevantes – levando em consideração o fator subjetividade, durante a análise de assunto ou criar uma categoria geral tópicos com essas cacarterísticas.

Ainda é possível encontrar tópicos contendo termos chave que servem como norte para o especialista explorar os documentos contidos no *corpus* de dados, como por exemplo os tópicos 6 e 7 que apresentam os termos “feed”, “feed_noticias” e “facebook”, constando assim somente um documento na coleção de documentos que aborde tais termos, podendo supor por exemplo que o tópico aborde assuntos eferente a Plataformas Digitais. Faz-se necessário ressaltar que o modelo apresentou termos iguais em tópicos diferentes, misturando assuntos e não possibilitando a extração de novos termos. Isso ocorre porque a quantidade de tópicos configurado para extração é superior ao necessário para o tamanho *corpus*.

Os termos que apresentam características específicas, diferentes dos termos generalistas encontrados no domínio de linguagem, acabam por apresentar maior facilidade na identificação dos tópicos e consequentemente na definição da suposição de um rótulo de maneira mais assertiva, como por exemplo nos tópicos 17 e 18 que apresentam os termos “jomgo” e “negro”, podendo ao especialista supor por exemplo que o tópico, junto com os demais termos, refere-se ao rótulo Memória e Patrimônio ou Patrimônio Imaterial. Outro termo chave encontrado está nos tópicos 11 e 14, sendo “namitec” que permite remeter estudos de e-science na área de Nanotecnologia.

Também é possível identificar termos como “*ontology*” encontrados nos tópicos 7, 9, 11 e 13 que se misturam a outros termos como “facebook” e “jomgo”, mas que não estão correlacionados, apresentando assim uma fragilidade do modelo ao misturar termos distintos em tópicos. Dentre os tópicos citados, o tópico 9 apresenta termos mais coesos como “ontology”, “ontologies” e “conhecimento”, podendo supor pelo especialista que o assunto esteja relacionado por exemplo a Representação do Conhecimento.

O Quadro 37 apresenta os resultados da modelagem de tópicos extraídos através do modelo LDA do *corpus* 6. São exibidos um conjunto de 26 tópicos contendo termos e pesos. O modelo realizou o processo de treinamento durante

34 minutos e 52 segundos. Os resultados contendo 10, 14, 18, 22, 30, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[66].

Quadro 37 – Tópicos extraídos do *corpus* 6 usando o modelo LDA

Tópico 0: 0.001*"editores" + 0.001*"livro_digital" + 0.001*"editorial" + 0.001*"revistas" + 0.000*"conduta" + 0.000*"mayer" + 0.000*"editoras" + 0.000*"multimídia" + 0.000*"maria_anjos" + 0.000*"livros_digitais";
Tópico 1: 0.000*"informação" + 0.000*"memória" + 0.000*"pesquisa" + 0.000*"social" + 0.000*"dados" + 0.000*"forma" + 0.000*"trabalho" + 0.000*"processo" + 0.000*"sociais" + 0.000*"meio";
Tópico 2: 0.010*"informação" + 0.004*"pesquisa" + 0.003*"conhecimento" + 0.003*"forma" + 0.003*"dados" + 0.002*"processo" + 0.002*"social" + 0.002*"relação" + 0.002*"produção" + 0.002*"trabalho";
Tópico 3: 0.000*"informação" + 0.000*"conhecimento" + 0.000*"pesquisa" + 0.000*"dados" + 0.000*"forma" + 0.000*"organização" + 0.000*"brasil" + 0.000*"uso" + 0.000*"trabalho" + 0.000*"relação";
Tópico 4: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"processo" + 0.000*"dados" + 0.000*"forma" + 0.000*"social" + 0.000*"signo" + 0.000*"relação" + 0.000*"espaço" + 0.000*"interpretante";
Tópico 5: 0.002*"exposição" + 0.001*"exposições" + 0.001*"documentação" + 0.001*"pinacoteca" + 0.001*"masp" + 0.001*"centros_memória" + 0.001*"museu_arte" + 0.001*"centros" + 0.000*"cubanos" + 0.000*"rhodia";
Tópico 6: 0.000*"informação" + 0.000*"documentos" + 0.000*"conhecimento" + 0.000*"gestão" + 0.000*"pesquisa" + 0.000*"arquivo" + 0.000*"processo" + 0.000*"meio" + 0.000*"forma" + 0.000*"arquivos";
Tópico 7: 0.005*"universidade" + 0.002*"revista" + 0.002*"universidade_federal" + 0.002*"humanas" + 0.002*"sudeste" + 0.002*"sul" + 0.002*"federal" + 0.001*"humanas_universidade" + 0.001*"sociais" + 0.001*"nordeste";
Tópico 8: 0.002*"namitec" + 0.001*"incts" + 0.001*"data" + 0.001*"inct" + 0.001*"nanocarbono" + 0.001*"nanotecnologia" + 0.001*"inglês" + 0.001*"nanobiotecnologia" + 0.001*"e-science" + 0.001*"colaborativas";
Tópico 9: 0.001*"cegueira" + 0.001*"tam" + 0.001*"jet" + 0.001*"olga" + 0.001*"claudia" + 0.001*"mirp" + 0.000*"excerto" + 0.000*"rodas" + 0.000*"rodas_conversa" + 0.000*"messias";
Tópico 10: 0.003*"museologia" + 0.001*"icofom" + 0.001*"icom" + 0.001*"jongo" + 0.001*"scheiner" + 0.000*"museology" + 0.000*"bourdieu" + 0.000*"museologia_patrimônio" + 0.000*"desvallées" + 0.000*"museum";
Tópico 11: 0.006*"des" + 0.006*"une" + 0.004*"nous" + 0.004*"est" + 0.003*"qui" + 0.003*"pour" + 0.003*"dans" + 0.002*"patrimoine" + 0.002*"sur" + 0.002*"par";
Tópico 12: 0.002*"santo" + 0.001*"festa" + 0.001*"santo_antônio" + 0.001*"antônio" + 0.001*"administrativo" + 0.001*"processo_administrativo" + 0.001*"caxias" + 0.001*"duque_caxias" + 0.001*"duque" + 0.001*"aldeia";
Tópico 13: 0.000*"informação" + 0.000*"processo" + 0.000*"museu" + 0.000*"pesquisa" + 0.000*"museologia" + 0.000*"social" + 0.000*"brasil" + 0.000*"forma" + 0.000*"campo" + 0.000*"produção";
Tópico 14: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"conhecimento" + 0.000*"forma" + 0.000*"dados" + 0.000*"processo" + 0.000*"relação" + 0.000*"social" + 0.000*"pessoas" + 0.000*"busca";

[66] Algoritmo de modelagem de tópicos. *Corpus* 6: teses e dissertações 2017. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_tesesdissertacoes_2017.ipynb/.

Tópico 15: 0.002*"taxonomia" + 0.001*"auditoria" + 0.001*"contábeis" + 0.001*"riscos" + 0.001*"procedimentos" + 0.001*"auditor" + 0.001*"distorções" + 0.001*"risco" + 0.001*"demonstrações" + 0.000*"inspeção";
Tópico 16: 0.001*"jurados" + 0.001*"loucura" + 0.001*"mental" + 0.000*"rancière" + 0.000*"bispo" + 0.000*"sic" + 0.000*"serviços_mental" + 0.000*"bispo_rosário" + 0.000*"arte_loucura" + 0.000*"jurado";
Tópico 17: 0.003*"jongo" + 0.002*"indígenas" + 0.002*"negro" + 0.001*"serrinha" + 0.001*"charrua" + 0.001*"imaterial" + 0.001*"indígena" + 0.001*"povos" + 0.001*"energia" + 0.001*"clarisse";
Tópico 18: 0.009*"informação" + 0.002*"ontology" + 0.001*"revista" + 0.001*"científica" + 0.001*"maria" + 0.001*"pesquisa" + 0.001*"conhecimento" + 0.001*"silva" + 0.001*"usabilidade" + 0.001*"produção_científica";
Tópico 19: 0.000*"informação" + 0.000*"brasil" + 0.000*"pesquisa" + 0.000*"revista" + 0.000*"conhecimento" + 0.000*"universidade" + 0.000*"produção" + 0.000*"forma" + 0.000*"dados" + 0.000*"des";
Tópico 20: 0.003*"feed" + 0.002*"contas" + 0.002*"feed_notícias" + 0.002*"facebook" + 0.002*"notícias" + 0.001*"texto_original" + 0.001*"original" + 0.001*"algoritmo" + 0.001*"quadrinhos" + 0.001*"tradução";
Tópico 21: 0.001*"ilha" + 0.000*"anchieta" + 0.000*"ilha_anchieta" + 0.000*"filhos_ilha" + 0.000*"correcional" + 0.000*"detentos" + 0.000*"instituto_correcional" + 0.000*"motim" + 0.000*"rebelião" + 0.000*"correcional_ilha";
Tópico 22: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"organização" + 0.000*"conhecimento" + 0.000*"processo" + 0.000*"forma" + 0.000*"trabalho" + 0.000*"brasil" + 0.000*"meio" + 0.000*"pessoas";
Tópico 23: 0.000*"informação" + 0.000*"processo" + 0.000*"conhecimento" + 0.000*"forma" + 0.000*"tempo" + 0.000*"meio" + 0.000*"comunicação" + 0.000*"pesquisa" + 0.000*"dados" + 0.000*"social";
Tópico 24: 0.001*"deficiência" + 0.001*"deficiência_visual" + 0.001*"cartas" + 0.001*"amor" + 0.001*"pessoas_deficiência" + 0.001*"audiodescrição" + 0.001*"acessibilidade" + 0.001*"audiodescritiva" + 0.001*"informação_audiodescritiva" + 0.001*"carta";
Tópico 25: 0.000*"forma" + 0.000*"informação" + 0.000*"social" + 0.000*"produção" + 0.000*"trabalho" + 0.000*"une" + 0.000*"relação" + 0.000*"meio" + 0.000*"tempo" + 0.000*"pesquisa".

Fonte: Elaborado pelo autor.

O modelo LDA apresentou junto aos resultados, tópicos caracterizados com um maior número de N-gramas do tipo bigramas e trigramas quando se comparado ao modelo LSI. Essa é uma das características que possibilita diminuir o esforço cognitivo do especialista ao realizar a suposição do nome do tópico, uma vez que, de acordo com o grupo de termos, não faz-se necessário consultar documentos externos como lista de bigramas ou trigramas, bem como o próprio *corpus* de dados. Além disso, os tópicos apresentam uma quantidade menor de termos generalistas do domínio da linguagem quando comparado ao modelo LSI. São exemplos de termos generalistas “processo” e “informação” que podem se compor para outros termos com outros significados como “processo_admistrativo”, “processo_busca_informação”, “informação_jurídica” ou “informação_documentação”. Uma característica fraca apresentada nos

resultados do modelo são os termos com pesos de baixo valor ou valores iguais a zero, dificultando a interpretação dos dados por não apresentar uma ordem de relevância entre os termos.

Os tópicos 2, 3, 22 e 23 apresentam termos generalistas ao domíno de linguagem, uma vez que todos os termos podem gerar diversas composições de significados. Nesses casos, cabe ao especialista decidir se explora os dados externos como a lista de N-gramas ou o próprio *corpus* de dados para realizar intepretações dos dados, se cria uma categoria de classificação geral para os tópicos com tais características ou se realiza o descarte dos tópicos.

O tópico 5 apresenta dois grupos de termos, sendo o primeiro formado por 8 termos e pesos agrupado por aproximação de assuntos e o segundo contendo 2 termos com os valores dos pesos zerados. Dessa forma, o especialista pode optar por definir o rótulo que melhor representa apenas o primeiro grupo de termos contidos no tópico, descartando assim os termos com pesos zerados. Os termos com maior representatividade deste tópico são “exposição”, “documentação”, “centros_memória”, “museu_arte” que permite ao especialista supor por exemplo que o tópico esteja relacionado a Museu, Patromonio e Informação além de termos chave como “pinacoteca” e “masp” que permite identificar o documento junto ao *corpus* de dados expandindo o rótulo do tópico para Documentação e Museologia.

O tópico 15 possui termos chave com pesos equilibrados que possibilita ao especialista realizar a definição da suposição de um rótulo para o tópico com pouco esforço cognitivo. Dentre esses termos estão “taxonomia”, “auditoria” e “contábeis” que podem ser classificados na área de Organização e Representação do Conhecimento.

APÊNDICE G - *Corpus* 8: artigos completos e resumos expandidos 2013

Também constituindo do segundo *corpora* de dados, o *corpus* 8 possui 303 documentos do tipo artigos completos e resumos expandidos publicados nos anais do ENANCIB no ano de 2013 e tamanho de 11.359kb. Os dados desse corpus resultam em 800.270 unigramas, 799.967 trigramas e 799.664. O Quadro 38 apresenta uma lista com os 50 termos mais frequentes de cada tipo de N-grama.

Quadro 38 – Lista de N-gramas por ordem de frequência do *corpus* 8

Unigramas
informação,17936; pesquisa,4982; conhecimento,4272; dados,2896; forma,2556; social,2413; processo,2348; museu,2039; comunicação,2014; uso,2002; organização,1923; biblioteca,1922; documentos,1909; relação,1801; produção,1775; trabalho,1768; sociais,1765; memória,1715; estudo,1694; sociedade,1642; brasil,1619; sistema,1584; desenvolvimento,1540; estudos,1536; contexto,1525; gestão,1504; usuários,1494; meio,1453; web,1446; termo,1443; científica,1439; campo,1425; busca,1377; tecnologia,1360; paulo,1326; instituições,1297; modelo,1278; relações,1273; rede,1233; construção,1232; autores,1208; tempo,1203; diferentes,1200; base,1191; nacional,1146; representação,1134; digital,1131; educação,1128; informacional,1111; sistemas,1098.
Bigramas
universidade_federal,2895; recuperação_informação,2728; redes_sociais,2542; uso_informação,2385; informação_conhecimento,2329; belo_horizonte,1943; gestão_informação,1825; biblioteca_universitárias,1804; ensino_superior,1794; ponto_vista,1757; coleta_dados,1694 muitas_vezes,1644 fontes_informação,1638; patrimônio_cultural,1613; dissertação_mestrado,1600; sistemas_informação,1579; informação_tecnologia,1572; gestão_documentos,1517; produção_científica,1494; comunicação_científica,1493; informação_science,1491; tecnologias_informação,1406; competência_informacional,1404; informação_informação,1396; tomada_decisão,1305; informação_comunicação,1303; minas_gerais,1236; arquitetura_informação,1231; bases_dados,1226; busca_informação,1204; sociedade_informação,1197; biblioteca_universitária,1185; organização_conhecimento,1182; porto_alegre,1151; dados_pesquisa,1117; base_dados,1106; necessidades_informação,1081; tendo_vista,1071; biblioteca_digitais,1060; organização_informação,1010; comportamento_informacional,993; sistema_informação,981; biblioteca_digital,979; memória_social,969; produtos_serviços,935; gestão_conhecimento,916; informação_brasília,913; fonte_dados,913; educação_distância,907; qualidade_informação,881.
Trigramas
tecnologias_informação_comunicação,145; fonte_dados_pesquisa,125; resource_description_framework,119; informação_comunicação_oral,107; gestão_informação_conhecimento,106; society_informação_science,97; fundação_oswaldo_cruz,94; american_society_informação,93; journal_american_society,90; organização_representação_conhecimento,85; nacional_pesquisa_informação,84; encontro_nacional_pesquisa,82; dissertação_mestrado_informação,82; portal_periódicos_capes,80; informação_science_technology,74; informação_belo_horizonte,66; instituição_ensino_superior,60; instituições_ensino_superior,56; busca_uso_informação,56;

informação_universidade_federal,54; informação_tecnologia_was,53;
modelo_conceitual_frbr,52; machine_readable_cataloging,52;
universidade_federal_paraíba,48; extensible_markup_language,47;
sistema_recuperação_informação,46; produção_comunicação_informação,46;
sistemas_organização_conhecimento,45; patrimônio_histórico_artístico,45;
circulação_apropriação_informação,44; brasília_briquet_lemos,44;
informação_sociedade_estudos,42; federal_minas_gerais,42;
universidade_federal_bahia,41; histórico_artístico_nacional,41;
instituições_arquivísticas_nacionais,41; universidade_federal_minas,40;
library_informação_science,39; sistemas_recuperação_informação,39;
perspectivas_informação_belo,39; poder_executivo_federal,39; ciclo_vida_dados,39;
fonte_elaboração_autores,38; mediação_circulação_apropriação,38;
datagramazero_revista_informação,38; universidade_federal_santa,38;
informação_conhecimento_organizações,37; educação_superior_distância,37;
sob_ponto_vista,36

Fonte: Elaborado pelo autor.

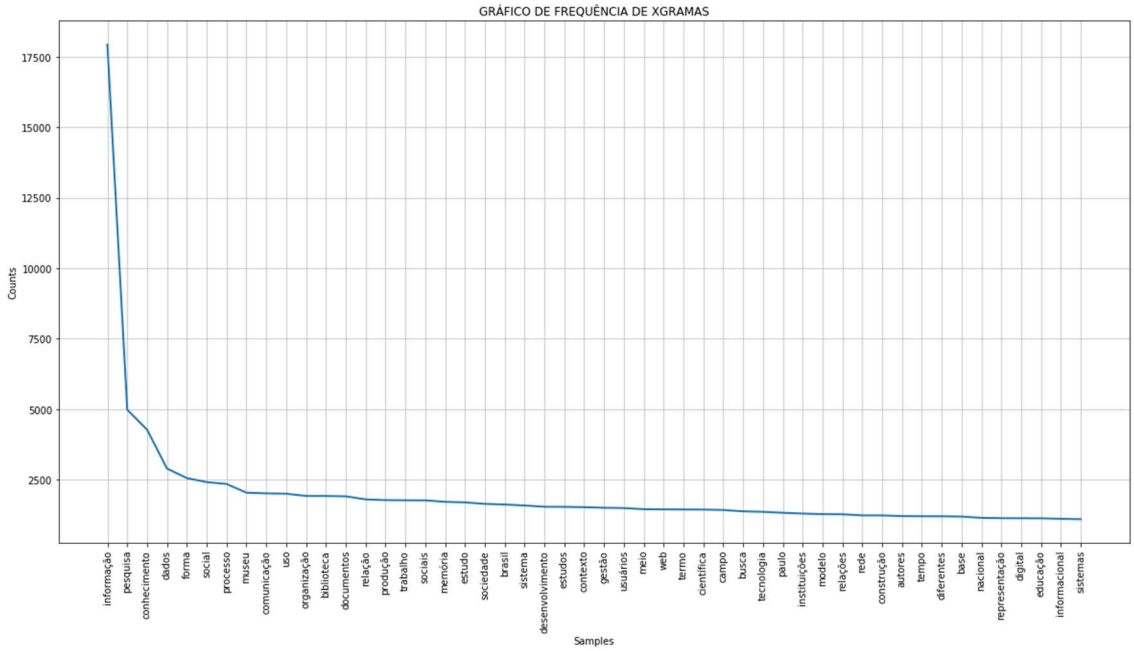
Em uma lista geral dos mil termos mais frequentes do *corpus* de dados pode-se destacar os bigramas mais frequentes, sendo eles, “informação_science” com frequência de 530 e ocupando a posição 175, “informação_tecnologia” com frequência de 496 e posição 196, “recuperação_informação” com frequência de 431 na posição 255, “redes_sociais” com frequência de 427 e posição 257 e “informação_conhecimento” com frequência de 418 na posição 267. Já todos os trigramas aparecem após ao milésimo termo desta lista, sendo os mais ranqueados “tecnologias_informação_comunicação” com frequência de 145, “fonte_dados_pesquisa” com frequência de 125, “resource_description_framework” com frequência de 119, “informação_comunicação_oral” com frequência de 107 e “gestão_informação_conhecimento” com frequência de 106.

Os termos do tipo unigramas apresentam uma maior frequência quando se comparado aos demais tipos de N-gramas, por exemplo, o primeiro unigrama representado pelo termo “informação” e com frequência de 17.936 possui uma diferença equivalente a 520% quando se comparado ao primeiro bigrama representado pelo termo “universidade_federal” e frequência de 2.895. Esse percentual alcança 12.272% quando comparado o primeiro unigrama com o primeiro trigrama representado pelo termo “tecnologias_informação_comunicação” e frequência de 145.

O Gráfico 58 apresenta os 50 termos com maior frequência extraído do *corpus* 8. Constam dentre os resultados somente N-gramas do tipo unigramas,

sendo os primeiros representados pelos termos “informação” com frequência de 17.936, “pesquisa” com frequência de 4.982, “conhecimento” com frequência de 4.272, “dados” com frequência de 2896; “forma” com frequência de 2.556 e “social” com frequência de 2.413.

Gráfico 58 – Termos mais frequentes do *corpus* 8



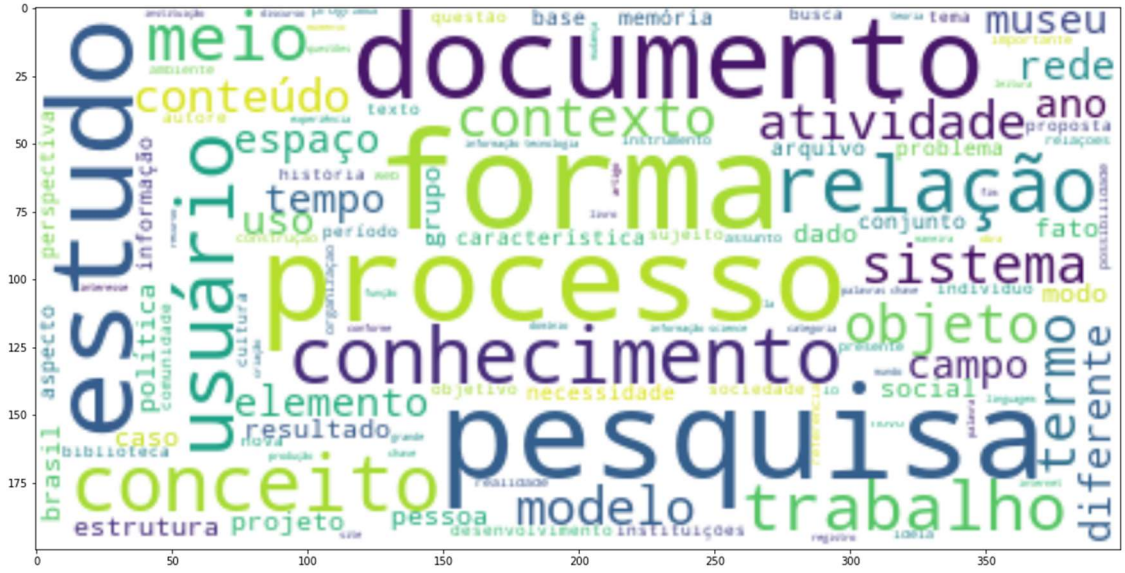
Fonte: Elaborado pelo autor.

Ainda no gráfico é possível perceber um crescimento constante de frequência entre o 50º e o 2º termo, sendo respectivamente “sistemas” com frequência de 1.098 e “pesquisa” com frequência de 4.892. Do segundo para o primeiro termo representado por “informação” e frequência de 17.936 existe um uma diferença que alcança 260%. Também é possível perceber a existência de termos fortes como “sociedade”, “desenvolvimento”, “gestão”, “memória” e “científica” que podem ser melhores interpretados pelo especialista ao utilizar os documentos externos gerados pelos algoritmos como as listas de bigramas e trigramas. Também é possível identificar termos fracos como “diferentes”, “autores”, “paulo’, “meio” e “brasil” que podem ser descartados por não fazem parte ou ter pouca relevância com a linguagem de domínio estudada.

Com base nos termos extraídos no *corpus* de dados foi possível elaborar uma nuvem de palavras contendo os 250 N-gramas mais frequentes conforme apresentado na Figura 19. Pode-se observar dentre os resultados a existêncis de termos fortes e fracos relacionados ao domínio de linguagem, sendo importante ressaltar que os dados representados na figura são quantitativos, não

sendo realizado quaisquer análises e ações qualitativas, como por exemplo a exclusão de termos irrelevantes ou inserção de termos que possuam maior significado junto ao dom[inio de linguagem.

Figura 19 – Nuvem de palavras do *corpus* 8



Fonte: Elaborado pelo autor.

São destacados como termos fortes os unigramas “estudo”, “conceito” e “política” que apresentam conceitos relevantes ao domínio da linguagem, mesmo os termos possuindo vertentes para outros significados como encontrados nos bigramas ou trigramas. Entre os termos fracos constam “meio”, “contexto” e “atividade” que não apresentam ou apresentam baixa relevância quando relacionamento com o domínio da linguagem. O *corpus* 8 possui uma maior quantidade de termos fortes quando se comparado ao *corpora* 1.

Após a fase de transformação, os dados foram conectados aos modelos de extração de tópicos. O Quadro 39 apresenta um conjunto de resultados constituídos por 30 tópicos e seus respectivos termos e pesos que foram extraídos utilizando o modelo LSI. Para alcançar o resultado, o modelo foi executado durante 30.5 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do *GitHub*[67].

[67] Algoritmo de modelagem de tópicos. *Corpus* 8: artigos completos e resumos expandidos 2013. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2013.ipynb/.

Quadro 39 – Tópicos extraídos do *corpus* 8 usando o modelo LSI

Tópico 0: 0.776*"informação" + 0.176*"pesquisa" + 0.172*"conhecimento" + 0.096*"dados" + 0.087*"social" + 0.086*"processo" + 0.085*"forma" + 0.076*"uso" + 0.073*"organização" + 0.071*"comunicação";
Tópico 1: 0.573*"museu" + -0.405*"informação" + 0.165*"memória" + 0.160*"patrimônio" + 0.113*"museologia" + 0.111*"brasil" + 0.108*"cultural" + 0.105*"nacional" + 0.104*"instituições" + 0.102*"história";
Tópico 2: 0.409*"museu" + -0.364*"conhecimento" + 0.343*"informação" + -0.267*"dados" + -0.178*"pesquisa" + -0.157*"web" + -0.143*"biblioteca" + -0.094*"organização" + -0.094*"documentos" + -0.087*"representação";
Tópico 3: 0.691*"conhecimento" + -0.203*"biblioteca" + -0.201*"web" + -0.191*"dados" + 0.165*"gestão" + -0.163*"usuários" + 0.103*"científico" + 0.094*"organização" + 0.093*"conhecimento_científico" + 0.091*"gestão_conhecimento";
Tópico 4: 0.480*"biblioteca" + -0.192*"documentos" + 0.150*"educação" + -0.149*"web" + 0.142*"pesquisa" + -0.128*"representação" + 0.124*"sociais" + 0.120*"biblioteconomia" + -0.119*"dados" + 0.119*"marketing";
Tópico 5: -0.380*"biblioteca" + 0.306*"pesquisa" + -0.215*"museu" + 0.214*"científica" + 0.185*"produção" + -0.181*"usuários" + 0.144*"artigos" + 0.136*"autores" + -0.127*"web" + -0.106*"marketing";
Tópico 6: 0.391*"museu" + -0.369*"memória" + 0.235*"dados" + -0.234*"documentos" + 0.178*"web" + 0.167*"pesquisa" + -0.159*"arquivo" + -0.121*"arquivos" + -0.120*"história" + 0.109*"científica";
Tópico 7: 0.271*"documentos" + -0.256*"sociais" + -0.252*"rede" + -0.226*"redes" + -0.214*"social" + 0.185*"arquivo" + 0.171*"gestão" + 0.162*"arquivos" + -0.136*"redes_sociais" + 0.118*"federal";
Tópico 8: -0.268*"biblioteconomia" + 0.196*"gestão" + 0.191*"redes" + 0.190*"rede" + 0.188*"web" + 0.149*"sociais" + 0.136*"arquivos" + -0.134*"curso" + 0.122*"política" + 0.121*"redes_sociais";
Tópico 9: 0.360*"memória" + 0.261*"pesquisa" + -0.198*"rede" + -0.193*"documentos" + -0.190*"redes" + -0.161*"classificação" + 0.160*"dados" + -0.152*"biblioteconomia" + -0.147*"sociais" + 0.128*"digital";
Tópico 10: -0.310*"biblioteconomia" + -0.249*"web" + 0.221*"usuários" + -0.213*"curso" + -0.179*"dados" + -0.159*"cursos" + 0.140*"estudos" + -0.140*"ensino" + 0.130*"científica" + 0.122*"biblioteca";
Tópico 11: -0.310*"pesquisa" + 0.278*"memória" + 0.248*"web" + 0.171*"citação" + 0.147*"autores" + 0.144*"conhecimento" + 0.132*"livros" + -0.129*"processo" + 0.127*"usuários" + -0.119*"curso";
Tópico 12: 0.358*"digital" + 0.311*"livros" + 0.258*"digitais" + 0.225*"editoras" + 0.202*"livro" + -0.192*"memória" + -0.162*"web" + 0.161*"livros_digitais" + -0.154*"pesquisa" + -0.119*"usuários";
Tópico 13: 0.378*"direito" + -0.178*"digital" + -0.168*"rede" + -0.154*"redes" + -0.143*"biblioteconomia" + 0.142*"biblioteca" + -0.138*"memória" + 0.134*"direito_informação" + 0.122*"política" + 0.122*"mundo";
Tópico 14: -0.206*"web" + -0.182*"pesquisa" + 0.165*"dados" + 0.142*"modelo" + -0.132*"classificação" + 0.131*"marketing" + 0.128*"biblioteca" + -0.126*"livros" + 0.123*"artigos" + 0.121*"patrimônio";
Tópico 15: 0.210*"pesquisa" + 0.166*"ontologias" + -0.165*"decisão" + -0.139*"direito" + -0.132*"processo" + 0.123*"biblioteca" + -0.119*"citação" + 0.110*"tecnologia" + -0.109*"autores" + -0.105*"knee";
Tópico 16: 0.203*"ontologias" + -0.185*"direito" + 0.167*"knee" + 0.158*"prosthesis" + 0.135*"termo" + 0.134*"knee_prosthesis" + 0.129*"ontologia" + 0.128*"nacional" + -0.122*"objeto" + 0.118*"mesh";

Tópico 17: 0.254*"direito" + 0.197*"classificação" + 0.167*"organização" + -0.166*"processo" + -0.158*"usuários" + 0.144*"memória" + -0.142*"modelo" + 0.134*"biblioteca" + 0.122*"museu" + -0.115*"tecnologia";
Tópico 18: 0.309*"dados" + -0.149*"ontologias" + -0.147*"modelo" + -0.139*"pesquisa" + 0.133*"digital" + -0.133*"frbr" + -0.123*"entidades" + 0.120*"objeto" + 0.118*"preservação" + 0.115*"vida";
Tópico 19: -0.216*"classificação" + 0.183*"arquivo" + 0.170*"knee" + 0.160*"prosthesis" + 0.147*"arquivos" + -0.146*"organização" + -0.146*"usuários" + 0.136*"knee_prosthesis" + -0.126*"nacional" + -0.126*"tempo";
Tópico 20: 0.262*"direito" + -0.189*"marketing" + -0.187*"processo" + 0.163*"usuários" + 0.155*"knee" + 0.146*"prosthesis" + -0.146*"ontologias" + 0.130*"usuário" + 0.124*"knee_prosthesis" + 0.121*"obra";
Tópico 21: 0.279*"direito" + 0.235*"ontologias" + -0.199*"classificação" + -0.162*"educação" + 0.135*"ontologia" + 0.133*"biblioteconomia" + 0.113*"usuários" + 0.108*"direito_informação" + 0.106*"redes" + 0.104*"marketing";
Tópico 22: -0.293*"marketing" + 0.181*"periódicos" + 0.159*"uso" + 0.152*"dados" + 0.141*"educação" + -0.137*"web" + 0.133*"portal" + -0.127*"knee" + 0.123*"usuários" + -0.120*"prosthesis";
Tópico 23: 0.187*"patrimônio" + -0.169*"científica" + 0.162*"leitura" + -0.137*"universidades" + 0.132*"livros" + 0.130*"pesquisa" + -0.129*"sistema" + 0.127*"usuários" + -0.127*"distância" + -0.120*"produção";
Tópico 24: -0.218*"patrimônio" + 0.217*"leitura" + -0.215*"direito" + -0.162*"preservação" + 0.142*"imagens" + -0.134*"classificação" + -0.133*"industrial" + 0.128*"imagem" + 0.125*"projetos" + -0.116*"processo";
Tópico 25: -0.223*"imagens" + -0.185*"imagem" + -0.166*"patrimônio" + 0.157*"memória" + 0.153*"social" + 0.144*"marketing" + 0.134*"classificação" + -0.132*"objetos" + -0.122*"objeto" + 0.119*"modelo";
Tópico 26: -0.177*"citação" + 0.162*"produção" + 0.142*"artigos" + -0.139*"teses" + -0.136*"política" + -0.127*"dados" + 0.126*"lei" + -0.118*"rede" + 0.107*"marketing" + 0.103*"científica";
Tópico 27: 0.164*"processo" + 0.156*"leitura" + -0.139*"competências" + 0.139*"direito" + -0.138*"ontologias" + -0.133*"competência" + 0.125*"sistema" + 0.119*"imagens" + 0.117*"submissão" + 0.111*"curso";
Tópico 28: -0.205*"tempo" + 0.195*"projetos" + 0.177*"sistema" + 0.156*"termo" + 0.143*"citação" + -0.126*"periódicos" + -0.126*"marketing" + -0.125*"informacional" + -0.112*"científica" + -0.101*"nacional";
Tópico 29: -0.218*"tempo" + 0.185*"capital" + -0.181*"leitura" + 0.165*"marketing" + 0.154*"imagens" + 0.141*"trabalho" + 0.131*"imagem" + 0.120*"autores" + -0.102*"lei" + 0.095*"campo".

Fonte: Elaborado pelo autor.

Os resultados obtidos através do modelo LSI apresentam dentre os seus tópicos, termos de relevância com pesos acima de 0.500 como “informação” no tópico 0, “museu” no tópico 1 e “conhecimento” no tópico 3. Os termos de cada tópico são ordenados de acordo com os valores dos pesos, sendo os maiores valores os mais representativos. O tópico 0 é constituído de termos generalistas encontrados no domínio da linhagem, bem como “pesquisa” que pode representar “pesquisa_cientifica”, “dados_pesquisa” ou “pesquisa_informação”. Cabe ao especialista explorar os documentos externos gerados a partir dos

algoritmos como lista de bigramas, trigramas ou mesmo o *corpus* de dados para definir uma suposição do nome do tópico de maneira mais assertiva. Também é possível criar uma categoria de classificação geral para o tópico com essas características ou mesmo realizar o descarte dos dados.

O tópico 1 possui características fortes em seus termos e pesos de forma que o especialista possa realizar a suposição do nome do tópico sem a necessidade de consultar documentos externos. Exemplos desses termos são “museu”, “informação” e “patrimônio”, além dos demais termos estarem coesos com os termos apresentados como “memória”, “museologia” e “cultural” acabam por remeter ao especialista a suposição de um rótulo como por exemplo Museu, Informação e Patrimônio na Ciência da Cnformação.

Termos específicos também possibilitam uma interpretação mais assertiva pelo especialista, como por exemplo os tópicos 16 e 19 que apresentam termos como “knee” e “prosthesis”, além de termos chave como “organização”, “classificação” e “ontologia” que remetem as áreas de Informação e Saúde e/ou Organização e Representação do Conhecimento.

O Quadro 40 apresenta os resultados da modelagem de tópicos utilizando o modelo LDA, apresentando assim um conjunto de 30 tópicos com os respectivos termos e pesos. O processo de treinamento foi realizado durante 6 minutos e 39 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do *GitHub*[68].

Quadro 40 – Tópicos extraídos do *corpus* 8 usando o modelo LDA

Tópico 0: 0.000*"informação" + 0.000*"museu" + 0.000*"memória" + 0.000*"pesquisa" + 0.000*"social" + 0.000*"forma" + 0.000*"conhecimento" + 0.000*"documentos" + 0.000*"nacional" + 0.000*"autores";
Tópico 1: 0.002*"livro" + 0.001*"eletrônico" + 0.001*"livro_eletrônico" + 0.001*"informação" + 0.001*"documentos" + 0.001*"programa" + 0.001*"ufes" + 0.001*"papel" + 0.001*"suporte" + 0.001*"e-book";
Tópico 2: 0.005*"informação" + 0.002*"tempo" + 0.001*"conhecimento" + 0.001*"pesquisa" + 0.001*"organização" + 0.001*"medicamentos" + 0.001*"atores" + 0.001*"produção" + 0.001*"rede" + 0.001*"campo";
Tópico 3: 0.002*"samba" + 0.000*"discursiva" + 0.000*"letra" + 0.000*"canto" + 0.000*"letras_samba" + 0.000*"letra_samba" + 0.000*"discursivas" + 0.000*"carnaval" + 0.000*"firme" + 0.000*"firme_forte";

[68] Algoritmo de modelagem de tópicos. *Corpus* 8: artigos completos e resumos expandidos 2013. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2013.ipynb/.

Tópico 4: 0.006*"informação" + 0.002*"usuários" + 0.002*"pesquisa" + 0.002*"estudos" + 0.001*"uso" + 0.001*"leitura" + 0.001*"biblioteca" + 0.001*"classificação" + 0.001*"processo" + 0.001*"dados";
Tópico 5: 0.001*"dados" + 0.001*"publicação" + 0.001*"publicações" + 0.001*"recursos" + 0.000*"publicações_ampliadas" + 0.000*"ampliadas" + 0.000*"modelo" + 0.000*"digitais" + 0.000*"publicação_ampliada" + 0.000*"ampliada";
Tópico 6: 0.001*"paisagem" + 0.001*"objetos" + 0.001*"material" + 0.001*"objeto" + 0.001*"geografia" + 0.001*"coisas" + 0.001*"usuários" + 0.000*"cultura" + 0.000*"educação" + 0.000*"ibge";
Tópico 7: 0.005*"biblioteca" + 0.003*"informação" + 0.002*"redes" + 0.002*"sociais" + 0.001*"redes_sociais" + 0.001*"educação" + 0.001*"cultura" + 0.001*"pesquisa" + 0.001*"distância" + 0.001*"rede";
Tópico 8: 0.005*"informação" + 0.002*"direito" + 0.001*"crecimiento" + 0.001*"direito_informação" + 0.001*"paradigma" + 0.001*"conhecimento" + 0.001*"processo" + 0.001*"brasil" + 0.001*"abordagem" + 0.001*"campo";
Tópico 9: 0.003*"informação" + 0.001*"sintagmas" + 0.001*"nominais" + 0.001*"sintagmas_nominais" + 0.001*"repositórios" + 0.001*"pesquisa" + 0.001*"deputados" + 0.001*"portugal" + 0.001*"repositório" + 0.001*"livre";
Tópico 10: 0.011*"informação" + 0.003*"conhecimento" + 0.002*"pesquisa" + 0.002*"social" + 0.001*"science" + 0.001*"organização" + 0.001*"processo" + 0.001*"forma" + 0.001*"estudos" + 0.001*"tecnologia";
Tópico 11: 0.002*"informação" + 0.001*"arquivos" + 0.001*"cfdt" + 0.001*"memória" + 0.001*"dados" + 0.001*"scielo" + 0.001*"arquivo" + 0.001*"documentos" + 0.000*"fonte" + 0.000*"acervo";
Tópico 12: 0.007*"informação" + 0.002*"pesquisa" + 0.002*"processo" + 0.002*"dados" + 0.001*"conhecimento" + 0.001*"modelo" + 0.001*"forma" + 0.001*"organização" + 0.001*"uso" + 0.001*"web";
Tópico 13: 0.002*"autores" + 0.001*"cocitação" + 0.001*"acessibilidade" + 0.001*"usuários" + 0.001*"relativos" + 0.001*"valores" + 0.001*"informação" + 0.001*"domínio" + 0.001*"citados" + 0.001*"estudos";
Tópico 14: 0.012*"informação" + 0.004*"pesquisa" + 0.002*"dados" + 0.002*"científica" + 0.001*"conhecimento" + 0.001*"sistema" + 0.001*"comunicação" + 0.001*"produção" + 0.001*"forma" + 0.001*"web";
Tópico 15: 0.001*"repositórios" + 0.001*"direitos" + 0.001*"dados" + 0.001*"informação" + 0.001*"caso" + 0.001*"corte" + 0.001*"memória" + 0.001*"humanos" + 0.001*"interamericana" + 0.001*"pesquisa";
Tópico 16: 0.003*"informação" + 0.003*"memória" + 0.002*"documentos" + 0.001*"pesquisa" + 0.001*"termo" + 0.001*"arte" + 0.001*"universidade" + 0.001*"história" + 0.001*"dados" + 0.001*"arquivo";
Tópico 17: 0.002*"conhecimento" + 0.001*"trabalho" + 0.001*"informação" + 0.001*"políticas" + 0.001*"museu" + 0.001*"produção" + 0.001*"sociedade" + 0.001*"sociais" + 0.001*"capitalismo" + 0.001*"engajamento";
Tópico 18: 0.008*"informação" + 0.001*"pesquisa" + 0.001*"comunicação" + 0.001*"política" + 0.001*"gestão" + 0.001*"sociedade" + 0.001*"social" + 0.001*"nacional" + 0.001*"uso" + 0.001*"construção";
Tópico 19: 0.002*"informação" + 0.001*"inteligência" + 0.001*"competência" + 0.001*"forma" + 0.001*"conhecimento" + 0.001*"ex-votos" + 0.001*"comunicação" + 0.001*"pesquisa" + 0.001*"bens" + 0.001*"competência_informação";
Tópico 20: 0.003*"pesquisa" + 0.001*"documentos" + 0.001*"brasil" + 0.001*"produção" + 0.001*"científica" + 0.001*"artigos" + 0.001*"federal" + 0.001*"tecnologia" + 0.001*"dados" + 0.001*"brasileira";

Tópico 21: 0.002*"verdade" + 0.001*"informação" + 0.001*"habermas" + 0.001*"mundo" + 0.001*"linguagem" + 0.001*"documentos" + 0.001*"prontuário" + 0.001*"memória" + 0.001*"vida" + 0.001*"conhecimento";
Tópico 22: 0.004*"informação" + 0.003*"dados" + 0.001*"linguagem" + 0.001*"vida" + 0.001*"ciclo" + 0.001*"ciclo_vida" + 0.001*"pesquisa" + 0.001*"conhecimento" + 0.001*"modelo" + 0.001*"processo";
Tópico 23: 0.010*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.002*"dados" + 0.001*"organização" + 0.001*"uso" + 0.001*"biblioteconomia" + 0.001*"trabalho" + 0.001*"forma" + 0.001*"contexto";
Tópico 24: 0.002*"conhecimento" + 0.002*"informação" + 0.001*"rede" + 0.001*"gestão" + 0.001*"científica" + 0.001*"redes" + 0.001*"dados" + 0.001*"colaboração" + 0.001*"produção" + 0.001*"temas";
Tópico 25: 0.004*"conhecimento" + 0.004*"informação" + 0.002*"pesquisa" + 0.001*"científico" + 0.001*"social" + 0.001*"rede" + 0.001*"leitura" + 0.001*"gestão" + 0.001*"conhecimento_científico" + 0.001*"tecnologia";
Tópico 26: 0.002*"informação" + 0.001*"objeto" + 0.001*"universidades" + 0.001*"arquivologia" + 0.001*"arquivo" + 0.001*"citação" + 0.001*"científico" + 0.001*"conexões" + 0.001*"produção" + 0.001*"ufrgs";
Tópico 27: 0.002*"web" + 0.002*"dados" + 0.001*"rankings" + 0.001*"recursos" + 0.001*"metadados" + 0.001*"rede" + 0.001*"digital" + 0.001*"semântica" + 0.001*"usuários" + 0.001*"web_semântica";
Tópico 28: 0.005*"museu" + 0.004*"informação" + 0.002*"memória" + 0.002*"social" + 0.001*"cultural" + 0.001*"cultura" + 0.001*"patrimônio" + 0.001*"forma" + 0.001*"processo" + 0.001*"sociedade";
Tópico 29: 0.005*"informação" + 0.002*"conhecimento" + 0.001*"biblioteca" + 0.001*"gestão" + 0.001*"forma" + 0.001*"entidades" + 0.001*"lei" + 0.001*"ontologias" + 0.001*"entidade" + 0.001*"documentos".

Fonte: Elaborado pelo autor.

Tem sido uma característica forte apresentada entre os resultados obtidos através do modelo LDA um maior quantitativo de N-gramas do tipo bigramas e trigramas quando se comparado ao modelo LSI, entretanto, para esse corpus de dados, o processo de aprendizagem do modelo apresentou uma quantidade maior de unigramas quando se comparado aos *corpus* anteriores que utilizaram o mesmo modelo. É possível observar entre os resultados unigramas generalistas, bem como apresentando no tópico 8 o termo “memória”, que possibilita gerar diversas composições de significados como como “memória_social” ou “memória_coletiva”, sendo necessário ao especialista consultar documentos externos como lista de bigramas e trigramas geradas a partir do algoritmo ou mesmo acessar ao *corpus* de dados para explorar dos documentos de maneira mais aprofundada. Também é possível encontrar entre os resultados unigramas especialistas que possibilitam ao profissional especialista no domínio da linguagem supor por meio de análise de assunto o nome ou rótulo para o tópico. Uma outra característica diferenciada no *corpus* 8

está na existência de valores de pesos para todos os termos extraídos, facilitando assim na ordem de relevância dos termos.

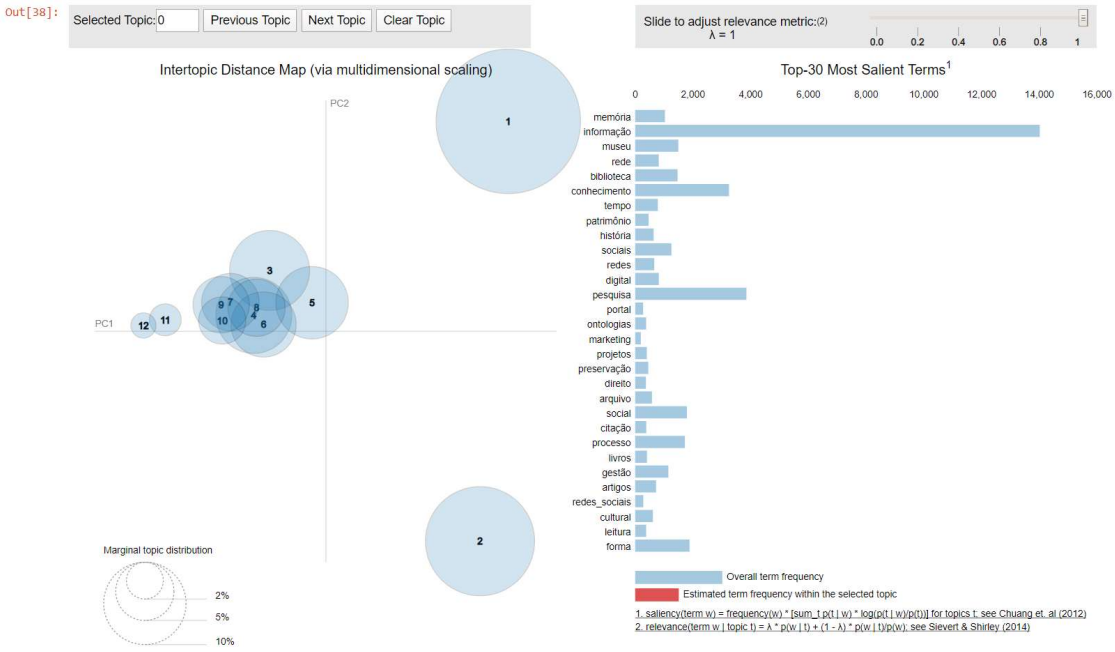
Exemplo de termos especialista podem ser encontrados no tópico 7 como “livro_eletrônico”, “suporte”, “papel” e “e-book” que ao serem associado a termos generalistas como “informação” e “documentos”, ambos os grupos contemplando pesos equivalentes, podem remeter ao especialista que o tópico esteja relacionado a área ou assunto de Informação e Tecnologia, entretanto, quando analisado por meio de acesso a documentos externos como o corpus de dados, acaba por remeter de maneira mais assertiva a área de Produção e Comunicação da Informação.

Ainda com as mesmas características que o tópico 7, os tópicos 3, 6 e 11 apresentam uma série de termos especialistas que possibilitam ao profissional, analista de assuntos, identificar os rótulos que melhor representam os tópicos por meio de exploração de documentos externos, bem como o *corpus* de dados que possibilita identificar de maneira rápida e assertiva por meio de busca por termos chave os documentos que contemplam termos como “canto”, ”samba”, “letra_samba” e “carnaval” contidos no corpus 3, “paisagem”, “geografia”, “ibge” no tópico 6 e “acervo”, “arquivo”, “memória” e “cfdt (Confédération Française Démocratique du Travail)” para o tópico 11. Dessa forma, o especialista poderá supor por exemplo que os termos dos tópicos 3 e 11 estão relacionados a Informação e Memória enquanto o termo do tópico 6 está relacionado a Organização e Representação do Conhecimento.

A Figura 20 apresenta os resultados obtidos através da visualização dinâmica gerada a partir do modelo LDA e da biblioteca pyLDAvis. O processamento para alcançar esse resultado foi de 3 horas 40 minutos e 12 segundos. A esquerda da imagem é possível observar os tópicos representados por círculos no plano bidimensional em um dimensionamento multidimensional como forma de projetar as distâncias intertópicas em duas dimensões. Já a direita da imagem estão os gráficos em barras que representam os termos individuais numa visão geral que são utilizados para interpretação dos tópicos.

O *download* do arquivo para a visualização dinâmica dos tópicos no formato HTML pode ser realizada através do GitHub[69].

Figura 20 – Visualização geral dos tópicos do *corpus* 8

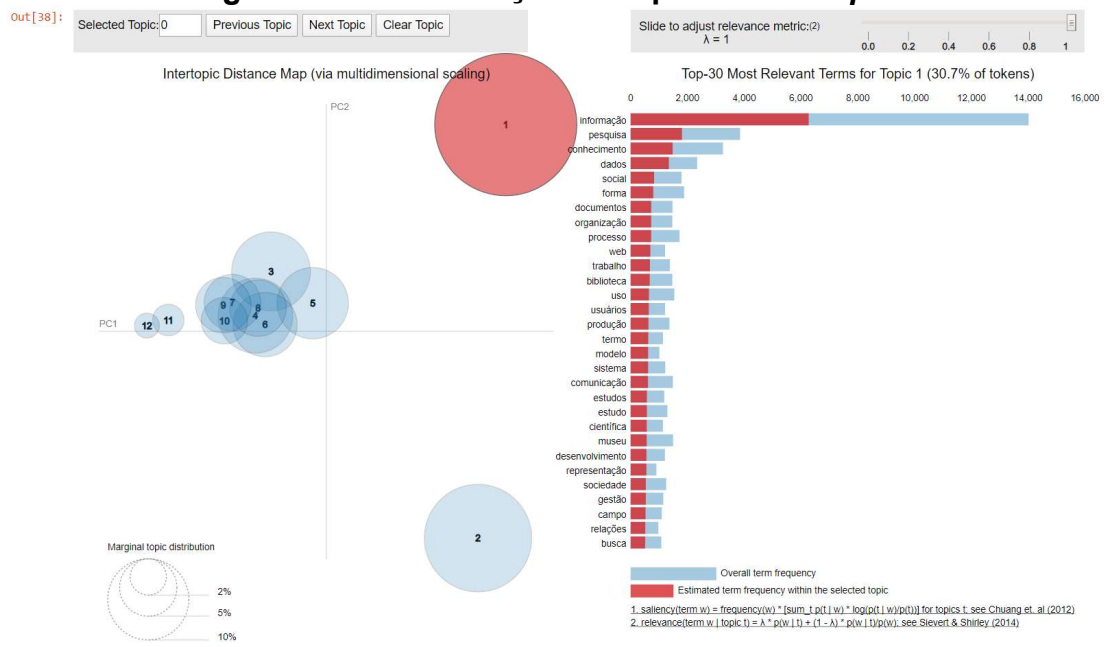


Fonte: Elaborado pelo autor.

A Figura 21 destaca ao lado esquerdo em vermelho o tópico 1 e a direita uma lista contendo 30 termos que melhor representam o tópico utilizando como ajuste de métrica de relevância valor igual a 1.0. Os termos do tópico 1 representam 30.7% dos *tokens*, sendo representados na cor azul para frequência geral e em vermelho para a frequência estimada do termo no tópico selecionado.

[69] Visualização dinâmica dos tópicos. *Corpus* 8: artigos completos e resumos expandidos 2013. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2013_gts.html/.

Figura 21 – Visualização do tópico 1 do *corpus* 8



Fonte: Elaborado pelo autor.

O tópico 1 apresenta um conjunto de termos fortes como “informação”, “pesquisa”, “conhecimento”, “dados”, “social” e “forma” no topo de sua lista quando a métrica ajustada para 1.0. Esses termos são ordenados mediante a sua frequência de termos encontrada no corpus de dados, além de também estarem contidos em outros tópicos. Quando ajustado a métrica para 0.2 surgem termos como “frbr (Functional Requirements for Bibliographic Records)”, “metadados”, “classificação”, “semântica”, “ontologia”, “representação” e “registros_bibliográficos” e para 0.1 surgem termos como “classificação_bibliográfica”, “modelo_conceitual_frbr”, “linked_data”, “machine_readable”, “ordenação_documentos” e “skos (Simple Knowledge Organization System)”, podendo ao especialista por exemplo, supor através da interpretação dos dados e análise de assunto que o tópico esteja relacionado ao rótulo Organização e Representação do Conhecimento. Tal suposição vai ao encontro da disciplina de Organização e Processamento da Informação e subdisciplina como Organização do Conhecimento / Representação da Informação do campo da Ciência da Informação conforme apontado por (PINHEIRO, 2006).

Torna-se possível perceber entre nos resultados uma maior aproximação entre os tópicos 3 ao 10, possuindo assim, termos comuns entre os tópicos e

sendo necessário realizar configurações de métricas para uma melhor interpretação dos dados.

APÊNDICE H: *Corpus* 9: artigos completos e resumos expandidos 2014

O *corpus* 9 que constitui o segundo *corpora* de dados é formado por 333 documentos do tipo de artigos completos e resumos expandidos publicados no ano de 2014. A coleção de documentos possui o tamanho de 12.985kb, 904.110 unigramas, 903.777 bigramas e 903.444 trigramas. O Quadro 41 apresenta os 50 termos mais frequentes de cada tipo de N-grama.

Quadro 41 – Lista de N-gramas por ordem de frequência do *corpus* 9

Unigramas
informação,19947; pesquisa,5488; conhecimento,4535; dados,3513; forma,2743; processo,2680; social,2611; biblioteca,2522; organização,2410; documentos,2316; trabalho,2309; uso,2297; produção,2233; comunicação,2116; termo,2067; estudo,1974; brasil,1945; relação,1868; sociais,1860; gestão,1808; estudos,1801; meio,1767; desenvolvimento,1750; usuários,1716; científica,1687; contexto,1630; sistema,1623; paulo,1585; sociedade,1574; web,1558; campo,1520; busca,1499; autores,1487; museu,1456; sistemas,1428; cultura,1420; informacional,1418; fonte,1382; processos,1351; artigos,1338; tecnologia,1328; nacional,1328; base,1319; diferentes,1316; resultados,1303; construção,1285; universidade,1273; atividades,1257; memória,1245; modelo,1212.
Bigramas
recuperação_informação,602; informação_tecnologia,529; informação_science,525; universidade_federal,498; redes_sociais,487; gestão_informação,484; produção_científica,479; organização_conhecimento,378; informação_conhecimento,376; fontes_informação,335; uso_informação,324; informação_comunicação,297; dados_pesquisa,286; ponto_vista,285; belo_horizonte,284; competência_informacional,280; organização_informação,273; comunicação_científica,265; coleta_dados,264; bases_dados,253; sistemas_informação,240; tecnologias_informação,237; arquitetura_informação,237; informação_informação,235; tendo_vista,234; base_dados,231; biblioteca_universitárias,225; dissertação_mestrado,221; direitos_humanos,219; muitas_vezes,215; ensino_superior,213; informação_brasília,212; fonte_dados,211; big_data,203; gestão_documentos,201; porto_alegre,199; tomada_decisão,197; biblioteca_nacional,194; gestão_conhecimento,191; sociedade_informação,190; profissionais_informação,190; minas_gerais,189; patrimônio_cultural,184; pesquisa_informação,183; teses_dissertações,178; estudo_caso,177; representação_informação,175; busca_informação,162; tecnologia_informação,162; representação_conhecimento,157.
Trigramas
fonte_dados_pesquisa,193; tecnologias_informação_comunicação,152; dissertação_mestrado_informação,106; nacional_pesquisa_informação,100; encontro_nacional_pesquisa,99; american_society_informação,97; society_informação_science,97; informação_science_technology,92; international_organization_standardization,90; resource_description_framework,81; journal_american_society,80; informação_universidade_federal,77; instituição_ensino_superior,75; informação_belo_horizonte,74; organização_representação_conhecimento,71; machine_readable_cataloging,71; universidade_federal_minas,69; universidade_federal_paraíba,68; instituições_ensino_superior,66; federal_minas_gerais,65; extensible_markup_language,65; sistemas_recuperação_informação,64;

livros_didáticos_digitais,64; informação_tecnologia_was,61; brasília_briquet_lemos,61; fundação_oswaldo_cruz,60; gestão_informação_conhecimento,60; sistemas_organização_conhecimento,58; fonte_elaborado_autores,57; portal_periódicos_capes,56; perspectivas_informação_belo,54; universidade_federal_santa,52; library_informação_science,50; informação_sociedade_estudos,49; sistema_recuperação_informação,47; universidade_federal_bahia,47; total_fonte_dados,46; república_federativa_brasil,46; world_wide_web,44; federal_santa_catarina,44; datagramazero_revista_informação,43; universidade_paulo_paulo,39; busca_uso_informação,39; produtos_serviços_informação,39; gestão_arquivística_documentos,39; paulo_martins_fontes,38; patrimônio_histórico_artístico,37; redes_sociais_internet,37; universidade_federal_fluminense,36; conselho_nacional_arquivos,36.

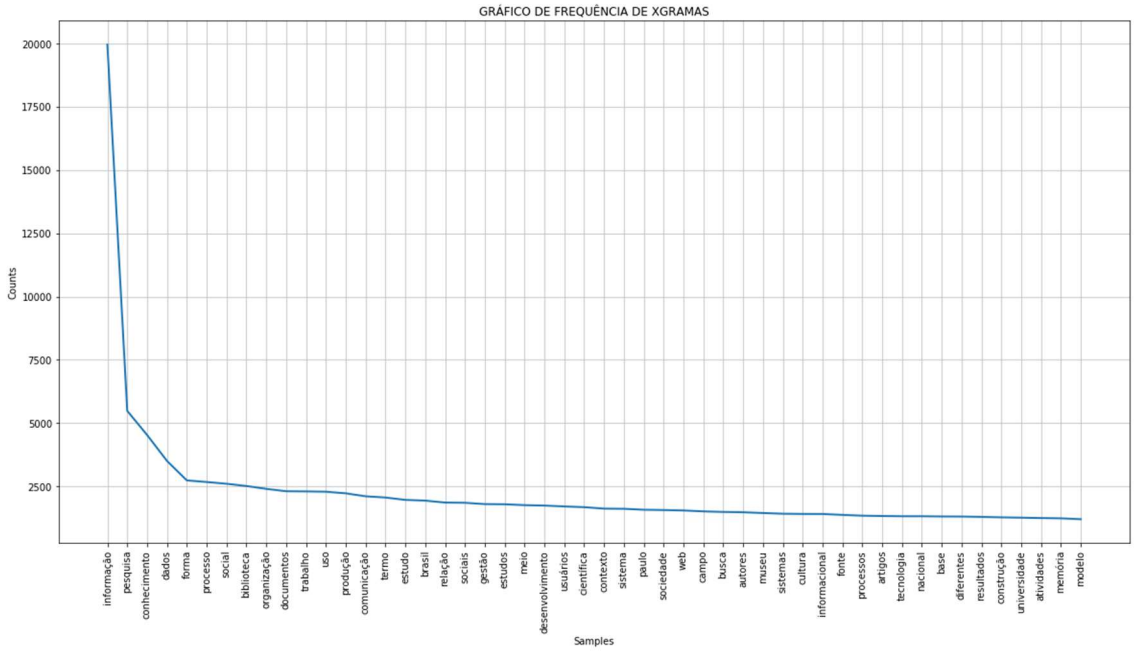
Fonte: Elaborado pelo autor.

Em uma lista geral de N-gramas contendo os mil termos mais frequentes extraídos do *corpus* de dados pode-se destacar os bigramas mais frequentes, sendo eles “recuperação_informação” com frequência de 602 na posição 186, “informação_tecnologia” com frequência de 529 e posição 235, “informação_science” com frequência de 525 e posição 239, “universidade_federal” com frequência de 498 e posição 255 e “redes_sociais” com frequência de 487 e posição 267. Entre os trigramas mais frequentes, apenas o primeiro representado pelo termo “fonte_dados_pesquisa” e com frequência de 193 está entre o milésimo termo ocupando a posição 853. Já os termos “tecnologias_informação_comunicação” com frequência de 152, “dissertação_mestrado_informação” com frequência de 106, “nacional_pesquisa_informação” com frequência de 100 e “encontro_nacional_pesquisa” com frequência de 99 aparecem após o milésimo termo mais frequente do *corpus* de dados.

Os unigramas apresentam maior frequência quando se comparado aos demais tipos de N-gramas como por exemplo “informação” com frequência de 19.947 é 3.213% maior que o bigrama com maior frequência, sendo “recuperação_informação” aparecendo 603 vezes no *corpus* de dados. O percentual chega a 10.235% quando comparado o unigrama com o primeiro trigrama mais frequente, sendo “fonte_dados_pesquisa” e frequência de 193.

O Gráfico 59 apresenta os 50 N-gramas mais frequentes do *corpus* 9. É possível perceber que todos os dados são do tipo unigrama, sendo os mais frequentes representados pelos termos “informação” com frequência de 19.947, “pesquisa” com frequência de 5.488, “conhecimento” com frequência de 4.535, “dados” com frequência de 3.513 e “forma” com frequência de 2.743.

Gráfico 59 – Termos mais frequentes do *corpus* 9

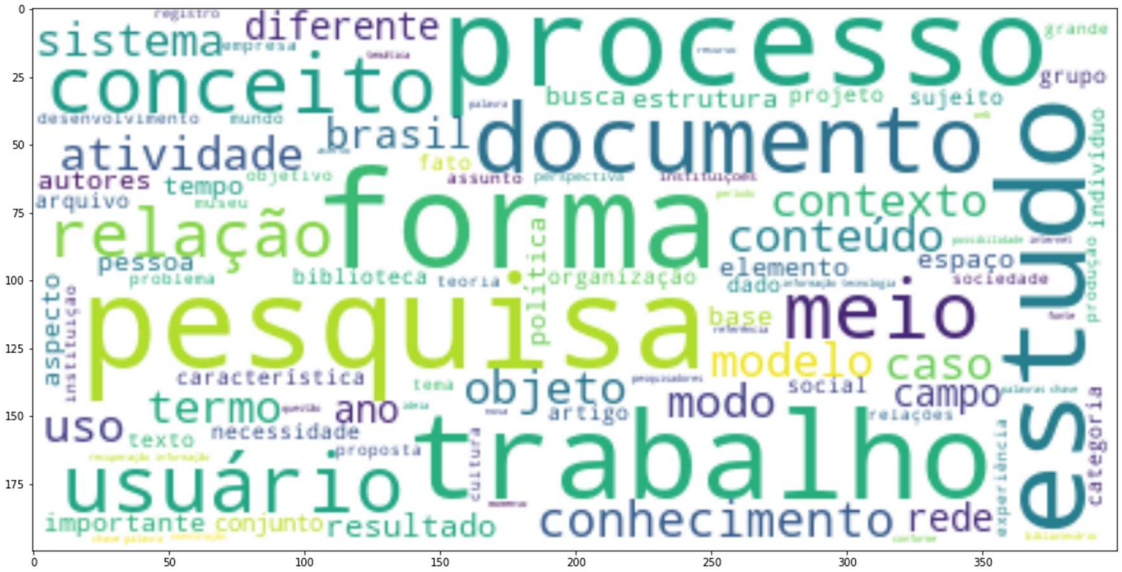


Fonte: Elaborado pelo autor.

Perceve-se no gráfico uma constância crescente entre o 50º e o 2º termo, respectivamente representado por “modelo” com frequência de 1.212 e “pesquisa” com frequência de 5.488. Já a diferença entre o segundo para o primeiro termo representado por “informação” e frequência de 19.947 chega a 266%. Também é possível identificar termos fortes como “cultura”, “informacional”, “fonte”, “processo” e “artigos” que podem ser melhores explorados pelo especialista por meio de recursos como a listagem de bigramas e trigramas para realizar a suposição do nome do tópico de forma mais assertiva. Além disso são encontrados tópicos fracos como “meio”, “brasil”, “paulo”, “universidade” e “diferentes” que podem ser descartados pelo especialista por apresentarem baixa relevância ao domínio de linguagem estudado.

A Figura 22 apresenta uma nuvem de palavras contendo os 250 termos mais frequentes do *corpus* de dados. Os dados são compostos por unigramas e bigramas, já que o primeiro trigrama aparece após o a quantidade de termos utilizada para construção automática desta figura. Além disso, os dados utilizados para criação da nuvem de palavras não passaram por quaisquer tipos de tratamento qualitativo.

Figura 22 – Nuvem de palavras do *corpus* 9



Fonte: Elaborado pelo autor.

Os N-gramas apresentados na imagem podem ser considerados fortes ou fracos quando se comparado ao domínio da linguagem ou realizada a análise de assunto apoiada aos bigramas e trigramas do *corpus* de dados. São exemplos de tópicos fortes os termos “uso”, “usuário” e “necessidade” que apresentam significância ao domínio de linguagem quando analisado os bigramas e trigramas. Já os termos “diferente”, “sujeito” e “caso” podem ser considerados fracos por não possuírem significados ou termos que correlacionem relacionados ao domínio da linguage. O *corpus* em questão apresenta mais tópicos fortes quando se comparado a todos os *corpus* contidos no primeiro *corpora* de dados.

Após realizado a fase de transformação, os dados são conectados aos modelos de tópicos que após treinados resultam em um conjunto de tópicos constituídos por termos e pesos. O Quadro 42 apresenta um conjunto contendo 30 tópicos extraídos do *corpus* de dados utilizando o modelo LSI. O processo de aprendizagem do modelo foi realizado em 35.3 segundos. Os resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos podem ser acessados através do GitHub[70].

Quadro 42 – Tópicos extraídos do *corpus* 9 usando o modelo LSI

Tópico 0: 0.789*"informação" + 0.162*"pesquisa" + 0.145*"conhecimento" + 0.101*"dados" + 0.087*"organização" + 0.086*"processo" + 0.082*"forma" + 0.080*"social" + 0.079*"uso" + 0.073*"biblioteca";

[70] Algoritmo de modelagem de tópicos. *Corpus* 9: artigos completos e resumos expandidos 2014. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2014.ipynb/.

Tópico 1: -0.510*"informação" + 0.222*"pesquisa" + 0.178*"documentos" + 0.159*"museu" + 0.154*"dados" + 0.146*"conhecimento" + 0.132*"biblioteca" + 0.126*"brasil" + 0.118*"arquivos" + 0.112*"científica";
Tópico 2: 0.562*"biblioteca" + -0.233*"documentos" + -0.221*"conhecimento" + 0.219*"usuários" + -0.163*"organização" + -0.157*"gestão" + 0.125*"usuário" + 0.115*"social" + 0.110*"bibliotecário" + -0.105*"metadados";
Tópico 3: 0.297*"documentos" + 0.259*"arquivos" + -0.241*"conhecimento" + -0.233*"dados" + 0.211*"direitos" + -0.200*"pesquisa" + 0.198*"biblioteca" + -0.186*"artigos" + -0.166*"científica" + 0.161*"arquivo";
Tópico 4: 0.326*"conhecimento" + -0.299*"dados" + 0.284*"museu" + 0.229*"cultura" + -0.197*"documentos" + -0.169*"pesquisa" + 0.161*"cultural" + 0.145*"social" + 0.129*"organização" + -0.126*"metadados";
Tópico 5: 0.298*"biblioteca" + 0.281*"conhecimento" + -0.208*"museu" + 0.186*"usuários" + 0.180*"organização" + -0.163*"direitos" + 0.148*"sistemas" + 0.138*"usuário" + 0.132*"metadados" + -0.132*"brasil";
Tópico 6: 0.345*"museu" + -0.343*"conhecimento" + 0.252*"dados" + 0.168*"web" + 0.139*"termo" + -0.136*"biblioteca" + 0.132*"objetos" + 0.125*"metadados" + -0.123*"direitos" + -0.118*"brasil";
Tópico 7: 0.378*"museu" + -0.316*"direitos" + -0.195*"sociais" + -0.186*"social" + -0.180*"humanos" + -0.168*"redes" + -0.163*"direitos_humanos" + 0.151*"biblioteca" + -0.144*"web" + -0.134*"dados";
Tópico 8: 0.445*"direitos" + 0.240*"direitos_humanos" + 0.235*"humanos" + -0.180*"redes" + -0.173*"documentos" + 0.172*"termo" + -0.169*"rede" + -0.166*"sociais" + -0.165*"social" + -0.162*"gestão";
Tópico 9: -0.214*"biblioteca" + 0.186*"dados" + 0.179*"trabalho" + 0.171*"museu" + 0.135*"gestão" + -0.127*"artigos" + -0.126*"termo" + 0.120*"processo" + -0.118*"representação" + -0.115*"web";
Tópico 10: 0.232*"biblioteca" + 0.197*"dados" + -0.193*"termo" + 0.171*"museu" + 0.157*"gestão" + 0.152*"metadados" + 0.148*"nacional" + -0.148*"trabalho" + 0.142*"conhecimento" + 0.134*"cultura";
Tópico 11: 0.293*"usuários" + -0.236*"dados" + 0.184*"redes" + 0.181*"usuário" + 0.165*"museu" + -0.135*"cultura" + 0.129*"rede" + 0.120*"sociais" + -0.116*"trabalho" + 0.106*"redes_sociais";
Tópico 12: -0.350*"metadados" + 0.212*"dados" + -0.147*"digital" + -0.145*"informacional" + -0.128*"registros" + -0.117*"digitais" + -0.113*"direitos" + -0.112*"competência" + 0.110*"biblioteca" + -0.108*"museu";
Tópico 13: 0.183*"artigos" + 0.174*"metadados" + 0.159*"produção" + -0.142*"web" + -0.135*"pesquisa" + 0.132*"trabalho" + 0.127*"periódicos" + 0.127*"vida" + -0.127*"redes" + -0.125*"organização";
Tópico 14: -0.234*"cultura" + -0.231*"dados" + 0.221*"termo" + 0.212*"ontologias" + -0.174*"usuários" + 0.173*"ontologia" + -0.153*"usuário" + -0.123*"arquivo" + -0.108*"semiótica" + 0.108*"gestão";
Tópico 15: 0.256*"informacional" + 0.230*"competência" + -0.192*"digital" + 0.181*"competência_informacional" + -0.143*"acessibilidade" + 0.125*"biblioteca" + 0.121*"social" + 0.117*"museu" + -0.116*"digitais" + -0.116*"sistema";
Tópico 16: -0.221*"trabalho" + 0.209*"informacional" + -0.183*"biblioteconomia" + -0.170*"curso" + -0.169*"cursos" + 0.147*"digital" + 0.142*"competência" + 0.131*"livros" + 0.128*"competência_informacional" + -0.113*"disciplinas";
Tópico 17: 0.302*"pesquisa" + -0.177*"assunto" + 0.170*"arquivos" + 0.163*"ontologias" + -0.152*"livros" + -0.149*"organização" + 0.142*"ontologia" + -0.117*"artigos" + 0.115*"preservação" + -0.105*"classificação";

Tópico 18: -0.263*"pesquisa" + 0.192*"trabalho" + 0.161*"teses" + 0.160*"dissertações" + 0.142*"teses_dissertações" + -0.136*"assunto" + -0.121*"curso" + 0.120*"digital" + -0.118*"rede" + -0.117*"livros";
Tópico 19: -0.202*"pesquisa" + 0.158*"nacional" + 0.153*"informacional" + -0.153*"gestão" + 0.147*"arquivo" + -0.143*"social" + 0.131*"arquivos" + 0.126*"memória" + -0.124*"documentos" + 0.123*"organização";
Tópico 20: -0.321*"teses" + -0.310*"dissertações" + -0.272*"teses_dissertações" + 0.192*"web" + 0.190*"ontologias" + 0.139*"ontologia" + -0.130*"programas" + 0.125*"multimídia" + -0.123*"registros" + 0.099*"periódicos";
Tópico 21: -0.192*"arquivos" + -0.166*"livros" + -0.143*"arquivo" + 0.140*"referência" + 0.139*"cultura" + 0.138*"lei" + -0.137*"digitais" + 0.136*"web" + 0.134*"transparência" + -0.115*"didáticos";
Tópico 22: 0.310*"referência" + 0.245*"processo" + 0.179*"conhecimento" + -0.158*"cultura" + -0.151*"informacional" + 0.148*"bibliotecário" + 0.135*"serviço" + 0.129*"educativo" + -0.127*"trabalho" + -0.125*"sistemas";
Tópico 23: -0.415*"termo" + -0.243*"cultura" + -0.153*"service" + 0.136*"vida" + -0.112*"brasil" + 0.107*"social" + -0.099*"transparência" + -0.095*"lei" + 0.095*"pesquisa" + -0.095*"cultural";
Tópico 24: 0.158*"ontologias" + 0.153*"cultura" + 0.148*"trabalho" + -0.141*"nacional" + 0.126*"processo" + -0.125*"informacional" + -0.122*"transparência" + 0.121*"mediação" + -0.118*"competência" + -0.117*"liberdade";
Tópico 25: -0.212*"registros" + 0.154*"mercado" + 0.150*"marketing" + 0.146*"metadados" + 0.119*"data" + -0.117*"pesquisa" + 0.112*"gestão" + -0.106*"trabalho" + -0.105*"inteligência" + 0.105*"conceito";
Tópico 26: -0.196*"inteligência" + 0.163*"referência" + 0.161*"digital" + -0.131*"liberdade" + 0.131*"termo" + 0.130*"imagem" + 0.129*"vida" + -0.124*"usuários" + 0.123*"inclusão" + -0.122*"produção";
Tópico 27: -0.252*"mercado" + 0.212*"memória" + -0.181*"empresas" + -0.178*"marketing" + 0.126*"gestão" + -0.124*"fontes" + -0.122*"clientes" + -0.111*"serviços" + 0.110*"rede" + 0.096*"processo";
Tópico 28: 0.172*"arquivos" + -0.166*"inteligência" + 0.138*"web" + -0.132*"nacional" + -0.124*"referência" + -0.117*"pesquisa" + -0.104*"biblioteca_nacional" + -0.103*"digital" + -0.100*"produção" + -0.096*"silva";
Tópico 29: 0.172*"curso" + 0.138*"evasão" + -0.127*"assunto" + 0.123*"data" + -0.121*"arquivos" + -0.121*"acidente" + -0.119*"trabalho" + -0.112*"rede" + -0.112*"transparência" + 0.111*"cartas".

Fonte: Elaborado pelo autor.

Os resultados obtidos através do modelo LSI apresentam uma série de tópicos fortes constituídos por termos e pesos de qualidade que contribuem para que o especialista realize a identificação da suposição do nome dos tópicos de maneira mais assertiva. O tópico 0 é a exceção desses resultados por apresentar termos generalistas como “informação” que pode se compor para outros termos de relevância para o domínio da linguagem como “informação_conhecimento”, “recuperação_informação” ou “uso_informação”. Dessa forma, cabe ao especialista utilizar de recursos externos como as listas de bigramas e trigramas gerados a partir do algoritmo e/ou explorar o *corpus* de dados para supor rótulos ao tópico. Uma alternativa para termos ou mesmo tópicos generalistas está em

criar um rótulo com características gerais ao domínio de linguagem mesmo ou realizar o descarte dos termos ou tópicos.

O tópico 2 é constituído de termos como “biblioteca”, “documentos’, “conhecimento” que quando analisados sozinhos podem apontar para rótulos diferentes, entretanto constam outros termos como “organização”, “gestão” e “metadados” que possibilita ao especialista supor por exemplo que o tópico esteja relacionado a Gestão da Informação ou Sistemas de Gestão da Informação.

Essas características na qualidade dos resultados também podem ser observadas no tópico 17 que apresenta uma combinação de termos como “arquivos”, “livros”, “preservação”, “organização” e “ontologia”, podendo supor por exemplo pelo especialista que o tópico esteja relacionado a área de Organização e Representação do Conhecimento pertinente ao domínio de linguagem, bem como o tópico 22 com os termos “informacional”, “educativo”, “referência”, “bibliotecário”, “serviço’ e “sistemas” podendo supor que o tópico esteja relacionado a área de Serviços Informacionais.

O Quadro 43 apresenta os resultados alcançados através da modelagem de tópicos utilizando o modelo LDA quando aplicado ao *corpus* 9. O modelo foi treinado durante 8 minutos e 5 segundos gerando um conjunto de 30 tópicos com os seus respectivos termos e pesos. Os resultados com 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[71].

Quadro 43 – Tópicos extraídos do *corpus* 9 usando o modelo LDA

Tópico 0: 0.001*"licitação" + 0.001*"documentos" + 0.001*"lei" + 0.001*"twitter" + 0.001*"candidatos" + 0.001*"rede" + 0.001*"candidato" + 0.001*"uff" + 0.001*"conversação" + 0.000*"processos_licitação";
Tópico 1: 0.013*"informação" + 0.002*"pesquisa" + 0.002*"gestão" + 0.001*"organização" + 0.001*"gestão_informação" + 0.001*"conhecimento" + 0.001*"dados" + 0.001*"serviços" + 0.001*"informacional" + 0.001*"mediação";
Tópico 2: 0.001*"indexação" + 0.001*"facetada" + 0.001*"analistas" + 0.001*"usabilidade" + 0.001*"classificação" + 0.000*"etiquetagem" + 0.000*"taxonomia" + 0.000*"taxonomia_facetada" + 0.000*"usuário" + 0.000*"faceted";
Tópico 3: 0.001*"cinema" + 0.001*"canais" + 0.001*"televisão" + 0.001*"audiovisual" + 0.001*"lei" + 0.001*"brasil" + 0.001*"globo" + 0.001*"filmes" + 0.000*"brasileiro" + 0.000*"programação";

[71] Algoritmo de modelagem de tópicos. *Corpus* 9: artigos completos e resumos expandidos 2014. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2014.ipynb/.

Tópico 4: 0.005*"informação" + 0.001*"curso" + 0.001*"evasão" + 0.001*"mercado" + 0.001*"pesquisa" + 0.001*"competências" + 0.001*"informais" + 0.001*"empresas" + 0.001*"comunicação" + 0.001*"sociedade";
Tópico 5: 0.003*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"ontologias" + 0.001*"científica" + 0.001*"forma" + 0.001*"ontologia" + 0.001*"termo" + 0.001*"documentos";
Tópico 6: 0.002*"informação" + 0.001*"nacional" + 0.001*"biblioteca" + 0.001*"tesauro" + 0.001*"silva" + 0.001*"pesquisa" + 0.001*"conhecimento" + 0.001*"biblioteca_nacional" + 0.001*"organização" + 0.001*"peregrino";
Tópico 7: 0.002*"registros" + 0.001*"conversão" + 0.001*"metadados" + 0.001*"machine_readable" + 0.001*"machine_readable_cataloging" + 0.001*"readable_cataloging" + 0.001*"readable" + 0.001*"machine" + 0.001*"cataloging" + 0.001*"teoria"
Tópico 8: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"memória" + 0.001*"patrimônio" + 0.001*"cultural" + 0.001*"documentos" + 0.001*"preservação" + 0.001*"biblioteca" + 0.001*"social" + 0.001*"cultura";
Tópico 9: 0.003*"informação" + 0.001*"vida" + 0.001*"memória" + 0.001*"pesquisa" + 0.001*"social" + 0.001*"objeto" + 0.001*"forma" + 0.001*"assunto" + 0.001*"processo" + 0.001*"qualidade";
Tópico 10: 0.003*"informação" + 0.002*"pesquisa" + 0.002*"museu" + 0.002*"campo" + 0.001*"científica" + 0.001*"conhecimento" + 0.001*"trabalho" + 0.001*"produção" + 0.001*"história" + 0.001*"brasil";
Tópico 11: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"biblioteca" + 0.001*"dados" + 0.001*"social" + 0.001*"usuários" + 0.001*"arquitetura" + 0.001*"comunicação" + 0.001*"arquitetura_informação" + 0.001*"biblioteconomia";
Tópico 12: 0.005*"informação" + 0.002*"conhecimento" + 0.001*"rede" + 0.001*"pesquisa" + 0.001*"social" + 0.001*"atores" + 0.001*"produção" + 0.001*"dados" + 0.001*"sociais" + 0.001*"assunto";
Tópico 13: 0.013*"informação" + 0.004*"conhecimento" + 0.004*"pesquisa" + 0.003*"dados" + 0.002*"uso" + 0.002*"forma" + 0.002*"web" + 0.002*"organização" + 0.002*"processo" + 0.001*"termo";
Tópico 14: 0.008*"informação" + 0.002*"biblioteca" + 0.002*"conhecimento" + 0.001*"direitos" + 0.001*"pesquisa" + 0.001*"social" + 0.001*"processo" + 0.001*"comunicação" + 0.001*"forma" + 0.001*"informacional";
Tópico 15: 0.002*"meg" + 0.001*"competências" + 0.001*"choo" + 0.001*"excelência" + 0.001*"complexidade" + 0.001*"informacionais" + 0.001*"organizações" + 0.001*"tomada" + 0.001*"decisão" + 0.001*"modelo";
Tópico 16: 0.004*"informação" + 0.002*"documentos" + 0.001*"conhecimento" + 0.001*"portfólio" + 0.001*"termo" + 0.001*"autores" + 0.001*"processo" + 0.001*"pesquisa" + 0.001*"documental" + 0.001*"dados";
Tópico 17: 0.003*"metadados" + 0.003*"documentos" + 0.002*"gestão" + 0.001*"informação" + 0.001*"arquivística" + 0.001*"domínio" + 0.001*"instrumentos" + 0.001*"sistemas" + 0.001*"gestão_arquivística" + 0.001*"arquivísticos";
Tópico 18: 0.001*"informação" + 0.001*"vigilância" + 0.001*"sociais" + 0.001*"redes" + 0.001*"conhecimento" + 0.001*"propriedade" + 0.001*"intelectual" + 0.001*"redes_sociais" + 0.001*"propriedade_intelectual" + 0.001*"pesquisa";
Tópico 19: 0.005*"informação" + 0.002*"pesquisa" + 0.001*"redes" + 0.001*"termo" + 0.001*"produção" + 0.001*"estudos" + 0.001*"rede" + 0.001*"dados" + 0.001*"sociais" + 0.001*"biblioteconomia";
Tópico 20: 0.008*"informação" + 0.002*"lei" + 0.002*"acessibilidade" + 0.002*"documentos" + 0.002*"transparência" + 0.002*"direito" + 0.001*"gestão" + 0.001*"brasil" + 0.001*"arquivo" + 0.001*"pesquisa";

Tópico 21: 0.001*"imagens" + 0.001*"imagem" + 0.001*"education" + 0.001*"superior" + 0.001*"higher" + 0.001*"patologia" + 0.001*"internacionalización" + 0.001*"higher_education" + 0.000*"journal" + 0.000*"international";
Tópico 22: 0.003*"informação" + 0.002*"arquivos" + 0.002*"arquivo" + 0.001*"documentos" + 0.001*"autoridade" + 0.001*"cultura" + 0.001*"registros" + 0.001*"processo" + 0.001*"aquisição" + 0.001*"inclusão_digital"
Tópico 23: 0.002*"museu" + 0.001*"cultural" + 0.001*"patrimônio" + 0.001*"informação" + 0.001*"arte" + 0.001*"objetos" + 0.001*"política" + 0.001*"social" + 0.001*"cultura" + 0.001*"paulo";
Tópico 24: 0.002*"informação" + 0.002*"museu" + 0.002*"conhecimento" + 0.001*"arte" + 0.001*"inteligência" + 0.001*"cultura" + 0.001*"organização" + 0.001*"pesquisa" + 0.001*"processo" + 0.001*"documentação";
Tópico 25: 0.000*"informação" + 0.000*"artigos" + 0.000*"pesquisa" + 0.000*"científica" + 0.000*"metadados" + 0.000*"dados" + 0.000*"produção" + 0.000*"periódicos" + 0.000*"brasil" + 0.000*"documentos";
Tópico 26: 0.002*"memória" + 0.001*"definições" + 0.001*"termo" + 0.001*"ontologias" + 0.001*"definição" + 0.001*"gramame" + 0.000*"leukemia" + 0.000*"retratos" + 0.000*"acervo" + 0.000*"fotografia";
Tópico 27: 0.003*"cartas" + 0.001*"monde" + 0.001*"leitores" + 0.001*"leitoras" + 0.001*"jornal" + 0.001*"jornais" + 0.001*"leitores_leitoras" + 0.001*"publicadas" + 0.000*"des" + 0.000*"leitores/as";
Tópico 28: 0.002*"informação" + 0.002*"artigos" + 0.002*"produção" + 0.001*"programas" + 0.001*"teses" + 0.001*"periódicos" + 0.001*"dissertações" + 0.001*"teses_dissertações" + 0.001*"científica" + 0.001*"pesquisa";
Tópico 29: 0.004*"informação" + 0.001*"trabalho" + 0.001*"dados" + 0.001*"processo" + 0.001*"referência" + 0.001*"termo" + 0.001*"pesquisa" + 0.001*"digital" + 0.001*"forma" + 0.001*"social.

Fonte: Elaborado pelo autor.

Os resultados obtidos através do modelo LDA apresentam um baixo quantitativo de termos do tipo bigramas e trigramas, além de possuírem unigramas generalistas e especialistas. Os unigramas generalistas são os termos comuns ao domínio de linguagem que podem apresentar diversas composições de significados, como por exemplo o termo “biblioteca” que pode abordar assuntos como “gestão_biblioteca” ou “biblioteca_universitária”. Já os termos especialistas permitem uma interpretação dos resultados mais assertiva por parte do profissional indexador ou mesmo um direcionamento para a exploração de documentos externos gerados a partir do processo de modelagem de tópicos. São exemplos de termos especialistas encontrados entre os resultados do modelo LDA os termos “acessibilidade”, “etiquetagem” ou “taxonomia”.

O tópico 0 apresenta uma série de termos especialistas que permite ao professional através de técnicas de análise de assunto investigar através dos documentos externos, como o *corpus* de dados, os documentos que melhor representam o tópico ou mesmo realizar a suposição de nome do tópico por meio

de sua subjetividade levando em consideração experiência e vivência na área. São exemplos os termos “licitação”, “documentos”, “lei”, “candidato”, e “rede” que permite supor o rótulo Política e Economia da Informação para o tópico.

Já no tópico 1 é possível perceber termos generalistas como “informação” e “dados” que quando associados a termos especialistas como “gestão_informação”, “organizações”, “ serviços”, “mediação” e “informacional” podem remeter ao especialista ao rótulo Gestão da Informação e do Conhecimento, sem mesmo realizar consultas a documentos externos como o *corpus* de dados. Quando os resultados apresentam um conjunto de termos especialistas e coesos relacionados a um determinado assunto, acaba por minimizar o esforço cognitivo do profissional especialista no domínio da linguagem durante a realização da análise de assuntos, um exemplo disso está no tópico 2 que apresenta termos especialistas característicos da área de Organização e Representação do Conhecimento como “indexação”, “taxonomia_facetada”, “etiquetagem” e “classificação”.

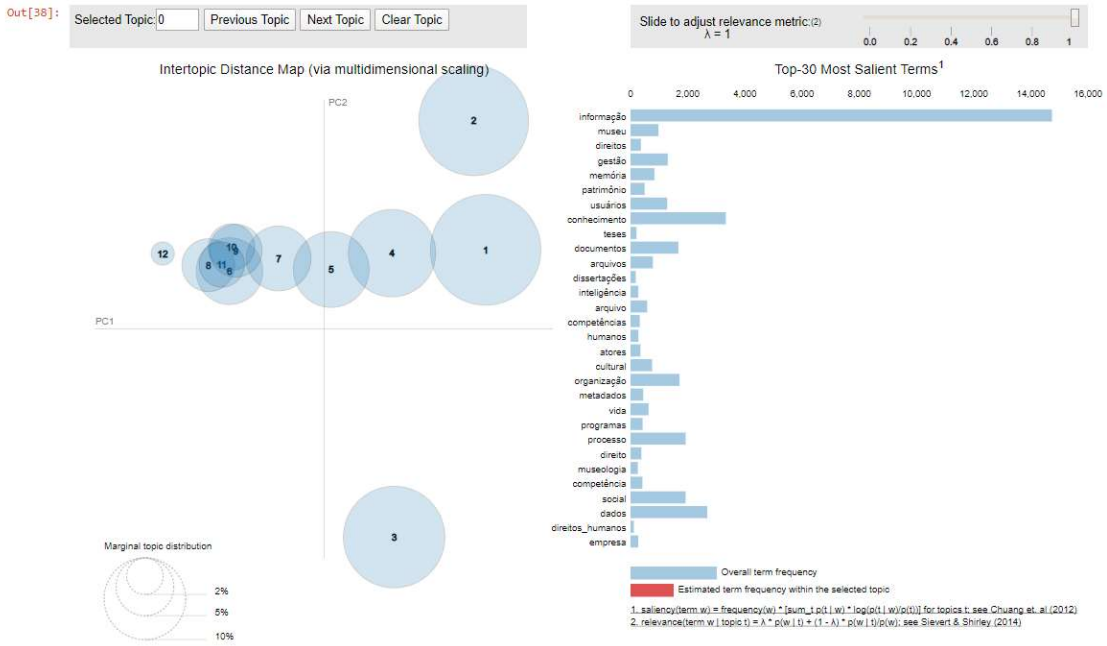
Os tópicos 3 e 27 também apresentam um conjunto de termos especialistas que possibilitam ao profissional indexador inferir por meio de análise de assunto a suposição dos nomes dos tópicos de maneira direta ou mesmo realizando a análise através documentos externos como o *corpus* de dados. O tópico 3 contempla os termos “audiovisual”, “programação”, “cinema”, “programação”, “filmes” e “lei” que, quando explorado o *corpus* de dados, permite localizar documentos referente ao tópico e identificar por meio de técnicas de análise de assunto outras informações como título, resumo ou grupo de trabalho contidas nos documentos, podendo supor por exemplo que o tópico esteja associado ao rótulo de Política e Economia da Informação.

Já o tópico 27 apresenta os termos “cartas”, “jornal”, “publicadas”, “leitoras”, além do termo chave “monde” que possibilita a identificação da suposição do nome do tópico, seja analisando os termos eu explorando o *corpus* de dados para identificação de mais informações que permitem aferir que o tópico esteja relacionado a Mediação, Circulação e Apropriação de Informações.

A visualização dinâmica dos tópicos permite ao especialista realizar uma navegação com maior interatividade entre os tópicos, explorando assim um maior quantitativo de termos por meio de configuração de métricas. A Figura 23 representa uma visualização com os dados extraídos por meio do modelo LDA

e construídos a partir da biblioteca pyLDAvis, sendo processado durante o tempo de 4 horas 11 minutos e 6 segundos para ser construída. A parte esquerda da imagem ilustra em formato de círculos os tópicos que estão em plano bidimensional, usando um dimensionamento multidimensional para projetar as distâncias intertópicas em duas dimensões. Já a direita da imagem estão os gráficos de barras que representam os termos individuais dos tópicos, sendo úteis no processo de interpretação dos tópicos por parte do especialista. O arquivo com a visualização dinâmica dos tópicos está desmobilizada para *download* no formato HTML através do GitHub[72].

Figura 23 – Visualização geral dos tópicos do *corpus* 9

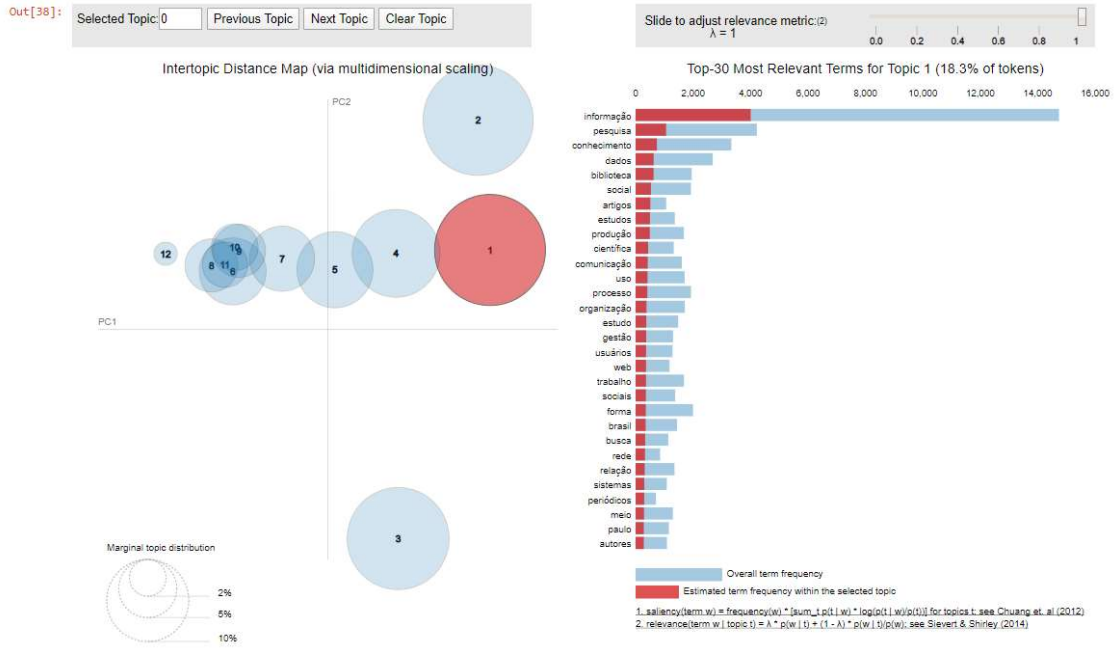


Fonte: Elaborado pelo autor.

A Figura 24 destaca o tópico 1 destacado de vermelho ao lado esquerdo da imagem e os termos mais relevantes representados ao lado direito. Os termos apresentam 18.3% dos *tokens*, sendo a cor azul referente a frequência geral e a cor vermelha a frequência estimada do termo no tópico selecionado.

[72] Visualização dinâmica dos tópicos. *Corpus* 9: artigos completos e resumos expandidos 2014. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2014_gts.html/.

Figura 24 – Visualização do tópico 1 do *corpus* 9



Fonte: Elaborado pelo autor.

Os tópicos encontrados no *corpus* 9 possuem um maior equilíbrio entre os termos dos tópicos 1 e 2, possuindo assim um quantitativo superior a 18% dos *tokens* para ambos os tópicos. Além disso, é possível observar uma proximidade entre os tópicos 4, 5 e 7, entretanto, possuindo pouca similaridade entre os termos dos tópicos. Já os tópicos 6, 8, 9, 10 e 11 apresentam uma concentração de tópicos próximos, exigindo assim do especialista um maior esforço cognitivo na interpretação dos tópicos, além de poder navegar entre outras possibilidades de resultados através dos ajustes de métricas.

O tópico 1 apresenta o conjunto de termos “informação”, “pesquisa”, “conhecimento”, “dados”, “biblioteca”, e “social” no topo de sua lista quando utilizado a métrica de relevância com o valor 1.0. Ao ajustar a métrica para 0.2 é possível identificar termos com características mais específicos como “altimétricos”, “altimetria”, “encontrabilidade_informação”, “etnometodologia”, “informação_pervasiva”, “arquitetura_informação_pervasiva” que auxilia ao especialista na interpretação do tópico por apresentar mais especificidade, podendo supor por exemplo que o rótulo do tópico esteja relacionado a Informação e Tecnologia. Os termos encontrados contêm assuntos das disciplinas de Fundamentos da Ciência da Informação, Tecnologias da Informação e Organização e processamento da informação, conforme apontado por (PINHEIRO, 2006).

APÊNDICE I - *Corpus* 10: artigos completos e resumos expandidos 2015

O *corpus* 10 pertencente ao segundo *corpora* de dados foi constituído por 280 documentos do tipo artigos completos e resumos expandidos publicados no ano de 2015. O corpus possui o tamanho de 10.974kb e o quantitativo de 772.021 unigramas, 771.741 bigramas e 771.461 trigramas. O Quadro 44 apresenta uma lista contendo os 50 termos mais frequentes separados por tipos de N-gramas.

Quadro 44 – Lista de N-gramas por ordem de frequência do *corpus* 10

Unigramas
informação,17934; pesquisa,5044; conhecimento,4580; dados,3019; biblioteca,2372; forma,2365; processo,2361; social,2221; museu,2197; gestão,2053; documentos,2019; produção,1969; trabalho,1961; organização,1960; uso,1933; comunicação,1813; memória,1713; meio,1664; brasil,1657; científica,1623; relação,1578; estudo,1542; estudos,1510; desenvolvimento,1467; termo,1465; sociais,1458; busca,1457; sociedade,1434; tecnologia,1378; resultados,1330; autores,1324; sistema,1311; contexto,1293; educação,1255; paulo,1247; usuários,1217; informacional,1209; conteúdo,1177; história,1163; digital,1143; relações,1127; construção,1125; nacional,1110; avaliação,1096; recuperação,1088; campo,1087; cultura,1082; diferentes,1079; universidade,1072; base,1069.
Bigramas
recuperação_informação,716; informação_tecnologia,560; gestão_informação,477; informação_conhecimento,459; produção_científica,452; informação_science,421; organização_conhecimento,372; universidade_federal,369; belo_horizonte,366; uso_informação,325; informação_comunicação,321; deste_artigo,311; redes_sociais,304; fontes_informação,295; gestão_documentos,289; autores_trabalho,282; conteúdo_textual,281; artigo_nomes,281; extraídos_metadados,281; textual_deste,280; nomes_e-mails,280; e-mails_extraídos,280; metadados_informados,280; informados_total,280; total_responsabilidade,280; responsabilidade_autores,280; xvi_issn,277; xvi_xvi,271; informação_informação,261; gestão_conhecimento,253; dissertação_mestrado,246; biblioteca_universitárias,237; ponto_vista,236; competência_informacional,232; inteligência_competitiva,227; dados_pesquisa,221; base_dados,217; comunicação_oral,212; competência_informação,212; tecnologias_informação,210; bases_dados,210; coleta_dados,205 sistemas_informação,198; sociedade_informação,193; busca_informação,192; pesquisa_informação,189; ensino_superior,186; periódicos_científicos,184; organização_representação,178; comunicação_científica,175.
Trigramas
conteúdo_textual_deste,280; textual_deste_artigo,280; deste_artigo_nomes,280; artigo_nomes_e-mails,280; nomes_e-mails_extraídos,280; e-mails_extraídos_metadados,280; extraídos_metadados_informados,280; metadados_informados_total,280; informados_total_responsabilidade,280; total_responsabilidade_autores,280; responsabilidade_autores_trabalho,280; xvi_xvi_issn,271; fonte_dados_pesquisa,145; gestão_informação_conhecimento,145; tecnologias_informação_comunicação,136; organização_representação_conhecimento,108; sistemas_recuperação_informação,98; dissertação_mestrado_informação,95;

encontro_nacional_pesquisa,94;	nacional_pesquisa_informação,93;
sistema_recuperação_informação,91;	xvi_issn_informação,85;
informação_comunicação_oral,68;	informação_tecnologia_was,68;
informação_belo_horizonte,67;	resource_description_framework,65;
portal_periódicos_capes,62;	instituições_ensino_superior,57;
sistemas_organização_conhecimento,55;	conhecimento_comunicação_oral,55;
universidade_federal_paraíba,55;	american_society_informação,54;
society_informação_science,54;	informação_sociedade_estudos,53;
periódicos_científicos_eletrônicos,53;	informação_universidade_federal,51;
informação_science_technology,49;	autores_trabalho_abstract,49;
unidade_prestadora_serviços,49; fonte_elaborado_autores,47;	journal_american_society,46;
international_organization_standardization,46;	instituição_ensino_superior,45;
datagramazero_revista_informação,45;	ensino_pesquisa_extensão,45;
museu_história_natural,44;	prestadora_serviços_informativos,44;
american_library_association,43;	perspectivas_informação_belo,41;
instituto_benjamin_constant,41.	

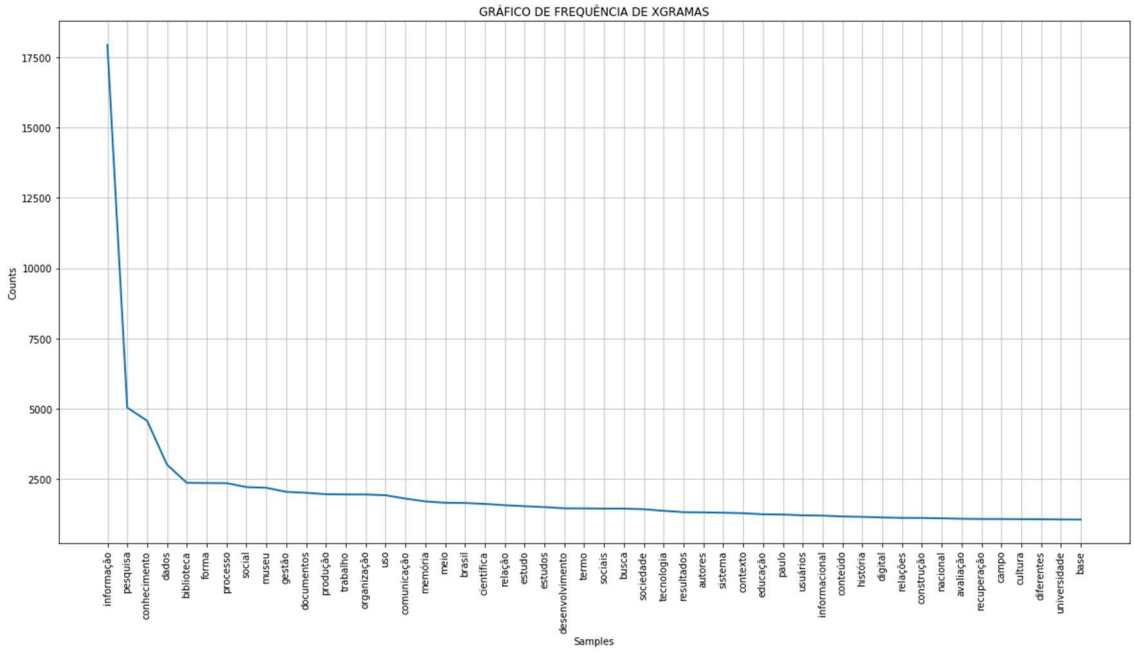
Fonte: Elaborado pelo autor.

Esse corpus de dados é o primeiro quando se comparado aos demais corpus de dados utilizados na pesquisa que apresenta em sua lista geral de mil N-gramas mais frequentes os 5 primeiros termos de cada tipo, sendo unigramas, bigramas e trigramas. Destaca-se na lista de bigramas os termos mais frequentes “recuperação_informação” com frequência de 716 e ocupando a posição 109, “informação_tecnologia” com frequência de 560 e posição 156, “gestão_informação” com frequência de 477 e posição 209, “informação_conhecimento” com frequência de 459 e posição 218 e “produção_científica” com frequência de 452 e posição 226. Já entre os trigramas mais frequentes estão os termos “conteúdo_textual_deste” na posição 462, “textual_deste_artigo” na posição 463, “deste_artigo_nomes” na posição 464, “artigo_nomes_e-mails” na posição 465 e “nomes_e-mails_extraídos” na posição 466, ambos com frequência de 280. É importante ressaltar que os termos mais frequentes não estão atrelados a qualidade dos N-gramas, sendo necessário ao especialista realizar a filtragem dos dados de acordo com sua necessidade.

Já os unigramas apresentam maior frequência quando comparado aos outros tipos de N-gramas. O unigrama “informação”, com frequência de 17.934 é 2.405% maior que o bigrama mais frequente “recuperação_informação” extraído 716 vezes do *corpus*. Esse percentual aumenta para 6.305% quando comparado o mesmo unigrama com o trigrama “conteúdo_textual_deste” e frequência de 280.

O Gráfico 60 apresenta os 50 termos mais frequentes extraídos do *corpus* de dados, sendo todos eles caraterizado por unigramas. Pode-se destacar os termos “informação” com frequência de 17.934, “pesquisa” com frequência de 5.044, “conhecimento” com frequência de 4.580, “dados” com frequência de 3.019 e “biblioteca” com frequência de 2.372.

Gráfico 60 – Termos mais frequentes do *corpus* 10



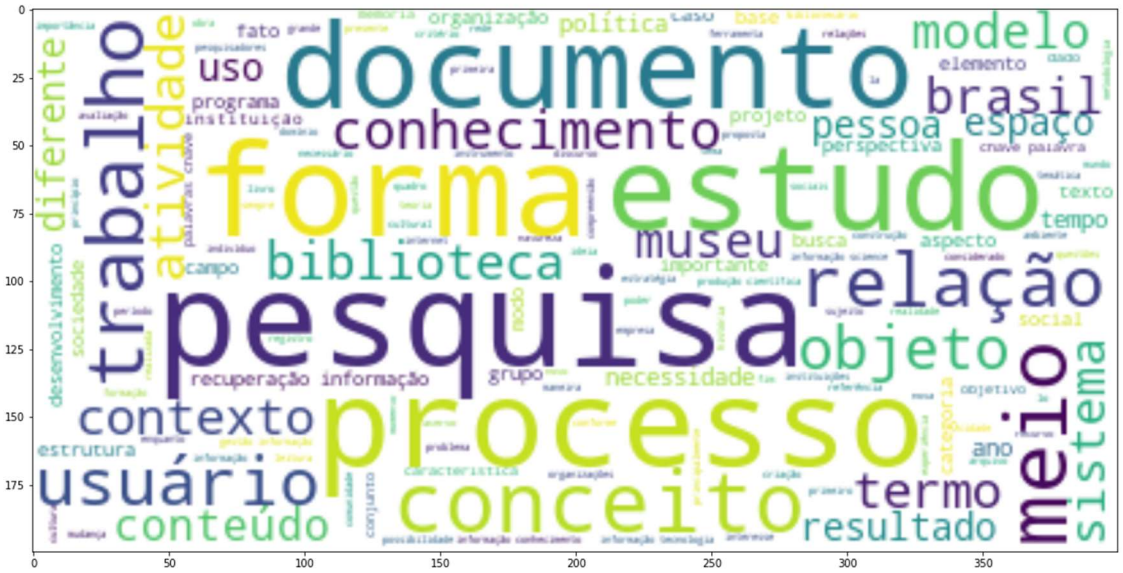
Fonte: Elaborado pelo autor.

Os resultados do gráfico apresentam um crescente de forma constante entre o 50º e o 2º termo representados respectivamente pelos termos “base” com frequência de 1.069 e “pesquisa” com frequência de 5.044. É possível perceber uma diferença de 256% entre o primeiro e o segundo termo mais frequentes, sendo “informação” com frequência de 17.934 e “pesquisa” com frequência 5.044. Além disso, é possível identificar tópicos fortes como “organização”, “uso”, “comunicação”, “sociedade” e “usuários” que podem ser explorados pelo especialista junto as listas de bigramas e trigramas. Constam também termos fracos como “brasil”, “meio”, “paulo”, “diferentes” e “universidade” que podem ser desncosiderados pelo profissional especialista por não estarem alinhados com domínio da linguagem.

Por meio das listas dos N-gramas gerados através no *corpus* de dados, foi criado uma nuvem de palavras contendo os 250 termos mais frequentes. A Figura 25 contempla bigramas e unigramas que destacam como termos fortes e

fracos contidos na coleção de documentos. Para a criação da imagem não foi utilizado quaiquer tipos de análises qualitativa.

Figura 25 – Nuvem de palavras do *corpus* 10



Fonte: Elaborado pelo autor.

Dentre os termos fortes, pode-se destacar “biblioteca”, “museu” e “documento” que apresentam representatividade alinhadas ao domínio da linguagem. Já os termos “elemento”, “ano” e “meio” são considerados termos fracos por não apresentar relevância ao domínio da linguagem. O *corpus* também possui uma maior quantidade de termos fortes quando se comparado ao *corpora* 1 constituido por teses e dissertações.

Após a fase de transformação os dados são conectados aos modelos de extração de dados. O Quadro 45 apresenta os resultados constituídos por 30 tópicos, sendo cada um deles contemplados por 10 termos e seus respectivos pesos obtidos através do modelo LSI quando aplicado ao *corpus* de dados. O tempo de treinamento do modelo para obtenção dos resultados foi de 23.9 segundos. Os demais resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponível através do GitHub[73].

Quadro 45 – Tópicos extraídos do *corpus* 10 usando o modelo LSI

Tópico 0: 0.789*"informação" + 0.173*"conhecimento" + 0.167*"pesquisa" + 0.102*"dados" + 0.085*"gestão" + 0.085*"processo" + 0.081*"forma" + 0.078*"organização" + 0.074*"social" + 0.074*"uso";

[73] Algoritmo de modelagem de tópicos. *Corpus* 10: artigos completos e resumos expandidos 2015. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2015.ipynb/.

Tópico 1: -0.467*"informação" + 0.407*"museu" + 0.235*"pesquisa" + 0.151*"história" + 0.145*"biblioteca" + 0.140*"conhecimento" + 0.125*"memória" + 0.116*"dados" + 0.109*"produção" + 0.100*"patrimônio";

Tópico 2: 0.670*"conhecimento" + -0.346*"museu" + 0.280*"organização" + 0.167*"organização_conhecimento" + -0.155*"informação" + -0.130*"biblioteca" + 0.091*"produção" + 0.084*"científica" + -0.084*"história" + 0.081*"gestão";

Tópico 3: -0.506*"museu" + 0.367*"biblioteca" + 0.252*"dados" + -0.247*"conhecimento" + 0.173*"pesquisa" + 0.170*"documentos" + -0.158*"informação" + -0.111*"organização" + -0.094*"história" + 0.085*"recuperação";

Tópico 4: -0.613*"biblioteca" + 0.378*"dados" + 0.175*"pesquisa" + -0.150*"conhecimento" + 0.142*"recuperação" + -0.106*"gestão" + 0.099*"recuperação_informação" + 0.098*"museu" + -0.098*"serviços" + -0.096*"organização";

Tópico 5: -0.517*"documentos" + -0.343*"gestão" + 0.222*"biblioteca" + -0.199*"arquivo" + 0.191*"dados" + -0.174*"arquivos" + -0.174*"gestão_documentos" + 0.129*"pesquisa" + -0.112*"documento" + -0.109*"documental";

Tópico 6: -0.293*"dados" + 0.250*"pesquisa" + -0.248*"biblioteca" + 0.244*"produção" + 0.236*"científica" + -0.164*"museu" + -0.151*"organização" + 0.145*"grupos" + -0.136*"termo" + 0.133*"pesquisadores";

Tópico 7: 0.310*"gestão" + -0.293*"memória" + 0.291*"museu" + 0.194*"serviços" + -0.175*"social" + 0.161*"pesquisa" + 0.134*"dados" + 0.120*"marketing" + -0.111*"termo" + -0.107*"representação";

Tópico 8: 0.410*"dados" + -0.227*"recuperação" + -0.225*"recuperação_informação" + -0.168*"biblioteca" + -0.145*"periódicos" + 0.138*"gestão" + -0.127*"documentos" + -0.123*"museu" + 0.121*"memória" + -0.116*"indexação";

Tópico 9: 0.425*"digital" + 0.283*"preservação" + -0.258*"fontes" + 0.223*"preservação_digital" + 0.210*"periódicos" + -0.191*"pesquisa" + 0.178*"digitais" + -0.133*"busca" + 0.131*"eletrônicos" + 0.121*"científicos";

Tópico 10: -0.305*"serviços" + -0.297*"marketing" + 0.214*"dados" + -0.205*"informativos" + -0.201*"serviços_informativos" + -0.164*"termo" + -0.134*"unidade" + -0.132*"orientação" + -0.124*"prestadora" + -0.119*"prestadora_serviços";

Tópico 11: -0.532*"fontes" + -0.173*"empresas" + -0.164*"fontes_informação" + -0.151*"periódicos" + -0.138*"preservação" + -0.136*"digital" + -0.123*"inteligência" + -0.118*"memória" + 0.116*"recuperação" + -0.109*"negócios";

Tópico 12: 0.243*"biblioteconomia" + 0.199*"curso" + 0.172*"digital" + 0.165*"profissional" + -0.147*"produção" + 0.146*"formação" + 0.136*"cursos" + -0.130*"memória" + 0.128*"ufmg" + -0.123*"social";

Tópico 13: -0.273*"memória" + 0.164*"gestão" + 0.151*"termo" + -0.145*"marketing" + -0.141*"serviços" + -0.137*"história" + -0.137*"organização_conhecimento" + 0.131*"avaliação" + 0.130*"processo" + -0.128*"recuperação";

Tópico 14: 0.245*"termo" + -0.226*"pesquisa" + 0.158*"lei" + 0.149*"indexação" + 0.143*"brasil" + -0.137*"busca" + 0.134*"palavras" + -0.133*"informacional" + -0.132*"gestão" + -0.131*"usuários";

Tópico 15: -0.313*"pesquisa" + -0.175*"lei" + 0.165*"produção" + 0.155*"curso" + 0.155*"biblioteconomia" + 0.153*"recuperação" + 0.129*"dados" + 0.128*"avaliação" + -0.125*"grupos" + -0.124*"digital";

Tópico 16: 0.207*"informacional" + 0.187*"competência" + 0.168*"avaliação" + 0.159*"periódicos" + 0.127*"usuários" + 0.123*"busca" + -0.118*"web" + -0.114*"relações" + 0.108*"competência_informacional" + -0.106*"rede";

Tópico 17: 0.224*"gestão" + 0.199*"memória" + -0.198*"informacional" + -0.192*"produção" + -0.162*"competência" + 0.160*"pesquisa" + 0.151*"termo" + -0.148*"científica" + 0.140*"palavras-chave" + 0.127*"auditoria";

Tópico 18: -0.215*"rede" + -0.196*"social" + -0.179*"redes" + -0.165*"usuários" + 0.158*"informacional" + 0.157*"história" + 0.156*"pesquisa" + -0.156*"sociais" + -0.138*"avaliação" + 0.137*"biblioteca";
Tópico 19: 0.187*"descrição" + 0.176*"memória" + -0.158*"recuperação" + 0.158*"usuários" + -0.154*"organização" + -0.152*"poder" + 0.151*"web" + -0.146*"recuperação_informação" + -0.144*"informacional" + 0.137*"arquivo";
Tópico 20: 0.194*"inteligência" + 0.180*"poder" + -0.174*"documentos" + 0.164*"lei" + -0.142*"biblioteconomia" + -0.142*"social" + 0.130*"cultural" + -0.127*"memória" + 0.125*"busca" + 0.125*"patrimônio";
Tópico 21: 0.234*"inteligência" + 0.228*"poder" + 0.202*"organização" + -0.178*"fontes" + -0.172*"cultural" + 0.161*"competitiva" + -0.157*"avaliação" + -0.153*"conhecimento" + -0.152*"educação" + -0.143*"gestão";
Tópico 22: 0.261*"avaliação" + 0.179*"organização" + 0.156*"periódicos" + -0.151*"palavras" + -0.142*"conhecimento" + 0.142*"poder" + -0.126*"documentos" + -0.124*"repositórios" + -0.120*"fontes" + 0.116*"organização_conhecimento";
Tópico 23: -0.183*"competência" + 0.176*"objetos" + -0.156*"descrição" + -0.151*"educação" + -0.133*"inteligência" + -0.127*"organização" + -0.123*"description" + -0.123*"memória" + 0.119*"tecnologia" + -0.113*"metadados";
Tópico 24: 0.204*"natural" + 0.200*"história" + -0.180*"cultural" + 0.178*"rede" + 0.171*"história_natural" + -0.146*"leitura" + 0.145*"memória" + -0.144*"mediação" + 0.136*"redes" + -0.133*"descrição";
Tópico 25: -0.209*"avaliação" + 0.191*"poder" + 0.184*"patrimônio" + -0.176*"memória" + 0.164*"arquivo" + -0.147*"inteligência" + -0.136*"leitura" + -0.132*"educação" + 0.124*"cultural" + 0.114*"termo";
Tópico 26: 0.220*"leitura" + -0.188*"memória" + -0.163*"sistema" + 0.163*"auditoria" + 0.146*"natural" + 0.144*"poder" + 0.133*"história" + 0.132*"autores" + -0.121*"educação" + 0.115*"história_natural";
Tópico 27: -0.268*"poder" + 0.178*"inteligência" + -0.171*"gestão" + 0.145*"municipal" + 0.135*"horizonte" + 0.134*"belo_horizonte" + 0.133*"belo" + 0.132*"patrimônio" + -0.126*"usuários" + 0.125*"arquivo";
Tópico 28: 0.184*"auditoria" + -0.182*"indexação" + -0.153*"palavras" + 0.145*"palavras-chave" + 0.135*"termo" + -0.127*"história" + 0.125*"artigos" + -0.117*"natural" + -0.113*"rede" + 0.113*"ontologia";
Tópico 29: 0.212*"patrimônio" + 0.196*"objetos" + 0.155*"palavras" + -0.143*"leitura" + 0.138*"competência" + 0.128*"autores" + -0.117*"produção" + -0.117*"história" + -0.114*"termo" + -0.113*"arquivo".

Fonte: Elaborado pelo autor.

Os resultados obtidos através do modelo LSI apresentaram conjuntos de termos e pesos que possibilitam ao especialista analisar os resultados e realizar a suposição dos nomes de cada tópico. São destacados termos fortes como “informação” no tópico 0, “conhecimento” no tópico 2, “biblioteca” no tópico 4 e “documentos” no tópico 5 que possuem valores em seus pesos superiores a 0.500, seja negativo ou positivo. O tópico 0 apresenta termos generalistas como “informação” e “conhecimento” que podem se composições para outros conceitos dentro do domínio de linguagem, entretanto, é possível observar outros termos dentro do tópico como “gestão”, “organização” e “dados” que possibilitam ao especialista supor através de uma análise subjetiva e levando em

consideração seus conhecimentos e vivência na área que o tópico esteja relacionado a Gestão da Informação e do Conhecimento.

Outros tópicos também permitem realizar a suposição dos noems dos rótulos sem a necessidade de consultar documentos externos como as listas de bigramas e trigramas ou mesmo ao próprio corpus de dados, dentre eles, o tópico 1 que apresenta termos chave como “informação”, “museu” e “patrimônio” além de termos como “produção”, “conhecimento” e “memória” que podem remeter ao especialista o rótulo Museu, Patrimônio e Informação.

O tópico 27 constituído dos termos “poder”, “patrimônio”, “inteligência”, “gestão” e “municipal” permitem ao especialista supor que o nome do tópico esteja relacionado a área da Política, Economia e Informação. Também é encontrado entre os resultados do tópico o termo chave “belo_horizonte” que possibilita ao especialista consultar o *corpus* de dados e encontrar outros indicadores mais assertivos.

O Quadro 46 apresenta os resultados obtidos através do modelo de extração LDA quando aplicado ao *corpus* de dados. Os resultados são constituídos por 30 tópicos com os respectivos termos e pesos que foram treinados durante 6 minutos e 25 segundos. Os outros grupos de resultados contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[74].

Quadro 46 – Tópicos extraídos do *corpus* 10 usando o modelo LDA

Tópico 0: 0.001*"inteligência" + 0.001*"acessibilidade" + 0.001*"competitiva" + 0.001*"inteligência_competitiva" + 0.001*"web" + 0.001*"fotografia" + 0.000*"erros" + 0.000*"memória" + 0.000*"sites" + 0.000*"acervos";
Tópico 1: 0.004*"informação" + 0.002*"museu" + 0.001*"pesquisa" + 0.001*"objetos" + 0.001*"virtual" + 0.001*"ontologias" + 0.001*"conhecimento" + 0.001*"conteúdo" + 0.001*"metadados" + 0.001*"profissional";
Tópico 2: 0.001*"acessibilidade" + 0.001*"sociais" + 0.001*"web" + 0.000*"acessibilidade_web" + 0.000*"wcag" + 0.000*"usuários" + 0.000*"prioridade" + 0.000*"avaliadores" + 0.000*"barreiras" + 0.000*"automáticos";
Tópico 3: 0.015*"informação" + 0.003*"conhecimento" + 0.003*"pesquisa" + 0.002*"dados" + 0.002*"processo" + 0.002*"trabalho" + 0.002*"forma" + 0.002*"uso" + 0.001*"social" + 0.001*"comunicação";
Tópico 4: 0.002*"informação" + 0.002*"pesquisa" + 0.001*"lei" + 0.001*"conhecimento" + 0.001*"políbio" + 0.001*"modelo" + 0.001*"web" + 0.001*"estudos" + 0.001*"educação" + 0.001*"grupos";

[74] Algoritmo de modelagem de tópicos. *Corpus* 10: artigos completos e resumos expandidos 2015. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2015.ipynb/.

Tópico 5: 0.005*"informação" + 0.001*"periódicos" + 0.001*"pesquisa" + 0.001*"dados" + 0.001*"mulheres" + 0.001*"digital" + 0.001*"preservação" + 0.001*"cidade" + 0.001*"patrimônio" + 0.001*"memória";
Tópico 6: 0.002*"informação" + 0.002*"autores" + 0.001*"redes" + 0.001*"memorial" + 0.001*"rede" + 0.001*"social" + 0.001*"livro" + 0.001*"sociais" + 0.001*"grau" + 0.001*"arquivística";
Tópico 7: 0.002*"educação" + 0.001*"científica" + 0.001*"ensino" + 0.001*"tecnológica" + 0.001*"distância" + 0.001*"instituições" + 0.001*"educação_profissional" + 0.001*"alunos" + 0.001*"divulgação" + 0.001*"multiple_document";
Tópico 8: 0.002*"produção" + 0.002*"pesquisa" + 0.002*"informação" + 0.002*"científica" + 0.001*"avaliação" + 0.001*"produção_científica" + 0.001*"periódicos" + 0.001*"memória" + 0.001*"bourdieu" + 0.001*"bric";
Tópico 9: 0.009*"informação" + 0.002*"social" + 0.002*"conhecimento" + 0.002*"biblioteca" + 0.002*"pesquisa" + 0.001*"uso" + 0.001*"forma" + 0.001*"cultura" + 0.001*"sociais" + 0.001*"comunicação";
Tópico 10: 0.005*"informação" + 0.002*"biblioteca" + 0.002*"busca" + 0.002*"pesquisa" + 0.002*"sistema" + 0.002*"recuperação" + 0.001*"arquivo" + 0.001*"documentos" + 0.001*"usuários" + 0.001*"processo";
Tópico 11: 0.010*"informação" + 0.002*"dados" + 0.002*"conhecimento" + 0.002*"pesquisa" + 0.002*"gestão" + 0.001*"documentos" + 0.001*"organização" + 0.001*"uso" + 0.001*"processo" + 0.001*"brasil";
Tópico 12: 0.006*"museu" + 0.004*"informação" + 0.002*"história" + 0.001*"pesquisa" + 0.001*"forma" + 0.001*"arte" + 0.001*"conhecimento" + 0.001*"social" + 0.001*"memória" + 0.001*"leitura";
Tópico 13: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"científica" + 0.000*"palavras" + 0.000*"forma" + 0.000*"conhecimento" + 0.000*"publicação" + 0.000*"dados" + 0.000*"tempo" + 0.000*"social";
Tópico 14: 0.001*"gestão" + 0.001*"conhecimento" + 0.001*"modelo" + 0.001*"documentos" + 0.001*"empresa" + 0.001*"organização" + 0.001*"gestão_conhecimento" + 0.001*"gad" + 0.001*"motivacional" + 0.001*"modelagem";
Tópico 15: 0.002*"ontologia" + 0.001*"ontologias" + 0.001*"domínio" + 0.001*"hemonto" + 0.001*"sangue" + 0.001*"etapa" + 0.001*"ontoforinfoscience" + 0.001*"ontology" + 0.001*"digitais" + 0.001*"construção";
Tópico 16: 0.001*"tecnologia" + 0.001*"periódicos" + 0.001*"objetos" + 0.001*"ufjf" + 0.001*"universidade" + 0.001*"artigos" + 0.001*"pesquisa" + 0.001*"museu" + 0.001*"matemática" + 0.001*"mdct";
Tópico 17: 0.002*"cultural" + 0.001*"patrimônio" + 0.001*"gestão" + 0.001*"ceu" + 0.001*"dispositivo" + 0.001*"culturais" + 0.001*"cultura" + 0.001*"paulo" + 0.001*"unesco" + 0.001*"valor";
Tópico 18: 0.006*"informação" + 0.004*"conhecimento" + 0.002*"pesquisa" + 0.002*"organização" + 0.002*"científica" + 0.002*"produção" + 0.001*"estudos" + 0.001*"organização_conhecimento" + 0.001*"pesquisadores" + 0.001*"comunicação";
Tópico 19: 0.001*"documento" + 0.001*"patrimônio" + 0.001*"citação" + 0.001*"referências" + 0.001*"estudos" + 0.000*"menções" + 0.000*"periódicos" + 0.000*"dissertações" + 0.000*"teses" + 0.000*"informação";
Tópico 20: 0.005*"informação" + 0.002*"conhecimento" + 0.002*"pesquisa" + 0.001*"documentos" + 0.001*"forma" + 0.001*"memória" + 0.001*"gestão" + 0.001*"processo" + 0.001*"produção" + 0.001*"dados";
Tópico 21: 0.002*"pesquisa" + 0.002*"informação" + 0.002*"museu" + 0.001*"palavras" + 0.001*"dados" + 0.001*"termo" + 0.001*"critério" + 0.001*"texto" + 0.001*"classificação" + 0.001*"indexação";

Tópico 22: 0.003*"biblioteca" + 0.002*"memória" + 0.002*"imagem" + 0.001*"arquivos" + 0.001*"social" + 0.001*"arquivo" + 0.000*"memórias" + 0.000*"história" + 0.000*"acervos" + 0.000*"acervo";
Tópico 23: 0.003*"informação" + 0.002*"conhecimento" + 0.002*"gestão" + 0.001*"rede" + 0.001*"memória" + 0.001*"pesquisa" + 0.001*"pesquisadores" + 0.001*"científica" + 0.001*"produção" + 0.001*"ambiente";
Tópico 24: 0.003*"serviços" + 0.002*"marketing" + 0.002*"informativos" + 0.002*"serviços_informativos" + 0.001*"orientação" + 0.001*"gestão" + 0.001*"prestadora" + 0.001*"lde" + 0.001*"unidade" + 0.001*"prestadora_serviços";
Tópico 25: 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"informação" + 0.001*"avaliação" + 0.001*"gestão" + 0.001*"museu" + 0.001*"termo" + 0.001*"ecm" + 0.001*"comunicação" + 0.001*"educação";
Tópico 26: 0.003*"biblioteca" + 0.001*"nacional" + 0.001*"nacionais" + 0.001*"acervo" + 0.001*"biblioteca_nacionais" + 0.001*"descrição" + 0.001*"biblioteca_nacional" + 0.001*"braille" + 0.001*"instituto" + 0.001*"objetos";
Tópico 27: 0.000*"informação" + 0.000*"pesquisa" + 0.000*"dados" + 0.000*"relações" + 0.000*"forma" + 0.000*"objetos" + 0.000*"conhecimento" + 0.000*"memória" + 0.000*"museu" + 0.000*"construção";
Tópico 28: 0.001*"portal" + 0.001*"usabilidade" + 0.001*"severidade" + 0.001*"problemas" + 0.001*"grau" + 0.001*"lai" + 0.001*"usuário" + 0.000*"grau_severidade" + 0.000*"usuários" + 0.000*"avaliadores";
Tópico 29: 0.008*"informação" + 0.003*"pesquisa" + 0.001*"mediação" + 0.001*"uso" + 0.001*"recuperação" + 0.001*"dados" + 0.001*"recuperação_informação" + 0.001*"conhecimento" + 0.001*"documentos" + 0.001*"forma".

Fonte: Elaborado pelo autor.

É possível perceber por meio dos resultados obtidos através do modelo LDA a existência de tópicos contendo termos do tipo unigramas generalistas e especialistas, além de bigramas. Os termos generalistas requerem por parte do especialista do domínio de linguagem um maior esforço cognitivo para realizar análise de assunto e interpretação dos dados de maneira mais assertiva, uma vez que tópicos como “dados” podem apresentar composições de termos com outros significados, como por exemplo “análise_dados” ou “ciclo_vida_dados”. Já os termos especialistas como “recuperação_informação”, “indexação” ou “ontology” possibilitam uma interpretação dos tópicos de maneira mais assertiva ou mesmo apresentando um norte para que o especialista possa realizar pesquisas em documentos externos que possibilitem uma melhor interpretação dos tópicos. Outra característica apresentada nos resultados desse *corpus* está na existência de pesos equilibrados entre os termos, entretanto, uma dificuldade encontrada está na interpretação de tópicos que possuam termos com valores iguais a 0, já que não é possível identificar uma ordem de relevância entre os termos.

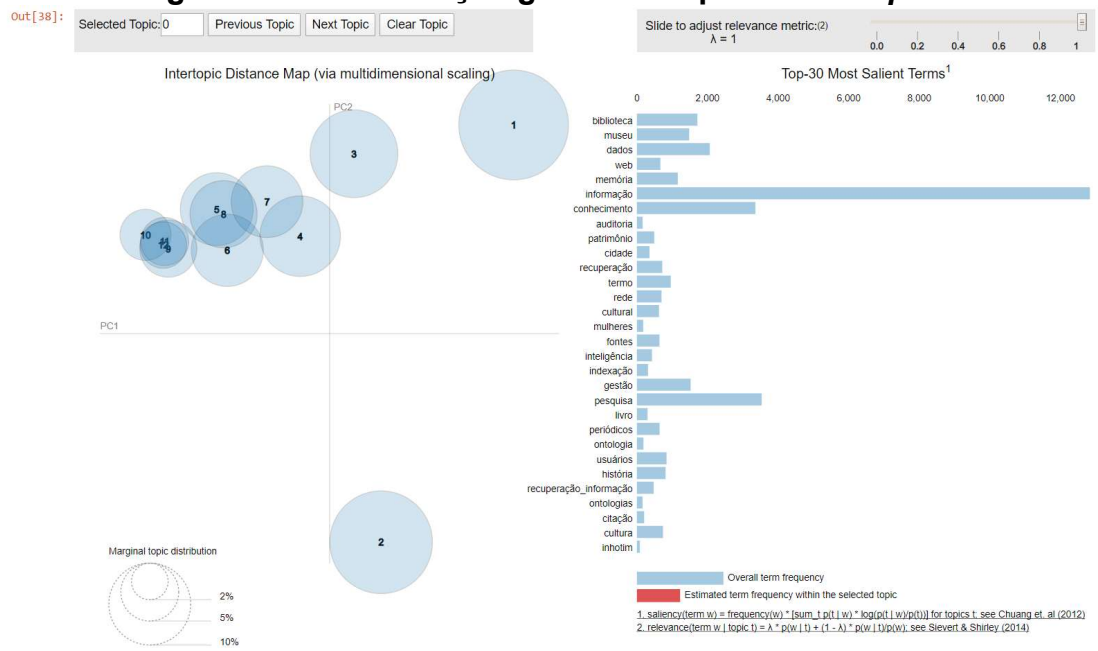
O tópico 1 apresenta termos generalistas e específicos com pesos representativos que contribuem para uma melhor interpretação dos dados a ser realizada pelo profissional especialista em análise de assunto e/ou no domínio da linguagem, dentre eles “informação” e “conhecimento” como termos generalistas e “ontologias”, “metadados”, “conteúdo”, e “virtual” como termos especialistas, remetendo ao rótulo Organização e Representação do Conhecimento.

Tópicos contendo termos chave possibilitam ao especialista realizar a identificação de assuntos através de auxílio de documentos externos, como por exemplo o *corpus* de dados. O tópico 4 apresenta termos coesos e com equilíbrio entre seus pesos, dentre eles, o termo “políbio” aparece somente em um único documento em todo o *corpus* analisado, remetendo assim a área de Informação e Memória. Já o tópico 14 apresenta os termos “gestão”, “conhecimento”, “modelo”, “documento”, “gad (Gestão Arquivística de Documentos)” e “modelagem”, podendo o especialista supor através de análise de assunto que o tópico esteja relacionado a área Gestão do Informação e do Conhecimento ou Informação e Tecnologia.

A visualização dinâmica dos tópicos é uma outra maneira que contribui para a identificação dos rótulos dos tópicos realizada pelo especialista do domínio da linguagem. A Figura 26 representa a visualização construída a partir dos resultados de extração de tópicos do *corpus* de dados utilizando o modelo LDA e da biblioteca pyLDAvis que foi processada durante 3 horas, 15 minutos e 43 segundos. A esquerda da imagem contempla os tópicos extraídos em formato de círculos em um plano bidimensional, sendo o centro de cada tópico utilizado para cálculos como forma de identificar o distanciamento entre os tópicos. A direita da imagem contempla o gráfico de barras com os termos mais frequentes de cada tópico, podendo ser alterados de acordo com a configuração de ajuste de relevância de métricas. O *download* para a visualização dinâmica dos tópicos no formato HTML pode ser realizada através do GitHub[75].

[75] Visualização dinâmica dos tópicos. *Corpus* 10: artigos completos e resumos expandidos 2015. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2015_gts.html/.

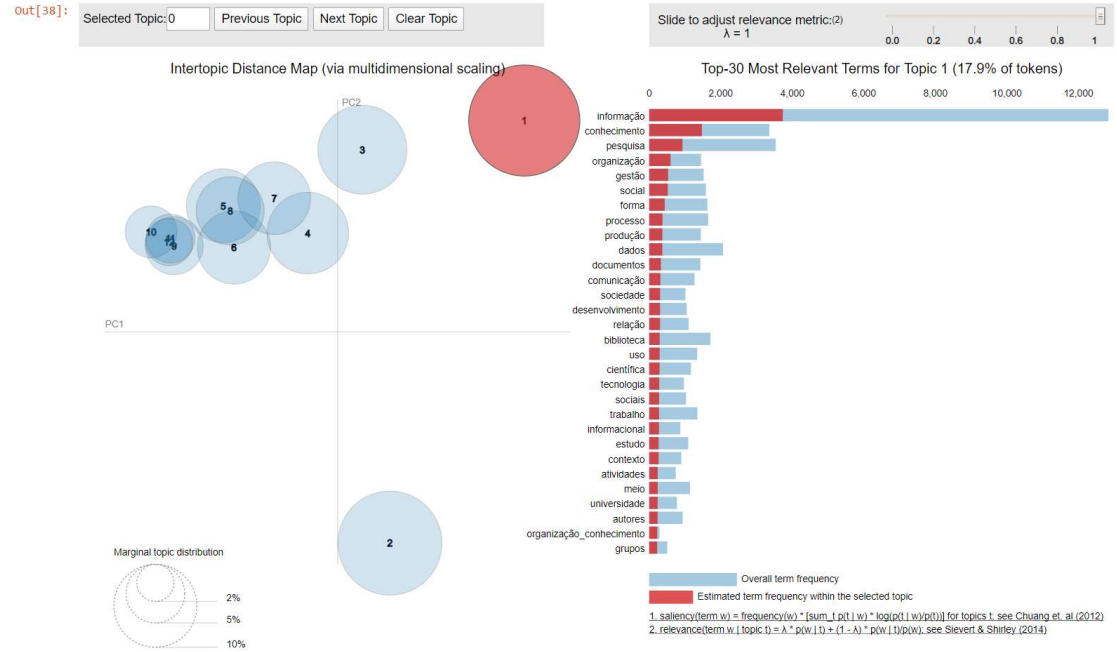
Figura 26 – Visualização geral dos tópicos do *corpus* 10



Fonte: Elaborado pelo autor.

O tópico 1 destacado a esquerda da Figura 27 através do círculo vermelho apresenta um conjunto de 30 tópicos que melhor lhe representam ao lado direito da imagem. Os termos representam 17.9% dos *tokens* e são representados na cor azul para a frequência geral e na cor vermelha para a frequência estimada do termo no tópico selecionado.

Figura 27 – Visualização do tópico 1 do *corpus* 10



Fonte: Elaborado pelo autor.

Torna-se possível identificar uma distância intertópica por meio de escala multidimensional entre os tópicos 1 e 2 representada pela diversidade dos termos encontrados em cada tópico. Mesmo existindo uma aproximação entre o percentual de *tokens* relevantes para cada tópico, sendo 17.9% para o tópico 1 e 15.7% para o tópico 2, a diferença está nas particularidades de cada tópico. Enquanto o tópico 1 apresenta os termos “informação”, “conhecimento”, “pesquisa”, “organização”, e “gestão” no topo de sua lista, o tópico 2 apresenta os termos “informação”, “pesquisa”, “produção”, “conhecimento” e “poder”, ambos configurado com métrica para 1.0. Entretanto, quando ajustado a métrica para 0.2, os resultados apresentam características divergentes, como “organização_conhecimento”, “conhecimento”, “ecm (Enterprise Content Management)”, “Ingwersen[76]” e “suporte_social” para o tópico 1 e “políbio”, “bric”, “políbio_alves[77]”, “Ide (Livro Digital Eletrônico)” e “poder” para o tópico 2. Pode-se supor por exemplo que o tópico 1 esteja relacionado a Ciência, Tecnologia & Inovação e/ou Produção e Comunicação da Informação enquanto o tópico 2 pode abordar conceitos sobre Informação e Memória.

Os tópicos 9 a 12 possuem um volume de termos próximos, assim como os tópicos 5 e 8, sendo necessário por parte de especialista um maior esforço cognitivo para identificação dos tópicos, podendo assim explorar as métricas de relevância e observando os resultados.

[76] Peter Ingwersen, professor da Universidade de Copenhagen – Dinamarca da área de recuperação da informação com pesquisas em aspectos cognitivos da interação usuário-sistema baseado em tarefas.
[77] Políbio Alves, autor e poeta brasileito.

APÊNDICE J - *Corpus* 11: artigos completos e resumos expandidos 2016

O *corpus* 11, também pertencente ao segundo *corpora* de dados é constituído por 389 documentos do tipo artigos completos e resumos expandidos publicados no ano de 2016, tendo assim o tamanho de 15.467kb. Os dados do corpus são constituídos de 1.072.464 unigramas, 1.072.075 bigramas e 1.071.686 trigramas. O Quadro 47 apresenta a lista dos 50 termos mais frequentes organizados por tipos de N-gramas.

Quadro 47 – Lista de N-gramas por ordem de frequência do *corpus* 11

Unigramas
informação,22559; pesquisa,6547; conhecimento,5893; dados,4028; biblioteca,3228; forma,3081; comunicação,3031; memória,3003; social,2996; processo,2888; produção,2718; uso,2653; documentos,2540; organização,2504; gestão,2421; brasil,2272; meio,2217; relação,2170; estudo,2133; científica,2049; universidade,2036; estudos,2031; trabalho,2012; paulo,1921; contexto,1896; desenvolvimento,1880; sociais,1870; tecnologia,1840; usuários,1812; termo,1777; digital,1689; sociedade,1689; cultura,1675; museu,1634; busca,1606; nacional,1606; sistema,1594; autores,1576; representação,1546; modelo,1524; resultados,1521; federal,1511; campo,1494; diferentes,1480; web,1478; relações,1454; cultural,1444; base,1429; sistemas,1429; processos,1428.
Bigramas
universidade_federal,884; informação_tecnologia,756; informação_science,627; produção_científica,593; recuperação_informação,564; informação_conhecimento,502; organização_conhecimento,495; belo_horizonte,404; gestão_informação,389; gestão_conhecimento,350; informação_universidade,334; competência_informação,332; arquitetura_informação,324; informação_comunicação,315; organização_informação,313; uso_informação,312; dissertação_mestrado,312; pesquisa_informação,306; comunicação_científica,302; bases_dados,295; apresentação_comunicação,286; redes_sociais,283; sintagmas_nominais,276; biblioteca_universitárias,269; informação_informação,266; patrimônio_cultural,264; ponto_vista,262; coleta_dados,262; dados_pesquisa,247; nacional_pesquisa,243; fontes_informação,242; tecnologias_informação,239; conhecimento_organization,239; encontro_nacional,233; base_dados,233; informação_brasília,226; modalidade_apresentação,226; representação_informação,225; santa_catarina,215; sistemas_informação,199; ensino_superior,199; tendo_vista,199; porto_alegre,196; preservação_digital,196; universidade_estadual,195; minas_gerais,193; web_semântica,193; produção_conhecimento,189; oswaldo_cruz,186; estudo_caso,184.
Trigramas
nacional_pesquisa_informação,198; encontro_nacional_pesquisa,197; informação_universidade_federal,189; modalidade_apresentação_comunicação,166; apresentação_comunicação_oralresumo,163; universidade_federal_paraíba,147; tecnologias_informação_comunicação,133; dissertação_mestrado_informação,132; international_organization_standardization,122; comunicação_oral_resumo,118; universidade_federal_santa,117; sistemas_organização_conhecimento,116; apresentação_comunicação_oral,115; informação_belo_horizonte,106; federal_santa_catarina,105; gestão_informação_conhecimento,104; fundação_oswaldo_cruz,104; federal_minas_gerais,102;

sintagmas_nominais_descritores,102;	universidade_federal_minas,101;
society_informação_science,99;	american_society_informação,97;
informação_tecnologia_was,94;	resource_description_framework,93
fonte_dados_pesquisa,91;	journal_american_society,88;
organização_representação_conhecimento,88;	informação_science_technology,84;
universidade_federal_bahia,80; brasília_briquet_lemos,75; extensible_markup_language,74;	
machine_readable_cataloging,72;	informação_sociedade_estudos,72;
sistemas_recuperação_informação,69;	conselho_nacional_arquivos,69;
ambientes_informacionais_digitais,67;	comunicação_web_social,67;
fonte_elaborado_autores,63;	library_informação_science,60;
informação_universidade_brasília,59;	dispositivos_comunicação_web,57;
instituição_ensino_superior,52;	arquitetura_organização_informação,52;
biblioteconomia_universidade_federal,51;	ppgci_universidade_federal,51;
produção_comunicação_informação,50;	perspectivas_informação_belo,48;
comunicação_informação_tecnologia,47;	instituto_brasileiro_informação,46;
portal_periódicos_capes,46.	

Fonte: Elaborado pelo autor.

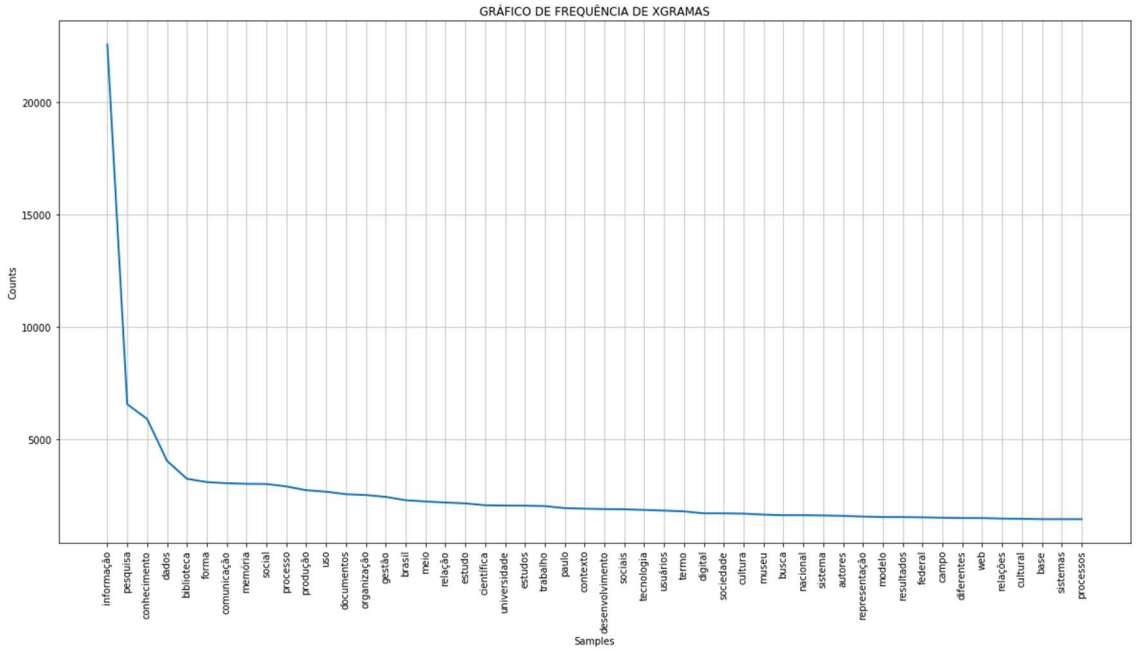
Destaca-se numa lista geral contendo os mil primeiros N-gramas mais frequentes do *corpus* de dados os bigramas “universidade_federal” com frequência de 884 e ocupando a posição de ranqueamento 114, seguido de “informação_tecnologia” com frequência 756 e posição 144, “informação_science” com frequência de 627 e posição 192, “produção_científica” com frequência 593 e posição 216 e “recuperação_informação” com frequência 564 e posição 235. Entre os trigramas mais frequentes estão “nacional_pesquisa_informação” com frequência 198 e posição 926, “encontro_nacional_pesquisa” com frequência 197 e posição 931, “informação_universidade_federal” com frequência 189 e posição 986, além de “modalidade_apresentação_comunicação” com frequência 166 e “apresentação_comunicação_oralresumo” com frequência de 163 que aparecem após ao milésimo termo mais frequente.

Entre os tipos de N-gramas, os unigramas apresentam maior frequência quando se comparado aos demais tipos de termos extraídos do *corpus* de dados. Percebe-se um distanciamento entre os tipos de N-gramas ao se comparar os termos mais frequentes encontrados em cada categoria, como o unigrama “informação” com frequência 22.559 sendo 2.452% maior que o bigrama “universidade_federal” com frequência de 884. Esse percentual aumenta para 11.293% quando comparado o unigrama com o trigrama “nacional_pesquisa_informação” com frequência de 198.

O Gráfico 61 apresenta os 50 termos mais frequentes extraídos do corpus de dados. É possível observar que todos os termos são do tipo unigramas,

destacando dentre os mais frequentes os termos “informação” com frequência de 22.559, “pesquisa” com frequência de 6.547, “conhecimento” com frequência de 5.893, “dados” com frequência de 4.028 e “biblioteca” com frequência de 3.228.

Gráfico 61 – Termos mais frequentes do *corpus* 11



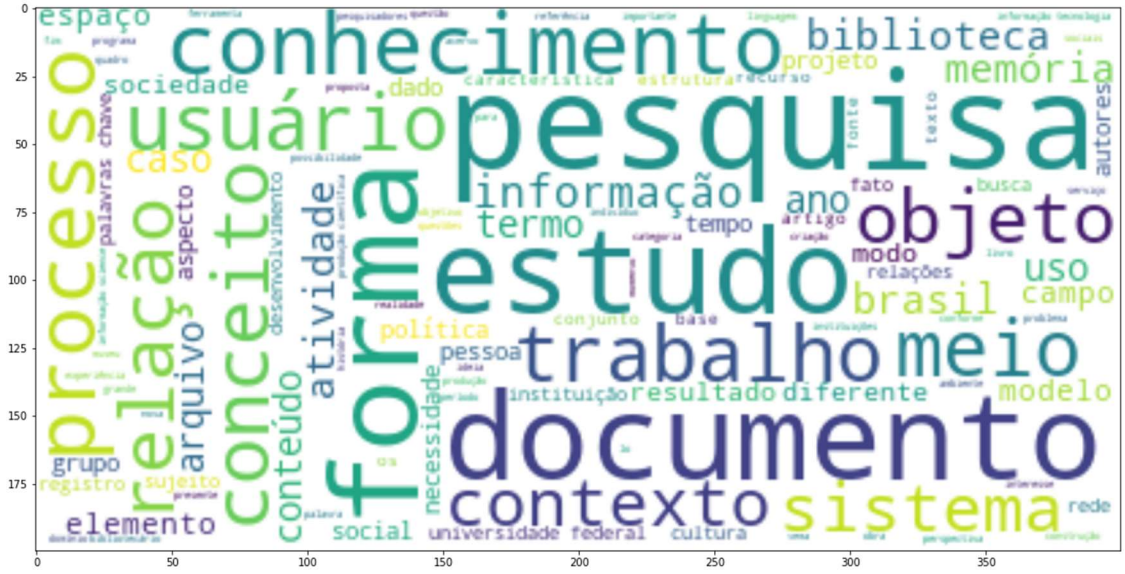
Fonte: Elaborado pelo autor.

É possível observar no gráfico um valor crescente de forma constante entre o 50º e o 2º termo, representados respectivamente pelos termos “processo” com frequência de 1.428 e “pesquisa” com frequência de 6.547 enquanto o primeiro termo “informação” possui frequência de 22.559. É possível perceber uma diferença de 345% entre o primeiro e o segundo termos mais frequentes. Também é possível identificar termos fortes como “conhecimento”, “busca”, “usuários”, “digital” e “gestão” a ser explorados pelo especialista com auxílio das listas de bigramas e trigramas, bem como termos fracos como “brasil”, “meio”, “paulo”, “diferentes” e “federal” podendo ser descartados por não estarem alinhados com domínio da linguagem.

A Figura 28 apresenta uma nuvem de palavras constituída pelos 250 termos mais frequentes do *corpus* de dados. Para construção dessa figura, não foi utilizado nenhum tipo de filtragem dos termos, não sendo realizado assim quaisquer tratamentos qualitativos para sua construção, entretanto, quando analisado o domínio do assunto, torna-se possível identificar tópicos fortes e fracos. Os N-gramas representativos na figura são do tipo unigrama e bigrama,

já que a frequência do primeiro trigrama não aparece dentro do quantitativo de termos estipulados para criação da ilustração.

Figura 28 – Nuvem de palavras do *corpus* 11



Fonte: Elaborado pelo autor.

Destaca-se entre os termos fortes os unigramas “pesquisa”, “relação” e “processo” que apresentam conceitos relacionados ao domínio da linguagem, sobre tudo, quando associados aos bigramas e trigramas. Já os termos “universidade_federal”, “elemento” e “palavra_chave” podem ser considerados fracos por não possuírem relevância ao domínio de linguagem quando realizado a análise de assunto apoiada nas listas de bigramas e trigramas. Como resultado, o *corpus* representado pela figura apresenta maior número de tópicos fortes que o *corpora* 1 constituído por teses e dissertações.

Após a etapa de transformação, os dados foram conectados aos modelos onde passaram pelo processo de treinamento e extração dos tópicos. O Quadro 48 apresenta os resultados do modelo LSI constituído por 30 tópicos e seus respectivos termos e pesos quando aplicado ao *corpus* de dados. O processo de treinamento foi realizado em 43.3 segundos. Os resultados dos conjuntos contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados através do GitHub[78].

[78] Algoritmo de modelagem de tópicos. *Corpus* 11: artigos completos e resumos expandidos 2016. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2016.ipynb/.

Quadro 48 – Tópicos extraídos do *corpus* 11 usando o modelo LSI

Tópico 0: 0.769*"informação" + 0.180*"pesquisa" + 0.168*"conhecimento" + 0.107*"dados" + 0.090*"social" + 0.088*"comunicação" + 0.086*"biblioteca" + 0.083*"forma" + 0.081*"processo" + 0.081*"uso";

Tópico 1: -0.529*"informação" + 0.209*"dados" + 0.202*"pesquisa" + 0.191*"documentos" + 0.169*"memória" + 0.166*"conhecimento" + 0.165*"museu" + 0.128*"biblioteca" + 0.102*"preservação" + 0.102*"patrimônio";

Tópico 2: -0.480*"sintagmas" + -0.434*"nominais" + -0.432*"sintagmas_nominais" + -0.240*"descritores" + -0.161*"nominais_descritores" + -0.161*"sintagmas_nominais_descritores" + -0.157*"critério" + 0.131*"memória" + 0.121*"biblioteca" + -0.105*"indexação";

Tópico 3: 0.430*"memória" + -0.334*"dados" + -0.321*"conhecimento" + 0.208*"biblioteca" + 0.156*"museu" + -0.139*"web" + 0.123*"social" + 0.120*"cultural" + 0.120*"sintagmas" + 0.119*"história";

Tópico 4: 0.508*"conhecimento" + -0.287*"biblioteca" + -0.264*"dados" + -0.237*"metadados" + -0.204*"digital" + -0.173*"preservação" + 0.155*"organização" + -0.130*"documentos" + -0.122*"digitais" + -0.117*"usuários";

Tópico 5: -0.559*"biblioteca" + 0.228*"metadados" + 0.224*"documentos" + 0.212*"memória" + 0.182*"preservação" + -0.138*"comunicação" + 0.134*"arquivos" + 0.131*"digital" + -0.127*"usuários" + -0.121*"pesquisa";

Tópico 6: 0.368*"biblioteca" + -0.292*"pesquisa" + 0.243*"conhecimento" + -0.241*"científica" + -0.207*"produção" + -0.199*"dados" + -0.167*"artigos" + 0.166*"metadados" + 0.136*"organização" + -0.132*"periódicos";

Tópico 7: -0.476*"dados" + -0.255*"memória" + 0.229*"gestão" + 0.169*"metadados" + 0.164*"documentos" + -0.149*"web" + 0.147*"preservação" + -0.145*"data" + 0.141*"digital" + 0.121*"comunicação";

Tópico 8: 0.494*"museu" + -0.391*"memória" + 0.202*"patrimônio" + -0.180*"documentos" + 0.175*"cultural" + -0.158*"arquivos" + 0.156*"objetos" + 0.146*"digital" + -0.124*"arquivo" + 0.105*"arte";

Tópico 9: -0.310*"indexação" + 0.273*"memória" + 0.263*"dados" + 0.218*"gestão" + 0.181*"comunicação" + 0.159*"social" + -0.149*"termo" + 0.148*"conhecimento" + -0.139*"representação" + -0.124*"usuários";

Tópico 10: -0.350*"memória" + 0.231*"documentos" + -0.211*"metadados" + 0.205*"arquivo" + 0.186*"dados" + -0.179*"pesquisa" + -0.169*"preservação" + -0.167*"digital" + 0.165*"arquivos" + 0.140*"gestão";

Tópico 11: 0.274*"usuários" + -0.213*"biblioteca" + 0.191*"gestão" + 0.170*"museu" + -0.155*"social" + -0.149*"linguagem" + -0.146*"metadados" + -0.141*"brasil" + 0.140*"uso" + 0.131*"memória";

Tópico 12: 0.550*"comunicação" + -0.224*"conhecimento" + 0.200*"social" + -0.188*"pesquisa" + -0.154*"patrimônio" + 0.147*"dispositivos" + 0.139*"científica" + 0.131*"indexação" + 0.125*"web" + 0.121*"periódicos";

Tópico 13: -0.207*"indexação" + 0.199*"pesquisa" + 0.185*"competência" + 0.164*"social" + 0.162*"documentos" + -0.155*"gestão" + 0.154*"mediação" + -0.153*"brasil" + 0.152*"usuários" + -0.138*"museu";

Tópico 14: -0.444*"indexação" + -0.179*"digital" + 0.166*"usuários" + 0.150*"arquitetura" + 0.143*"web" + -0.140*"processo" + 0.137*"organização" + -0.137*"termo" + -0.129*"pesquisa" + -0.121*"preservação";

Tópico 15: 0.273*"patrimônio" + 0.240*"web" + 0.188*"cultural" + 0.188*"social" + 0.151*"pesquisa" + 0.133*"dispositivos" + 0.130*"indexação" + 0.130*"documentos" + -0.124*"forma" + -0.113*"leitura";

Tópico 16: -0.230*"qualidade" + -0.199*"objeto" + -0.168*"gestão" + -0.161*"objetos" + 0.152*"indexação" + -0.140*"serviços" + -0.133*"periódicos" + -0.131*"biblioteca" + 0.125*"termo" + -0.121*"digital";
Tópico 17: -0.231*"digital" + -0.203*"qualidade" + 0.183*"objeto" + 0.175*"objetos" + 0.160*"comunicação" + -0.141*"periódicos" + -0.130*"cultural" + -0.122*"mediação" + -0.115*"serviços" + 0.113*"elementos";
Tópico 18: 0.395*"museu" + -0.274*"cultural" + -0.254*"cultura" + -0.204*"patrimônio" + 0.157*"competência" + -0.127*"mediação" + 0.102*"web" + 0.101*"educação" + -0.098*"organização" + -0.095*"organizacional";
Tópico 19: 0.248*"web" + -0.211*"museu" + -0.188*"organização" + -0.183*"digital" + -0.173*"dados" + 0.171*"mediação" + 0.160*"qualidade" + -0.146*"comunicação" + 0.137*"gestão" + 0.134*"cultural";
Tópico 20: -0.222*"arquitetura" + -0.216*"digital" + -0.206*"produção" + 0.168*"comunicação" + -0.162*"arquitetura_informação" + 0.153*"metadados" + 0.142*"competência" + -0.141*"documento" + 0.128*"patrimônio" + 0.110*"transparência";
Tópico 21: 0.338*"social" + -0.202*"digital" + -0.195*"comunicação" + -0.175*"competência" + 0.162*"qualidade" + -0.147*"organização" + 0.135*"metadados" + -0.131*"patrimônio" + 0.130*"modelo" + 0.120*"sociais";
Tópico 22: -0.194*"qualidade" + -0.177*"social" + 0.162*"cultura" + 0.155*"biblioteca" + -0.144*"universidade" + -0.130*"semiótica" + 0.130*"periódicos" + -0.124*"universidades" + 0.115*"documentos" + 0.113*"conhecimento";
Tópico 23: -0.227*"digital" + 0.194*"metadados" + -0.169*"transparência" + 0.164*"museu" + 0.156*"cultura" + -0.146*"social" + -0.143*"uso" + 0.136*"organizacional" + 0.134*"qualidade" + -0.131*"usuários";
Tópico 24: 0.216*"arquitetura" + 0.189*"revisão" + -0.167*"cultura" + -0.159*"empresas" + 0.159*"arquitetura_informação" + 0.158*"editor" + 0.153*"artigo" + 0.146*"processo" + 0.142*"pares" + -0.133*"serviços";
Tópico 25: -0.238*"mediação" + -0.153*"web" + 0.153*"classificação" + 0.150*"tecnologia" + -0.144*"semiótica" + 0.141*"termo" + -0.138*"museu" + 0.133*"patrimônio" + -0.132*"registros" + -0.132*"conhecimento";
Tópico 26: -0.176*"linguagem" + -0.172*"semiótica" + 0.168*"cultura" + 0.141*"social" + -0.140*"patrimônio" + 0.130*"arquitetura" + 0.125*"pesquisa" + -0.123*"tecnologia" + -0.122*"empresas" + 0.118*"arquitetura_informação";
Tópico 27: 0.265*"mediação" + -0.225*"web" + 0.165*"registros" + -0.159*"pesquisa" + -0.142*"cultura" + 0.135*"cultural" + 0.131*"arquivo" + 0.129*"repositório" + -0.121*"comportamento" + -0.120*"uso";
Tópico 28: -0.187*"registros" + -0.147*"linguagem" + 0.134*"redes" + 0.133*"empresas" + 0.129*"semiótica" + 0.119*"rede" + 0.117*"usuários" + 0.114*"artigo" + 0.112*"revisão" + 0.110*"folksonomia";
Tópico 29: 0.219*"registros" + -0.181*"usuários" + 0.157*"revisão" + 0.153*"classificação" + 0.152*"pesquisa" + 0.142*"editor" + -0.126*"indexação" + 0.124*"cultura" + -0.121*"brasil" + 0.120*"pares".

Fonte: Elaborado pelo autor.

Os resultados alcançados através do modelo LSI apresentam grupos de termos específicos que permitem ao especialista realizar a suposição dos nomes dos tópicos sem a necessidade de consultar documentos externos como lista de bigramas e trigramas, além de acesso ao próprio *corpus* de dados, entretanto, existe uma proximidade entre um grupo de tópicos que dificulta a análise de assunto. Existem termos considerados fortes por possuírem pesos elevados

quando se comparado ao demais termos, como por exemplo o termo “informação” nos tópicos 0 e 1 que possuem valores maiores que 0.500. Os termos são ordenados nos tópicos por ordem de relevância com base nos pesos extraídos.

O tópico 1 apresenta os termos generalistas ”informação” e “conhecimento” que podem apresentar outros termos através de composição com significados diferentes, entretanto, constam também os termos chave como “preservação”, “património” e “museu” que permitem ao especialista supor que o rótulo do tópico esteja relacionado a Museologia, Patrimônio e Informação.

Já o tópico 2 apresenta termos especialistas como “sintagmas_nominais”, “descritores” e “indexação” que permitem ao especialista supor que o rótulo do tópico esteja relacionado a Organização e Representação do Conhecimento. Outro termo especialista é “folksonomia”, resultado do tópico 28 que analisado pelo especialista juntamente com termos como “linguagem”, “rede”, “usuários’ e “semiótica” também permite inferir o mesmo rótulo do tópico 2.

Os tópicos 14, 20, 24 e 26 apresentam correlações entre seus termos, sobre tudo, no termo “arquitetura_informação” que é resultado em todos esses tópicos. Dessa forma, o especialista pode optar por unificar o conjunto de tópicos ou selecionar os mais representativos e descartar o de menores representação. Através da análise de assunto o especialista poderá supor por exemplo que os tópicos estejam relacionados a Informação e Tecnologia.

O Quadro 49 apresenta os resultados da extração de tópicos realizada no *corpus* de dados por meio do modelo LDA. O tempo de treinamento do algoritmo foi de 10 minutos e 11 segundos, apresentando assim um conjunto de 30 tópicos e seus respectivos termos e pesos. Os outros resultados com os conjuntos de 10, 14, 18, 22, 26, 34, 38 e 42 tópicos estão disponibilizados no GitHub[79].

Quadro 49 – Tópicos extraídos do corpus 11 usando o modelo LDA

Tópico 0: 0.003*"informação" + 0.002*"web" + 0.001*"conhecimento" + 0.001*"pesquisa" + 0.001*"histórias" + 0.001*"gestão" + 0.001*"modelo" + 0.001*"usuários" + 0.001*"busca" + 0.001*"projetos";
Tópico 1: 0.003*"comunicação" + 0.002*"informação" + 0.002*"conhecimento" + 0.001*"organizacional" + 0.001*"gestão" + 0.001*"organização" + 0.001*"processo" + 0.001*"estudo" + 0.001*"museu" + 0.001*"comunicação_organizacional";

[79] Algoritmo de modelagem de tópicos. *Corpus* 11: artigos completos e resumos expandidos 2016. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2016.ipynb/.

Tópico 2: 0.001*"brics" + 0.000*"artigos" + 0.000*"lattes" + 0.000*"arqueologia" + 0.000*"temáticas" + 0.000*"crescimento" + 0.000*"mapeamento" + 0.000*"mpeg" + 0.000*"international_organization_standardization" + 0.000*"organization_standardization";
Tópico 3: 0.000*"dados" + 0.000*"informação" + 0.000*"pesquisa" + 0.000*"nominais" + 0.000*"conhecimento" + 0.000*"sintagmas_nominais" + 0.000*"sintagmas" + 0.000*"digital" + 0.000*"biblioteca" + 0.000*"comunicação";
Tópico 4: 0.001*"cultural" + 0.001*"brasil" + 0.001*"acessibilidade" + 0.001*"científica" + 0.001*"deficiência" + 0.001*"disciplina" + 0.000*"museu" + 0.000*"mediação" + 0.000*"mediação_cultural" + 0.000*"alunos";
Tópico 5: 0.003*"dados" + 0.001*"reuso" + 0.001*"conteúdo" + 0.001*"data" + 0.001*"reuso_dados" + 0.001*"arquétipos" + 0.001*"sistemas" + 0.001*"openehr" + 0.001*"sentenças" + 0.000*"metadados";
Tópico 6: 0.002*"biblioteca" + 0.001*"informação" + 0.001*"cultural" + 0.001*"repositórios" + 0.001*"profissão" + 0.001*"bolso" + 0.001*"dados" + 0.001*"usuários" + 0.001*"mediação" + 0.001*"serviços";
Tópico 7: 0.001*"obra" + 0.001*"citação" + 0.001*"cruz" + 0.001*"frbr" + 0.001*"oswaldo" + 0.001*"oswaldo_cruz" + 0.000*"atributos" + 0.000*"entidades" + 0.000*"texto" + 0.000*"repositório";
Tópico 8: 0.002*"qualidade" + 0.001*"serviços" + 0.001*"gestão" + 0.001*"gamificação" + 0.001*"arquivo" + 0.001*"ambiente" + 0.001*"controle" + 0.001*"gap" + 0.001*"avaliação" + 0.001*"gestão_qualidade";
Tópico 9: 0.012*"informação" + 0.003*"memória" + 0.003*"conhecimento" + 0.003*"social" + 0.002*"pesquisa" + 0.002*"história" + 0.001*"processo" + 0.001*"forma" + 0.001*"documentos" + 0.001*"sociais";
Tópico 10: 0.005*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.002*"documentos" + 0.001*"dados" + 0.001*"linguagem" + 0.001*"classificação" + 0.001*"organização" + 0.001*"forma" + 0.001*"processo";
Tópico 11: 0.008*"informação" + 0.001*"conhecimento" + 0.001*"pesquisa" + 0.001*"produção" + 0.001*"organização" + 0.001*"comunicação" + 0.001*"cultura" + 0.001*"gestão" + 0.001*"forma" + 0.001*"representação";
Tópico 12: 0.000*"documentos" + 0.000*"fotografia" + 0.000*"audiovisuais" + 0.000*"documentos_audiovisuais" + 0.000*"sonoros" + 0.000*"iconográficos" + 0.000*"ctdais" + 0.000*"iconográficos_sonoros" + 0.000*"audiovisuais_iconográficos" + 0.000*"documentos_audiovisuais_iconográficos";
Tópico 13: 0.002*"conhecimento" + 0.002*"informação" + 0.001*"dados" + 0.001*"organização" + 0.001*"organização_conhecimento" + 0.001*"domínio" + 0.001*"conceitos" + 0.001*"relações" + 0.001*"organization" + 0.001*"sistemas";
Tópico 14: 0.001*"referências" + 0.000*"redes" + 0.000*"eduardo" + 0.000*"conhecimento" + 0.000*"convites" + 0.000*"rede" + 0.000*"literárias" + 0.000*"nonaka" + 0.000*"compartilhamento" + 0.000*"autores";
Tópico 15: 0.015*"informação" + 0.005*"pesquisa" + 0.003*"dados" + 0.003*"conhecimento" + 0.002*"biblioteca" + 0.002*"uso" + 0.002*"produção" + 0.002*"forma" + 0.001*"usuários" + 0.001*"processo";
Tópico 16: 0.001*"sublimação" + 0.001*"coleção" + 0.001*"definição" + 0.001*"dados" + 0.001*"definições" + 0.001*"termo" + 0.000*"conceito" + 0.000*"livros" + 0.000*"uti" + 0.000*"impulso";
Tópico 17: 0.003*"biblioteca" + 0.003*"informação" + 0.002*"conhecimento" + 0.002*"pesquisa" + 0.001*"comunicação" + 0.001*"social" + 0.001*"usuários" + 0.001*"atividades" + 0.001*"uso" + 0.001*"gestão";
Tópico 18: 0.001*"família" + 0.001*"lai" + 0.001*"lei" + 0.001*"arquivos" + 0.000*"lei_informação" + 0.000*"paulo" + 0.000*"cidadão" + 0.000*"fazenda" + 0.000*"regulamentação" + 0.000*"secretarias";

Tópico 19: 0.003*"informação" + 0.002*"pesquisa" + 0.002*"museu" + 0.001*"metadados" + 0.001*"produção" + 0.001*"pesquisadores" + 0.001*"científica" + 0.001*"documentos" + 0.001*"patrimônio" + 0.001*"preservação";
Tópico 20: 0.005*"informação" + 0.002*"preservação" + 0.002*"digital" + 0.001*"metadados" + 0.001*"pesquisa" + 0.001*"instituições" + 0.001*"digitais" + 0.001*"mediação" + 0.001*"preservação_digital" + 0.001*"folksonomia";
Tópico 21: 0.002*"informação" + 0.001*"cultura" + 0.001*"informacional" + 0.001*"informacionais" + 0.001*"arquitetura_informação" + 0.001*"arquitetura" + 0.001*"violência" + 0.001*"flor" + 0.001*"ambientes" + 0.001*"cultura_informacional";
Tópico 22: 0.001*"museu" + 0.001*"informação" + 0.001*"indexação" + 0.001*"interdisciplinaridade" + 0.001*"fotografias" + 0.001*"memória" + 0.001*"avaliação" + 0.001*"pesquisa" + 0.001*"cultura" + 0.001*"usuário";
Tópico 23: 0.001*"microbiologia" + 0.001*"instituto" + 0.001*"educação" + 0.001*"paulo" + 0.001*"patente" + 0.001*"biblioterapia" + 0.001*"rossini" + 0.001*"universidade" + 0.001*"recursos" + 0.001*"biblioteca";
Tópico 24: 0.001*"objetos" + 0.001*"museu" + 0.001*"arte" + 0.000*"art" + 0.000*"objeto" + 0.000*"grafite" + 0.000*"instrumentos" + 0.000*"documentação" + 0.000*"tecnologia" + 0.000*"mast";
Tópico 25: 0.007*"informação" + 0.003*"conhecimento" + 0.002*"comunicação" + 0.002*"pesquisa" + 0.001*"científica" + 0.001*"processo" + 0.001*"social" + 0.001*"objeto" + 0.001*"campo" + 0.001*"forma";
Tópico 26: 0.002*"leitura" + 0.002*"biblioteca" + 0.001*"digital" + 0.001*"cultura" + 0.001*"nacional" + 0.001*"livro" + 0.001*"presença" + 0.001*"periódicos" + 0.001*"brasil" + 0.001*"presença_digital";
Tópico 27: 0.001*"jurema" + 0.001*"periódicos" + 0.001*"doaj" + 0.000*"publicação" + 0.000*"umbanda" + 0.000*"religiosidade" + 0.000*"aberto" + 0.000*"religião" + 0.000*"culto" + 0.000*"seal";
Tópico 28: 0.002*"museu" + 0.001*"memória" + 0.001*"informação" + 0.001*"patrimônio" + 0.001*"pesquisa" + 0.001*"cultural" + 0.001*"instituições" + 0.001*"produção" + 0.001*"governança" + 0.001*"brasil";
Tópico 29: 0.002*"dados" + 0.001*"imagens" + 0.001*"descrição" + 0.001*"imagem" + 0.001*"museologia" + 0.001*"revistas" + 0.001*"informação" + 0.001*"base" + 0.001*"documental" + 0.001*"literatura".

Fonte: Elaborado pelo autor.

Os resultados obtidos através do modelo LDA apresentam tópicos contendo termos generalistas e especialistas. Os termos generalistas exigem do especialista um maior esforço cognitivo para realizar a interpretação dos tópicos, uma vez que esses termos podem gerar diversas composições de significados, como por exemplo “arte” que pode abortar assuntos como “arte_conteporanea”, “obras_arte” ou “arte_visual”.

Um dos recursos a ser utilizado pelo especialista do domínio da linguagem está na exploração de documentos externos como lista de bigramas e trigramas ou mesmo acesso ao *corpus* de dados para uma melhor exploração das informações. Para tópicos contento todos os termos generalistas, sugere-se ao especialista a criação de uma categorial geral para classificação ou mesmo o descarte do tópico. Entretanto, o modelo apresenta em seus resultados termos

especialistas como “religiosidade”, “microbiologia”, “folksonomia” considerados termos chave para realizar a interpretação dos tópicos ou analisar documentos externos de forma que possa identificar rótulos mais assertivo para os tópicos. Os bigramas “preservação_digital” e “arquitetura_informação” também contribuem para atividade.

Outro exemplo de suposição de rótulo para conjunto de termos está no tópico 3, que apresenta uma coesão entre os termos generalistas “dado”, “informação” e “conhecimento” e termos especialistas como “nominais” e “sintagmas_nominais” podendo estar relacionado a área de Organização e Representação do Conhecimento. Mesmo que haja um conjunto de termos contendo todos os pesos com valores iguais a 0, faz-se necessário ao especialista do domínio da linguagem realizar a análise de assunto do tópico, pois os valores dos pesos podem estar associados ao tamanho do corpus de dados ou mesmo na configuração do modelo.

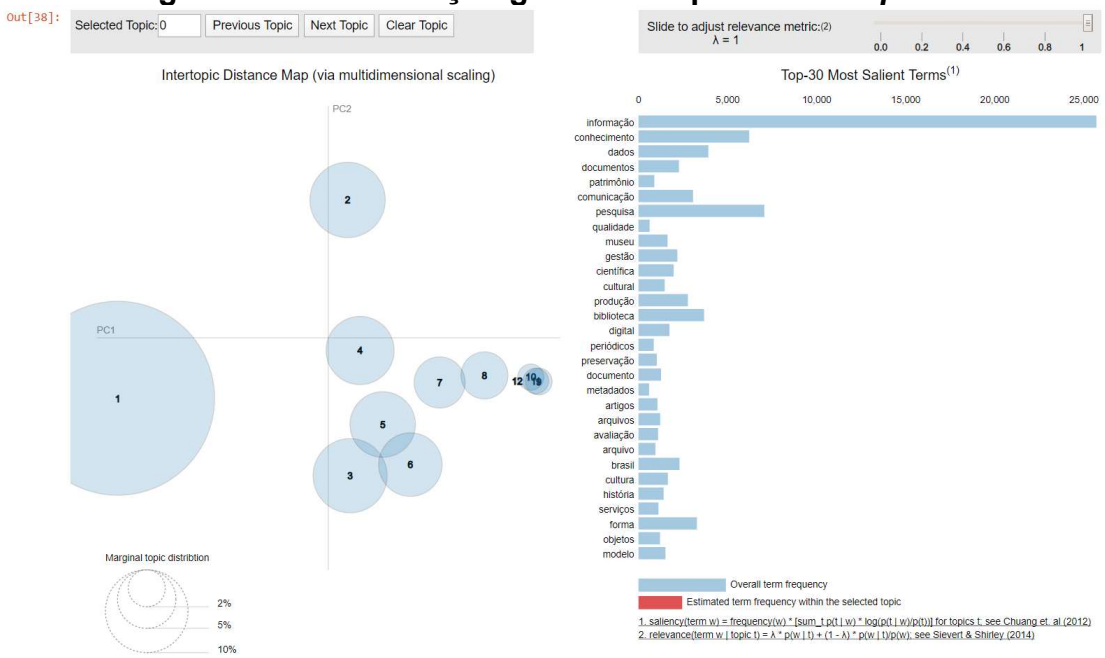
O tópico 11 apresenta termos especialistas como “openehr”, “arquétipos” e “reuso_dados” que possibilita ao profissional especialista explorar documentos externos aos resultados textuais gerados pelo algorítmo, como por exemplo o *corpus* de dados que é ideal para identificação dos documentos que melhor representam os termos especialistas e posteriormente realizar análise através de informações como título, resumo ou grupo de trabalho. Nesse caso, o especialista pode supor que os temas e assuntos do tópico esteja relacionado a Informação e Saúde.

Também é possível encontra entre os resultados da modelagem de extração tópicos fracos que apresentam diferentes temas ou assuntos, como por exemplo, no tópico 21 que contempla o termo “arquitetura_informação” e pode ser associado a Informação e Tecnologia, além de tópicos especializados como “cultura_informacional”, “violência” e “flor” que pode ser associado a área ou tema de Informação e Memória.

A Figura 29 apresenta os resultados obtidos através da visualização dinâmica dos tópicos extraídos do corpus de dados a partir do modelo LDA e usando a biblioteca pyLDAvis. A imagem está dividia em dois grupos, sendo o primeiro, da esquerda, representando os tópicos no formato de círculos no plano bidimensional em um dimensionamento multidimensional para projetar as distâncias intertópicas em duas dimensões, e o segundo, a direita,

representando os termos individualmente contidos nos tópicos no formato de gráfico de barras utilizados para identificação dos tópicos. O processo para gerar a visualização dinâmica dos tópicos foi realizado durante 5 horas, 4 minutos e 4 segundos. O *download* do arquivo no formato HTML contendo a visualização dinâmica dos tópicos pode ser realizada através do GitHub[80].

Figura 29 – Visualização geral dos tópicos do *corpus* 11

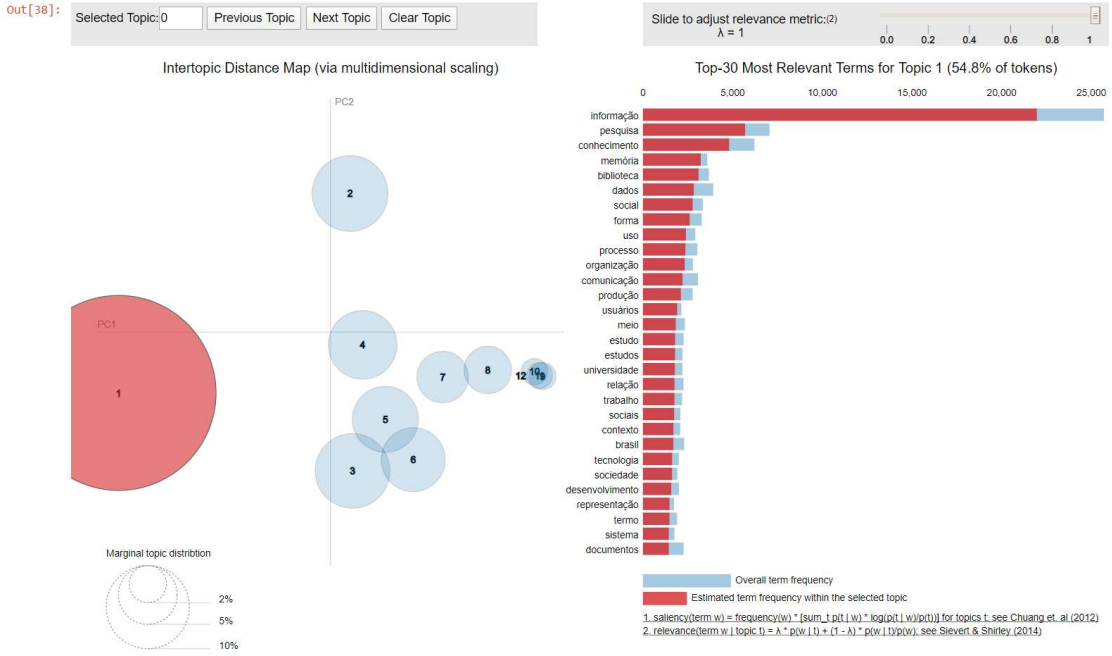


Fonte: Elaborado pelo autor.

A Figura 30 destaca à esquerda o tópico 1, representado pelo círculo vermelho como o tópico mais relevante do *corpus* de dados. Já a esquerda da figura estão as 30 palavras mais relevantes do tópico, representando assim 54.8% dos *tokens* como apresentado no gráfico de barras. A cor azul representa a frequência geral e a cor vermelha representa a frequência estimada do termo no tópico selecionado.

[80] Visualização dinâmica dos tópicos. *Corpus* 11: artigos completos e resumos expandidos 2016. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2016_gts.html/.

Figura 30 – Visualização do tópico 1 do *corpus* 11



Fonte: Elaborado pelo autor.

O tópico 1 do *corpus* de dados apresentas características generalistas por representar através dos seus termos o percentual de 54.8% dos *tokens* mais relevantes para o tópico. Comparando por exemplo aos demais primeiros tópicos de outro *corpus* que utilizam a visualização dinâmica, estão o corpus 13 com 31.8%, corpus 8 com 30.7%, corpus 7 com 28.5%, corpus 9 com 18.3%, corpus 10 com 17,9% e corpus 12 com 15.8%. Mediante a esse percentual elevado, os termos se tornam de difícil interpretação por parte do especialista ao realizar a análise de assuntos para a suposição do nome do tópico, uma vez que os termos do topo da lista como “informação”, “pesquisa”, “conhecimento”, “memória”, “biblioteca”, “dados”, “social”, “forma”, “uso”, e “processo” utilizando como métrica o valor 1.0 podem ser encontrados em diversas áreas da Ciência da Informação por serem termos generalistas.

Com o percentual elevado dos *tokens*, pôde-se obter termos específico somente quando configurado a métrica para valor igual a 0.0, surgindo assim, termos no topo da lista como “bim”, “folksonomia”, “jurema”, “bamidelê”, “eye_tracking” que permite supor rótulos como Organização e Representação do Conhecimento, Informação e Tecnologia e Informação e Memória.

APÊNDICE K - *Corpus* 12: artigos completos e resumos expandidos 2017

O *corpus* 12 é formado por 387 documentos do tipo artigos completos e resumos expandidos publicados nos anais do ENANCIB no ano de 2017 e faz parte dos dados constituintes do *segundo* corpora. O *corpus* possui o tamanho de 15.610kb gerando em seu conteúdo o quantitativo de 1.047.521 unigramas, 1.047.134 bigramas e 1.046.747 trigramas. O Quadro 50 apresenta a lista dos 50 termos mais frequentes organizados por tipos de N-gramas.

Quadro 50 – Lista de N-gramas por ordem de frequência do *corpus* 12

Unigramas
informação,23111; pesquisa,7415; conhecimento,7135; dados,5606; biblioteca,3477; forma,3291; social,2967; processo,2922; gestão,2916; documentos,2758; produção,2757; organização,2660; uso,2651; meio,2505; comunicação,2467; relação,2450; estudo,2274; trabalho,2269; sociais,2256; estudos,2214; contexto,2185; termo,2129; universidade,2110; autores,2093; brasil,2073; desenvolvimento,2050; científica,1995; sociedade,1922; tecnologia,1915; digital,1830; artigos,1787; paulo,1779; resultados,1739; representação,1725; museu,1695; cultural,1681; busca,1677; relações,1633; fonte,1622; memória,1614; sistema,1613; conceito,1610; campo,1581; construção,1579; cultura,1572; conceitos,1562; processos,1536; data,1521; diferentes,1518; web,1485.
Bigramas
informação_tecnologia,1007; universidade_federal,853; produção_científica,675; informação_science,654; recuperação_informação,632; organização_conhecimento,610; informação_conhecimento,581; gestão_conhecimento,540; belo_horizonte,535; gestão_informação,510; competência_informação,449; pesquisa_informação,443; dados_pesquisa,438; redes_sociais,394; modalidade_apresentação,385; big_data,381; uso_informação,376; gestão_documentos,368; informação_informação,352; universidade_estadual,346; informação_comunicação,344; comunicação_oral,332; representação_conhecimento,332; minas_gerais,327; coleta_dados,302; fontes_informação,300; fonte_elaborado,293; bases_dados,289; nacional_pesquisa,287; fonte_dados,285; apresentação_comunicação,281; oral_resumo,278; ponto_vista,272; encontro_nacional,271; base_dados,269; conhecimento_organization,269; porto_alegre,268; tecnologias_informação,266; patrimônio_cultural,266; organização_representação,263; sistemas_informação,256; dissertação_mestrado,249; biblioteca_universitárias,243; tomada_decisão,242; ensino_superior,230; tendo_vista,229; comunicação_científica,224; artigos_publicados,223; muitas_vezes,215; linked_data,214.
Trigramas
apresentação_comunicação_oral,279; modalidade_apresentação_comunicação,278; comunicação_oral_resumo,278; nacional_pesquisa_informação,268; encontro_nacional_pesquisa,248; fonte_dados_pesquisa,248; organização_representação_conhecimento,205; fonte_elaborado_autores,183; tecnologias_informação_comunicação,182; universidade_federal_minas,165; federal_minas_gerais,165; international_organization_standardization,148; universidade_estadual_unesp,130; informação_tecnologia_was,130; gestão_informação_conhecimento,126; universidade_federal_paraíba,118; resource_description_framework,111; informação_belo_horizonte,108; modalidade_apresentação_resumo,107; informação_universidade_federal,99;

society_informação_science,90; american_society_informação,88; library_informação_science,87; customer_relationship_management,87; informação_science_technology,86; federal_paraíba_ufpb,84; journal_american_society,83; dissertação_mestrado_informação,82; classificação_decimal_universal,80; universidade_federal_santa,76; informação_sociedade_estudos,75; universidade_federal_fluminense,74; produção_comunicação_informação,72; federal_santa_catarina,71; sistema_integrado_biblioteca,69; minas_gerais_ufmg,68; universidade_federal_grande,65; pesquisa_brasileira_informação,64; sistemas_organização_conhecimento,63; perspectivas_informação_belo,62; instituições_ensino_superior,60; gerais_belo_horizonte,59; gt-2_organização_representação,59; minas_gerais_belo,58; doc_doc_doc,58; mediação_circulação_apropriação,57; sistema_recuperação_informação,57; informação_comunicação_tic,56; informação_gestão_conhecimento,56; comunicação_informação_tecnologia,56.

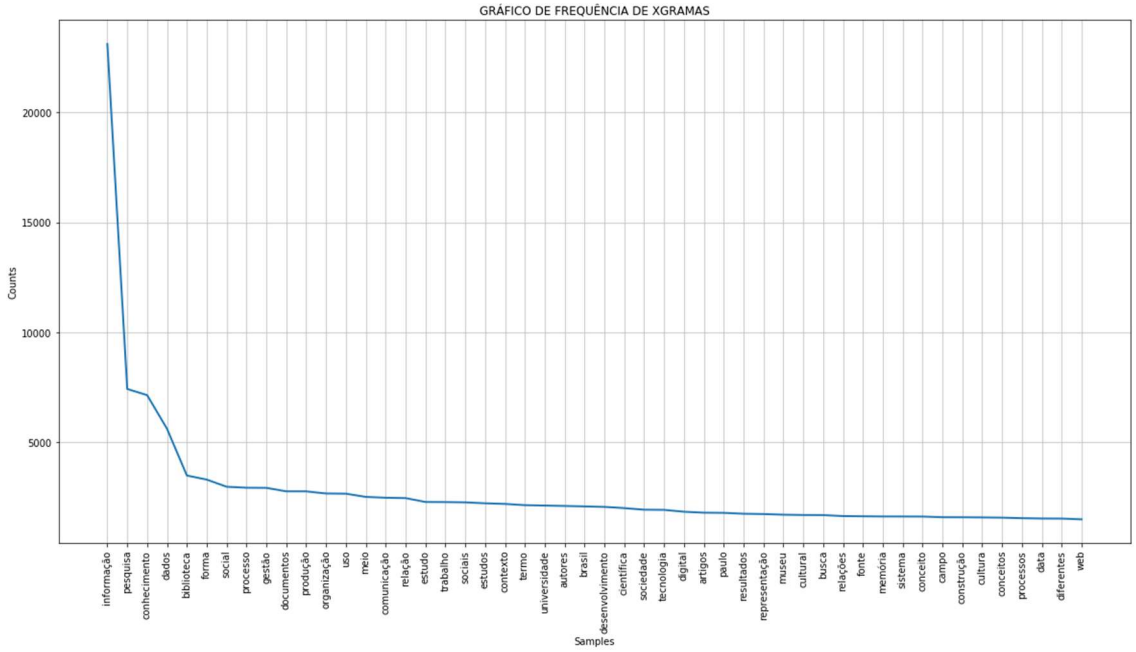
Fonte: Elaborado pelo autor.

Entre os mil termos com maior frequência extraídos do *corpus* de dados estão os bigramas “informação_tecnologia” com frequência de 1.007 e ocupando a posição 98 de ranqueamento, “universidade_federal” com frequência de 853 e posição 134, “produção_científica” com frequência de 675 e posição 188, “informação_science” com frequência de 654 e posição 188 e “recuperação_informação” com frequência de 632 e posição 212. Entre os trigramas mais frequentes estão os termos “apresentação_comunicação_oral” com frequência de 279 e posição 670, “modalidade_apresentação_comunicação” com frequência de 278 e posição 674, “comunicação_oral_resumo” com frequência de 278 e posição 675, “nacional_pesquisa_informação” com frequência de 268 e posição 713 e “encontro_nacional_pesquisa” com frequência de 248 e posição 780. Faz-se necessário ressaltar que são apresentados os termos mais frequentes sem levar em consideração critérios de qualidade a ser realizado pelo especialista do domínio da linguagem. Todos os primeiros termos de cada tipo de N-gramas estão entre o milésimo termo da lista geral de N-gramas.

Os termos do tipo unigramas apresentam maior frequência quando se comparado aos bigramas e trigramas, sendo possível perceber um distanciamento entre os primeiros termos de cada tipo de N-gramas. O unigrama “informação” com frequência de 23.111 é 2.115% maior que o bigrama “informação_tecnologia” com frequência de 1.007. Esse percentual chega a 8.184% quando comparado o unigrama com o trigrama “apresentação_comunicação_oral” com frequência de 279.

O Gráfico 62 apresenta os 50 N-gramas mais frequentes extraídos do *corpus* de dados. É possível observar através dos resultados que todos os termos são do tipo unigramas. Entre os termos mais frequentes estão “informação” com frequência de 23.111, “pesquisa” com frequência de 7.415, “conhecimento” com frequência de 7.135, “dados” com frequência de 5.606 e “biblioteca” com frequência de 3.477.

Gráfico 62 – Termos mais frequentes do *corpus* 12



Fonte: Elaborado pelo autor.

Do 50º termo “web” com frequência de 1.485 ao 2º termo “pesquisa” com frequência de 7.415 é possível observar uma frequência crescente de forma constante entre os termos, entretanto, do 2º ao 1º termo “informação” com frequência de 23.111 percebe-se uma variação no volume de frequência, alcançando assim uma diferença de 223%. Também é possível perceber nesse grupo de dados a existência de termos fortes como “uso”, “sociais”, “cultural”, “processo” e “produção” que podem ser aprofundados pelo especialista ao utilizar as listas de bigramas e trigramas, além de termos fracos como “diferentes”, “paulo”, “brasil”, “autores” e “trabalho” que podem ser descartados pelo especialista por não estarem alinhados ao domínio de linguagem.

Com base na frequência dos N-gramas geradas pelo algoritmo a partir do *corpus* de dados foi possível gerar uma nuvem de palavras contendo 250 termos do tipo unigrama e bigrama apresentado na Figura 31. Faz importante ressaltar

que não foi utilizando nenhum tipo de tratamento qualitativo para criação da ilustração.

Figura 31 – Nuvem de palavras do *corpus* 12

Fonte: Elaborado pelo autor.

Na figura é possível identificar exemplos de termos fortes e fracos, sendo o primeiro possuindo significado de relevância para o domínio da linguagem estudada e o segundo possuindo pouco ou nenhuma relevância ou significado. Pode-se utilizar a lista de bigramas e trigramas extraídos do *corpus* 12 para realizar uma melhor interpretação dos termos do tipo unigramas. Exemplo de termos fortes são “sociedade”, “termo” e “informação_tecnologia” e de fracos “contexto”, “universidade_federal” e “palavra_chave”. O *corpus* apresenta um quantitativo maior de tópicos fortes quando se comparado ao *corpora* 1 constituído de documentos do tipo teses e dissertações.

Os dados foram conectados aos modelos de extração que resultaram em conjuntos de tópicos, termos e pesos. O Quadro 51 apresenta os resultados da modelagem de tópicos contendo 30 desses conjuntos que foram extraídos utilizando o modelo LSI e levou 37.9 segundos para ser treinado. Os resultados dos conjuntos de 10, 14, 18, 22, 26, 34, 38 e 42 tópicos podem ser acessados através do GitHub[81].

[81] Algoritmo de modelagem de tópicos. *Corpus* 12: artigos completos e resumos expandidos 2017. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2017.ipynb/.

Quadro 51 – Tópicos extraídos do *corpus* 12 usando o modelo LSI

Tópico 0: 0.762*"informação" + 0.216*"conhecimento" + 0.192*"pesquisa" + 0.140*"dados" + 0.090*"gestão" + 0.086*"forma" + 0.080*"processo" + 0.076*"organização" + 0.076*"biblioteca" + 0.074*"social";

Tópico 1: 0.531*"informação" + -0.483*"dados" + -0.250*"conhecimento" + -0.197*"pesquisa" + -0.163*"data" + -0.162*"biblioteca" + -0.132*"documentos" + -0.116*"gestão" + -0.085*"museu" + -0.080*"artigos";

Tópico 2: 0.732*"conhecimento" + -0.420*"dados" + 0.198*"gestão" + 0.160*"organização" + -0.118*"informação" + -0.117*"data" + -0.117*"pesquisa" + 0.116*"organizacional" + 0.106*"gestão_conhecimento" + -0.067*"web";

Tópico 3: 0.435*"biblioteca" + -0.434*"dados" + -0.255*"conhecimento" + 0.168*"cultural" + 0.165*"documentos" + -0.163*"data" + 0.157*"museu" + -0.148*"informação" + 0.119*"memória" + 0.109*"social";

Tópico 4: 0.551*"biblioteca" + -0.453*"documentos" + -0.185*"arquivos" + -0.181*"arquivo" + 0.163*"conhecimento" + -0.161*"museu" + -0.151*"gestão" + -0.147*"documento" + -0.120*"gestão_documentos" + -0.105*"classificação";

Tópico 5: 0.317*"biblioteca" + -0.290*"pesquisa" + 0.248*"gestão" + 0.248*"documentos" + -0.211*"produção" + -0.204*"artigos" + -0.190*"científica" + -0.153*"autores" + 0.145*"dados" + -0.140*"social";

Tópico 6: 0.364*"gestão" + -0.223*"representação" + 0.216*"pesquisa" + -0.182*"termo" + 0.161*"documentos" + -0.157*"conceitos" + -0.141*"web" + 0.125*"artigos" + -0.124*"conceito" + -0.119*"indexação";

Tópico 7: 0.446*"museu" + 0.207*"cultural" + 0.200*"patrimônio" + -0.173*"indexação" + 0.170*"memória" + -0.168*"biblioteca" + -0.147*"classificação" + 0.146*"gestão" + -0.143*"termo" + -0.140*"documentos";

Tópico 8: 0.405*"leitura" + -0.357*"museu" + 0.271*"leitor" + 0.156*"jornal" + -0.150*"biblioteca" + 0.149*"dia" + 0.139*"social" + 0.120*"bom" + 0.119*"mediação" + 0.112*"discurso";

Tópico 9: 0.277*"digital" + -0.194*"biblioteca" + -0.156*"cultural" + -0.155*"dados" + 0.145*"uso" + -0.144*"informação" + 0.141*"pesquisa" + 0.136*"avaliação" + 0.128*"rede" + 0.125*"processo";

Tópico 10: -0.312*"museu" + -0.295*"leitura" + 0.255*"memória" + -0.230*"leitor" + 0.191*"social" + 0.161*"digital" + -0.156*"gestão" + 0.150*"sociais" + -0.129*"jornal" + -0.127*"artigos";

Tópico 11: -0.521*"pesquisa" + -0.192*"digital" + 0.183*"artigos" + 0.174*"web" + 0.163*"social" + 0.150*"sociais" + 0.136*"gestão" + 0.120*"produção" + 0.106*"autores" + 0.105*"relações";

Tópico 12: 0.307*"sintagmas" + 0.296*"indexação" + 0.284*"nominais" + 0.281*"sintagmas_nominais" + 0.160*"memória" + 0.155*"automática" + 0.143*"indexação_automática" + -0.138*"classificação" + 0.136*"social" + 0.135*"sociais";

Tópico 13: -0.297*"digital" + 0.174*"pesquisa" + -0.154*"artigos" + -0.148*"digitais" + -0.145*"web" + 0.145*"classificação" + -0.139*"produção" + -0.138*"leitor" + 0.126*"social" + -0.125*"leitura";

Tópico 14: 0.261*"gestão" + -0.256*"avaliação" + -0.185*"competência" + 0.149*"leitor" + 0.134*"termo" + 0.125*"pesquisa" + -0.125*"arquivo" + -0.125*"sintagmas" + -0.116*"nominais" + -0.114*"sintagmas_nominais";

Tópico 15: -0.283*"cultural" + -0.257*"pesquisa" + 0.223*"digital" + 0.193*"museu" + -0.187*"mediação" + 0.156*"classificação" + -0.143*"cultura" + -0.140*"inovação" + -0.137*"gestão" + 0.130*"dados";

Tópico 16: -0.375*"memória" + 0.180*"museu" + 0.167*"documento" + 0.167*"documentos" + -0.165*"avaliação" + -0.154*"sistema" + 0.154*"rede" + -0.142*"tesauro" + -0.123*"direitos" + -0.116*"direito";

Tópico 17: -0.330*"digital" + 0.205*"web" + -0.200*"cultural" + -0.151*"big" + 0.151*"sistema" + -0.150*"data" + -0.146*"big_data" + 0.142*"pesquisa" + 0.135*"memória" + -0.129*"cultura";
Tópico 18: 0.244*"memória" + -0.229*"cultural" + -0.213*"classificação" + -0.208*"direito" + 0.197*"big" + 0.189*"big_data" + -0.173*"web" + 0.157*"data" + 0.156*"documento" + 0.121*"representação";
Tópico 19: 0.290*"big" + 0.280*"big_data" + 0.272*"data" + 0.261*"inovação" + -0.202*"gestão" + -0.169*"dados" + -0.128*"competência" + 0.120*"brasil" + -0.120*"memória" + 0.108*"direito";
Tópico 20: -0.270*"classificação" + 0.249*"arquivo" + 0.230*"belo" + 0.226*"horizonte" + 0.225*"belo_horizonte" + -0.191*"memória" + 0.173*"tesauro" + 0.164*"termo" + 0.153*"municipal" + -0.144*"inovação";
Tópico 21: 0.305*"avaliação" + -0.198*"memória" + -0.188*"web" + 0.175*"sistema" + -0.161*"competência" + 0.138*"processo" + 0.129*"cultural" + 0.124*"mediação" + -0.122*"direitos" + 0.120*"cultura";
Tópico 22: -0.256*"brazil" + 0.199*"classificação" + -0.198*"maps" + -0.170*"mapas" + -0.165*"brasil" + 0.159*"artigos" + -0.141*"avaliação" + -0.123*"produção" + 0.122*"usuários" + 0.121*"big";
Tópico 23: 0.227*"brazil" + 0.209*"autores" + 0.187*"maps" + 0.154*"mapas" + 0.151*"classificação" + 0.132*"artigos" + -0.132*"produção" + 0.127*"cultural" + -0.114*"programas" + -0.112*"universidade";
Tópico 24: 0.444*"inovação" + -0.325*"avaliação" + -0.135*"qualidade" + 0.125*"serviços" + 0.121*"dados" + -0.118*"memória" + 0.117*"documento" + -0.114*"data" + -0.110*"classificação" + 0.104*"empresas";
Tópico 25: -0.209*"produção" + -0.187*"recuperação" + 0.165*"avaliação" + 0.153*"memória" + -0.149*"cultural" + -0.149*"recuperação_informação" + 0.139*"digital" + -0.132*"competência" + -0.132*"data" + -0.131*"big";
Tópico 26: -0.198*"mediação" + 0.187*"inovação" + -0.172*"usuários" + -0.167*"sistema" + 0.127*"produção" + 0.121*"organização" + 0.102*"classificação" + 0.098*"biblioteca" + 0.097*"científica" + 0.095*"recuperação";
Tópico 27: 0.232*"autores" + -0.178*"artigos" + -0.161*"modelo" + 0.160*"trabalhos" + -0.144*"capital" + 0.133*"mediação" + -0.122*"periódicos" + -0.120*"científico" + -0.118*"rda" + -0.117*"sistema";
Tópico 28: 0.249*"documento" + 0.175*"tesauro" + -0.173*"gestão" + -0.159*"organização" + 0.125*"termo" + -0.123*"trabalhos" + 0.120*"avaliação" + -0.117*"autores" + 0.108*"cultural" + 0.107*"artigos";
Tópico 29: -0.248*"organização" + 0.154*"classificação" + 0.144*"histórias" + -0.143*"representação" + -0.143*"organização_conhecimento" + -0.138*"cultural" + -0.135*"leitor" + 0.129*"termo" + -0.118*"social" + -0.109*"jornal".

Fonte: Elaborado pelo autor.

É possível identificar através dos resultados alcançados por meio do modelo LSI um grupo de tópicos que contemplam termos fortes quando comparado aos demais termos, dentre eles se destacam “informação” nos tópicos 0 e 1 e “conhecimento” no tópico 2 como peso acima de 0.500. Além disso, a existência de tópicos que contemplam termos generalistas como “organização” pode ser encontrada no tópico 0. Esse tipo de termo possibilita composições para outros termos com significados diferentes como “organização_conhecimento” e “organização_informação”, cabendo assim ao especialista do domínio da linguagem realizar uma análise de assunto

explorando documentos externos como listas de bigramas e trigramas gerada pelo algoritmo ou mesmo acessando o próprio *corpus* de dados. Além disso, o especialista poderá optar por criar uma categoria geral para tópicos desse tipo ou excluí-los da análise.

O tópico 2 apresentar termos generalistas e comuns ao domínio de linguagem de forma que possam apresentar diversas composições de significados como “conhecimento” e “informação”, entretanto, quando analisado o contexto com termos mais específicos como “dados”, “web”, “organizacional” e “gestão_conhecimento” permite-se ao especialista inferir uma suposição que o rótulo do termo esteja associado a Gestão da Informação e do Conhecimento.

Os tópicos 24 e 25 apresentam um conjunto de termos específicos como “big_data”, “recuperação_informação” e generalistas como “documento”, ”memória”, “qualidade” e “produção” que também possibilita ao especialista realizar a análise de contexto dos termos e supor que o rótulo do tópico esteja relacionado a Informação e Tecnologia. Outro exemplo de termo específico está no tópico 28 que apresenta dentre seus resultados o termo “tesauro”, entretanto, os demais termos não possibilitam identificar de forma clara a área de assunto do tópico, podendo estar relacionado as áreas de Informação e Tecnologia ou Organização e Representação do Conhecimento, dessa forma, cabe ao especialista explorar os documentos externos para realizar uma suposição do rótulo de maneira mais assertiva.

O Quadro 52 apresenta os resultados da modelagem de tópicos utilizando o modelo de extração LDA. Os resultados são constituídos por um conjunto de 30 tópicos com os seus respectivos termos e pesos que foram processados durante 9 minutos e 2 segundos. Os demais resultados treinados pelo modelo contendo 10, 14, 18, 22, 26, 34, 38 e 42 tópicos podem ser acessados através do GitHub [82].

Quadro 52 – Tópicos extraídos do *corpus* 12 usando o modelo LDA

Tópico 0: 0.001*"museu" + 0.001*"data" + 0.001*"dados" + 0.001*"web" + 0.001*"leitura" + 0.001*"informação" + 0.001*"barroso" + 0.001*"linked" + 0.001*"linked_data" + 0.001*"cidoc";
Tópico 1: 0.002*"leitura" + 0.001*"livro" + 0.001*"municipal" + 0.001*"clubes" + 0.001*"literatura" + 0.001*"memória" + 0.001*"paulo" + 0.001*"cidade" + 0.001*"livro_leitura" + 0.001*"cultural";

[82] Algoritmo de modelagem de tópicos. *Corpus* 12: artigos completos e resumos expandidos 2017. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2017.ipynb/.

Tópico 2: 0.014*"informação" + 0.003*"pesquisa" + 0.002*"conhecimento" + 0.002*"produção" + 0.001*"comunicação" + 0.001*"dados" + 0.001*"forma" + 0.001*"científica" + 0.001*"sociedade" + 0.001*"social";
Tópico 3: 0.004*"informação" + 0.002*"museu" + 0.002*"pesquisa" + 0.002*"biblioteca" + 0.001*"memória" + 0.001*"arquivo" + 0.001*"sociais" + 0.001*"conhecimento" + 0.001*"gestão" + 0.001*"documentos";
Tópico 4: 0.002*"cultural" + 0.002*"patrimônio" + 0.002*"museu" + 0.001*"culturais" + 0.001*"observatórios" + 0.001*"observatório" + 0.001*"imaterial" + 0.001*"cultura" + 0.001*"patrimônio_cultural" + 0.001*"unesco";
Tópico 5: 0.005*"informação" + 0.003*"conhecimento" + 0.002*"pesquisa" + 0.001*"autores" + 0.001*"brasil" + 0.001*"direito" + 0.001*"dados" + 0.001*"termo" + 0.001*"organização" + 0.001*"estudo";
Tópico 6: 0.002*"conhecimento" + 0.002*"organização" + 0.002*"organização_conhecimento" + 0.001*"almanaques" + 0.001*"coleção" + 0.001*"organization" + 0.001*"almanaque" + 0.001*"isko-brasil" + 0.001*"conhecimento_organization" + 0.001*"ancib";
Tópico 7: 0.003*"informação" + 0.002*"dados" + 0.002*"biblioteca" + 0.001*"pesquisa" + 0.001*"manuais" + 0.001*"uso" + 0.001*"deficiência" + 0.001*"indicadores" + 0.001*"artigos" + 0.001*"desenvolvimento";
Tópico 8: 0.004*"informação" + 0.003*"conhecimento" + 0.002*"museu" + 0.002*"dados" + 0.002*"pesquisa" + 0.001*"web" + 0.001*"organização" + 0.001*"representação" + 0.001*"forma" + 0.001*"autores";
Tópico 9: 0.000*"gestão" + 0.000*"informação" + 0.000*"museu" + 0.000*"qualidade" + 0.000*"dados" + 0.000*"pesquisa" + 0.000*"data" + 0.000*"publicações" + 0.000*"base" + 0.000*"inteligentes";
Tópico 10: 0.013*"informação" + 0.003*"conhecimento" + 0.002*"pesquisa" + 0.002*"dados" + 0.002*"gestão" + 0.002*"biblioteca" + 0.001*"uso" + 0.001*"forma" + 0.001*"termo" + 0.001*"organização";
Tópico 11: 0.004*"informação" + 0.002*"trabalho" + 0.001*"digital" + 0.001*"rda" + 0.001*"dados" + 0.001*"serviços" + 0.001*"conhecimento" + 0.001*"pesquisa" + 0.001*"forma" + 0.001*"marx";
Tópico 12: 0.004*"documentos" + 0.003*"arquivo" + 0.003*"arquivos" + 0.002*"dados" + 0.001*"informação" + 0.001*"gestão" + 0.001*"municipal" + 0.001*"arquivística" + 0.001*"belo" + 0.001*"belo_horizonte";
Tópico 13: 0.005*"informação" + 0.002*"dados" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"documentos" + 0.001*"avaliação" + 0.001*"data" + 0.001*"processo" + 0.001*"classificação" + 0.001*"termo";
Tópico 14: 0.001*"cidade" + 0.001*"digital" + 0.001*"bairro" + 0.001*"inclusão" + 0.001*"aplicativo" + 0.000*"nomen" + 0.000*"assunto" + 0.000*"inclusão_digital" + 0.000*"cidadãos" + 0.000*"frsad";
Tópico 15: 0.001*"programas" + 0.001*"indicadores" + 0.001*"conceito" + 0.001*"qualis" + 0.001*"avaliação" + 0.001*"livro" + 0.001*"didático" + 0.001*"educação" + 0.001*"total" + 0.001*"produção";
Tópico 16: 0.005*"informação" + 0.005*"pesquisa" + 0.004*"dados" + 0.004*"conhecimento" + 0.002*"gestão" + 0.001*"social" + 0.001*"capital" + 0.001*"artigos" + 0.001*"produção" + 0.001*"processos";
Tópico 17: 0.000*"feministas" + 0.000*"feminismo" + 0.000*"estudos_feministas" + 0.000*"periódico" + 0.000*"estudos" + 0.000*"mulheres" + 0.000*"feminista" + 0.000*"revista_estudos_feministas" + 0.000*"revista_estudos" + 0.000*"produtivos";
Tópico 18: 0.004*"informação" + 0.002*"conhecimento" + 0.002*"representação" + 0.002*"documento" + 0.001*"pesquisa" + 0.001*"relação" + 0.001*"sistema" + 0.001*"relações" + 0.001*"classificação" + 0.001*"documentos";

Tópico 19: 0.004*"conhecimento" + 0.003*"informação" + 0.001*"gestão" + 0.001*"organização" + 0.001*"conceitos" + 0.001*"processo" + 0.001*"competências" + 0.001*"meio" + 0.001*"seleção" + 0.001*"trabalho";
Tópico 20: 0.001*"redes" + 0.001*"amado" + 0.001*"jorge" + 0.001*"científica" + 0.001*"redes_sociais" + 0.001*"indicadores" + 0.001*"sociais" + 0.001*"jorge_amado" + 0.001*"objetos" + 0.001*"escritor";
Tópico 21: 0.002*"tesauro" + 0.001*"software" + 0.001*"anotação" + 0.001*"termo" + 0.001*"multimídia" + 0.001*"requisitos" + 0.001*"synaptica" + 0.001*"softwares" + 0.001*"multites" + 0.001*"uso";
Tópico 22: 0.011*"informação" + 0.004*"pesquisa" + 0.003*"biblioteca" + 0.002*"social" + 0.002*"conhecimento" + 0.002*"forma" + 0.002*"processo" + 0.002*"sociais" + 0.002*"dados" + 0.002*"memória";
Tópico 23: 0.001*"marvel" + 0.001*"quadrinhos" + 0.000*"cinema" + 0.000*"filme" + 0.000*"filmes" + 0.000*"super-heróis" + 0.000*"studios" + 0.000*"marvel_studios" + 0.000*"narrativa" + 0.000*"histórias_quadrinhos";
Tópico 24: 0.002*"informação" + 0.002*"documentos" + 0.002*"biblioteca" + 0.001*"gestão" + 0.001*"cultura" + 0.001*"pesquisa" + 0.001*"arquivos" + 0.001*"modelo" + 0.001*"records" + 0.001*"museu";
Tópico 25: 0.004*"informação" + 0.002*"digital" + 0.002*"pesquisa" + 0.001*"inovação" + 0.001*"inclusão" + 0.001*"conhecimento" + 0.001*"social" + 0.001*"inclusão_digital" + 0.001*"autores" + 0.001*"dados";
Tópico 26: 0.002*"periódicos" + 0.001*"informação" + 0.001*"docentes" + 0.001*"científica" + 0.001*"produção" + 0.001*"artigos" + 0.001*"pesquisa" + 0.001*"documentos" + 0.001*"dados" + 0.001*"científicos";
Tópico 27: 0.004*"conhecimento" + 0.001*"universidades" + 0.001*"patentes" + 0.001*"documentos" + 0.001*"pesquisa" + 0.001*"patente" + 0.001*"explícito" + 0.001*"domínio" + 0.001*"depósitos" + 0.001*"implícito";
Tópico 28: 0.000*"raras" + 0.000*"santa_catarina" + 0.000*"catarina" + 0.000*"santa" + 0.000*"obras_raras" + 0.000*"obras" + 0.000*"critérios" + 0.000*"raridade" + 0.000*"raro" + 0.000*"acervo";
Tópico 29: 0.001*"mental" + 0.001*"caps" + 0.000*"dsc" + 0.000*"atenção" + 0.000*"psicossocial" + 0.000*"trecho" + 0.000*"atenção_psicossocial" + 0.000*"rede_biblioteca" + 0.000*"rede" + 0.000*"trecho_dsc".

Fonte: Elaborado pelo autor.

Os tópicos extraídos do *corpus* de dados a partir do modelo LDA apresentam termos com características generalistas e especialistas junto ao domínio da linguagem. Os termos generalistas requerem do profissional especialista um maior esforço cognitivo para realizar a interpretação dos dados e posteriormente criar a suposição dos nomes dos tópicos. São exemplos de termos generalistas “produção” e “conhecimento” contidos no tópico 2 que podem se compor para outros termos pertencentes ao domínio de linguagem como “processo_produção”, “produção_documentos”, “campo_conhecimento” ou “compartilhamento_conhecimento”. O especialista poderá explorar documentos externos como listas de bigramas ou trigramas além de acesso ao próprio *corpus* de dados como recurso para buscar realizar uma interpretação

dos dados mais assertiva. Além disso, poderá criar uma categoria geral para classificar tópicos generalistas ou mesmo fazer o descarte dos tópicos.

Os termos especialistas como “histórias_quadrinhos”, “revista_estudos_feministas” e “observatório” permite ao profissional especialista do domínio de linguagem, realizar a interpretação dos dados com um menor esforço cognitivo ou mesmo norteá-lo para buscas de informações em documentos externos como o corpus de dados, para, a partir daí, criar suposições de nomes ou assuntos dos tópicos de maneira mais assertiva. Uma característica fraca dos termos está no equilíbrio dos pesos, em muitos casos com valores iguais, o que dificulta por exemplo a identificação de relevância dos termos perante ao tópico.

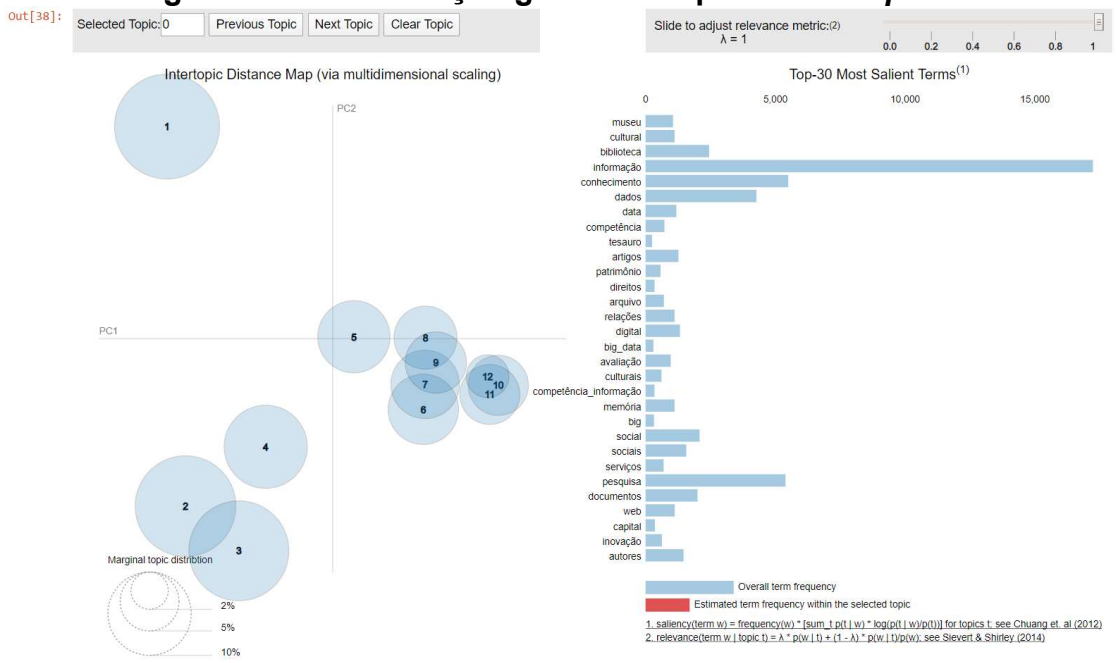
Também é possível encontrar entre os resultados, tópicos que possuem termos com mais de uma área de conhecimento, como por exemplo, o tópico 4 que contempla os termos “patrimônio”, “cultural”, “observatório”, “imaterial” e “unesco”, podendo supor por exemplo pelo especlista que o tópico aborde assuntos relacionados a mediação, patrimônio e informação ou mediação, circulação e apropriação da informação.

O tópico 17 apresenta um conjunto de termos específicos como “mulheres”, “estudos_feministas”, “periódicos” e “revista_estudos_feministas” que permite ao especialista do domínio de linguagem supor por exemplo através da análise de assunto que o tópico esteja relacionado a área ou tema de Produção e Comunicação da Informação. O tópico 21 também apresenta um conjunto de termos especialistas como “*softwares*”, “*synaptica*” e “tesauro” que possibilita ao profissional especialista realizar a análise de assunto a partir de documentos externos como o *corpus* de dados, identificando assim, o documento que melhor representa o conjunto de termos bem como outras informações como título, resumo ou grupo de trabalho. Através dessas informações, o especialista pode inferir por exemplo que o tópico esteja relacionado a área ou tema de Organização e Representação do Conhecimento.

Uma alternativa para identificação da suposição dos nomes tópicos, realizada por especialistas com conhecimento ao domínio da linguagem, está na visualização dinâmica dos tópicos construída a partir dos resultados do modelo LDA e da biblioteca pyLDAvis. A Figura 32 ilustra a visão geral dos tópicos extraídos do *corpus* de dados que foi processado durante 4 horas 30 minutos e

46 segundos. A esquerda da imagem estão os tópicos no formato de círculos em um no plano bidimensional e possuindo dimensionamento multidimensional para projetar as distâncias intertópicas em duas dimensões. O lado esquerdo da imagem ilustra por meio de gráfico de barras os termos contidos em cada tópico utilizados por especialistas na intepretação dos tópicos. O *download* do arquivo para a visualização dinâmica dos tópicos no formato HTML pode ser realizada através do GitHub[83].

Figura 32 – Visualização geral dos tópicos do *corpus* 12

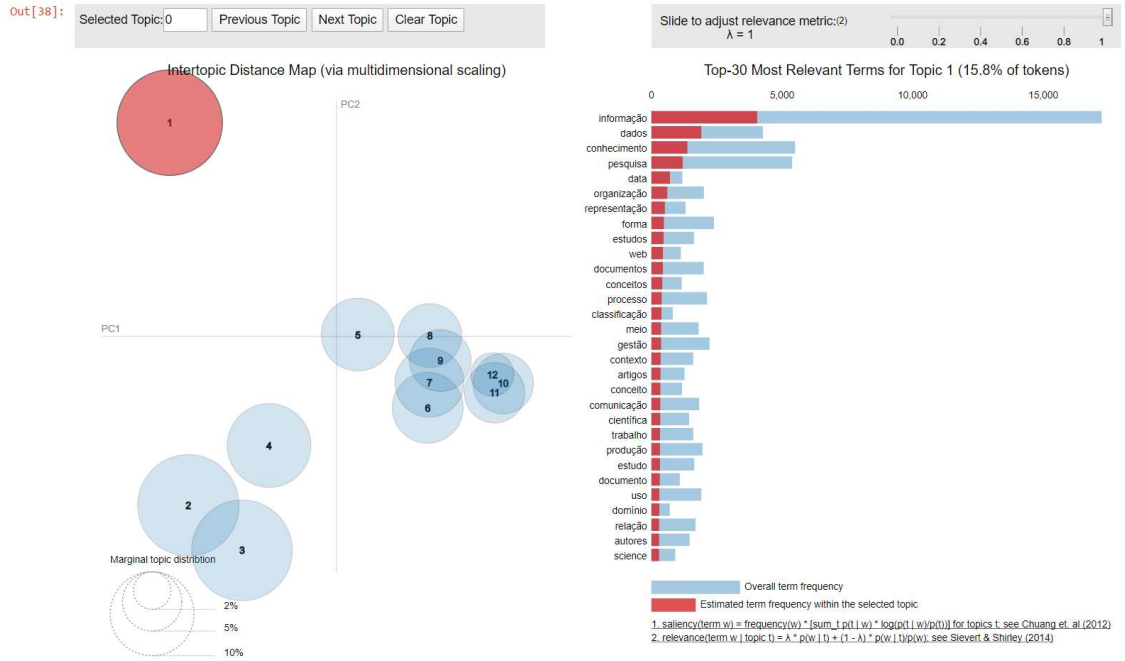


Fonte: Elaborado pelo autor.

Destaca-se a esquerda da Figura 33 o tópico 1 representado pelo círculo vermelho e a direita da imagem um conjunto contendo 30 termos que melhor representam o tópico. Os termos apresentados no gráfico de barras representam 15.8% dos *tokens*, sendo a cor azul considerado a frequência geral e a cor vermelha a cor estimada dos termos.

[83] Visualização dinâmica dos tópicos. *Corpus* 12: artigos completos e resumos expandidos 2017. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2017_gts.html/.

Figura 33 – Visualização do tópico 1 do *corpus* 12



Fonte: Elaborado pelo autor.

Dentre os *tokens* mais relevantes do corpus de dados estão no topo da lista os termos “informação”, “dados”, “conhecimento”, “pesquisa” e “data”. Já na parte baixa da lista estão os termos “uso”, “domínio”, “relação”, “autores” e “science”, ambos os termos são exibidos com ajuste de relevância de métrica com valor de 1.0. Os termos no topo da lista possuem características generalistas que podem ser encontrados em diversas áreas da Ciência da Informação enquanto os termos da parte de baixo da lista apresentam característica mais específica, entretanto, ainda necessita de uma análise mais aprofundada por parte do especialista para definição da suposição do nome do tópico.

Ao reduzir o ajuste de relevância de métrica para 0.2 é possível encontrar termos como “aboutness”, “data”, “dados”, “customer_relationship”, “relationship_management”, “customer_relationship_management”, “orcid”, “herbários”, “dga (Dados Governamentais Abertos)” e “linked”, podendo o especialista assim realizar a análise de assunto por meio dos termos mais específicos e estimar a suposição do tópico. Entretanto, para essa métrica ainda são apresentados resultados que abrangem mais de uma área de estudos da Ciência da Informação como por exemplo Organização e Representação do Conhecimento, Política e Economia da Informação, Gestão da Informação e do Conhecimento e Informação e Tecnologia. Para esse caso, a utilização do valor

da métrica ajustado para 0.0 reduz a quantidade de áreas, porém, os termos do tópico continuam abrangendo mais de uma área da Ciência da Informação.

Outro comportamento apresentado entre os resultados está na intercessão de termos correlacionados entre os tópicos 2 e 3, tais como “informação”, “pesquisa”, “conhecimento”, “gestão” e “dados” quando ajustado as métricas de relevância para 1.0, porém, ao ajustar as métricas para 0.2 os tópicos apresentam características específicas, facilitando assim ao especialista realizar a interpretação dos tópicos por meio de análise de assuntos. O tópico 2, por exemplo, apresenta termos como “almanaques”, “observatórios”, “vulnerabilidades”, “rid (Recursos de Informação Digitais)” e “observatório”, porém, com um número de frequência é inferior quando se comparado a frequência dos termos “sigaa”, “gestão_documentos”, “departamento_cultura” e “incaper” do tópico 3. Nesses dois conjuntos de termos é possível identificar assuntos distintos da Ciência da Informação.

APÊNDICE L - *Corpus* 13: artigos completos e resumos expandido 2018

O último *corpus* que também constitui o segundo *corpora* de dados contempla 387 documentos do tipo artigos completos e resumos expandidos publicados nos anais do ENANCIB no ano de 2018. O corpus possui o tamanho de 17.516kb e a quantidade de 1.174.611 unigramas, 1.174.167 bigramas e 1.173.723. O Quadro 53 apresenta a lista dos 50 termos mais frequentes organizados por tipos de N-gramas.

Quadro 53 – Lista de N-gramas por ordem de frequência do *corpus* 13

Unigramas
informação,27146; pesquisa,8355; conhecimento,6720; dados,5824; biblioteca,3505; forma,3488; social,3349; gestão,3197; documentos,3157; processo,3151; uso,2784; brasil,2766; produção,2750; meio,2730; organização,2713; trabalho,2647; comunicação,2644; desenvolvimento,2634; estudo,2542; memória,2524; museu,2447; contexto,2446; universidade,2433; relação,2429; sociais,2396; tecnologia,2270; termo,2209; sociedade,2160; científica,2139; estudos,2091; federal,2083; informacional,1924; autores,1873; campo,1859; paulo,1855; usuários,1793; cultura,1782; resultados,1781; sistemas,1780; fonte,1760; digital,1736; sistema,1712; instituições,1706; processos,1698; construção,1682; busca,1676; cultural,1663; nacional,1655; diferentes,1618; artigos,1614.
Bigramas
informação_tecnologia,1153; universidade_federal,1144; informação_science,766; competência_informação,703; produção_científica,589; informação_conhecimento,571; gestão_conhecimento,569; recuperação_informação,559; gestão_informação,549; organização_conhecimento,525; dados_pesquisa,508; belo_horizonte,471; modalidade_apresentação,445; redes_sociais,409; uso_informação,405; patrimônio_cultural,390; sistemas_informação,374; informação_informação,369; dissertação_mestrado,367; pesquisa_informação,365; comunicação_oral,350; informação_comunicação,347; universidade_estadual,333; comunicação_científica,329; tendo_vista,321; regime_informação,310; fontes_informação,308; apresentação_comunicação,305; fonte_elaborado,305; bases_dados,304; oral_resumo,303; minas_gerais,302; coleta_dados,301; fonte_dados,290; ensino_superior,288; gestão_documentos,286; representação_informação,284; arquitetura_informação,283; ponto_vista,279; base_dados,265; santa_catarina,263; portal_periódicos,262; nacional_pesquisa,261; encontro_nacional,256; sociedade_informação,255; tecnologias_informação,253; organização_informação,252; produtos_serviços,242; porto_alegre,242; biblioteconomia_informação,239.
Trigramas
modalidade_apresentação_comunicação,304; apresentação_comunicação_oral,304; comunicação_oral_resumo,302; fonte_dados_pesquisa,242; encontro_nacional_pesquisa,229; nacional_pesquisa_informação,227; portal_periódicos_capes,204; tecnologias_informação_comunicação,171; universidade_federal_paraíba,164; sistemas_organização_conhecimento,162; fonte_elaborado_autores,156; universidade_federal_santa,154; international_organization_standardization,145; gestão_informação_conhecimento,142; federal_santa_catarina,140; modalidade_apresentação_resumo,139; organização_representação_conhecimento,134; dissertação_mestrado_informação,134;

universidade_federal_minas,133;	federal_minas_gerais,132;
informação_tecnologia_was,127;	informação_universidade_federal,118;
informação_belo_horizonte,109;	library_informação_science,101;
instituto_federal_educação,96;	federal_educação_tecnologia,93;
plano_desenvolvimento_institucional,92;	universidade_federal_grande,89;
república_federativa_brasil,76;	informação_science_technology,75;
instituição_ensino_superior,72;	informação_sociedade_estudos,70;
instituições_ensino_superior,70;	universidade_federal_bahia,69;
universidade_federal_pernambuco,69	perspectivas_informação_belo,68;
mediação_circulação_apropriação,68;	linked_open_data,68;
desenvolvimento_competência_informação,68;	universidade_federal_fluminense,67;
resource_description_framework,67;	circulação_apropriação_informação,65;
ensino_pesquisa_extensão,62;	artigos_publicados_periódicos,59;
pesquisa_brasileira_informação,59;	mestrado_informação_universidade,59;
produção_comunicação_informação,58;	american_society_informação,58;
society_informação_science,58; arquivologia_biblioteconomia_museologia,58.	

Fonte: Elaborado pelo autor.

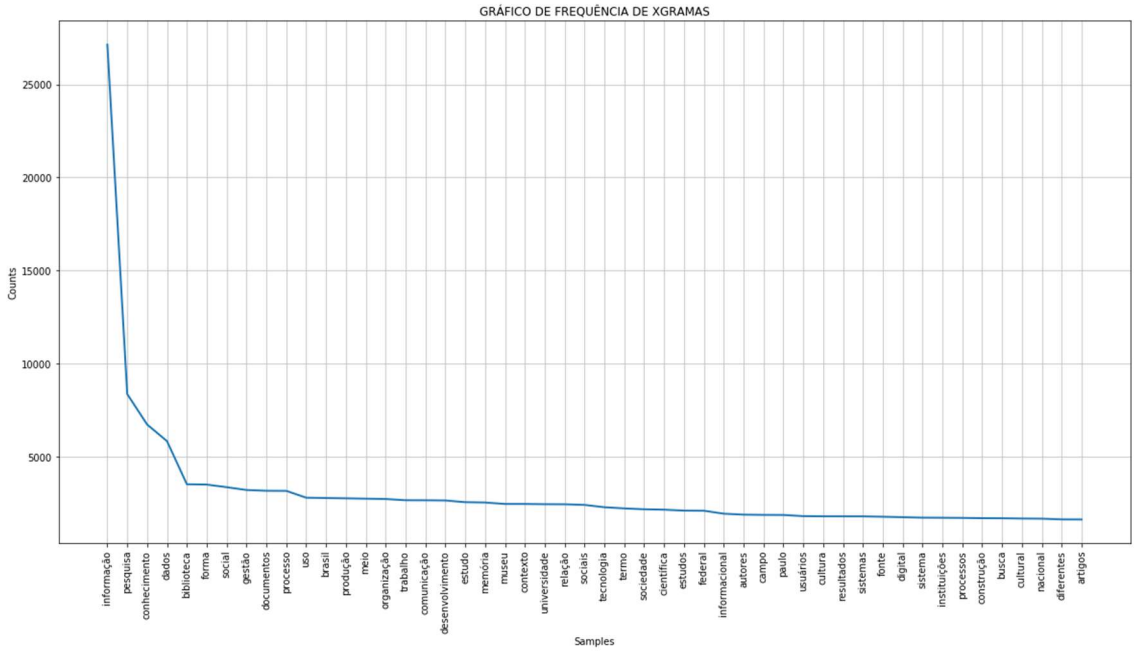
Em uma lista geral contendo mil termos mais frequentes extraídos do *corpus* de dados destaca-se os bigramas “informação_tecnologia” com frequência de 1.153 e ocupando a posição 92 de ranqueamento, seguido de ”universidade_federal” com frequência de 1.144 na posição 96, “informação_science” com frequência de 766 e posição 187, “competência_informação” com frequência de 703 e posição 210 e “produção científica” com frequência de 589 e posição 290. Entre os trigramas mais frequentes estão “modalidade_apresentação_comunicação” com frequência de 304 e posição 661, “apresentação_comunicação_oral” com frequência também de 304 e posição 662, “comunicação_oral_resumo” com frequência de 302 e posição 676, “fonte_dados_pesquisa” com frequência de 242 e posição 903 e “encontro_nacional_pesquisa” com frequência 229 e posição 952. Todos os primeiros termos de cada tipo de N-grama estão entre o milésimo termo da lista de N-gramas.

Os termos do tipo unigramas apresentam maior frequência quando comparado aos demais tipos de N-gramas. Além disso, é possível encontrar um distanciamento entre os primeiros termos de cada tipo. O unigrama “informação” com frequência de 27.146 é 2.254% maior que o bigrama “informação_tecnologia” com frequência de 1.153. Esse distanciamento aumenta para 8.830% ao comparar o unigrama com o trigrama “modalidade_apresentação_comunicação” com frequência de 304.

O Gráfico 63 apresenta os 50 termos mais frequentes extraídos do corpus de dados, sendo todos do tipo unigramas. Entre os termos mais frequentes estão

“informação” com frequência de 27.146, “pesquisa” com frequência de 8.355, “conhecimento” com frequência de 6.720, “dados” com frequência de 5.824 e “biblioteca” com frequência de 3.505.

Gráfico 63 – Termos mais frequentes do *corpus* 13

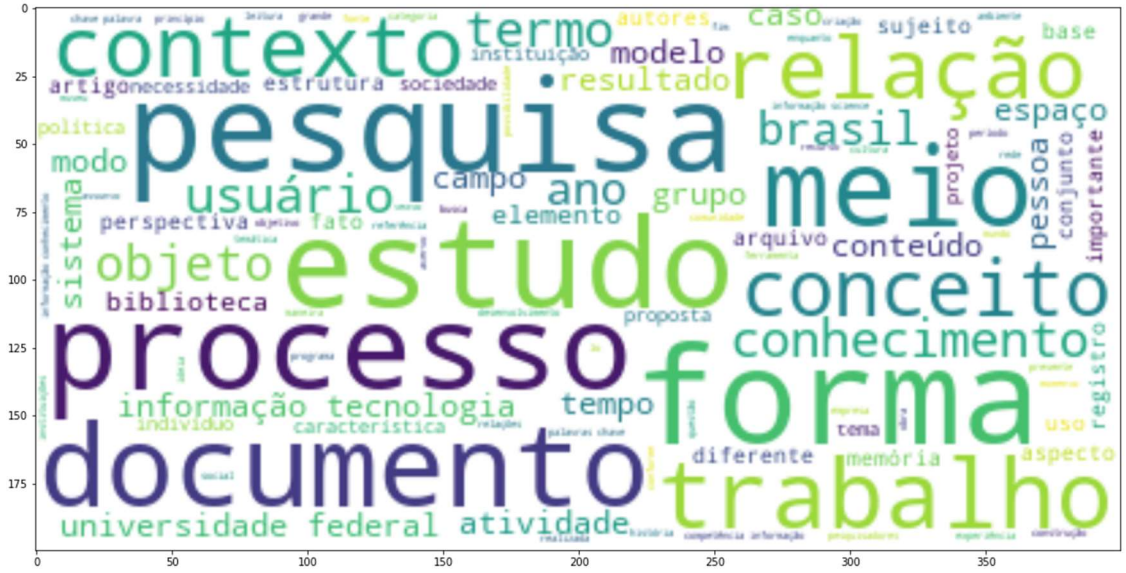


Fonte: Elaborado pelo autor.

Do 50º termo “artigos” com frequência de 1.614 ao 2º termo “pesquisa” com frequência de 8.334 é possível observar o crescimento da frequência de forma constância entre os termos, entretanto, do 2º ao 1º termo “informação” com frequência de 27.146 percebe-se uma variação no volume de frequências, alcançando assim uma diferença de 226% entre os termos. Entre os resultados é possível perceber a existência de termos fortes e generalistas como “organização”, “memória”, “museu”, “tecnologia” e “comunicação” que podem ser aprofundados pelo especialista ao utilizar os bigramas e trigramas, além dos termos fracos como “nacional”, “federal”, “brasil”, “autores” e “meio” que podem ser descartados pelo especialista por não estarem alinhados ao domínio de linguagem.

A Figura 34 apresenta uma nuvem de palavras contendo 250 N-gramas mais frequentes do *corpus* de dados. Para a criação da imagem, não foi utilizado quaisquer tipos de análises qualitativas.

Figura 34 – Nuvem de palavras do *corpus* 13



Fonte: Elaborado pelo autor.

Torna-se possível identificar junto a imagem termos fortes e fracos quando se comparado ao domínio de linguagem e realizado a análise de assunto. São exemplo de termos fortes “pesquisa”, “biblioteca” e “informação_tecnológica”, que apresentam significado junto ao domínio de linguagem, seja analisando os termos individualmente ou através de listas de bigramas e trigramas. Exemplos de termos fracos estão “importante”, “diferente” e “universidade_federal” que não apresentam significados quando associados ao domínio de linguagem.

Os dados conectados aos modelos de extração passam por treinamentos de aprendizagem que resultam em um conjunto de tópicos, termos e pesos. O Quadro 54 apresenta 30 desses conjuntos extraídos do corpus de dados utilizando o modelo LSI, que por sua vez, foi realizado em 43.7 segundos. Os resultados dos conjuntos de 10, 14, 18, 22, 26, 34, 38 e 42 tópicos podem ser acessados através do GitHub[84].

Quadro 54 – Tópicos extraídos do *corpus* 13 usando o modelo LSI

Tópico 0: 0.787*"informação" + 0.186*"pesquisa" + 0.158*"conhecimento" + 0.127*"dados" + 0.078*"forma" + 0.078*"gestão" + 0.075*"social" + 0.073*"processo" + 0.072*"organização" + 0.071*"biblioteca";
Tópico 1: 0.502*"informação" + -0.470*"dados" + -0.308*"pesquisa" + -0.168*"conhecimento" + -0.159*"museu" + -0.120*"biblioteca" + -0.104*"data" + -0.092*"brasil" + -0.090*"documentos" + -0.081*"científica";

[84] Algoritmo de modelagem de tópicos. *Corpus* 13: artigos completos e resumos expandidos 2018. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2018.ipynb/.

Tópico 2: -0.613*"dados" + 0.254*"biblioteca" + 0.236*"conhecimento" + 0.211*"museu" + -0.173*"informação" + 0.133*"memória" + -0.132*"data" + 0.104*"documentos" + 0.097*"educação" + 0.094*"cultura";
Tópico 3: -0.675*"conhecimento" + 0.239*"museu" + 0.228*"biblioteca" + -0.206*"gestão" + -0.169*"organização" + -0.141*"inovação" + 0.123*"memória" + -0.117*"gestão_conhecimento" + 0.094*"cultural" + 0.091*"brasil";
Tópico 4: -0.502*"biblioteca" + 0.328*"documentos" + 0.290*"museu" + 0.150*"arquivos" + -0.144*"educação" + -0.139*"pesquisa" + 0.139*"arquivo" + 0.117*"preservação" + 0.114*"memória" + 0.109*"documento";
Tópico 5: 0.409*"documentos" + -0.346*"museu" + 0.286*"biblioteca" + 0.209*"gestão" + -0.166*"pesquisa" + 0.142*"arquivos" + -0.127*"museologia" + 0.124*"federal" + 0.121*"documental" + -0.119*"científica";
Tópico 6: -0.268*"periódicos" + -0.262*"pesquisa" + 0.214*"conhecimento" + -0.209*"documentos" + 0.207*"dados" + -0.196*"científica" + -0.182*"artigos" + -0.165*"capes" + -0.160*"produção" + 0.153*"museu";
Tópico 7: -0.509*"museu" + 0.265*"memória" + 0.226*"social" + -0.195*"gestão" + -0.169*"museologia" + 0.167*"leitura" + 0.139*"sociais" + -0.110*"federal" + -0.099*"educação" + -0.095*"informação";
Tópico 8: -0.211*"digital" + 0.193*"gestão" + -0.163*"metadados" + -0.162*"biblioteca" + 0.160*"dados" + -0.159*"indexação" + -0.157*"termo" + -0.150*"cloud" + -0.150*"preservação" + -0.139*"representação";
Tópico 9: -0.330*"preservação" + -0.274*"cloud" + -0.258*"digital" + -0.199*"services" + -0.186*"cloud_services" + -0.174*"metadados" + 0.167*"organização" + 0.152*"termo" + -0.147*"inovação" + -0.145*"pesquisa";
Tópico 10: 0.345*"portal" + 0.302*"periódicos" + -0.262*"pesquisa" + 0.236*"capes" + 0.183*"gestão" + 0.174*"usuários" + 0.153*"portal_periódicos" + -0.121*"competência" + 0.119*"uso" + 0.114*"periódicos_capes";
Tópico 11: -0.247*"competência" + 0.218*"memória" + 0.195*"federal" + 0.182*"conhecimento" + -0.165*"leitura" + -0.158*"usuários" + 0.152*"instituto" + -0.151*"pesquisa" + -0.149*"informacional" + 0.144*"preservação";
Tópico 12: -0.342*"competência" + 0.227*"pesquisa" + -0.210*"informacional" + 0.203*"leitura" + 0.177*"gestão" + -0.157*"competência_informação" + 0.157*"biblioteca" + 0.153*"inovação" + -0.153*"portal" + -0.134*"periódicos";
Tópico 13: 0.292*"leitura" + -0.230*"memória" + 0.182*"brasil" + 0.167*"biblioteca" + 0.148*"conhecimento" + -0.146*"instituto" + -0.140*"instituto_federal" + 0.129*"lei" + -0.128*"pesquisa" + 0.128*"cloud";
Tópico 14: -0.264*"memória" + 0.204*"usuários" + 0.169*"inovação" + -0.165*"portal" + 0.150*"social" + 0.148*"sociais" + 0.137*"web" + -0.135*"pesquisa" + -0.130*"leitura" + -0.121*"cultural";
Tópico 15: -0.289*"pesquisa" + -0.212*"memória" + -0.212*"usuários" + -0.184*"verdade" + 0.171*"leitura" + -0.136*"comissão" + 0.125*"dga" + -0.122*"portal" + 0.122*"modelo" + -0.120*"repositórios";
Tópico 16: 0.232*"verdade" + 0.228*"patrimônio" + -0.204*"museu" + 0.189*"cultural" + 0.172*"comissão" + 0.161*"edm" + 0.151*"arquivos" + 0.143*"inovação" + -0.137*"documentos" + -0.134*"dga";
Tópico 17: 0.234*"leitura" + 0.197*"inovação" + 0.176*"memória" + -0.174*"biblioteca" + 0.157*"indexação" + 0.157*"dga" + -0.145*"social" + -0.144*"sociais" + -0.140*"repositórios" + -0.128*"científica";
Tópico 18: 0.236*"memória" + 0.172*"biblioteconomia" + -0.155*"pesquisa" + 0.154*"universidade" + 0.138*"brasil" + -0.137*"competência" + -0.136*"portal" + 0.136*"curso" + -0.136*"leitura" + 0.127*"museologia";

Tópico 19: -0.229*"termo" + 0.200*"leitura" + -0.191*"competência" + 0.189*"verdade" + -0.185*"memória" + -0.180*"biblioteca" + -0.155*"pessoais" + 0.129*"comissão" + -0.127*"inovação" + -0.124*"artigos";
Tópico 20: 0.232*"verdade" + -0.198*"pesquisa" + 0.159*"leitura" + 0.152*"comissão" + 0.150*"museu" + 0.139*"arquivos" + -0.130*"classificação" + 0.128*"competência" + 0.126*"artigos" + 0.126*"repositórios";
Tópico 21: -0.322*"repositórios" + -0.203*"digitais" + -0.191*"leitura" + -0.178*"repositório" + -0.174*"inovação" + 0.165*"web" + -0.134*"indexação" + 0.124*"citação" + 0.123*"memória" + -0.121*"digital";
Tópico 22: 0.312*"inovação" + -0.237*"gestão" + -0.183*"indexação" + -0.154*"imagens" + -0.132*"patrimônio" + 0.131*"biblioteca" + 0.127*"dga" + -0.120*"leitura" + -0.119*"gestão_conhecimento" + 0.114*"portal";
Tópico 23: -0.288*"inovação" + 0.215*"dga" + 0.201*"verdade" + 0.144*"comissão" + -0.140*"leitura" + 0.138*"trabalho" + -0.131*"memória" + 0.130*"gestão" + 0.129*"patrimônio" + 0.121*"organização";
Tópico 24: -0.244*"dga" + -0.217*"imagens" + -0.190*"indexação" + 0.187*"trabalho" + 0.138*"arquitetura" + -0.136*"arquivos" + 0.134*"leitura" + -0.126*"usuários" + 0.123*"sistemas" + 0.123*"arquitetura_informação";
Tópico 25: -0.164*"arquitetura" + 0.155*"modelo" + 0.152*"competência" + -0.136*"usuários" + -0.135*"arquitetura_informação" + 0.131*"organização" + 0.129*"social" + 0.120*"competência_informação" + -0.114*"patrimônio" + -0.111*"conhecimento";
Tópico 26: -0.203*"imagens" + -0.170*"indexação" + 0.170*"web" + -0.168*"verdade" + -0.134*"social" + -0.133*"imagem" + 0.129*"dga" + 0.127*"informacional" + -0.124*"sociais" + 0.122*"usuários";
Tópico 27: -0.186*"termo" + -0.174*"dga" + -0.174*"arquitetura" + -0.167*"portal" + 0.142*"produção" + -0.142*"arquitetura_informação" + -0.141*"social" + -0.134*"autores" + 0.129*"processo" + -0.123*"sociais";
Tópico 28: -0.236*"termo" + 0.186*"livro" + 0.184*"raridade" + 0.142*"livros" + -0.138*"arquivos" + -0.136*"arquitetura" + 0.128*"raro" + 0.123*"livro_raro" + -0.119*"leitura" + 0.116*"citação";
Tópico 29: -0.253*"citação" + 0.195*"termo" + 0.188*"arquitetura" + -0.174*"valor" + -0.168*"organização" + -0.166*"autores" + 0.155*"arquitetura_informação" + 0.155*"livro" + 0.145*"raridade" + -0.138*"patente".

Fonte: Elaborado pelo autor.

Os resultados extraídos através do modelo LSI apresentam termos fortes quando se comparado aos demais termos dos tópicos. Exemplos são os termos “informação” nos tópicos 0 e 1, “dados” no tópico 2, “conhecimento” no tópico 3 e “biblioteca” no tópico 4 que apresentam pesos acima de 0.500. Os termos são ordenados por ordem de relevância de acordo com os seus respectivos pesos. É possível encontrar entre os resultados termos considerados generalistas como “social” que podem se compor para outros termos com significados diferentes como “responsabilidade_social”, “rede_social” ou “inclusão_social”, cabendo assim ao especialista analisar os demais termos do tópico. Caso haja somente termos generalistas no tópico, o especialista poderá explorar documentos externos como a listas de bigramas e trigramas geradas através algoritmo ou mesmo acessar o *corpus* para realizar uma análise de assunto mais assertiva.

Além disso, o especialista poderá criar tópico geral para realizar a classificação ou mesmo descartar o tópico com essa característica.

Os tópicos 8 e 9 apresentam termos com características específicas e que se correlacionam, tais como "*cloud*", "metadados", "preservação" e "digital". Esses termos atrelado aos demais contidos nos tópicos possibilitam ao especialista realizar através da análise de assunto a suposição do nome do tópico, como por exemplo Informação e Tecnologia. Outro exemplo dessa especificidade pode ser encontrado nos tópicos 23, 24, 26 e 27 ao apresentarem o termo "dga (Dados Governamentais Abertos)" que remete a área de estudos relacionada Política, Economia e Informação. Para esse caso, faz-se necessário observar os demais termos, pois o modelo pode misturar assuntos não correlacionados mediante o tamanho do corpus e a quantidade de tópicos configurados para extração.

O Quadro 55 apresenta os resultados alcançados por meio do modelo de extração LDA que foi treinado durante 9 minutos e 50 segundos. O modelo extraiu um conjunto de 30 tópicos constituídos por termos e pesos. Os resultados dos conjuntos de 10, 14, 18, 22, 26, 34, 38 e 42 tópicos podem ser acessados através do GitHub[85].

Quadro 55 – Tópicos extraídos do *corpus* 13 usando o modelo LDA

Tópico 0: 0.002*"pesquisa" + 0.001*"palavra" + 0.001*"informe" + 0.001*"sertão" + 0.001*"integridade" + 0.001*"integridade_pesquisa" + 0.001*"conduta" + 0.001*"rosa" + 0.001*"iumforme" + 0.000*"científica";
Tópico 1: 0.001*"informação" + 0.001*"valor" + 0.001*"pesquisa" + 0.001*"patentes" + 0.001*"patente" + 0.001*"dados" + 0.001*"inovação" + 0.001*"memória" + 0.001*"patrimônio" + 0.001*"redes";
Tópico 2: 0.002*"biblioteca" + 0.002*"pesquisa" + 0.002*"dados" + 0.002*"federal" + 0.001*"artigos" + 0.001*"instituto" + 0.001*"desenvolvimento" + 0.001*"educação" + 0.001*"informação" + 0.001*"instituto_federal";
Tópico 3: 0.001*"periódicos" + 0.001*"zona" + 0.001*"bradford" + 0.001*"artigos" + 0.001*"bahia" + 0.001*"consumo" + 0.001*"documentos" + 0.001*"igc" + 0.001*"docentes" + 0.000*"produção";
Tópico 4: 0.004*"informação" + 0.002*"museu" + 0.002*"portal" + 0.001*"pesquisa" + 0.001*"periódicos" + 0.001*"capes" + 0.001*"patrimônio" + 0.001*"biblioteca" + 0.001*"trabalho" + 0.001*"portal_periódicos";
Tópico 5: 0.003*"museu" + 0.001*"objetos" + 0.001*"cultural" + 0.001*"patrimônio" + 0.001*"pesquisa" + 0.001*"valor" + 0.001*"informação" + 0.001*"social" + 0.001*"valores" + 0.001*"imagem";

[85] Algoritmo de modelagem de tópicos. *Corpus* 13: artigos completos e resumos expandidos 2018. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/03_lda_lsi_artigosresumos_2018.ipynb/.

Tópico 6: 0.008*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.002*"biblioteca" + 0.002*"social" + 0.001*"produção" + 0.001*"científica" + 0.001*"processo" + 0.001*"serviços" + 0.001*"informacional";
Tópico 7: 0.002*"dados" + 0.002*"pesquisa" + 0.001*"museu" + 0.001*"conhecimento" + 0.001*"documentos" + 0.001*"identificação" + 0.001*"gestão" + 0.001*"grafo" + 0.001*"prontuário" + 0.001*"fonte";
Tópico 8: 0.001*"bourdieu" + 0.001*"palavras-chave" + 0.001*"livros" + 0.001*"artigos" + 0.001*"biblioteca" + 0.001*"campo" + 0.000*"comunidade" + 0.000*"pesquisa" + 0.000*"secretariado" + 0.000*"educação";
Tópico 9: 0.003*"informação" + 0.002*"conhecimento" + 0.001*"pesquisa" + 0.001*"gestão" + 0.001*"brasil" + 0.001*"bibframe" + 0.001*"processo" + 0.001*"modelo" + 0.001*"estatísticas" + 0.001*"biblioteca";
Tópico 10: 0.010*"informação" + 0.002*"pesquisa" + 0.002*"biblioteca" + 0.001*"processo" + 0.001*"forma" + 0.001*"dados" + 0.001*"busca" + 0.001*"uso" + 0.001*"termo" + 0.001*"contexto";
Tópico 11: 0.003*"documentos" + 0.001*"documental" + 0.001*"pesquisa" + 0.001*"museu" + 0.001*"arquivo" + 0.001*"identificação" + 0.001*"arquivística" + 0.001*"informação" + 0.001*"museologia" + 0.001*"gestão";
Tópico 12: 0.002*"museu" + 0.002*"dados" + 0.002*"biblioteca" + 0.001*"pesquisa" + 0.001*"conhecimento" + 0.001*"fair" + 0.001*"curso" + 0.001*"museologia" + 0.001*"bliss" + 0.001*"blockchain";
Tópico 13: 0.004*"informação" + 0.001*"pesquisa" + 0.001*"gestão" + 0.001*"dados" + 0.001*"documentos" + 0.001*"inovação" + 0.001*"modelo" + 0.001*"descrição" + 0.001*"conhecimento" + 0.001*"edm";
Tópico 14: 0.002*"informação" + 0.002*"pesquisa" + 0.001*"dados" + 0.001*"cultural" + 0.001*"sociais" + 0.001*"mediação" + 0.001*"brasil" + 0.001*"social" + 0.001*"conhecimento" + 0.001*"artigos";
Tópico 15: 0.005*"dados" + 0.002*"dga" + 0.001*"dados_pessoais" + 0.001*"pessoais" + 0.001*"memória" + 0.001*"repositórios" + 0.001*"proteção" + 0.001*"preservação" + 0.001*"termo" + 0.001*"lei";
Tópico 16: 0.001*"conceitos" + 0.001*"conceitual" + 0.001*"mapas" + 0.001*"conceituais" + 0.001*"mapas_conceituais" + 0.000*"mapa" + 0.000*"website" + 0.000*"mapeamento_conceitual" + 0.000*"mapeamento" + 0.000*"mapa_conceitual";
Tópico 17: 0.001*"memória" + 0.001*"arquivo" + 0.001*"arquivos" + 0.001*"cidade" + 0.001*"privacidade" + 0.001*"municipal" + 0.001*"arte" + 0.001*"privacy" + 0.001*"arquivísticas" + 0.001*"portal";
Tópico 18: 0.002*"conhecimento" + 0.002*"museu" + 0.001*"biblioteca" + 0.001*"gestão" + 0.001*"inovação" + 0.001*"pesquisa" + 0.001*"ferroviário" + 0.001*"políticas" + 0.001*"uso" + 0.001*"ferramentas";
Tópico 19: 0.018*"informação" + 0.004*"pesquisa" + 0.004*"conhecimento" + 0.003*"dados" + 0.002*"forma" + 0.002*"organização" + 0.002*"social" + 0.002*"documentos" + 0.002*"processo" + 0.002*"comunicação";
Tópico 20: 0.001*"periódicos" + 0.001*"metadados" + 0.001*"publicação" + 0.001*"pesquisa" + 0.001*"ferreira" + 0.001*"fachin" + 0.001*"campo" + 0.001*"estivals" + 0.001*"avaliadores" + 0.001*"avaliação";
Tópico 21: 0.002*"informação" + 0.001*"web" + 0.001*"arquivos" + 0.001*"pesquisa" + 0.001*"recursos" + 0.001*"avaliação" + 0.001*"biblioteca" + 0.001*"dados" + 0.001*"pessoais" + 0.001*"modelo";
Tópico 22: 0.002*"cloud" + 0.002*"preservação" + 0.001*"lei" + 0.001*"informação" + 0.001*"services" + 0.001*"cloud_services" + 0.001*"documentos" + 0.001*"digital" + 0.001*"metadados" + 0.001*"cultura";

Tópico 23: 0.002*"conhecimento" + 0.001*"científica" + 0.001*"capes" + 0.001*"comunicação" + 0.001*"gestão" + 0.001*"pesquisa" + 0.001*"divulgação" + 0.001*"relacionamentos" + 0.001*"arquivos" + 0.001*"museologia";
Tópico 24: 0.003*"informação" + 0.002*"biblioteca" + 0.002*"pesquisa" + 0.001*"comunicação" + 0.001*"memória" + 0.001*"científica" + 0.001*"usuários" + 0.001*"dados" + 0.001*"social" + 0.001*"desenvolvimento";
Tópico 25: 0.001*"imagens" + 0.001*"imagem" + 0.000*"instagram" + 0.000*"uso" + 0.000*"fotográficas" + 0.000*"imagens_fotográficas" + 0.000*"uso_imagens" + 0.000*"reputação" + 0.000*"sociais" + 0.000*"marcação";
Tópico 26: 0.001*"museu" + 0.001*"brasil" + 0.001*"francisca" + 0.001*"francisca_arruda" + 0.001*"arruda" + 0.001*"professora" + 0.001*"ramalho" + 0.001*"arruda_ramalho" + 0.001*"francisca_arruda_ramalho" + 0.001*"pesquisadores";
Tópico 27: 0.002*"verdade" + 0.002*"comissão" + 0.001*"comissões" + 0.001*"comissões_verdade" + 0.001*"estadual" + 0.001*"arquivos" + 0.001*"comissão_verdade" + 0.001*"relatório" + 0.001*"comissão_estadual" + 0.001*"final";
Tópico 28: 0.001*"documentos" + 0.001*"leitura" + 0.001*"ordenação" + 0.001*"arquivos" + 0.001*"classificação" + 0.001*"ordenação_documentos" + 0.001*"complexidade" + 0.001*"social" + 0.001*"econômica" + 0.001*"paulo";
Tópico 29: 0.011*"informação" + 0.003*"conhecimento" + 0.002*"pesquisa" + 0.001*"informacional" + 0.001*"biblioteconomia" + 0.001*"organização" + 0.001*"arquivologia" + 0.001*"relação" + 0.001*"gestão" + 0.001*"uso".

Fonte: Elaborado pelo autor.

É possível observar uma diferença nas características entre os termos extraídos do *corpus* 7 ao *corpus* 13, ambos pertencentes ao segundo *corpora* de dados que contemplam entre seus documentos artigos completos e resumos expandidos do ENANCIB. Entre os primeiros *corpus* surgiram um maior número maior de N-gramas do tipo bigrama e trigrama que foi reduzindo esse quantitativo com o passar das análises dos *corpus*. Dentre os resultados alcançados por meio do modelo LDA junto ao *corpus* 13 é possível identificar termos generalistas como “memória” que podem conter composições de significados dentro do domínio de linguagem como “memória_social” ou “memória_coletiva”, exigindo assim do especialista um maior esforço cognitivo para realizar a interpretação dos dados. O especialista pode utilizar de recursos externos como listas de bigramas ou trigramas geradas a partir do algoritmo ou mesmo explorar o *corpus* de dados para realizar uma melhor interpretação dos dados. Para tópicos que contemplam todos os termos generalistas, o especialista poderá por exemplo criar uma categoria geral para alocar os tópicos com essa característica ou mesmo descarta-los.

É possível identificar também a existência de tópicos especialistas como “mapeamento_conceitual” e “*cloud_services*” que quando associado a outros termos do tópico possibilita realizar a identificação do assunto de forma mais

assertiva. Além disso, os termos especialistas podem apresentar menor esforço cognitivo ao especialista durante a interpretação dos tópicos. Nota-se também entre os resultados a existências de termos equilibrados ou com valores iguais a 0, o que dificultar ao especialista identificar a ordem de relevância dos termos.

Entre os resultados extraídos através do modelo LDA é possível identificar um conjunto de tópicos como 2, 3, 4 e 5 que contemplam uma aproximação entre alguns dos termos comuns, tais como “biblioteca”, “conhecimento”, “documentos”, “memória”. Além disso, os tópicos contemplam termos especialistas como “dados”, “gestão”, “gestão_informação”, “arquivos”, “preservação”, “documental” e “científica”, característicos de cada tópico. Esse conjunto de tópicos apresentam termos contendo pesos que determinam a ordem de relevância para o tópico. Através da análise de assunto, o especialista poderá decidir por unificar os tópicos e supor um único tema ou assunto, como por exemplo, Gestão da Informação.

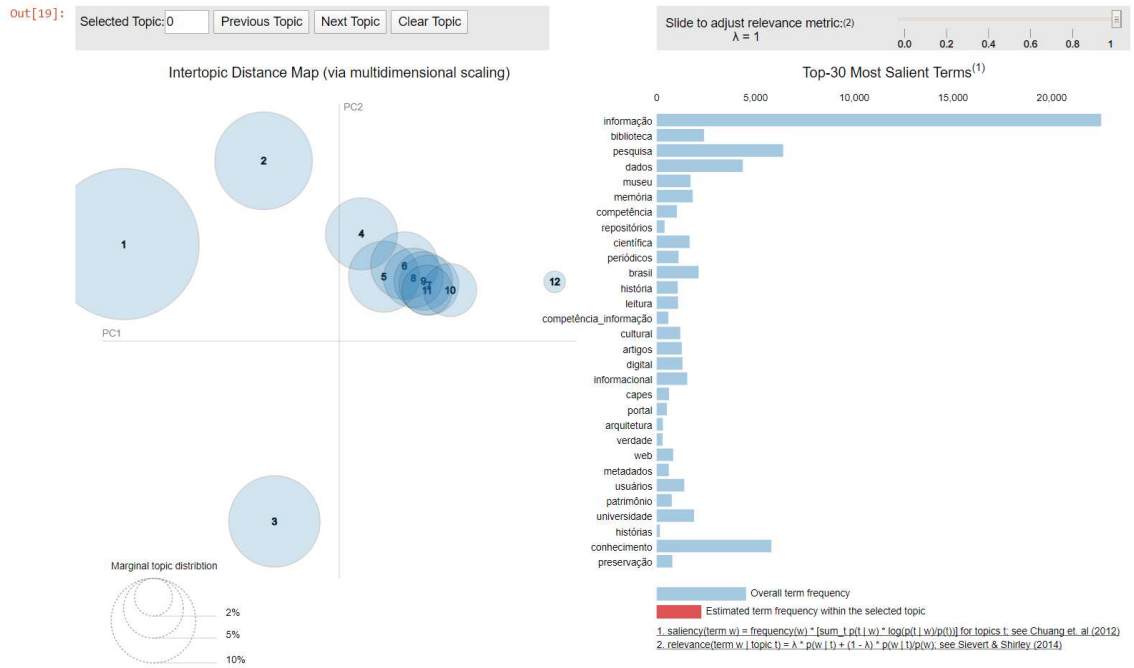
O tópico 10 apresenta um conjunto de termos específicos e pesos de relevância que apresentam ordem de importância, tais como “portal”, “pesquisa”, “capes”, “gestão”, “usuários”, “uso” “portal_periódicos”. Esses temos, quando analisado pelo especialista do domínio da linguagem pode remeter por exemplo a suposição de um rótulo na área de Gestão da Informação e do Conhecimento.

Também é possível identificar entre os resultados termos específicos como “dga (Dados Governamentais Abertos)” encontrados nos tópicos 15, 16, 17, 22, 23, 24 e 27, entretanto, é possível perceber a existência de termos fracos nos tópicos que não estão coesos ao termo específico, além disso, ao explorar os documentos externos como o *corpus* de dados é possível identificar a existência de um único documento que aborde o assunto. Para esse tipo de caso, o especialista poderá analisar termos secundários para identificar o assunto ou mesmo realizar o descarte dos tópicos.

A visualização dinâmica dos tópicos auxilia ao especialista do domínio da linguagem na interpretação e suposição dos nomes dos tópicos extraídos do *corpus* de dados. A figura 35 representa essa visualização criada a partir dos tópicos extraídos utilizando o modelo LDA e da biblioteca pyLDAvis que foi processada durante 5 horas e 5 minutos. A esquerda da imagem estão os tópicos extraídos representados por círculos num plano bidimensional. Seu dimensionamento multidimensional é utilizado para projetar as distâncias

intertópicas em duas dimensões. À direita da imagem estão os termos encontrados em cada tópico no formato de gráfico de barras, utilizados para interpretação dos tópicos. O *download* do arquivo para a visualização dinâmica dos tópicos no formato HTML pode ser realizado através do GitHub[86].

Figura 35 – Visualização geral dos tópicos do *corpus* 13

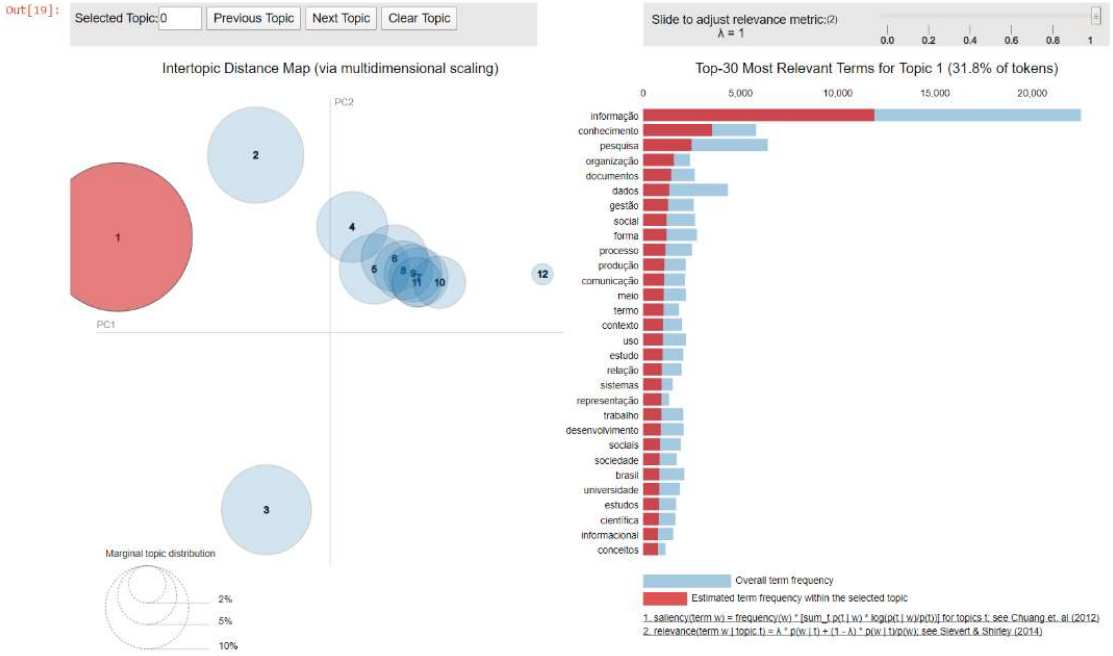


Fonte: Elaborado pelo autor.

A Figura 36 apresenta em seu lado esquerdo o tópico 1 representado pelo círculo vermelho. Já a direita da imagem está um conjunto de 30 termos no formato de gráfico de barras que melhor representam o tópico selecionado. Esses termos representam 31.8% dos *tokens* do corpus de dados, sendo a cor azul a frequência geral e a cor vermelha a frequência estimada do termo no tópico selecionado.

[86] Visualização dinâmica dos tópicos. *Corpus* 13: artigos completos e resumos expandidos 2018. Disponível em: https://github.com/marcosdesouza82/topic-model-tese/blob/master/lda_enancib_2018_gts.html/.

Figura 36 – Visualização do tópico 1 do *corpus* 13



Fonte: Elaborado pelo autor.

O tópico 1 apresenta no topo de sua lista os termos “informação”, “conhecimento”, “pesquisa”, “organização” e “documentos” com ajuste de relevância de métrica no valor de 1.0. Esses termos considerados genéricos estão contidos em 11 dos 12 tópicos gerados, dificultando assim na intepretação dos tópicos por parte do especialista. Ao reduzir o valor da métrica para 0.2 surgem termos mais específicos do tópico como “organização_conhecimento”, “fake_news”, “classificação”, “tesauro”, “ontologias” e “sistemas_organização_conhecimento” no qual pode-se supor pelo especialista que o tópico esteja relacionado a Organização e Representação do Conhecimento. Do 4º ao 10º tópico é possível identificar correlações através das intercessões entre os tópicos, entretanto, somente com o ajuste de métricas é possível explorar os dados e realizar a suposição dos nomes dos tópicos de maneira mais assertiva.

APÊNDICE M – Equivalência de termos – Expressões regulares

Novos termos podem ser inseridos junto ao algoritmo de acordo as necessidades. O conhecimento na linguagem de domínio estuda contribui para um aprimoramento dos resultados.

Quadro 56 – Equivalência de termos por meio de expressões regulares

Nº	Termo	Novo termo
1	museus	museu
2	abc	academia brasileira de ciência
3	ai	arquitetura da informação
4	ala	american library association
5	american society for information science and Technology	arist
6	aoi	arquitetura e organização da informação
7	associação brasileira de educação em ciência da informação	abecin
8	associação brasileira de ensino de biblioteconomia e documentação	abebd
9	associação brasileira de normas técnicas	abnt
10	associação nacional de pesquisa e pós-graduação em ciência da informação	ancib
11	base de dados referencial de artigos de periódicos em ciência da informação	brapci
12	bci	biblioteconomia e ciência da informação
13	biblioteca digital de teses e dissertações	bdtd
14	bn	biblioteca nacional
15	bu	biblioteca universitária
16	c&t	ciência e tecnologia
17	cc	ciência da computação
18	ccn	catálogo coletivo nacional
19	cdd	classificação decimal de dewey
20	cdu	classificação decimal universal

21	cepe	conselho de extensão e pesquisa
22	ci	ciência da informação
23	cms	content management system
24	cne	conselho nacional de educação
25	conselho nacional de desenvolvimento científico e tecnológico	cnpq
26	coordenação de aperfeiçoamento de pessoal de nível superior	capes
27	crm	customer relationship management
28	dc	dublin core
29	descriptive ontology for linguistic and cognitive engineering	dolce
30	ead	educação à distância
31	eci	escola de ciência da informação
32	enade	exame nacional de desempenho
33	encontro nacional de pesquisa em ciência da informação	enancib
34	fgv	fundação getúlio vargas
35	fiocruz	fundação oswaldo cruz
36	ftp	file transfer protocol
37	fundação brasileira à pesquisa do estudo de minas gerais	fapemig
38	fundação brasileira à pesquisa do estudo de são paulo	fapesp
39	fundação brasileira à pesquisa do estudo do rio de janeiro	faperj
40	fundação de amparo às pesquisas	faps
41	gic	gestão da informação e conhecimento
42	gpl	general public licence
43	hypertext markup language limguagem	html
44	ia	inteligência artificial
45	ibpc	instituto brasileiro do patrimônio cultural
46	ibpc	instituto brasileiro do patrimônio cultural
47	ics	informação cultura e sociedade
48	ict	informação, ciência e tecnologia
49	idh	índice de desenvolvimento humano
50	ies	instituição de ensino superior

51	information science and technology abstracts	ista
52	instituto brasileiro de bibliografia e documentação	ibbd
53	instituto brasileiro de geografia e estatística	ibge
54	instituto brasileiro de informação em ciência e tecnologia	ibict
55	instituto de ciência da informação	ici
56	instituto nacional de estudos e pesquisas educacionais anísio teixeira	inep
57	instituto universitário de pesquisa do rio de janeiro	iuoerj
58	international federation of library associations and institutions	ifla
59	international standard book number	isbn
60	international standard serial number	issn
61	isi	institute for scientific information
62	iso	international organization for standardization
63	it	informação e tecnologia
64	jcr	journal citation reports
65	jstor	journal storage
66	kos	sistemas de organização do conhecimento
67	lc	linguagem cinzenta
68	ld	linguagem documentária
69	ldb	lei de diretrizes e bases
70	library and informations science abstracts	lisa
71	ln	linguagem natural
72	lod	linked open data
73	marc	machine readable cataloging
74	mdi	multiple document interface
75	mec	ministério da educação
76	npd	núcleo de pesquisa e documentação
77	oai	open archives iniciative
78	oc	organização do conhecimento
79	ocr	optical character recognition

80	oi	organização da informação
81	ojs	open journal systems
82	ong	organização não-governamental
83	onu	organização das nações unidas
84	organização das nações unidas para a educação e cultura	unesco
85	osi	open society institute
86	oui	organização e uso da informação
87	owl	web ontology language
88	pln	processamento de linguagem natural
89	plos	public library of science
90	poi	produção e organização da informação
91	pontifícia universidade católica de minas gerais	pucmg
92	pontifícia universidade católica de são paulo	pucsp
93	pontifícia universidade católica do rio de janeiro	pucrj
94	pontifícia universidade católica do rio grande do sul	pucrs
95	ppg	programa de pós-graduação
96	programa de pós-graduação em ciência da informação	ppgci
97	programa de pós-graduação em ciências sociais	ppgcs
98	programa de pós-graduação em sociologia	ppgs
99	rc	representação do conhecimento
100	rdf	resource description framework
101	ri	recuperação da informação
102	sad	sistema de apoio à decisão
103	sci	science citation index
104	scientific electronic library online	scielo
105	sdr	zona de desenvolvimento real
106	serviço central de informação bibliográfica	scib
107	serviço nacional de aprendizagem comercial	senac
108	serviço nacional de aprendizagem industrial	senai
109	serviço social da indústria	sesi

110	serviço social do comércio	sesc
111	sibi	sistema integrado de bibliotecas
112	sig	sistema de informação gerencial
113	snad	secretária nacional de políticas anti-drogas
114	snpq	sistema nacional de pós-graduação
115	soc	sistemas de organização do conhecimento
116	sri	sistema de recuperação da informação
117	ssd	sistema de suporte à decisão
118	tcc	trabalho de conclusão de curso
119	universidade federal de minas gerais	ufmg
120	universidade federal de ouro preto	ufop
121	universidade federal de pernambuco	ufpe
122	universidade federal de santa caratina	ufsc
123	universidade federal de são carlos	ufscar
124	universidade federal de sergipe	ufs
125	universidade federal de viçosa	ufv
126	universidade federal do ceará	ufc
127	universidade federal do estado do rio de janeiro	unirio
128	universidade federal do maranhão	ufma
129	universidade federal do pará	ufpa
130	universidade federal do paraná	ufpr
131	universidade federal do rio de janeiro	ufrj
132	universidade federal do rio grande do norte	ufrn
133	universidade federal do rio grande do sul	ufrgs
134	universidade federal fluminense	uff
135	universidade federal rural do rio de janeiro	ufrrj
136	uri	uniform resource identifier
137	world wide web consortium	w3c
138	xml	extensible markup language
139	zpd	zona de desenvolvimento proximal
140	citações	citação
141	links	link

142	museus	museu
143	literacy	alfabetização
144	bibliotecas	biblioteca
145	informações	informação
146	information	informação
147	bibliographia	bibliografia
148	termos	termo
149	knowledge	conhecimento
150	accessibility	acessibilidade
151	research	pesquisa
152	image	imagem
153	archivists	arquivistas

Fonte: Elaborado pelo autor.

APÊNDICE N – Lista adicional de *stop words*

Quadro 57 – Lista adicional de *stop words*

'tais', 'xviii', 'jan.', 'jul.', 'jan', 'and', 'the', 'acho', 'lo', 'pra', 'é', 'onde', 'senão', 'quanto', 'outros', 'sobre', 'sobretudo', 'ser', 'ainda', 'quais', 'desse', 'assim', 'tal', 'através', 'podemos', 'portanto', 'pode', 'tanto', 'alguns', 'possível', 'p.', 'v', 'p', '-se', 'se', 'se', 'nesse', 'nessa', 'neste', 'nesta', 'http', 'https', 'disponível', 'acesso', 'sendo', 'marília', 'rio', 'dessa', 'diz', 'respeito', 'finais', 'considerações', 'desta', 'belo horizointe' ,'janeiro',' fevereiro', 'março', 'abril', 'maio', 'junho', 'julho', 'agosto', 'setembro', 'outubro', 'novembro', 'dezembro', 'porto alegre', 'outro', 'xix', 'londrina', 'paulista', 'xvii', 'sentido', 'maior', 'bem', 'ter', 'deve', 'devem', 'entretanto', 'fazer', 'todo', 'tipo', 'exemplo', 'pois', 'apenas', 'utilizado', 'utilizados', 'acordo', 'casa', 'figura', 'cada', 'tese', 'acesso', 'partir', 'parte','segundo', 'autor' ,'outras', 'podem', 'vez', 'todos', 'então', 'sim', 'todas', 'apresenta', 'algumas', 'outra', 'lo','la', '-lo', '-la', 'porque', 'por que', 'os', 'etc', 'jan', 'fev', 'mar', 'abr', 'mai', 'jun', 'jul', 'ago', 'set', 'out', 'nov', 'dez', 'então', 'aqui', 'enancib'.

Fonte: Elaborado pelo autor.