

**UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO**

AMARILDO MARTINS DE MAGALHÃES

**ORGANIZAÇÃO DA INFORMAÇÃO:
UM MODELO SEMIAUTOMÁTICO DE CLASSIFICAÇÃO DE ATRAÇÕES EM
PERFIS TURÍSTICOS USANDO APRENDIZADO DE MÁQUINA**

Belo Horizonte

2021

AMARILDO MARTINS DE MAGALHÃES

**ORGANIZAÇÃO DA INFORMAÇÃO:
UM MODELO SEMIAUTOMÁTICO DE CLASSIFICAÇÃO DE ATRAÇÕES EM
PERFIS TURÍSTICOS USANDO APRENDIZADO DE MÁQUINA**

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: Dra. Renata Maria Abrantes Baracho

Coorientador: Dr. Thomas Mandl

Belo Horizonte

2021

M188o

Magalhães, Amarildo Martins de.

Organização da informação [recurso eletrônico]: um modelo semiautomático de classificação de atrações em perfis turísticos usando aprendizado de máquina / Amarildo Martins de Magalhães. - 2021.

1 recurso eletrônico (178 f. : il., color): pdf.

Orientadora: Renata Maria Abrantes Baracho.

Coorientador: Thomas Mandl

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

Referências: f. 163-178.

Exigências do sistema: Adobe Acrobat Reader.

1. Ciência da Informação – Teses. 2. Aprendizado do computador – Teses. 3. Turismo – Teses. 4. Organização da informação – Teses. I. Título. II. Baracho, Renata Maria Abrantes. III. Mandl, Thomas Leonhard. IV. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4:338.9



UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO
CONHECIMENTO



FOLHA DE APROVAÇÃO

ORGANIZAÇÃO DA INFORMAÇÃO: UM MODELO SEMIAUTOMÁTICO DE CLASSIFICAÇÃO DE ATRAÇÕES EM PERFIS TURÍSTICOS USANDO APRENDIZADO DE MÁQUINA

AMARILDO MARTINS DE MAGALHÃES

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 09 de março de 2021, pela banca constituída pelos membros:

Prof(a). Renata Maria Abrantes Baracho Porto (Orientadora)
Escola de Arquitetura/UFMG

Prof(a). Thomas Leonhard Mandl (Coorientador)
University of Hildesheim

Prof(a). Lorenzo Cantoni
USI - Suíça

Prof(a). Fernando Silva Parreiras
FUMEC

FERNANDO SILVA PARREIRAS:03073186646
2021.03.09 13:43:26 -03'00'

Prof(a). Renato Rocha Souza
FGV/RJ

Prof(a). Marcos de Souza

Belo Horizonte, 9 de março de 2021.



ATA DA DEFESA DE TESE DO ALUNO AMARILDO MARTINS DE MAGALHÃES

Realizou-se, no dia 09 de março de 2021, às 09:00 horas, todos por videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *ORGANIZAÇÃO DA INFORMAÇÃO: UM MODELO SEMIAUTOMÁTICO DE CLASSIFICAÇÃO DE ATRAÇÕES EM PERFIS TURÍSTICOS USANDO APRENDIZADO DE MÁQUINA*, apresentada por AMARILDO MARTINS DE MAGALHÃES, número de registro 2016712133, 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). Renata Maria Abrantes Baracho Porto - Escola de Arquitetura/UFMG (Orientadora), Prof(a). Thomas Leonhard Mandl - University of Hildesheim (Coorientador), Prof(a). Renato Rocha Souza - FGV/RJ, Prof(a). Fernando Silva Parreiras - FUMEC, Prof(a). Lorenzo Cantoni - USI - Suíça, Prof(a). Marcos de Souza.

A Comissão considerou a tese:

Aprovada

Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Belo Horizonte, 09 de março de 2021.

Prof(a). Renata Maria Abrantes Baracho Porto

Prof(a). Thomas Leonhard Mandl

Prof(a). Lorenzo Cantoni

Prof(a). Fernando Silva Parreiras

FERNANDO SILVA PARREIRAS:03073186646
2021.03.09 13:43:26 -03'00'

Prof(a). Renato Rocha Souza

Prof(a). Marcos de Souza

DEDICATÓRIA

A meus pais (*in memoriam*), aos meus filhos Laura Borges Magalhães, Pedro Borges Magalhães e à minha esposa Luiza Borges Barbosa.

AGRADECIMENTOS

À Deus por me proteger e possibilitar saúde para seguir sempre em frente.

Agradeço do fundo do meu coração aos meus filhos Laura, Pedro e minha esposa Luiza. Aos quais tive que abdicar de estar junto muitas vezes para que pudesse atingir essa conquista. Obrigado. Muito obrigado por todo carinho, paciência e apoio. Essa conquista é nossa, vocês são parte dela, e são minha vida.

À minha orientadora professora Dra. Renata Maria Abrantes Baracho, por toda a compreensão em meio às adversidades, ensinamentos e trabalho em conjunto por mais de oito anos de pesquisa.

Ao meu coorientador professor Dr. Thomas Mandl, por todo o apoio, trabalho e ensinamentos durante a pesquisa.

Aos membros da banca examinadora, que certamente vão contribuir com a melhoria da pesquisa. Obrigado pela disponibilidade e pelo interesse.

Aos colegas do Programa de Pós-Graduação em Gestão & Organização do Conhecimento (PPGGOC), equipe administrativa e docentes, parte essencial nessa caminhada.

Ao Instituto Federal de Minas Gerais (IFMG), o qual financiou parte da pesquisa e a todos os meus colegas de trabalho que me apoiaram na conclusão do trabalho.

A todos que que contribuíram de certa forma, direta ou indiretamente, para que esse objetivo se realizasse.

Muito obrigado!

“A dúvida é o princípio da sabedoria”
(Aristóteles).

RESUMO

O paradigma da evolução tecnológica trouxe uma mudança disruptiva no comportamento das pessoas, que agora tomam decisões baseando-se no conteúdo que consomem na Internet. Esse aspecto não é diferente na indústria do Turismo, em que novas tecnologias e o compartilhamento de avaliações permitem que usuários busquem informações para apoio em decisões como a escolha do destino, da hospedagem, das atrações, da alimentação dentre outras. Essas avaliações fornecem uma fonte importante de informações, no entanto, seu volume pode dificultar a extração de conhecimento e seu uso efetivo. Como descobrir se uma determinada atração com mais de 100 mil opiniões publicadas em texto não-estruturado possui similaridade com o que um turista procura? Essa indagação motiva o desenvolvimento desta pesquisa, que possui como objetivo geral a criação de um modelo que permita transformar as avaliações feitas pelos usuários em classes ou perfis turísticos. Além disso, na literatura observam-se trabalhos de classificação de destinos e atrações turísticas com base nas avaliações. O uso de perfis no turismo é comum como forma de classificar destinos e turistas. Nesse sentido, esse estudo oferece uma visão adicional sobre os dois aspectos, ao passo que permite a junção de perfis turísticos com informações contidas nas avaliações. O trabalho apresenta uma pesquisa aplicada, tendo como base o Pragmatismo, de natureza híbrida com objetivo exploratório. Utiliza-se a organização das avaliações, conteúdo qualitativo para exploração quantitativa e qualitativa. A metodologia apresenta a criação e validação de um modelo de classificação em três níveis. O Nível Conceitual inclui a exploração de conhecimento de especialistas do domínio, com a criação de um conjunto de 12 perfis turísticos e definição de destinos. Nesse nível também, ocorre a coleta de 3.4 milhões de avaliações turísticas escritas em português. No Nível Tecnológico, as informações são organizadas, representadas e um processo de classificação automático de texto é realizado usando diferentes técnicas de Aprendizado de Máquina. O Nível Validação apresenta uma comparação entre os métodos automáticos e a classificação realizada pelos especialistas. O método com melhor desempenho é utilizado para explorar a compatibilidade entre destinos, atrações, estados, países e perfis, assim como as diferenças entre a popularidade e similaridade de destinos perante um perfil. Explora-se a similaridade entre destinos e a variação de perfis nos destinos mais visitados. Os resultados específicos apresentam

descobertas para o turismo, como a identificação dos melhores destinos para cada perfil, destinos mais populares que não são os mais relevantes para um perfil ou a identificação de um grau de similaridade muito alto entre destinos nacionais e internacionais. Os resultados do modelo apresentam acurácia superior à 70%, usando tecnologia e especialistas oferecem uma alternativa importante para modelos de organização do conhecimento, principalmente devido ao dinamismo e crescimento exponencial de conteúdo na Internet. Os resultados podem ajudar turistas que procuram certas experiências, governos a fomentar o turismo para um público específico ou entidades privadas que visam ofertar produtos e serviços direcionados. Independente do ator no processo, a organização e classificação de informações turísticas exerce um facilitador no processo decisório.

Palavras-chave: Classificação automática de texto. Aprendizado de Máquina. Extração de conhecimento. Avaliações turísticas. Perfis no Turismo.

ABSTRACT

The technological evolution paradigm has brought a disruptive change people's behavior, who now make decisions based on the content they consume on the Internet. This aspect is no different in the Tourism industry, where new technologies and the sharing of reviews allow users to seek information to support decisions such as choosing a destination, accommodation, attractions, food, among others. Reviews provide an important source of information; however, their volume can make it difficult to extract knowledge and use it effectively. How to find out if a particular point of interest with more than 100,000 opinions written in unstructured text is similar to what a tourist is looking for? This question motivates the development of this research, which has as its direct objective the creation of a model that allows transforming the reviews made by users into tourist classes (profiles). In the literature, some works try to address the problem of point of interest classification using reviews. In addition, the use of profiles in tourism is common, as a way of classifying destinations and tourists. In this sense, this study can present an additional view on both aspects, while allowing the joining of tourist profiles with review's information. The work presents an applied research, based on Pragmatism, of a hybrid nature with an exploratory objective. It uses the reviews organization as they quality nature as a source for a quantitative exploration analysis. The methodology presents the creation and validation of a classification model at three levels. At the Conceptual Level, knowledge is explored from domain experts, such as the creation of a set of 12 tourist profiles and definition of destinations to be used in the research. At this level, 3.4 million tourist reviews written in Portuguese are also collected. At the Technological Level, information is organized, represented and an automatic text classification process is carried out using different Machine Learning techniques. The Validation Level presents a comparison between automatic methods and a classification carried out by specialists. The best performing method is used to explore compatibility between destinations, attractions, states, countries and profiles, as well as the differences between the popularity and similarity of destinations with a profile. It also explores the similarity between destinations and the profile variation of the most visited destinations. The specific results present interesting discoveries in tourism, such as the identification of the best destinations for each profile, the most popular destinations that are not the most relevant for a profile or the identification of a very high degree of similarity between national and international destinations. The

model performance above 70% accuracy, using technology and specialists offer an important alternative for models of knowledge organization, mainly due to the dynamism and exponential growth of content on the Internet. The results can help tourists looking for certain experiences, governments to promote tourism for a specific audience or private companies that aim to offer targeted products and services. Regardless of the actor in the process, the organization and classification of tourist information turn the decision-making process easier.

Keywords: Automatic Text Classification. Machine Learning. Knowledge Discovery. Tourism reviews. Tourism profiles.

LISTA DE FIGURAS

Figura 1 –	Mapa de literatura	29
Figura 2 –	Tipos de Aprendizado de Máquina	36
Figura 3 –	Processo básico de classificação automática de texto	38
Figura 4 –	Processo de classificação automática	40
Figura 5 –	Exemplo de <i>crawler</i>	40
Figura 6 –	Exemplo pré-processamento utilizando Python.....	42
Figura 7–	Diferença entre abordagens CBOW e Skip-Gram	44
Figura 8 –	Termos próximos à Esqui	44
Figura 9 –	Nuvens com importância de palavras para os tópicos.....	50
Figura 10 –	Avaliação de usuário do TripAdvisor.....	53
Figura 11 –	Avaliação com texto pequeno	53
Figura 12 –	Avaliação com texto grande.....	54
Figura 13 –	Categorias atração Cataratas do Iguaçu	55
Figura 14 –	Framework de pesquisa exibindo os três níveis	64
Figura 15 –	Página de Nova York no TripAdvisor.....	68
Figura 16 –	Página da atração Plataforma de Observação Top of the Rock.....	69
Figura 17 –	Arquitetura do processo de extração de dados	70
Figura 18 –	Estrutura da classe	71
Figura 19 –	Estratégia de classificação automática	78
Figura 20 –	Exemplo de limpeza de avaliações.....	82
Figura 21 –	Exemplo de remoção de <i>stop-words</i>	82
Figura 22 –	Processo de remoção de adjetivos, verbos e nomes próprios	83
Figura 23 –	Exemplo de processo de Stemização.....	84
Figura 24 –	Processo de criação de bigramas e trigramas.....	84
Figura 25 –	Exemplo resultado modelo Universidade de São Paulo (USP)	88
Figura 26 –	Exemplo resultado usando o método próprio	89
Figura 27 –	Exemplo de representação <i>bag-of-words (BOW)</i>	90
Figura 28 –	Arquitetura processo de classificação por similaridade de representação	93
Figura 29 –	Criação de <i>corpus</i> para as avaliações de atração e perfis	95
Figura 30 –	Representação do <i>corpus</i> usando Termo Frequência – Inverso Documento Frequência (TF-IDF).....	96

Figura 31 – Similaridade entre Maragogi e o perfil Cultura.....	96
Figura 32 – Similaridade entre Maragogi e o perfil Praia.....	96
Figura 33 – Resultado de similaridade entre a atração e os perfis.....	97
Figura 34 – Carregando modelo <i>Word2vec</i>	98
Figura 35 – <i>Corpus</i> dos perfis Praia e Cultura com índices.....	99
Figura 36 – Função para classificação usando algoritmo <i>Word2vec</i>	99
Figura 37 – Comparação do perfil Praia com a avaliação exemplo.....	100
Figura 38 – Comparação do perfil Cultura com a avaliação exemplo.....	100
Figura 39 – Arquitetura do processo de classificação supervisionada.....	101
Figura 40 – Processamento modelo <i>Support Vector Machine (SVM)</i>	103
Figura 41 – Desempenho do modelo <i>Support Vector Machine (SVM)</i>	103
Figura 42 – Matriz de Confusão do modelo <i>Support Vector Machine (SVM)</i>	104
Figura 43 – Principais termos do perfil Vida Noturna.....	122
Figura 44 – Principais termos do perfil Religioso.....	128
Figura 45 – Comparativo de perfis entre Bruxelas e Ouro Preto.....	145

LISTA DE GRÁFICOS

Gráfico 1 – Crescimento do turismo pelo número internacional de chegadas	21
Gráfico 2 – Hiperplanos do algoritmo <i>Support Vector Machine (SVM)</i>	46
Gráfico 3 – Hiperplanos do algoritmo <i>Support Vector Machine (SVM)</i>	48
Gráfico 4 – Resultado de representação Termo Frequência – Inverso Documento Frequência (TF-IDF)	86
Gráfico 5 – Similaridade entre documentos por cosseno do ângulo	92
Gráfico 6 – Desempenho dos modelos de classificação teste com cinco <i>k-fold</i> ..	106
Gráfico 7 – Países com atrações mais similares à Vida Noturna.....	120
Gráfico 8 – Estados com atrações mais similares à Vida Noturna.....	121
Gráfico 9 – Frequência X Peso dos principais termos do perfil Vida Noturna.....	123
Gráfico 10 – Países com atrações mais similares ao perfil Religioso	126
Gráfico 11 – Estados com atrações mais similares ao perfil Religioso	127
Gráfico 12 – Frequência X Peso dos principais termos do perfil Religioso	128
Gráfico 13 – Popularidade X Relevância de destinos para o perfil Cultura.....	137
Gráfico 14 – Popularidade X Relevância de destinos para o perfil Natureza/Exóticos	139
Gráfico 15 – Variação de perfis entre os dez destinos nacionais mais visitados ...	148
Gráfico 16 – Variação de perfis entre os dez destinos internacionais mais visitados	150

LISTA DE QUADROS

Quadro 1 – Métodos de aprendizado de máquina.....	37
Quadro 2 – Perfis turísticos.....	56
Quadro 3 – Esquema de classificação (perfis turísticos).....	76
Quadro 4 – Atrações usadas para geração de <i>corpus</i> do perfil.....	79
Quadro 5 – Exemplo de avaliações	86

LISTA DE TABELAS

Tabela 1 – Destinos turísticos utilizados no estudo	72
Tabela 2 – Termos mais frequentes no <i>corpus</i> de cada perfil	91
Tabela 3 – Desempenho dos modelos de classificação	105
Tabela 4 – Desempenho dos modelos de classificação usando <i>word embeddings</i>	107
Tabela 5 – Classificação da atração Museu do Louvre com o método SVM.....	108
Tabela 6 – Classificação manual do destino Paris pelos especialistas	110
Tabela 7 – Classificação do destino Paris usando o método TF-IDF	112
Tabela 8 – Validação dos métodos automáticos – distâncias estatísticas	113
Tabela 9 – Validação dos métodos automáticos – relevância do perfil para o Destino	114
Tabela 10 – Atrações mais similares ao perfil Vida Noturna	118
Tabela 11 – Destinos mais similares ao perfil Vida Noturna.....	119
Tabela 12 – Atrações mais similares ao perfil Religioso.....	124
Tabela 13 – Destinos mais similares ao perfil Religioso.....	125
Tabela 14 – Quantidade de categorias do TripAdvisor por perfil.....	136
Tabela 15 – Similaridade entre destinos considerando os perfis turísticos	141
Tabela 16 – Destinos nacionais mais visitados entre 2011 e 2019	147
Tabela 17 – Relevância de perfis dos dez destinos mais visitados entre 2011 e 2019	147
Tabela 18 – Destinos internacionais mais visitados entre 2011 e 2019	149
Tabela 19 – Relevância de perfis dos dez destinos mais visitados entre 2011 e 2019	149

LISTA DE SIGLAS E ABREVIATURAS

AM	Aprendizado de Máquina
BOW	Bag of Words
BPS	Escala de Personalidade da Marca
CI	Ciência da Informação
e-WOM	boca-a-boca digital
IA	Inteligência Artificial
KOS	Sistemas de Organização do Conhecimento
LDA	Latent Dirichlet Allocation
MIT	Instituto de Tecnologia de Massachusetts
MT	Mineração de Texto
NLTK	Conjunto de Ferramentas para Linguagem Natural
OWL	Linguagem Ontológica da Web
PIB	Produto Interno Bruto
PLN	Processamento de Linguagem Natural
POI	Ponto de Interesse
PPGGOC	Programa de Pós-Graduação em Gestão & Organização do Conhecimento
RI	Recuperação da Informação
SCM	<i>Soft Cosine Measure</i> = Medida de Cosseno Leve
SVM	Support Vector Machine
TF	Termo Frequência
TF-IDF	Termo Frequência – Frequência Inversa do Documento
TI	Tecnologia da Informação
USP	Universidade de São Paulo
WOM	word of mouth
WTTC	Conselho Mundial de Viagem e Turismo

SUMÁRIO

1	INTRODUÇÃO	20
1.1	<i>Delimitação do problema</i>	22
1.1.1	Questões de pesquisa	25
1.2	<i>Objetivos</i>	25
1.3	<i>Inovação e contribuições da pesquisa</i>	26
1.4	<i>Estrutura da tese</i>	27
2	REFERENCIAL TEÓRICO	28
2.1	<i>Mapa de Literatura</i>	28
2.2	<i>Gestão da informação</i>	29
2.2.1	Organização da informação	30
2.2.2	Recuperação da informação	31
2.2.3	Explosão de dados e informações	32
2.2.4	Conteúdo Não Estruturado – Avaliações	33
2.3	<i>Tecnologia da informação</i>	34
2.3.1	Inteligência artificial	34
2.3.2	Aprendizado de máquina	35
2.3.2.1	<i>Abordagens do aprendizado de máquina</i>	36
2.3.3	Classificação automática de texto	38
2.3.3.1	<i>Aquisição dos textos</i>	40
2.3.3.2	<i>Definição de objetivos e métodos</i>	41
2.3.3.3	<i>Pré-processamento dos textos</i>	41
2.3.3.4	<i>Representação dos textos</i>	42
2.3.3.5	<i>Modelo de classificação supervisionado</i>	45
2.3.3.5.1	Algoritmos supervisionados	45
2.3.3.5.2	Avaliação da classificação	46
2.3.3.6	<i>Modelo de classificação não supervisionado</i>	48
2.3.3.6.1	Modelagem de tópicos	49
2.4	<i>Informação no Turismo</i>	50
2.4.1	Avaliações no turismo	51
2.4.2	Avaliações em uma plataforma de turismo: o exemplo do TripAdvisor	52
2.4.3	O uso de perfis no turismo	55

2.4.4	Estado da arte.....	58
3	METODOLOGIA E DESENVOLVIMENTO.....	60
3.1	<i>Aspectos gerais.....</i>	60
3.2	<i>Metodologia da pesquisa</i>	61
3.3	<i>Nível conceitual</i>	65
3.3.1	Participação de especialistas	65
3.3.2	Destinos estudados	66
3.3.3	Coleta de dados.....	67
3.3.4	Coletando avaliações.....	68
3.3.5	Arquitetura de organização dos dados.....	70
3.3.6	Criação de perfis turísticos.....	74
3.3.7	Estratégia de classificação	76
3.4	<i>Nível tecnológico.....</i>	80
3.4.1	Pré-processamento dos textos.....	80
3.4.2	Representação dos textos.....	85
3.4.3	Representação com TF e TF-IDF	85
3.4.4	Representação com <i>word embeddings</i>	87
3.4.5	Representação com <i>bag-of-words (BOW)</i>	89
3.4.6	Classificação por similaridade da representação.....	91
3.4.6.1	<i>Verificação de similaridade</i>	92
3.4.6.2	<i>Arquitetura do processo classificatório.....</i>	93
3.4.6.3	<i>Classificação usando Termo Frequência (TF) e Termo Frequência – Inverso Documento Frequência (TF-IDF)</i>	95
3.4.6.4	<i>Classificação usando casamento de palavras</i>	97
3.4.7	Classificação supervisionada.....	101
3.4.8	Arquitetura do processo de classificação supervisionada.....	101
3.4.9	Modelo de classificação usando Support Vector Machine (SVM)...	102
3.4.10	Desempenho dos modelos de classificação supervisionados	105
3.4.11	Modelo supervisionado com representação <i>word embeddings</i>	106
3.4.12	Aplicação dos modelos no <i>corpus</i> de atração não classificado	107
3.5	<i>Nível de validação</i>	109
3.5.1	Classificação manual de destinos nos perfis.....	109
3.5.2	Validação estatística entre classificação humana e automática	111

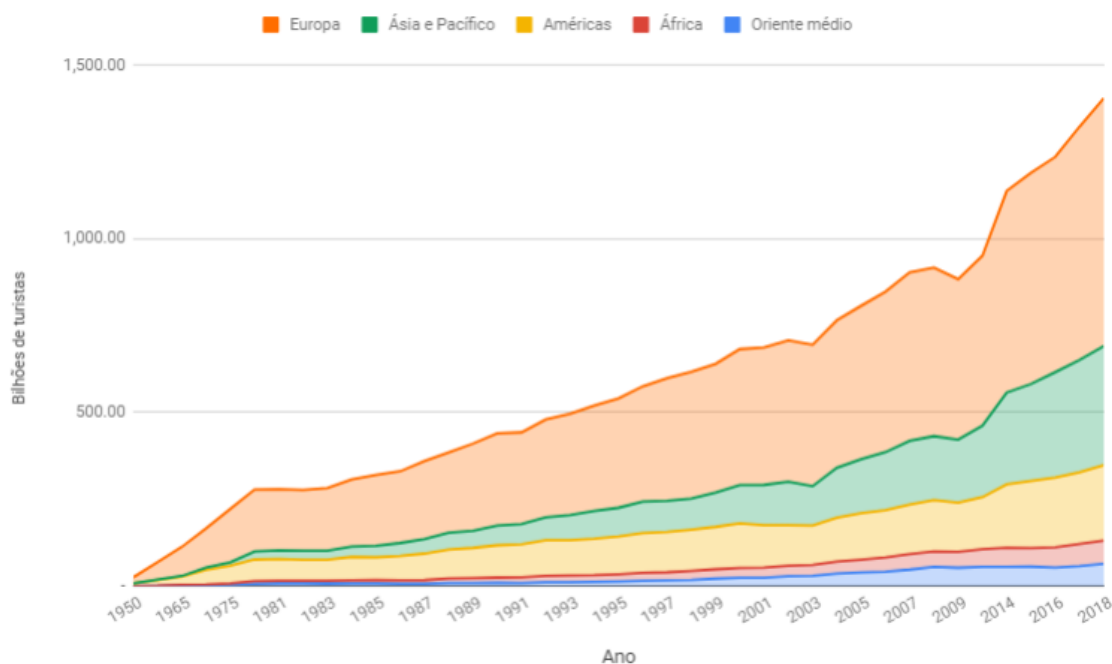
3.5.3	Validação qualitativa entre classificação humana e automática.....	113
4	RESULTADOS E DISCUSSÕES	116
4.1	Resultados analíticos por perfil	116
4.1.1	Resultado do perfil Vida Noturna	117
4.1.1.1	<i>Atrações mais similares ao perfil Vida Noturna</i>	117
4.1.1.2	<i>Destinos mais similares ao perfil Vida Noturna</i>	119
4.1.1.3	<i>Países mais similares ao perfil Vida Noturna</i>	120
4.1.1.4	<i>Estados nacionais mais similares ao perfil Vida Noturna</i>	121
4.1.1.5	<i>Análise linguística do perfil Vida Noturna</i>	122
4.1.2	Resultado do perfil Religioso	124
4.1.2.1	<i>Atrações mais similares ao perfil Religioso</i>	124
4.1.2.2	<i>Destinos mais similares ao perfil Religioso</i>	125
4.1.2.3	<i>Países mais similares ao perfil Religioso</i>	126
4.1.2.4	<i>Estados nacionais mais similares ao perfil Religioso</i>	126
4.1.2.5	<i>Análise linguística do perfil Religioso</i>	127
4.2	Resultados sintéticos por perfil	129
4.2.1	Perfil Cultura	129
4.2.2	Perfil Paisagem/Arquitetura	130
4.2.3	Perfil Família	130
4.2.4	Perfil Gastronomia	131
4.2.5	Perfil Aventura	132
4.2.6	Perfil Praia	132
4.2.7	Perfil Compras	133
4.2.8	Perfil Relaxante	133
4.2.9	Perfil Romântico	134
4.2.10	Perfil Natureza/Exóticos	135
4.3	Popularidade x relevância do perfil	135
4.4	Similaridade entre destinos	140
4.5	Perfil turístico de destinos mais visitados	146
5	CONCLUSÕES	152
6	CONSIDERAÇÕES FINAIS	162
	REFERÊNCIAS	163

1 INTRODUÇÃO

O paradigma da evolução tecnológica trouxe uma mudança disruptiva no comportamento das pessoas, que agora tomam decisões baseando-se no conteúdo que consomem na Internet. As decisões tomadas podem ser simples como a escolha de um destino turístico ou mais complexas como a escolha de um presidente. Na era da informação, esse insumo não é produzido apenas por veículos oficiais, haja vista que agora, todos são atores no processo de geração e propagação de informações. Essa realidade possibilita um avanço ao passo que as informações conseguem atingir mais pessoas em menos tempo, fator essencial num mundo globalizado. Há de se ressaltar que esse conteúdo produzido carece de organização e validação, porque influencia diretamente a vida das pessoas. Isso evidencia o papel da Ciência da Informação (CI) como fator chave no progresso da sociedade, permitindo que o conteúdo seja organizado com efetiva geração de conhecimento. Muitos segmentos têm sido afetados por esse novo comportamento, entre eles a indústria do Turismo. Simples ou complexas, muitas decisões se dirigem ao turismo, como a escolha de um hotel, de um carro para alugar, a escolha de uma atração para visitar, um destino turístico, um restaurante ou até mesmo o menu que será servido no jantar.

Em 2019, o Turismo foi responsável por 10,3% do Produto Interno Bruto (PIB) global e movimentou 330 milhões de empregos, representando 10% dos empregos no mundo (WORLD TRAVEL & TOURISM COUNCIL, 2019). Nos últimos anos, o setor vem crescendo consideravelmente e ganhando ainda mais importância no cenário econômico mundial. Em 2019, houve um crescimento de 3,5%, superior ao crescimento geral da economia, que representou 2,5%, e isso ocorre pelo nono ano consecutivo (WTTC, 2020). O Gráfico 1 exibe o crescimento do Turismo medido pelo número internacional de chegadas de turistas por região.

Gráfico 1 – Crescimento do turismo pelo número internacional de chegadas



Fonte: WTTC (2020).

O Gráfico 1 apresenta o número de turistas internacionais chegando em cada região. É possível observar que o número de turistas cresce exponencialmente desde 1950. Há uma evolução de 25 milhões para 1,4 bilhão de turistas por ano. Essa evolução foi possível devido ao investimento em infraestrutura, melhoria de acesso à informação e principalmente, o contínuo crescimento da classe média. Além disso, há forte influência de fatores ligados à melhoria das atrações de destinos (WTTC, 2019). No Brasil, o turismo expandiu 7% e movimentou mais de 1,6 trilhão de dólares em 2017 e foi responsável por 8% do PIB nacional em 2016 (TOMÉ, 2018). Durante o ano de 2019, 6,3 milhões de pessoas desembarcaram no Brasil de voos internacionais, número próximo ao de 6,5 milhões durante o ano de 2016 quando houve o evento das Olimpíadas, ano em que o país recebeu mais turistas estrangeiros¹.

A Internet tem sido um pilar importante nesse crescimento, pois permite a oferta de serviços de turismo em um formato diferenciado, capaz de atingir mais pessoas em menos tempo. Essa realidade permite que consumidores se orientem melhor, buscando e compartilhando informações de suas viagens na Internet. Permite também que consumidores adquiram produtos e serviços por preços melhores e facilita a

¹ Disponível em: <https://www.gov.br/turismo/pt-br>. Acesso em: 4 fev. 2021.

comunicação entre cliente e fornecedor. Apesar da Internet ser o fator tecnológico principal, a evolução tecnológica no geral permite uma mudança no cenário do turismo. Os turistas vêm alterando seu perfil ao longo dos últimos anos, principalmente diminuindo a dependência de agências de turismo para planejar e executar suas viagens. Com o volume de informações disponível, aplicativos, sites, soluções são criados com o propósito de subsidiar as ações independentes do turista. É possível comprar passagens online, reservar hotéis, carros, restaurantes, descobrir as atrações principais de determinado local, rastrear a bagagem, converter moedas ou resolver problemas com o idioma de outro país, tudo isso disponível em tempo real na palma da mão do usuário.

Soluções como Google, Airbnb, TripAdvisor, Booking, Yelp, Trivago e Expedia oferecem serviços baseados em dados e informações que facilitam a tomada de decisão dos usuários. O surgimento dessas plataformas ocorre em escala paralela e natural à evolução de serviços ofertados com base em um vetor único: a Internet. Além desses portais, as redes sociais permitem a disseminação de conteúdo textual não-estruturado na forma de avaliações sobre produtos e serviços. A pesquisa de Askalidis e Malthouse (2016) aponta que há um aumento de 270% de conversão (venda) de um produto que acumula mais avaliações, e mostra que a maior parte desse efeito, vem apenas das 10 primeiras avaliações. Um dos motivadores é a quantidade de conteúdo existente e a dificuldade dos usuários de classificar, organizar e recuperar a informação de forma personalizada. A pesquisa da Nielsen Company (2016) aponta que 63% das pessoas que compram serviços relacionados ao turismo, fazem pesquisa online. O conteúdo é um facilitador, no entanto, o volume de dados e informações vem crescendo exponencialmente e necessita de evolução dos meios de organização do conhecimento como forma de facilitar o seu entendimento, interpretação e, conseqüentemente, a tomada de decisão.

1.1 Delimitação do problema

O compartilhamento de informações em plataformas de serviços relacionados ao turismo é uma forma de agregar valor a tais portais, e como consequência, ocorre o crescimento exponencial de informações supracitado. As opiniões de turistas são registradas em forma textual de avaliações e *rating* (até 5 estrelas na escala Likert); sendo geralmente direcionadas a três categorias: hotéis, restaurantes e atrações.

Esse conteúdo pode ser positivo ou negativo e permite a redução de incerteza no processo decisório dos usuários (FANG *et al.*, 2016).

A atração Cataratas do Iguaçu na cidade de Foz do Iguaçu possui 44 mil avaliações no site TripAdvisor². No Google, essa mesma atração possui 63 mil avaliações³. Observa-se também que um turista geralmente escolhe um destino com base num conjunto de atrações e não apenas em uma (ARENTZE; KEMPERMAN; AKSENOV, 2018). Ou seja, ao escolher um destino, um turista além de analisar uma atração, como no exemplo de Foz do Iguaçu, procura por outras atrações, assim, aumentando o grau de dificuldade na análise das avaliações e na sua decisão.

Se observamos a cidade de Foz do Iguaçu, apenas no TripAdvisor, possui 19 atrações listadas, somando um total de 112 mil avaliações. Esse número gera um paradoxo, se por um lado, há uma fonte rica de informação, por um outro, como um usuário pode se beneficiar do melhor que há nesse conteúdo? Guy *et al.* (2017) mostrou que devido a esse número elevado de informações, os usuários não conseguem absorver corretamente o conteúdo e há uma perda de informações importantes. É necessário investir em formas de classificar e organizar esse conteúdo, permitindo aos usuários explorar melhor o conhecimento existente. Dada a relevância das avaliações no processo de decisão do usuário, como Chevalier e Mayzlin apontam em seu estudo já em 2006, o número de estudos voltados à organização desse conteúdo vem crescendo a cada ano.

Como um turista, em processo de escolha do seu próximo destino, poderia classificar se determinada atração é interessante para toda a família, ou romântica, ou relacionada à natureza? Se o turista pudesse identificar os perfis de uma determinada atração ou destino, quais outras atrações ou destinos ofereceriam experiências semelhantes que talvez estariam mais adequadas ao orçamento ou expectativa de viagem? Gretzel *et al.* (2006) demonstra que o uso de perfis turísticos pode ser importante para ajudar a recomendar destinos e atrações. As escolhas no turismo geralmente são complexas e relacionadas a experiências emocionais, e uma das formas de solucionar isso é por meio de soluções baseadas em modelos de perfis Sertkan, Neidhardt e Werthner (2018). Burke, Ramezani e Göker (2011) argumentam

² Disponível em: <https://www.similarweb.com/website/tripadvisor.com#search>. Acesso em: 4 fev. 2021.

³ Disponível em: <https://www.tripadvisor.com.br/>. Acesso em: 4 fev. 2021.

que a melhor forma para essa recomendação é baseada em conhecimento e/ou conteúdo.

A Internet permite que o turista se torne mais independente (LEE; LAW; MURPHY, 2011), e a possibilidade de obter perfis de atrações e destinos fornece um facilitador a mais no processo decisório desse usuário. Não somente turistas podem se beneficiar dessa classificação, mas também empresas privadas podem explorar nichos de mercado por meio da oferta de serviços e soluções para um público segmentado. Governos federal, estadual e municipal podem explorar limitações de oferta turística e adaptar, remover ou adicionar novas atrações com o objetivo de fomentar o turismo para um público específico. Tais dados processados, poderiam ser usados em sistemas de recomendação para usuários específicos, como por exemplo: De acordo com os usuários, quais os melhores destinos nacionais para quem gosta de gastronomia? Quais atrações são mais relevantes culturalmente no estado de Minas Gerais? Independente do usuário, seja ele turista, governo ou setor privado, o conteúdo das avaliações pode ser mais explorado se as informações são organizadas.

O conceito de dados está intrinsicamente ligado ao conceito de informação, porque dados são a base da informação, e informação é materializada pelo processamento de dados (WANG, 2018). A CI possui duas grandes vertentes: de um lado, a perspectiva técnica, a qual busca classificar, representar, organizar, curar e armazenar a informação; e de outro, uma perspectiva social, que busca o estudo da interação humana com a informação, aspectos legais e políticos de seu uso (MARCHIONINI, 2017). A situação apresenta um desafio da CI, ou seja, a dificuldade de organizar e extrair conhecimento de uma massa cada vez maior e variada de dados. Apesar de não haver consenso na conceituação dessa massa de dados, o termo *big data* é usado para descrever tal realidade. De acordo com Souza, Almeida e Baracho (2013), o *Big Data* oferece uma oportunidade única para a CI, haja vista que suas técnicas nunca foram tão necessárias como antes. Os recursos técnicos que fundamentam a CI em conjunto com tecnologia aplicada permitem que soluções de classificação automáticas ou semiautomáticas sejam implementadas, trazendo dessa forma agilidade no processo de classificação, redução de custo e melhoria na extração do conhecimento. A CI é inexoravelmente ligada à Tecnologia da Informação (TI) (SARACEVIC, 1996). Por ser um campo interdisciplinar, fornece um arcabouço metodológico para que se torne possível melhorar a exploração do conhecimento

existente nas avaliações. O problema pode ser classificado em CI nas subáreas de Classificação Automática, Mineração de Textos (MT) ou Descoberta de Conhecimento, observando que tais métodos visam transformar dados brutos e dispersos em conhecimento acionável (MA; SONG; ZHANG, 2008).

A necessidade de se classificar destinos e atrações turísticas em perfis a partir das avaliações dos usuários, a dificuldade nesse processo de classificação com consequente exploração do conhecimento, bem como as indagações apresentadas, evidenciam a razão e motivam a presente pesquisa.

1.1.1 Questões de pesquisa

Considerando a impossibilidade de organizar o conhecimento manualmente e, conforme o tema e o problema apresentados, surge a seguinte questão primária de pesquisa: Como descobrir se uma determinada atração com mais de 100 mil opiniões escritas em texto não-estruturado possui similaridade com o que o turista procura? De forma secundária, por meio da metodologia desenvolvida, é possível extrair as informações a seguir:

- a) qual a similaridade entre os destinos estudados?
- b) os destinos mais populares são os mais relevantes para um perfil?
- c) quais os 10 principais destinos por perfil?
- d) quais as 10 principais atrações por perfil?
- e) qual o perfil turístico das cidades mais visitadas?

As questões secundárias emergem como entendimento por parte do autor, que um dos papéis da CI é gerar conhecimento científico por meio da solução de problemas reais, organizando a informação e extraíndo conhecimento que permita um facilitador no processo decisório sobre o tema proposto.

1.2 Objetivos

Espera-se que os resultados ajudem turistas, governo ou empresas a melhorarem seu processo decisório usando o conhecimento existente nas opiniões dos próprios turistas. A análise linguística das avaliações está dentro do conceito de processamento de linguagem natural à luz da CI, e espera-se contribuir com as técnicas de organização do conhecimento.

1.2.1. Objetivo geral

A presente pesquisa tem como objetivo geral classificar o conhecimento existente nas avaliações sobre atrações turísticas feitas por brasileiros, mapeando atrações e destinos em perfis turísticos. Como consequência, procura explorar e descobrir novas informações que possam surgir desta nova realidade.

1.2.2. Objetivos específicos

Pautando-se pelo objetivo principal, objetivos específicos detalham etapas que precisam ser cumpridas em prol de atingir o escopo total da pesquisa. A seguir apresentamos os objetivos específicos:

- a) levantar os principais destinos turísticos dos brasileiros;
- b) criar um conjunto de perfis turísticos;
- c) identificar melhores métodos de classificação de atrações e destinos;
- d) explorar as informações a partir dessa classificação;
- e) descobrir o perfil dos destinos mais visitados por brasileiros;
- f) identificar similaridades entre destinos e atrações turísticas;
- g) explorar linguisticamente os termos que compõem cada perfil;
- h) identificar popularidade e relevância de destinos.

1.3 Inovação e contribuições da pesquisa

Esta pesquisa apresenta como inovação, criar e mapear perfis turísticos com base nas avaliações dos usuários. São consideradas apenas as avaliações feitas por brasileiros que visitaram determinada atração. Os resultados permitem explorar e obter uma visão holística e técnica sobre o perfil do turista, destinos e atrações de acordo com a opinião dos brasileiros. Esta pesquisa contribui técnica e conceitualmente ao campo de CI, exibindo as etapas do processo de aquisição, armazenagem, curadoria, classificação e recuperação da informação, tendo como dados as avaliações feitas pelos usuários.

A possibilidade de identificar perfis turísticos de destinos e atrações pode ser um fator importante no processo decisório do turista. O modelo desenvolvido poderá ser utilizado num sistema de recomendação de atrações e destinos turísticos com base na identificação prévia de variáveis do turista como orçamento, localização, objetivo

de viagem, estado civil e grupo familiar. Uma vez identificado o perfil do turista é possível recomendar destinos e atrações que ofereçam experiências turísticas similares com o seu perfil. Exemplo, se o turista procura por destinos românticos, mas seu orçamento não permite ir a Paris, qual destino nacional ofereceria atrações e experiências mais próximas à Cidade da Luz? Cohen (1972) aponta que a demanda, necessidade e expectativa do turista variam consideravelmente de acordo com sua idade, ciclo familiar e renda familiar.

O turismo representa uma parcela considerável da economia global com ramificação em países, estados e municípios. O governo em geral pode se beneficiar dessa pesquisa, sendo possível traçar, como por exemplo, uma matriz SWOT⁴: instrumento de gestão que permite avaliar Forças, Fraquezas, Oportunidades e Ameaças com base no perfil de suas cidades. Diversas decisões de fomento ao turismo poderiam ser segmentadas, como o direcionamento de marketing para nichos específicos de público interessado nos perfis turísticos mais relevantes da cidade, ou a criação e ampliação de experiências turísticas para melhorar perfis de baixa relevância. Instituições privadas poderiam direcionar marketing para um público específico ou aproveitar nichos para a criação de atrações direcionadas a suprimirem as limitações de determinado segmento dentro da cidade ou estado.

1.4 Estrutura da tese

O trabalho é dividido em 7 capítulos. O referencial teórico, descrito no capítulo 2, apresenta quatro aspectos macros: o mapa de literatura, o processo de gestão da informação no contexto da explosão informacional, o uso de tecnologia como forma de tentar solucionar o problema e o uso de informações e avaliações no turismo. O capítulo 3 apresenta a metodologia desenvolvida em seus três níveis: conceitual, tecnológico e validação dos modelos. O processo de validação dos modelos apresenta resultados do processo classificatório. A partir do processo realizado, uma exploração de informações turísticas com foco nos objetivos sob uma ótica qualitativa é abordada no capítulo 4. O capítulo 5 apresenta-se as conclusões da pesquisa com base nos objetivos e questões levantadas. O capítulo 6 expõe um olhar sobre o conhecimento e seu impacto no turismo. Por fim, o capítulo 7 apresenta as referências utilizadas.

⁴ Strengths, weaknesses, opportunities, threats.

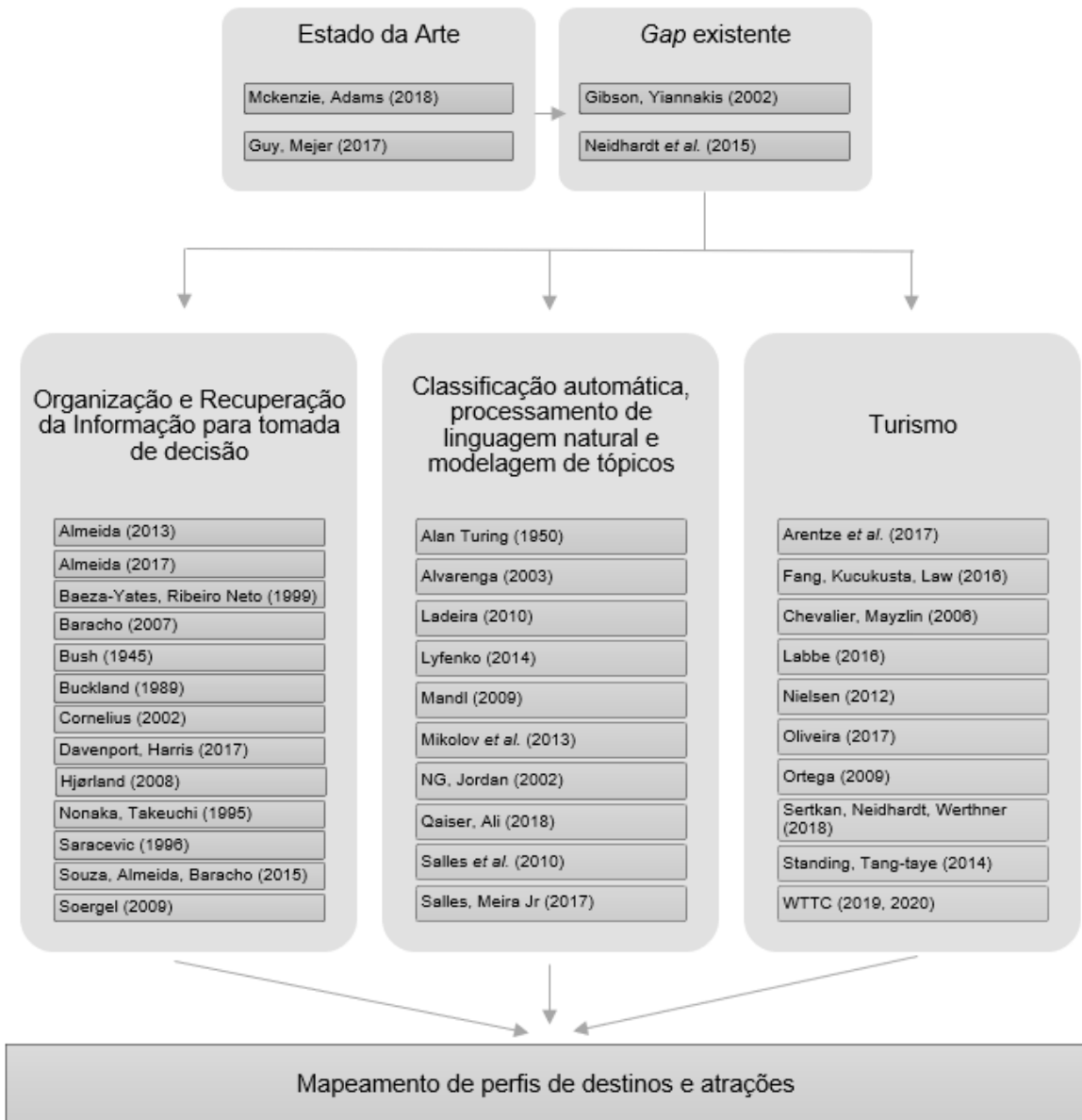
2 REFERENCIAL TEÓRICO

Calvin Mooers (1951) introduziu o termo Recuperação da Informação (RI) envolvendo tanto aspectos técnicos como sociais. A demanda pelo insumo informação sempre existiu, e com o seu aumento exponencial, cresce de forma proporcional à necessidade de métodos e tecnologias capazes de lidar com essa realidade. Esse trabalho utiliza o conceito preconizado por Buckland (1991) sobre informação como coisa e informação como conhecimento. A Informação como coisa, por sua qualidade informativa, a partir da qual é possível extrair *insights* e conhecimento e a informação como conhecimento, pois, por meio do seu uso, é possível reduzir a incerteza no processo decisório (BUCKLAND, 1991). A informação sempre foi um dos ativos mais importantes em qualquer esfera de decisão. Outro fundamento desse trabalho é o conceito de Cornelius (2002), em que, a informação alimenta e altera a estrutura de conhecimento em um receptor. A informação classificada e processada permite ao receptor um novo ponto de vista para ajudá-lo a interpretar eventos ou objetos, sendo elemento fundamental para construção do conhecimento humano (NONAKA; TAKEUCHI, 1997).

2.1 Mapa de Literatura

A fundamentação teórica desse trabalho é composta pela Gestão da Informação e Conhecimento, com foco na Organização e Recuperação de Informação. O problema oriundo da explosão de dados e informações culmina na necessidade de investir em técnicas de Inteligência Artificial (IA) como ferramenta, dessa forma, apresenta-se uma revisão literária de métodos de Aprendizado de Máquina (AM) com foco na Classificação Automática de Texto. Em seguida, apresenta-se um panorama sobre o uso de informações no turismo, a importância dos perfis, o *gap* (lacuna existente no objeto de pesquisa), e o estado da arte de pesquisas relacionadas.

Figura 1 – Mapa de literatura



Fonte: elaborado pelo autor (2021).

2.2 Gestão da informação

O processo de organização e transição da informação é classificado por alguns como Arquitetura da Informação e ocorre em nível tático e operacional fornecendo os fundamentos metodológicos e funcionais que dá suporte à tomada de decisão (LIMA-MARQUES; MACEDO, 2006). De acordo com Davenport (1998), a gestão da informação envolve quatro etapas: a) determinação da demanda; b) obtenção; c) distribuição; d) utilização. Demanda, distribuição e utilização dependem dos envolvidos, da estrutura, dos canais de comunicação e envolvem aspectos subjetivos

culturais e sociais. Já a etapa de obtenção é onde a esfera técnica atua, permitindo coletar, explorar, organizar e recuperar.

2.2.1 Organização da informação

O objetivo da organização da informação é dar suporte ao fluxo de tratamento e recuperação dos objetos informacionais estruturados, semiestruturados e não-estruturados (VICTORINO; BRÄSCHER, 2009). Aristóteles aparentemente foi o primeiro a propor um sistema de categorias, indicando que objetos do conhecimento humano deveriam ser classificados em categorias. Os objetos deveriam ser classificados perante as categorias com base na similaridade de seus atributos (ALMEIDA, 2013). Sistemas de Organização do Conhecimento (KOS) devem permitir que o conhecimento expresso em informações seja efetivamente organizado para que as pessoas consigam recuperá-lo. De acordo com Hjørland (2008), existem dois aspectos diferentes de KOS: a) um aspecto mais amplo, a partir do qual considera que o conhecimento perpassa todas as esferas da sociedade, tendo em sua base a organização e disseminação do conhecimento como parte essencial à sociedade; e b) um aspecto mais específico, o qual geralmente refere-se a itens funcionais para organizar e gerenciar a informação, tornando sua manipulação e recuperação mais fáceis. Soergel (2008) cita que entre as técnicas de KOS destacam-se as Ontologias, Tesouros, Classificação, Taxonomia dentre outras, responsáveis por permitir essa organização.

Com uma quantidade de dados e informações produzidos em escala exponencial, o uso de técnicas de CI com técnicas de computação vem permitindo avanços significativos na organização da informação. De acordo com Nickel *et al.* (2016), o uso de modelos que considerem relações e regras entre entidades, como as ontologias, pode melhorar significativamente o desempenho de algoritmos de IA. Métodos tradicionais de KOS representam uma abordagem *top-down*, ou seja, universal, normativa, enquanto métodos de classificação automática representam uma abordagem *bottom-up*, ou seja, mais descritiva e dinâmica (IBEKWE-SANJUAN; BOWKER, 2017). O uso de métodos tradicionais de KOS em conjunto com técnicas de computação pode ser um fator importante para representar o conhecimento na era do *Big Data*. O conceito de “melhor dos dois mundos”, usando uma abordagem *bottom-up* permite um progresso no campo pragmático e epistemológico, amparando-

se em técnicas de KOS e computação para organizar a informação e facilitar a recuperação. Embora exista uma distinção de técnicas usadas em cada etapa, alguns autores conceituam o termo Recuperação da Informação (RI) como o meio envolvendo todo o processo de organização e sua recuperação, exatamente por ser o objetivo final das etapas anteriores (ALMEIDA *et al.*, 2017).

2.2.2 Recuperação da informação

Recuperação da Informação (RI) refere-se ao processo armazenamento, representação e retorno de informação para um problema específico do usuário (MANDL, 2009). A comparação de similaridade de uma *query* (demanda de informação) digitada pelo usuário e a representação de documentos constituem a base de RI. No entanto, esse processo envolve diversas outras etapas, como: armazenamento, modelagem, classificação e categorização, arquitetura do sistema, interface do usuário e visualização dos dados. Apesar de o grande volume de dados e informações, Mutula (2016) chama a atenção que apenas 0,5% são efetivamente analisados e usados, isso ocorre principalmente por limitação das técnicas de organização e recuperação da informação. A Internet é o principal canal para busca de informação e considerando o *Big Data*, a recuperação da informação carece de evolução e exploração para efetivamente usar o conhecimento produzido (HJØRLAND, 2016).

Embora a demanda de conhecimento seja semelhante perante dados e informação, há distinção com relação a sua recuperação com resultados diferentes aos usuários. Recuperar dados significa retornar ao usuário objetos que coincidam exatamente com os termos de busca, como por exemplo: liste os nomes e salários dos funcionários admitidos esse mês. Nesse caso, o uso de expressões regulares ou álgebra satisfazem com precisão a demanda do usuário. Por um outro lado, recuperar informação significa trabalhar com linguagem natural e muitas vezes com texto não estruturado. Isso pode levar a problemas semânticos como imprecisão e ambiguidade, como por exemplo: com base em suas publicações, liste todos os pesquisadores que conheçam sobre IA. Nesse caso, o retorno não é preciso, mas aproximado, exibindo os documentos mais relevantes perante os termos informados (BAEZA-YATES; RIBEIRO-NETO, 1999). O desafio de RI em face ao *Big Data* é retornar informação

relevante ao usuário num cenário crescente em complexidade com dados não organizados e a produção exponencial de informação não estruturada.

2.2.3 Explosão de dados e informações

Bush em 1945 classificou o fenômeno da explosão informacional como crítico e propôs como solução a utilização de tecnologias da informação como forma de tornar o conhecimento mais acessível. Este problema, nos últimos anos, vem se intensificando principalmente devido ao avanço tecnológico e inclusão digital, que provocou uma mudança de paradigma na produção e uso da informação. Produzida e disseminada por diversos atores, a informação teve com essa evolução digital, duas grandes alterações, tais como: o crescimento exponencial do volume de informação e a velocidade de sua disseminação, que se tornou extremamente rápida principalmente devido à Internet e redes sociais. Assim como informação, a utilização de dispositivos inteligentes, ou seja, dispositivos que conseguem produzir e transmitir dados, tornam esse cenário mais complexo. Essa realidade apresenta algumas características comuns conforme aponta Maluvaru, Shiram e Sugumaram (2014):

- a) *volume*: dados e informações coletados em grandes quantidades de uma variedade de locais como empresas, sistemas, redes sociais, dispositivos de usuário e serviços na web;
- b) *velocidade*: dados e informações gerados em tempo real por meio de mais conexões entre usuários e dispositivos;
- c) *variedade*: dados e informações estruturados ou não de diferentes formas: texto, vídeo, tabelas, banco de dados, imagens;
- d) *veracidade*: dados e informações de diferentes fontes; jornais, revistas, usuários das redes sociais, sites, blogs.

Alguns autores classificam essa nova realidade como *big data*, a qual seria caracterizada pela produção de informações e dados dos dispositivos (Internet das Coisas) considerando as características supracitadas. Nesse estudo não se procura discutir a conceituação ou nomenclatura usada para classificar tal realidade, pois, como apresenta (ZHAN; WIDÉN, 2017), não há um consenso sobre tais definições, ao invés disso, procura-se a partir de sua existência, apresentar técnicas para solucionar tal problema. Se por um lado, há a evidência de um problema informacional,

por um outro, há uma enorme oportunidade: a exploração do conhecimento existente nesse conteúdo.

2.2.4 Conteúdo Não Estruturado – Avaliações

O conteúdo disponível no *Big Data* pode ser estruturado, como em banco de dados relacionais, no qual dados são mapeados em atributos. Além disso, pode ser também não estruturado, com dados e informações dentro de texto, vídeo, imagem ou áudio sem estrutura definida, ou seja, sua recuperação requer prévia organização. Paul *et al.* (2014) apontam que 80% da informação armazenada em computadores (sem considerar imagens, áudios e vídeos) consiste em texto. Considerando que a informação disponível em áudios, vídeos e imagens também precisa ser processada para poder ser recuperada, nota-se que a maioria do conteúdo disponível na *web* carece de classificação para seu efetivo uso. A não ser que sejam aplicados métodos ontológicos para organização conteúdo, como a Linguagem Ontológica da Web (OWL), geralmente o conteúdo está disperso entre marcações HTML.

Um exemplo de conteúdo não estruturado importante em diversas esferas de decisão, são as avaliações. Uma avaliação expressa a opinião de determinado indivíduo sobre certo produto, serviço ou local. Essa opinião pode ser imparcial, negativa ou positiva e geralmente acompanha uma classificação de estrelas indicando esse sentimento na escala Likert⁵ de 1 a 5. As opiniões dos usuários estão expressas na Internet em portais como Facebook, Twitter, Instagram, Blogs, Google e diversos outros sites. No ramo do Turismo, é comum usuários escreverem uma avaliação sobre determinado hotel, restaurante, ponto ou destino turístico. Esse conteúdo não é estruturado e está disperso na *web* em sites como Google, Amazon, TripAdvisor, Booking e Trivago. Quando um consumidor em potencial, consome as avaliações online de um hotel por exemplo, ele está propenso a formar impressões sobre o conteúdo, como sua credibilidade, utilidade, precisão ou viés (SEN; LERMAN, 2017). Essas impressões exercem influência sobre suas decisões (SPARKS; PERKINS; BUCKLEY, 2013). Embora esteja em formato não estruturado, se processado e analisado, esse conteúdo pode ser uma importante fonte de conhecimento.

⁵ Escala Likert: exhibe o grau de concordância que um respondente afirma sobre determinado item entre Discordo Totalmente e Concordo Totalmente (LIKERT..., 2021).

Explorar esse conhecimento significa entender suas características e a escolha de ferramentas apropriadas. Na indústria ou academia, é crescente a implementação de técnicas de IA, como: Redes Neurais, Reconhecimento de Padrões, AM, em problemas de organização ou recuperação da informação (HASSANI *et al.*, 2020). Essas subáreas de estudo são diretamente ligadas a CI, como apontam Souza, Almeida e Baracho (2013). A mineração de texto tem sido utilizada, seja para análise política de *Tweets*, análise da opinião dos consumidores sobre determinado serviço ou produto, a audiência de um filme ou até mesmo para achar os ativos mais interessantes para determinado tipo de investidor (DAVE; LAWRENCE; PENNOCK, 2003). Independente da área de aplicação, o objetivo geralmente passa por entender um segmento (público) e direcionar decisões de acordo com esse segmento.

2.3 Tecnologia da informação

A Tecnologia da Informação permite a evolução do meio físico para o digital, assim como oferece ferramentas capazes de nortear e ajudar nos processos de gestão da informação. Souza (2006) afirma que a melhoria dos sistemas de organização da informação depende de esforço mútuo de várias linhas de pesquisa, entre eles o Aprendizado de Máquina, uma subárea de IA que vem oferecendo importantes avanços no problema de organização de grandes massas de dados.

2.3.1 Inteligência artificial

As máquinas podem pensar? Essa foi a questão de pesquisa de Alan Turing (1950), que propôs um jogo de imitação capaz de simular o raciocínio humano, exibindo já naquela época, que era possível criar artefatos inteligentes. Ao longo do tempo, houve muito progresso principalmente devido à evolução tecnológica e produção em massa de dados. Mandl (2009) aponta que lidar com esse volume de conteúdo requer métodos inteligentes, e muitas técnicas de IA têm sido aplicadas em sistemas de RI. Robôs e ficção à parte, existe uma forma de IA que vem ganhando atenção da academia e indústria. Uma pesquisa com gestores de diferentes corporações no mundo, conduzido pelo Instituto de Tecnologia de Massachusetts (MIT) e o Google enfatizou que 95% das empresas já implementaram ou pretendem implementar técnicas de IA como forma de melhorar seus processos informacionais e inteligência competitiva (MIT TECHNOLOGY REVIEW; GOOGLE CLOUD, 2017). IA é a

habilidade de treinar um computador ou programa para realizar tarefas tipicamente reservadas à inteligência humana (MEHR, 2017).

Inteligência Artificial foi inicialmente desenvolvida numa tentativa de produzir aplicações com um conjunto de regras e inferência para simular a mente humana, no entanto, a complexidade do processo cognitivo humano sempre foi um obstáculo. Saracevic (1996) classifica a IA dentro da CI como IA Forte (tenta entender como a mente humana processa informação) e IA Fraca (implementa programas e algoritmos para solucionar problemas). Com o *Big Data* e a redução de custo de infraestrutura de computação, evolui o conceito de IA Fraca (LOURIDAS; EBERT, 2016). Nessa nova forma baseada no aprendizado humano por indução, um programa não necessariamente precisa aprender como a mente funciona, e sim aprender baseado em dados e situações empíricas. Essa nova subárea gerou um desmembramento dentro de IA, chamado Aprendizado de Máquina (AM) (DAVENPORT; HARRIS, 2017). IA tem sido aplicada em problemas distintos dentro da CI, como classificação automática da informação, processamento de linguagem natural, mineração de dados e texto ou descoberta de conhecimento.

2.3.2 Aprendizado de máquina

Em Lógica ou Estatística, indução é o raciocínio a partir da observação do treino de casos perante regras gerais, que então são aplicados em casos de teste (DU; SWAMY, 2019). Em outras palavras, é a aplicação do aprendizado adquirido durante um treinamento numa coleção de dados semelhantes, em novos dados para teste ou inferência. A transferência de aprendizado ocorre quando a experiência adquirida durante a execução de uma tarefa é aperfeiçoada em relação a execução de uma outra tarefa (YANG; HANNEKE; CARBONELL, 2013). A ideia geral é que um programa aprende a executar uma tarefa estudando um conjunto de exemplos e, então, executa essa tarefa num conjunto de dados ainda não vistos anteriormente. De certa forma, não há um consenso sobre uma distinção entre IA e AM, mas de fato, essa área tem sido objeto de estudo e aplicação na academia e na indústria.

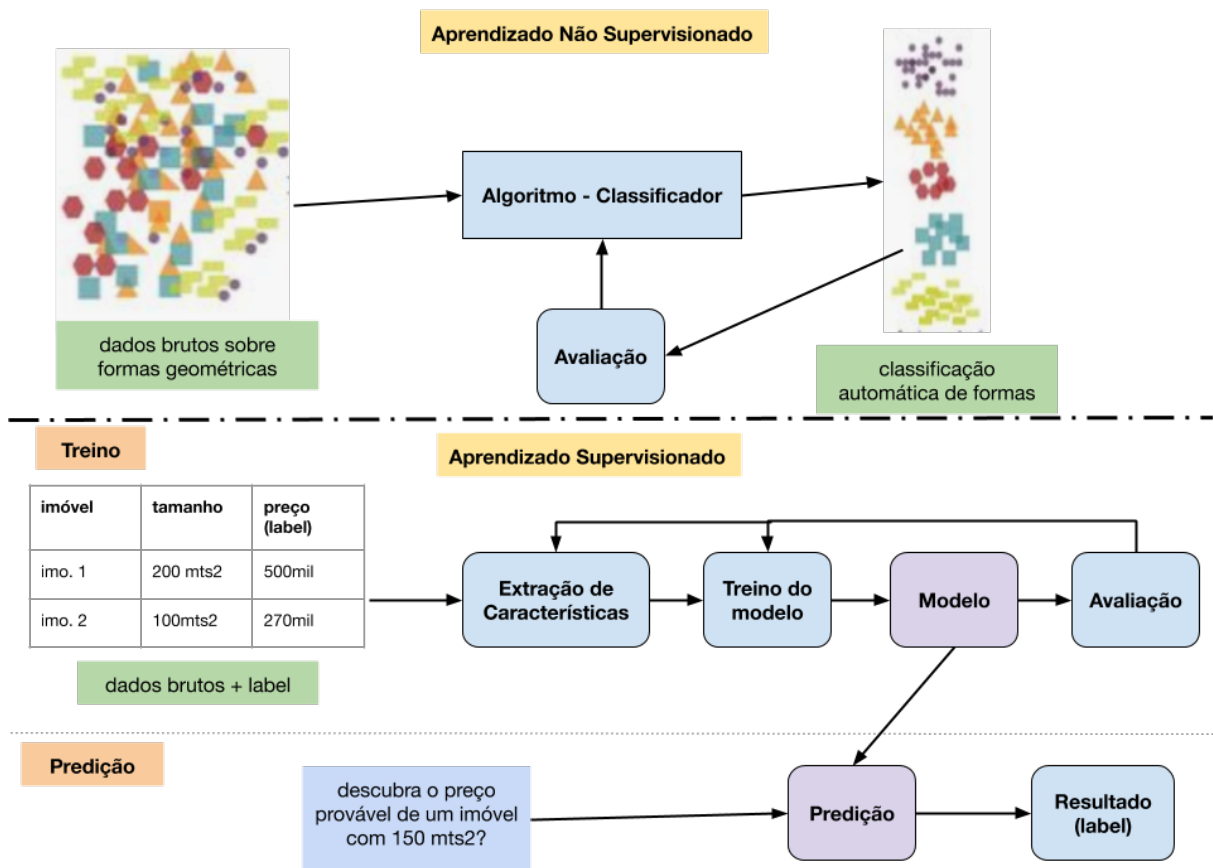
Os modelos de IA ou AM estão ligados em praticamente todas as áreas que podem envolver computação e dados, como segurança, medicina, marketing, economia, turismo, mercado financeiro, judiciário entre outros. Esse fator tem gerado um impacto

profundo em negócios, governo e sociedade (CIOFFI *et al.*, 2020). Dentre as aplicações, é possível destacar assistentes virtuais (Alexa, Google Now, Siri), previsão de tráfego, monitoramento de vídeo (reconhecimento facial), serviços em redes sociais (pessoas que você pode conhecer, filmes e livros similares ao que você gosta), suporte ao usuário (resposta automática às dúvidas), recuperação da informação (Google), recomendação de produtos e serviços, detecção de fraude *online*, dentre diversas outras.

2.3.2.1 Abordagens do aprendizado de máquina

A resposta mais precisa sobre qual o melhor algoritmo seria: depende do objetivo do estudo, dos dados, dos recursos computacionais e dos métodos. Geralmente os algoritmos são classificados em dois tipos: supervisionados e não supervisionados (SEADLE, 2007). A Figura 2 exibe a distinção entre esses dois tipos.

Figura 2 – Tipos de Aprendizado de Máquina



Fonte: elaborado pelo autor (2021).

No aprendizado não supervisionado, existe a tentativa de associar os dados brutos por meio de uma redução intrínseca de sua dimensionalidade ou clusterização, procurando agrupar os dados de acordo com suas características. Nesse modelo, não há dados previamente classificados, ou seja, é um processo todo automático de classificação e organização. Já no aprendizado supervisionado, os dados brutos possuem classificação e características previamente definidas. Com base nessas características, um modelo de classificação ou regressão é gerado e aplicado em dados existentes, ou seja, esses algoritmos aprendem com base nos exemplos. O modelo então pode ser aplicado para classificar dados ainda não classificados. O objetivo da regressão geralmente é prever valores, como: o preço provável de um imóvel, como demonstrado na Figura 2. O objetivo da classificação é agrupar textos, documentos, objetos, imagens, como: classificar a área de conhecimento de um conjunto de documentos científicos (LOURIDAS; EBERT, 2016). O Quadro 1 exibe alguns exemplos de algoritmos separados de acordo com o tipo.

Quadro 1 – Métodos de aprendizado de máquina

APRENDIZADO DE MÁQUINA			
SUPERVISIONADO		NÃO SUPERVISIONADO	
CLASSIFICAÇÃO	REGRESSÃO	CLUSTERIZAÇÃO	REDUÇÃO DA DIMENSIONALIDADE
Árvores de classificação	Árvores de decisão	Clusterização hierárquica	Decomposição tensorial
Florestas randômicas	Classificação difusa	Algoritmos genéticos	Alocação de Dirichlet Latente
Máquinas e vetores de suporte	Redes Bayesianas	Modelos de mistura gaussiana	Estatística multidimensional
Regressão logística	Regressão linear	Clusterização K-means	Análise de componentes principais

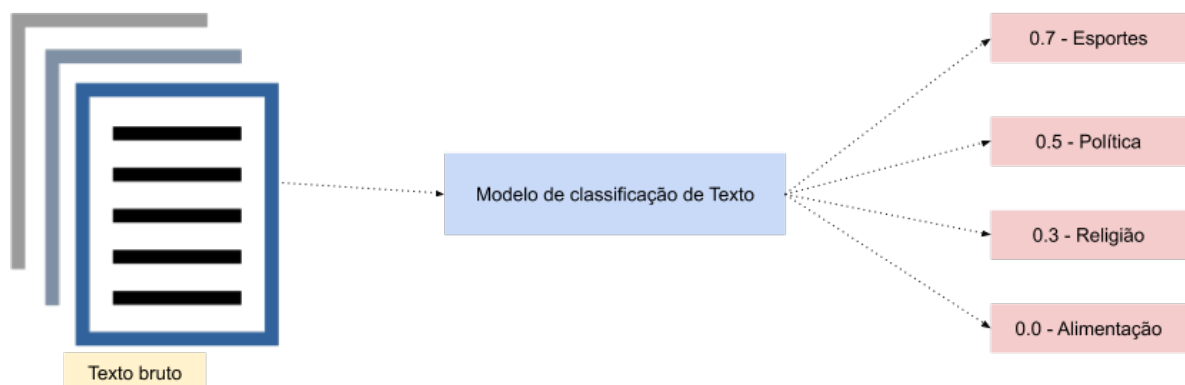
Fonte: elaborado pelo autor (2021).

Esses métodos da computação possuem ligação com a CI, pois, boa parte do processo de AM compreende as etapas como: extração de dados, limpeza dos dados, transformação dos dados, organização ou extração de características. Somente após essas etapas é realizado o processamento dos algoritmos, melhorias nos modelos e avaliação dos resultados (DODDS, 2020). Souza (2020) aponta que tarefas como Similaridade de documentos, Classificação de documentos, Análise de sentimento, Classificação de imagens, Modelagem de tópicos, Mineração de textos e dados e RI são tarefas típicas de CI que geralmente podem ser automatizadas com algoritmos de AM.

2.3.3 Classificação automática de texto

Sobre os processos de Organização do Conhecimento, Smiraglia e Cai (2017) apontam que o termo Indexação Automática é quase sempre relacionado ao termo Classificação Automática. Os autores identificam a ênfase da área com RI e que os termos Clusterização e Classificação Automática têm sido usados por muitos autores para classificar seus trabalhos dentro da área de Organização do Conhecimento. Por muito tempo usou-se o termo Documento como forma de descrever o processo de classificação. Isso se deve principalmente pela origem do problema, ou seja, a necessidade de se criar métodos que permitissem que livros e revistas fossem classificados. No entanto, com a evolução desses documentos para o formato digital, houve também a criação de outras formas de documentos, como por exemplo: um *post* na rede social, uma avaliação de uma atração turística, um comentário em um *blog*. A classificação de textos compreende a associação de uma ou mais classes (rótulos) ou categorias, com um determinado documento (texto). Salles *et al.* (2017) apontam que a classificação de texto é um dos maiores problemas no processo de RI, haja vista as características não estruturadas e o volume de dados na Web. A Figura 3 exibe um processo básico de Classificação Automática de Texto.

Figura 3 – Processo básico de classificação automática de texto



Fonte: elaborado pelo autor (2021).

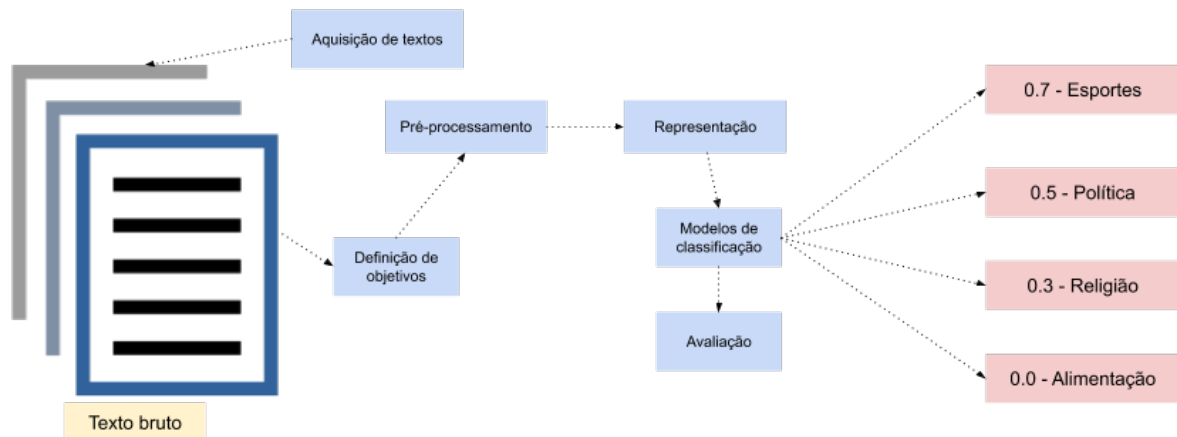
No processo básico, cada texto (documento) é associado com um grau de pertinência a um conjunto de classes. Essa pertinência indica a probabilidade daquele texto fazer parte daquela classe. Forman (2003) aponta que o processo de classificação de texto está inserido dentro do processo de mineração de texto.

A Mineração de Texto (MT) refere-se a descoberta de conhecimento não trivial, antes não conhecido e potencialmente importante de uma coleção de textos. Conforme apontam Altinel e Ganiz (2018), a partir de MT é possível classificar textos, filtragem de documentos, sumarização, análise de sentimento/opinião. Esse conceito é análogo ao de mineração de dados, sendo que quase sempre seu objetivo conceitual é a descoberta de conhecimento. Embora existam pequenas diferenças, os conceitos realmente são muito similares. Hotho, Nürnberger e Paaß (2005) aponta que é possível diferenciar três perspectivas de mineração de texto: a) extração de informação, b) mineração de dados, e c) descoberta de conhecimento. Como boa parte do conteúdo na web é expresso em texto não estruturado, as aplicações em MT são muitas, como: Governo e grupos militares usam MT como forma de monitorar cidadãos por questões de segurança; políticos usam MT para entender eleitores e determinar melhor seu formato de comunicação; aplicações biomédicas para estratificar e indexar eventos clínicos comuns em pacientes com os mesmos sintomas e efeitos colaterais; Opiniões de usuários são mineradas para identificar aspectos importantes ao negócio, suporte ao usuário, adequação e melhoria dos serviços ofertados.

De acordo com Alvarenga (2003), uma classificação de documento a priori requer tempo e esforço de especialistas, no entanto, o próprio documento possui características que, se extraídas corretamente, permitem classificá-lo perante uma coleção. Dada sua natureza não estruturada e o volume de informações na Web, as técnicas de Classificação Automática de Texto oferecem uma alternativa importante na organização do conhecimento. Métodos como Processamento de Linguagem Natural (PLN), AM e Mineração de Dados são aplicados em conjunto de forma a identificar padrões nos textos e permitir sua classificação.

Independente do modelo de classificação escolhido, geralmente as etapas envolvidas no processo compreendem: Aquisição dos textos, Definição de objetivos e métodos, Pré-processamento, Representação dos textos, Classificação e Avaliação. A Figura 4 exhibe o processo completo de classificação automática.

Figura 4 – Processo de classificação automática



Fonte: elaborado pelo autor (2021).

2.3.3.1 Aquisição dos textos

A aquisição dos dados (textos) depende do local e formato que estão armazenados. A coleção de textos geralmente é conceituada como *corpus*. Os textos podem estar armazenados em documentos PDF, num banco de dados relacional ou na Web. Como a maior parte dos documentos está disponível na Web, o uso de *Crawlers* tem se mostrado com uma opção interessante para coletar esses dados. Esses *Crawlers* podem ser usados para extrair textos no meio de marcações HTML. De posse dos dados, os mesmos podem ser salvos em arquivos XML, CSV ou banco de dados para que possam ser manipulados posteriormente pelos algoritmos. A Figura 5 exibe um exemplo de um pequeno *Crawler* criado na linguagem Python para extrair automaticamente texto do site www.harvard.edu.

Figura 5 – Exemplo de *crawler*

```

1 import requests
2 from bs4 import BeautifulSoup
3
4 url = 'https://www.harvard.edu/'
5 html = requests.get(url).text
6 soup = BeautifulSoup(html, "html.parser")
7 corpus = [d.get_text().strip() for d in soup.find_all('div', attrs={'class': 'card_text'})]
8 print(corpus)

```

```

['Harvard expert Olga Jonas compares the 1918 influenza pandemic and the current coronavirus pandemic', 'A Harvard researcher says that during the coronavirus pandemic dreams full of bugs, masks, and natural disasters are common', 'Alumni offer the Class of 2020 advice on the value and future of a career giving back to communities', 'For Kennedy School graduate Ingrid Olea, a journey that started with a career change has led to achievements in education policymaking', 'With a residency in oral and maxillofacial surgery, Dental School graduate Jeffrey Taylor hopes to one day fix life-altering facial deformities', 'For her senior project, graduate Cathy Wang sought to help accelerate tendon healing and engineered a degradable hydrogel scaffold', 'Growing up in Kenya, graduate Billy Koech had the opportunity to take an electricity class, which inspired him to study electrical engineering at Harvard']

```

Fonte: elaborado pelo autor (2021).

2.3.3.2 Definição de objetivos e métodos

Nessa etapa é importante definir o escopo de estudo, os métodos que serão utilizados, algoritmos e se haverá ou não a participação de especialistas no processo. Os métodos de classificação automática seguem a linha dos tipos de algoritmos de AM conforme descrito na seção 2.3.2.1, podem ser supervisionados ou não supervisionados. Nessa etapa, é importante avaliar a participação de especialistas no processo. Dickman (2003) identificou melhor desempenho nos processos de classificação ao envolver a participação de especialistas do domínio. Chen, Müller e Stenberg (2006) apresentam um trabalho de classificação de texto em uma taxonomia, a partir de regras criadas com participação humana. Os resultados se mostram superiores aos métodos puramente automáticos de classificação.

2.3.3.3 Pré-processamento dos textos

A etapa de pré-processamento usa PLN para limpar e extrair do texto somente o que é importante e relevante para o objeto em estudo. O objetivo principal é trazer mais qualidade aos dados, limpando inconsistências e eliminando valores que possam comprometer o resultado do estudo. Acuña (2011) classifica essa etapa como responsável por trazer mais desempenho e eficiência às etapas seguintes do processo de classificação. Essa etapa pode envolver ações como:

- a) remoção de *stop-words* como preposições, conjunções, artigos e advérbios que não contribuem para o significado semântico do texto;
- b) limpeza de dados com erro, remoção de duplicados, incompletos, ausentes;
- c) remoção de verbos e adjetivos depende do escopo de estudo pode ser interessante (BAEZA-YATES; RIBEIRO-NETO, 2013);
- d) remoção de acentos e caracteres especiais;
- e) transformação de palavras do plural para singular e maiúscula em minúscula;
- f) transformação dos textos em *tokens*, ou seja, termos separados;
- g) stemização para reduzir a palavra a sua raiz (radical), exemplo: sonhador => sonh (nesse caso a palavra reduzida pode não ser parte do idioma);
- h) lematização para reduzir a palavra ao seu lemma, exemplo: foi => ir (nesse caso sempre retorna uma palavra válida do idioma);

- i) bigramas e Trigramas referem-se a união de palavras que ocorrem com frequência em conjunto, exemplo: vale_pena ou compra_ingresso_internet;
- j) separação dos dados entre treino e teste.

Essas técnicas de PLN são muito utilizadas em estudos que envolvem classificação ou recuperação automática de informação. Dependendo do domínio de aplicação, etapas como a extração de características dos dados e normalização de valores podem ser aplicadas, com o objetivo de facilitar a representação e comparação (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006). A Figura 6 exibe um exemplo de pré-processamento utilizando a linguagem Python.

Figura 6 – Exemplo pré-processamento utilizando Python

```

1 doc = corpus_atracoas[0][3][13]
2 print(doc)

('Um banho de arte Museu de acervo muito rico e diversificado, pois contempla vários estilos e tipos de arte produzidos em épocas remotas (algumas com milhares de anos) e culturas diferentes (como a chinesa e n oa), além de concentrar obras de artistas renomados, como Van Gogh, Degas, Gauguin, Monet,...',)

1 doc_clean = corpus_label_proc['Cultura'][13]
2 print(doc_clean)

['arte', 'museu', 'acervo', 'rico', 'arte', 'epoca', 'remota', 'cultura', 'chinesa', 'noa', 'artista', 'van_gogh', 'dega', 'gauguin', 'monet']

```

Fonte: elaborado pelo autor (2021).

Na Figura 6 pode-se observar o pré-processamento do documento doc, gerando a saída doc_clean. É possível observar que artigos, caracteres especiais, acentos, preposições foram removidos. O bigrama van_gogh foi criado por ocorrer não somente nesse documento, mas sim com frequência na coleção. O texto foi transformado em tokens que são usados nas próximas etapas do processo de classificação. O conjunto de diferentes palavras obtido pela junção de todos os documentos em uma coleção (*corpus*) é chamado de Dicionário (HOTHO; NÜRNBERGER; PAAß, 2015).

2.3.3.4 Representação dos textos

Um texto não estruturado contém uma vasta quantidade de informações, dados e conhecimento intrínseco. Diferente de um banco de dados, que possui uma estrutura definida, nesse caso, é necessário criar uma estrutura. Algoritmos e computadores conseguem manipular apenas números, ou seja, é necessário representar o texto de forma que seja possível que um algoritmo consiga processá-lo. “[...] Assim, a

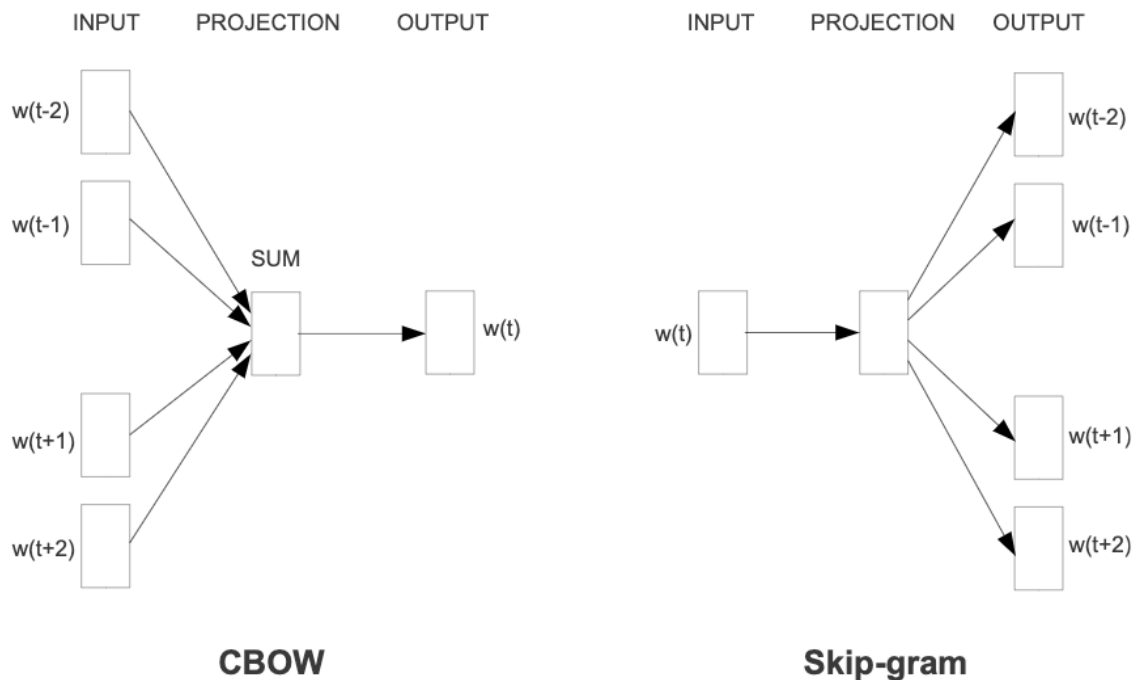
informação precisa ser representada de maneira que possa ser posteriormente manipulada e extraída por processos automatizados, o que exige que a mesma seja convertida em uma estrutura lógica” (LADEIRA, 2010, p. 17).

Essa representação visa não somente transformar texto em número, como também identificar o peso dos termos perante um documento e coleção e identificar a similaridade entre eles (SONG; LIU; YANG, 2005). Além disso, conforme provado por Song, Liu e Yang (2005), o termo é a melhor unidade para representação e classificação.

Os métodos mais comuns são *Bag of Words* (BOW), Termo Frequência (TF) e Termo Frequência – Frequência Inversa do Documento (TF-IDF). Esses métodos consideram o modelo espaço vetor, ou seja, é criado um vetor representando a frequência de cada termo dentro de um documento. Enquanto o método BOW apenas conta a quantidade de palavras, o método TF atribui peso para a quantidade de termos considerando o tamanho do documento. Já o método TF-IDF atribui peso para os termos considerando sua frequência no documento e sua importância na coleção (QAISER; ALI, 2018). Harish, Guru e Manjunath (2010) relatam que tais métodos possuem como desvantagem a dimensão da representação e a perda de semântica entre os termos.

Como forma de resolver o problema semântico das representações supracitadas, Mikolov *et al.* (2013) apresentam uma representação usando Rede Neural com o conceito de casamento de termos (*word embeddings*). Como resultado dessa pesquisa, o algoritmo *Word2Vec* foi desenvolvido, que permite diversas aplicações de PLN. Com esse algoritmo, cada termo na verdade se torna um vetor, mantendo sua relação com as palavras mais próximas, podendo assumir uma abordagem CBOW ou Skip-gram. A Figura 7 exibe a diferença entre os dois modelos.

Figura 7– Diferença entre abordagens CBOW e Skip-Gram



Fonte: Mikolov *et al.* (2013, p. 5).

Na Figura 7, é possível ver que no modelo CBOW, o termo $w(t)$ sendo projetado com base nos seus dois termos anteriores $w(t-2)$, $w(t-1)$ e nos dois posteriores $w(t+1)$, $w(t+2)$, ou seja, considerando uma janela fixa de 2. Já no modelo Skip-gram é possível ver que a partir do termo $w(t)$, o algoritmo procura encontrar o termo semelhante, com base nos seus dois termos anteriores $w(t-2)$, $w(t-1)$ e nos dois posteriores $w(t+1)$, $w(t+2)$, considerando também uma janela fixa de 2. Como a representação mantém a proximidade entre os termos, é possível realizar operações matemáticas entre os termos, como exemplo: Rei – Homem + Mulher resultaria em Rainha. Para que se atinja esse nível de precisão, é necessário ter uma massa grande de dados e realizar um bom pré-processamento. A Figura 8 exibe os termos próximos à Esqui com base numa representação usando *word2vec*.

Figura 8 – Termos próximos à Esqui

```

1 word = 'esqui'
2 model.most_similar(word,topn=5)

[('snowboard', 0.933731198310852),
 ('ski', 0.9156543016433716),
 ('esqui_snowboard', 0.8937780857086182),
 ('ski_snowboard', 0.8795595169067383),
 ('snowboard_esqui', 0.8778543472290039)]

```

Fonte: elaborado pelo autor (2021).

Independentemente do método utilizado, uma vez que os textos de um *corpus* estejam representados, se torna possível aferir a similaridade entre eles, usando medidas como cosseno do ângulo dos vetores de representação de cada texto. A partir das representações, é possível realizar também uma classificação supervisionada ou não supervisionada dos textos.

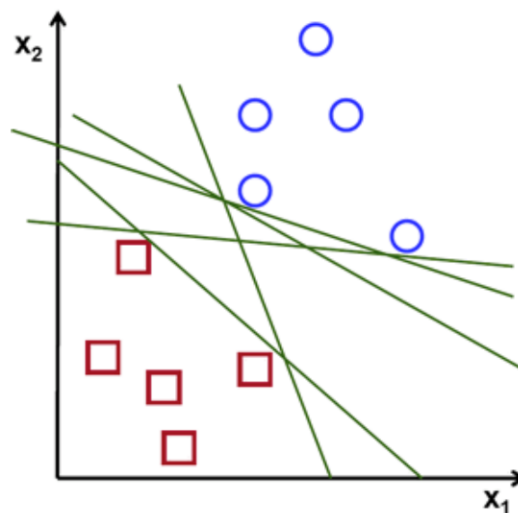
2.3.3.5 Modelo de classificação supervisionado

Os modelos automáticos de classificação de texto supervisionados implementam técnicas de AM supervisionadas. De acordo com Salles *et al.* (2010), primeiro um modelo é construído baseando-se em algum algoritmo como Naive Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression ou outro mais apropriado para o estudo. Em seguida, esse modelo é treinado em um conjunto de textos classificados (com classes), numa fase conhecida como treino. Em seguida, o modelo é aplicado a um conjunto de textos ainda não classificados (sem classes), ainda não conhecido pelo modelo, numa fase conhecida como predição. A tarefa principal é construir um modelo capaz de estimar e prever a(s) classe(s) de um objeto avaliando suas características (NASTESKI, 2017).

2.3.3.5.1 Algoritmos supervisionados

Algoritmos discriminativos usam distribuição de probabilidade condicional, ou seja, eles tentam classificar diretamente os textos em classes, enquanto os algoritmos generativos usam distribuição de probabilidade conjunta, ou seja, primeiro tentam encontrar predições de probabilidade dos textos perante as classes e então escolhem a classe mais provável para representar o texto (NG; JORDAN, 2001).

O algoritmo SVM é um algoritmo discriminativo e procura encontrar um hiperplano num espaço dimensional N (N = número de classes) que separe os textos entre as classes. O hiperplano com a maior distância entre os elementos classificados e separados exibe a melhor configuração de classificação possível para o modelo. O Gráfico 2 exibe um exemplo de hiperplanos encontrados na classificação de textos em duas classes (Círculo e Quadrado).

Gráfico 2 – Hiperplanos do algoritmo *Support Vector Machine* (SVM)

Fonte: elaborado pelo autor (2021).

De acordo com Harish, Guru e Shantharamu (2010), o aprendizado no SVM quase independe do espaço dimensional (quantidade de classes), assim, esse método apresenta bom desempenho para classificações de textos em muitas classes. O método Naive Bayes é um método de classificação generativo baseado no Teorema de Bayes. Uma das características desse algoritmo é que ele assume que cada termo em um texto é independente de outro. Exemplo: “Festa estava animada” é o mesmo que “Animada essa festa” (NASTESKI, 2017). O método usando Logistic Regression é similar ao Naive Bayes, porém discriminativo, ou seja, o algoritmo realiza uma regressão para identificar binariamente se um texto pertence a uma classe ou não (NG; JORDAN, 2001). Já o método Random Forest é discriminativo e cria várias árvores de decisão combinando-as para obter uma predição com maior acurácia. De acordo com Esteves (2016), para classificar um novo texto, o mesmo é testado com todas as árvores geradas. Cada árvore gera uma classificação, e a classificação que gera o maior número de votos é considerada correta para aquele texto.

2.3.3.5.2 Avaliação da classificação

As métricas mais conhecidas para validar um sistema de classificação e recuperação da informação são *Precision* e *Recall*. *Recall* se refere a habilidade de retornar todos os resultados relevantes, enquanto *Precision* representa a fração dos documentos corretos perante os retornados. O conceito pode ser exemplificado como: imagine um sistema de classificação que procure encontrar cavalos identifique 8 elementos numa

lista contendo 12 cavalos e burros. Dos oito elementos, cinco realmente são cavalos e três são burros, dessa forma, seu valor de *Precision* então é 5/8, ou seja, 62,5%, enquanto seu valor de *Recall* é 5/12, ou seja, 41,66%.

De acordo com Chen, Müller e Stenberg (2006), uma medida muito utilizada para avaliar os sistemas de classificação supervisionados de texto é o *F-measure*, pois ela combina *Precision* e *Recall* e representa uma média harmoniosa entre ambas métricas, sendo que seu resultado está sempre entre 0 e 1, e quanto mais próxima a 1, melhor a acurácia do modelo de classificação. A Equação 1 apresenta a fórmula para se calcular *F-measure*.

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (1)$$

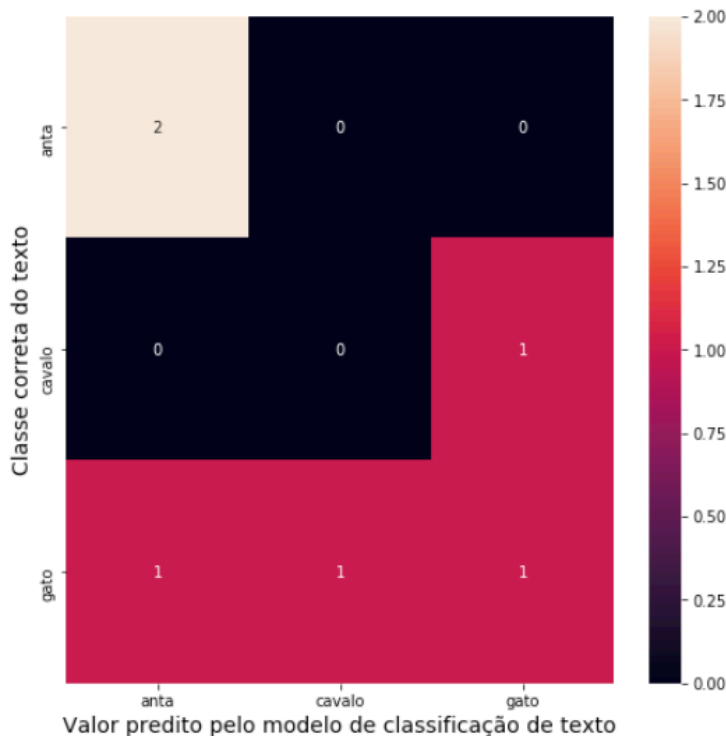
Uma forma comum de se avaliar os sistemas de classificação de texto é separando os dados em treino e teste, num método conhecido como validação cruzada. Nesse caso, o modelo treina apenas em textos com classes definidas e somente conhece os demais textos no momento do teste (GEISSER, 1993). Esse é um importante fator dos sistemas de classificação automática, haja vista que um teste feito em textos já treinados poderia comprometer completamente a qualidade do resultado (COUSSEMENT, 2014). A partir desse teste é possível extrair *Precision*, *Recall*, *F-measure*, ou Matriz de Confusão, que é uma medida visual de acurácia do modelo de classificação que permite visualizar um comparativo entre acertos e erros do modelo por classe. O Gráfico 3 exibe um exemplo de uma Matriz de Confusão.

Gráfico 3 – Hiperplanos do algoritmo *Support Vector Machine* (SVM)

```

1 from sklearn.metrics import confusion_matrix
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4 classes_corretas = ["gato", "anta", "gato", "gato", "anta", "cavalo"]
5 classes_preditas = ["anta", "anta", "gato", "cavalo", "anta", "gato"]
6 fig, ax = plt.subplots(figsize=(8,8))
7 sns.heatmap(confusion_matrix(classes_corretas, classes_preditas), annot=True, fmt='d',
8             xticklabels=["anta","cavalo","gato"], yticklabels=["anta","cavalo","gato"])
9 plt.ylabel('Classe correta do texto',fontsize=14)
10 plt.xlabel('Valor predito pelo modelo de classificação de texto',fontsize=14)
11 plt.show()

```



Fonte: elaborado pelo autor (2021).

No Gráfico 3 percebe-se que o conjunto de classes corretas dos textos `classes_corretas` e o conjunto de classes que foram preditas durante a classificação por determinado algoritmo `classes_preditas`. A Matriz de Confusão oferece uma visualização interessante, pois pode-se identificar que duas antas foram identificadas corretamente como antas, um gato foi identificado como anta, um gato como cavalo, um gato como gato e um cavalo como gato. Dessa forma, é possível identificar os acertos e erros do processo de classificação em nível de classe.

2.3.3.6 Modelo de classificação não supervisionado

No modelo de classificação não supervisionado não há dados já previamente classificados, ou seja, o algoritmo precisa encontrar padrões agrupando os dados com

características semelhantes em *clusters*. Louridas e Ebert (2016) relatam que existem dois tipos: os algoritmos de *clustering*, que particiona um *corpus* entre vários grupos “*clusters*” satisfazendo certo critério; E os algoritmos de redução da dimensionalidade, que tentam captar os aspectos fundamentais dos dados com base num processo de redução de suas dimensões. Du e Swamy (2019) afirmam que os métodos não supervisionados permitem uma análise puramente baseada na correlação entre os dados, um processo de análise linguística para encontrar padrões ou características sem a ajuda de um especialista.

Essa forma de classificação tem sido importante, devido à natureza não estruturada de boa parte do conteúdo na Web e principalmente o custo envolvido nos processos manuais de classificação. A natureza desse processo é também de descoberta, pois, ao passo que milhões de dados e informações são avaliados é possível descobrir padrões, características imperceptíveis ao olho humano. Existem diferentes aplicações dos métodos não supervisionados, como: detecção de anomalias, reconhecimento de entidade, reconhecimento de imagens, segmentação de mercado, sistemas de recomendação, para sugerir itens relevantes para os usuários, como livros, filmes ou produtos a comprar ou a modelagem de tópicos dos textos de um corpus.

2.3.3.6.1 Modelagem de tópicos

A modelagem de tópicos envolve geralmente abordagem não supervisionada e análise linguística de um determinado *corpus*. O método *Latent Dirichlet Allocation* (LDA), é um método probabilístico de PLN completamente não supervisionado que modela cada documento como um conjunto de tópicos. Conforme aponta Ramage *et al.* (2009), um texto frequentemente refere-se a mais de um tópico (assunto). Knispelis (2016) mostra que a modelagem de um texto com LDA envolve o entendimento sobre os termos, o contexto e os assuntos daqueles textos. LDA procura observar os textos para inferir uma estrutura de tópicos escondida (BLEI, 2012). Esse modelo tem sido aplicado com sucesso na análise textual de avaliações de usuários em portais, seja para análise linguística de texto ou análise de sentimento, como demonstrado por Nawangsari, Kusumaningrum e Wibowo (2019) e McKenzie e Adams (2018). A partir dos resultados do processamento LDA, é possível descobrir os principais termos e tópicos que compõem determinado corpus e ainda comparar a similaridade entre os

tópicos. A Figura 9 exibe um exemplo de nuvens de palavras (as mais importantes) dentro de cada tópico de um modelo LDA gerado.

Figura 9 – Nuvens com importância de palavras para os tópicos



Fonte: elaborado pelo autor (2021).

No primeiro tópico formado, é possível visualizar que muitos documentos foram agrupados em características relacionadas a Arquitetura, já no segundo é possível observar a presença de termos mais ligados a Museu, como coleção, artista, pintura e escultura. Naturalmente esses termos tendem a ocorrerem juntos dentro do mesmo contexto. Termos com mais frequência assumirão posições mais proeminentes dentro de cada tópico (BLEI, 2012).

2.4 Informação no Turismo

Ortega (2009) afirma que o dilema de Hamlet no turismo não é: “ser ou não ser” e sim, “aonde ir, como ir e o que fazer”. Analogia a parte, o fato é que essas ações necessitam de um insumo comum para tomada de decisão: a informação. A tecnologia vem exercendo um impacto importante no crescimento contínuo do turismo perante o PIB mundial. A Internet tem sido um vetor importante nesse crescimento, principalmente por permitir que mais pessoas tenham mais acesso à informação. Conforme apontado no relatório do Conselho Mundial de Viagem e Turismo (WTTC) de 2019, o aumento da faixa de renda e o acesso à informação são fatores essenciais a esse crescimento. Ao planejar uma viagem, turistas usam diferentes tipos de informações em ações como: a escolha do destino, a escolha do local de estadia, a escolha de atrações turísticas a visitar, a consulta de condições climáticas, a escolha do meio de transporte, a consulta sobre moeda, hábitos e cultura local. Sites como Yelp, Google, TripAdvisor, Booking, Airbnb, Trivago ajudam turistas nesse processo decisório. O estudo do Google (2014) sobre o impacto da Internet no Turismo, aponta

que 65% dos respondentes usam a Internet como fonte de inspiração para suas viagens, sendo que 42% da inspiração ocorre a partir de sites e aplicativos de avaliações de turismo. Pesquisas anteriores também reforçam que os turistas se orientam para sua viagem com base em informações de visitantes anteriores (GRETZEL; YOO, 2008; VERMEULEN; SEEGERS, 2009).

2.4.1 Avaliações no turismo

O termo boca-a-boca “word of mouth” (WOM) sempre um dos fatores importantes no processo decisório de clientes com relação a compra de determinado produto ou serviço. De acordo com a Nielsen Company (2012), 92% das pessoas no mundo dizem confiar mais na recomendação de amigos e família sobre qualquer outra forma de marketing. Com as plataformas de compartilhamento de informações entre usuários em plataformas digitais criou o termo (e-WOM), ou seja, “boca-a-boca digital”. Assim como em outras áreas, a quantidade de dados e informações referentes a hotéis, restaurantes, voos, turistas, pontos turísticos cresce exponencialmente. Usuários expressam seu grau de satisfação com determinado serviço, produto ou Ponto de Interesse (POI) dentro das plataformas de compartilhamento digital.

O estudo de Chevalier e Mayzlin (2006) avalia o impacto de WOM na compra de livros no site Amazon.com. Os autores evidenciam que o número de avaliações, bem como sua característica positiva ou negativa impacta diretamente no número de vendas de determinado livro. O estudo de Gupta e Harris (2010) mostra que consumidores mais motivados a ler e processar informação, estão mais propensos a tomar decisões com base nessas recomendações; já os consumidores com menos motivação para processar as informações, tomam menos decisões baseadas no conteúdo do e-WOM. Essa descoberta oferece um *insight* importante, pois, se há consumidores menos propensos a consumir o conteúdo, essa realidade pode se tornar pior se observarmos a quantidade enorme de avaliações existentes.

As avaliações são feitas em restaurantes, hotéis, atrações turísticas, companhias aéreas ou pode ser feita em qualquer outro tipo de experiência que um turista vivencia durante sua viagem. Esse conteúdo não é rico apenas para usuários, mas importante para que empresas e governo entendam a opinião dos seus consumidores e consigam, a partir disso, adequar produtos e serviços com essa oferta.

Consumidores usando essas avaliações se confrontam com um número grande de informação, muitas vezes conflitantes (SEN; LERMAN, 2017), sendo que sua percepção, intenção e decisão depende desse conteúdo e da forma como é apresentado (SPARKS; PERKINS; BUCKLEY, 2013). Por exemplo, a atração Central Parque em Nova Iorque possui 205 mil avaliações no Google, e 132 mil avaliações no TripAdvisor. Há muita informação relevante e conhecimento dentro desse conteúdo, no entanto, para que o usuário faça uso efetivo, é necessário técnicas para simplificar o seu uso.

2.4.2 Avaliações em uma plataforma de turismo: o exemplo do TripAdvisor

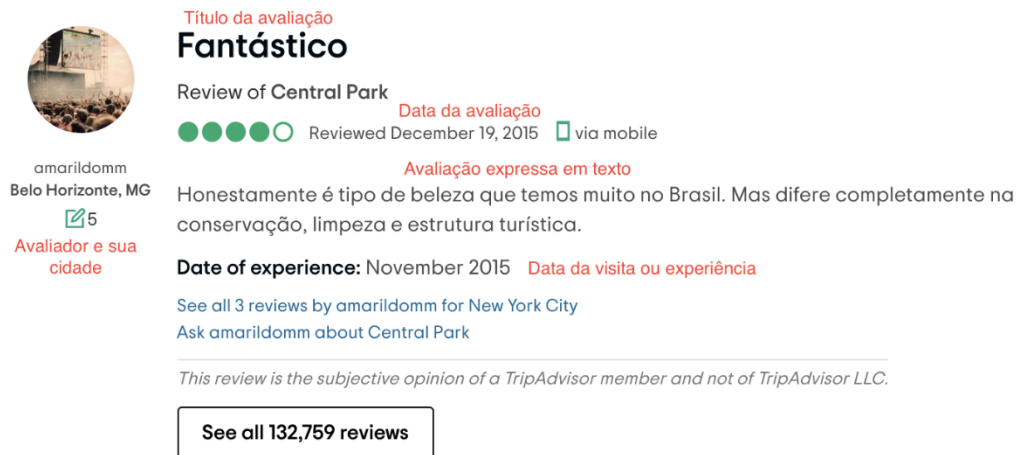
O TripAdvisor possui cerca de 54 milhões de visitas por mês, sendo que 99,13% desse acesso é orgânico, ou seja, os usuários chegaram nas páginas do site a partir de conteúdo e não marketing pago na Internet⁶. Com cerca de 570 milhões de avaliações cobrindo 7,3 milhões de atrações turísticas, hotéis, companhias aéreas e restaurantes, o site explora esse conteúdo para oferecer serviços aos usuários⁷. Isso reforça a importância desse site como fonte de conteúdo para turistas, empresas e governo. Oliveira (2017) aponta que os usuários podem comparar preços, reservar hotéis, casas de temporadas, voos e passeios de forma online e principalmente avaliar produtos e destinos turísticos a partir de seu conteúdo.

Quando uma avaliação é feita, ela passa por uma triagem da equipe do TripAdvisor, que pode aprovar ou não determinada avaliação. Embora não seja possível identificar se esse processo é feito por especialistas ou por algoritmos, trata-se de um importante fator para dar legitimidade ao conteúdo do site. Conforme aponta Oliveira (2017), o site incentiva o processo de compartilhamento de conteúdo por meio de gameificação, ou seja, atribuição de pontos de acordo com o conteúdo gerado pelo usuário no site. A Figura 10 exibe um exemplo de avaliação dentro do TripAdvisor.

⁶ Disponível em: <https://www.similarweb.com/website/tripadvisor.com#search>. Acesso em: 23 maio 2020.


⁷ Disponível em: <https://expandedramblings.com/index.php/tripadvisor-statistics/>. Acesso em: 23 maio 2020.

Figura 10 – Avaliação de usuário do TripAdvisor



Título da avaliação
Fantástico

Review of Central Park

Data da avaliação
Reviewed December 19, 2015 

Avaliação expressa em texto

Honestamente é tipo de beleza que temos muito no Brasil. Mas difere completamente na conservação, limpeza e estrutura turística.

Date of experience: November 2015 **Data da visita ou experiência**

[See all 3 reviews by amarildomm for New York City](#)
[Ask amarildomm about Central Park](#)

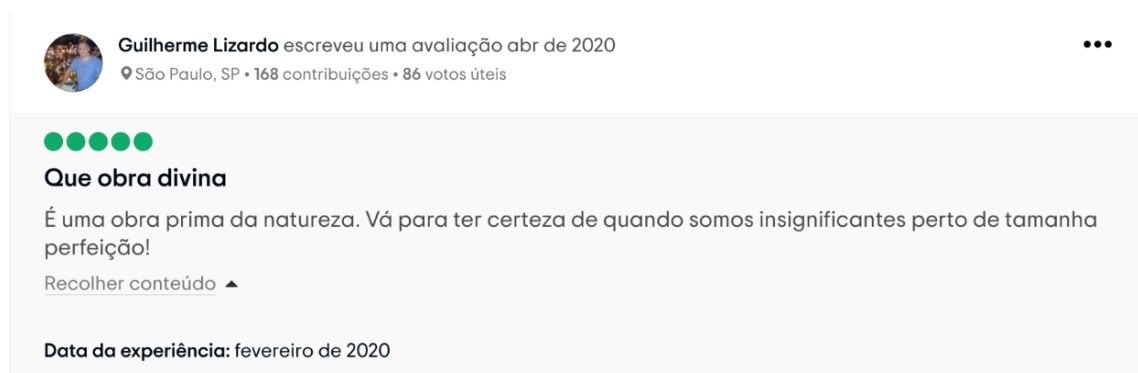
This review is the subjective opinion of a TripAdvisor member and not of TripAdvisor LLC.

[See all 132,759 reviews](#)

Fonte: TripAdvisor⁸.

Cada avaliação é composta por título, quantidade de estrelas 1-5 (sendo 1 = horrível, 2 = ruim, 3 = razoável, 4 = muito bom e 5 = excelente), data da avaliação, data de experiência ou visita e a própria avaliação em si, podendo incluir também imagens e vídeos. O site guarda a informação do usuário e de sua cidade junto com a avaliação. Os títulos das avaliações podem possuir até 121 caracteres, enquanto as avaliações possuem apenas tamanho mínimo de 100 caracteres, ou seja, com isso o tamanho dos textos das avaliações pode variar muito. Embora essa variação possa ocorrer, conforme aponta Korfiatis, García-Bariocanal e Sánchez-Alonso (2012), a facilidade de leitura do texto possui mais influência do que o seu tamanho. As Figura 11 e Figura 12 exibem essa diferença de tamanho.

Figura 11 – Avaliação com texto pequeno



Guilherme Lizardo escreveu uma avaliação abr de 2020

📍 São Paulo, SP • 168 contribuições • 86 votos úteis

Que obra divina

É uma obra prima da natureza. Vá para ter certeza de quando somos insignificantes perto de tamanha perfeição!

[Recolher conteúdo](#) ▲


Data da experiência: fevereiro de 2020


Fonte: TripAdvisor⁹.

⁸ Disponível em: https://www.tripadvisor.com.br/Attraction_Review-g303444-d312332-Reviews-Iguazu_Falls-Foz_do_Iguacu_State_of_Parana.html. Acesso em: 20 maio 2020.

⁹ Disponível em: https://www.tripadvisor.com.br/ShowUserReviews-g60763-d105127-r333687617-Central_Park-New_York_City_New_York.html. Acesso em: 5 maio 2020.

Figura 12 – Avaliação com texto grande

 **@turistainiciante** escreveu uma avaliação 18 de mai
 📍 Juazeiro do Norte, CE • 1.063 contribuição • 32 votos úteis



●●●●●

Fantástico

As cataratasdoiguacu é uma das sete maravilhas da natureza, e o título não é à toa, as Cataratas são incríveis o lugar é de uma beleza sem tamanhos 😍 . As Cataratas ficam dentro do Parque Nacional do Iguazu, que tem uma área de preservação gigantesca, é tudo muito lindo lá dentro 😍 . Dentro do parque tem vários pontos onde você pode ver as Cataratas, tem alguns pontos onde você fica bem próximo das quedas de água, é tudo muito emocionante 😍 . Na entrada do Parque você pega um ônibus que te leva até as @cataratasdoiguacu, ele para em um determinado ponto, depois é só ir caminhando e ir apreciando toda a beleza do lugar. Nós ficamos simplesmente maravilhados com tudo, foi uma experiência incrível 😍 😍 😍 ! Quem quiser pode conferir minha viagem completa no instagram, lá também tem dicas, fotos, vídeos e os preços dos principais passeios e restaurantes de Foz do Iguazu, dá uma conferida lá: @turistainiciante

🎫 41 reais (Entrada das Cataratas/adulto)
 🎫 11 reais (Entrada para crianças de 2 a 11 anos)
 🎫 11 reais (Entrada para pessoas acima de 59 anos)

🕒 Horário de funcionamento: 09 às 17 horas.

📍 O Parque funciona todos dias.

📍 Os ingressos podem ser comprados pelo site ou mesmo lá na entrada do parque.

Recolher conteúdo ▲

Data da experiência: outubro de 2019

Tipo de viagem: Viajou em casal

Fonte: TripAdvisor¹⁰.

Avaliar se uma atração turística de um destino, é compatível com uma demanda de viagem, é uma tarefa complexa ao se considerar a quantidade de conteúdo existente. Conforme exibido na Figura 10, apenas a atração Central Parque em Nova Iorque, possui 132 mil avaliações. Há muita informação relevante nesse conteúdo, com potencial para geração de conhecimento, no entanto, organizar tais textos é tarefa essencial para isso. O TripAdvisor oferece uma categorização de atrações, como Natureza e Parques, Cachoeiras, Museus, Diversão e Jogos etc. A Figura 13 exibe um exemplo da atração Cataratas do Iguazu.

¹⁰ Disponível em: https://www.tripadvisor.com.br/Attraction_Review-g303444-d312332-Reviews-Iguazu_Falls-Foz_do_Iguacu_State_of_Parana.html. Acesso em: 5 maio 2020.

Figura 13 – Categorias atração Cataratas do Iguaçu

Cataratas do Iguaçu
 ●●●●● 44.038 avaliações
 N.º 1 de 54 atividades em Foz do Iguaçu
 Cachoeiras

Visão geral
 PARQUE NACIONAL DO IGUAÇU E CATARATAS DO IGUAÇU: O Parque nacional do Iguaçu, criado em 1939, abriga o maior remanescente de floresta Atlântica da região sul do Brasil. O Parque protege uma riquíssima biodiversidade, constituída por espécies... mais

🕒 **Aberto agora:** 09:00 - 17:00 ⓘ
 ⌚ **Duração sugerida:** 2-3 horas
 📍 **Endereço:** Foz do Iguaçu, Paraná 85855-750 Brasil [Mapa](#)
 ✎ [Aprimore o perfil](#)

Ingressos
 2 opções a partir de **R\$ 106,48**
[Verificar disponibilidade](#)

📷 **Todas as fotos (30.071)**

Fonte: TripAdvisor¹¹.

Na Figura 13, a atração Cataratas do Iguaçu está na categoria Cachoeiras e possui 44 mil avaliações. De acordo com Oliveira (2017), essa classificação pode ser gerada pelos próprios usuários ou proprietários dos locais. De acordo com Oliveira (2017) embora as categorias forneçam um importante esquema de classificação, não há evidência de grau de pertinência das atrações com relação a essas categorias, dessa forma, uma atração bem menos relevante a essa categoria é igual a uma muito relevante.

2.4.3 O uso de perfis no turismo

As características do turista influenciam as suas escolhas na busca por experiências, atrações ou destinos turísticos. Atrações e destinos turísticos geralmente não são escolhidos por critérios racionais, mas sim por preferências implícitas do turista (SERTKAN; NEIDHARDT; WERTHNER, 2018). O estudo de perfis turísticos não é um tema novo. Cohen (1972), Hamilton-Smith (1987) já exibiram que a escolha e comportamento do turista, necessidades e expectativas variam consideravelmente. Crompton (1979) relata sete motivos sociológicos, incluindo fuga e relaxamento e dois motivos culturais pertinentes na escolha de destinos. O estudo de Gibson e Yiannakis (2002) mostra que o interesse e hábitos do turista mudam de acordo com localização

¹¹ Disponível em: https://www.tripadvisor.com.br/Attraction_Review-g303444-d312332-Reviews-Iguazu_Falls-Foz_do_Iguacu_State_of_Parana.html. Acesso em: 5 maio 2020.

geográfica, educação, estágio de vida, gênero, faixa de renda e estado civil. Os autores criam uma lista com 17 perfis turísticos e aplicam um questionário para verificar a aderência de turistas perante a estes perfis. Não somente condições físicas ou materiais influenciam esse processo de escolha, como também condições psicológicas, como mostra o estudo de Labbe (2016) sobre a influência dos traços de personalidades do *Big 5* (Abertura à experiência, Conscienciosidade/Perseverança, Extroversão, Simpatia/Amabilidade e Estabilidade Emocional) com experiências turísticas.

Neidhardt *et al.* (2014) apresentam um *framework* para identificar características de turistas com base em sete perfis turísticos. Os autores usam fotos para capturar o processo de escolha do usuário e afirmam que com base nisso, é possível captar traços que nem os próprios usuários conhecem. Para criar o conjunto dos sete perfis, os autores tomam de base os 17 perfis de Gibson e Yiannakis (2002) e os traços de personalidade do *Big 5*, porém objetivando reduzir a quantidade de perfis com foco na aderência de características. Os autores realizam uma análise de fator para reduzir a quantidade de perfis, chegando a sete perfis. O Quadro 2 exibe os sete perfis resultado desse trabalho.

Quadro 2 – Perfis turísticos

PERFIL	DESCRIÇÃO
Ação e Diversão	Gosta de ação, festas, exclusividade e evita lugares quietos.
Conhecimento	Mente aberta, educativo e bem organizado, turista frequente, que gosta de viajar em grupos, adquirir conhecimento, diferente de preguiçoso.
Cultura	Extrovertido, cultural, histórico. Amante de arte, consumidor ativo de comida e vinho.
Independente e Histórico	Buscando sentido da vida, interessado em história e tradições. Prefere viajar sozinho ao invés de tour em grupos.
Natureza e Recreação	Amante da natureza e do silêncio, quer escapar do cotidiano evitando aglomerações.
Social e Esportes	Mente aberta a esportes, que adora se socializar com povo local e não gosta de áreas de turismo intenso.
Sol e Relaxamento	Amante de locais quentes e sol. Não gosta de locais frios, chuva e cheio de pessoas.

Fonte: elaborado pelo autor (2021).

Nota: baseado em Neidhardt *et al.* (2014).

Os autores afirmam que seria possível integrar Atrações no seu modelo, de forma a gerar um sistema de recomendação com base no perfil do usuário x perfil da atração.

Werthner e Ricci (2004) apontam que, no turismo o processo de decisão é complexo, pois, não envolve somente atrações turísticas, mas sim, acomodação, transporte e alimentação. Os sistemas de recomendação oferecem ao turista um facilitador nesse processo decisório, no entanto, é difícil imaginar a classificação manual de uma quantidade cada vez maior de pontos turísticos, hotéis e restaurantes. Como forma de resolver tal problema, Glatzer, Neidhardt, Werthner (2018) apresenta um modelo baseado em mineração de texto para classificar hotéis nos sete perfis.

Sertkan, Neidhardt e Werthner (2018) afirmam que as técnicas de Criação de Perfis e Personalização podem ajudar no processo decisório do usuário quanto à seleção de atrações e destinos turísticos. Os autores realizam um estudo com o objetivo de identificar características dos destinos que possam influenciar os sete perfis do turista — criados por Neidhardt *et al.* (2014) — sob duas abordagens, uma supervisionada e outra não supervisionada. Os autores usam um banco de dados de uma empresa alemã de Turismo e realizam um processo manual de classificação de 561 destinos turísticos com a participação de especialistas. Esse processo é usado para treinar um algoritmo de regressão linear para classificar destinos. Apesar dos algoritmos apresentarem bom desempenho de classificação, os autores citam que são necessários mais dados para ter uma visão melhor sobre o problema. Apontam o exemplo da cidade do Rio de Janeiro, um destino que oferece diferentes atrações turísticas para diferentes perfis, como praias, natureza, calor, vida noturna e que a associação dessa cidade com um único perfil não seria correta. Isso reforça a tese de Gretzel *et al.* (2006), em que as pessoas tendem a escolher destinos por uma combinação de características.

A literatura evidencia a importância do tema, bem como abre espaço para uma investigação com relação ao mapeamento de perfis turísticos considerando turistas, atrações e destinos. Para isso, é importante observar que dentro de um mesmo destino, há ofertas distintas de atrações para públicos variados, pois conforme aponta Shin *et al.* (2017), apesar de um ponto turístico pertencer a um destino, pode possuir características distintas do destino, e como tal, necessita de representação própria. É importante observar também um conjunto de perfis abstrato capaz de mapear tanto turistas quanto destinos e atrações, ou seja, sem considerar aspectos como faixa de renda, gênero ou estado civil, que são propriedades particulares do turista e poderiam afetar sua utilização em destinos ou atrações. Lawton e Kallai (2002) apontam que

indivíduos reconhecem atrações turísticas de forma diferente. Portanto, ao se classificar um destino ou atração turística em perfis, é importante observar um conjunto de opiniões, e como tal, as avaliações podem ser importantes, tanto por sua diversidade quanto quantidade, conforme limitação exposta no trabalho de Sertkan, Neidhardt e Werthner (2018).

2.4.4 Estado da arte

Dentro do Turismo, pesquisas buscando extrair conhecimento e organizar o conteúdo de avaliações vem sendo objeto de estudo de muitos pesquisadores. Mckenzie e Adms (2018) classificam esses trabalhos em duas categorias: a) pesquisas relacionadas ao comportamento do usuário e sistemas de recomendação; b) veracidade e credibilidade de avaliações *online*. Alguns estudos também vêm buscando entender o poder de influência das avaliações no processo decisório do usuário. Seja pela sua credibilidade (FANG *et al.*, 2016), ou por sua habilidade de influenciar (SHIN *et al.*, 2016), pesquisadores procuram estudar o uso e explorar o conteúdo das avaliações. Hochmeister, Gretzel e Werthner (2013) relevam que a reputação do criador da avaliação também exerce impacto significativo na percepção sobre sua credibilidade.

Considerando o crescimento exponencial do número de avaliações, três categorias de estudos vêm procurando melhorar a experiência do usuário e o processo decisório com base nessa realidade. A primeira categoria é sobre criação de *ranking* de avaliações, a partir do qual cada avaliação recebe uma pontuação de acordo com sua utilidade e uso (KIM *et al.*, 2006; LU *et al.*, 2010). A segunda é sobre seleção das avaliações mais relevantes e úteis que cobrem mais aspectos possíveis (LAPPAS; GROVELLA; TERZI, 2012; LAPPAS; GUNOPULOS, 2010). A terceira refere-se a sobre o resumo de avaliações, com a classificação de avaliações entre positivas e negativas (CREMONESI *et al.*, 2014; HU; LIU, 2004; VERMEULE; SEEGERS, 2009).

Mckenzie e Adms (2018) desenvolvem um estudo sobre cidades e atrações turísticas usando avaliações do TripAdvisor por meio de uma análise linguística. A partir de dados extraídos do TripAdvisor e usando o algoritmo LDA para explorar os textos, os autores procuram identificar categorias e atrações mais características para cada destino. O trabalho dos autores é completo ao passo que avalia diferenças linguísticas

de percepção de um mesmo ponto turístico por idioma. De acordo com os autores, o uso de avaliações pode ser um importante fator no turismo, pois, permite conhecer melhor atrações e destinos turísticos com base na opinião de milhares de visitantes.

A literatura evidencia o potencial das avaliações como fonte de informação e conhecimento, no entanto, mostra também que tal conteúdo precisa ser organizado para que o conhecimento possa efetivamente ser explorado seja pela indústria do turismo ou usuários. Nesse sentido, o trabalho de Guy *et al.* (2017) busca extrair dicas importantes de avaliações extraídas do TripAdvisor, do tipo: “Entrada grátis toda última sexta-feira do mês.”, considerando também uma análise linguística das avaliações. Os autores usam algoritmos como SVM e *Logistic Regression* para classificar as dicas gerando assim um ranking por atração, podendo então extrair as 3 dicas mais importantes de cada atração. Os autores usam a participação de profissionais para ajudar no processo de classificação e validação. Com esse método, os autores conseguiram que 73% das dicas extraídas fossem marcadas como úteis. O trabalho de Shin *et al.* (2017) é similar a esse, porém, ao invés de perfis turísticos os autores usam o conceito de Escala de Personalidade de Destino, que é baseada na Escala de Personalidade da Marca (BPS) de Jennifer Aaker (1997).

O presente estudo exhibe um processo de mapeamento de atrações e destinos em perfis turísticos a partir do texto das avaliações. Até onde se sabe, esse é o primeiro estudo voltado à classificação de atrações e destinos em perfis turísticos automaticamente com base em avaliações. Pelo fato de considerar avaliações como fonte de informação, esta pesquisa se difere de métodos tradicionais de pesquisa baseados em questionário, pois objetiva-se extrair da opinião dos turistas, experiências e emoções que classifiquem determinada atração. Como há um recorte de opiniões apenas de turistas brasileiros, o trabalho pode oferecer uma fonte interessante para extração de conhecimento sobre atrações nacionais e internacionais com base na ótica dos brasileiros.

3 METODOLOGIA E DESENVOLVIMENTO

Essa seção apresenta os métodos de classificação de pesquisa quanto a sua natureza, objetivo ou orientação filosófica, com um aprofundamento nos métodos híbridos de pesquisa. Em seguida apresenta a metodologia da presente pesquisa em conjunto com framework metodológico utilizado. Na sequência, apresenta-se uma associação do método de pesquisa adotado perante as questões de pesquisa.

3.1 Aspectos gerais

De acordo com Creswell (2003), é possível classificar uma pesquisa quanto a sua natureza em três tipos: qualitativa, quantitativa e híbrida (quantitativa e qualitativa). Baracho (2007) relata que o método híbrido envolve as características qualitativa e quantitativa, por meio da coleta e análise de dados em um único estudo. Newman e Benz (1998) apresentam uma reflexão importante sobre métodos de pesquisa, a partir da qual os métodos quantitativo e qualitativo não poderiam ser vistos como dicotomias, como polos opostos, pois, os métodos podem se associar durante a pesquisa, representando uma associação contínua. Para Creswell (2003), os métodos mistos ou pesquisa híbrida residem no meio dessa associação, porque incorporam elementos qualitativos e quantitativos. Santos *et al.* (2017) apresentam uma abordagem interessante sobre a relação entre os dois métodos e relatam que os métodos mistos ou híbridos são geralmente escolhidos para promover um entendimento sobre um fenômeno não passível de explicação com apenas um método (qualitativo ou quantitativo).

Uma pesquisa híbrida pode também fazer uso concomitante dos métodos qualitativo ou quantitativo (SANTOS *et al.*, 2017). Existem variações na forma como os tipos de pesquisa se associam, uma pesquisa pode iniciar com a obtenção de dados qualitativos, sequenciada por uma análise quantitativa ou vice-versa. Para Creswell (2003), esse sequenciamento entre os métodos pode ser Explanatório Sequencial, no qual um pesquisador conduz uma pesquisa quantitativa para gerar resultados, que posteriormente são explicados por meio de uma abordagem qualitativa. Já no tipo Exploratório Sequencial, um pesquisador utiliza dados qualitativos sobre as opiniões dos participantes, para gerar uma segunda fase quantitativa que explora os resultados da primeira etapa.

Gil (2008) apresenta que as pesquisas podem ser classificadas quanto ao seu objetivo como: a) exploratórias, quando procuram desenvolver, esclarecer conceitos e ideias tendo em vista a formulação de problemas mais precisos ou construção de hipóteses; b) descritivas, quando procuram descrever as características de determinada população ou fenômeno relacionando ou não suas variáveis; ou c) explicativas, quando procuram explicar o porquê das coisas, tendo em sua preocupação principal explicar atores que determinam ou contribuem para a ocorrência dos fenômenos.

No campo epistemológico, Creswell (2003) classifica as pesquisas quanto a sua orientação filosófica em quatro dimensões: a) pós-positivismo, a qual reconhece que não podemos ser positivistas sobre nossa visão de conhecimento e que procura desafiar verdades absolutas sobre esse; b) construtivista, na qual o pesquisador procura entender a complexidade do problema, uma visão mais holística, mais abrangente e menos específica; c) transformativa, a qual procura focar em grupos e indivíduos que possam ser marginalizados em nossa sociedade, incluindo aspectos políticos e sociais; e d) pragmatismo, na qual existe uma preocupação com ações e aplicações para solução de problemas, procurando entender melhor o problema em estudo ao invés de focar nos métodos.

3.2 Metodologia da pesquisa

O primeiro aspecto a se considerar na classificação da presente pesquisa é quanto a orientação filosófica. O estudo se enquadra dentro do pragmatismo, pois busca apresentar uma solução para o problema gerado pelo excesso de informações turísticas e a consequente dificuldade de turistas, empresas ou governo em utilizar tais informações para tomada de decisão. Com relação ao seu objetivo, essa pesquisa se classifica como exploratória, pois procura explorar informações geradas a partir de um processo de classificação de avaliações turísticas. No campo técnico, a pesquisa classifica-se em um processo de mineração de texto, ou descoberta de conhecimento em base de dados, pois esse é um conceito utilizado para denominar a exploração de informações implícitas em grandes volumes de dados (BIGOLIN, BOGORNÝ, ALVARES, 2003). A formulação de novos problemas ou hipóteses se torna possível a partir dos resultados dessa pesquisa. Como pesquisa aplicada, espera-se que os resultados possam ser úteis no processo decisório de diferentes atores do ramo do

turismo, como turista, governo e setor privado, reduzindo sua incerteza sobre uma massa bruta de dados.

Com referência a sua natureza, o trabalho se enquadra no método híbrido, mais precisamente do tipo Exploratório Sequencial, pois, utiliza a organização de dados qualitativos (opiniões de turistas) para exploração quantitativa. O método qualitativo não se evidencia puramente pelo uso de dados com opiniões de turistas, mas também pela participação de dois agentes de turismo que fornecem informações qualitativas sobre perfis e destinos turísticos, como forma de filtrar e definir o escopo de investigação. Os modelos de classificação utilizam análise estatísticas quantitativas sobre os textos como forma de inferir a melhor forma de classificá-los. Os resultados da pesquisa são gerados perante uma perspectiva quantitativa, com base no resultado do processamento. Contudo, tais resultados são avaliados sob uma ótica híbrida envolvendo tanto aspectos qualitativos quanto quantitativos.

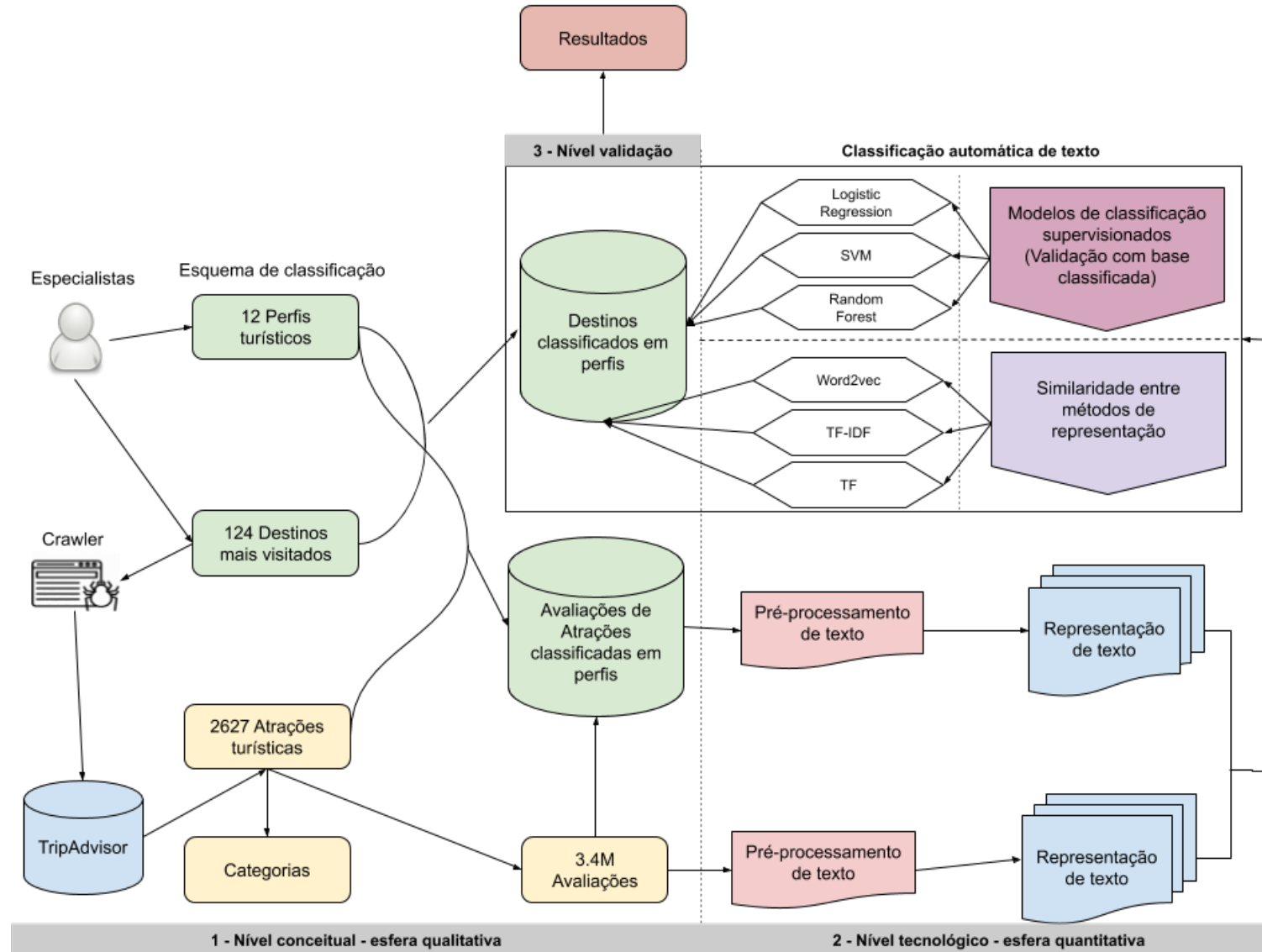
O objetivo da presente pesquisa é identificar a possibilidade de classificar destinos e atrações em perfis turísticos usando como fonte de dados as avaliações dos usuários. A arquitetura da pesquisa está dividida em duas etapas: a) construção de um modelo de classificação de destinos e atrações em perfis turísticos; b) análise exploratória de informações oriundas desse processo classificatório. A construção do modelo e procedimentos metodológicos são apresentadas nessa seção. A seção 4 exibe os resultados e uma análise dos mesmos sob uma perspectiva exploratória.

A construção do modelo envolve aspectos técnicos, conceituais e a participação de especialistas. O trabalho é caracterizado como multidisciplinar por envolver aspectos da CI, como os métodos de organização e recuperação da informação e aspectos da Ciência da Computação. Saracevic (1995) aponta que a CI é naturalmente um campo multidisciplinar. Essa interconexão com Tecnologia é vital para a CI dada a impossibilidade de organizar, classificar e recuperar informações usando os métodos tradicionais.

Nessa seção apresenta-se o processo metodológico envolvido na construção do modelo. O artefato produzido é responsável por classificar atrações em perfis turísticos de forma automática. Construção do modelo de classificação em três níveis: a) nível conceitual, no qual ocorre a participação de agentes especializados em

turismo e extração dos dados; b) nível tecnológico, no qual ocorre a construção técnica dos modelos de classificação automáticos; e c) nível validação, que exhibe o processo de validação dos modelos automáticos perante a classificação feita pelos especialistas. A Figura 14 exhibe o framework da pesquisa contemplando os três níveis.

Figura 14 – Framework de pesquisa exibindo os três níveis



Fonte: elaborado pelo autor (2021).

3.3 Nível conceitual

O nível conceitual envolve as etapas de definição de variáveis do estudo, coleta de dados, estratégia adotada e classificação manual de dados por especialistas. Como produto final, o objetivo é ter um modelo capaz de classificar atrações e destinos turísticos em perfis de acordo com suas características. Espera-se que os dados sejam coletados, filtrados e armazenados de forma que possam ser usados no nível tecnológico de forma mais fácil. O processo desenvolvido nesse nível é de natureza qualitativa, pois, os perfis são construídos com base na experiência dos especialistas, assim como definição dos destinos turísticos que vão participar do estudo. O processo de classificação manual também envolve aspectos subjetivos e a extração do conhecimento dos especialistas. Para que ocorra o processo classificatório, são necessários dados sobre destinos e atrações turísticas, perfis turísticos e avaliações de usuários perante as atrações.

3.3.1 Participação de especialistas

Muitos estudos adotam abordagens puramente automáticas de classificação, ou seja, com base nos próprios dados, as informações são organizadas e recuperadas para que o conhecimento possa ser gerado. Métodos completamente não supervisionados usam apenas o conhecimento do autor e técnicas de computação para organizar dados e atingir o objetivo do estudo. Tais métodos têm sido importantes, haja vista a dificuldade em organizar o volume de dados e informações. Embora apresentem ótimos resultados, como no trabalho de Mckenzie e Adams (2018) e Fang *et al.* (2016), depende do objetivo do estudo o envolvimento ou não de atores externos. O envolvimento de especialistas do domínio em estudo no trabalho pode enriquecer os resultados, haja vista que naturalmente há um processo de extração manual do conhecimento desses atores. Cleverley e Burnett (2015) apresentam que a sinergia existente usando os métodos mistos (manual e automático) geram resultados melhores que um único método no processo de organização do conhecimento. O conhecimento dos especialistas combina-se com o conhecimento gerado pelo processo automático.

A escolha dos agentes para participação nos processos se deu pela proximidade do autor com um agente, residente na cidade de Formiga, em Minas Gerais. Esse agente trabalha em agências de turismo há sete anos. A partir desse, foi indicado um outro agente que também poderia participar da pesquisa, residente na cidade de Belo Horizonte, que trabalha em agências de turismo há 12 anos. O trabalho no dia a dia dos agentes envolve entre outros fatores, o entendimento da demanda de viagem do turista e a compatibilização desta demanda com oferta de atrações e destinos turísticos de acordo com o perfil do turista.

Na indicação de destinos turísticos, diversas variáveis são consideradas, como orçamento para a viagem, experiências que o turista procura, quem participará da viagem, entre outros fatores, conforme apontado por Gibson e Yiannakis (2002). O processo de seleção de um destino turístico envolve fatores como a escolha de atrações, escolha de hospedagem, voos, alimentação, entre outros fatores. Nesse trabalho, apresentamos um recorte com relação a escolha das atrações, pois possuem maior peso na escolha dos turistas, segundo os próprios agentes. Com base na experiência desses agentes seria possível definir um conjunto de perfis turísticos, ou seja, um conjunto que compreenda as características mais buscadas pelos turistas perante as atrações de um destino turístico.

3.3.2 Destinos estudados

A primeira etapa de seleção foi com relação aos destinos turísticos. Um destino possui diversas atrações turísticas, sendo que cada atração pode possuir características diferentes. Por exemplo, na cidade do Rio de Janeiro, há atrações voltadas ao público que gosta de arquitetura, como também há atrações para o público que gosta de aventura ou praia. Partindo desse princípio, o método define os destinos mais visitados por brasileiros e em seguida coleta dados sobre as atrações desses destinos. Para seleção dos destinos, foi solicitado aos agentes a criação de uma lista com o nome do país e nome dos destinos mais visitados pelos brasileiros, conforme sua experiência. Para definição da lista final, os dois agentes entram em consenso em relação aos destinos.

Como todo o trabalho foi realizado remotamente, os agentes preencheram uma planilha *online* no Google Sheets com os nomes dos destinos.

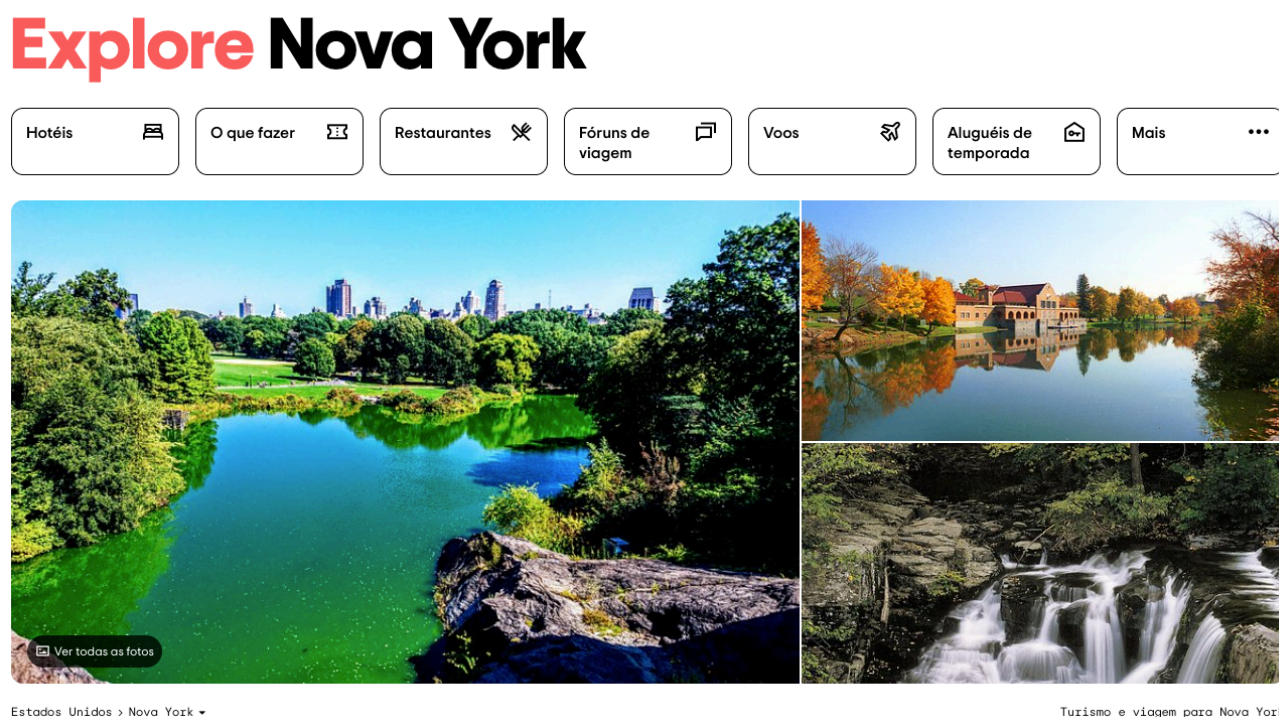
Um destino turístico pode ser uma cidade, um distrito, ou até mesmo um Parque, como o Parque Estadual do Jalapão, localizado no estado do Tocantins. A lista final apresentou 124 destinos, sendo 72 na América do Sul, com predominância de destinos nacionais, totalizando 63 destinos no Brasil. Em segundo lugar com mais destinos apareceu a Europa, com 25 destinos, seguida da América Central com 9 destinos e América do Norte com 8 destinos. Ásia, África e Oceania aparecem nas últimas posições, com quatro, três e três destinos respectivamente.

3.3.3 Coleta de dados

A partir da lista de destinos, o próximo passo dentro do nível conceitual é a aquisição e organização de dados sobre atrações e avaliações de cada destino turístico. Como o foco do estudo é avaliar a possibilidade de classificar atrações em perfis turísticos, seria necessário buscar essas atrações e também as avaliações de cada atração. Devido a diversidade e quantidade de dados, foi escolhido o site TripAdvisor para obtenção dos dados.

O TripAdvisor possui uma página para cada destino turístico, na qual é possível obter um conjunto de informações sobre o destino como: hotéis, restaurantes e atrações. O foco do estudo é a utilização de dados apenas de atrações. As atrações estão listadas dentro do menu “O que fazer” dentro da página de cada destino. A Figura 15 exibe um exemplo das opções do TripAdvisor para o destino Nova York.

Figura 15 – Página de Nova York no TripAdvisor



Fonte: TripAdvisor¹².

3.3.4 Coletando avaliações

No TripAdvisor, dentro da página do destino é possível listar todas as atrações na opção “O que Fazer”. O objetivo foi extrair o máximo de atrações possíveis dentro de cada uma dessas páginas. Para cada atração, seriam extraídas as suas avaliações e categorias. A Figura 16 exhibe a página da atração Plataforma de Observação *Top of the Rock*.

¹²Disponível em: https://www.tripadvisor.com.br/Tourism-g60763-New_York_City_New_York-Vacations.html. Acesso em: 5 maio 2020.

Figura 16 – Página da atração Plataforma de Observação Top of the Rock

Plataforma de Observação Top of the Rock
 ●●●●● 79.053 avaliações
 N.º 7 de 1.272 atividades em Nova York
 Mirantes, Deques e torres de observação

O que dizem os viajantes

” Já conhecia o **empire state** e pude conhecer no passado as torres gêmeas, o **top of the rock** me surpreendeu, pois está em uma localização central e permite uma visão 360 graus da cidade!

” Já conhecia o **empire state** e pude conhecer no passado as torres gêmeas, o **top of the rock** me surpreendeu, pois está em uma localização central e permite uma visão 360 graus da cidade!

✎ Aprimore o perfil

Mirante Top of the Rock, Nova York a partir de **R\$ 210,22**

Verificar disponibilidade

Certificado de Excelência

Todas as fotos (39.519)

Fonte: TripAdvisor¹³.

Na Figura 16 é possível visualizar que essa atração possui 79.053 avaliações feitas por usuários e está classificada nas categorias: mirantes, deques e torres de observação. Conforme exibido na seção 2.4.2, cada avaliação possui uma estrutura composta por título, avaliação (comentário), data de visita, data da avaliação, o *Rating* (quantidade de estrelas), usuário que fez o comentário e sua cidade.

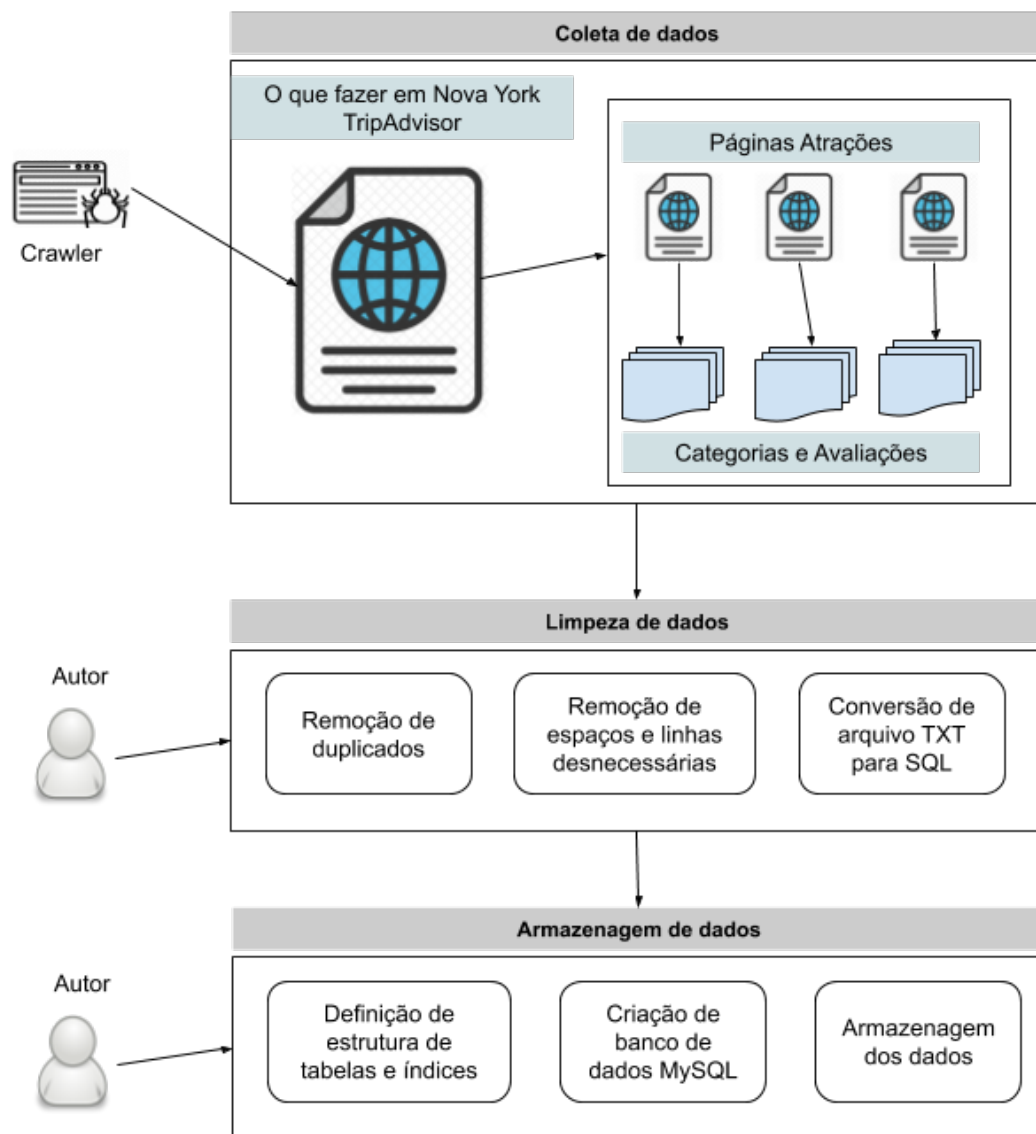
Foi definido como escopo a utilização de avaliações realizadas por brasileiros escritas no idioma português. Esse recorte ocorre por dois motivos: a) a impossibilidade de tradução das avaliações escritas em outro idioma, sem que seu conteúdo pudesse ser comprometido; b) um recorte conceitual na pesquisa, que busca classificar atrações em perfis com base na opinião dos brasileiros. O conteúdo das avaliações é qualitativo, subjetivo e envolve as experiências de um determinado turista perante uma atração. No total, 124 destinos e 3.263 atrações turísticas com 3.458.668 avaliações foram coletados no período de 2011 à 2019. As categorias de atrações totalizaram 111 e o conjunto formado pela relação (atração x categoria) totalizou 6.371 registros.

¹³Disponível em: https://www.tripadvisor.com.br/Attraction_Review-g60763-d587661-Reviews-Top_of_the_Rock-New_York_City-New_York.html. Acesso em: 5 maio 2020.

3.3.5 Arquitetura de organização dos dados

Com base na lista de destinos e na análise da estrutura de páginas do TripAdvisor, foi desenvolvido um *Crawler* para extrair os dados. Para cada atração, o *Crawler* extrai suas categorias e avaliações. Após o *Crawler* extrair os dados das páginas do TripAdvisor, foi necessário realizar um processo de limpeza nos mesmos e em seguida, o armazenamento no banco de dados para facilitar futuras manipulações. A Figura 17 exibe a arquitetura do processo.

Figura 17 – Arquitetura do processo de extração de dados



Fonte: elaborado pelo autor (2021).

Para desenvolvimento do *Crawler*, foi utilizado o *Scrapy*¹⁴, que é um framework grátis escrito em Python para extração de páginas da Web. Com o *Scrapy* é possível navegar entre as páginas existentes e outras referenciadas por esta. Permite assim, um aprofundamento maior no nível de extração. Os dados foram extraídos e salvos em arquivos com dados separados por vírgula (CSV) com o nome de cada Atração. Cada avaliação contém: usuário que fez a avaliação, cidade do usuário, destino turístico, título da avaliação, avaliação, quantidade de estrelas *Rating*, data da visita, data da avaliação, URL da avaliação, e categorias da atração. Foi realizada a criação de uma classe com estes atributos na linguagem Python, para que os mesmos fossem mapeados e posteriormente salvos em memória (arquivo CSV). A Figura 18 exibe a estrutura dessa classe.

Figura 18 – Estrutura da classe

```

9  import scrapy
10
11
12  class ReviewItem(scrapy.Item):
13      # Items to get
14      user          = scrapy.Field()
15      city_user     = scrapy.Field()
16      city          = scrapy.Field()
17      rev_title     = scrapy.Field()
18      review        = scrapy.Field()
19      place         = scrapy.Field()
20      dt_review     = scrapy.Field()
21      rating        = scrapy.Field()
22      dt_stay       = scrapy.Field()
23      source_url    = scrapy.Field()
24      categorias    = scrapy.Field()

```

Fonte: elaborado pelo autor (2021).

Após realizar a extração e salvar os arquivos CSV localmente, foi realizado um processo de limpeza dos dados eliminando avaliações duplicadas e aquelas com atributos ausentes. Esse processo foi importante para trazer mais consistência aos dados e conseqüentemente ao estudo. Como haveria consultas frequentes aos dados, o que se torna inviável usando arquivos CSV, foi utilizado o banco de dados relacional MySQL para armazená-los. Os dados foram separados nas tabelas: destinos, atrações, categorias, categorias das atrações e avaliações. Dessa forma, é possível utilizar

¹⁴ Disponível em: <https://scrapy.org/>. Acesso em: 5 maio 2020.

indexação nas tabelas, tornando as consultas mais rápidas. A Tabela 1 exibe a lista final de destinos usados na pesquisa. Para cada destino turístico é exibido o país, o número de atrações e o número de avaliações.

Tabela 1 – Destinos turísticos utilizados no estudo

Continua

PAIS	DESTINO	ATRAÇÕES	AVALIAÇÕES	PAIS	DESTINO	ATRAÇÕES	AVALIAÇÕES
África do Sul	Joanesburgo	20	1149	Brasil	Chapada Diamantina	7	1961
África do Sul	Cidade do Cabo	29	5532	Brasil	Chapada dos Veadeiros	26	14378
Alemanha	Berlim	30	17623	Brasil	Curitiba	30	105353
Alemanha	Frankfurt	25	5928	Brasil	Diamantina	25	3719
Alemanha	Munique	30	8748	Brasil	Fernando de Noronha	30	33364
Argentina	Buenos Aires	30	82048	Brasil	Florianópolis	25	53978
Argentina	Córdoba	17	1327	Brasil	Fortaleza	30	67353
Argentina	San Carlos de Bariloche	28	13785	Brasil	Foz do Iguaçu	28	108483
Austrália	Brisbane	16	514	Brasil	Garibaldi	12	1868
Austrália	Melbourne	22	816	Brasil	Goiânia	29	15668
Bahamas	Nassau	5	546	Brasil	Gramado	30	169406
Bélgica	Bruxelas	16	5574	Brasil	Holambra	10	2575
Brasil	Angra dos Reis	41	28550	Brasil	Ilhabela	30	19842
Brasil	Aracaju	30	30990	Brasil	Itaipava	7	1845
Brasil	Arraial d'Ajuda	15	66581	Brasil	Jericoacoara	12	35628
Brasil	Arraial do Cabo	24	22804	Brasil	João Pessoa	30	30335
Brasil	Balneário Camboriú	30	29121	Brasil	Joinville	30	6942
Brasil	Belém	27	32317	Brasil	Maceió	30	33050
Brasil	Belo Horizonte	30	66662	Brasil	Maragogi	20	19706
Brasil	Bento Gonçalves	30	32858	Brasil	Morro de São Paulo	13	16553
Brasil	Blumenau	30	14674	Brasil	Natal	26	67752
Brasil	Bombinhas	28	12703	Brasil	Navegantes	6	509
Brasil	Bonito	25	34000	Brasil	Olinda	30	10135
Brasil	Brasília	30	79635	Brasil	Ouro Preto	24	22108
Brasil	Cabo Frio	30	17889	Brasil	Palmas	28	5816
Brasil	Caldas Novas	20	16289	Brasil	Paraty	26	24671
Brasil	Campos do Jordão	28	61907	Brasil	Parque Estadual de Jalapão	2	152
Brasil	Canela	29	56689	Brasil	Petrópolis	16	36573
Brasil	Canoa Quebrada	21	8978	Brasil	Pirenópolis	29	11521
Brasil	Caraíva	1	1812	Brasil	Pomerode	22	4325

Tabela 1 – Destinos turísticos utilizados no estudo

				Conclusão			
PAÍS	DESTINO	ATRAÇÕES	AVALIAÇÕES	PAÍS	DESTINO	ATRAÇÕES	AVALIAÇÕES
Brasil	Porto Alegre	30	45809	Federação Russa	São Petersburgo	32	3316
Brasil	Porto de Galinhas	23	45230	França	Paris	30	84421
Brasil	Porto Seguro	31	105061	Grécia	Atenas	27	7823
Brasil	Praia do Forte	9	29250	Holanda	Amsterdã	30	25363
Brasil	Recife	30	52394	Hungria	Budapeste	25	9500
Brasil	Rio de Janeiro	30	203441	Irlanda	Dublim	26	5973
Brasil	Rio Quente	4	8608	Israel	Jerusalém	26	3961
Brasil	Salvador	30	68346	Itália	Florença	58	19805
Brasil	Santos	30	21275	Itália	Milão	58	14298
Brasil	São Luís	30	14902	Itália	Roma	30	40729
Brasil	São Paulo	30	181526	Itália	Veneza	30	13197
Brasil	Tiradentes	24	12150	Japão	Tóquio	26	2120
Brasil	Trindade	4	8295	México	Cancun	25	7770
Brasil	Ubatuba	30	24058	México	Cidade do México	30	10299
Brasil	Vitória	30	19142	Nova Zelândia	Auckland	14	1017
Canadá	Montreal	28	4418	Panamá	Panamá	26	7575
Canadá	Toronto	28	9642	Peru	Cusco	29	15123
Colômbia	Bogotá	28	11582	Peru	Lima	30	20094
Colômbia	Cartagena	29	17714	Peru	Machu Picchu	19	5281
Colômbia	San Andres	10	5497	Portugal	Lisboa	30	67299
Curacao	Willemstad	12	3801	Portugal	Porto	30	31679
Dinamarca	Copenhague	22	3527	Reino Unido	Londres	31	30251
Egito	Cairo	23	1899	República Dominicana	Punta Cana	17	9505
Emirados Árabes Unidos	Dubai	27	9185	República Dominicana	Santiago	10	17594
Espanha	Barcelona	30	39059	República Tcheca	Praga	28	12484
Espanha	Madrid	30	28637	Rússia	Moscou	18	3826
Estados Unidos	Boston	29	3572	Suécia	Estocolmo	23	3127
Estados Unidos	Las Vegas	28	21479	Suíça	Zurique	26	2497
Estados Unidos	Los Angeles	30	9558	Tailândia	Bangcoc	22	7023
Estados Unidos	Miami	27	13381	Turquia	Istambul	25	10610
Estados Unidos	Nova Iorque	30	86119	Uruguai	Montevidéu	30	43980
Estados Unidos	Orlando	30	76563	Uruguai	Punta del Este	27	21316

Fonte: elaborado pelo autor (2021).

3.3.6 Criação de perfis turísticos

O principal objetivo dos perfis turísticos criados é que possam ser capazes de classificar atrações turísticas e conseqüentemente destinos turísticos. Fundamentado pelo trabalho de McKenzie e Adams (2018), a partir do entendimento das atrações turísticas é possível traçar um perfil do destino. Embora o principal objetivo seja a classificação de atrações turísticas, o modelo foi criado para que pudesse também ser capaz de classificar o turista, assim, o conjunto de perfis deveria ser abstrato ao ponto de poder classificar ambas as esferas (turista x atração). Este requisito foi identificado durante as reuniões com os agentes, pois a partir de um modelo único capaz de classificar atrações, destinos e turistas tornaria mais fácil uma possível análise de similaridade.

No processo de construção dos perfis turísticos, foram utilizados dois fatores: experiência dos agentes e fundamentação da literatura. Com relação a fundamentação da área, foram avaliados o modelo de sete perfis turísticos de Neidhardt *et al.* (2014) e o estudo de Gibson e Yiannakis (2002) englobando 22 perfis turísticos (17 papéis + os 5 traços de personalidade). O modelo de Neidhardt *et al.* (2014) é uma apresentação reduzida e sintetizada do modelo de Gibson e Yiannakis (2002). Ambos os modelos foram avaliados com o objetivo de verificar a possibilidade de sua utilização como o esquema de classificação que seria utilizado nessa pesquisa. O modelo de Gibson e Yiannakis (2002) foi descartado por envolver características que eram exclusivas do turista. Conforme supracitado, o modelo considera também aspectos de personalidade. Dentro desse modelo, um dos perfis é *escapista*, ou seja, aquele que busca escapar do cotidiano. Essa classificação é interessante do ponto de vista do turista, no entanto, não apresenta o grau de escapismo que determinado destino ou atração possuem, sendo, portanto, incompatível com o foco do estudo. Por um outro lado, o modelo de sete perfis é muito genérico, porque envolve em um mesmo perfil situações muito amplas, como por exemplo inserir dentro do perfil Independente e Histórico o amor pela arte e também por vinhos.

Inexistindo na literatura esquema de classificação capaz de atingir o objetivo do estudo, decidiu-se pela criação de um conjunto de perfis capaz de representar tanto turista quanto

atração turística. Variáveis relacionadas a clima como neve, frio e calor foram removidas porque sofrem muitas alterações de acordo com a estação do ano. Exemplo: A cidade de Nova York pode oscilar entre 35° e -35° Celsius entre verão e inverno. Assim, a variável clima está diretamente ligada ao período de viagem e não necessariamente uma característica única do destino. As variáveis não consideradas como faixa de renda e clima são importantes no processo decisório do turista, em um sistema de recomendação, seria importante considerá-las como forma de filtro para o usuário, porém, não para avaliar o grau de similaridade entre destinos ou com turistas.

Os agentes realizaram reuniões com e sem a participação do autor para que a lista de perfis turísticos fosse finalizada. Foram apresentados aos agentes o estudo em tela, seu objetivo e o que se esperaria de seus trabalhos. Foram apresentadas também os esquemas de classificação existentes de Gibson e Yannkis (2002) e Neidhardt *et al.* (2014). Na criação dos perfis, os agentes deveriam considerar sua experiência cotidiana atendendo diferentes demandas de turistas. Nas reuniões, foi possível identificar que devido ao local de trabalho e experiência dos agentes, ambos já trabalharam com classes sociais distintas. Os agentes deveriam entrar num consenso com relação a escolha dos perfis turísticos. O Quadro 3 apresenta o esquema de classificação (perfis turísticos) definidos para a presente pesquisa.

Quadro 3 – Esquema de classificação (perfis turísticos)

PERFIL	DESCRIÇÃO
Aventura	Destinos com boa oferta ou turistas que gostam de mergulho, rapel, trilhas, entre outras atividades de aventura.
Compras	Destinos com boa oferta ou turistas que gostam de shoppings, feiras e lojas.
Cultura	Destinos com boa oferta ou turistas que gostam de cultura conhecimento, história etc.
Família	Destinos com boa oferta ou turistas que gostam de diversão com crianças, parques etc.
Gastronomia	Destinos com boa oferta ou turistas que gostam de vinhos, cervejas e culinária.
Natureza/Exóticos	Destinos com boa oferta ou turistas que natureza como rios, cachoeiras, animais, e opções exóticas.
Paisagem/Arquitetura	Destinos com boa oferta ou turistas que gostam de montes, monumentos ou construções modernas ou antigas (castelos).
Praia	Destinos com boa oferta ou turistas que gostam de praias.
Relaxante	Destinos com boa oferta ou turistas que procuram descanso, reflexão, fugir de aglomerações e movimento.
Religioso	Destinos com boa oferta ou turistas que procuram por igrejas, catedrais e história religiosa.
Romântico	Destinos com boa oferta ou turistas que procuram por locais para curtir a dois.
Vida Noturna	Destinos com boa oferta ou turistas que gostam de casinos, bares, boates.

Fonte: elaborado pelo autor (2021).

3.3.7 Estratégia de classificação

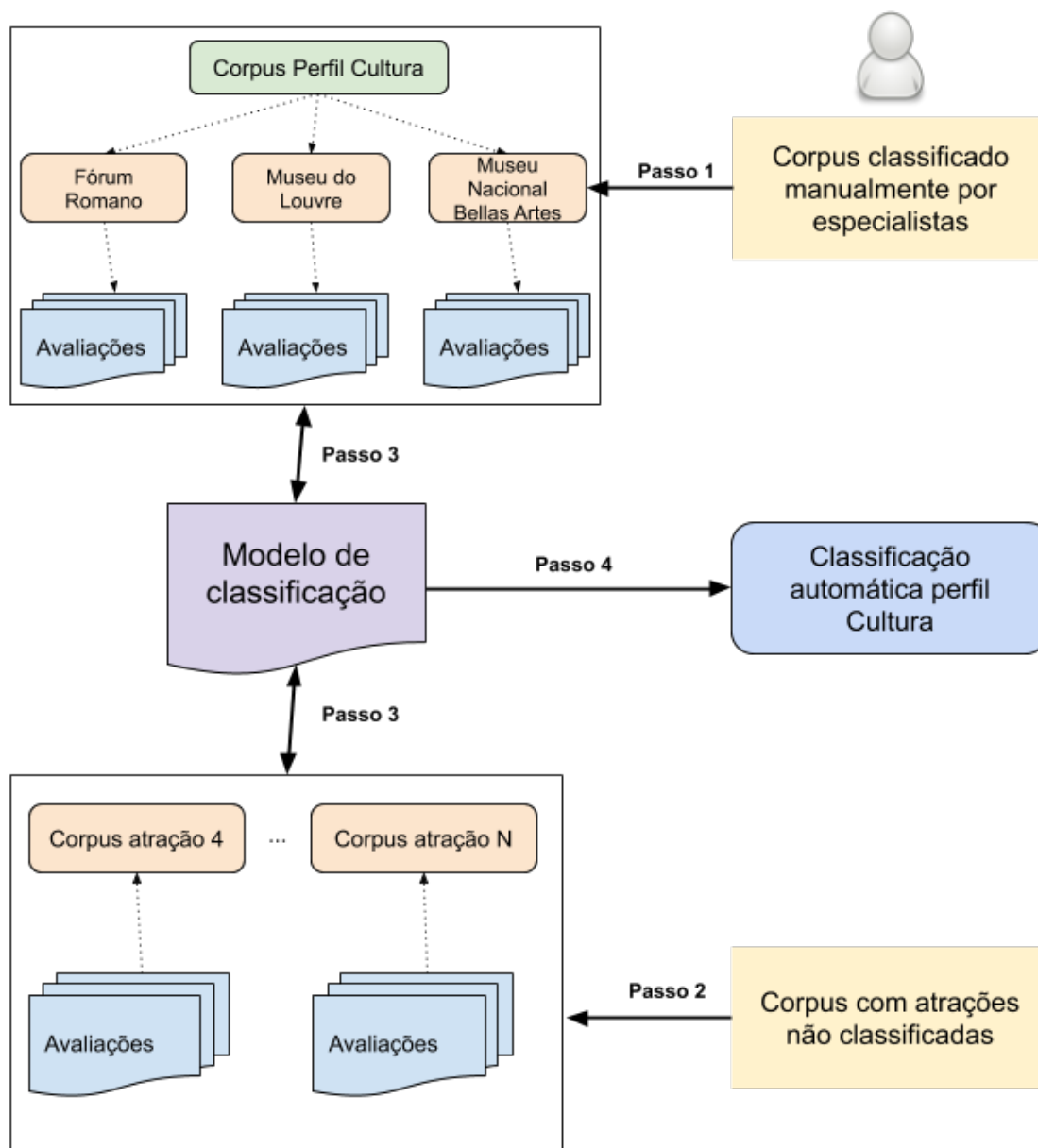
O trabalho para classificar todas as atrações turísticas (3.263) dos destinos, de forma manual, seria dispendioso. Por exemplo, se considerar que são 12 perfis turísticos, então seriam 31.524 classificações envolvendo (atração x perfil) para cada agente. Se cada agente demorasse três minutos para avaliar a atração e estabelecer corretamente seu nível de compatibilidade com determinado perfil, então seriam necessários 63.048 minutos para finalizar, ou seja, considerando oito horas diárias ininterruptas de trabalho, seriam necessários 6,56 meses ininterruptos para que cada agente concluísse a classificação manual das atrações. A matemática ratifica a importância de construir

alternativas automáticas ou semiautomáticas para classificar e organizar o conteúdo do *big data*.

Como forma de classificar as 3.263 atrações, adotou-se uma heurística na qual algumas atrações foram classificadas manualmente, e a partir dessas, as demais seriam classificadas automaticamente. Conforme apontam Askalidis e Malthouse (2016), as avaliações podem expressar uma opinião positiva ou negativa sobre determinado produto ou serviço, no entanto, raramente envolvem fatores externos ao item em avaliação. Com base nesse princípio, adotou-se a abordagem de identificar as atrações que melhor representassem determinado perfil turístico. Nesse sentido, as atrações funcionariam como uma classe de documentos, ou seja, uma atração seria como uma área de pesquisa científica, enquanto suas avaliações seriam documentos publicados com termos e palavras-chave dentro daquela área.

Para classificar manualmente as atrações melhores para cada perfil turístico, foi elaborado uma lista de atrações por destino, filtrando apenas atrações com mais de mil avaliações positivas (*rating* ≥ 4) em uma escala até cinco. Esse filtro ocorre com base no trabalho de Guy *et al.* (2017) que aponta que avaliações com *rating* acima de quatro apresentam mais relevância e utilidade perante os usuários. Uma avaliação negativa é importante para o usuário, no entanto, ela pode não representar corretamente o que há de melhor naquela atração. A amostra classificada manualmente deveria ter documentos (avaliações) capazes de representar o que há de melhor naquela coleção (atração). Em um sistema de recomendação, não seria interessante indicar uma atração com baixa reputação em Cultura para uma turista que procura boas experiências de Cultura. A partir da lista gerada, os agentes escolheram em consenso as três atrações que melhor representassem cada perfil. Para evitar que uma atração tivesse mais relevância perante as outras duas e enviesar o corpus, utilizou-se um tamanho fixo de documentos (avaliações) para cada atração, as 1.000 avaliações mais recentes com *rating* maior ou igual a 4. A Figura 19 exibe a estratégia de classificação utilizada.

Figura 19 – Estratégia de classificação automática



Fonte: elaborado pelo autor (2021).

Na Figura 19 é possível visualizar o processo de classificação automática no exemplo do perfil *Cultura*. No passo 1, ocorre a classificação manual das principais atrações para cada perfil, gerando um *corpus* do perfil *Cultura*. No passo 2, um *corpus* para cada atração turística é gerado com suas avaliações. Os dois *corpora* são usados nos modelos de classificação (passo 3). No passo 4, um grau de relevância entre o perfil *Cultura* e cada atração ainda não classificada é gerado. O texto de cada avaliação é

formado pela concatenação das colunas Título da Avaliação e Comentário. Para geração do *corpus* de cada perfil, os agentes usaram seu conhecimento e experiência prática com base em relatos de turistas que visitaram para tais atrações. O Quadro 4 apresenta as atrações usadas para gerar o *corpus* de cada perfil.

Quadro 4 – Atrações usadas para geração de *corpus* do perfil

PERFIL	DESTINOS	ATRAÇÕES
Aventura	Canela, Natal, Las Vegas	Alpen Park, Dunas de Genipabu, Stratosphere Tower
Compras	Orlando, Miami, Miami	Orlando International Premium Outlets, Disney Springs, Dolphin Mall
Cultura	Roma, Paris, Buenos Aires	Forum romano, Museu do Louvre, Museu Nacional de Belas Artes
Família	Orlando, Arraial d'Ajuda, Orlando	Walt Disney World Resort, Eco Parque, Universal Orlando Resort
Gastronomia	Bento Gonçalves, Petrópolis, Belo Horizonte	Vinícola Casa Valduga, Cervejaria Bohemia, Edifício Maletta
Natureza/Exóticos	Bonito, São Paulo, Dubai	Gruta do Lago Azul, Jardim Botânico de São Paulo, Burj Khalifa
Paisagem/Arquitetura	Madri, Machu Picchu, Praga	Templo de Debod, Machu Picchu, Castelo de Praga
Praia	San Andres, Maceió, Florianópolis	Playa de san luiz, Praia de Ponta Verde, Praia dos Ingleses
Relaxante	Paris, Veneza, Jericoacoara	Seine River, Canal Grande, Duna do Pôr do Sol
Religioso	São Paulo, Roma, Paris	Solo Sagrado De Guarapiranga, Basílica di Santa Maria Maggiore, Catedral de Notre-Dame
Romântico	Gramado, Punta cana, Buenos Aires	Lago Negro, Saona Island, Puerto Madero
Vida Noturna	Las Vegas, Porto Seguro, Lisboa	Casino at Bellagio, Rua do Mucugê, Bairro Alto

Fonte: elaborado pelo autor (2021).

As principais atrações de cada perfil envolvem destinos nacionais e internacionais. Embora possam existir atrações que pudessem representar com mais precisão a essência de cada perfil, a seleção se deu com base nos dados existentes, então se uma atração não tivesse ao menos 1.000 avaliações com rating maior ou igual a 4, a mesma

seria descartada. A apresentação da estratégia de classificação encerra o nível conceitual da metodologia.

3.4 Nível tecnológico

Conforme exibido na seção anterior, o Nível Conceitual compreende o planejamento, definição da participação de especialistas, coleta de dados, organização de dados e a estratégia de classificação. O Nível Tecnológico compreende as etapas de pré-processamento dos dados, representação dos dados e dois métodos de classificação: classificação por similaridade dos documentos representados e classificação usando abordagens supervisionadas de AM. O objetivo desse nível é atingir modelos de classificação automáticos capazes de classificar com boa acurácia as atrações em perfis turísticos, para que possam ser avaliados na próxima etapa, o Nível de Validação. Dentro do Nível tecnológico, a linguagem Python, na versão 3.5.2 foi escolhida para tratamento das etapas, por ser largamente utilizada em projetos de classificação textual. O ambiente Jupyter Notebook¹⁵, na versão 6.0.0 foi utilizado para execução dos códigos, pois, permite fácil organização do conteúdo.

A base conceitual para a classificação automática é a comparação de similaridade de um *corpus* do perfil, ou seja, documentos (avaliações) que melhor representam as características do perfil, com um corpus da atração, ou seja, documentos (avaliações) que compõem as características daquela atração. O produto desse processo é um grau de pertinência de uma determinada atração perante um perfil, ou seja, quão aquela atração é relevante perante aquele perfil, com base nas opiniões expressas nas avaliações.

3.4.1 Pré-processamento dos textos

Acuña (2011) afirma que mais de 60% do tempo gasto num processo de mineração de dados ou classificação de texto é direcionado na etapa de preparação dos dados, que contribui significativamente para o sucesso do projeto. O *corpus* do perfil é formado por

¹⁵ Disponível em: <https://jupyter.org/>. Acesso em: 5 maio 2020.

3.000 avaliações das três atrações escolhidas pelos agentes. Com relação a atração, cada uma apresenta uma quantidade de avaliações diferente, de acordo com sua popularidade (MCKENZIE; ADAMS, 2018). Ambas as coleções são compostas por documentos (avaliações) em estado bruto, ou seja, da forma exata como foram extraídas do TripAdvisor. Para que os algoritmos possam comparar a similaridade entre os textos de forma eficiente, é necessário extrair do seu conteúdo apenas o que há de relevante para o estudo.

O pré-processamento refere-se a uma etapa de PLN e como tal, a linguagem Python oferece alguns recursos para esse processamento em português. Nesse trabalho, utilizou-se os recursos disponíveis no Conjunto de Ferramentas para Linguagem Natural (NLTK) na versão 3.4.4, das bibliotecas Sklearn na versão 0.21.3 e Gensim na versão 3.8.0. Embora os recursos dessas ferramentas sejam abundantes em inglês, em português ainda há certa limitação, como a não inclusão de diversos artigos, preposições e conjunções na lista de *stop-words* do NLTK. Conjunções, advérbios, preposições, artigos podem ser considerados *stop-words*, porque não acrescentam valor significativo à representação do texto. Dessa forma, o processo envolveu a utilização de tais ferramentas em conjunto com códigos desenvolvidos pelo autor para tratamento adequado dos dados perante o estudo.

Como forma de evitar que termos com o mesmo significado expressos em diferentes formas (plural ou singular) tivessem conotação diferente, foi criada uma função para transformar os termos no plural para singular. Dessa forma, o termo museus seria alterado no processamento para museu, ou seja, seria compatibilizado com outros locais que o termo é expresso já no singular. Não se encontrou funções que pudessem fazer isso no idioma português dentro das bibliotecas avaliadas. Acentos, números, quebras de linha, caracteres especiais foram removidos de todos os textos. Cada palavra de cada avaliação foi transformada em *tokens*, criando assim um vetor de *tokens* que representasse cada avaliação. A Figura 20 exhibe um exemplo desse processo.

Figura 20 – Exemplo de limpeza de avaliações

```
In [25]: 1 corpus_exemplo = [['Jardins enormes. Visitei um jardim em 2017!.'],
2           ['O passeio de teleférico é bem agradável no morro dos urubus']]
3 print("\n\033[94m\033[1m Avaliações de exemplo:\033[0m",corpus_exemplo)
4 print("-----")
5 corpus_limpo = [[amm.clear_text(d[0])] for d in corpus_exemplo]
6 print("\n\033[94m\033[1m Após limpeza acentuação e caracteres especiais:\033[0m",corpus_limpo)
7 print("-----")
8 corpus_singular = [amm.plural2_singular(d[0].split(" ")) for d in corpus_limpo]
9 print("\n\033[94m\033[1m Após conversão de plural para singular e tokenização:\033[0m",corpus_singular)
10 print("-----")

Avaliações de exemplo: [['Jardins enormes. Visitei um jardim em 2017!.'], ['O passeio de teleférico é bem agradável
no morro dos urubus']]
-----
Após limpeza acentuação e caracteres especiais: [['jardins enormes visitei um jardim em'], ['o passeio de teleferico
e bem agradável no morro dos urubus']]
-----
Após conversão de plural para singular e tokenização: [['jardim', 'enorme', 'visitei', 'um', 'jardim', 'em'], ['o',
'passeio', 'de', 'teleferico', 'e', 'bem', 'agradavel', 'no', 'morro', 'do', 'urubu']]
-----
```

Fonte: elaborado pelo autor (2021).

Após as avaliações estarem limpas de números, caracteres especiais, acentuação e convertidas para o singular, iniciou-se o processo de remoção de *stop-words*. O pacote NLTK contém uma lista de *stop-words* com 204 termos como artigos, preposições, advérbios e conjunções. Nessa lista, não é possível localizar diversos termos que são considerados *stop-words* em português, como por exemplo a conjunção “*porque*” e o advérbio “*ainda*”. Como forma de melhorar a lista no idioma português, uma lista de *stop-words* adicional foi gerada contendo mais 1.527 artigos, preposições, advérbios e conjunções. Essa lista foi criada a partir da extração automática de listas do idioma português na Internet usando *crawlers*. A Figura 21 exibe o mesmo *corpus* usado no exemplo anterior com a remoção de *stop-words*. No exemplo, os termos (e, bem, no, de, o, dos) são removidos do texto.

Figura 21 – Exemplo de remoção de *stop-words*

```
1 texto = 'o passeio de teleferico e bem agradável no morro dos urubus'
2 amm.proc_text(text = texto, verbs=False, string=False, text_add = '',adjectives=False, plural=False)

['passeio', 'teleferico', 'agradavel', 'morro', 'urubus']
```

Fonte: elaborado pelo autor (2021).

Lyfenko (2014) relata que dentro de um trabalho de classificação automática de documentos, dependendo do objeto do estudo, é interessante proceder com a remoção de verbos e adjetivos. No exemplo das avaliações de atrações turísticas, entendeu-se que os verbos e adjetivos acrescentariam pouca informação relevante para caracterizar

um destino, por isso, optou-se pela sua remoção. Além dos verbos e adjetivos, o nome da atração e o nome do destino se mostraram termos muito frequentes nas avaliações de determinada atração, como no exemplo dessa avaliação do Museu do Louvre em Paris: “O Louvre não se resume só as suas obras, das quais não preciso nem mencionar o valor, mas o prédio todo, o entorno, a pirâmide e principalmente, à noite...Se for a Paris, mesmo que numa passagem rápida, não deixe de passear pelo entorno”¹⁶. Nesse sentido, os nomes próprios da atração e destino não acrescentam valor para extração de características que identifiquem a atração com o propósito de comparação, sendo assim, adotou-se a remoção desses termos de cada corpus. Ao entender a necessidade de se remover os verbos e adjetivos, se deparou com outra realidade: a inexistência de bases livres de verbos e adjetivos em português. Para contornar essa situação, novamente foram criadas duas listas, uma lista com 260.524 verbos e outra com 6.626 adjetivos. As listas foram criadas com base em extrações de dados da Internet usando *crawlers*. Para criar a lista de verbos, usou-se o verbo no seu infinitivo e todas as suas conjugações possíveis. A Figura 22 exibe um exemplo desse processo. No exemplo, é possível visualizar que os adjetivos (itinerantes, gigantesco e belo) foram removidos do texto final. Os verbos (reserve, estiver, visitar, existe e deixe) também foram removidos. Os nomes (Paris e Louvre) também foram removidos do texto final.

Figura 22 – Processo de remoção de adjetivos, verbos e nomes próprios

```

1 avaliacao = "Quando estiver em Paris, reserve pelo menos um dia para visitar o belo Museu do Louvre. \
2 Existe um acervo gigantesco de obras de arte (pinturas e esculturas). Nao deixe de visitar o \
3 apartamento do imperador Napoleao, alem das exposicoes itinerantes"
4
5 sem_verbos_e_adjectives = amm.proc_text(text = avaliacao, verbs=True, string=False, text_add = 'Paris Louvre',
6     adjectives=True, plural=True)
7
8 print(sem_verbos_e_adjectives)
9
['museu', 'acervo', 'arte', 'pintura', 'escultura', 'apartamento', 'imperador', 'napoleao', 'exposicao']

```

Fonte: elaborado pelo autor (2021).

O próximo passo dentro do pré-processamento foi a análise de viabilidade da aplicação de regras de Lematização e Stemização nos textos. Avaliando-se as opções disponíveis em português, não foram encontradas bibliotecas com processamento adequado dessas

¹⁶Disponível em: https://www.tripadvisor.com.br/Attraction_Review-g187147-d188757-Reviews-Louvre_Museum-Paris_Ile_de_France.html. Acesso em: 8 fev. 2021.

técnicas. A Figura 23 exibe um exemplo de Stemização da palavra paisagem usando o pacote de português do NLTK. No exemplo, é possível ver que a palavra foi contraída para sua raiz *país*, porém considerando a remoção de acento, essa nova palavra possui dois significados semânticos diferentes da palavra base. Dessa forma, entendeu-se que tais processos, apesar de importantes, poderiam comprometer os resultados.

Figura 23 – Exemplo de processo de Stemização

```
1 import nltk
2 stemmer = nltk.stem.RSLPStemmer()
3 print(stemmer.stem("paisagem"))
```

país

Fonte: elaborado pelo autor (2021).

O último passo no pré-processamento foi a criação de bigramas e trigramas. Esse processo permite concatenar dois termos que ocorrem com muita frequência juntos (bigrama) e três termos que ocorrem com muita frequência juntos (trigrama). Utilizou-se o parâmetro mínimo de cinco ocorrências, ou seja, para formar os bigramas e trigramas, os termos deveriam aparecer juntos no mínimo cinco vezes. Para se avaliar essa frequência, todos os documentos (avaliações) do *corpus* de cada atração são considerados. A Figura 24 exibe um exemplo de bigramas gerados por esse processo.

Figura 24 – Processo de criação de bigramas e trigramas

```
1 print(corpus_atracoes[0][3][99])
```

('Maravilhoso! Um espetáculo este museu! Incrível mesmo, parada obrigatória para quem aprecia artes e também para os "simp atizantes". Tem obras de Vincent van Gogh, Picasso, TARSILA DO AMARAL (Exposição só dela no segundo andar!), Frida Kahlo, Salvador Dalí, Henri Matisse, Claude Monet, entre muitos outros. Super...')

exemplo de teste processado do perfil cultura

```
1 print(corpus_label_proc['Cultura'][99])
```

['museu', 'parada_obrigatoria', 'arte', 'simpatizante', 'vincent_van_gogh', 'picasso', 'tarsila_amaral', 'frida_kahlo', 'salvador_dali', 'henri_matisse', 'claudes_monet']

Fonte: elaborado pelo autor (2021).

Os procedimentos adotados na etapa de pré-processamento foram fundamentais para as próximas etapas, pois, permitiu-se extrair dos textos, os detalhes mais importantes para determinar as características de uma atração ou perfil. O trabalho técnico desenvolvido é importante também para futuros estudos no idioma português, que podem fazer uso das listas e recursos gerados. Com o *corpus* das atrações e dos perfis

processados, o próximo passo consistiu em representar os textos de forma que o computador conseguisse interpretá-los.

3.4.2 Representação dos textos

Um computador consegue processar apenas números e para que o processo de análise textual ocorra, é necessário transformar os termos (tokens) em números por meio de um processo de representação. Um modelo de representação apropriado é essencial para obter sucesso num processo de classificação (HARISH; GURU; MANJUNATH, 2010). Nesse processo de representação, a frequência das palavras pode ser considerada ou não, seu peso no documento ou na coleção também e até mesmo sua semântica. No modelo espaço-vetor, cada avaliação se torna um vetor de números, ou seja, o conjunto de avaliações de uma atração se torna um conjunto de vetores que representam tais avaliações.

No trabalho foram utilizadas duas abordagens, uma considerando os métodos TF e TF-IDF e outra considerando o método BOW e *word embeddings* (casamento de palavras). Os métodos de representação foram aplicados usando as bibliotecas Sklearn e Gensim. Para normalizar o tamanho dos vetores, foi utilizada a norma Euclidiana. Essa norma, também chamada de *l2-norm* é a mais comum para medir um tamanho de um vetor (PERONE, 2020b).

3.4.3 Representação com TF e TF-IDF

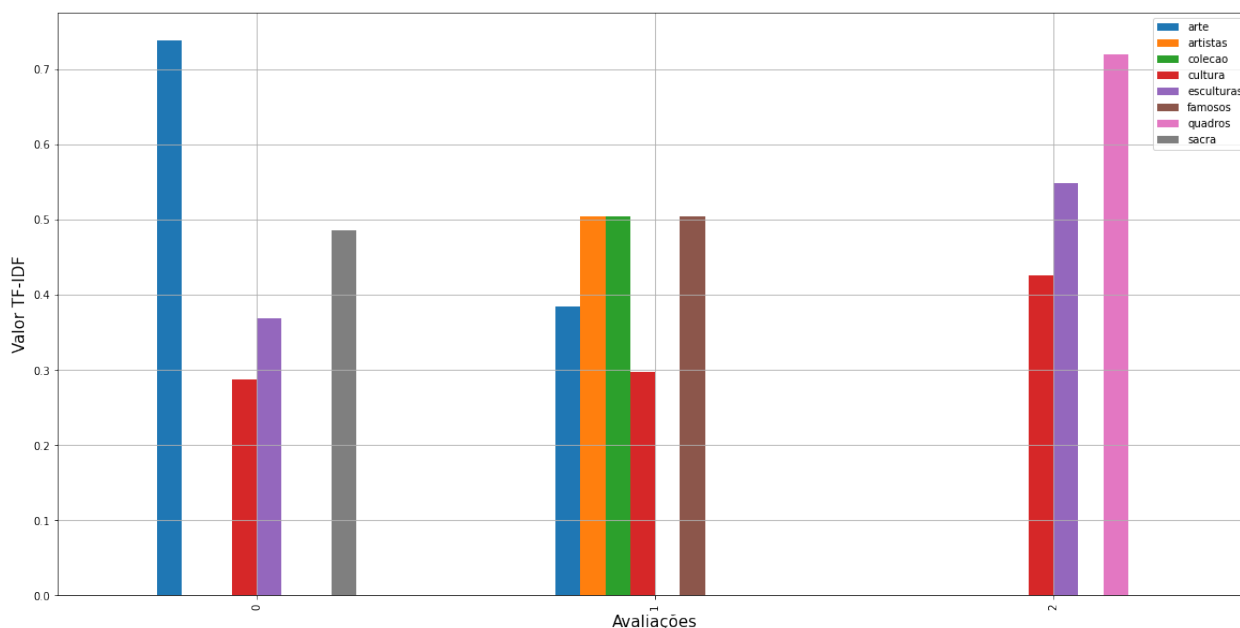
Na primeira abordagem, o corpus de cada perfil e cada atração foram representados usando os métodos TF e TF-IDF. Essa representação é usada tanto nos algoritmos de classificação supervisionada quanto na comparação de representações. Para ilustrar a representação do método TF-IDF, o Quadro 5 apresenta três exemplos de avaliações.

Quadro 5 – Exemplo de avaliações

ID AVALIAÇÃO	TEXTO AVALIAÇÃO
0	arte e cultura – esculturas e arte sacra
1	coleção artistas famosos cultura e arte
2	quadros e esculturas cultura

Fonte: elaborado pelo autor (2021).

A partir dos dados do Quadro 5, foi realizado uma representação usando o método TF-IDF com a ferramenta Sklearn. O Gráfico 4 exibe uma visualização gráfica do resultado.

Gráfico 4 – Resultado de representação Termo Frequência – Inverso Documento
Frequência (TF-IDF)

Fonte: elaborado pelo autor (2021).

No exemplo do Gráfico 4, é possível visualizar o resultado TF-IDF para cada avaliação usando a normalização l_2 -norm. Para cada termo, um valor é calculado considerando sua frequência na avaliação e sua frequência no *corpus*. O termo *cultura* que possui frequência em todas as avaliações apresenta valor menor de relevância TF-IDF. A lógica é, se o termo é encontrando com frequência no *corpus*, então ele não seria um bom termo para representar determinado documento. Diferente do termo *arte*, que se repete duas vezes no primeiro documento e não aparece em todos os documentos, recebendo assim

um valor de representação maior. Seguindo esses modelos, foram geradas as representações TF, quanto TF-IDF de cada *corpus*.

3.4.4 Representação com *word embeddings*

A representação TF-IDF e TF considera a frequência dos termos nos documentos e coleção, no entanto, não considera aspectos semânticos, como a proximidade de palavras em um texto (QAISER; ALI, 2018). Como forma de contornar esse tipo de situação, outra abordagem foi utilizada, o método de *word embeddings*. Se *a* e *b* são duas palavras e *x* e *y* são suas representações *word embeddings*, então espera-se que a distância entre *x* e *y* seja um indicativo da relação semântica entre as duas palavras (ROY *et al.*, 2016). O método foi aplicado usando o algoritmo *Word2vec* da biblioteca Gensim. O algoritmo é uma rede neural e como tal, requer uma grande quantidade de dados para aprender a semântica entre os termos (MIKOLOV *et al.*, 2013).

Inicialmente avaliou-se a utilização de modelos já existentes, como o repositório de *word embeddings*¹⁷ da Universidade de São Paulo (USP). Os modelos disponíveis foram treinados com base em 17 *corpora* com diferentes textos em português, considerando fontes como Wikipédia, GoogleNews, Portal G1 da Globo entre outros. Esse repositório possui diferentes versões dos modelos com treinos usando a lógica SKIP-GRAM e CBOW, com dimensões (quantidade de palavras de cada vetor) distintas. Foram avaliadas as versões com SKIP-GRAM e CBOW com dimensão 100. Para avaliar a qualidade dos modelos, foi adotada uma análise qualitativa do resultado semântico retornado pelos modelos da USP perante alguns termos comuns do turismo. A Figura 25 exibe um exemplo de termos semânticos retornados pelo modelo SKIP-GRAM para a palavra *turista*.

¹⁷Disponível em: <http://www.nilc.icmc.usp.br/embeddings>. Acesso em: 5 fev. 2021.

Figura 25 – Exemplo resultado modelo Universidade de São Paulo (USP)

```

1 #retorno word2vec processado usp
2 model.most_similar("turista")

[('lojista', 0.8308435678482056),
 ('atleta', 0.8226292133331299),
 ('jogador', 0.8161439895629883),
 ('molecote', 0.8098851442337036),
 ('mergulhador', 0.8092334866523743),
 ('cliente', 0.8082325458526611),
 ('joalheiro', 0.8081656098365784),
 ('viajante', 0.8071156740188599),
 ('ouvinte', 0.8071004152297974),
 ('estelionatário', 0.8060380220413208)]

```

Fonte: elaborado pelo autor (2021).

O retorno do modelo é com base no que a rede neural aprendeu durante o treinamento. Depende de dois fatores essenciais: a quantidade de dados, que no caso são muitos e o conteúdo dos dados, que no caso, são variados, pois, uma base de notícias pode envolver o uso da mesma palavra em contextos distintos (FERNÁNDEZ-REYES; HERMOSILLO-VALADEZ; MONTES-y-GÓMEZ, 2018). É importante avaliar esse fator, pois uma base com conteúdo direcionado, exemplo: artigos da PubMed utilizados num trabalho de medicina, pode apresentar resultados superiores por ser um conteúdo específico.

Com base nesse pressuposto, foi realizada a construção de um modelo de *word embeddings* específico para o turismo. Um dos fatores importantes nessa construção é o fato de existir uma base específica do turismo (avaliações turísticas) em uma grande quantidade (3,4 milhões). Como o corpus de atração já havia sido pré-processado e estava pronto para ser representado, foi realizada uma junção de 3 milhões de avaliações de turistas para se tornar a fonte da construção do modelo. Utilizou-se uma janela de quatro palavras (quatro antes e quatro depois) do termo em análise tanto na lógica SKIP-GRAM, quanto CBOV. O modelo foi gerado com um dicionário com 104702 palavras únicas extraídas do corpus de avaliações. A Figura 26 exhibe um exemplo de termos semânticos retornados para a palavra *turista*.

Figura 26 – Exemplo resultado usando o método próprio

```

1 #retorno wor2vec processado local
2 model.most_similar("turista")

[('visitante', 0.7681383490562439),
 ('viajante', 0.7342207431793213),
 ('turismo', 0.7265468835830688),
 ('extrangeiro', 0.7189979553222656),
 ('turistica', 0.7138558626174927),
 ('tursita', 0.7137143015861511),
 ('turisca', 0.7072365880012512),
 ('turistada', 0.700389564037323),
 ('contingente', 0.697627067565918),
 ('atrae', 0.6929479837417603)]

```

Fonte: elaborado pelo autor (2021).

Analisando o resultado dos modelos da USP (genérico) e dos modelos específicos (turismo), concluiu-se que os modelos gerados com conteúdo específico apresentaram resultados mais apropriados para o estudo atual. É importante afirmar que o objeto do estudo impacta diretamente nessa decisão, pois, se fosse outra área de estudo, os modelos da USP poderiam ser mais indicados. Os modelos gerados foram testados tanto na abordagem SKIP-GRAM quanto CBOW, e não se notou grandes diferenças nas duas, o método SKIP-GRAM foi escolhido para utilização no trabalho, porque procura encontrar termos relacionados com base num termo específico.

3.4.5 Representação com *bag-of-words* (BOW)

A representação pelo método BOW foi utilizada numa situação específica. Na comparação entre *corpus* representados com o método *word embeddings*. O método BOW considera a frequência de termos dentro de uma avaliação independente de sua posição, tamanho da avaliação ou frequência no *corpus*. Exatamente por isso, a tradução do termo BOW para português é “saco de palavras”, ou seja, não importa a ordem. A Figura 27 exibe um exemplo de um *corpus* representado com o método BOW.

Figura 27 – Exemplo de representação *bag-of-words* (BOW)

```

1 from sklearn.feature_extraction.text import CountVectorizer
2 import pandas as pd
3 corpus = ["arte e cultura - arte sacra", "colecão com artistas famosos cultura e arte",
4          "quadros e esculturas cultura"]
5 count = CountVectorizer()
6 bag_of_words = count.fit_transform(corpus)
7 df = pd.DataFrame(bag_of_words.toarray(), columns=count.get_feature_names())
8 df.index.name = 'Avaliações'
9 df.head()

```

	arte	artistas	colecão	com	cultura	esculturas	famosos	quadros	sacra
Avaliações									
0	2	0	0	0	1	0	0	0	1
1	1	1	1	1	1	0	1	0	0
2	0	0	0	0	1	1	0	1	0

Fonte: elaborado pelo autor (2021).

No exemplo, o termo *arte* aparece duas vezes na avaliação 0. A representação não mantém a ordem das palavras em cada avaliação. Embora o método tenha algumas desvantagens, ele é importante para se obter uma representação rápida de um *corpus* considerando apenas frequência de palavras em cada documento. A Tabela 2 apresenta os cinco termos mais frequentes encontrados no corpus de cada perfil usando o método BOW.

Tabela 2 – Termos mais frequentes no *corpus* de cada perfil

PERFIL	TERMO	FREQUÊNCIA	PERFIL	TERMO	FREQUÊNCIA
Aventura	Brinquedo	909	Paisagem / Arquitetura	Sol	561
	Atração	803		Atração	286
	Parque	801		Foto	285
	Buggy	551		Parque	276
	Emoção	423		Catedral	259
Compras	Loja	2805	Praia	Mar	1024
	Preço	1202		Água	939
	Restaurante	900		Restaurante	705
	Outlet	874		Praia	450
	Shopping	806		Opção	367
Cultura	Arte	573	Relaxante	Barco	710
	Acervo	465		Sena	626
	Ruína	328		Gondola	356
	Coliseu	321		Margem	337
	Artista	251		Jeri	310
Família	Atração	1036	Religioso	Igreja	995
	Criança	902		Paz	628
	Parques	708		Arquitetura	324
	Brinquedo	556		Beleza	274
	Diversão	501		Natureza	253
Gastronomia	Cerveja	839	Romântico	Restaurante	873
	Vinho	789		Pedalinho	626
	Degustação	786		Mar	421
	Bar	742		Água	408
	Restaurante	708		Foto	336
Natureza / Exóticos	Natureza	486	Vida Noturna	Restaurante	1314
	Foto	436		Bar	885
	Prédio	323		Noite	882
	Água	281		Cassino	794
	Beleza	271		Loja	562

Fonte: elaborado pelo autor (2021).

Com as coleções representadas, o próximo passo dentro do Nível Tecnológico é o processo de classificação. Para a classificação usou-se duas abordagens: a primeira considerando a comparação de similaridade entre os documentos representados e a segunda usando uma abordagem clássica supervisionada.

3.4.6 Classificação por similaridade da representação

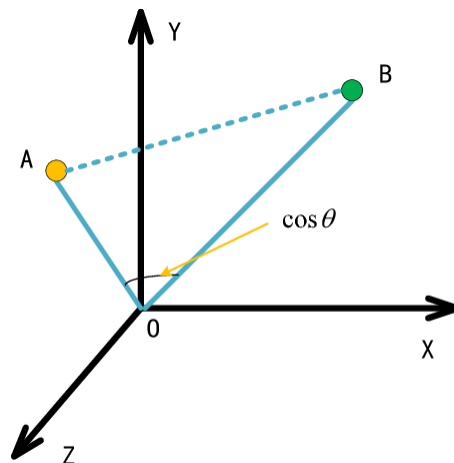
No Nível Tecnológico, as etapas anteriores foram preparatórias para o processo dessa etapa. No trabalho adotou-se duas formas de classificação. A primeira é considerando a

similaridade apenas das representações geradas na etapa anterior. A segunda será exibida na sessão 3.4.7 e refere-se a uma abordagem mais clássica usando AM. Conforme aponta Lyfenko (2014), num processo de classificação, é possível verificar o grau de aderência de determinado documento perante uma classe, com base na comparação da representação dos vetores da classe e do documento. Essa forma de comparação procura a similaridade de todos os documentos diante de uma determinada classe medindo o cosseno do ângulo entre dois vetores. A comparação de similaridade entre as representações é também muito utilizada em processos de recuperação da informação, pois, apenas o objetivo muda, ou seja, retornar informação relevante ao usuário e não classificar. Entretanto, a necessidade de medir a similaridade de dois documentos é a mesma.

3.4.6.1 Verificação de similaridade

A similaridade entre dois documentos pode ser avaliada pelo cosseno do ângulo entre dois vetores, assim como outras medidas, como distância Jaccard, a distância Euclidiana e divergência Jensen Shannon (LJUBESIC *et al.*, 2008). A escolha do método depende do estudo em análise. Nessa pesquisa, adotou-se a comparação de similaridade entre os vetores considerando o cosseno do ângulo destes, por geralmente apresentar bons resultados (NÁTHER, 2005; CAVNAR; TRENKLE, 1994). O Gráfico 5 exibe um exemplo de similaridade por cosseno.

Gráfico 5 – Similaridade entre documentos por cosseno do ângulo



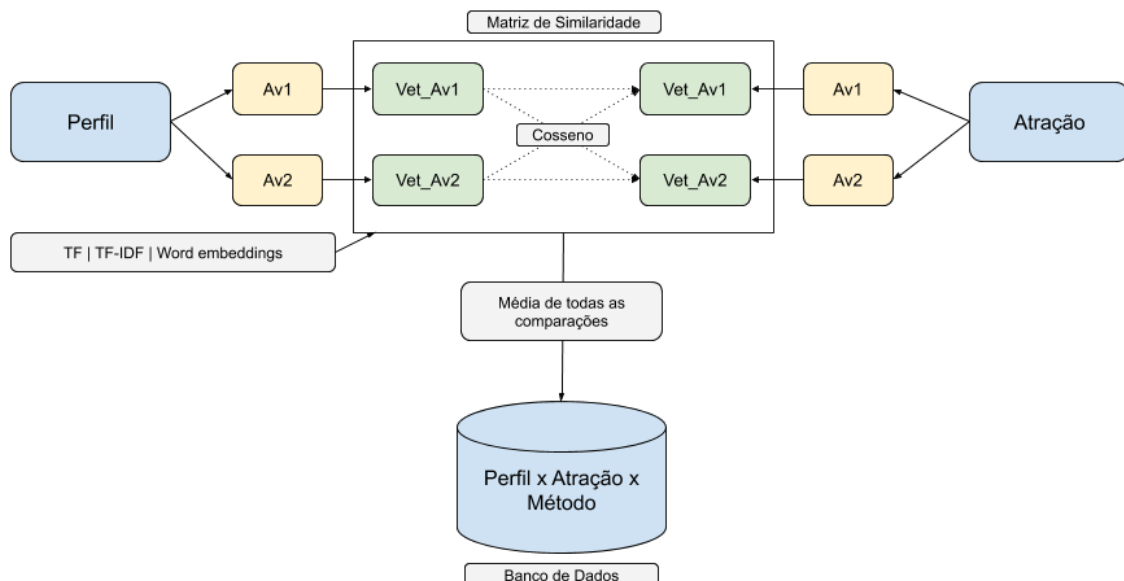
Fonte: Havan (2019).

No Gráfico 5, é possível visualizar dois documentos A e B projetados no espaço multidimensional Y, X e Z. É possível observar que a distância entre os vetores A e B que representam o texto de duas avaliações R1 e R2 é medida pelo cosseno do ângulo entre esses vetores $\cos\theta$. Assim, a similaridade entre as avaliações R1 e R2 é representada pelo valor de $\cos\theta$ e à medida que o ângulo entre os vetores diminui, a similaridade das avaliações aumenta. Quando duas avaliações são completamente iguais, o seu valor de cosseno é igual a 1, dessa forma, quanto maior o cosseno, maior a similaridade entre duas avaliações (PERONE, 2020a).

3.4.6.2 Arquitetura do processo classificatório

A comparação de similaridade entre as representações vetoriais de duas avaliações forma a base do processo. De um lado, tem-se um *corpus* para cada um dos 12 perfis e de outro lado um *corpus* para cada uma das 3263 atrações. Essas coleções são então representadas usando os métodos TF, TF-IDF e *word embeddings*. Uma matriz de similaridade considerando todas as avaliações do perfil perante todas as avaliações da atração foi gerada para cada par (Perfil x Atração). O resultado final de similaridade entre o Perfil e a Atração foi dado pela média aritmética dessa matriz de similaridade. A Figura 28 exibe a arquitetura desse processo de classificação.

Figura 28 – Arquitetura processo de classificação por similaridade de representação



Fonte: elaborado pelo autor (2021).

Na Figura 28, é possível observar as avaliações Av1 e Av2 de determinado perfil sendo representadas pelos vetores Vet_Av1 e Vet_Av2 usando TF, TF-IDF e *word embeddings*. O mesmo ocorre para a atração. Para gerar a matriz de similaridade, comparações múltiplas são realizadas entre os vetores de todas as avaliações da atração e do perfil. A média de todas essas comparações é o resultado final de similaridade entre (Perfil x Atração). Após o processamento, todos os resultados são salvos no banco de dados considerando o perfil, a atração e o método que gerou aquele resultado. Salvar esses dados facilitou o Nível 3 – Validação e análise dos resultados.

Após realizar uma análise sobre a quantidade de avaliações por atração, notou-se que havia muitas atrações com apenas uma, duas ou três avaliações. Após alguns testes de classificação com ambos os métodos, percebeu-se esse número baixo de avaliações estava comprometendo o resultado. Além disso, não é possível extrair padrões de um número tão pequeno de opiniões. Foram realizados testes com o objetivo de encontrar um número mínimo adequado de avaliações para que cada atração pudesse ser classificada. Após essa análise, identificou-se o número mínimo de 30 avaliações por cada atração. Com isso, 636 atrações foram removidas do estudo, por não satisfazer os critérios mínimos para qualidade da classificação. No total, 2.627 avaliações estariam aptas para o estudo, sendo que o corpus de cada uma foi comparado com o corpus de cada perfil.

A similaridade entre o perfil e a atração foi dada pela comparação entre todas as avaliações do perfil perante todas as avaliações da atração. Ou seja, uma atração com 100 avaliações gera 300 mil comparações para cada perfil (100 avaliações da atração x 3.000 do perfil). Assim, para analisar o grau de relevância de uma atração perante os 12 perfis foram necessárias 3,6 milhões de comparações. Embora a comparação usando o cosseno do vetor seja um processo simples logicamente, em termos de computação, esse processo se mostrou muito demorado. A seguir apresenta-se mais detalhes sobre o processo de classificação usando TF, TF-IDF e *word embeddings*.

3.4.6.3 Classificação usando Termo Frequência (TF) e Termo Frequência – Inverso Documento Frequência (TF-IDF)

Uma vez definida a estratégia de classificação e estando os textos representados, foi possível iniciar a classificação das atrações perante os perfis. O método TF e TF-IDF possui implementações praticamente similares dentro da ferramenta Sklearn. A diferença é a atribuição do parâmetro `use_idf` que deve ser configurado para `True`, quando o objetivo é usar TF-IDF. Se esse parâmetro é configurado para o valor `False`, então o algoritmo usa o cálculo de TF para similaridade. Para calcular a similaridade do cosseno foi utilizado a biblioteca `Pairwise` do Sklearn. As figuras 29 a 33 exibem uma exemplificação do processo de classificação usando a representação TF-IDF. A Figura 29 exhibe o processo de criação de um *corpus* para os perfis *Praia* e *Cultura* e para uma atração *Maragogi*.

Figura 29 – Criação de *corpus* para as avaliações de atração e perfis

```

1 from sklearn.feature_extraction.text import TfidfVectorizer
2 from sklearn.metrics.pairwise import cosine_similarity
3 import pandas as pd
4
5 #criando corpus de avaliações
6 maragogi = ["O ceu é azul", "O sol é brilhante", "O sol no céu é brilhante",
7            "Podemos ver o sol brilhante e reluzente"]
8 praia    = ["Praia linda com céu azul", "O sol estava quente e brilhando", "Agua azul e brilhante"]
9 cultura  = ["Museus com vários artistas", "Sol quente, mas apreciamos a arquitetura do prédio"]
10
11 corpus_completo = maragogi + praia + cultura
12 corpus_completo

['O ceu é azul',
'O sol é brilhante',
'O sol no céu é brilhante',
'Podemos ver o sol brilhante e reluzente',
'Praia linda com céu azul',
'O sol estava quente e brilhando',
'Agua azul e brilhante',
'Museus com vários artistas',
'Sol quente, mas apreciamos a arquitetura do prédio']

```

Fonte: elaborado pelo autor (2021).

Na Figura 29, é possível visualizar que foi necessário fazer uma junção do *corpus* dos perfis com o *corpus* da atração exibido na variável *corpus_completo*. Após a junção do *corpus*, o próximo passo foi representá-lo usando TF-IDF, conforme exibido na Figura 30.

Figura 30 – Representação do *corpus* usando Termo Frequência – Inverso Documento Frequência (TF-IDF)

```

1 # #representando corpus
2 tfidf_vectorizer = TfidfVectorizer(norm='l2',use_idf=True, ngram_range=(1,3))
3 tfidf_matrix     = tfidf_vectorizer.fit_transform(corpus_completo)
4 tfidf_matrix.shape

(9, 68)

```

Fonte: elaborado pelo autor (2021).

O método TF-IDF cria uma matriz que representa documentos x palavras. A matriz TF-IDF gerada a partir do *corpus_completo* possuiu nove documentos e 68 palavras. Com o *corpus* tanto da atração quanto dos perfis representado em TF-IDF, o próximo passo é o cálculo da similaridade de cada perfil com a atração. A Figura 31 exibe a comparação de similaridade da atração Maragogi com o perfil Cultura:

Figura 31 – Similaridade entre Maragogi e o perfil Cultura

```

: 1 #calculado similaridade avaliacoes de atracao com avalicoes do perfil cultura
2 matriz_similaridade_cultura_maragogi = cosine_similarity(tfidf_matrix[7:9], tfidf_matrix[:4])
3 print("A similaridade entre o perfil cultura e a atração Maragogi é:",
4       matriz_similaridade_cultura_atracao.mean())

A similaridade entre o perfil cultura e a atração Maragogi é: 0.015221068575408904

```

Fonte: elaborado pelo autor (2021).

O método *cosine_similarity* da biblioteca Sklearn foi utilizado para encontrar o cosseno dos vetores. Como todas as avaliações estão no mesmo *corpus*, foi necessário dividir o *corpus* para comparar cada perfil. Para chegar no resultado final, é feita a média aritmética. A Figura 32 exibe o mesmo processo de comparação da atração, porém agora com o perfil Praia.

Figura 32 – Similaridade entre Maragogi e o perfil Praia

```

1 #calculado similaridade avaliacoes de atracao com avalicoes do perfil praia
2 matriz_similaridade_praia_atracao = cosine_similarity(tfidf_matrix[4:7], tfidf_matrix[:4])
3 print("A similaridade entre o perfil praia e a atração Maragogi é", matriz_similaridade_praia_atracao.mean())

A similaridade entre o perfil praia e a atração Maragogi é 0.06614337023832124

```

Fonte: elaborado pelo autor (2021).

O mesmo processo é adotado apenas separando as avaliações específicas do perfil Praia. A Figura 33 exibe o resultado de todo o processo de classificação.

Figura 33 – Resultado de similaridade entre a atração e os perfis

```

1 #exibindo resultado
2 media_cosseno = [matriz_similaridade_praia_atracao.mean(),matriz_similaridade_cultura_atracao.mean()]
3 resultado = pd.DataFrame(media_cosseno, index=['Praia', 'Cultura'], columns = ['Similaridade'] )
4 resultado.index.name = 'Perfil'
5 resultado

```

Similaridade	
Perfil	
Praia	0.066143
Cultura	0.015221

Fonte: elaborado pelo autor (2021).

Analisando o resultado expresso na Figura 33, é possível visualizar que as avaliações da atração Maragogi são classificadas em dois perfis (Praia e Cultura). Os valores indicam uma similaridade muito maior da atração Maragogi em relação ao perfil Praia do que Cultura. Assim, a atração Maragogi é mais relevante ou pertinente para o perfil Praia.

Embora a classificação usando os métodos TF e TF-IDF tenha sido simples de implementação e entendimento, o processo foi demorado em termos computacionais. Usando um servidor hospedado na Amazon São Paulo, com quatro núcleos e 16GB de memória, o processamento da classificação demorou cerca de 5 horas para cada atração, totalizando 60 horas de processamento, ou seja, 2,5 dias. À medida que cada atração era classificada, seu resultado era salvo no banco de dados, pois se havendo qualquer falha no processamento, a retomada do processo se daria a partir do ponto de parada. As classificações usando TF e TF-IDF não consideram a semântica entre as palavras. Como forma de sanar esse problema, o próximo passo exhibe um processo e classificação usando *word embeddings*.

3.4.6.4 Classificação usando casamento de palavras

A comparação de similaridade usando cosseno do ângulo dos vetores não considera a semântica textual. Nesse caso se existem dois documentos A e B, respectivamente com os termos *software computador* e *programa computador*, os dois documentos teriam valor de cosseno igual a 0, ou seja, completamente diferentes. Como forma de solucionar esse problema, Sidorov *et al.* (2014) apresentam o conceito de Medida de Cosseno Leve “*Soft Cosine Measure*” (SCM), que permite avaliar a similaridade entre objetos considerando a semântica do texto. Os autores usam um novo modelo matemático para atingir a

similaridade entre os textos, porém usando representações dos mesmos com casamento de palavras “*word embeddings*”. Os resultados dos autores superam o método tradicional de similaridade por cosseno.

Nesse trabalho, usou-se esse método como forma de identificar a qualidade de uma classificação semântica sobre as demais. Para realizar o processo, foi utilizada a ferramenta Gensim. Para que fosse possível realizar essa classificação, o modelo de casamento de palavras criado com o algoritmo *Word2vec*, cujo desenvolvimento relatado na seção 3.4.4, foi utilizado. As próximas figuras exibem uma exemplificação desse processo de classificação. Na Figura 34, o modelo *Word2vec* é carregado e os termos mais similares a palavra arquitetura são exibidos.

Figura 34 – Carregando modelo *Word2vec*

```

1 %%time
2 #word2vec
3 model_name = 'w2v_cities'
4 model_save = '../..../datasets/fugiu/labels/'+model_name+'.bin'
5 model = KeyedVectors.load_word2vec_format(model_save, binary=True, unicode_errors='ignore')
6 model.init_sims(replace=True) # replace raw vectors in-place with unit-normed ones
7 model.most_similar("arquitetura")

```

CPU times: user 3.44 s, sys: 68 ms, total: 3.5 s
Wall time: 3.5 s

```

[('arquitetonica', 0.8422379493713379),
 ('construcao', 0.8371604681015015),
 ('edificacao', 0.8103249073028564),
 ('arquitetuta', 0.7993497848510742),
 ('arquiteura', 0.7925085425376892),
 ('arquiterura', 0.7882814407348633),
 ('aquitetura', 0.7491632103919983),
 ('arquietura', 0.7444842457771301),
 ('arquitertura', 0.7420250177383423),
 ('arquitura', 0.7370285391807556)]

```

Fonte: elaborado pelo autor (2021).

No exemplo, é possível visualizar que os dois termos mais próximos a palavra *arquitetura* são *arquitetônica* e *construção*. A Figura 35 exhibe os documentos e seus respectivos índices, de 0 a 8. Tais índices são utilizados pela função de similaridade para retornar um par (índice, SCM) para cada documento que o método encontra similaridade.

Figura 35 – *Corpus* dos perfis Praia e Cultura com índices

```

1 #criando corpus de avaliações
2 praia = ["Praia linda com céu azul", "O sol estava quente e brilhando", "Água azul e brilhante"]
3 cultura = ["Museu com vários artistas", "Sol quente, mas apreciamos a arquitetura do prédio"]
4
5 corpus_completo = praia + cultura
6 documentos = [d.split(" ") for d in corpus_completo]
7 aux=0
8 for d in corpus_completo:
9     print(aux,d)
10    aux+=1

```

0 Praia linda com céu azul
1 O sol estava quente e brilhando
2 Água azul e brilhante
3 Museu com vários artistas
4 Sol quente, mas apreciamos a arquitetura do prédio

Fonte: elaborado pelo autor (2021).

Os documentos com índices de 0 a 2 representam o *corpus* do perfil Praia. Os documentos com índice 3 a 4 representam o *corpus* do perfil Cultura. A Figura 36 exibe a função desenvolvida para classificar um texto (parâmetro avaliação) perante um *corpus* (parâmetro perfil_corpus) usando o modelo *Word2vec* (parâmetro *word2vec_model*).

Figura 36 – Função para classificação usando algoritmo *Word2vec*

```

1 def softcosinesim(word2vec_model,perfil_corpus,avaliacao):
2     termsim_index = WordEmbeddingSimilarityIndex(word2vec_model)
3     dictionary = corpora.Dictionary(perfil_corpus+avaliacao)
4     bow_corpus = [dictionary.doc2bow(document) for document in perfil_corpus]
5     similarity_matrix = SparseTermSimilarityMatrix(termsim_index, dictionary)
6     docsim_index = SoftCosineSimilarity(bow_corpus, similarity_matrix,num_best=10)
7     bow_query = [dictionary.doc2bow(document) for document in avaliacao]
8     sims = docsim_index[bow_query]
9     return sims

```

Fonte: elaborado pelo autor (2021).

Essa função recebe os parâmetros modelo *Word2vec*, *corpus* do perfil, o texto de uma avaliação e retorna uma lista com pares (índice e valor SCM). Uma matriz de similaridade é criada considerando o modelo *Word2vec* e um dicionário formado pela junção dos textos do perfil e da avaliação. Conforme exibido na seção 3.4.5, o modelo BOW é usado para representar os textos com base na frequência de palavras. Embora a representação seja feita com BOW, a função projeta os textos no espaço vetor considerando uma base não ortogonal, ou seja, o ângulo entre os vetores (similaridade) é derivado do ângulo dos vetores do *Word2vec* correspondentes a cada palavra (SIDOROV *et al.*, 2014). Na Figura 37, um exemplo de texto de avaliação é comparado com as avaliações do perfil Praia. Nenhuma das palavras usadas nesse texto de exemplo de avaliação existe tanto no perfil Praia quanto Cultura.

Figura 37 – Comparação do perfil Praia com a avaliação exemplo

```

1 #comparação com perfil praia
2 avaliacao = [['azulzinho', 'arquitetonica', 'astro_rei']]
3 resultado = softcosinesim(word2vec_model, documentos[0:3], avaliacao)[0]
4 print("A similaridade com o perfil Praia é: ", sum([d[1] for d in resultado])/5)

```

A similaridade com o perfil Praia é: 0.05095949172973633

Fonte: elaborado pelo autor (2021).

Na Figura 37 é possível visualizar que os termos *azulzinho* e *astro_rei* remetem aos termos céu e sol que existem nas duas avaliações do perfil Praia. O modelo de classificação retorna uma compatibilidade de 0,0509 de valor SCM para a similaridade entre o perfil Praia e essa avaliação. A Figura 38 exibe a comparação do perfil Cultura com a mesma avaliação.

Figura 38 – Comparação do perfil Cultura com a avaliação exemplo

```

1 #comparação com perfil cultura
2 avaliacao = [['azulzinho', 'arquitetonica', 'astro_rei']]
3 resultado = softcosinesim(word2vec_model, documentos[4:6], avaliacao)[0]
4 print("A similaridade com o perfil Cultura é: ", sum([d[1] for d in resultado])/5)

```

A similaridade com o perfil Cultura é: 0.027233907580375673

Fonte: elaborado pelo autor (2021).

Nesse exemplo, o valor retornado de SCM é menor, pois apenas um documento do perfil Cultura possui relevância para essa avaliação. O documento com índice 4: “Sol quente, mas apreciamos a arquitetura do prédio”. Na avaliação de exemplo, há o termo *arquitetônica*, que conforme exibido na Figura 38 é o termo mais próximo retornado pelo modelo Word2vec para a palavra *arquitetura*.

É importante ressaltar que se algum desses exemplos fossem aplicados sem considerar a abordagem de casamento de palavras, a similaridade seria zero. Essa demonstração comprova a validade do método para um modelo de classificação semântico. Embora os resultados sejam promissores ocorreram limitações computacionais na aplicação desse método. O processo de classificação apresentou-se extremamente lento, exatamente por considerar a semântica entre os textos, cada palavra se torna um vetor no espaço e cada documento se torna um vetor com vetores de palavras. Diversas tentativas de otimizar a performance foram realizadas, como contratar servidores mais potentes, melhorar os

parâmetros do modelo de classificação e uso de outras implementações, como da classe `matutils` do Gensim. O melhor cenário foi adotado no trabalho e o processamento demorou 10 dias para ser concluído com os respectivos resultados salvos em banco de dados.

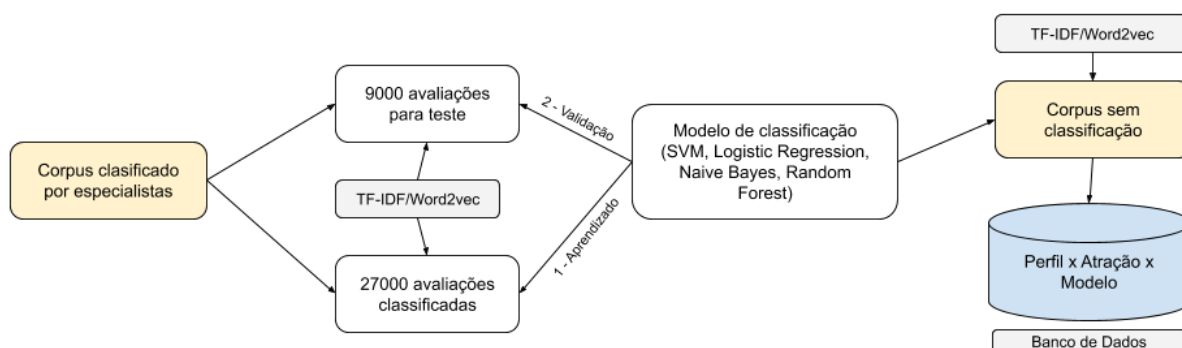
3.4.7 Classificação supervisionada

A seção anterior descreveu o processo de classificação usando a similaridade de representação dos documentos. Nessa seção, outro método de classificação utilizado na pesquisa é descrito. O processo de classificação supervisionada segue uma linha tradicional usando algoritmos de AM. A lógica de classificação é similar à anterior, ou seja, a partir de um corpus com atrações classificadas pelos especialistas, procura-se classificar o restante das atrações.

3.4.8 Arquitetura do processo de classificação supervisionada

A arquitetura do processo de baseia no modelo de validação cruzada, descrito na seção 2.3.3.5.2. Esse modelo divide o *corpus* de atrações classificadas em duas partes: uma utilizada para treino (aprendizado) e outra utilizada para validação. Os algoritmos supervisionados reconhecem padrões nos dados durante o aprendizado, e em seguida sua precisão é medida durante a etapa de validação. Essa metodologia permite aferir a qualidade do modelo de classificação e ajustar parâmetros nos algoritmos para que uma precisão melhor seja avaliada. O método utiliza as representações TF-IDF/Word2vec criadas anteriormente. A Figura 39 exibe esse processo.

Figura 39 – Arquitetura do processo de classificação supervisionada



Fonte: elaborado pelo autor (2021).

Independente do algoritmo utilizado no modelo de classificação, a arquitetura é a mesma. O conjunto de atrações classificadas pelos especialistas representa 36 mil avaliações (12 perfis x 3000 avaliações em cada perfil). Os dados são separados usando um percentual de 75% para treino e 25% para avaliação. Não existe ordem nessa seleção, ou seja, as amostras para teste são selecionadas randomicamente, mas obedecendo uma quantidade mínima por perfil. O percentual de 25% destinado a teste e validação também faz parte do *corpus* classificado, porém, durante a etapa de validação, a classe (perfil) desses dados é removida. Assim, o modelo de classificação precisa prever o perfil correto de cada avaliação. Essa predição do modelo é comparada com o perfil verdadeiro daquela avaliação e dessa forma é possível avaliar o desempenho do modelo. As métricas de recuperação da informação *F-measure*, *Precision* e *Recall* são usadas para avaliar o desempenho dos modelos. Conforme aponta BERRAR (2019), é muito comum se construir um modelo de AM capaz de se adaptar muito bem aos dados classificados, porém ter resultados não satisfatórios em dados ainda não vistos pelo modelo (não classificados).

Após a validação, uma etapa de aprimoramento dos modelos é realizada a fim de melhorar seu desempenho. Em seguida, os modelos são enfim aplicados no *corpus* de cada atração (dados ainda não conhecidos pelos modelos). Para cada par (perfil x atração), um valor é retornado pelos algoritmos que reflete o grau de similaridade entre o conjunto de avaliações do perfil x o conjunto de avaliações da atração. Considerando trabalhos correlatos, nessa pesquisa, utilizou-se os algoritmos SVM, Logistic Regression, Random Forest e Naive Bayes. Esses algoritmos foram utilizados a partir de sua implementação Sklearn. As próximas seções descrevem o processo elaborado usando a abordagem supervisionada.

3.4.9 Modelo de classificação usando Support Vector Machine (SVM)

O modelo SVM foi aplicado usando a implementação LinearSVC do pacote Sklearn. O treino do modelo foi realizado durante 3.56 segundos e o modelo já estava criado. A variável *X_train* representa o conjunto de avaliações classificadas (27000). Já a variável *y_train* representa os perfis de cada uma dessas avaliações (27000). Os dados foram

representados usando TF-IDF. A Figura 40 exibe o código usado para esse processamento.

Figura 40 – Processamento modelo *Support Vector Machine (SVM)*

```

1 %%time
2
3 model = LinearSVC()
4 X_train, X_test, y_train, y_test, indices_train, indices_test = train_test_split(features, labels,
5                                     df.index, test_size=0.25, random_state=0)
6 model.fit(X_train, y_train)
7 y_pred = model.predict(X_test)

```

CPU times: user 2.35 s, sys: 1.2 s, total: 3.56 s
Wall time: 3.39 s

Fonte: elaborado pelo autor (2021).

A variável *y_pred* representa a tentativa do modelo de classificar o conjunto de avaliações para teste (9000). Nesse caso, o valor da classe (perfil) dessas 9000 avaliações não é passado ao algoritmo. A partir desse momento, é possível visualizar o desempenho do modelo perante a classificação das 9 mil avaliações. A Figura 41 exibe esse resultado.

Figura 41 – Desempenho do modelo *Support Vector Machine (SVM)*

```

1 print(classification_report(y_test, y_pred, target_names=
2                               category_id_df['Profile Name'].values))

```

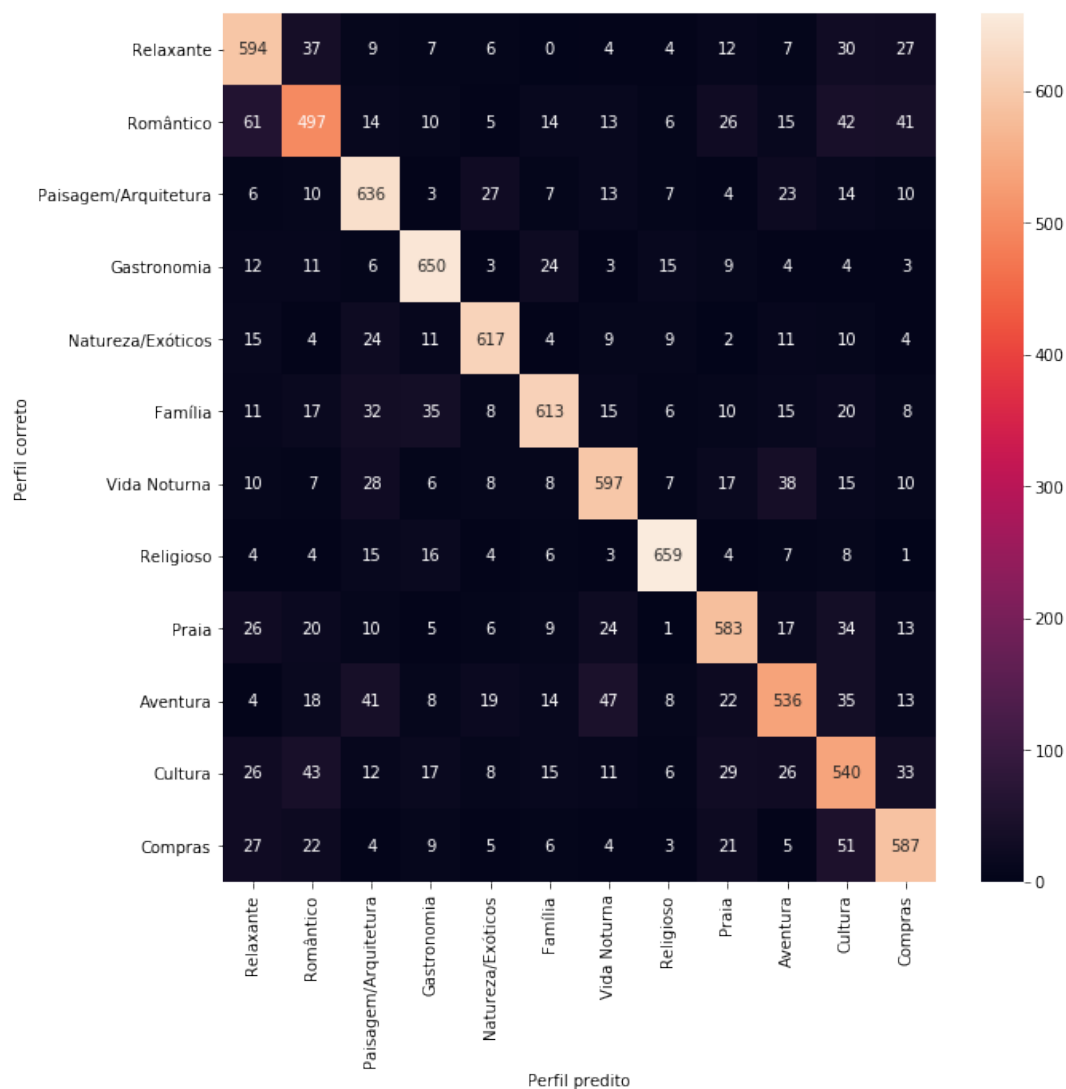
	precision	recall	f1-score	support
Relaxante	0.75	0.81	0.77	737
Romântico	0.72	0.67	0.69	744
Paisagem/Arquitetura	0.77	0.84	0.80	760
Gastronomia	0.84	0.87	0.85	744
Natureza/Exóticos	0.86	0.86	0.86	720
Família	0.85	0.78	0.81	790
Vida Noturna	0.80	0.79	0.80	751
Religioso	0.90	0.90	0.90	731
Praia	0.79	0.78	0.78	748
Aventura	0.76	0.70	0.73	765
Cultura	0.67	0.70	0.69	766
Compras	0.78	0.79	0.79	744
accuracy			0.79	9000
macro avg	0.79	0.79	0.79	9000
weighted avg	0.79	0.79	0.79	9000

Fonte: elaborado pelo autor (2021).

A Figura 41 exibe os resultados das métricas de classificação e recuperação da informação *Precision*, *Recall* e *F1-Score* (measure). A variável *support* representa a quantidade de testes realizados para cada perfil. Os resultados são exibidos numa escala entre 0 e 1, sendo 0 o menor grau de acerto e 1 o maior. O resultado final do algoritmo é dado pelo valor F1-Score. O modelo SVM apresentou F1-Score de 0.79, ou seja, em

9000 avaliações classificou corretamente o perfil de 7110 avaliações. Na Figura 41 é possível visualizar também que o perfil *Cultura* obteve o maior número de erros no processo classificatório, enquanto o perfil *Religioso* obteve o maior número de acertos. Como forma de visualizar graficamente os erros e acertos do algoritmo, a Figura 42 exibe a Matriz de Confusão.

Figura 42 – Matriz de Confusão do modelo *Support Vector Machine (SVM)*



Fonte: elaborado pelo autor (2021).

A Matriz de Confusão permite visualizar os perfis corretos (eixo y) e os perfis preditos pelo modelo de classificação (eixo x). Assim, é possível visualizar por exemplo que 27 avaliações do perfil *Relaxante* foram classificadas no perfil *Compras*. Por essa

visualização, o perfil *Romântico* foi “confundido” 61 vezes com o perfil *Relaxante*, ou seja, o modelo classificou 61 avaliações como *Relaxante*, que na verdade eram *Romântico*. Por um outro lado, o perfil *Relaxante* não foi confundido nenhuma vez com o perfil *Família*, enquanto o perfil *Praia* foi confundido uma única vez com o perfil *Religioso*.

3.4.10 Desempenho dos modelos de classificação supervisionados

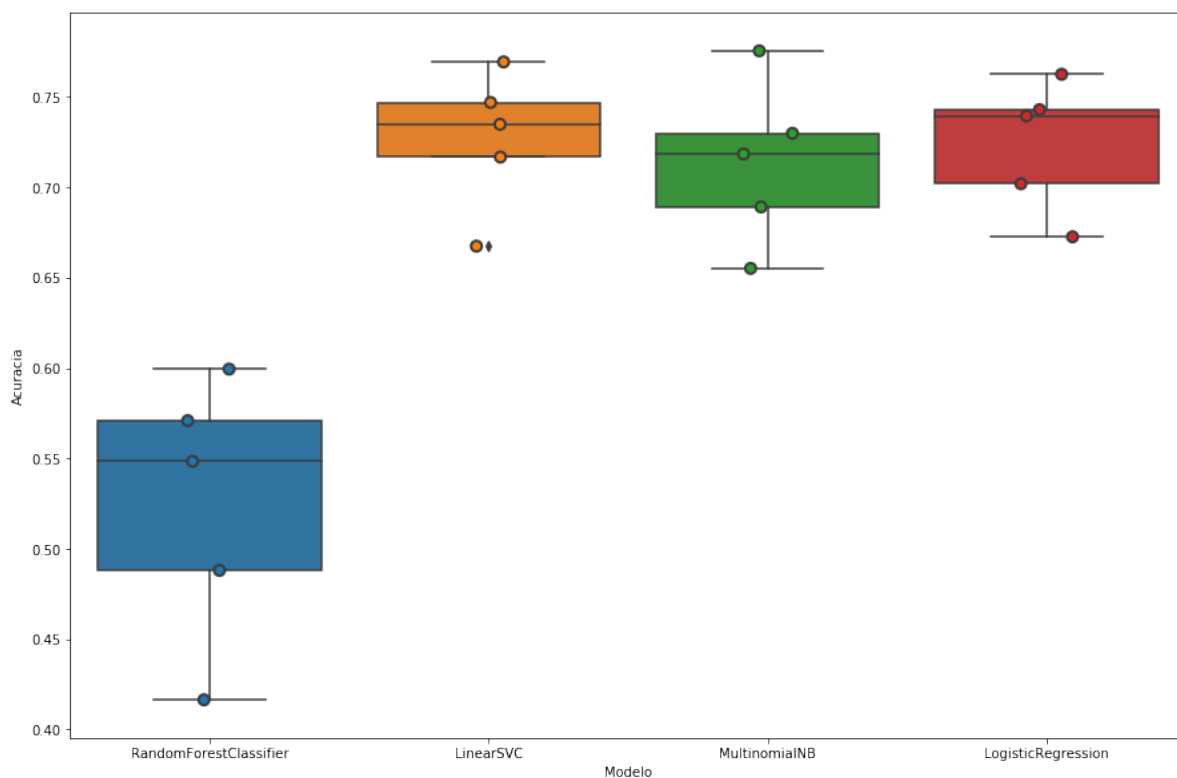
O processo descrito para o modelo SVM foi realizado também para os modelos Naive Bayes, Logistic Regression e Random Forest. Os modelos foram treinados usando a mesma lógica com 75% dos dados classificados e validação com 25% dos dados. Novamente a representação TF-IDF foi utilizada. Os resultados dos modelos também foram avaliados usando as métricas de desempenho *Precision*, *Recall* e *F1-Score*. O resultado final de acurácia é a média de F1-Score perante a classificação de todos os perfis. A Tabela 3 exibe o resultado de todos os modelos.

Tabela 3 – Desempenho dos modelos de classificação

MODELO	F1-SCORE
Logistic Regression	0.79
Naive Bayes	0.78
Random Forest	0.70
Support Vector Machine (SVM)	0.79

Fonte: elaborado pelo autor (2021).

O resultado apresentado no QUADRO é fruto de processos de otimização realizados na parametrização dos algoritmos. O método Random Forest por exemplo, inicialmente apresentou resultado de 0,61, após alterar o parâmetro *max_depth* (número de nós da árvore) para 10, o resultado subiu para 0,70. Nos testes realizados individualmente, os modelos SVM e Logistic Regression tiveram resultados iguais. Uma outra validação foi realizada usando o método *k-fold*. Esse método permite a divisão de todo o corpus classificado em subgrupos de teste do mesmo tamanho. Em seguida, realiza validações específicas em cada subgrupo de forma circular (BERRAR, 2019). O Gráfico 6 exibe o resultado do processamento usando cinco subgrupos de avaliação.

Gráfico 6 – Desempenho dos modelos de classificação teste com cinco *k-fold*

Fonte: elaborado pelo autor (2021).

O eixo y representa a acurácia de cada subgrupo testado. Para cada modelo é possível ver cinco pontos, cada um representando um F1-score médio do subgrupo. O eixo x representa o modelo utilizado em cada teste. Embora os resultados sejam inferiores ao primeiro teste de validação realizado, os mesmos são próximos. Avaliando-se a média de F1-Score dos subgrupos de teste, o modelo SVM apresenta relativa superioridade com acurácia média de 0,7270, seguido por Logistic Regression com 0,7238, Naive Bayes com 0,7135 e Random Forest com 0,5246. Uma vez que todos os modelos foram criados, parametrizados e validados, os mesmos estariam prontos a serem aplicados no corpus de dados não classificados, para assim, gerar a classificação de todas as atrações.

3.4.11 Modelo supervisionado com representação *word embeddings*

As validações apresentadas anteriormente foram processadas com a representação T-IDF. Para avaliar o desempenho do método de representação *word-embeddings* usando

o modelo supervisionado, foram realizados testes usando essa representação. A Tabela 4 exibe o resultado do processamento.

Tabela 4 – Desempenho dos modelos de classificação usando *word embeddings*

MODELO	F1-SCORE
Support Vector Machine (SVM)	0.73
Logistic Regression	0.72
Naive Bayes	0.72
Random Forest	0.72

Fonte: elaborado pelo autor (2021).

Os valores do processamento usando *word embeddings* foram inferiores à representação TF-IDF. Apenas o método Random Forest apresentou resultado superior. Nadbor (2016) realiza um estudo comparando as representações de *word embeddings* e TF-IDF em abordagem supervisionada. O resultado do autor é semelhante ao dessa pesquisa, ou seja, os métodos TF-IDF apresentaram resultados superiores. De acordo com o autor, uma das causas pode ser o fato do método *word embeddings* apresentar melhor performance quando há um número muito grande de dados não classificados. Esse resultado deve ser passível de uma investigação posterior e mais aprofundada com objetivo específico de comparar os métodos, haja vista, que tal comparação em nível mais analítico e aprofundado não está inserida no escopo desse estudo.

3.4.12 Aplicação dos modelos no *corpus* de atração não classificado

Considerando os testes realizados na abordagem supervisionada, optou-se por utilizar os modelos classificados por TF-IDF. Após a otimização dos modelos de classificação, os mesmos foram aplicados no *corpus* de todas as atrações (dados não classificados). Cada modelo de classificação gerou um valor de similaridade entre atração e perfil considerando o F1-score. Atrações com menos de 30 avaliações foram descartadas por não possuírem um conjunto de texto grande o suficiente para extrair suas características. O resultado do processamento foi salvo no banco de dados para cada par (atração, perfil). Os valores foram normalizados entre 0 e 1 com base em todo o resultado de cada método. Após esse processo classificatório foi possível extrair os perfis por atração e por

destino. A Tabela 5 exibe o resultado da classificação de perfis da atração Museu do Louvre em Paris usando o método SVM.

Tabela 5 – Classificação da atração Museu do Louvre com o método SVM

PERFIL	SIMILARIDADE	SIM_NORMALIZADA
Cultura	0.29756	0.40350
Paisagem/Arquitetura	0.09474	0.12650
Religioso	0.08191	0.10890
Natureza/Exóticos	0.07056	0.09340
Romântico	0.06261	0.08260
Relaxante	0.06106	0.08040
Compras	0.06093	0.08030
Família	0.05814	0.07650
Vida Noturna	0.05731	0.07530
Aventura	0.05598	0.07350
Gastronomia	0.05600	0.07350
Praia	0.04319	0.05600

Fonte: elaborado pelo autor (2021).

O perfil Cultura é o mais similar e relevante para a atração Museu do Louvre com uma similaridade de 0.40350 ou se considerar o valor normalizado 0,40350 em 1. Em segundo lugar aparece o perfil Paisagem/Arquitetura com valor normalizado de 0,12650. É relevante esses dois perfis aparecerem em primeiro usando o método SVM. O Museu do Louvre oferece peças, quadros, arte, esculturas, pinturas e diversos tipos de coleções para públicos variados do perfil cultura. Por um outro lado, chama atenção a imponente arquitetura do prédio que abriga o acervo, construída inicialmente para ser uma fortaleza e se tornando residência real no Século XVI, assim como a pirâmide de vidro no seu interior semelhante à de Quéops, em Gizé no Egito (MURTINHO, 2018). O fato de o perfil Praia aparecer em última posição mostra mais um acerto do algoritmo. Na verdade, o valor desse perfil deveria ser zero, porém, praticamente não existe similaridade zero. Isso porque quando um turista descreve uma atração, usa palavras específicas e comuns que também poderiam ser usadas para descrever uma experiência relacionada a praia.

O corpus composto por 2627 atrações com mais de 3,4 milhões de avaliações foi classificado automaticamente usando as duas abordagens: Similaridade de

representação dos textos e modelos supervisionados de AM. A próxima seção apresenta uma comparação dos modelos de classificação automáticos com a classificação manual dos especialistas.

3.5 *Nível de validação*

O processo metodológico dessa pesquisa envolve três níveis, o Nível Conceitual com a proposta conceitual de organização das informações; O Nível Tecnológico com construção e aplicação de métodos técnicos para se atingir o objetivo da pesquisa; E o Nível de Validação, responsável por apresentar a os resultados de comparação dos métodos automáticos de classificação com o processo de classificação manual realizado pelos especialistas. O processo de validação envolve novamente a participação de um produto intelectual produzido pelos especialistas. Essa classificação manual é fruto do conhecimento desses agentes de turismo durante anos de prática no ramo.

3.5.1 *Classificação manual de destinos nos perfis*

Para verificar o grau de confiabilidade dos resultados dos algoritmos seria necessário possuir atrações classificadas usando os mesmos perfis. No entanto, os modelos automáticos classificaram 2.627 atrações em 12 perfis, ou seja, se cada par (atração, perfil), um especialista demorasse 3 minutos, considerando um trabalho ininterrupto de 8 horas diárias, seriam necessários sete meses para concluir essa classificação. A inviabilidade de processar uma grande quantidade de informação em pouco tempo é um motivador pela busca de processos automatizados de organização da informação. Por tal dificuldade e fundamentando-se no trabalho de Mckenzie e Adams (2018), que classifica os destinos ao invés de atrações, a classificação manual foi realizada considerando o destino. Assim cada agente de turismo deveria sintetizar o conhecimento sobre determinado destino com base em suas atrações.

Os destinos foram classificados com base nas suas atrações. Considerando a lista de 124 destinos e os 12 perfis, para cada par (destinos, perfil) foi atribuído um valor entre 0 e 10 por cada especialista, sendo 0 quando o destino não possuir nenhuma relevância para aquele perfil e 10 quando tiver grau máximo de relevância. Quando um agente de

turismo vai indicar determinado destino, além de características do destino, o mesmo observa características de suas atrações. Assim, esses agentes possuem experiência para classificar quando um destino oferece alguma experiência para determinado tipo de perfil. Cada agente inseriu em uma planilha um valor para cada par (perfil, destino). A média dos dois agentes foi considerada como o resultado da classificação dos especialistas para cada destino perante os perfis. O trabalho intelectual dos especialistas foi salvo no banco de dados, e um processo de normalização dos valores foi realizado. A Tabela 6 exibe o resultado da classificação feita pelos especialistas para o destino *Paris*.

Tabela 6 – Classificação manual do destino Paris pelos especialistas

Destino	Label	Similaridade	norm
Paris	Cultura	10	1.00000
Paris	Paisagem/Arquitetura	10	1.00000
Paris	Romântico	10	1.00000
Paris	Família	8	0.80000
Paris	Gastronomia	7	0.70000
Paris	Compras	6	0.60000
Paris	Relaxante	6	0.60000
Paris	Religioso	5	0.50000
Paris	Vida Noturna	0	0.00000
Paris	Aventura	0	0.00000
Paris	Praia	0	0.00000
Paris	Natureza/Exóticos	0	0.00000

Fonte: dados da pesquisa (2021).

Os perfis *Cultura* e *Paisagem/Arquitetura* também aparecem nas primeiras posições, similar ao resultado da atração *Museu do Louvre*. No entanto, a classificação do destino não considera apenas uma atração. O agente de turismo ao classificar determinada atração com nota máxima em algum perfil, entende que aquele destino pode oferecer uma ou mais atrações que forneçam o maior grau possível de compatibilidade com esse perfil. O que se observa da classificação dos especialistas é que ao contrário da automática, ela não é granular, ou seja, um destino que não tem praia não aparece com nenhum valor relacionado a *Praia*, como exibido no exemplo de Paris. A classificação

manual dos destinos foi realizada entre o tempo estipulado de dois meses, por não estarem totalmente dedicados a essa tarefa. Esse tempo foi acordado por se encaixar no cronograma da pesquisa.

Na próxima seção, a classificação manual dos especialistas é comparada aos modelos automáticos de classificação.

3.5.2 Validação estatística entre classificação humana e automática

É importante frisar que o processo chamado de automático é semiautomático, pois foi construído a partir de amostras classificadas pelos especialistas. No entanto, esse processo apresenta uma heurística para construção de modelos automáticos. Esses modelos automáticos podem economizar tempo no trabalho de organização das informações, no entanto, precisam ser validados.

Os especialistas classificaram destinos, já os algoritmos classificaram atrações. Para que fosse possível uma avaliação dos dois resultados, ambos deveriam estar na mesma escala. Nesse sentido, foi criado um método para classificar um destino com base nas suas atrações classificadas automaticamente. Nessa criação observou-se dois fatores: quantidade e qualidade. Isso porque, uma média simples poderia comprometer os resultados, haja vista que um destino com cinco atrações nota 7 no perfil *Aventura* teria a mesma relevância de um destino com apenas uma única atração nota 7 em *Aventura*. Assim, para gerar o resultado de um destino perante a classificação de suas atrações nos perfis, foi utilizado um somatório das 5 atrações mais relevantes para cada perfil. A Tabela 7 exibe o resultado do método TF-IDF para o destino Paris.

Tabela 7 – Classificação do destino Paris usando o método TF-IDF

Destino	Label	Similaridade	norm
Paris	Relaxante	1.7250	0.471896
Paris	Religioso	1.6319	0.444002
Paris	Compras	1.5792	0.428212
Paris	Paisagem/Arquitetura	1.5514	0.419883
Paris	Romântico	1.5431	0.417396
Paris	Cultura	1.5139	0.408647
Paris	Aventura	1.4515	0.389951
Paris	Vida Noturna	1.3499	0.359510
Paris	Natureza/Exóticos	1.3458	0.358281
Paris	Gastronomia	1.3181	0.349982
Paris	Praia	1.2708	0.335810
Paris	Família	1.2250	0.322088

Fonte: dados da pesquisa (2021).

Já no resultado do método TF-IDF é possível notar uma diferença grande perante o resultado da classificação manual dos especialistas. A coluna *norm* representa o valor normalizado, que seria utilizado na comparação. A normalização foi realizada usando o método *MinMaxScaler* do pacote *Sklearn*.

Todos os modelos de classificação automáticos usados no trabalho TF, TF-IDF, Word2vec, SVM, Logistic Regression, Random Forest e Naive Bayes foram comparados com a classificação manual. Os conjuntos formados por (modelo – destino – perfil) foram comparados entre si. Exemplo: SVM-Paris-Cultura foi comparado com MANUAL-Paris-Cultura. Para verificar a similaridade entre os conjuntos, duas medidas de similaridade foram utilizadas, a distância Euclidiana (L2) e a distância Manhattan (L1). A distância Euclidiana exibe quão similares são dois conjuntos normalizados (HOTHO; NÜRNBERGER; PAAß, 2015). Quanto maior a distância Euclidiana ou Manhattan, maior a diferença entre os dois conjuntos (KHOSLA; RAJPAL; SINGH, 2014). Para se obter o valor total entre determinado método e o valor manual, a classificação de todos os 124 destinos foi comparada entre os dois métodos. A média aritmética das distâncias

Euclidianas e Manhattan aferidas em cada destino foi utilizada como o resultado geral de divergência entre a classificação manual e a classificação automática.

Tabela 8 – Validação dos métodos automáticos – distâncias estatísticas

MÉTODO	DISTÂNCIA EUCLIDIANA MÉDIA	DISTÂNCIA MANHATTAN
Método de representação word embeddings	1.1730	3.5761
Método de classificação supervisionada SVM	1.1802	3.6304
Método de representação TF-IDF	1.1990	3.6808
Método de representação TF	1.2523	3.6809
Método de classificação supervisionada Logistic Regression	1.2795	3.6383
Método de classificação supervisionada Naive Bayes	1.2852	3.6400
Método de classificação supervisionada Random Forest	1.2924	3.7179

Fonte: elaborado pelo autor (2021).

Considerando tanto a distância Euclidiana quanto a distância Manhattan, o método que se aproximou mais dos especialistas foi o de *word embeddings*. O método SVM ficou em segundo lugar considerando tanto a distância Euclidiana quanto a de Manhattan. O método Random Forest que obteve o pior resultado no teste cruzado da abordagem supervisionada foi também o que exibiu classificação mais diferente dos especialistas, tanto na distância Euclidiana quanto Manhattan. Os métodos TF-IDF e TF tiveram um desempenho bom (terceiro e quarto) considerando a distância Euclidiana, porém, considerando a distância Manhattan, ficaram apenas na frente do método Random Forest.

3.5.3 Validação qualitativa entre classificação humana e automática

Considera-se prudente ressaltar que a avaliação de qualidade parte do pressuposto de que os especialistas realizaram a melhor classificação, dessa forma, o modelo que melhor se aproximasse destes, seria mais apropriado para classificação. Os resultados da validação estatística fornecem uma base importante para a pesquisa, pois a partir desta, é possível visualizar os modelos automáticos mais próximos dos especialistas em termos matemáticos. Entretanto, uma outra avaliação interessante seria considerar a relevância de determinado perfil para um destino específico. Assim, a avaliação

estatística observa para cada destino, a distância entre os valores de cada perfil, desconsiderando sua posição no ranking.

Como forma de complementar a análise estatística, foi realizada uma nova análise qualitativa, dessa vez considerando a posição de relevância dos perfis para cada destino. Novamente, a classificação dos 124 destinos nos 12 perfis foi avaliada considerando uma comparação entre especialistas e métodos automáticos. Nesse novo teste, a ordem dos perfis para um destino foi considerada. Assim, se para o destino Paris, os especialistas classificaram que o perfil mais relevante é o Cultura (maior nota), logo, os métodos que também classificassem Cultura como o perfil mais relevante para Paris teriam resultados melhores. A comparação foi realizada considerando o método de Kendall's Tau por meio de sua implementação no pacote Spicy. De acordo com Shieh (1998), por meio desse método é possível avaliar a simetria entre dois rankings considerando a posição dos elementos. Quanto maior o resultado de Kendall's Tau, maior a similaridade entre os dois rankings. A Tabela 9 exibe o resultado da nova comparação.

Tabela 9 – Validação dos métodos automáticos – relevância do perfil para o Destino

MÉTODO	KENDALL'S TAU
Método de classificação supervisionada SVM	0.1102
Método de classificação supervisionada Random Forest	0.0831
Método de representação TF-IDF	0.0801
Método de classificação supervisionada Logistic Regression	0.0508
Método de representação Word2vec	0.0412
Método de classificação supervisionada Naive Bayes	0.0401
Método de representação TF	-0.0322

Fonte: elaborado pelo autor (2021).

O resultado considerando a posição dos perfis foi relativamente diferente do resultado puramente estatístico. O método com mais proximidade dos especialistas foi o SVM, com 0,1102, que no caso da análise estatística ficou em segundo lugar considerando a distância de Manhattan e Euclidiana. Em compensação, o método *word2vec* caiu, ficando em antepenúltimo lugar.

Considerando que a ordem dos perfis, ou seja, sua relevância em relação à uma atração ou destino possui maior peso do que a medida estatística entre as duas distribuições, escolheu-se para análise dos resultados dessa pesquisa, o método de classificação supervisionada SVM. Esse método foi o segundo mais bem posicionado na análise estatística e o primeiro na análise considerando a relevância dos perfis com base no destino. O método SVM também obteve melhor resultado na etapa de validação cruzada, que realiza um teste relevante de qualidade do modelo, apresentando acurácia de 79%. Ratifica-se que a abordagem utilizada considera que o melhor método de classificação é o aquele produzido pelos especialistas, ou seja, procura-se utilizar o método automático que mais se aproximou dessa classificação manual.

Na próxima seção da pesquisa apresenta-se os resultados obtidos da exploração de informações que foram geradas a partir da classificação das atrações e destinos em perfis turísticos.

4 RESULTADOS E DISCUSSÕES

Por sua natureza híbrida exploratória, os resultados são apresentados sob uma ótica qualitativa, explorando as informações oriundas do processo de classificação com base nos objetivos da pesquisa. A partir da metodologia desenvolvida e dos modelos utilizados, se torna possível explorar informações como: *Quais os perfis da atração Museu do Louvre? Quais os perfis do destino Belo Horizonte? Quais os 10 melhores destinos para Aventura? Quais as 10 atrações mais românticas?* Os rankings e informações geradas permitem um entendimento melhor sobre atrações, destinos e turistas com base no processo classificatório realizado. O uso de avaliações feitas por usuários, permite uma ampliação dos resultados, pois, os mesmos são construídos e formam uma síntese de milhões de opiniões.

Os resultados são apresentados em cinco categorias: a) resultados analíticos por perfil; b) resultados sintéticos por perfil; c) popularidade e relevância por perfil; d) similaridade entre destinos; e e) perfil turístico dos destinos mais visitados. A primeira categoria procura explorar analiticamente o resultado de dois perfis, exibindo *rankings* de destinos e atrações turísticas mais relevantes (maior similaridade) por perfil. Como os resultados se tornariam muito extensos, a segunda categoria apresenta uma análise sintética sobre os demais perfis. A terceira categoria busca apresentar um comparativo entre a popularidade e similaridade de destinos para alguns perfis. A quarta categoria procura identificar similaridade entre destinos e atrações turísticas. A quinta categoria procura explorar o perfil dos destinos mais visitados pelos brasileiros. Todas as informações exploradas nessa seção são obtidas com base no método de classificação SVM, que obteve resultado mais próximo à classificação dos especialistas.

4.1 Resultados analíticos por perfil

Os rankings permitem uma análise qualitativa sob determinada atração ou destino turístico considerando a opinião dos próprios turistas. Como há um recorte considerando apenas avaliações positivas sobre determinada atração, os resultados exibem uma classificação geral de qualidade.

Os *rankings* apresentados apresentam os 10 primeiros destinos ou atrações turísticas mais relevantes para cada perfil. Na esfera de atração, computou-se o resultado com base no valor de similaridade (valor SCM) entre a atração e determinado perfil. Na esfera destino, usou-se o somatório de similaridade das cinco primeiras atrações mais relevantes para cada perfil. Optou-se por exibir apenas os dez primeiros, porque no caso de atrações, para cada perfil, 2.627 foram classificadas; já no caso de destinos, 124 destinos foram classificados. Como a análise se tornaria muito extensa, nessa seção, procurou-se analisar os resultados de dois perfis: Vida Noturna e Religioso com mais profundidade. Uma visão sintética sobre os demais perfis é apresentada na seção 4.2. A escolha desses perfis deu-se pelo fato de possuírem mais atrações visitadas, conforme a informação de data de visita extraída do TripAdvisor.

4.1.1 Resultado do perfil Vida Noturna

Os resultados são divididos em cinco etapas de análise: a) as atrações mais similares ao perfil Vida Noturna; b) os destinos mais similares ao perfil; c) os estados mais similares ao perfil; d) análise linguística sobre o perfil; e por fim, e) uma conclusão sobre os resultados do perfil.

4.1.1.1 Atrações mais similares ao perfil Vida Noturna

Como especificado na Tabela 10, o perfil Vida Noturna destina-se a turistas com interesse em bares, cassinos, boates, e em geral, atividades noturnas. A Tabela 10 exibe as dez atrações mais relevantes para o perfil Vida Noturna em ordem decrescente de similaridade, ou seja, quanto maior o valor de Relevância, maior a similaridade entre a atração e o perfil.

Tabela 10 – Atrações mais similares ao perfil Vida Noturna

ATRAÇÃO	DESTINO	RELEVÂNCIA	CATEGORIAS TRIPADVISOR
Sofitel Montevideo Casino	Montevideú	60.33	Cassinos
Conrad Casino	Punta del Este	59.96	Cassinos
Istiklal Street	Istambul	52.12	Pontos de interesse
Casino at Bellagio	Las Vegas	51.45	Cassinos
Newbury Street	Boston	51.38	Pontos de interesse
Niederdorf	Zurique	50.92	Bairros, Pontos de interesse
Hard Rock Casino	Punta Cana	50.50	Cassinos
Rua Padre Chagas	Porto Alegre	50.02	Pontos de interesse
Rue St-Paul	Montreal	49.95	Pontos de interesse
Strøget	Copenhague	48.82	Bairros, Pontos de interesse

Fonte: elaborado pelo autor (2021).

Na Tabela 10, a coluna Relevância apresenta a média do resultado SCM de comparação do *corpus* do perfil Vida Noturna com as avaliações de cada atração. No ranking das dez atrações mais relevantes para Vida Noturna, apenas a atração *Casino at Bellagio* consta na lista de atrações mais características desse perfil, de acordo com os especialistas (QUADRO 4). As atrações Rua do Mucugê em Porto Seguro e Bairro Alto em Lisboa também indicadas pelos especialistas como as mais características do perfil Vida Noturna aparecem na posição 13 e 12 respectivamente.

Analisando as páginas das dez primeiras atrações no TripAdvisor, em geral, é possível perceber que se referem a casinos ou ruas movimentadas com boa oferta de bares, lojas, pubs e restaurantes específicos para o público adulto. No TripAdvisor, essas atrações são classificadas entre Cassinos e Pontos de interesse. Esse resultado permite visualizar a distinção dessa pesquisa em comparação ao modelo de classificação existente no TripAdvisor. A categoria Cassinos remete a vida noturna, no entanto, a categoria Pontos de interesse é genérica. Apenas uma atração nacional aparece na lista das 10 primeiras, a Rua Padre Chagas, uma atração boêmia na cidade de Porto Alegre. Analisando o perfil das 10 primeiras atrações, realmente são voltadas para o público objeto do perfil. Não houve erro de classificação do modelo classificatório com relação a alguma atração completamente dissonante do perfil em estudo, como uma igreja por exemplo.

4.1.1.2 Destinos mais similares ao perfil Vida Noturna

Para se considerar o resultado geral de relevância do destino para o perfil, foi considerado o somatório das cinco atrações de cada destino mais relevantes para Vida Noturna. Em seguida, ordenou-se os resultados por ordem decrescente de relevância e extraiu-se os 10 primeiros destinos. A Tabela 11 exibe os dez destinos mais relevantes para o perfil Vida Noturna.

Tabela 11 – Destinos mais similares ao perfil Vida Noturna

DESTINO	RELEVÂNCIA	MELHOR ATRAÇÃO
Las Vegas	2.1293	Casino at Bellagio
Lisboa	2.0021	Bairro Alto
Arraial d'Ajuda	1.8246	Rua do Mucugê
Montreal	1.7771	Rue St-Paul
Cartagena	1.7764	Bairro Getsemani
Montevideú	1.7615	Sofitel Montevideo Casino
Punta del Este	1.7447	Conrad Casino
Madrid	1.6937	La Latina
Barcelona	1.6873	El Born
Cidade do México	1.6292	La Condesa

Fonte: elaborado pelo autor (2021).

O fato de não se utilizar a média de todas as atrações é pela necessidade de uma análise que considere quantidade e qualidade das atrações com base em um determinado perfil. A média poderia favorecer destinos com poucas atrações, como por exemplo: um destino com cinco atrações nota 10 no perfil Vida Noturna, obteria a mesma nota de um destino com apenas uma atração nota 10 no perfil Vida Noturna. Como um somatório total de todas as atrações, privilegiaria destinos com mais atrações turísticas do que outros, adotou-se a soma das cinco principais atrações porque é um número mínimo de atração que todos os destinos possuem.

Os três primeiros destinos são exatamente aqueles escolhidos pelos especialistas como os mais característicos do perfil Vida Noturna. As melhores atrações desses destinos também coincidem com a escolha dos agentes. Entre as cinco atrações mais relevantes de Las Vegas, duas são cassinos (Bellagio e Wynn) e duas são ruas com cassinos (The

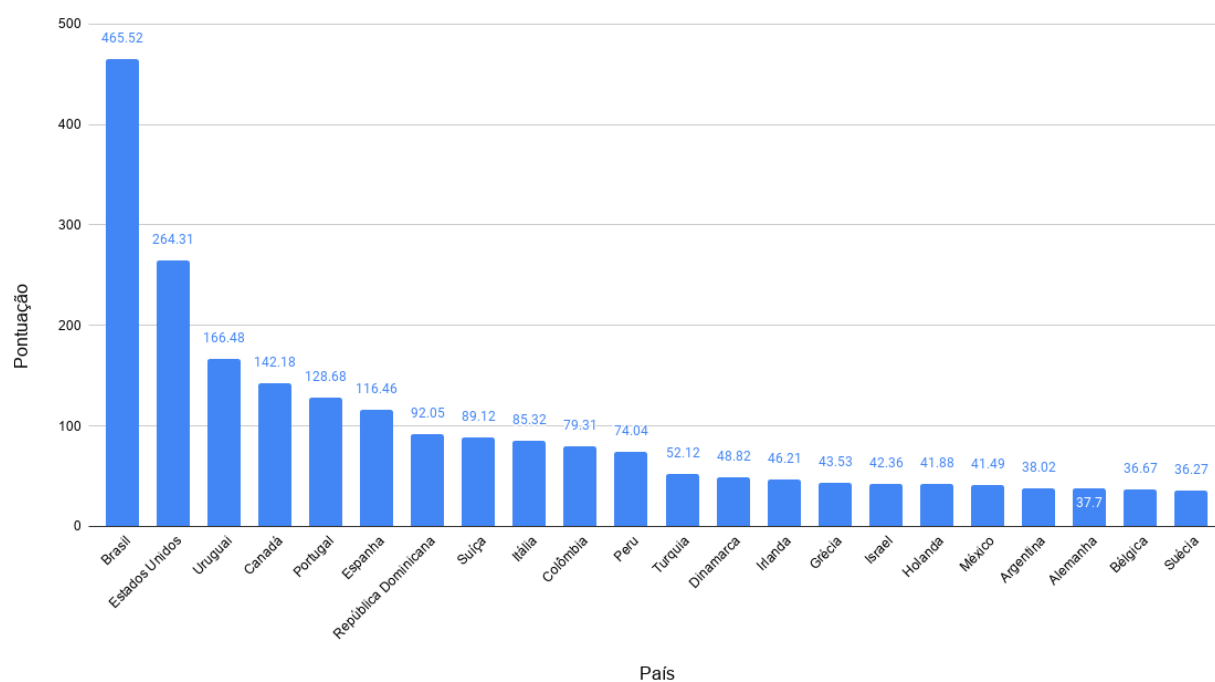
Strip e Fremont Street). Novamente apenas um destino nacional apareceu na lista dos 10 primeiros, no caso, a cidade de Arraial d'Ajuda.

4.1.1.3 Países mais similares ao perfil Vida Noturna

Considerando que apenas uma atração e um destino nacional apareceram nos primeiros colocados de similaridade do perfil, procurou-se analisar a relevância de países em relação ao perfil Vida Noturna. O Gráfico 7 exibe o resultado.

Gráfico 7 – Países com atrações mais similares à Vida Noturna

Países com atrações mais similares à Vida Noturna



Fonte: dados da pesquisa (2021).

Embora o Brasil não esteja muito frequente entre as atrações e destinos mais similares à Vida Noturna, se observarmos as 50 atrações mais bem avaliadas nesse perfil, o país é o que apresenta maior similaridade com o perfil Vida Noturna. É importante ressaltar que o Brasil possui o maior número de destinos estudados, ou seja, é natural que o país tenha mais chance de obter mais relevância em qualquer perfil, pelo maior número de atrações nos dados coletados. No Brasil, as três cidades mais similares são Arraial

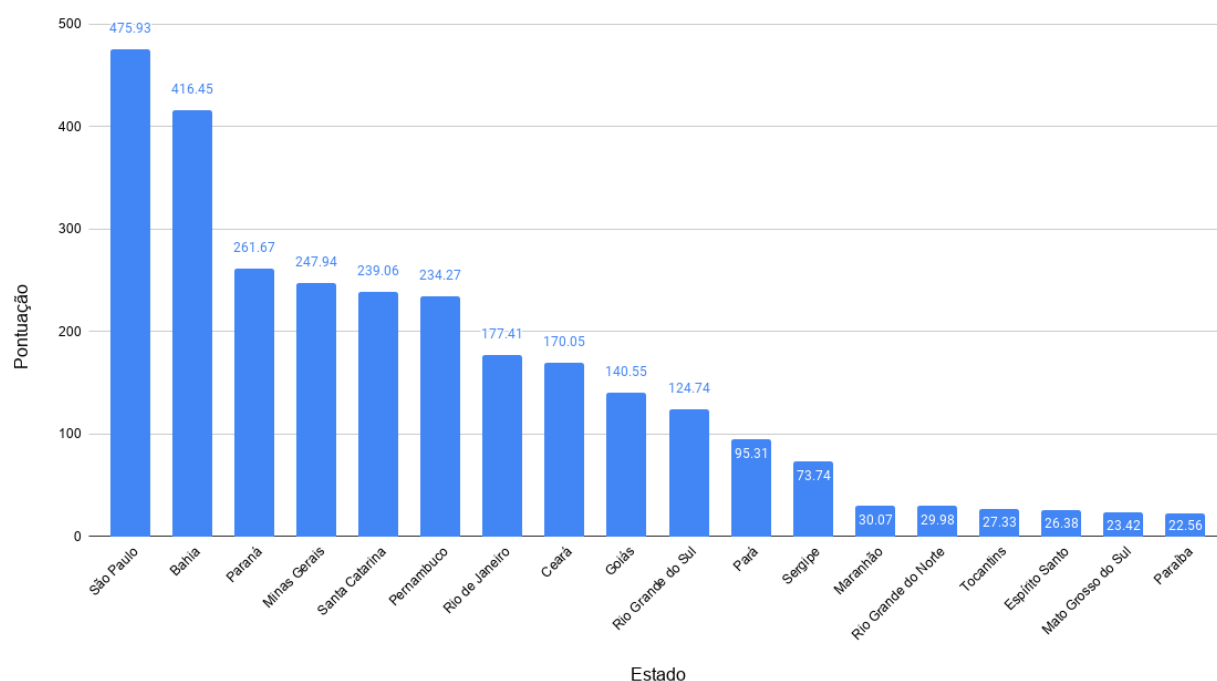
d'Ajuda, Tiradentes e Porto Alegre, respectivamente. Os Estados Unidos ficam em segundo lugar com as cidades Las Vegas e Boston entre as duas primeiras.

4.1.1.4 Estados nacionais mais similares ao perfil Vida Noturna

Procurando avaliar o cenário interno no Brasil e uma diferença entre o resultado geral e nacional, realizou-se uma consulta para identificar os estados nacionais com atrações mais similares ao perfil Vida Noturna. A análise foi feita com base nas 50 atrações mais relevantes para esse perfil. O Gráfico 8 exibe o resultado.

Gráfico 8 – Estados com atrações mais similares à Vida Noturna

Estados com atrações mais similares à Vida Noturna



Fonte: dados da pesquisa (2021).

Embora em termos qualitativos não esteja nas primeiras posições tanto de destino, quanto de atração, o Brasil apresenta boa oferta de atrações se observamos as 50 atrações mais relevantes para o perfil. As regiões sul e sudeste concentram a maioria das atrações mais relevantes para o perfil, sendo que entre as primeiras, aparece apenas o estado da Bahia da região do nordeste. O estado de São Paulo possui mais atrações compatíveis com o perfil Vida Noturna. Entre as principais, destacam-se a atração Rua

do Meio na cidade de Ilhabela com 40,5 pontos, Vila Capivari na cidade de Campos do Jordão com 34,45% de relevância, e a atração Bairro Jardins na cidade de São Paulo com 33,98%. Em segundo lugar, aparece o estado da Bahia, conhecido por ser um destino com boa oferta de atrações para agitação e vida noturna. As atrações Rua do Mucugê com 46,92% e Beco das Cores com 44,78% de relevância, se destacam como as primeiras nesse estado. No estado de Minas Gerais, quarto colocado, as cidades históricas de Ouro Preto e Tiradentes dominam as cinco primeiras posições, superando a capital mineira Belo Horizonte.

4.1.1.5 Análise linguística do perfil Vida Noturna

Com o objetivo de identificar os termos mais frequentes em avaliações relevantes ao perfil, realizou-se uma análise linguística usando LDA. Criou-se um modelo com apenas um tópico sobre um *corpus* formado por todas as avaliações das 50 atrações mais relevantes do perfil Vida Noturna. A Figura 43 exibe uma nuvem com os termos mais importantes para o perfil Vida Noturna.

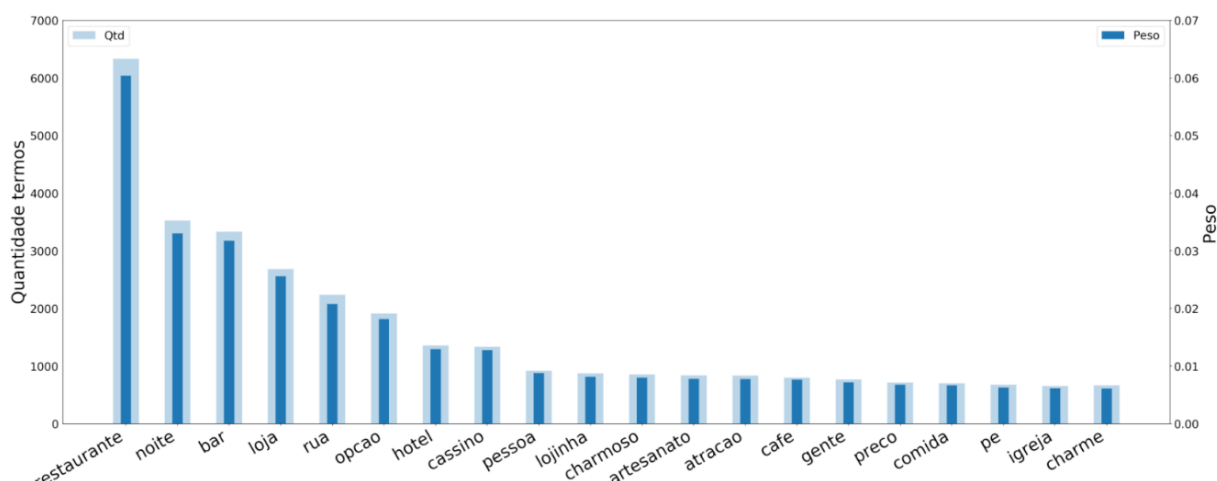
Figura 43 – Principais termos do perfil Vida Noturna



Fonte: elaborado pelo autor (2021).

Quanto maior a palavra é exibida na Figura 43, maior a sua relevância para o perfil. Entre as principais palavras destacam-se restaurante, noite, rua, cassino e bar. Essas palavras ocorrem com maior frequência nas avaliações dos usuários e são as que mais caracterizam o perfil Vida Noturna. No Gráfico 9 é possível visualizar um comparativo entre a frequência de cada termo e seu peso para um determinado tópico LDA.

Gráfico 9 – Frequência X Peso dos principais termos do perfil Vida Noturna



Fonte: dados da pesquisa (2021).

O Gráfico 9 exhibe os 20 termos mais frequentes no *corpus* considerando as 50 principais atrações do perfil Vida Noturna. O termo *restaurante* é o mais frequente nas avaliações ocorrendo mais de 6 mil vezes, recebendo conseqüentemente um maior peso em relação aos demais. O termo *restaurante* ocorre duas vezes mais do que o segundo e terceiro termo, *noite* e *bar* respectivamente. A ocorrência dos termos *restaurante* e *bar* sugere que o perfil Vida Noturna esteja muito ligado a esse tipo de atração. É possível visualizar que o adjetivo *charmoso* apareceu entre os principais, ou seja, esse termo não existia no dicionário de adjetivos criado para remoção de adjetivos. O termo apresentado, não acrescenta valor à classificação, ao contrário dos substantivos *restaurante* ou *cassino*, o que ratifica a importância da etapa de pré-processamento realizada com remoção de verbos ou adjetivos.

Os termos *restaurante* e *bar* alguns dos principais do perfil Vida Noturna, não se considerou esse tipo de atração no estudo. No TripAdvisor, bares e restaurantes ficam em outra categoria, diferente da categoria Atrações, que manifestou um ponto crítico de qual direção deveria ser dada. Assim, em uma avaliação sobre uma atração turística Rua Mucugê, o turista pode relatar aspectos sobre os restaurantes e bares da região, mas não necessariamente sobre um específico, o que não seria o foco do estudo. Ou seja, o objetivo é entender quão boa é a oferta de determinado tipo de atração num determinado destino para um perfil específico. Um estudo interessante a ser realizado posteriormente,

poderia ser a inclusão desse tipo de atração de forma a observar o comportamento dos perfis gerais dos destinos.

4.1.2 Resultado do perfil Religioso

Os resultados são divididos em cinco etapas de análise: a) as atrações mais similares ao perfil Religioso; b) os destinos mais similares ao perfil; c) os estados mais similares ao perfil; d) análise linguística sobre o perfil; e por fim, e) uma conclusão sobre os resultados do perfil.

4.1.2.1 Atrações mais similares ao perfil Religioso

Como especificado na Tabela 12, o perfil Religioso destina-se a turistas com interesse em igrejas, catedrais ou história religiosa em geral. A Tabela 12 exhibe as dez atrações mais relevantes para o perfil em ordem decrescente de similaridade, ou seja, quanto maior o valor de Relevância, maior a similaridade entre a atração e o perfil.

Tabela 12 – Atrações mais similares ao perfil Religioso

ATRAÇÃO	DESTINO	RELEVÂNCIA	CATEGORIAS TRIPADVISOR
Basílica Santo Antônio do Embaré	Santos	79.78	Igrejas e Catedrais
Church of All Nations (Basílica of the Agony)	Jerusalém	78.20	Pontos sagrados e religiosos
Paróquia Divino Espírito Santo	Holambra	77.85	Locais de interesse
Arcibasílica di San Giovanni in Laterano	Roma	77.80	Igrejas e Catedrais
Basílica Nossa Senhora de Lourdes	Belo Horizonte	77.64	Igrejas e Catedrais
Basílica of the National Shrine of Mary, Queen of the Universe	Orlando	76.55	Igrejas e Catedrais, Locais sagrados e religiosos
Santa Maria della Vittoria	Roma	76.32	Igrejas e Catedrais
Parroquia de Cristo Resucitado	Cancun	76.07	Igrejas e Catedrais, Locais sagrados e religiosos, Construções Arquiteturais
Chiesa di Sant'Ignazio di Loyola	Roma	75.67	Igrejas e Catedrais, Locais sagrados e religiosos.
Iglesia de San Francisco	Bogotá	74.62	Igrejas e Catedrais

Fonte: elaborado pelo autor (2021).

Na Tabela 12, a coluna Relevância apresenta a média do resultado SCM de comparação do *corpus* do perfil Religioso com as avaliações de cada atração. No ranking das dez atrações mais relevantes para o perfil, nenhuma das atrações escolhidas pelos especialistas como as mais relevantes apareceu. No entanto, entre as 10 primeiras atrações, três são na cidade de Roma na Itália. O Brasil possui três atrações entre as 10 mais relevantes. Todas as atrações estão categorizadas no TripAdvisor em categorias pertinentes ao perfil.

4.1.2.2 Destinos mais similares ao perfil Religioso

Para se considerar o resultado geral de relevância do destino para o perfil, foi considerado o somatório das 5 atrações de cada destino mais relevantes para o perfil Religioso. Em seguida, ordenou-se os resultados por ordem decrescente de relevância e extraiu-se os 10 primeiros destinos. A Tabela 13 exibe os dez destinos mais relevantes para o perfil.

Tabela 13 – Destinos mais similares ao perfil Religioso

DESTINO	RELEVÂNCIA	MELHOR ATRAÇÃO
Roma	3.7701	Arcibasilica di San Giovanni in Laterano
Veneza	2.8311	Basilica dei Santi Giovanni e Paolo (San Zanipolo)
Florença	2.7776	Church of Santa Maria Novella
Milão	2.7542	Chiesa di Santa Maria presso San Satiro
Cartagena	2.7398	Iglesia de Santo Toribio
Jerusalém	2.7183	Church of All Nations (Basilica of the Agony)
Montreal	2.6687	St. Patricks Basilica
João Pessoa	2.5579	Paróquia Santo Antônio de Lisboa
Brasília	2.5191	Santuário Dom Bosco
Praga	2.5105	Kostel Panny Marie Vítězné a Pražské Jezulátko

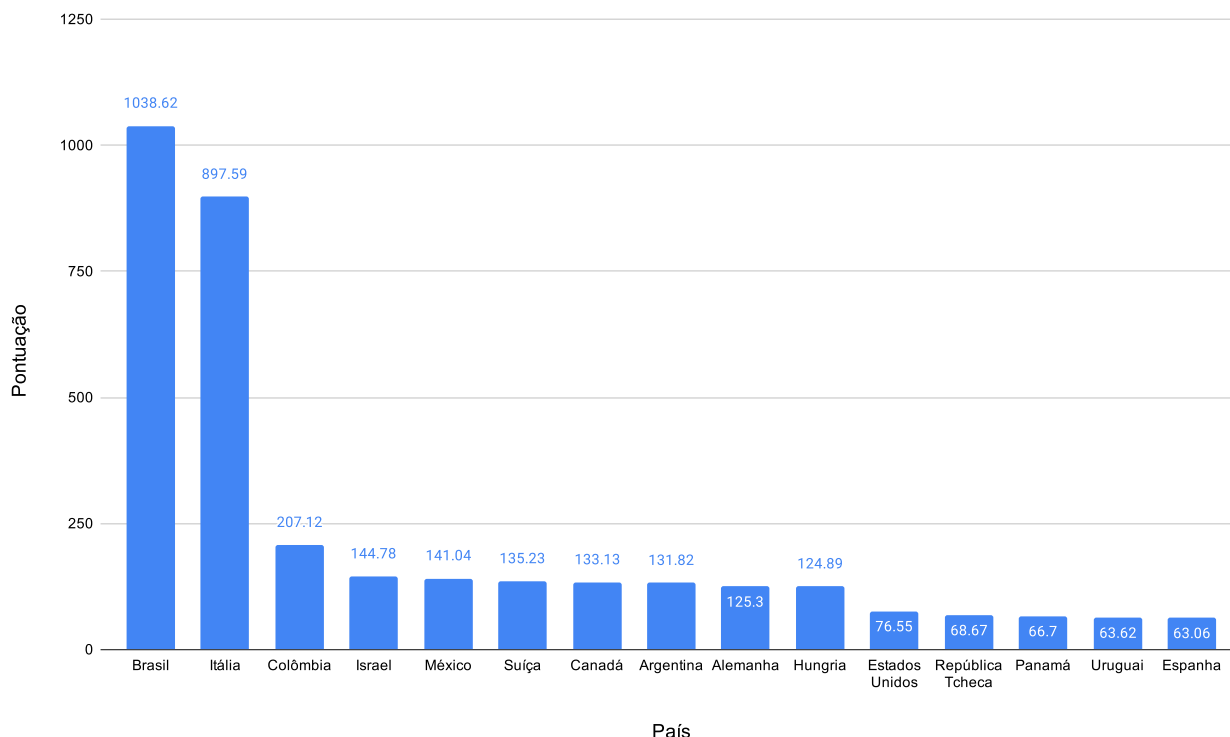
Fonte: elaborado pelo autor (2021).

A Itália é o país com mais destinos presentes entre os 10 primeiros. O Brasil aparece na lista com dois destinos (João Pessoa e Brasília). Das 3 cidades que possuem as atrações mais características do perfil, apenas Roma apareceu na listagem, porém a atração mais relevante para essa cidade foi diferente da atração escolhida pelos especialistas no Quadro 4.

4.1.2.3 Países mais similares ao perfil Religioso

Nessa seção procurou-se avaliar a relevância dos países perante o perfil Religioso considerando as 50 atrações mais relevantes. O Gráfico 10 exibe o resultado.

Gráfico 10 – Países com atrações mais similares ao perfil Religioso



Fonte: dados da pesquisa (2021).

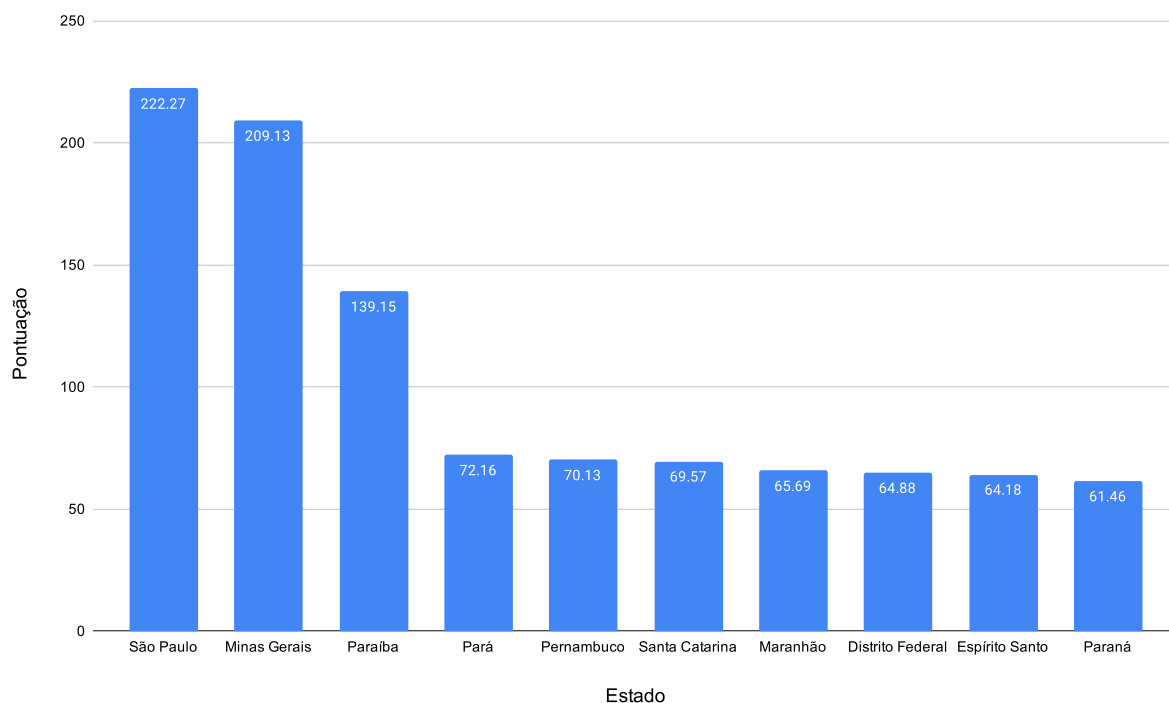
No Brasil, o país com mais relevância para o perfil, as três cidades mais similares são João Pessoa, com 2,55 pontos de relevância, seguida por Brasília e Salvador com 2,51 e 2,23 pontos respectivamente. Na lista das 50 atrações mais relevantes, o Brasil possui 15 atrações e Itália 13. A Itália aparece em segundo lugar com a maior pontuação, seguida por Colômbia e Israel.

4.1.2.4 Estados nacionais mais similares ao perfil Religioso

Procurando avaliar o cenário interno no Brasil e uma diferença entre o resultado geral e nacional, realizou-se uma consulta para identificar os estados nacionais com atrações

mais similares ao perfil Religioso. A análise foi feita com base nas 50 atrações mais relevantes para esse perfil. O Gráfico 11 exibe o resultado.

Gráfico 11 – Estados com atrações mais similares ao perfil Religioso



Fonte: dados da pesquisa (2021).

O estado de São Paulo possui maior relevância com o perfil Religioso seguido por Minas Gerais. Ambos estados possuem três atrações na lista das 50 mais relevantes. As duas atrações mais relevantes no Brasil estão no estado de São Paulo, a Basílica Santo Antônio do Embaré em Santos com 79,78% e a Paróquia Divino Espírito Santo em Holambra com 77,85% de relevância. A terceira atração mais relevante é a Basílica Nossa Senhora de Lourdes em Belo Horizonte com 77,74% de relevância. As três atrações também estão na lista das dez principais atrações.

4.1.2.5 Análise linguística do perfil Religioso

Com o objetivo de identificar os termos mais frequentes em avaliações relevantes ao perfil, realizou-se uma análise linguística usando LDA. Criou-se um modelo com apenas um tópico sobre um corpus formado por todas as avaliações das 50 atrações mais

relevantes do perfil Religioso. A Figura 44 exibe uma nuvem com os termos mais importantes para o perfil.

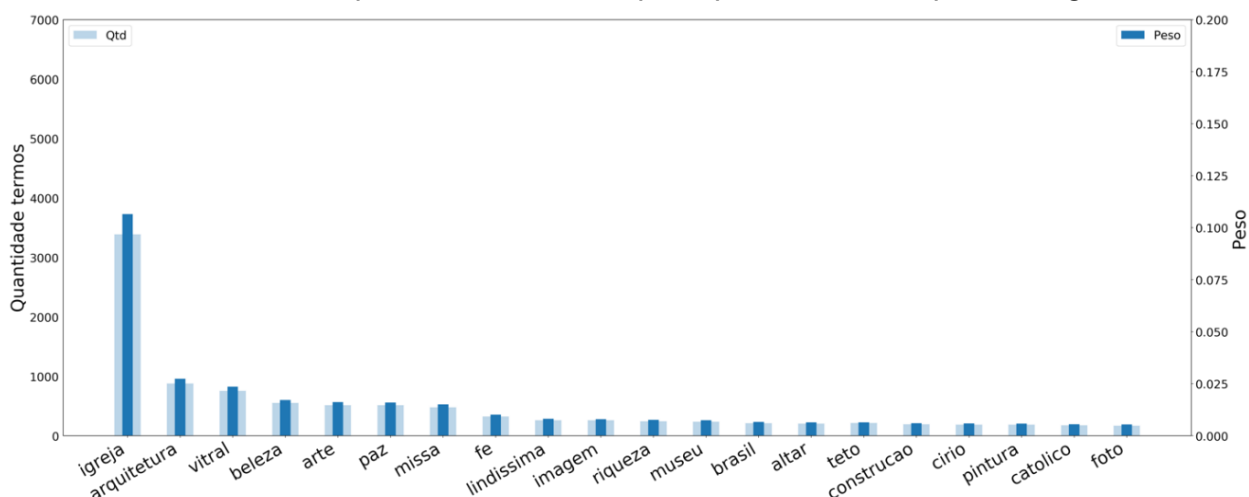
Figura 44 – Principais termos do perfil Religioso



Fonte: elaborado pelo autor (2021).

Quanto maior a palavra é exibida na Figura 44, maior a sua relevância para o perfil. As palavras que mais se destacam são: *igreja*, *missa* e *arquitetura*. Essas palavras ocorrem com maior frequência nas avaliações dos usuários e são que mais caracterizam o perfil Religioso. No Gráfico 12 é possível visualizar um comparativo entre a frequência de cada termo e seu peso para um determinado tópico LDA.

Gráfico 12 – Frequência X Peso dos principais termos do perfil Religioso



Fonte: dados da pesquisa (2021).

O Gráfico 12 exibe os 20 termos mais frequentes no corpus considerando as 50 principais atrações do perfil Religioso. O termo igreja é o mais frequente nas avaliações ocorrendo mais de 3 mil vezes, recebendo conseqüentemente um maior peso em relação aos demais. O termo igreja ocorre três vezes mais do que o segundo e terceiro termo, arquitetura e vitral respectivamente. É importante observar também que ao descrever uma experiência do perfil Religioso, os turistas relatam com frequência termos relacionados ao perfil Paisagem/Arquitetura, como *arquitetura* e *construção*. É interessante observar que embora o termo arquitetura seja o segundo mais relevante no corpus, nenhuma atração que não tenha relação com o perfil foi retornada, como por exemplo: alguma atração relacionada a Paisagem/Arquitetura e não a Religioso. Isso reforça a classificação usando um conjunto de perfis, porque uma mesma atração oferece múltiplas experiências.

4.2 Resultados sintéticos por perfil

Uma visão sintética sobre os demais perfis é apresentada na seção 4.1, esse tópico busca traçar um overview sobre cada um dos demais perfis. Para cada perfil, uma análise sobre localidade dos destinos e um panorama do Brasil considerando os 30 destinos e as 100 atrações turísticas com melhor resultado no perfil. A análise nos 30 destinos é medida em pontos, ou seja, o somatório de relevância das atrações de um destino. A análise das 100 atrações turísticas é medida em percentual de relevância entre a atração e o perfil.

4.2.1 Perfil Cultura

O perfil Cultura é dominado por países da Europa, apresentando um total de 29.09 pontos de relevância em comparação ao segundo América do Sul, com 12.40 pontos de relevância. Dos 30 destinos mais relevantes para o perfil, 16 se localizam na Europa, oito na América do Sul, três na América do Norte, dois na América Central e um na África. O Brasil possui seis destinos entre os 30 mais relevantes para Cultura: São Luís, São Paulo, Joinville, Curitiba, Recife e Tiradentes. A cidade de Roma na Itália é a mais relevante para o perfil com 2.35 pontos, seguida por Paris na França com 2.03 pontos.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a Europa possui 55 atrações, seguida da América do Sul com 23 atrações. Roma é a cidade com mais atrações, sete no total, seguida por Paris e Amsterdã com seis e cinco atrações respectivamente. No Brasil, São Paulo e Pernambuco são os estados que possuem mais atrações entre as 100 primeiras, com três cada um. As duas atrações mais bem avaliadas no perfil são: State Tretyakov Gallery em Moscou com 58,61% e Coleção Peggy Guggenheim em Veneza com 58,42% de relevância para o perfil. No Brasil, a atração mais relevante para o perfil é o Museu da Fotografia em Fortaleza com 46,84%, seguida pelo Museu de Arte Sacra Pierre Chalita em Maceió com 41,93%.

4.2.2 Perfil Paisagem/Arquitetura

A Europa apresenta maior relevância para o perfil Paisagem/Arquitetura, com 30,31 pontos, seguida por América do Sul com 11,42 pontos. Dos 30 destinos mais relevantes, 18 se localizam na Europa, sete na América do Sul, três na América Central e dois na África. O Brasil possui três destinos entre os 30 mais relevantes para o perfil: Brasília, Aracajú e Petrópolis. A cidade de Praga na República Tcheca é a mais relevante para o perfil com 2.34 pontos, seguida por Madri na Espanha com 2.19 pontos.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a Europa possui 54 atrações, seguida da América do Sul com 35 atrações. Praga é a cidade com mais atrações, sete no total, seguida por Machu Picchu e Madri com oito e sete atrações respectivamente. No Brasil, Pernambuco e Tocantins são os estados que possuem mais atrações entre as 100 primeiras, com dois cada um. As duas atrações mais bem avaliadas no perfil são: Charlottenburg Palace em Berlim com 56,49% e Wallenstein Palace Gardens em Praga com 56,23% de relevância para o perfil. No Brasil, a atração mais relevante para o perfil é o Palácio Araguaia, em Palmas, com 39,56%, seguida pelo Forte São Pedro do Boldró em Fernando de Noronha com 38,01%.

4.2.3 Perfil Família

A América do Sul apresenta maior relevância para o perfil Família com 23,60 pontos, seguida por América do Norte com 5,9 pontos. Dos 30 destinos mais relevantes, 23 se

localizam na América do Sul, cinco na América do Norte, um na Europa, e um na América Central. O Brasil possui 22 destinos entre os 30 mais relevantes para o perfil, ou seja, 73% dos destinos. A cidade de Orlando nos Estados Unidos é a mais relevante para o perfil com 2.16 pontos, seguida por Caldas Novas no Brasil com 1.78 pontos.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 68 atrações, seguida da América do Norte com 23 atrações. Orlando é a cidade com mais atrações, 15 no total, seguida por Caldas Novas e Ubatuba com sete atrações cada uma. No Brasil, São Paulo e Goiás são os estados que possuem mais atrações entre as 100 primeiras, com 17 e 13 respectivamente. As duas atrações mais relevantes para o perfil são: Eco Parque em Arraial do Ajuda com 50,56% e Universal's Islands of Adventure em Orlando com 48,14%.

4.2.4 Perfil Gastronomia

A América do Sul apresenta maior relevância para o perfil Gastronomia com 35,12 pontos, seguida por Europa com 7,58 pontos. Dos 30 destinos mais relevantes, 20 se localizam na América do Sul, seis na Europa, dois na América do Norte, um na África, e um na América Central. O Brasil possui 16 destinos entre os 30 mais relevantes para o perfil, ou seja, 53,33% dos destinos. A cidade de Bento Gonçalves é a mais relevante para o perfil com 4.77 pontos, seguida por Garibaldi no Brasil com 3,84 pontos.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 76 atrações, seguida da Europa com 17 atrações. Bento Gonçalves é a cidade com mais atrações, 26 no total, seguida por Garibaldi e Gramado com sete e seis atrações respectivamente. No Brasil, o estado do Rio Grande do Sul aparece disparado com 42 atrações entre as 100 primeiras, seguido por Rio de Janeiro e Santa Catarina com quatro e seis respectivamente. As duas atrações mais relevantes para o perfil são: Vinícola Don Laurindo em Bento Gonçalves com 100% e Lidio Carraro Vinícola Boutique com 96,81% de relevância, ambas na cidade de Bento Gonçalves. Avaliando os resultados há uma predominância de vinícolas e cervejarias, o que é normal, considerando as atrações mais características escolhidas para esse perfil.

4.2.5 Perfil Aventura

A América do Sul apresenta maior relevância para o perfil Aventura, com 19,41 pontos, seguida por América do Norte com 7,53 pontos. Dos 30 destinos mais relevantes, 16 se localizam na América do Sul, cinco América do Norte, quatro na Europa, três na América Central, um na Ásia, e um na África. O Brasil possui 15 destinos entre os 30 mais relevantes para o perfil, ou seja, 50% dos destinos. A cidade de Orlando nos Estados Unidos é a mais relevante para o perfil com 2,53 pontos, seguida por Natal no Brasil com 1,85 pontos.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 48 atrações, seguida da América do Norte com 25 atrações. Orlando é a cidade com mais atrações, 12 no total, seguida por Las Vegas e Natal com seis e cinco atrações respectivamente. No Brasil, o estado de Goiás aparece com sete atrações entre as 100 primeiras, seguido por Rio Grande do Norte e Rio Grande do Sul com cinco e quatro atrações respectivamente. As duas atrações mais relevantes para o perfil são: Kitakas em Caldas Novas com 70% e Alpen Park em Canela com 69,17%.

4.2.6 Perfil Praia

Entre os 30 destinos mais relevantes para Praia, constam apenas destinos na América do Sul (27) e América Central (3). A América do Sul apresenta relevância de 61,26 pontos, enquanto a América Central 6,08 pontos. O Brasil concentra 26 destinos dos 30 mais relevantes para o perfil, ou seja, 86% do total. A cidade de Maceió no Brasil é a mais relevante para o perfil com 2,87 pontos, seguida por Ubatuba, também no Brasil com 2,69 pontos.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 94 atrações, seguida da América Central com cinco atrações. O Brasil aparece disparado com o mais relevante para o perfil, com 92% das 100 primeiras atrações. Ubatuba é a cidade com mais atrações, 15 no total, seguida por Maceió e Florianópolis com dez atrações cada uma. No Brasil, o estado de Santa Catarina aparece com 23 atrações entre as 100 primeiras, seguido por São Paulo e Alagoas com 21 e 12 atrações

respectivamente. As duas atrações mais bem avaliadas para o perfil são: Praia de Guaxuma com 59,95% e Praia de Ponta Verde com 58,7%, ambas na cidade de Maceió.

4.2.7 Perfil Compras

A América do Sul apresenta maior relevância para o perfil Compras, com 22,95 pontos, seguida por América do Norte com 13,78 pontos. Dos 30 destinos mais relevantes, 14 se localizam na América do Sul, seis América do Norte, quatro na América Central, três na Ásia e três na Europa. O Brasil possui 14 destinos dos 30 mais relevantes para o perfil, seguido por Estados Unidos com cinco destinos. As cidades de Miami e Orlando nos Estados Unidos são as mais relevantes para o perfil com 3,15 e 3,00 pontos respectivamente. No Brasil, os destinos mais relevantes para Compras são: Belo Horizonte com 2,39 pontos e Goiânia com 2,22 pontos de relevância.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 42 atrações, seguida da América do Norte com 21 atrações. Miami é a cidade com mais atrações, cinco no total, seguida por Orlando e Cidade do Panamá com três atrações cada uma. No Brasil, o estado do Paraná aparece com cinco atrações entre as 100 primeiras, seguido por Santa Catarina e Goiás com quatro atrações cada um. As duas atrações mais bem avaliadas para o perfil são: The Florida Mall em Orlando com 91,75% e Citadel Outlets em Las Vegas com 86,36%. No Brasil, as duas atrações mais relevantes para o perfil são: Midway Mall e Rio Mar Recife Mall com 81,99 e 75,36 pontos respectivamente.

4.2.8 Perfil Relaxante

A América do Sul apresenta maior relevância para o perfil Relaxante, com 14,40 pontos, seguida por Europa com 11,25 pontos. Dos 30 destinos mais relevantes, 12 se localizam na América do Sul, nove na Europa, quatro na América do Norte, dois na Ásia, um na África e um na Oceania. O Brasil possui 10 destinos dos 30 mais relevantes para o perfil, seguido por Estados Unidos com quatro destinos. As cidades de Veneza na Itália e Arraial do Cabo no Brasil são as mais relevantes para o perfil com 1,81 e 1,76 pontos

respectivamente. Considerando apenas o Brasil, os destinos mais relevantes após Arraial do Cabo são: Jericoacoara com 1,65 pontos e Paraty com 1,24 pontos de relevância.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui metade das atrações, seguida da Europa com 28 atrações. Veneza é a cidade com mais atrações, nove no total, seguida por Jericoacoara e Arraial do Cabo com sete e cinco atrações respectivamente. No Brasil, o estado do Rio de Janeiro aparece com 19 atrações entre as 100 primeiras, seguido por Ceará e São Paulo com nove e seis atrações respectivamente. As duas atrações mais bem avaliadas para o perfil são: Rio Sena em Paris com 66,95% e Canal Grande em Veneza com 58,51%. No Brasil, as duas atrações mais relevantes para o perfil são: Pedra Perfil do Gorila e Fenda de Nossa Senhora de Assunção, ambas em Arraial do Cabo com 47,60 e 44,11 pontos respectivamente.

4.2.9 Perfil Romântico

A América do Sul apresenta maior relevância para o perfil Romântico, com 32,21 pontos, seguida por América Central com 7,8 pontos. Dos 30 destinos mais relevantes, 24 se localizam na América do Sul, cinco na América Central e um na América do Norte. O Brasil possui dez destinos dos 30 mais relevantes para o perfil, seguido por Estados Unidos com quatro destinos. As cidades de Punta Cana na República Dominicana e Cancun no México são as mais relevantes para o perfil com 2,15 e 1,76 pontos respectivamente. Considerando apenas o Brasil, os destinos mais relevantes são: Paraty com 1,66 pontos e Angra dos Reis com 1,61 pontos de relevância.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 68% das atrações, seguida pela América Central com 23 atrações. Maragogi é a cidade com mais atrações, nove no total, seguida por Paraty e Angra dos Reis com sete atrações cada uma. No Brasil, o estado do Rio de Janeiro aparece com 18 atrações entre as 100 primeiras, seguido por Alagoas e Bahia com dez e cinco atrações respectivamente. As duas atrações mais bem avaliadas para o perfil são: Lago Negro em Gramado com 65,71% e Saona Island em Punta Cana com 55,55%. No Brasil, a atração mais relevante para o perfil após o Lago Negro é a Ilha dos Namorados em Aracaju com 40,52 pontos.

4.2.10 Perfil Natureza/Exóticos

A América do Sul apresenta maior relevância para o perfil Natureza/Exóticos, com 28,24 pontos, seguida por América do Norte com 5,07 pontos. Dos 30 destinos mais relevantes, 21 se localizam na América do Sul, quatro na América do Norte, três na Europa, um na Ásia e um na Oceania. O Brasil possui 19 destinos dos 30 mais relevantes para o perfil, seguido por Canadá com dois destinos. As cidades de Bonito no Brasil e Dubai nos Emirados Árabes Unidos são as mais relevantes para o perfil com 1,90 e 1,76 pontos respectivamente. Considerando apenas o Brasil, os destinos mais relevantes após Bonito são: Chapada Diamantina com 1,54 pontos e São Paulo com 1,52 pontos de relevância.

Considerando as 100 primeiras atrações mais relevantes para o perfil, a América do Sul possui 66 atrações, seguida pela Europa com nove atrações. Bonito é a cidade com mais atrações, dez no total, seguida de Dubai com cinco atrações. No Brasil, o estado do Mato Grosso do Sul aparece com dez atrações entre as 100 primeiras, seguido por São Paulo e Santa Catarina com oito e seis atrações respectivamente. As duas atrações mais bem avaliadas para o perfil são: Gruta de São Miguel em Bonito com 45,35% e Orquidário Binot em Petrópolis com 43,82%.

4.3 Popularidade x relevância do perfil

Essa categoria busca responder à questão: Os destinos mais populares no TripAdvisor são de fato os mais relevantes para determinado perfil? Seguindo o modelo de McKenzie e Adams (2018), a popularidade de um destino é medida com base no número de visitas ou número de estrelas dos usuários. No entanto, considerando um determinado perfil, os destinos mais populares são realmente os que apresentam maior relevância com relação ao perfil? Entender essa diferença com base na opinião dos turistas é importante, pois, permite identificar lacunas de destinos que possam oferecer experiências mais relevantes para determinado público, porém, talvez não sejam tão populares. Os resultados foram avaliados para todos os perfis, porém, optou-se por exibir nessa pesquisa apenas os resultados dos perfis Natureza/Exóticos e Cultura, os quais tiveram mais categorias do TripAdvisor mapeadas nos perfis turísticos do estudo.

4.3.1 Mapeamento categoria TripAdvisor x perfil turístico

Conforme exibido na seção 2.4.2, quando um usuário realiza uma avaliação no TripAdvisor, o mesmo insere a data de visita à atração turística. A partir dessa informação é possível identificarmos os destinos mais visitados por período e por categoria do TripAdvisor. Cada atração possui suas categorias que podem ser inseridas pelo usuário ou especialistas da plataforma. Para que fosse possível comparar a popularidade de determina atração x sua relevância para determinado perfil, foi necessário realizar um mapeamento entre as categorias do TripAdvisor e os perfis. Esse mapeamento buscou apontar um perfil turístico criado no trabalho para cada categoria do TripAdvisor. Esse processo foi desenvolvido em conjunto com os especialistas e buscou identificar o perfil mais próximo de determinada categoria, exemplo: as categorias do TripAdvisor Museus, Museus especializados, Locais históricos, Galerias de arte, entre outras foram mapeadas no perfil Cultura.

No total, 111 categorias foram extraídas das atrações do TripAdvisor, porém, algumas muito abstratas ou específicas e não se encaixaram nos perfis, como: transportes, ao ar livre, outras atividades ao ar livre, navios e cemitérios entre outras. Ressalta-se que uma mesma atração pode possuir mais de uma categoria no TripAdvisor. Por fim, 94 categorias do TripAdvisor foram mapeadas no seu respectivo perfil turístico. A Tabela 14 exhibe o total de categorias mapeadas por cada perfil.

Tabela 14 – Quantidade de categorias do TripAdvisor por perfil

PERFIL	QUANTIDADE
Cultura	17
Natureza/Exóticos	16
Relaxante	12
Família	11
Aventura	11
Paisagem/Arquitetura	7
Compras	5
Gastronomia	5
Vida Noturna	4
Romântico	3
Religioso	2
Praia	1

Fonte: elaborado pelo autor (2021).

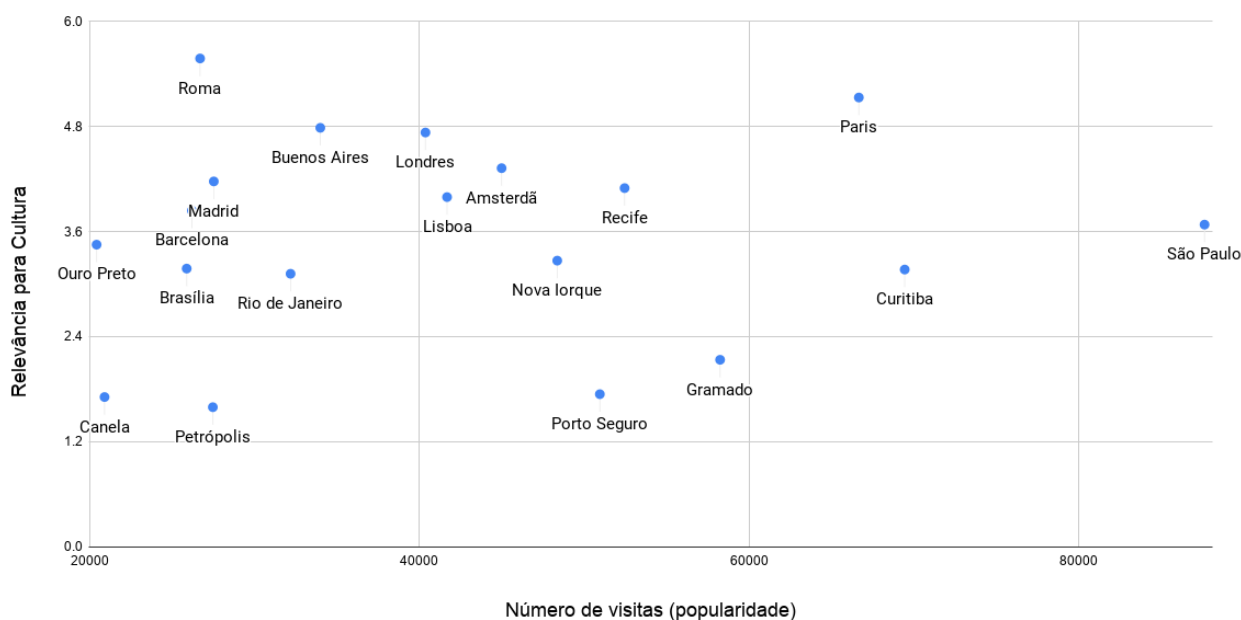
No TripAdvisor, apenas a categoria Praia não possui outras relacionadas, ou seja, toda atração do tipo praia, é simplesmente classificada na categoria Praia, que se tornou o perfil Praia.

Para identificar a popularidade do destino, considerou-se o somatório do número de visitas em suas atrações ligadas ao perfil em estudo. Para a análise, usou-se os 20 destinos com maior número de visitas em cada perfil, ou seja, os 20 mais populares nas categorias do TripAdvisor que são ligadas àquele perfil. Para cada um desses destinos, buscou-se o seu grau de relevância para o perfil, considerando o somatório das 30 atrações mais pertinentes para aquele perfil. Dessa forma, o resultado de relevância considera não apenas as cinco atrações mais pertinentes para cada perfil, e sim um somatório de todas as atrações do destino.

4.3.2 Popularidade X Relevância de destinos para o perfil Cultura

A lista de destinos mais populares para Cultura inclui destinos nacionais e internacionais. O Gráfico 13 exibe uma comparação na qual é possível avaliar as duas dimensões estudadas (Popularidade x Relevância).

Gráfico 13 – Popularidade X Relevância de destinos para o perfil Cultura



Fonte: dados da pesquisa (2021).

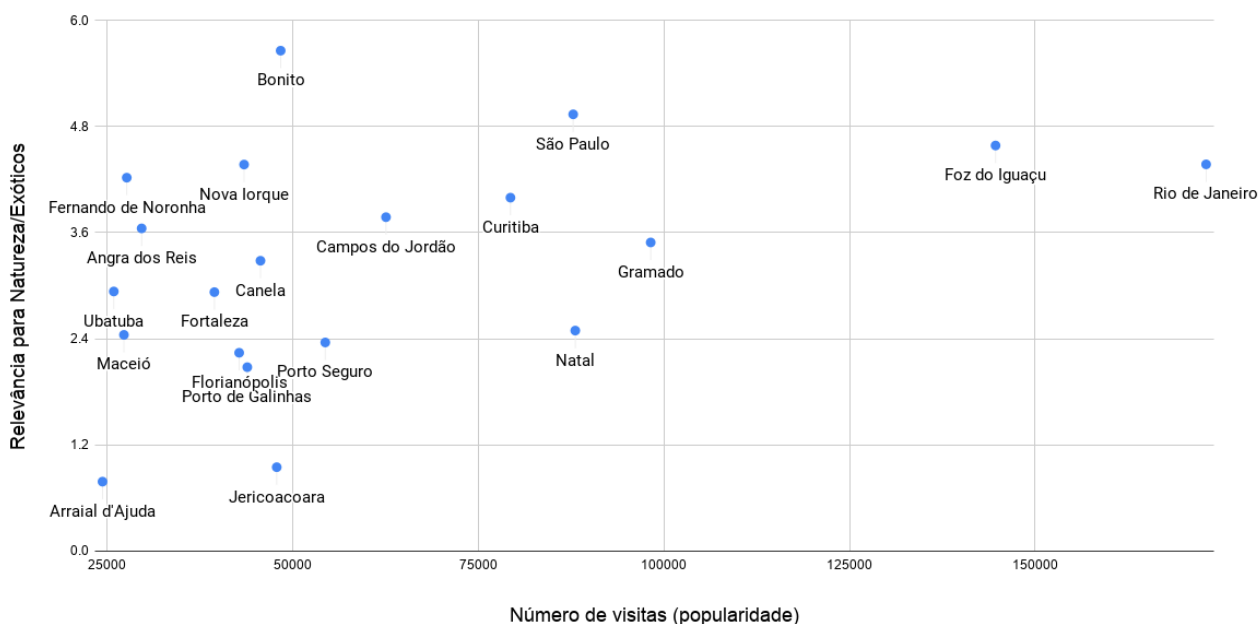
No Gráfico 13 é possível visualizar São Paulo como o destino mais popular entre todos, com mais de 80 mil visitas entre 2013 e junho de 2019. Conforme supracitado, a popularidade é avaliada pelo somatório de visitas em atrações nas categorias mapeadas ao perfil Cultura, ou seja, considera as visitas pertinentes para a análise e não todas as visitas ao destino no período. Apesar de ser o destino mais popular, São Paulo não é o destino mais relevante se considerarmos Cultura, ficando na posição 10° com 3,68 pontos. O destino mais similar ao perfil Cultura foi a cidade de Roma, na Itália com pontuação de 5,57 pontos. No entanto, se observarmos a escala de popularidade, Roma aparece na 15° posição, com cerca de 26 mil visitas no período estudado. Paris aparece como a cidade mais popular e com maior relevância para Cultura, ficando na segunda posição de popularidade com cerca de 66mil visitas e também na segunda posição considerando a sua relevância para o perfil, com 5,13 pontos. Outras cidades também relevantes para Cultura como Buenos Aires, Londres e Amsterdã com 4,78, 4,73 e 4,32 pontos respectivamente, são intermediárias na lista de popularidade, ficando nas posições 11°, 10° e 8°. Considerando destinos nacionais, Recife apresenta o melhor resultado para as duas variáveis, com 4,09 pontos de relevância para Cultura e cerca de 52mil visitas no período estudado.

O resultado do perfil Cultura demonstra que, embora existam destinos entre os mais populares e relevantes como Paris por exemplo, há também destinos interessantes e relevantes para Cultura, porém com menor popularidade. Considerando que o escopo de análise considera avaliações em português, é natural que cidades brasileiras sejam mais populares pela facilidade de visitação a tais destinos em detrimento a destinos internacionais. No entanto, o fato de destinos como Nova Iorque e Paris aparecerem entre os mais populares é importante para se observar um equilíbrio entre os dados, ou seja, considerando também destinos internacionais. Esse fator pode ter relação com um determinado tipo de público que faz esse tipo de avaliação e pode envolver fatores como escolaridade e nível socioeconômico. Se outras fontes de dados são consideradas, os resultados podem ser diferentes, assim, é importante ratificar que o estudo apresenta resultados sob a ótica dos visitantes e opinião dos brasileiros que realizam avaliações *online*.

4.3.3 Popularidade X Relevância de destinos para o perfil Natureza/Exóticos

Na lista de destinos mais populares para Natureza/Exóticos há uma predominância de destinos nacionais. O Gráfico 14 exibe o gráfico Scatter da comparação, no qual é possível avaliar as duas dimensões estudadas (Popularidade x Relevância).

Gráfico 14 – Popularidade X Relevância de destinos para o perfil Natureza/Exóticos



Fonte: dados da pesquisa (2021).

Com relação ao perfil Natureza/Exóticos, a cidade do Rio de Janeiro é o destino mais visitado, com mais de 175mil visitas entre 2013 e junho de 2019. Esse destino aparece na quarta posição considerando a relevância de suas atrações para Natureza/Exóticos. O destino mais similar ao perfil é Bonito no Mato Grosso do Sul com 5,65 pontos, destino esse também que teve uma atração escolhida como característica do perfil. Embora seja um destino com muita relevância, é um destino pouco visitado, com cerca de 48 mil visitas no período. A cidade de Nova Iorque, única representante internacional na lista, possui poucas visitas (cerca de 50 mil) e boa relevância (4,37 pontos). Analisando-se as atrações de Nova Iorque, as principais com maior pontuação no perfil foram One World Observatory, Intrepid Sea, Air & Space Museum e Empire State Building. O Central Parque aparece apenas na quinta posição. Isso ocorre, porque o perfil envolve tanto

aspectos de natureza quanto exóticas. Portanto, se um turista relata uma experiência exótica ao visitar o Empire State, isso seria relevante ao classificar tal atração perante o perfil.

É interessante visualizar na imagem, que embora o Rio de Janeiro seja conhecido por sua natureza, a cidade de São Paulo, obteve maior relevância com relação ao perfil Natureza/Exóticos. As cinco primeiras atrações mais relevantes ao perfil no destino São Paulo, são Jardim Botânico com 37,67 pontos, Farol Santander com 29,48 pontos, Zoológico de São Paulo com 28,96 pontos, Parque Estadual da Cantareira com 28,15 pontos e Aquário de São Paulo com 28 pontos. Observando a comparação das duas cidades, é importante ressaltar que o perfil em análise envolve aspectos de Natureza e Exóticos, ou seja, o processo classificatório considerou aspectos (termos) que representassem as duas categorias. Os especialistas escolheram a atração exótica Burj Khalifa como característica desse perfil, assim, é esperado que o perfil seja um pouco misto. Isso justifica a presença da atração Farol Santander entre as primeiras de São Paulo. Outro fator, é que natureza não é somente paisagem, sim com relação a fauna e flora, por isso a presença de atrações como Aquário e Zoológico de São Paulo. Na cidade do Rio de Janeiro, a atração mais relevante para o perfil também é o Jardim Botânico do Rio com 30,93 pontos, seguida pelo Parque Lage com 27,42 pontos e o AquaRio com 24,31 pontos, superando atrações conhecidas como Corcovado – Cristo Redentor e Pão de Açúcar. A forma como um turista relata sua experiência exerce completo impacto no processo de classificação, assim, nota-se que um turista relata muito mais aspectos relacionados a Natureza numa visita ao Jardim botânico, do que em uma visita ao Pão de Açúcar. Os resultados apontam a existência de atrações mais relevantes para o perfil, que não são tão populares, como Bonito por exemplo.

4.4 Similaridade entre destinos

A possibilidade em identificar destinos que ofereçam experiências similares é importante para o turista. Por exemplo, responder à questão de: qual destino nacional oferece experiências mais próximas a Paris? Essa informação pode melhorar o processo decisório do turista, que poderia optar por experiências semelhantes em destinos que se

encaixem melhor em sua condição, considerando variáveis como distância, custo, locomoção, clima entre outras.

Considerando que 124 destinos estavam classificados nos 12 perfis, o objetivo é avaliar a similaridade entre todos os destinos, ou seja, cada destino seria comparado com outros 123, para encontrar os destinos mais similares. Utilizou-se a distância Euclidiana para avaliar o grau de similaridade entre os destinos. A Tabela 15 exibe cada destino da pesquisa e os dois destinos mais similares de acordo com os 12 perfis.

Tabela 15 – Similaridade entre destinos considerando os perfis turísticos

Continua

DESTINO	DESTINOS SIMILARES	SIMILARIDADE	DESTINO	DESTINOS SIMILARES	SIMILARIDADE
Amsterdã	Londres	0.84	Bento Gonçalves	Garibaldi	1.09
	Cidade do México	0.93		Gramado	2.18
Angra dos Reis	Ilhabela	0.67	Berlim	Londres	0.67
	Morro de São Paulo	0.78		Moscou	0.72
Aracaju	Porto Seguro	1.01	Blumenau	Campos do Jordão	0.89
	Fernando de Noronha	1.09		Curitiba	0.98
Arraial d'Ajuda	Porto Seguro	0.93	Bogotá	Tiradentes	0.68
	Cabo Frio	1.07		Lima	0.70
Arraial do Cabo	Angra dos Reis	0.78	Bombinhas	Maragogi	0.67
	Ilhabela	0.90		Porto de Galinhas	0.74
Atenas	Bruxelas	0.77	Bonito	Chapada dos Veadeiros	0.98
	Cairo	0.87		Chapada Diamantina	0.99
Auckland	Brisbane	0.78	Boston	Frankfurt	0.70
	Chapada dos Veadeiros	1.01		Montevideú	0.81
Balneário Camboriú	Cabo Frio	0.67	Brasília	Belém	0.86
	Ilhabela	0.83		Lima	1.01
Bangcoc	Panamá	0.75	Brisbane	Auckland	0.78
	Los Angeles	1.26		Chapada Diamantina	0.88
Barcelona	Buenos Aires	0.84	Bruxelas	Ouro Preto	0.75
	Ouro Preto	0.87		Holambra	0.75
Belo Horizonte	Curitiba	0.89	Budapeste	Cusco	0.80
	Blumenau	1.06		Istambul	0.95
Belém	Brasília	0.86	Buenos Aires	Porto Alegre	0.76
	Diamantina	0.88		Cidade do México	0.83

Tabela 15 – Similaridade entre destinos considerando os perfis turísticos

Continua

DESTINO	DESTINOS SIMILARES	SIMILARIDADE	DESTINO	DESTINOS SIMILARES	SIMILARIDADE
Cabo Frio	Balneário Camboriú	0.67	Diamantina	Tiradentes	0.73
	Ilhabela	0.68		Belém	0.88
Cairo	Moscou	0.84	Dubai	Foz do Iguaçu	1.18
	Atenas	0.87		Los Angeles	1.37
Caldas Novas	Foz do Iguaçu	1.36	Dublim	Munique	0.83
	San Carlos de Bariloche	1.41		Petrópolis	1.04
Campos do Jordão	Porto Alegre	0.70	Estocolmo	Copenhague	0.83
	Boston	0.82		Istambul	0.92
Cancun	Fortaleza	0.83	Fernando de Noronha	Ilhabela	0.72
	Natal	1.12		Morro de São Paulo	0.77
Canela	Santiago	1.26	Florença	Milão	0.69
	Gramado	1.31		Jerusalém	0.91
Canoa Quebrada	Morro de São Paulo	0.82	Florianópolis	Ilhabela	0.56
	Praia do Forte	0.93		Ubatuba	0.68
Caraíva	Parque Estadual de Jalapão	0.57	Fortaleza	Cancun	0.83
	Nassau	1.08		Maceió	1.09
Cartagena	Montreal	0.98	Foz do Iguaçu	Porto Alegre	1.02
	Zurique	1.19		Tóquio	1.03
Chapada Diamantina	Chapada dos Veadeiros	0.87	Frankfurt	Nova Iorque	0.58
	Brisbane	0.88		Boston	0.70
Chapada dos Veadeiros	Palmas	0.86	Garibaldi	Bento Gonçalves	1.09
	Chapada Diamantina	0.87		Gramado	2.04
Cidade do Cabo	Palmas	0.83	Goiânia	Foz do Iguaçu	1.07
	San Carlos de Bariloche	0.84		Los Angeles	1.09
Cidade do México	Buenos Aires	0.83	Gramado	Canela	1.31
	Madrid	0.84		Santiago	1.65
Copenhague	Londres	0.80	Holambra	Bruxelas	0.75
	Istambul	0.80		Melbourne	0.84
Curitiba	Belo Horizonte	0.89	Ilhabela	Florianópolis	0.56
	Toronto	0.95		Angra dos Reis	0.67
CUSCO	Budapeste	0.80	Istambul	Copenhague	0.80
	Zurique	0.81		Estocolmo	0.92
Córdoba	Jerusalém	1.01	Itaipava	Pomerode	1.15
	Bruxelas	1.06		Joanesburgo	1.44

Tabela 15 – Similaridade entre destinos considerando os perfis turísticos

Continua

DESTINO	DESTINOS SIMILARES	SIMILARIDADE	DESTINO	DESTINOS SIMILARES	SIMILARIDADE
Jericoacoara	Arraial do Cabo	1.10	Montreal	Cartagena	0.98
	Canoa Quebrada	1.14		Bogotá	1.06
Jerusalém	Florença	0.91	Morro de São Paulo	San Andres	0.61
	Ouro Preto	0.93		Praia do Forte	0.70
Joanesburgo	Santiago	0.87	Moscou	Berlim	0.72
	Pomerode	1.1		Cairo	0.84
Joinville	Tiradentes	0.91	Munique	Dublin	0.83
	Diamantina	0.92		Berlim	1.00
João Pessoa	Salvador	0.74	Nassau	Trindade	0.79
	Santos	0.86		Rio Quente	0.79
Las Vegas	Dubai	1.53	Natal	Balneário Camboriú	1.01
	Foz do Iguaçu	1.57		Fortaleza	1.11
Lima	Bogotá	0.7	Navegantes	San Andres	0.56
	Cidade do México	1.57		Trindade	0.66
Lisboa	Barcelona	1.05	Nova Iorque	Frankfurt	0.58
	Madrid	1.15		Porto Alegre	0.81
Londres	Berlim	0.67	Olinda	Salvador	0.68
	Copenhague	0.8		Vitória	0.75
Los Angeles	Goiânia	1.09	Orlando	Goiânia	2.22
	Nova Iorque	1.12		Dubai	2.25
Maceió	Fortaleza	1.09	Ouro Preto	Tiradentes	0.45
	Cancun	1.15		Bruxelas	0.75
Machu Picchu	São Petersburgo	1.14	Palmas	Cidade do Cabo	0.83
	Estocolmo	1.16		Chapada dos Veadeiros	0.86
Madrid	Cidade do México	0.84	Panamá	Bangcoc	0.75
	Barcelona	1.05		Porto	1.35
Maragogi	Bombinhas	0.67	Paraty	Ilhabela	1.02
	Morro de São Paulo	0.76		Fernando de Noronha	1.08
Melbourne	Pomerode	0.74	Paris	Milão	1.02
	Bruxelas	0.79		Veneza	1.1
Miami	Los Angeles	1.14	Parque Estadual de Jalapão	Caraíva	0.57
	Goiânia	1.28		Nassau	1.03
Milão	Florença	0.69	Petrópolis	Dublin	1.04
	Bogotá	0.93		Bruxelas	1.09
Montevideu	Boston	0.81	Pirenópolis	Palmas	1.05
	Campos do Jordão	1.09		Cidade do Cabo	1.16

Tabela 15 – Similaridade entre destinos considerando os perfis turísticos

			Conclusão		
DESTINO	DESTINOS SIMILARES	SIMILARIDADE	DESTINO	DESTINOS SIMILARES	SIMILARIDADE
Pomerode	Melbourne	0.74	San Carlos de Bariloche	Cidade do Cabo	0.84
	Bruxelas	0.84		Chapada dos Veadeiros	1.07
Porto	Zurique	0.83	Santiago	Joanesburgo	0.87
	Frankfurt	0.97		Canela	1.26
Porto Alegre	Campos do Jordão	0.70	Santos	Vitória	0.70
	Frankfurt	0.72		Salvador	0.80
Porto Seguro	Cabo Frio	0.76	São Luís	Salvador	0.85
	Ilhabela	0.81		Olinda	0.90
Porto de Galinhas	Bombinhas	0.74	São Paulo	Porto Alegre	0.86
	Ilhabela	0.78		Campos do Jordão	0.92
Praga	Cusco	1.06	São Petersburgo	Cairo	1.01
	Budapeste	1.15		Machu Picchu	1.14
Praia do Forte	San Andres	0.47	Tiradentes	Ouro Preto	0.45
	Morro de São Paulo	0.70		Bogotá	0.68
Punta Cana	Willemstad	1.02	Toronto	Curitiba	0.95
	Balneário Camboriú	1.38		Frankfurt	0.99
Punta del Este	Montevideú	1.32	Trindade	Navegantes	0.66
	Pirenópolis	1.33		Nassau	0.79
Recife	Porto	1.05	Tóquio	Bruxelas	0.77
	Santos	1.16		Nova Iorque	0.86
Rio Quente	Nassau	0.79	Ubatuba	Florianópolis	0.68
	Brisbane	1.15		Bombinhas	0.75
Rio de Janeiro	Palmas	0.86	Veneza	Paris	1.10
	Porto de Galinhas	0.87		Florença	1.32
Roma	Florença	1.26	Vitória	Santos	0.70
	Milão	1.30		Olinda	0.75
Salvador	Olinda	0.68	Willemstad	Porto de Galinhas	0.93
	João Pessoa	0.74		Balneário Camboriú	0.96
San Andres	Praia do Forte	0.47	Zurique	Cusco	0.81
	Navegantes	0.56		Porto	0.83

Fonte: elaborado pelo autor (2021).

Para cada um dos 124 destinos, os dois destinos mais similares foram exibidos em conjunto com o resultado do cálculo da distância Euclidiana. Avaliando os resultados é possível perceber que a similaridade pode ser unidirecional ou bidirecional. O destino mais similar a Londres por exemplo, é Berlim com distância Euclidiana de 0,67 e o destino

mais similar a Berlim é Londres, também com distância de 0,67. Já o destino mais similar a Paris é Milão, com distância 1,02, no entanto, o destino mais similar a Milão é Florença com 0,69. Observou-se também que os destinos mais similares podem ser internacionais ou nacionais. O destino mais similar a Ouro Preto por exemplo, é Tiradentes, um resultado esperado, porém o segundo mais similar é Bruxelas, na Bélgica. A Figura 45 exibe um comparativo do perfil de Ouro Preto e Bruxelas.

Figura 45 – Comparativo de perfis entre Bruxelas e Ouro Preto

```
In [21]: 1 df_tfidf.loc[df_tfidf['Destino'] == 'Ouro Preto'].sort_values(by='norm', ascending=False)
```

Out[21]:

	Destino	Perfil	Similaridade	norm
1043	Ouro Preto	Religioso	1.4862	0.400348
1033	Ouro Preto	Paisagem/Arquitetura	1.4348	0.384947
1037	Ouro Preto	Aventura	1.3070	0.346656
1041	Ouro Preto	Romântico	1.3028	0.345398
1039	Ouro Preto	Compras	1.2945	0.342911
1038	Ouro Preto	Praia	1.2833	0.339555
1042	Ouro Preto	Natureza/Exóticos	1.2791	0.338297
1032	Ouro Preto	Cultura	1.2431	0.327511
1040	Ouro Preto	Relaxante	1.2263	0.322477
1036	Ouro Preto	Gastronomia	1.2236	0.321668
1034	Ouro Preto	Vida Noturna	1.1944	0.312919
1035	Ouro Preto	Família	0.8959	0.223484

```
In [22]: 1 df_tfidf.loc[df_tfidf['Destino'] == 'Bruxelas'].sort_values(by='norm', ascending=False)
```

Out[22]:

	Destino	Perfil	Similaridade	norm
263	Bruxelas	Religioso	1.6416	0.446908
252	Bruxelas	Cultura	1.2694	0.335391
253	Bruxelas	Paisagem/Arquitetura	1.2098	0.317534
257	Bruxelas	Aventura	1.1571	0.301744
261	Bruxelas	Romântico	1.1319	0.294193
259	Bruxelas	Compras	1.1278	0.292965
256	Bruxelas	Gastronomia	1.1069	0.286703
255	Bruxelas	Família	1.0861	0.280471
262	Bruxelas	Natureza/Exóticos	1.0695	0.275497
254	Bruxelas	Vida Noturna	1.0653	0.274239
260	Bruxelas	Relaxante	1.0028	0.255513
258	Bruxelas	Praia	0.8833	0.219709

Fonte: dados da pesquisa (2021).

Entre os cinco perfis mais relevantes para ambos os destinos, quatro são os mesmos: (Religioso, Paisagem/Arquitetura, Aventura e Romântico), seguindo exatamente a mesma ordem de relevância para o destino. Com isso, nota-se que os visitantes descrevem ambos os locais de forma semelhante, por possuírem características ou

proporcionarem experiências semelhantes. O resultado de similaridade entre os destinos aponta que é possível que o turista encontre experiências similares em destinos nacionais para destinos internacionais, como no caso de Ouro Preto. Num trabalho futuro, seria interessante aprofundar nos resultados dessa comparação para permitir um entendimento melhor desse relacionamento com foco na exploração das informações ou tomada de decisão.

4.5 Perfil turístico de destinos mais visitados

Essa categoria busca analisar o perfil dos 10 destinos mais visitados pelos brasileiros em todo período estudado. Os resultados são apresentados em dois níveis: primeiro, busca-se explorar a variação dos perfis dos dez destinos nacionais mais visitados por período. Em seguida, apresenta-se a variação dos perfis dos dez destinos internacionais mais visitados por período. Entender o perfil desses destinos ajuda a entender as preferências dos brasileiros com relação ao perfil de destinos e como essa preferência tem se alterado ao longo do tempo.

A base de dados contém avaliações sobre visitas às atrações entre os anos de 2011 a 2019. Utilizou-se a soma das cinco das atrações mais relevantes para cada perfil dos dez destinos nacionais mais visitados, ou seja, se São Paulo e Rio de Janeiro foram as duas primeiras cidades que mais receberam visitas, o resultado de cada um dos 12 perfis em determinado ano seria o somatório do resultado individual de cada uma das dez cidades.

Buscou-se separar os resultados entre nacional e internacional com o objetivo de avaliar a variação ano a ano para diferentes tipos de turistas. Embora aqui seja apresentado somente essa separação com a base de dados classificada, é possível extrair a mesma informação, por continente, país específico, por estado ou até mesmo a variação de perfil considerando as atrações de um mesmo destino ao longo do tempo.

4.5.1 Variação dos perfis turísticos de destinos nacionais

A primeira análise compreende a variação dos perfis dos dez destinos nacionais mais visitados entre 2011 e 2019. A Tabela 16 apresenta os dez destinos mais visitados no período.

Tabela 16 – Destinos nacionais mais visitados entre 2011 e 2019

POSIÇÃO	DESTINO	VISITAS
1	Rio de Janeiro	203178
2	São Paulo	181397
3	Gramado	169272
4	Foz do Iguaçu	108405
5	Curitiba	105310
6	Porto Seguro	104925
7	Brasília	79599
8	Salvador	68269
9	Natal	67632
10	Fortaleza	67308

Fonte: elaborado pelo autor (2021).

No Brasil, a cidade do Rio de Janeiro foi a mais visitada no período com cerca de 2 milhões de visitas. Entre as dez primeiras, 40% são cidades do nordeste brasileiro. A Tabela 17 exibe a relevância de perfis dos dez destinos mais visitados no período.

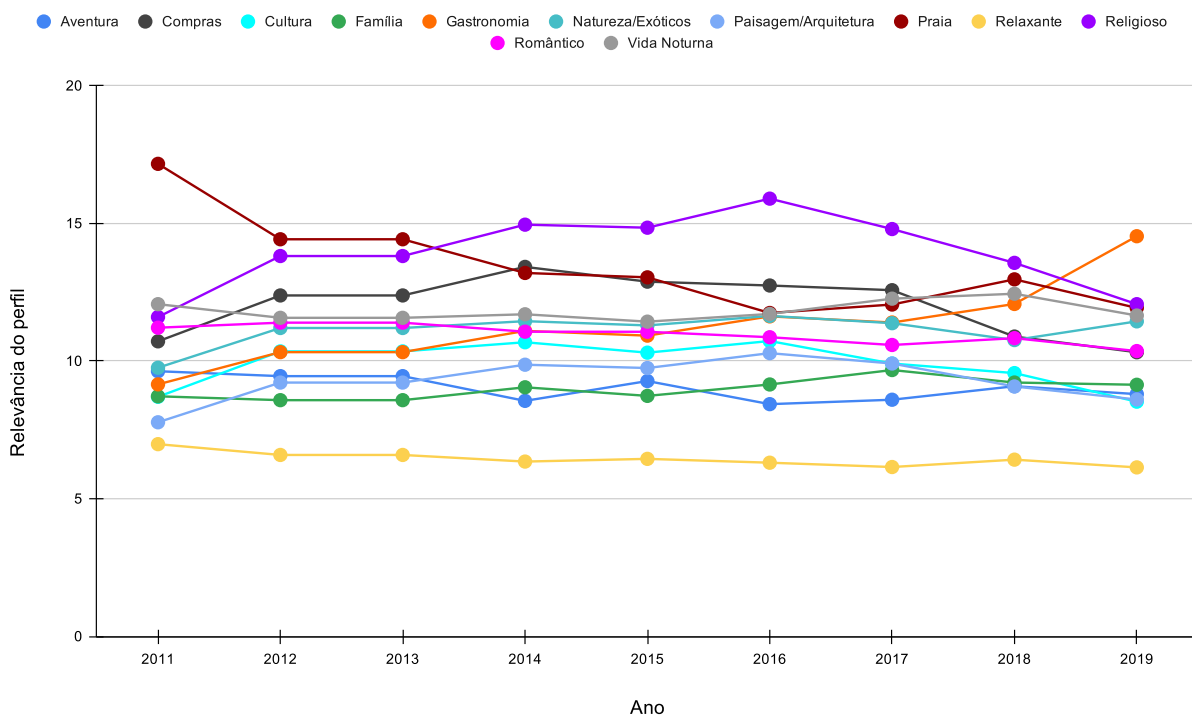
Tabela 17 – Relevância de perfis dos dez destinos mais visitados entre 2011 e 2019

POSIÇÃO	PERFIL	RELEVÂNCIA
1	Religioso	125.3671
2	Praia	120.9638
3	Compras	108.3314
4	Vida Noturna	106.4303
5	Gastronomia	101.4669
6	Natureza/Exóticos	100.1398
7	Romântico	98.7913
8	Cultura	89.1356
9	Paisagem/Arquitetura	83.7248
10	Aventura	81.2961
11	Família	80.8619
12	Relaxante	57.9929

Fonte: elaborado pelo autor (2021).

O perfil *Religioso* e *Praia* aparecem como os mais relevantes entre os dez destinos mais visitados no período. A próxima análise apresenta a variação entre a relevância dos perfis perante os dez destinos nacionais mais visitados em cada ano. O Gráfico 15 apresenta o resultado.

Gráfico 15 – Variação de perfis entre os dez destinos nacionais mais visitados



Fonte: dados da pesquisa (2021).

O perfil Praia aparece em 2011 como principal entre os dez destinos com mais visitadas nesse ano. É possível observar que esse perfil se mantém entre os principais, sendo o segundo perfil preferido no período estudado. Esse é um ponto interessante, pois, imaginava-se que esse perfil seria o principal entre os destinos mais visitados. O perfil Religioso foi o mais relevante em todo o período, no entanto, nota-se que houve um acréscimo de relevância entre os anos 2011 e 2016, mas desde então a relevância desse perfil vem apresentando queda. Como a análise envolve a pertinência da atração perante o perfil e a visita, pode-se sugerir que atrações com esse tipo de perfil vêm sendo menos visitadas. A visita de pontos com boa relevância para o perfil Gastronomia apresenta aumento durante o período estudado, atingindo 2019 como o perfil com mais relevância entre os dez destinos mais visitados. Comportamento semelhante se observa para o perfil Vida noturna. O perfil Cultura se manteve estável até 2016, e desde então, nota-se que tem caído. O perfil Relaxante aparece bem abaixo dos demais, como o perfil com menor relevância entre os dez destinos mais visitados ao longo do período.

4.5.2 Variação dos perfis turísticos de destinos internacionais

A segunda análise compreende a variação dos perfis dos dez destinos internacionais mais visitados entre 2011 e 2019. A Tabela 18 apresenta os dez destinos mais visitados no período.

Tabela 18 – Destinos internacionais mais visitados entre 2011 e 2019

POSIÇÃO	DESTINO	VISITAS
1	Nova Iorque	85977
2	Paris	84341
3	Buenos Aires	81953
4	Orlando	76449
5	Lisboa	67196
6	Montevideu	43922
7	Roma	40683
8	Barcelona	39021
9	Porto	31655
10	Londres	30221

Fonte: elaborado pelo autor (2021).

A Europa é o continente mais visitado por brasileiros, com 60% dos destinos, no entanto, o destino mais visitado é Nova Iorque. Ao menos entre os turistas que fazem avaliações sobre as visitas, há mais visitas para Europa e Estados Unidos, do que para a América do Sul, que teoricamente seria mais fácil para os brasileiros visitarem, devido a fatores como custo ou proximidade. A Tabela 19 exibe a relevância de perfis dos dez destinos mais visitados no período.

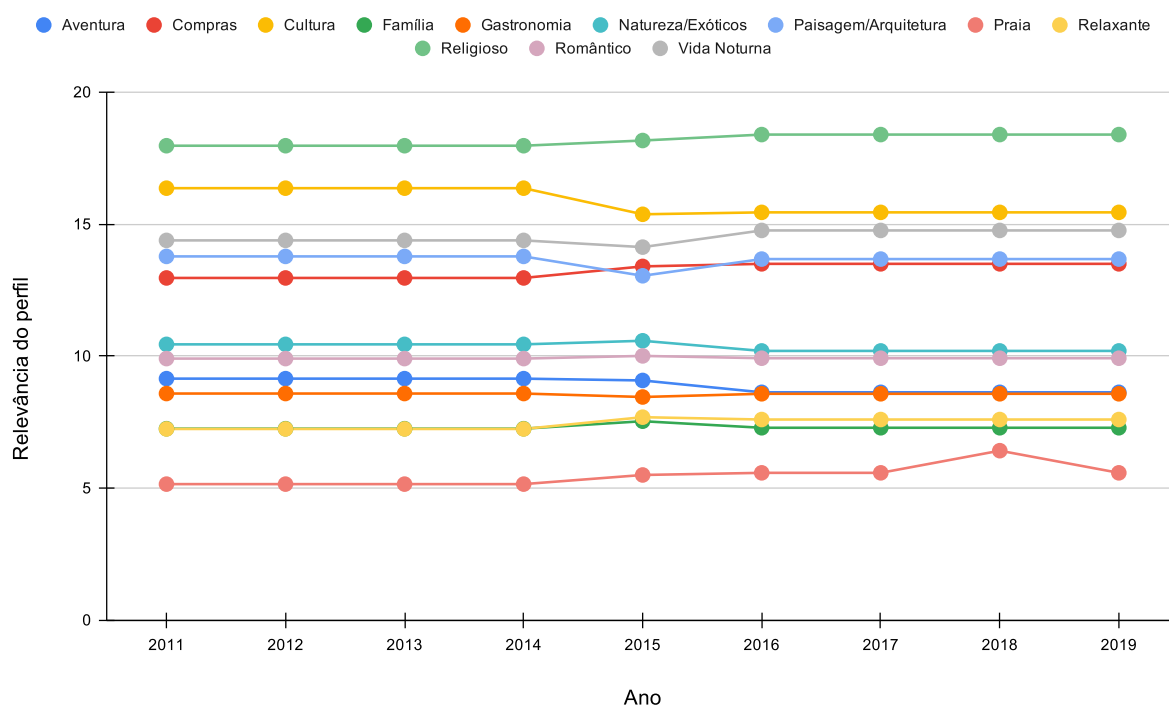
Tabela 19 – Relevância de perfis dos dez destinos mais visitados entre 2011 e 2019

POSIÇÃO	PERFIL	RELEVÂNCIA
1	Religioso	163.6462
2	Cultura	142.6323
3	Vida Noturna	130.739
4	Paisagem/Arquitetura	122.8867
5	Compras	119.2556
6	Natureza/Exóticos	93.1814
7	Romântico	89.3457
8	Aventura	80.2124
9	Gastronomia	77.104
10	Relaxante	67.0717
11	Família	65.7125
12	Praia	49.2865

Fonte: elaborado pelo autor (2021).

O perfil *Religioso* aparece novamente como o mais relevante para os dez destinos internacionais mais visitados no período. No entanto, há uma mudança brusca no perfil internacional. O perfil *Cultura* que é apenas o oitavo entre os destinos nacionais, aparece como o segundo perfil mais relevante entre os destinos internacionais. O perfil *Praia*, que é o segundo mais relevante para os destinos nacionais, aparece em último lugar para de relevância para os destinos internacionais. A próxima análise apresenta a variação entre a relevância dos perfis para os dez destinos internacionais mais visitados em cada ano. O Gráfico 16 apresenta o resultado.

Gráfico 16 – Variação de perfis entre os dez destinos internacionais mais visitados



Fonte: dados da pesquisa (2021).

Observa-se uma variação menor entre os perfis, ou seja, ocorre pouca variação entre os destinos internacionais mais visitados. O perfil *Cultura* apresenta queda após 2014 enquanto o perfil *Religioso* apresenta aumento após 2014. O perfil *Praia* é o que apresenta menor relevância entre os destinos internacionais.

4.5.3 Resultado das variações dos perfis

O perfil *Religioso* é o mais relevante entre os destinos mais visitados nacionais e internacionais. Nos últimos anos o perfil de destinos mais visitados nacionais, tem variado um pouco, como no exemplo de *Religioso* e *Gastronomia*. Os perfis *Relaxante* e *Família* são os que apresentam menor relevância entre os destinos nacionais e internacionais (desconsiderando *Praia*). Já para destinos internacionais, ocorre variação mínima os perfis mais visitados. A diferença entre o perfil nacional e internacional pode ter relação com o fato de os brasileiros buscarem experiências diferentes nacionalmente e internacionalmente. A análise de perfil é interessante, ao passo que os destinos mais visitados podem variar ao longo do tempo. No entanto o perfil exibe uma outra escala, ou seja, independente do destino, exibe a preferência do turista.

Conhecer perfis turísticos é importante para que ações do governo possam ser desenvolvidas, como por exemplo, incentivo à Cultura entre destinos nacionais, tendo em vista que os destinos internacionais mais visitados possuem relevância maior nesse perfil em 57,7%. Além disso, entender essas preferências, permite que empresas turísticas possam oferecer produtos com experiências que atendam as preferências nacionais ou internacionais, como por exemplo: *Praia + Religioso*; ou internacional, como: *Religioso + Vida Noturna*.

5 CONCLUSÕES

As próximas seções descrevem as conclusões sobre a pesquisa, procurando responder as questões de pesquisa com foco nos objetivos. Nesse capítulo procura-se apresentar também as contribuições da pesquisa em cada esfera.

5.1 *Uso de avaliações*

A Internet permitiu a oferta e globalização de serviços por meio de plataformas digitais. Essa ruptura alterou formatos tradicionais de entrega de produtos e serviços, como reservar um hotel, reservar um carro ou escolher o que fazer em um determinado local turístico. Isso permitiu que novos “*players*” entrassem no mercado. Embora isso venha sofrendo alterações, uma das características da Internet é que usuários possuem preocupação com relação a segurança ou credibilidade sobre determinado serviço. Nesse sentido, as avaliações feitas em portais sobre produtos ou serviços aparecem como peça-chave para que determinado usuário, veja sob a ótica de outras pessoas, o grau de confiabilidade, qualidade ou relevância sobre determinada atração, serviço ou produto.

As avaliações são a nova forma de “Boca-a-Boca”, conceito que mais influencia os consumidores na sua tomada de decisão. As avaliações são uma fonte importante de informação sobre um produto, um serviço ou uma atração turística, no entanto, devido ao grande volume de dados, um usuário comum não consegue explorar todo o conhecimento existente nesse conteúdo. A riqueza desse conteúdo abre espaço para pesquisas e aplicação no mercado, como vem se mostrando nos últimos anos. Como se tratam de opiniões, se devidamente organizadas e processadas, diversos insights e descobertas podem ser realizadas nessa fonte de dados. Nessa pesquisa, utilizou-se a opinião de 3.4 milhões de pessoas sobre pontos de interesse turísticos para entender como eles seriam classificados. Portanto, seguindo o rigor metodológico de uma pesquisa, o conteúdo existente nas avaliações pode fornecer uma fonte inestimável de informações e conhecimento para ajudar a compreender determinado campo do saber.

Como demonstrado, as avaliações fornecem uma fonte importante de informações, no entanto, pesquisas futuras considerando outras fontes de dados para mapear os perfis de turistas, atrações e destinos podem ser interessante. Essa pesquisa apresenta resultados com base nas avaliações online de usuários, ou seja, reflete uma classificação turística com base num determinado tipo de turista. Os resultados foram apresentados de acordo com a metodologia e recortes propostos, os dados foram extraídos das bases de dados. Uma conferência sobre influências externas ao resultado podem ser assunto de pesquisas futuras e explorar informações dos usuários como perfil socioeconômico, idade, gênero entre outras variáveis que poderiam exercer implicações no resultado.

5.2 O papel da Ciência da Informação

A Internet permitiu um crescimento exponencial da produção e compartilhamento de informação. A maioria do conteúdo produzido precisa de estruturação, ou seja, os usuários apenas produzem e compartilham, sem se preocupar com a organização. Por um outro lado, dispositivos inteligentes como carros, relógios, assistentes virtuais, sensores ou rastreadores produzem dados na mesma velocidade, sem um padrão definido. O futuro é cada vez mais presente, não é ilógico pensar que cada dispositivo irá possuir um chip ou Bluetooth para transmissão de dados. Esse volume de dados e informações possui uma característica comum: precisa ser organizado com foco na recuperação e exploração de informações para geração de conhecimento. Nesse sentido, a CI exerce papel vital na solução desse problema.

No lado social, embora não faça parte do escopo dessa pesquisa, a discussão filosófica sobre como o conhecimento é organizado, usado e seu impacto no ser humano é vital e extremamente relevante, desde o período socrático. No lado técnico, é inconcebível desassociar da CI sua conexão com a tecnologia, conforme bem colocado por Saracevic (1995). Associar técnicas convencionais de organização da informação com tecnologias é o pilar na busca incessante de tornar o conhecimento efetivo. É necessário observar que o problema não é apenas organizar uma grande massa de dados e sim, que além de grande, esse conteúdo é completamente dinâmico. Se hoje são 3.4 milhões de

avaliações, semana que vem serão 3.8 milhões, assim, modelos que possam ser adaptados e que aprendam dinamicamente são essenciais.

5.3 Classificação semiautomática – modelo híbrido

No aspecto conceitual, embora o desempenho de práticas puramente automáticas venha sendo aperfeiçoadas ao longo do tempo, os estudos e aplicações mostram que tais métodos carecem quase que sempre de suporte de especialistas. As máquinas e os algoritmos são ferramentas, que nas mãos de especialistas do domínio podem trazer uma melhora significativa nos resultados. Os algoritmos de aprendizagem de máquina se favorecem do conhecimento dos especialistas, que em contrapartida se favorecem do processamento automático. Uma relação ganha-ganha. Nessa pesquisa, apresentou-se um método híbrido, semiautomático, envolvendo a participação de especialistas do domínio com métodos automáticos. Para que tais modelos obtenham sucesso é essencial considerar a validação dos modelos perante a ótica dos especialistas, como exibido nesse trabalho na comparação entre a classificação automática e a manual. Essa validação e refinamento em conjunto com o aprendizado dinâmico dos modelos podem ser a fator-chave no sucesso. O envolvimento dos agentes na pesquisa enriquece o trabalho, pois permite uma extração de conhecimento especializado sobre destinos, atrações e perfis turísticos. O conhecimento dos agentes é rico, fruto de suas experiências, tácito, e como tal, possivelmente não seria atingido por meio de um processo puramente automático.

No aspecto técnico, seguindo a teoria do aspecto conceitual, buscou-se a aplicação de métodos supervisionados de classificação, com base no conteúdo classificado pelos especialistas. Nessa pesquisa não se procurou empregar técnicas de aprendizado não supervisionado, ou seja, puramente automático. Os resultados do processo classificatório exibem uma variação entre os perfis, como o perfil Religioso que teve desempenho de 90% de acurácia, e o perfil Romântico, que teve 69%. Alguns fatores podem implicar nisso, como por exemplo, a escolha das atrações mais relevantes para cada perfil, ou ainda, a composição e características dos perfis definidos. Os resultados dos modelos de classificação exibidos nesse trabalho com desempenho acima de 70% sugerem um

sucesso na abordagem utilizada no problema de classificação de perfis turísticos de pontos de interesse usando avaliações.

Uma contribuição técnica dessa pesquisa está na melhoria dos repositórios de PLN para o idioma português, com a adição de 1.527 conjunções, 6.626 adjetivos e 260.524 verbos. Futuros trabalhos podem se beneficiar desses repositórios. A remoção de verbos, conjunções, adjetivos, criação de bigramas e trigramas, assim como todo o pré-processamento realizado foi fator importante no desempenho obtido. Afinal, tais termos acrescentam nada ou muito pouco para caracterizar uma atração turística.

A representação do texto usando um modelo de casamento de palavras treinado localmente no *corpus* de avaliações apresentou resultado muito superior a um modelo treinado com fontes de notícias tradicionais. Assim, pesquisas mais aprofundadas representando o texto com casamento de palavras precisam ser realizadas, dada a importância da semântica do texto. Os resultados da classificação supervisionada foram bons, principalmente se considerar que houve uma heurística na classificação do texto. Ao invés de classificar a avaliação propriamente, classificou-se a atração, assumindo que se uma avaliação fosse positiva, ela definitivamente refletia a característica do ponto de interesse. Pesquisas semelhantes usando a classificação da própria avaliação, apesar de possuírem um custo alto, talvez possam apresentar resultados ainda melhores.

A resposta da questão principal da pesquisa é realizada com base na análise qualitativa e quantitativa. Na esfera quantitativa, usando a validação cruzada, o algoritmo SVM classificou com precisão média de 73% as avaliações nos perfis corretos. Em alguns casos, como no caso do perfil Religioso, essa precisão atingiu 90%, ou seja, 9 acertos em 10 tentativas. Os resultados da análise qualitativa foram construídos sob o modelo que obteve melhor desempenho em relação a classificação manual dos especialistas, no caso, o método SVM. Na esfera qualitativa, nas listas e *rankings* utilizados, não houve nenhum caso de classificação incorreta, como o caso de uma Igreja sendo classificada como principal no perfil Vida Noturna. O fato de considerar grau de pertinência de uma atração perante um perfil é relevante, pois, conforme aponta a literatura, um mesmo destino apresenta características para distintos grupos de turistas. Os resultados sugerem sucesso na solução do problema com a metodologia adotada.

5.4 Perfis no turismo

Essa pesquisa apresenta um conjunto de 12 perfis turísticos como uma contribuição conceitual para a área do Turismo, pois, embora o conjunto tenha sido utilizado para classificar atrações com foco na opinião dos brasileiros, o mesmo não tem qualquer especificidade com relação ao Brasil, podendo ser aplicado também em outras pesquisas ou aplicações na área do Turismo. O grupo de perfis criado difere dos trabalhos existentes na literatura, porque oferece numa única escala, a possibilidade de classificar turista, atrações e destinos turísticos. Dessa forma, se torna possível a combinação entre atrações ou destinos turísticos com o perfil do turista na mesma escala. Para aproximar mais a experiência do turista com o destino em análise, uma pesquisa futura poderia considerar além dos pontos de interesse, outros itens de uma viagem como alimentação ou estadia. Outro fator relevante dos perfis é que proporcionam uma junção de experiências, ou seja, um casal cujo marido gosta de Aventura e a esposa gosta de Gastronomia, pode tomar uma decisão melhor a partir de destinos ou atrações que sejam compatíveis com ambos os perfis.

A pesquisa apresenta de forma inédita o uso de avaliações para formação de perfis turísticos de atrações. Classificar e agrupar as informações de mais de 3.4 milhões de opiniões sobre 124 destinos e 2.627 atrações turísticas contribui também para o próprio turismo, como uma nova forma de visualizar destinos e atrações. O uso de perfis favorece o esquema de classificação, pois segmenta destinos e atrações para um grupo de interesse específico. Assim, uma mesma atração pode representar 90% de experiência para o perfil Religioso, 70% para Arquitetura e 20% para Família. A classificação do destino com base em pontos de interesse é comum, no entanto, nessa pesquisa apresentou uma abordagem considerando a qualidade e quantidade de determinadas atrações como forma de gerar a classificação de um destino no perfil. Isso evita um problema de média, em que um destino com apenas uma atração nota 10 em Cultura, receba a mesma pontuação de outro destino com cinco atrações nota 10 para Cultura. Pelos resultados, nota-se talvez a importância de separar o perfil Natureza/Exóticos em dois, e assim, evitar que atrações, como o The Empire State Building, apareçam numa categoria que envolve Natureza. Considerando os resultados analíticos e sintéticos da

análise exploratória qualitativa, os demais perfis apresentaram coerência com relação aos destinos e retornados.

5.5 Popularidade x similaridade de destinos para um perfil turístico

Uma das questões indiretas da pesquisa é procurar identificar se as cidades mais visitadas por determinado perfil de turista, são de fato as mais relevantes ou similares para esse grupo específico. A partir da classificação de atrações e destinos em perfis turísticos isso se tornou possível. Os resultados apontam que apesar de alguns destinos serem os mais visitados por determinado grupo de interesse, eles não são os mais relevantes para determinado perfil. São Paulo é a cidade mais visitada para o perfil Cultura de acordo com o TripAdvisor, no entanto, Recife apresenta maior similaridade com esse perfil, ou seja, a soma de relevância das principais atrações de Recife no perfil Cultura é superior a São Paulo. O Rio de Janeiro é a cidade mais visitada no perfil Natureza/Exótico, porém, fica apenas em quarto lugar se considerar o grau de relevância para o perfil. Alguns resultados ratificam a relevância e popularidade de alguns destinos para determinados perfis, como por exemplo Paris, que é a segunda cidade mais visitada e a segunda também mais relevante para o perfil Cultura.

A identificação de destinos com boa relevância para um determinado grupo de interesse, porém menos visitados, abre espaço para que os turistas possam escolher tais destinos para obter experiências semelhantes em detrimento aos principais destinos. Essa escolha pode envolver fatores como orçamento, proximidade do destino, clima entre outros, no entanto, até então não existe uma ferramenta ou processo que informe presente ao turista esse tipo de informação.

5.6 Similaridade entre destinos

Outra variável estudada foi a similaridade entre destinos para um determinado grupo de interesse (perfil). A similaridade avaliada entre destinos com base nos perfis é interessante, porque preza pela característica de um mesmo destino oferecer múltiplas experiências, avaliando o quão parecidos são dois destinos. Os resultados apontam situações esperadas, como o fato de Londres ser o destino mais similar a Amsterdã, o

fato de Brisbane ser o destino mais similar a Auckland, a proximidade entre Blumenau e Campos do Jordão ou a proximidade entre Frankfurt e Nova Iorque. Apontam também situações inesperadas, como o fato de Curitiba ser a cidade mais similar a Toronto ou Ouro Preto ser a cidade mais similar a Bruxelas.

As aplicações práticas dessa descoberta são importantes para os turistas, porque um grupo interessado em experiências culturais pode verificar os destinos mais próximos e mais similares a uma determinada cidade e assim planejar e traçar uma rota de viagem com maior riqueza cultural. Além disso, um turista interessado em experiências que pudesse vivenciar em Bruxelas, poderia visitar Ouro Preto, por variáveis como orçamento, clima e diversas outras que possam influenciar sua decisão. Por um outro lado, ações de marketing podem ser direcionadas para públicos específicos usando as informações de similaridade para atrair novos turistas, como: Quer viver uma experiência de Bruxelas no Brasil? Visite Ouro Preto.

Em trabalhos futuros, seria interessante explorar os motivos pelo qual determinados destinos são mais similares, isso permitirá um entendimento maior sobre o tema, certamente descobrindo novas informações, as quais não foram objetivo dessa pesquisa.

5.7 Perfil turístico dos destinos mais visitados

Essa variável procurou visualizar o perfil dos destinos mais visitados pelos brasileiros. Nota-se, que a lista de destinos nacionais mais visitados entre 2011 e 2019 sofre bem mais alterações do que a lista internacional. Certamente o fator facilidade de visitação contribui para isso, no entanto, alguns pontos revelados são interessantes, como um aumento contínuo na procura de destinos nacionais com boa oferta de atrações de Gastronomia nos últimos anos. Embora, em queda nos últimos anos a nível nacional, o perfil Religioso é o que apresenta maior relevância tanto nos destinos nacionais quanto internacionais. Dos destinos nacionais mais visitados, o perfil Praia é o segundo mais relevante, enquanto na esfera internacional é Cultura. Acredita-se que isso tenha mais relação com as características dos destinos, pois os perfis Vida Noturna e Religioso aparecem na lista dos mais relevantes tanto nos destinos nacionais, quanto internacionais.

Como as avaliações são feitas por brasileiros, é possível inferir que os destinos mais visitados retratam de certa forma o perfil turístico não só dos destinos, como também a preferência dos brasileiros. Entre os cinco primeiros perfis tanto na esfera nacional quanto internacional, aparecem os perfis Religioso, Compras e Vida Noturna. A escolha de um determinado destino, envolve uma série de fatores, inclusive psicológicos conforme supracitado, no entanto, entender o comportamento e diferenças entre esses perfis, pode permitir a descoberta de conhecimento novo e relações não imaginadas anteriormente.

5.8 *Rankings de pontos de interesse e destinos por perfil*

Os rankings apresentam as atrações ou destinos turísticos mais relevantes para determinado grupo de interesse (perfil). O objetivo não foi identificar se uma atração ou destino é melhor ou pior do que outro e sim, encontrar a relevância destes para um certo perfil. Assim, para cada um dos 2.627 POIs avaliados, um grau de pertinência foi atribuído para um determinado perfil. A relevância dos destinos para os perfis foi calculada observando quantidade e qualidade de suas atrações. A partir dessa classificação, uma série de informações em forma de *rankings* se tornam possível de serem extraídas, como: os continentes mais relevantes para Cultura ou os países com melhor oferta de atrações românticas ou as atrações mais relevantes para Aventura no estado de Minas Gerais ou ainda o destino com melhor oferta de atrações para Vida Noturna.

Conforme os resultados analíticos do perfil Vida Noturna e Religioso exibidos, nos rankings apresentados, não houve nenhuma atração entre as primeiras que não fosse realmente pertinente ao perfil a se observar por sua categorização no TripAdvisor. Com relação aos destinos, a avaliação é mais complexa, porque conforme já relatado, um destino possui características para diferentes perfis. Os resultados sintéticos dos demais perfis apresentam desempenho similar. Para cada perfil, apresenta-se seu desempenho a nível de continente, país, estado e cidades. No nível de atração turística, apresenta-se quais atrações nacionais e internacionais são mais relevantes. Um comparativo com a realidade do Brasil em cada perfil é apresentado, o qual permite ver a necessidade de se

investir em atrações culturais no Brasil, devido à baixa relevância, assim como a oportunidade de explorar pontos fortes em perfis como Praia e Natureza.

O conhecimento sobre fraquezas e forças de atrações ou destinos turísticos permite que governos municipais promovam ações de fomento e melhoria de pontos fracos ou explore seus pontos fortes, segmentando conteúdo direcionado para um público específico. O setor privado também pode se beneficiar, ao passo que, conhecendo melhor as características de cada atração ou destino, poderia direcionar marketing específico para criar, adaptar ou remover atrações. Um destino turístico relatado pelos turistas com boa relevância para Vida Noturna poderia investir em novas experiências para o público do perfil ou público semelhante e assim atrair mais turistas. Do outro lado da equação está o turista, que poderia descobrir novos destinos ou atrações exatamente compatíveis com sua demanda. Ou ainda, visualizar dentro do seu estado, quais destinos oferecem a melhor experiência para quem procura por Aventura, ou quais atrações mais relevantes para diversão em Família. Independente do ator responsável pela decisão, a transformação de atrações e destinos em perfis turísticos oferece uma nova perspectiva no processo decisório sobre onde ir e o que fazer.

5.9 Limitações da pesquisa

Considera-se como limitação da pesquisa a fonte de dados utilizada. Embora seja uma fonte de dados considerável, de 3.4 milhões de avaliações escritas sobre atrações turísticas, a mesma reflete o universo de um público específico, aquele que realiza avaliações de forma online. Assim, ao identificar os perfis de uma atração ou destino e seu grau de relevância, os algoritmos podem apresentar um viés desse público, não refletindo todo o universo do local em análise.

5.10 Projetos futuros

Um objetivo futuro do autor é a criação de um sistema de recomendação de atrações turísticas ou destinos. Com base em algumas informações do turista como grupo familiar ou preferências de atividades ou clima, seria possível combinar tais dados com as características de atrações e destinos. Dessa forma, seria possível recomendar atrações

ou destinos mais compatíveis com a realidade do turista, como uma recomendação de filme do Netflix. Ou seja, um agente de turismo digital que poderia comparar 2.627 atrações e encontrar dentro delas, as mais similares de acordo com a demanda. Além disso, o turista poderia também visualizar quais destinos e atrações são mais similares, como por exemplo: Qual é o museu mais similar ao museu do Louvre no Brasil? Uma aplicação digital assim poderia facilitar o processo decisório do usuário na seleção do destino (aonde ir) e o que fazer (atrações).

6 CONSIDERAÇÕES FINAIS

Mario Quintana afirma que “*Viajar é trocar a roupa da alma*”. Quando se pensa em realização na vida, na maioria das vezes, isso está atrelado a algum tipo de viagem. Um turista diversas vezes pode deixar de visitar o destino dos sonhos, por inviabilidade financeira, distância ou uma série de fatores. Mas e se este turista pudesse obter uma experiência semelhante considerando suas características e limitações? Essa indagação apoiada pela paixão por viagens fez com o que o autor buscasse se aprofundar nesse campo de pesquisa. Além disso, por acreditar que toda pesquisa deveria ter um retorno à sociedade de forma a melhorar a vida das pessoas promovendo e disseminando conhecimento.

Atualmente, a informação é o ativo mais importante e ironicamente também o mais abundante, sendo a principal matéria-prima do conhecimento. A tecnologia tem um papel transformador e essencial ao permitir que a informação seja organizada e recuperada de forma a gerar o conhecimento efetivo. O conhecimento, por sua vez, promove a quebra de barreiras, de paradigmas, de pré-conceitos. Permite que o intransponível seja superado e é o principal fator na evolução de uma sociedade. Evoluir proporcionalmente à quantidade de informações disponíveis, não depende necessariamente da construção de um robô automático e sim primariamente de como essa informação atingirá as pessoas alterando seu repertório e conseqüentemente suas ações e decisões.

REFERÊNCIAS

- AAKER, J. L. Dimensions of brand personality. **Journal of Marketing Research**, [Chicago], v. 34, n. 3, p. 347-356, 1997. DOI: <https://doi.org/10.2307/3151897>. Disponível em: <https://www.jstor.org/stable/3151897>. Acesso em: 26 jan. 2021.
- ACUÑA, E. Preprocessing in Data Mining. *In*: LOVRIC, M. (ed). **International Encyclopedia of Statistical Science**. Berlin: Springer, 2011. DOI: https://doi.org/10.1007/978-3-642-04898-2_51. Disponível em: https://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_51. Acesso em: 25 jan. 2021.
- ALMEIDA, M. B. *et al.* A formação em Ciência da Informação no modelo do movimento i-School: o Programa de Pós-Graduação em Gestão e Organização do Conhecimento. *In*: ENCONTRO IBÉRICO ASSOCIAÇÃO DE EDUCAÇÃO E INVESTIGAÇÃO EM CIÊNCIA DA INFORMAÇÃO DA IBEROAMÉRICA E CARIBE, 8., 2017, Coimbra. **Anais [...]**. Faculdade de Letras da Universidade de Coimbra, 2017. p. 655-663. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=6598995>. Acesso em: 25 jan. 2021.
- 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, Aug. 2013. DOI: <https://doi.org/10.1002/asi.22861>. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/asi.22861>. Acesso em: 25 jan. 2021.
- ALTINEL, B.; GANIZ, M. C. Semantic text classification: a survey of past and recent advances. **Information Processing & Management**, Elmsford, v. 54, n. 6, p. 1129-1153, Nov. 2018. DOI: <https://doi.org/10.1016/j.ipm.2018.08.001>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457317305757?via%3Dihub>. Acesso em: 25 jan. 2021.
- ALVARENGA, L. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli**, Florianópolis, v. 8, n. 15, p. 18-40, 2003. DOI: <https://doi.org/10.5007/1518-2924.2003v8n15p18>. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2003v8n15p18>. Acesso em: 25 jan. 2021.
- ARENTZE, T.; KEMPERMAN, A.; AKSENOV, P. Estimating a latent-class user model for travel recommender systems. **Information Technology & Tourism**, Elmsford, v. 19, p. 61-82, Feb. 2018. DOI: <https://doi.org/10.1007/s40558-018-0105-z>. Disponível em: <https://link.springer.com/article/10.1007/s40558-018-0105-z>. Acesso em: 22 jan. 2021.
- ASKALIDIS, G.; MALTHOUSE, E. C. The value of online customer reviews. ACM CONFERENCE ON RECOMMENDER SYSTEMS, 10., 2016, Boston. **Proceedings [...]**. New York: ACM, 2016. p. 155-158. DOI: <https://doi.org/10.1145/2959100.2959181>. Disponível em: <https://dl.acm.org/doi/10.1145/2959100.2959181>. Acesso em: 25 jan. 2021.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information retrieval**. Essex: ACM Press, 1999.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information retrieval: the concepts and technology behind search**. New York: ACM Press, 2013.

BARACHO, R. M. A. **Sistema de recuperação de informação visual em desenhos técnicos de engenharia e arquitetura**: modelo conceitual, esquema de classificação e protótipo. 2007. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2007. Disponível em: <http://hdl.handle.net/1843/ECID-79CP7G>. Acesso em: 25 jan. 2021.

BERRAR, D. Cross-Validation. *In*: RANGANATHAN, S. *et al.* (ed.). **Encyclopedia of bioinformatics and computational biology**. [S. l.]: Elsevier, 2019. p. 542-545. DOI: <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012809633820349X?via%3Dihub>. Acesso em: 25 jan. 2021.

BIGOLIN, N. M.; BOGORNY, V.; ALVARES, L. O. Uma linguagem de consulta para mineração de dados em banco de dados geográficos orientado a objetos. *In*: CONFERÊNCIA LATINOAMERICANA DE INFORMÁTICA, 29., 2003, La Paz. **Anais [...]**. [S. l.: s. n.], 2003. p. 23-35.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, [Baltimore], v. 55, n. 4, Apr. 2012. DOI: <https://doi.org/10.1145/2133806.2133826>. Disponível em: <https://dl.acm.org/doi/10.1145/2133806.2133826>. Acesso em: 25 jan. 2021.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science**, Washington, v. 42, n. 5, p. 351-360, June 1991. DOI: [https://doi.org/10.1002/\(SICI\)1097-4571\(199106\)42:5<351::AID-ASI5>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3). Disponível em: <http://ppggoc.eci.ufmg.br/downloads/bibliografia/Buckland1991.pdf>. Acesso em: 25 jan. 2021.

BURKE, R.; FELFERNIG, A.; GÖKER, M. H. Recommender systems: an overview. **AI Magazine**, [Palo Alto], v. 32, n. 3, p. 13-18, June 2011. DOI: <https://doi.org/10.1609/aimag.v32i3.2361>. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/2361>. Acesso em: 25 jan. 2021.

BUSH, V. As we may think. **The Atlantic**, [Boston], July 1945. Disponível em: <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>. Acesso em: 25 jan. 2021.

CAVNAR, W. B.; TRENKLE, J. M. Ngrambased text categorization. *In*: ANNUAL SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL, 3., 1994, Las Vegas. **Proceedings [...]**. Las Vegas: Information Science Research Institute, University of Nevada, 1994. p. 161-175.

CHEN, D.; MÜLLER, H. M.; STERNBERG, P. W. Automatic document classification of biological literature. **BMC Bioinformatics**, [London], v. 7, 2006. DOI: <https://doi.org/10.1186/1471-2105-7-370>. Disponível em: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-370>. Acesso em: 25 jan. 2021.

CHEVALIER, J.; MAYZLIN, D. The Effect of word of mouth on sales: online book reviews. **Journal of Marketing Research**, [Chicago], v. 43, n. 3, p. 345-354, Aug. 2006. Disponível em: <https://www.jstor.org/stable/30162409>. Acesso em: 25 jan. 2021.

CHOO, W. C. **A organização do conhecimento**. Tradução Eliana Rocha. São Paulo: Senac, 2003.

CIOFFI, R. *et al.* Artificial Intelligence and machine learning applications in smart production: progress, trends, and directions. **Sustainability**, Basel, v. 12, n. 2, 2020. DOI: <https://doi.org/10.3390/su12020492>. Disponível em: <https://www.mdpi.com/2071-1050/12/2/492>. Acesso em: 25 jan. 2021.

CLEVERLEY, P. H.; BURNETT, S. The Best of Both Worlds: Highlighting the Synergies of Combining Manual and Automatic Knowledge Organization Methods to Improve Information Search and Discovery. **Knowledge Organization**, Wurzburg, v. 42, n. 6, p. 428-444, 2015. DOI: <https://doi.org/10.5771/0943-7444-2015-6-428>. Disponível em: <http://hdl.handle.net/10059/1364>. Acesso em: 22 jan. 2021.

COHEN, E. Towards a sociology of international tourism. **Social Research**, Baltimore, v. 39, n. 1, p. 164-182, Spring 1972. Disponível em: <https://www.jstor.org/stable/40970087>. Acesso em: 22 jan. 2021.

CORNELIUS, I. Theorizing information for information science. **Annual Review of Information Science and Technology**, [Medford], v. 36, n. 1, p. 392-425, 2002. DOI: <https://doi.org/10.1002/aris.1440360110>. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/aris.1440360110>. Acesso em: 25 jan. 2021.

COUSSEMENT, K. Improving customer retention management through cost-sensitive learning. **European Journal of Marketing**, Bradford, v. 48, n. 3/4, p.477-495, 2014. DOI: <https://doi.org/10.1108/EJM-03-2012-0180>. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/EJM-03-2012-0180/full/html>. Acesso em: 26 jan. 2021.

CREMONESI, P. *et al.* Polarized review summarization as decision making tool. *In: INTERNATIONAL WORKING CONFERENCE ON ADVANCED VISUAL INTERFACES*, 12. 2014, Como. **Proceedings** [...]. New York: ACM, 2014. p. 355-356. DOI: <https://doi.org/10.1145/2598153.2600047>. Disponível em: <https://dl.acm.org/doi/10.1145/2598153.2600047>. Acesso em: 26 jan. 2021.

CRESWELL, J. W. **Research design: qualitative, quantitative, and mixed methods approaches**. 2nd ed. Thousand Oaks: Sage Publications, 2003. 245 p.

CROMPTON, J. L. Motivations for pleasure vacation. **Annals of Tourism Research**, New York, v. 6, n. 4, p. 408-424, Oct./Dez. 1979. DOI: [https://doi.org/10.1016/0160-7383\(79\)90004-5](https://doi.org/10.1016/0160-7383(79)90004-5). Disponível em: <https://www.sciencedirect.com/science/article/pii/0160738379900045>. Acesso em: 26 jan. 2021.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *In*: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 12., 2003, [Budapeste]. **Proceedings** [...]. New York: ACM, 2003. p. 519-528. DOI: <https://doi.org/10.1145/775152.775226>. Disponível em: <https://dl.acm.org/doi/10.1145/775152.775226>. Acesso em: 26 jan. 2021.

DAVENPORT, T. H. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da informação. Tradução Bernadette Siqueira Abrão. São Paulo: Futura, 1998.

DAVENPORT, T. H.; HARRIS, J. G. **Competing in analytics**: the new science of winning. Boston: Harvard Business Review Press, 2007.

DICKMAN, S. Tough mining: the challenges of searching the scientific literature. **PLoS Biology**, San Francisco, v. 1, n. 2, p. E43, Nov. 2003. DOI: <https://doi.org/10.1371/journal.pbio.0000048>. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/14624250/>. Acesso em: 26 jan. 2021.

DODDS, L. Do data scientists spend 80% of their time cleaning data? Turns out, no? *In*: DODDS, L. **Lost Boy**: The blog of @ldodds. [S. l.], 31 Jan. 2020. Disponível em: <https://blog.ldodds.com/2020/01/31/do-data-scientists-spend-80-of-their-time-cleaning-data-turns-out-no/>. Acesso em: 26 jan. 2021.

DU, K. L.; SWAMY, M. N. S. Fundamentals of machine learning. *In*: DU, K. L.; SWAMY, M. N. S. **Neural networks and statistical learning**. London: Springer, 2019. cap. 2.

ESTEVES, G. C. **Churn prediction in the telecom business**. 2016. Dissertação (Mestrado Integrado em Engenharia Informática e Computação) – Faculdade de Engenharia, Universidade do Porto, Porto, 2016.

FANG, B. *et al.* Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics. **Tourism Management**, Guildford, v. 52, p. 498-506, Feb. 2016. DOI: <https://doi.org/10.1016/j.tourman.2015.07.018>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0261517715001715>. Acesso em: 22 jan. 2021.

FERNÁNDEZ-REYES, F. C.; HERMOSILLO-VALADEZ, J.; MONTES-y-GÓMEZ, M. A Prospect-Guided global query expansion strategy using word embeddings. **Information Processing and Management**, Elmsford, v. 54, n. 1, p. 1-13, Jan. 2018. DOI: <https://doi.org/10.1016/j.ipm.2017.09.001>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457317301140?via%3Dihub/>. Acesso em: 26 jan. 2021.

FORMAN, G. An extensive empirical study of feature selection metrics for text classification. **Journal of Machine Learning Research**, Cambridge, MA, v. 3, p. 1289-1305, 2003. Disponível em: <https://dl.acm.org/doi/10.5555/944919.944974>. Acesso em: 26 jan. 2021.

GEISSER, S. **Predictive inference**: an introduction. New York: Chapman & Hall, 1993. (Monographs on statistics and applied probability, 55).

GIBSON, H.; YIANNAKIS, A. Tourist roles needs and the lifecourse. **Annals of Tourism Research**, New York, v. 29, n. 2, p. 358-383, 2002. DOI: [https://doi.org/10.1016/S0160-7383\(01\)00037-8](https://doi.org/10.1016/S0160-7383(01)00037-8). Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0160738301000378>. Acesso em: 22 jan. 2021.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GLATZER, L.; NEIDHARDT, J.; WERTHNER, H. Automated assignment of hotel descriptions to travel behavioural patterns. *In*: INTERNATIONAL CONFERENCE IN JÖNKÖPING, 2018, Jönköping. **Proceedings** [...]. Cham: Springer, 2018. p. 409-421. DOI: https://doi.org/10.1007/978-3-319-72923-7_31. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-72923-7_31. Acesso em: 26 jan. 2021.

GOOGLE. **The 2014 traveler's road to decision**. [S. l.]: Think with Google, 2014. Disponível em: <https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/2014-travelers-road-to-decision/>. Acesso em: 26 jan. 2021.

GRETZEL, U. *et al.* Travel personality testing for destination recommendation systems. *In*: FESENMAIER, D. R.; WÖBER, K. W.; WERTHNER, H. (ed). **Destination recommendation systems: behavioural foundations and applications**. [S. l.]: CABI, 2006. DOI: <http://dx.doi.org/10.1079/9780851990231.0121>. Disponível em: <https://www.cabi.org/cabebooks/ebook/20063136637>. Acesso em: 26 jan. 2021.

GRETZEL, U.; YOO, K. H. Use and impact of online travel reviews. *In*: O'CONNOR, P.; HÖPKEN, W.; GRETZEL, U. (ed.). **Information and communication technologies in tourism 2008**. Vienna: Springer, 2008. DOI: https://doi.org/10.1007/978-3-211-77280-5_4. Disponível em: https://link.springer.com/chapter/10.1007%2F978-3-211-77280-5_4. Acesso em: 26 jan. 2021.

GUPTA, P.; HARRIS, J. How e-WOM recommendations influence product consideration and quality of choice: a motivation to process information perspective. **Journal of Business Research**, Athens, v. 63, n. 9/10, p. 1041-1049, Sept. 2010. Disponível em: <https://ideas.repec.org/a/eee/jbrese/v63yi9-10p1041-1049.html>. Acesso em: 26 jan. 2021.

GUY, I. *et al.* Extracting and ranking travel tips from user-generated reviews. *In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 26.*, 2017, Perth. **Proceedings** [...]. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2017. p. 987-996. DOI: <https://doi.org/10.1145/3038912.3052632>. Disponível em: <https://dl.acm.org/doi/abs/10.1145/3038912.3052632>. Acesso em: 22 jan. 2021.

HAMILTON-SMITH, E. Four kinds of tourism? **Annals of Tourism Research**, New York, v. 14, n. 3, p. 332-344, 1987. DOI: [https://doi.org/10.1016/0160-7383\(87\)90106-X](https://doi.org/10.1016/0160-7383(87)90106-X). Disponível em: <https://www.sciencedirect.com/science/article/pii/016073838790106X>. Acesso em: 26 jan. 2021.

HARISH, B. S.; GURU, D. S.; MANJUNATH, S. Representation and classification of text documents: a brief review. **International Journal of Computer Applications**, [Anaheim], número especial, p. 110-119, 2010. Disponível em: <https://www.ijcaonline.org/specialissues/rtippr/number2/984-107>. Acesso em: 26 jan. 2021.

HASSANI, H. *et al.* Text mining in big data analytics. **Big Data and Cognitive Computing**, Basel, v. 4, n. 1, 2020. DOI: <https://doi.org/10.3390/bdcc4010001>. Disponível em: <https://www.mdpi.com/2504-2289/4/1/1>. Acesso em: 26 jan. 2021.

HAVAN, E. Recommender system application development: with cosine similarity, rating thresholding, and other custom techniques. *In: TOWARDS Data Science. [S. l.]*, 6 Dec. 2019. Disponível em: <https://towardsdatascience.com/recommender-system-application-development-part-1-of-4-cosine-similarity-f6dbcd768e83>. Acesso em: 27 jan. 2021.

HJØRLAND, B. Knowledge Organization (KO). **Knowledge Organization**, Wurzburg, v. 43, n. 7, p. 475-484, 2016. Disponível em: https://www.ergon-verlag.de/isko_ko/downloads/ko_43_2016_6_j.pdf. Acesso em: 22 jan. 2021.

HJØRLAND, B. What is Knowledge Organization (KO)? **Knowledge Organization**, Wurzburg, v. 35, n. 2/3, p. 86-101, July 2008. Disponível em: <http://hdl.handle.net/10150/106183>. Acesso em: 26 jan. 2021.

HOCHMEISTER, M.; GRETZEL U.; WERTHNER H. Destination expertise in online travel communities. *In: CANTONI, L.; XIANG, Z. (ed.). Information and communication technologies in tourism 2013*. Berlin: Springer, 2013. p. 218-229. DOI: https://doi.org/10.1007/978-3-642-36309-2_19. Disponível em: https://link.springer.com/chapter/10.1007%2F978-3-642-36309-2_19. Acesso em: 26 jan. 2021.

HOTH, A.; NÜRNBERGER, A.; PAAß, G. A brief survey of text mining. **LDV forum**, [S. l.], v. 20, p. 19-62, 2005. Disponível em: <https://www.researchgate.net/publication/215514577>. Acesso em: 26 jan. 2021.

HU, M.; LIU, B. Mining and summarizing customer reviews. *In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2004, Seattle. Proceedings [...].* New York: ACM, 2004. p. 168-177. DOI: <https://doi.org/10.1145/1014052.1014073>. Disponível em: <https://dl.acm.org/doi/10.1145/1014052.1014073>. Acesso em: 26 jan. 2021.

IBEKWE-SANJUAN, F.; BOWKER, G. C. Implications of big data for knowledge organization. **Knowledge Organization**, Wurzburg, v. 44, n. 3, p. 187-198, 2017. Disponível em: <https://hal.archives-ouvertes.fr/hal-01489030>. Acesso em: 26 jan. 2021.

KHOSLA, G.; RAJPAL, N.; SINGH, J. Evaluation of euclidean and manhattan metrics in content based image retrieval system. **Journal of Engineering Research and Applications**, [s. l.], v. 4, n. 9, p.43-49, Sept. 2014. Disponível em: <http://www.ijera.com/pages/v4no9.html>. Acesso em: 26 jan. 2021.

KIM, S. M. *et al.* Automatically assessing review helpfulness. *In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 11., 2006., [Sydney]. Proceedings [...].* Stroudsburg: Association for Computational Linguistics, 2006. p. 423-430. Disponível em: <https://dl.acm.org/doi/10.5555/1610075.1610135>. Acesso em: 26 jan. 2021.

KNISPELIS, A. LDA Topic Models. [S. l.: s. n.], 8 July 2016. 1 vídeo (20 min). Disponível em: <https://www.youtube.com/watch?app=desktop&v=3mHy4OSyRf0>. Acesso em: 26 jan. 2021.

KORFIATIS, N.; GARCÍA-BARIOCANAL, E.; SÁNCHEZ-ALONSO, S. Evaluating content quality and helpfulness of online product reviews: the interplay of review helpfulness vs. review content. **Electronic Commerce Research and Applications**, [Amsterdam], v. 11, n. 3, p. 205-217, May/June 2012. DOI: <https://doi.org/10.1016/j.elerap.2011.10.003>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1567422311000639>. Acesso em: 26 jan. 2021.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Data preprocessing for supervised learning. **International Journal of Computer Science**, [s. l.], v. 1, n. 1, p. 111-117, 2006.

LABBE, M. **A study of personality traits of travel and cultural awareness**. 2016. Thesis (Degree in Psychology) – Honors College, University of Maine, [Orono], 2016. Disponível em: https://digitalcommons.library.umaine.edu/honors/392/?utm_source=digitalcommons.library.umaine.edu%2Fhonors%2F392&utm_medium=PDF&utm_campaign=PDFCoverPages. Acesso em: 26 jan. 2021.

LADEIRA, A. P. **Processamento de linguagem natural**: caracterização da produção científica dos pesquisadores brasileiros. 2010. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010. Disponível em: <http://hdl.handle.net/1843/ECID-8B3Q6C>. Acesso em: 26 jan. 2021.

LAPPAS, T.; CROVELLA, M. E.; TERZI, E. Selecting a characteristic set of reviews. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 18., 2012, Beijing. **Proceedings** [...]. New York: ACM, 2012. p. 832-840. DOI: <https://doi.org/10.1145/2339530.2339663>. Disponível em: <https://dl.acm.org/doi/10.1145/2339530.2339663>. Acesso em: 27 jan. 2021.

LAPPAS, T.; GUNOPULOS, D. Efficient confident search in large review corpora. *In*: JOINT EUROPEAN CONFERENCE ON MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES, 2010, Würzburg. **Proceedings** [...]. Berlin: Springer, 2010. DOI: https://doi.org/10.1007/978-3-642-15883-4_13. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-15883-4_13. Acesso em: 27 jan. 2021.

LAWTON, C. A.; KALLAI, J. Gender differences in wayfinding strategies and anxiety about wayfinding: a cross-cultural comparison. **Sex Roles**, New York, v. 47, n. 9/10, p. 389-401, 2002. DOI: <https://doi.org/10.1023/A:1021668724970>. Disponível em: <https://link.springer.com/article/10.1023/A:1021668724970>. Acesso em: 22 jan. 2021.

LEE, H. A.; LAW, R.; MURPHY, J. Helpful reviewers in Tripadvisor, an online travel community. **Journal of Travel & Tourism Marketing**, Binghamton, v. 28, n. 7, p. 675-688, Oct. 2011. DOI: <https://doi.org/10.1080/10548408.2011.611739>. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/10548408.2011.611739>. Acesso em: 27 jan. 2021.

LIKERT scale. *In*: OXFORD Reference. Oxford: Oxford University Press, ©2021. Disponível em: <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100105644>. Acesso em: 27 jan. 2021.

LIMA-MARQUES, M.; MACEDO, F. L. O. Arquitetura da informação: base para a gestão do conhecimento. *In*: TARAPANOFF, K. (org.). **Inteligência, informação e conhecimento em corporações**. Brasília, DF: IBICT: UNESCO, 2006. p. 241-255. Disponível em: <http://livroaberto.ibict.br/handle/1/465>. Acesso em: 27 jan. 2021.

LJUBESIC, N. *et al.* Comparing Measures of Semantic Similarity. *In*: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY INTERFACES, 30., 2008, Dubrovnik. **Proceedings** [...]. [Piscataway]: IEEE, 2008. p. 675-682. DOI: <https://doi.org/10.1109/ITI.2008.4588492>. Disponível em: <https://ieeexplore.ieee.org/document/4588492>. Acesso em: 27 jan. 2021.

LOURIDAS, P.; EBERT, C. Machine Learning. **IEEE Software**, Los Alamitos, v. 33, p. 110-115, 2016. DOI: <https://doi.org/10.1109/MS.2016.114>. Disponível em: <https://ieeexplore.ieee.org/document/7548905>. Acesso em: 22 jan. 2021.

LU, Y. *et al.* Exploiting social context for review quality prediction. *In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB*, 19., 2010, Raleigh. **Proceedings** [...]. New York: ACM, 2010. p. 691-670. DOI: <https://doi.org/10.1145/1772690.1772761>. Disponível em: <https://dl.acm.org/doi/10.1145/1772690.1772761>. Acesso em: 27 jan. 2021.

LYFENKO, N. D. Automatic classification of documents in a natural language: a conceptual model. **Automatic Documentation and Mathematical Linguistics**, [New York], v. 48, p. 158-166, 2014. DOI: <https://doi.org/10.3103/S0005105514030030>. Disponível em: <https://link.springer.com/article/10.3103/S0005105514030030>. Acesso em: 27 jan. 2021.

MA, F. C.; SONG, N. M.; ZHANG, Q. **IRM-KM paradigm and development of Information Science**. Wuhan: Wuhan University Press, 2008.

MALUVARU, D.; SHRIRAM, R.; SUGUMARAN, V. Big data analytics in information retrieval: promise and potential. *In: INTERNATIONAL RESEARCH FORUM CONFERENCE*, 8., 2014, Bengaluru. **Proceedings** [...]. [Bengaluru]: IRF, 2014. p. 41-46. Disponível em: https://www.digitalxplore.org/up_proc/pdf/87-140479834241-46.pdf. Acesso em: 27 jan. 2021.

MANDL, T. Artificial Intelligence for information retrieval. *In: RABUÑAL DOPICO, J.; DORADO, J.; PAZOS, A. (ed.). Encyclopedia of artificial intelligence*. Hershey: IGI Global, 2009. p. 151-156. DOI: <http://doi:10.4018/978-1-59904-849-9.ch023>. Disponível em: <https://www.igi-global.com/chapter/artificial-intelligence-information-retrieval/10240>. Acesso em: 27 jan. 2021.

MARCHIONINI, G. Information Science roles in the emerging field of data science. **Journal of Data and Information Science**, [Warsaw], v. 1, n. 2, p. 1-6, June 2017. DOI: <http://dx.doi.org/10.20309/jdis.201609>. Disponível em: <http://ir.las.ac.cn/handle/12502/8593>. Acesso em: 27 jan. 2021.

MCKENZIE, G.; ADAMS, B. A data-driven approach to exploring similarities of tourist attractions through online reviews. **Journal of Location Based Services**, Abingdon, v. 18, n. 2, p. 94-118, Aug. 2018. DOI: <https://doi.org/10.1080/17489725.2018.1493548>. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/17489725.2018.1493548>. Acesso em: 22 jan. 2021.

MEHR, H. **Artificial Intelligence for citizen services and government**. Cambridge, MA: Harvard Ash Center Technology & Democracy Fellow, 2017.

MIKOLOV, T. *et al.* Efficient estimation of word representations in vector space. *In: CORNELL UNIVERSITY. Computation and language*. [Ithaca], 2013. Disponível em: <https://arxiv.org/abs/1301.3781>. Acesso em: 27 jan. 2021.

MIT TECHNOLOGY REVIEW; GOOGLE CLOUD. **Aprendizado de máquina (ML): o novo campo de testes para a vantagem competitiva.** [S. l.], 2017. Disponível em: https://lp.google-mkto.com/rs/248-TPC-286/images/MIT_ML_Competitive_Advantage_PTBR.pdf. Acesso em: 27 jan. 2021.

MOOERS, C. N. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, [s. l.], v. 2, n. 1, p. 20-32, Jan. 1951. DOI: <https://doi.org/10.1002/asi.5090020107>. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090020107>. Acesso em: 25 jan. 2021.

MURTINHO, V. Louvre: um museu com história (parte 1). **Metálica**, Coimbra, ano 19, n. 50, p. 18-24, jun. 2018. DOI: https://doi.org/10.30779/cmm_metalica_50_01. Disponível em: <http://hdl.handle.net/10316/86999>. Acesso em: 27 jan. 2021.

MUTULA, S. Big Data Industry: Implication for the Library and Information Sciences. **African Journal of Library, Archives and Information Science**, Ibadan, v. 26, n. 2, p. 93-96, 2016. Disponível em: <https://www.ajol.info/index.php/ajlais/article/view/167425>. Acesso em: 22 jan. 2021.

NADBOR. Text classification with Word2Vec. *In*: NADBOR. **DS Lore: words about stuff.** [S. l.]: Octopress, 20 May 2016. Disponível em: <http://nadbordrozd.github.io/blog/2016/05/20/text-classification-with-word2vec/>. Acesso em: 27 jan. 2021.

NASTESKI, V. An overview of the supervised machine learning methods. **HORIZONS B**, Bitola, v. 4, p. 51-62, 2017. DOI: <https://doi.org/10.20544/HORIZONS.B.04.1.17.P05>. Disponível em: <https://uklo.edu.mk/filemanager/HORIZONTI%202017/Serija%20B%20br.%204/6.An%20overview%20of%20the%20supervised.pdf>. Acesso em: 27 jan. 2021.

NÁTHER, P. **N-gram based text categorization.** 2005. Thesis (Diploma thesis) – Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, 2005. Disponível em: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.7820&rep=rep1&type=pdf>. Acesso em: 27 jan. 2021.

NAWANGSARI, P.; KUSUMANINGRUM, R.; WIBOWO, A. Word2Vec for Indonesian sentiment analysis towards hotel reviews: an evaluation study. **Procedia Computer Science**, [Amsterdam], v. 157, p. 360-366, 2019. DOI: <https://doi.org/10.1016/j.procs.2019.08.178>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050919310968?via%3Dihub>. Acesso em: 27 jan. 2021.

NEIDHARDT, J. *et al.* Eliciting the users' unknown preferences. *In: ACM CONFERENCE ON RECOMMENDER SYSTEMS*, 8., 2014, Foster City. **Proceedings** [...]. New York: ACM, 2014. p. 309-312. DOI: <https://doi.org/10.1145/2645710.2645767>. Disponível em: <https://dl.acm.org/doi/10.1145/2645710.2645767>. Acesso em: 22 jan. 2021.

NEWMAN, I.; BENZ, C. R. **Qualitative-quantitative research methodology**: exploring the interactive continuum. Carbondale: University of Illinois Press, 1998.

NG, A. Y.; JORDAN, M. I. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. *In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS: NATURAL AND SYNTHETIC*, 14., 2001, [s. l.]. **Proceedings** [...]. Cambridge, MA: MIT Press, 2001. p. 841-848. Disponível em: <https://dl.acm.org/doi/10.5555/2980539.2980648>. Acesso em: 27 jan. 2021.

NICKEL, M. *et al.* A Review of Relational Machine Learning for Knowledge Graphs. **Proceedings of the IEEE**, [New York], v. 104, n. 1, p. 11-33, Jan. 2016. DOI: <https://doi.org/10.1109/JPROC.2015.2483592>. Disponível em: <https://dl.acm.org/doi/10.5555/2980539.2980648>. Acesso em: 27 jan. 2021.

NIELSEN COMPANY. Consumer trust in online, social and mobile advertising grows. **Nielsen Newswire**, New York, 4 Nov. 2012. Disponível em: <https://www.nielsen.com/us/en/insights/article/2012/consumer-trust-in-online-social-and-mobile-advertising-grows/>. Acesso em: 22 jan. 2021.

NIELSEN COMPANY. Global connected commerce survey. **Nielsen Newswire**, New York, 20 Jan. 2016. Disponível em: <https://www.nielsen.com/us/en/insights/report/2016/global-connected-commerce/>. Acesso em: 22 jan. 2021.

NONAKA, I.; TAKEUCHI, H. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. 2. ed. Rio de Janeiro: Campus, 1997.

OLIVEIRA, E.; BRANQUINHO FILHO, D. Automatic classification of journalistic documents on the Internet. **Transinformação**, Campinas, v. 29, n. 3, p. 245-255, Dec. 2017. DOI: <https://doi.org/10.1590/2318-08892017000300003>. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862017000300245&lng=en&nrm=iso. Acesso em: 27 jan. 2021.

OLIVEIRA, R. A. **Extração de dados web como suporte na elaboração de indicadores do turismo de Minas Gerais**: uma iniciativa em Big Data. 2017. Dissertação (Mestrado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2017. Disponível em: <http://hdl.handle.net/1843/ECIP-AN2PRB>. Acesso em: 27 jan. 2021.

ORTEGA, E. D. The internet effects on tourism industry. **SSRN**, [Amsterdan?], 12 May 2009. DOI: <http://dx.doi.org/10.2139/ssrn.1403087>. Disponível em: <https://ssrn.com/abstract=1403087>. Acesso em: 27 jan. 2021.

PAUL, S. K. *et al.* An Information Retrieval (IR) Techniques for text Mining on web for Unstructured data. **International Journal of Advanced Research in Computer Science and Software Engineering**, [Jaunpur], v. 4, n. 2, p. 67-70, Mar. 2014. Disponível em: https://www.researchgate.net/publication/289521268_An_Information_RetrievalIR_Techniques_for_text_Mining_on_web_for_Unstructured_data. Acesso em: 22 jan. 2021.

PERONE, C. S. Machine learning: cosine similarity for vector space models (part III). *In*: PERONE, C. S. **Blog Terra Incognita**. [S. l.], 22 nov. 2020a. Disponível em: <https://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/>. Acesso em: 26 jan. 2021.

PERONE, C. S. Talk: gradient-based optimization for deep learning. *In*: PERONE, C. S. **Blog Terra Incognita**. [S. l.], 22 nov. 2020b. Disponível em: <https://blog.christianperone.com/2020/11/optimization-deep-learning/>. Acesso em: 26 jan. 2021.

QAISER, S.; ALI, R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. **International Journal of Computer Applications**, New York, v. 181, n. 1, p. 25-29, 2018. DOI: <https://doi.org/10.5120/ijca2018917395>. Disponível em: <https://www.ijcaonline.org/archives/volume181/number1/29681-2018917395>. Acesso em: 22 jan. 2021.

RAMAGE, D. *et al.* Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2009, Singapore. **Proceedings** [...]. Stroudsburg: Association for Computational Linguistics, 2009. p. 248-256. Disponível em: <https://dl.acm.org/doi/10.5555/1699510.1699543>. Acesso em: 26 jan. 2021.

ROY, D. *et al.* Using word embeddings for automatic query expansion. *In*: SIGIR WORKSHOP ON NEURAL INFORMATION RETRIEVAL, 2016, Pisa. **Proceedings** [...]. [S. l.: s. n.], 2016. Disponível em: <https://arxiv.org/abs/1606.07608>. Acesso em: 26 jan. 2021.

SALLES, T. *et al.* A Two-stage machine learning approach for temporally-robust text classification. **Information Systems**, Elmsford, v. 69, p. 40-58, Sept. 2017. DOI: <https://doi.org/10.1016/j.is.2017.04.004>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306437917301801?via%3Dihub>. Acesso em: 27 jan. 2021.

SALLES, T. *et al.* Automatic document classification temporally robust. **Journal of Information and Data Management**, [Belo Horizonte], v. 1, n. 2, p. 199-211, Sept. 2010. Disponível em: <https://periodicos.ufmg.br/index.php/jidm/article/view/41>. Acesso em: 27 jan. 2021.

SANTOS, J. L. G. *et al.* Integração entre dados quantitativos e qualitativos em uma pesquisa de métodos mistos. **Texto contexto: enfermagem**, Florianópolis, v. 26, n. 3, p. e1590016, 2017. DOI: <https://doi.org/10.1590/0104-07072017001590016>. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-07072017000300330&lng=pt&nrm=iso. Acesso em: 27 jan. 2021.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235>. Acesso em: 27 jan. 2021.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**, Brasília, DF, v. 24, n. 1, 1995. Disponível em: <http://revista.ibict.br/ciinf/article/view/608>. Acesso em: 27 jan. 2021.

SEADLE, M. The new mission of a new i-school. **Library Hi Tech**, [Ann Arbor], v. 25, n. 1, p. 5-9, 2007. DOI: <https://doi.org/10.1108/07378830710735803>. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/07378830710735803/full/html>. Acesso em: 27 jan. 2021.

SEN, S.; LERMAN, D. Why are you telling me this? An examination into negative consumer reviews on the web. **Journal of Interactive Marketing**, [New York], v. 21, n. 4, p. 76-94, 2007. DOI: <https://doi.org/10.1002/dir.20090>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1094996807700397>. Acesso em: 22 jan. 2021.

SERTKAN, M.; NEIDHARDT, J.; WERTHNER, H. What is the “Personality” of a tourism destination? **Information Technology & Tourism**, Elmsford, v. 21, p. 105-133, 2018. DOI: <https://doi.org/10.1007/s40558-018-0135-6>. Disponível em: <https://link.springer.com/article/10.1007/s40558-018-0135-6>. Acesso em: 22 jan. 2021.

SHIEH, G. S. A weighted Kendall's tau statistic. **Statistics & Probability Letters**, Amsterdam, v. 39, n. 1, p. 17-24, July 1998. DOI: [https://doi.org/10.1016/S0167-7152\(98\)00006-6](https://doi.org/10.1016/S0167-7152(98)00006-6). Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167715298000066>. Acesso em: 27 jan. 2021.

SHIN, S. *et al.* How far, how near psychological distance matters in online travel reviews: a test of construal-level theory. *In*: INVERSINI, A.; SCHEGG, R. (ed.). **Information and communication technologies in tourism 2016**. Cham: Springer, 2016. p. 355-368. DOI: https://doi.org/10.1007/978-3-319-28231-2_26. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-28231-2_26. Acesso em: 22 jan. 2021.

SHIN, S. H. *et al.* Conceptual foundations of a landmark personality scale based on a destination personality scale: Text mining of online reviews. **Information Systems Frontiers**, Boston, v. 19, p. 743-752, 2017. DOI: <https://doi.org/10.1007/s10796-016-9725-z>. Disponível em: <https://link.springer.com/article/10.1007/s10796-016-9725-z>. Acesso em: 22 jan. 2021.

SIDOROV, G. *et al.* Soft similarity and soft cosine measure: similarity of features in vector space model. **Computacion y Sistemas**, [Ciudad de México], v. 18, n. 3, p. 491-504, 2014. DOI: <https://doi.org/10.13053/CyS-18-3-2043>. Disponível em: <http://www.scielo.org.mx/pdf/cys/v18n3/v18n3a7.pdf>. Acesso em: 27 jan. 2021.

SMIRAGLIA, R. P.; CAI, X. Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization. **Knowledge Organization**, Wurzburg, v. 44, n. 3, p. 215-233, 2017. DOI: <https://doi.org/10.5771/0943-7444-2017-3-215>. Disponível em: <https://www.nomos-elibrary.de/10.5771/0943-7444-2017-3-215/tracking-the-evolution-of-clustering-machine-learning-automatic-indexing-and-automatic-classification-in-knowledge-organization-volume-44-2017-issue-3>. Acesso em: 27 jan. 2021.

SMITH, C. 35 TripAdvisor statistics and facts (2020): by the numbers. *In*: DMR: Business Statistics. [S. l.], 11 July 2020. Disponível em: <https://expandedramblings.com/index.php/tripadvisor-statistics/>. Acesso em: 26 jan. 2021.

SOERGEL, D. **Knowledge Organization Systems**: overview. [S. l.], 2008. Disponível em: <http://www.dsoergel.com/SoergelKOSOverview.pdf>. Acesso em: 27 jan. 2021.

SONG, F.; LIU, S.; YANG, J. A comparative study on text representation schemes in text categorization. **Pattern Analysis and Applications**, [London], v. 8, p. 199-209, 2005. DOI: <https://doi.org/10.1007/s10044-005-0256-3>. Disponível em: <https://link.springer.com/article/10.1007/s10044-005-0256-3>. Acesso em: 27 jan. 2021.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 2, p. 161-173, maio/ago. 2006. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/320>. Acesso em: 27 jan. 2021.

SOUZA, R. R. **The role of machine learning for knowledge organization**. [S. l.], 2020. 102 slides.

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, DF, v. 42, n. 2, p.159-173, 2013. Disponível em: <http://revista.ibict.br/ciinf/article/view/1379/1557>. Acesso em: 27 jan. 2021.

SPARKS, B. A.; PERKINS, H.; BUCKLEY, R. Online travel reviews as persuasive communication: the effects of content type, source, and certification logos on consumer behavior. **Tourism Management**, Guildford, v. 39, p. 1-9, 2013. DOI: <https://doi.org/10.1016/j.tourman.2013.03.007>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0261517713000861>. Acesso em: 22 jan. 2021.

TOMÉ, L. M. Panorama do turismo no Brasil e oportunidades para a região nordeste. **Caderno Setorial ETENE**, [s. l.], ano 3, n. 59, p. 1-11, dez. 2018. Disponível em: https://www.bnb.gov.br/documents/80223/4296541/59_turismo.pdf/abf03c9a-5d73-3b1e-42d4-7cbfcfa47fd7. Acesso em: 26 jan. 2021.

TURING, A. M. Computing machinery and intelligence. **Mind**, Oxford, v. 59, n. 236, p. 433-460, Oct. 1950. DOI: <https://doi.org/10.1093/mind/LIX.236.433>. Disponível em: <https://academic.oup.com/mind/article/LIX/236/433/986238>. Acesso em: 25 jan. 2021.

VERMEULEN, I. E.; SEEGER, D. Tried and tested: the impact of online hotel reviews on consumer consideration. **Tourism Management**, Guildford, v. 30, n. 1, p. 123-127, Feb. 2009. DOI: <https://doi.org/10.1016/j.tourman.2008.04.008>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0261517708000824>. Acesso em: 27 jan. 2021.

VICTORINO, M. C.; BRÄSCHER, M. Organização da informação e do conhecimento, engenharia de software e arquitetura orientada a serviços: uma abordagem holística para o desenvolvimento de sistemas de informação computadorizados. **DataGramZero**, [Rio de Janeiro], v. 10, n. 3, 2009. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/6517>. Acesso em: 10 jan. 2021.

WANG, L. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, London, v. 74, n. 6, p.1243-1257, Oct. 2018. DOI: <https://doi.org/10.1108/JD-02-2018-0036>. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/JD-02-2018-0036/full/html>. Acesso em: 27 jan. 2021.

WERTHNER, H.; RICCI, F. E-Commerce and tourism. **Communications of the ACM**, [New York], v. 47, n. 12, p. 101-105, Dec. 2004. DOI: <https://doi.org/10.1145/1035134.1035141>. Disponível em: <https://cacm.acm.org/magazines/2004/12/6347-e-commerce-and-tourism/fulltext>. Acesso em: 27 jan. 2021.

WORLD TRAVEL & TOURISM COUNCIL. **Travel & tourism economic impact 2019 world**. London, 2019.

WORLD TRAVEL & TOURISM COUNCIL. **Travel & tourism: economic impact 2020**. London, 2020. Disponível em: <https://wtcc.org/Research/Economic-Impact>. Acesso em: 27 jan. 2021.

YANG, L.; HANNEKE, S.; CARBONELL, J. A theory of transfer learning with applications to active learning. **Machine Learning**, [Boston], v. 90, p. 161-189, 2013. DOI: <https://doi.org/10.1007/s10994-012-5310-y>. Disponível em: <https://link.springer.com/article/10.1007/s10994-012-5310-y>. Acesso em: 27 jan. 2021.

ZHAN, M.; WIDÉN, G. Understanding big data in librarianship. **Journal of Librarianship and Information Science**, London, v. 51, n. 2, p. 561-576, 2017. DOI: <https://doi.org/10.1177/0961000617742451>. Disponível em: <https://journals.sagepub.com/doi/10.1177/0961000617742451#articleCitationDownloadContainer>. Acesso em: 27 jan. 2021.