INTEGRAÇÃO DE INFORMAÇÕES DE USUÁRIOS PARA INFERÊNCIA DE LOCALIZAÇÃO GEOGRÁFICA NO TWITTER

SÍLVIO SOARES RIBEIRO JÚNIOR

INTEGRAÇÃO DE INFORMAÇÕES DE USUÁRIOS PARA INFERÊNCIA DE LOCALIZAÇÃO GEOGRÁFICA NO TWITTER

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Gisele Lobo Pappa

Belo Horizonte

Agosto de 2016

Soares Ribeiro Júnior, Sílvio

Integração de informações de usuários para inferência de localização geográfica no Twitter / Sílvio Soares Ribeiro Júnior. — Belo Horizonte, 2016
xx, 91 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientadora: Gisele Lobo Pappa

1. . I. Título.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Integração de informações de usuários para inferência de localização geográfica no twitter

SILVIO SOARES RIBEIRO JUNIOR

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. GISELE LOBO PAPPA - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 29 de junho de 2015.

Agradecimentos

Ser quem eu sou hoje e finalizar esse processo não seria possível sem a ajuda, companhia e inspiração vinda de várias pessoas.

Obrigado, primeiramente, a toda minha família e em especial aos meus pais Sílvio Soares Silva e Cleide Antônia Ribeiro que foram a minha primeira escola, porto seguro e que sempre me apoiaram nas minhas decisões e me incentivaram a seguir meus sonhos. Obrigado a minha irmã Vitória Aparecida Soares Ribeiro pelo seu apoio incondicional, conversas e pequenos conselhos.

Foram várias as pessoas que me ensinaram lições valiosas durante esse processo. Obrigado à minha orientadora Gisele Pappa, que me incentivou a iniciar essa jornada, pela sua paciência, orientação, conselhos e amizade. Obrigado ao meu companheiro de Observatório Walter Santos que me ensinou muito mais do que eu poderia ter aprendido sozinho. Obrigado ao meu co-piloto de mestrado e para-sempre-hóspede, Alessandro Sena, pela sua empolgação inspiradora mesmo nos momentos mais difíceis (e também pelos mojitos). Obrigado à minha amiga Bruna Neuenschwander, que mesmo com todos os motivos para ficar de saco cheio com minhas divagações ainda me olha com doçura quando me reencontra. Obrigado ao meu amigo Alexandre Davis que me ajudou a caminhar por caminhos desconhecidos por ambos e a superar o mi-mi-mi. Obrigado ao meu amigo Filipe De Lima Arcanjo, que me inspira com sua disciplina e desejo de ir um pouco mais além. Contem sempre comigo.

E obrigado aos meus amigoles que me acompanham desde a minha graduação e que espero que me acompanhem até o asilo: Thiago Brito, por encarnar o espírito da espontaneidade e me ensinar lições valiosas sobre a vida, ser meu consultor e professor particular; Paula Franciolli, por me ouvir tantas vezes, me proporcionar os melhores almoços e ser esse exemplo de perseverança; Vinícius Luiz, por me dar a oportunidade de ser amigo de uma pessoa que não canso de admirar pela lucidez, humor e sensibilidade; Brenno Pereira, que nunca hesitou em desviar do seu caminho para me alegrar e sem o qual a vida teria menos cores (e brilho). Presentes de Deus.

Um lar não é feito somente de paredes. Obrigado aos meus companheiros de

noites intermináveis de conversas que me ajudaram a conviver com as mais diversas intempéries: Aline Hoehne, que esbanja carinho onde passa; Rafael Bicalho, que sempre me espanta com sua criatividade; Marcos Paulo, guru dos assuntos sentimentais. Iuri e Saulo, dupla dinâmica e fonte inesgotável de energia.

A vida às vezes pode ser muito irônica e difícil, mas sempre melhor quando não estamos sozinhos. Obrigado Matheus P. de Oliveira pelo carinho, suporte e paciência nos meus momentos de loucura para acabar essa dissertação.

Nesses últimos anos, encontrei amigos que espero levar para vida inteira. Obrigado aos meus amigos Brunah Schall, Davi Fraga, Suzane de Sá, Juliana Braga que foram colocados no meu caminho por acaso mas continuam presentes. Obrigado pela companhia e amizade dos meus colegas do Departamento de Computação Mariane Moreira, Denise Eb, Paulo Bicalho, Mariane Moreira, Alex de Sá, Bruno Coutinho, Camila Araújo, Diogo Rennó, Elverton Fazzion, Fernando Carvalho, Gabriel Poesia, Hélio Almeida, Júlio Albinati, Luam Totti, Luiz Oliveira, Natália Tereza, Osvaldo Fonseca, Pedro Calais e Samuel Sérvulo. Obrigado aos meus colegas de boates e boatos Paulo Borges, Rafael Moreira, Marcelo Dias, Fábio Gomes, Alan Albuquerque, Jeferson Baeta, Pedro Nogueira, Leonardo Silva, Frederico Paixão, Daniel Nogueira, Marcus Martins.

Resumo

A localização geográfica de um usuário de redes sociais é essencial para várias aplicações que utilizam informações provenientes dessas redes, como detecção de eventos e epidemias.

Neste trabalho, investigamos a tarefa de inferir a localização de um usuário do Twitter, reconhecendo que além da marcação do GPS presente em algumas mensagens, a localização do usuário pode estar presente implícita ou explicitamente em outros atributos como seus tweets, campos do perfil e rede de usuários com os quais interage. Ao mesmo tempo que trabalhos anteriores exploraram as possibilidades de utilizar essas diversas fontes de informação, não existe uma técnica que de fato seja um consenso, mas sim a oferta de várias técnicas com vantagens e desvantagens próprias.

Inicialmente discutimos como contornar a esparsidade de usuários em certas regiões - o que pode dificultar a tarefa de inferência pela escassez de dados disponíveis. Uma forma de superar esse desafio é agrupando localizações de forma que o número de usuários seja suficiente para caracterizá-las, embora assim estejamos abrindo mão da precisão. Propomos uma métrica para avaliação da qualidade desses agrupamentos e apresentamos uma nova estratégia para agrupar cidades.

Trabalhos anteriores utilizam diversas métricas para análise da eficácia dos métodos de geoinferência, tornando difícil a comparação direta dos mesmos. Utilizamos as mesmas métricas em uma base de dados única e comparamos os métodos que utilizam a rede social do usuário, suas publicações e informações do perfil providos voluntariamente pelo mesmo.

Avaliamos diferentes métodos existentes de predição de localização baseadas nas redes de usuários e, pela primeira vez, comparamos seus resultados quando aplicados a dois tipos diferentes de rede: a de amizades e de menções do usuário. Buscamos, dessa forma, evidenciar suas diferenças e semelhanças ao serem utilizadas na tarefa de inferir a localização do usuário.

Após evidenciar a mudança no uso do vocabulário das publicações dos usuários com o passar do tempo, propomos um novo modelo de inferência de localização baseado

no texto, que leva em conta o tempo da publicação de uma mensagem, e comparamos com o modelo atemporal existente na literatura, mostrando seus benefícios.

Consideramos a utilização dos diferentes campos do perfil usuário, bem como a atualização dos modelos de inferência ao longo do tempo utilizados para prever a localização de novos usuários.

Por fim, utilizando-nos das informações resultantes de cada método, avaliamos e mostramos os benefícios da combinação dos resultados dos algoritmos estudados, inferindo a localização de 99% dos usuários e geolocalizando 78% desses em um raio de até 100km de distância da sua localização real.

Palavras-chave: Geolocalização, redes sociais, Twitter, Classificação.

Abstract

Knowing the location of a social-network user is essential to many applications that rely on information extracted from these networks, including event detection andtracking epidemics.

In this work, we investigate the task of inferring the location of Twitter users by recognizing that it may be encoded not only in their GPS-tagged messages, but also in tweets content, profile information and the network of users they interact with. While previous works have exploited the possibility of using these different sources of information, there is not a *de facto* technique that is a consensus as the best, but a range of many techniques with specific pros and cons.

The sparsity of users in a geographic area may be a challenge to many inference methods due to the lack of data to characterize it. Therefore we discuss how to overcome the sparsity of users in many geographic regions. Grouping and merging locations is one of the main approaches to solve this challenge, although by doing it we are giving up the precision of the inferred location. We propose a metric to evaluate the quality of generated clusters and present a new strategy to group cities.

Previous works have used different performance metrics to evaluate their geo-inferencing methods, making it difficult to compare them directly. We use standardized metrics in the same dataset, comparing methods that use users' social network, their publications and the information provided voluntarilly in their profiles.

We evaluate different network-based methods and compare their performance in two different types of users' network: the friendship and the mentions network. We make evident the differences and similarities in using those two networks in the geo-inference task.

After pointing out changes in vocabulary usage in the stream of publication along time, we propose a new method of infering the a user's location based not only in the text but also considering the publication date of her/his messages. We achieve better precision than the non-temporal similar approach from the literature.

We also evaluate the performance of using the profile fields provided by the users

and tested periodic updates of the models created to infer the location of new users.

Finally, we evaluate the benefits of combining the information output by each method to produce a more robust geo-inference model. By combining the base models, we achieved 99% recall and geolocated 78% of the users within 100km of distance from their real location.

Keywords: Geo-inference, social networks, Twitter, Classification.

Lista de Figuras

3.1 Distribuição dos tweets das bases Geo e NotGeo por mês 16
3.2 Número de usuários por quantidade de tweets postados 16
3.3 Para a maior parte dos usuários, a sua cidade principal aparece em mais de 90% dos seus tweets 17
3.4 As 30 cidades brasileiras com maior número de usuários geolocalizados . . 18
3.5 As 30 cidades brasileiras com maior número de habitantes 18
3.6 Comportamento dos usuários de acordo com seu padrão de movimentação 19
3.7 Função de distribuição acumulada para número de amigos, seguidos e seguidores 20
3.8 Dinâmica da rede de seguidos/seguidores no Twitter 20
3.9 A distância média entre um usuário e seus amigos é menor que 1,000km para 63% dos usuários 21
3.10 Função de distribuição acumulada da distância por tipo de conexão 21
3.11 A distribuição de conexões por usuário pouco muda com a alteração da janela de tempo 23
3.12 A distribuição da distância das conexões entre usuários pouco muda com a alteração da janela de tempo 24
3.13 Número de diferentes descrições, localizações declaradas e fusos horários dos usuários 25
3.14 Evolução do vocabulário considerando as bases de Tweets geolocalizados e não-geolocalizados 26
3.15 O número de palavras disponíveis muda em cada mês, independente da base utilizada 27
3.16 Perplexidade do modelo de unigrama criado em Jan-2013 aplicado nos meses seguintes 28
3.17 Evolução do conjunto de termos com maior χ^2 em cada mês 30

4.1 Locais de referência são citados mais frequentemente em tweets na própria cidade 35
4.2 Expressões idiomáticas podem ser utilizadas para identificar a localização do usuário 35
4.3 Tweets com informações geográficas 37
4.5 Exemplo de MDT: a decisão entre o classificador KNN e Naive Bayes é feita levando em consideração a probabilidade atribuída à classe por cada um . 44
4.6 Distribuição de usuários e habitantes por cidade 45
4.7 Distribuição de usuários e habitantes por cidade 46
4.8 As 500 cidades com maior número de usuários geolocalizados abriga mais de 80% dos mesmos 46
4.9 Comparação dos vários conjuntos de agrupamentos gerados 50
4.10 Comparação dos vários conjuntos de agrupamentos gerados 51
4.11 Comparação dos vários conjuntos de agrupamentos gerados 52
4.12 Comparação dos agrupamentos gerados pelas duas técnicas apresentadas . 54

5.1 A revocação de todos os métodos atinge seu máximo e se estabiliza a partir do tamanho de 120 dias 60
5.2 A ASC-g é a mesma para os dois tipos de agrupamento 61
5.3 Avaliação da ASC-g para classificação utilizando Naive Bayes e tweets geolocalizados. 63
5.4 Avaliação da ASC-c para classificação utilizando Naive Bayes e tweets geolocalizados. 64
5.5 Avaliação da ASC-g para classificação utilizando Naive Bayes e tweets geolocalizados e não-geolocalizados. 64
5.6 Avaliação da ASC-c para classificação utilizando Naive Bayes e tweets geolocalizados e não-geolocalizados. 65
5.7 Avaliação da distância entre o agrupamento predito e o agrupamento do usuário 67
5.8 Avaliação da distância entre o a cidade do usuário e o centro do agrupamento inferido 67
5.9 Avaliação quanto a distância entre o agrupamento inferido e o agrupamento do usuário 68
5.10 Avaliação quanto a distância entre a cidade do usuário e o centro do agrupamento inferido 69

5.11 Variação de **ASC-c** de acordo com o número de conexões geolocalizadas do usuário e a distância média entre a localização inferida para o usuário e as localizações dos usuários com os quais conecta 76
5.12 Variação da ASC-c de acordo com a probabilidade média dos algoritmos e o número de períodos em que os algoritmos classificaram tweets para o usuário 77
5.13 Para probabilidades abaixo de 0.6, o algoritmo de Naive Bayes apresenta bruscas variações quanto ao ASC-c para o campo de localização 77

Lista de Tabelas

3.1 Tabela de contingência para palavra e co-ocorrência na cidade c 29
3.2 Palavras mais descritivas em Agosto de 2013 e palavras que sempre se mantiveram como descritivas durante o período 31
3.3 Concordância entre o local mais frequente entre os tweets de cada usuário por mês e mais frequente durante o ano 31

5.1 Avaliação dos métodos sobre a rede de menções 61
5.2 Avaliação dos métodos de inferência aplicados na rede de amizades 62
5.3 O GroupMinUsers apresenta melhores resultados para ASC-g que o GroupMinDistUsers, quando Naive Bayes é aplicado a base de somente tweets geolocalizados 65
5.4 Os métodos que utilizam partição de tempo possuem resultados melhores para **ASC-g, ASC-c, Acc@100 e Mediana** do que o método sem partição de tempo. No entanto, esse último apresenta maior **revocação** 66
5.5 Os resultados não são sensíveis quanto janela de atualização de tempo. Os resultados são apresentados com os melhores conjuntos de atributos para cada partição de tempo. 68
5.6 A atualização do campo de descrição periodicamente não afeta a inferência 69
5.7 Casamento exato de nomes de cidade com os campos de descrição e localização 70
5.8 Comparação da melhor configuração de cada método. Mediana dada em km. 71
5.9 A diagonal inferior mostra a concordância segundo Kappa de Cohen, enquanto a diagonal superior mostra a porcentagem dos usuários cujas localizações puderam ser inferidas pelos dois métodos. 73
5.10 Atributos ordinários fornecidos para meta árvore de decisão 78
5.11 Combinando todos os métodos de inferência, 99% da base é coberta sem perda na precisão 79
5.12 Pesos dos votos em **Votos-GA**. Somente os pesos diferentes de zero são mostrados 80

5.13 Pesos dos votos em **Votos-GA S/Amigos**, sem a rede de amigos. Somente os pesos diferentes de zero 80
5.14 Fração de usuários cujas localizações foram inferidas por cada método de acordo com a meta árvore de decisão 81

Sumário

Agradecimentos **vii**

Resumo **ix**

Abstract **xi**

Lista de Figuras **xiii**

Lista de Tabelas **xvii**

1 Introdução **1**

1.1 Objetivos 3

1.2 Contribuições 3

1.3 Organização do trabalho 4

2 Trabalhos Relacionados **5**

2.1 Métodos para inferência da localização de usuários 6

2.1.1 Inferência da localização de usuários a partir das informações provenientes dos perfis 6

2.1.2 Inferência de localização a partir do texto dos tweets 7

2.1.3 Inferência da localização de usuários a partir das suas relações . 9

2.1.4 Técnicas híbridas para inferência da localização do usuário . . . 11

3 Caracterização da base de dados **13**

3.1 Coleta 13

3.1.1 Processo de coleta 15

3.2 A base em números 16

3.3 Rede de usuários 19

3.4 Atributos do perfil do usuário 22

3.5 Texto dos tweets 25

3.6 *Ground truth* 30

4 Métodos para inferência da localização do usuário 33
4.1 Inferência a partir das relações entre usuários 33
4.2 Inferência textual 34
4.2.1 Modelo linguístico estático e Modelo linguístico temporal 36
4.2.2 Criação de modelos linguísticos: seleção de atributos e treino . . 38
4.3 Utilizando atributos do perfil do usuário 41
4.4 Combinação de modelos 43
4.5 Desafio: escassez de dados 45

5 Experimentos e resultados 55
5.1 Métricas de sucesso 55
5.2 Experimentos 58
5.2.1 Inferência através da rede de relacionamentos 59
5.2.2 Rede de menções 59
5.2.3 Rede de amizades 62
5.2.4 Inferência a partir do texto dos tweets 62
5.2.5 Inferência a partir dos atributos 66
5.3 Análise comparativa dos métodos de inferência 70
5.4 Combinação de resultados 74

6 Conclusão e trabalhos futuros 83

Referências Bibliográficas 87

Capítulo 1

Introdução

Serviços de microblog como o Twitter permitem que pesquisadores, anunciantes, ativistas, governantes, entre outros interessados na análise de padrões sociais, tenham acesso a uma quantidade sem precedentes de informação compartilhada por usuários de forma voluntária na Web. O Twitter tem se tornado uma das principais ferramentas de comunicação da atualidade, com seu caráter híbrido de *news media* e rede social (Kwak et al. [2010]). Com aproximadamente meio bilhão de tweets gerados diariamente[1], o serviço de microblog transformou-se em uma base de dados sociais que tem mostrado o seu potencial em cobrir desastres naturais, como terremotos (Sakaki et al. [2010]), detectar eventos (Weng & Lee [2011]) ou monitorar a evolução de epidemias como a dengue (Gomide et al. [2011]).

Em muitos contextos é importante descobrir não somente o que foi dito, mas de onde foi dito. Conhecer a região de onde o usuário interage pode ajudar a reconhecer padrões de comportamento de usuários dessa região nas mídias sociais. Dessa forma, a localização do usuário é de fundamental importância para melhorar o desempenho de aplicações que dependem desse tipo de informação, seja para detectar epidemias, alertar usuários, autoridades e agências de mídia sobre desastres naturais ou recomendar conteúdo/produtos com base na localização do usuário.

Definimos a **localização do usuário** como aquela onde ele reside e passa a maior parte do seu tempo. A localização do usuário nem sempre coincide com a localização de todos os seus tweets - uma vez que o mesmo pode se movimentar de uma região para outra ocasionalmente. Mas, uma vez que partimos do pressuposto de que o usuário publica a maior parte dos seus tweets a partir da cidade onde reside, uma forma de lhe atribuir uma localização é levando em conta o local mais frequente dentre seus tweets contendo *tags* de geolocalização - ou **tweets geolocalizados**. Em tais tweets,

[1] https://about.twitter.com/company

a informação da localização no momento da sua publicação está contida em forma de coordenadas geográficas. Tweets geolocalizados são fáceis de manipular usando algoritmos, uma vez que contêm informação estruturada e fácil de ser interpretada por computadores. No entanto, somente uma pequena porção de usuários publica sua localização anexada ao seus tweets. Cerca de 0.7% dos tweets coletados por Graham et al. [2013] possuíam informação de geolocalização. É improvável, portanto, como supõem Graham et al. [2013], que essa amostra seja representativa do universo do conteúdo produzido: a divisão entre usuários que usam geolocalização ou não é quase certamente enviesada por fatores como classe socioeconômica, localização e educação.

Determinar a localização de um usuário do Twitter quando o mesmo não possui tweets geolocalizados é um desafio importante a ser vencido a fim de aumentar o volume de tweets disponíveis para aplicações que necessitam de informações geográficas sobre o usuário. Uma das formas mais simples de se inferir a localização do usuário é a partir da **localização declarada**, que é fornecida pelo mesmo ao registrar seu perfil na rede social. Esse campo, no entanto, permite a entrada de texto livre. É difícil esperar, portanto, que o usuário forneça sua localização correta e sem ambiguidades - como 'Belo Horizonte, Minas Gerais, Brasil' e, não raro, encontra-se o campo preenchido de forma jocosa, como demonstrado por Hecht et al. [2011].

Além das *tags* geográficas e localização declarada, outros autores trabalharam inferindo a localização do usuário através da sua rede de amizades [Davis Jr et al., 2011; Crandall et al., 2010], conteúdo publicado [Kinsella et al., 2011; Cheng et al., 2010] e outros atributos como descrição, fuso horário e nome do usuário [Schulz et al., 2013; Han et al., 2014]. Não existe, no entanto, um método para atribuição de localização que apresente, isoladamente, os melhores resultados tanto em precisão quanto em revocação. Cada método sofre de suas limitações: os métodos que utilizam o texto dos tweets se baseiam em modelos linguísticos para cada cidade. Esses modelos podem ser criados manualmente, mas com esforço e, quase sempre, com baixa revocação na inferência. Por outro lado, modelos linguísticos podem ser obtidos a partir de tweets publicados por usuários cujas localizações são conhecidas. No entanto, muitas cidades não possuem tweets suficientes para viabilizar a criação desses modelos. Métodos que utilizam a rede de amizades ou menções também sofrem com a escassez de usuários em determinadas cidades, quando as conexões dos usuários são distantes do usuário que se pretende localizar.

Este trabalho baseia-se nas seguintes hipóteses:

1. A informação sobre a localização do usuário pode estar implícita ou explicitamente contida no texto dos seus tweets e tem caráter temporal, isto é, as evidên-

cias mudam com o passar do tempo.

2. As informações preenchidas pelo usuário na sua conta, como a descrição do perfil e localização declarada, podem ser utilizadas para inferir a localização do mesmo.

3. As diferentes redes formadas entre usuários do Twitter podem ser utilizadas como indicadores da localização dos mesmos, com resultados potencialmente diferentes. Dentre as redes mais utilizadas na literatura, temos a **rede de menções** e **rede de amizades**, explicadas na seção 3.3.

4. A combinação de diferentes métodos de inferência de localização é uma forma mais robusta de prever a localização dos usuários do que cada método considerado isoladamente.

1.1 Objetivos

Para validar as hipóteses levantadas na seção anterior, com o objetivo de inferir corretamente a localização de um grande número de usuários, este trabalho visa responder às seguintes perguntas:

1. Como comparar os diversos métodos de inferência de localização dos usuários?

2. É possível melhorar os métodos atuais de inferência de localização textual através de um modelo linguístico de caráter temporal para diversas regiões ou cidades?

3. Como tirar vantagem dos dados declarados pelo usuário para inferir sua localização?

4. A forma como os usuários se relacionam através do Twitter tem alguma correspondência com suas localizações? Se sim, como derivar a partir daí um modelo para inferir a localização de usuários que não possuem tweets geolocalizados?

5. Combinando a localização inferida através de diversos atributos (texto dos tweets, localização declarada, descrição do perfil do usuário etc), é possível inferir a localização de um número maior de usuários de forma mais precisa do que utilizando cada método de inferência de localização isoladamente?

1.2 Contribuições

A partir dos objetivos descritos na Seção anterior, esta dissertação traz as seguintes contribuições:

1. Discussão e proposta de uma métrica para avaliar agrupamentos de cidades com a finalidade de contornar o problema de escassez de usuários em algumas regiões.

2. Comparação dos resultados dos métodos propostos por Backstrom et al. [2010], Davis Jr et al. [2011], Kong et al. [2014], Rout et al. [2013] que fazem inferência da localização do usuário através de sua rede. Neste trabalho, analisamos as diferenças dos resultados obtidos ao se aplicar tais métodos à rede de amizade e à rede de menções, mostrando que cada uma leva a um resultado diferente.

3. Análise da variação do uso do vocabulário com o passar do tempo e proposta de uso de um método de inferência de localização geográfica que leve em conta o tempo da publicação do tweet.

4. Análise da concordância entre os métodos baseados na rede de relacionamentos entre os usuários (Backstrom et al. [2010], Davis Jr et al. [2011], Kong et al. [2014], Rout et al. [2013]), métodos baseados no uso de classificadores no campo de descrição, localização declarada e textos dos tweets (como em Han et al. [2014]) e nosso modelo de inferência baseada no tempo proposto neste trabalho.

5. Proposta e análise da combinação dos diferentes métodos de inferência de localização analisados neste trabalho através de sistemas de votação e de uma meta-árvore de decisão.

1.3 Organização do trabalho

Este trabalho organiza-se da seguinte forma: no Capítulo 2 apresentamos um apanhado da literatura com os atuais avanços para resolução do problema. O Capítulo 3 dedica-se a explicar nossa base de dados: desde sua coleta até suas principais características. No Capítulo 4 explicamos nosso modelo de inferência a partir dos tweets dos usuários levando em consideração a passagem do tempo, bem como os outros modelos de inferência analisados que exploram diferentes atributos dos usuários. Ainda no Capítulo 4 detalhamos as formas como combinamos os resultados gerados a partir desses métodos de inferência base. O Capítulo 5 traz os resultados dos nossos experimentos acompanhados de discussões e, por fim, o Capítulo 6 fecha a dissertação com nossas conclusões e direções para trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Conhecer o local de onde um tweet se origina possibilita relacioná-lo a fatores socioeconômicos, eventos e acontecimentos locais e, dessa forma, contextualizá-lo para extrair mais informações do seu conteúdo. Existem exemplos na literatura do uso da informação geográfica para diversas aplicações, tanto a nível de localização de tweets quanto de usuários.

A localização dos usuários foi fundamental para análise dos movimentos e discussões entre os interlocutores virtuais durante a Primavera Árabe (2011) e as eleições Iranianas (2009), como mostrado por Lotan et al. [2011]. Em Yardi & Boyd [2010], os autores mostram que a percepção e disseminação de acontecimentos locais ocorre primeiro para indivíduos que o presenciam ou estão em sua proximidade e logo depois pela mídia convencional que possui uma área de cobertura muito menor do que a abrangida pelos usuários de mídias sociais. Gomide et al. [2011] propõem uma forma de monitorar e prever epidemias de dengue através da localização de usuários que publicam mensagens contendo termos relacionados à doença.

A partir da localização de tweets, Lee & Sumiya [2010] deduzem regularidades geográficas em padrões usuais de comportamento de multidões e as usam para detectar comportamentos não usuais que se tornam indicadores de acontecimentos locais, como festivais de música, por exemplo. Sakaki et al. [2010] propõem e implementam um sistema para de alerta sobre terremotos baseado no monitoramento de tweets com o foco na detecção do evento.

A eficácia desses trabalhos depende da disponibilidade de dados geográficos sobre os usuários do Twitter. A fim de aumentar a quantidade de dados geográficos disponíveis, diversas técnicas foram propostas na literatura. Nas próximas próximas seções descreveremos as principais abordagens para inferência da localização do usuário.

2.1 Métodos para inferência da localização de usuários

Identificar a localização geográfica de documentos digitais não é uma necessidade nova (Smith & Crane [2001]), mas que evolui com a criação e mudança de tendência de uso das ferramentas de publicação utilizadas por usuários. Métodos para localização de páginas na Web (Zong et al. [2005]), blogs e notícias (Lieberman et al. [2010]) e fotos no Flickr (Popescu & Grefenstette [2010]) se tornaram, muitas vezes, base para trabalhos relacionados à predição da localização de usuários nas novas mídias, como Facebook, Twitter, Instagram etc.

Os métodos que buscam identificar a localização de usuários/tweets baseiam-se na hipótese de que existem indicadores geográficos, explícitos ou não, em diferentes atributos do usuário ou no conteúdo das suas mensagens. Com base nos atributos utilizados, podemos explicitar os quatro tipos principais de técnicas utilizadas: **baseadas nos campos de informação do perfil preenchidos pelo usuário**, **baseadas na rede de usuários**, **baseadas no texto dos tweets** e **híbridas**, que utilizam um ou mais dos atributos anteriores. As próximas seções mostram exemplos de técnicas baseadas nessas quatro abordagens.

2.1.1 Inferência da localização de usuários a partir das informações provenientes dos perfis

Utilizar **o campo de localização** declarado pelo usuário é uma forma simples de estimar a localização das suas mensagens.

Hecht et al. [2011] oferecem uma análise extensa do uso do campo localização declarada e mostram a necessidade de se ter cuidado ao escolher esse método como abordagem principal na inferência de localização das mensagens/usuários. Examinando tweets manualmente, os autores constataram que apenas 66% dos perfis continham uma informação geográfica válida, 18% estavam em branco e 16% possuíam uma informação não-geográfica - a maioria com referências a cultura popular. Isso explica porque **Takhteyev et al. [2012]** rotularam manualmente vários usuários depois das altas taxas de erro encontradas ao tentar geolocalizá-los automaticamente através da localização declarada.

Graham et al. [2013], por outro lado, utilizam diferentes geocodificadores conhecidos, como o disponível pela API do Yahoo! e o Google, para identificar as coordenadas geográficas do usuário a partir do seu campo de localização. Tais métodos, no entanto, dependem de serviços cujo funcionamento é desconhecido. Nesse trabalho, as precisões obtidas variam ao redor de 45% para as regiões analisadas.

Em Han et al. [2014], os autores utilizam o campo de localização declarada, a descrição do perfil do usuário e seu fuso horário como atributos de entrada para o classificador Naive Bayes, gerando um modelo de inferência para localização do usuário.

Embora Han et al. [2014] tenham apontado pequenos erros nas inferências feitas a partir de modelos treinados em épocas diferentes aplicados a novos dados, nenhum trabalho analisou como o tempo de atualização dos modelos interfere na inferência da localização de novos usuários. Neste trabalho, testamos diferentes janelas de tempo entre a atualização dos modelos treinados no campo de descrição e localização declarada do usuário.

2.1.2 Inferência de localização a partir do texto dos tweets

Levando em consideração o **texto dos tweets** como atributo de localização, existem técnicas baseadas na construção de modelos linguísticos para regiões ou cidades, como propões Cheng et al. [2010], ou que utilizam *gazetteers* e reconhedores de entidades [Gelernter & Mushegian, 2011].

Os autores de Cheng et al. [2010] criaram um método baseado puramente no conteúdo de tweets para inferir a localização do usuário. Após encontrar palavras com alta localidade geográfica através de um método probabilístico que leva em conta a dispersão geográfica das mesmas, os autores usaram-nas para treinar classificadores como Support Vector Machine, Naive Bayes e AdaBoost. O método consegue colocar 51% dos usuários em um raio de 100 milhas do seu local verdadeiro, numa base com aproximadamente 130 mil usuários. Para superar o fato de que tweets geolocalizados são geralmente concentrados em grandes centros urbanos e escassos em centros menores, os autores propõem técnicas para suavização da vizinhança dos tweets geolocalizados.

Ikawa et al. [2012] utilizam somente as palavras provindas de mensagens publicadas no twitter através de serviços de geolocalização como o Foursquare[1] como documentos de treino para um algoritmo de classificação, filtrando, dessa forma, palavras que não possuem alta localidade geográfica. Um dos problemas dessa abordagem é que ela precisa que os usuários mantenham as mensagens publicadas pelos serviços de geolocalização padronizadas, para extração de termos locais. Além do mais, grande parte das postagens publicadas não são provenientes de serviços de geolocalização, diminuindo a quantidade de dados a que se tem acesso para treino.

Alguns autores consideram que determinados tópicos são mais comuns em algumas regiões e partem desse pressuposto parainferir a localização dos usuários. Esse é o caso de Eisenstein et al. [2010] que adicionam uma variável geográfica r na forma

[1] https://foursquare.com/

padrão do algoritmo de modelagem de tópicos utilizando **LDA** (*Latent Dirichlet Allocation*) proposto por Blei et al. [2003]. As localizações observadas, portanto, são geradas a partir de regiões geográficas e a variável regional r é atribuída ao usuário. A geração desses modelos, no entanto, tem alto custo computacional e sendo ineficiente ao ser aplicada em um grande volume de dados, além de não levar em consideração a natureza temporal dos tópicos, como destacam Han et al. [2014].

Roller et al. [2012] propõem um método supervisionado que deriva um modelo linguístico a partir de uma coleção de documentos cuja geolocalização (latitude e longitude) é conhecida. Diferente de Wing & Baldridge [2011], que utilizam *grids* de tamanho uniforme e fixo, nesse trabalho os autores agrupam documentos em *grids* não-uniformes, cujos tamanhos se adaptam ao número de documentos presentes no treino e a dispersão dos mesmos pela sua área geográfica. O método foi testado em uma base de dados com 449.694 usuários, com aproximadamente 85 tweets cada, e em artigos geolocalizados da Wikipédia. O método conseguiu no máximo 34% de acurácia ao estimar a localização dos usuários da base do Twitter.

Utilizando *gazetteers* e algoritmos de reconhecimento de entidade, Gelernter & Mushegian [2011]; Sultanik & Fink [2012] e Paradesi [2011] detectam e desambiguam nomes de locais citados no conteúdo do tweet - utilizando, dessa forma, uma abordagem que tira proveito da forma explícita como locais podem ser citados pelos usuários. O Reconhecedor de Entidades de Stanford [Finkel et al., 2005] foi utilizado nessas abordagens, porém esbarrou no problema de a linguagem no Twitter ser bastante informal, conter erros ortográficos e diversas formas de abreviação não padronizadas.

Em Kinsella et al. [2011], os autores propõem a criação de um modelo linguístico para cada cidade, de acordo com a frequência das palavras que seus usuários utilizam. O método infere a localização do usuário através de um *ranking*, criado a partir da probabilidade de um tweet ter sido gerado pelo modelo de uma cidade ou pela medida de divergência de Kullback-Leibler. Utilizando tweets das 10 cidades que produzem mais tweets geolocalizados, os autores conseguiram acerto igual a 65% utilizando os modelos linguísticos criados. O uso limitado do número de cidades, porém, deixa o trabalho pouco interessante sob o ponto de vista de aplicações que buscam inferir a localização dos usuários em grandes extensões territoriais - como países - ou globais e não analisa como fatores temporais podem afetar o modelo.

Han et al. [2014] também abordam o problema de inferência da localização do usuário como um problema de classificação em que cada localização é uma classe. Os autores se diferenciam por utilizar diversas técnicas de seleção de atributos para escolha de termos a serem usados pelos classificadores, além de utilizarem e mostrarem a eficácia em se usar os tweets não-geolocalizados dos usuários como treino.

No nosso trabalho, ao contrário dos demais, mostramos como termos associados a localizações variam com o tempo e avaliamos o impacto de diferentes janelas de tempo na atualização dos modelos linguísticos gerados pelos classificadores Naive Bayes e Regressão Logística.

2.1.3 Inferência da localização de usuários a partir das suas relações

No Twitter, um usuário pode escolher seguir outro usuário caso esteja interessado no conteúdo que ele publica no serviço de microblog. Essa relação nem sempre é mútua, isto é, um usuário pode ser seguido por diversos outros mas não segui-los. Um usuário **A** que escolhe receber as publicações de **B** é seu *follower* ou **seguidor**, enquanto **B** é seu *followee* ou **seguido**. Se usuários seguem um ao outro, daremos a eles a denominação de **amigos**. A existência de relações unidirecionais é evidente, principalmente ao encontrarmos perfis de artistas, jornalistas e instituições, por exemplo, que possuem um grande número de usuários seguidores que têm interesse no conteúdo gerado pelos primeiros e não são seguidos, por sua vez, por esses usuários importantes.

Como estudado por Crandall et al. [2010], usuários que possuem relações sociais fora do mundo virtual geralmente moram em regiões próximas e seguem-se mutuamente no Twitter. Backstrom et al. [2010] reconhecem que a probabilidade de existir uma relação de amizade entre dois usuários quaisquer do Facebook é inversamente proporcional à distância entre eles. A partir de um modelo probabilístico criado para representar os laços de amizade de quaisquer usuários dadas suas distâncias geográficas, para cada usuário com localização desconhecida, a mesma é inferida identificando qual das localizações entre seus amigos que maximiza a função de verosimilhança para o mesmo. O *ground truth* é estabelecido através dos endereços declarados voluntariamente pelos usuários ao preencher seus perfis.

Davis Jr et al. [2011] propõem um dos primeiros e mais simples métodos para inferência da localização de usuários a partir da sua rede de relacionamentos. Os autores definem o local mais frequente na rede do usuário como sua possível localização. O trabalho aponta como principais desafios para inferência a existência de usuários com muitos amigos, assim como aqueles que possuem poucos deles. Segundo os autores, localizações que não possuem um número mínimo de ocorrências na rede do usuário também apresentam pouca confiança para inferência. Dessa forma, os autores definem três parâmetros para calibração do seu algoritmo: número mínimo de amigos, número máximo de amigos e número mínimo de ocorrências da localização mais frequente.

Kong et al. [2014] assumem que nem todos os amigos de um determinado usuário devem possuir o mesmo peso na inferência de sua localização. Nesse trabalho, os

autores assumem que dois usuários a e b são amigos caso ambos se mencionem em pelo menos dois tweets. Através do conceito de coeficiente de proximidade social, calculado a partir do cosseno de similaridade da rede de amigos dos dois usuários, as localizações dos amigos do usuário são ponderadas durante a inferência de sua localização. Diferente de Backstrom et al. [2010], os autores atribuem localizações para usuários durante passadas pela rede, prevenindo inferência baseada em usuários com poucos amigos cuja localização é conhecida. Após múltiplas passagens na rede, um maior número de usuários é rotulado com a possibilidade da redução de precisão.

Rout et al. [2013] utilizam-se do classificador SVM com kernel RBF para treinar uma rede com relações de seguido e seguidor. Os principais atributos para descrever cada usuário para o classificador são: 1) as cidades dos amigos do usuário, 2) o número de amigos na rede social do usuário que moram na mesma cidade e finalmente 3) o número de relações recíprocas entre os usuários da rede de amizades do usuário com localização desconhecida. Na implementação deste trabalho, utilizamos como Jurgens et al. [2015], o SVM com kernel linear.

Jurgens [2013] usa um procedimento iterativo para inferir a localização de usuários. Durante cada passada do algoritmo, a localização de cada usuário é atualizada para a média geométrica da localização de todos os seus vizinhos no grafo obtido a partir da rede de seguidos e seguidores do Twitter. O uso de múltiplas passadas permite utilizar localizações inferidas para inferir a localização de novos usuários, potencialmente aumentando o número de inferências feitas mesmo em grafos muito esparsos.

Nos trabalhos anteriores, os autores aplicaram os métodos de inferência de localização citado somente em um tipo de rede. Neste trabalho, aplicaremos os métodos de inferência propostos por Backstrom et al. [2010], Davis Jr et al. [2011], Rout et al. [2013] e Kong et al. [2014] sobre dois tipos de rede: a rede de amizade e a rede de menções entre usuários, mostrando que ambas não possuem as mesmas propriedades e, consequentemente, levam a resultados diferentes. Em especial, mostramos que embora tenha tido maior acurácia durantes os testes, a rede de menções tem revocação significativamente menor do que a obtida aplicando os mesmos métodos na rede de amizades. Além disso, analisamos o impacto de restringir o tempo mínimo entre as menções de dois usuários para que consideremos uma conexão entre ambos na rede de menções.Finalmente, comparamos a concordância entre as inferência feitas pelos métodos mostrando e mostramos a redundância entre alguns deles.

2.1.4 Técnicas híbridas para inferência da localização do usuário

Uma coleção de técnicas híbridas - que tentam usar a informação da rede do usuário e do seu conteúdo - pode ser encontrada na literatura. Chandra et al. [2011] estendem os métodos de modelagem linguística, inspirados principalmente por Kinsella et al. [2011], adicionando o fator de interação social entre os usuários através de *tweets conversacionais*, isto é, tweets entre usuários. O modelo criado, no entanto, embora mais complexo, não consegue resultados significativamente superiores aos métodos que o inspiram.

Métodos utilizando **multi-indicadores**, como a localização indicada no perfil do usuário, volume de tweets, *hyperlinks* e fuso horário informado nos tweets dos usuários foram propostos recentemente em Schulz et al. [2013] e Bouillot et al. [2012]. Nesse trabalho, os autores extraem referências explícitas do perfil e tweets dos usuários (citação a lugares, coordenadas de GPS etc) e consideram a localização do usuário como aquela com maior número de referências agregadas. Essa técnica, no entanto, será tão proveitosa quanto a qualidade dos recursos disponíveis, como *gazetteers*, usados para inferência da localização, o que pode ser um impeditivo para alguns sistemas e áreas geográficas sobre as quais podemos não ter tantas informações, como nomes de pontos de interesse.

Já Ren et al. [2012] e Rodrigues et al. [2013], baseiam-se em modelos probabilísticos criados a partir dos laços sociais do usuário e do conteúdo dos seus tweets. Tais trabalhos apontam resultados promissores, mas ambos com experimentos em bases com menos de 10.000 usuários.

Em outro trabalho, Mahmud et al. [2012] utilizam-se de um comitê de classificadores treinados a partir de tweets originados das 100 cidades mais populosas dos Estados Unidos. Tendo como atributos as palavras, *hashtags* e nomes de locais retirados de um *gazetteer*, o método atribui a um usuário seu local mais provável com base no texto dos seus últimos 200 tweets, de forma hierárquica: primeiro atribuindo-lhe seu **fuso horário/estado** e, em um segundo passo, sua **cidade**. O método consegue uma acurácia igual a 54% e 66% para estimativa de cidade e estado, respectivamente.

No nosso trabalho, embora também consideremos um comitê de métodos de inferência para determinar a localização do usuário, tais métodos base utilizam atributos diversos, como tweets, descrição, localização declarada e rede do usuário. Além disso, ponderamos o voto de cada método-base utilizando um algoritmo genético e comparamos os resultados com os obtidos através da combinação dos métodos-base utilizando uma meta-árvore de decisão.

Capítulo 3

Caracterização da base de dados

Neste capítulo descrevemos o processo de coleta que deu origem à base de dados que utilizamos nos nossos experimentos e como se comportam alguns dos atributos utilizados para inferência da localização do usuário, incluindo a rede de amizade, rede de menções, texto e informações do perfil como descrição e localização declarada.

Na seção 3.1, descrevemos brevemente as APIs oferecidas pelo Twitter e suas limitações para, então, detalharmos como criamos as bases utilizadas neste trabalho. Na Seção 3.2, descrevemos as características gerais das bases de dados, para nas Seções 3.3, 3.4 e 3.5 aprofundarmo-nos nas caracterizações dos atributos de rede de relacionamentos, informações do perfil e tweets do usuário, respectivamente. Por fim, a Seção 3.6 explica a escolha do nosso *Ground Truth*.

3.1 Coleta

O Twitter oferece a API Streaming[1] aos desenvolvedores, que permite a recuperação em tempo real de até 10% do total dos **tweets postados publicamente** no intervalo de 15 minutos. Há três formas de especificar quais os requisitos que um tweet deve atender para que seja filtrado pela API:

- **Área geográfica**: tweets são recebidos caso estejam dentro dos limites de *bounding boxes* declaradas. **Bounding boxes** são área geográficas definidas por dois pares de latitude e longitude que descrevem um retângulo. A utilização de *bounding boxes* é a forma mais eficaz de receber tweets que foram gerados exclusivamente dentro de uma área geográfica sobre a qual se tem interesse e são marcados

[1] https://dev.twitter.com/streaming

com informação geográfica gerada por aparelhos contendo GPS ou através de GeoIP. Os serviços de GeoIP associam a localização do usuário de acordo com seu IP que está contido em uma faixa associada a alguma cidade. Nem sempre os tweets marcados através de GeoIP representam a real localização do usuário devido a mais de uma cidade compartilhar a mesma faixa de IP ou a uma desatualização da lista utilizada.

- **Termos**: tweets são recuperados em tempo real caso contenham algum dos termos fornecidos pelo desenvolvedor, independente do local em que foram publicados. É possível listar até 300 termos diferentes para serem monitorados através da API Streaming. Essa é uma opção amplamente utilizada para monitorar tweets publicados a respeito de algum assunto específico, como doenças ([Gomide et al., 2011]) ou políticos ([Tumasjan et al., 2010]).

- **Usuários**: tweets são recebidos caso tenham sido publicados por algum dos usuários listados pelo desenvolver. É possível declarar até 3.000 usuários diferentes para receber seus tweets públicos em tempo real. Essa opção pode ser utilizada para monitorar o comportamento de determinado grupo de usuários, por exemplo.

Além da API Streaming, o Twitter também oferece acesso à sua API REST[2] cujas diferentes interfaces limitam o número de requisições feitas dentro de um intervalo de tempo. Dentre as interfaces expostas, utilizamos:

- **user_timeline:** Através dessa interface, é possível recuperar os 3.200 últimos tweets públicos de um usuário. É possível fazer até 300 requisições no período de 15 minutos e cada requisição pode recuperar até 200 tweets. Essa interface é importante para recuperar tweets publicados no passado por algum usuário sobre o qual se tenha interesse.

- **followers/ids** e **friends/ids**: Através dessas interfaces, é possível recuperar os seguidores e seguidos de um usuário especificado pelo desenvolvedor, respectivamente. No presente momento, é possível fazer 15 requisições a cada 15 minutos e cada requisição recupera o **id** de 5.000 seguidores/seguidos. Essa interface é importante para amostrar a rede de usuários no Twitter.

[2]https://dev.twitter.com/rest

3.1.1 Processo de coleta

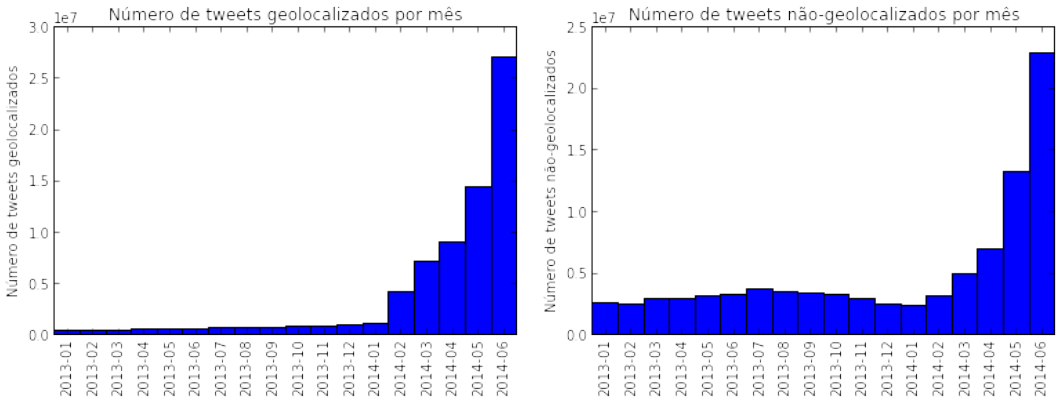
Tweets podem conter informação sobre a latitude e longitude da localização do usuário no momento da sua publicação. Para que um tweet contenha tal informação, é necessário, primeiramente, que o usuário altere suas configurações de privacidade. Uma vez que o usuário permita que sua localização seja disponibilizada através dos seus tweets, a mesma é derivada através do serviço de GPS do aparelho utilizado para postar a mensagem no microblog ou através da tradução do seu endereço de IP para informação geográfica. Para coletarmos somente tweets geolocalizados no Brasil, definimos várias *Bounding boxes* contendo o país e coletamos os tweets gerados dentro dos seus limites pelo período de novembro de 2013 até março de 2014 utilizando a API Streaming, removendo tweets provindos de países vizinhos que pudessem ter parte do seu território dentro das *bounding boxes* definidas para o Brasil.

Enumeramos todos os usuários que publicaram esses tweets, obtendo aproximadamente 601 mil usuários. De março de 2014 a junho de 2014 continuamos a coleta de tweets geolocalizados e, através da interface *user_ timeline*, coletamos os 200 tweets mais recentes de cada um dos 601 mil usuários, removendo tweets anteriores a janeiro de 2013. Essa nova coleta através da interface *user_ timeline* recuperou tanto tweets geolocalizados quanto tweets não-geolocalizados. Dessa forma, aumentamos nossa base de tweets geolocalizados para mais de 71 milhões de tweets. Além de novos tweets geolocalizados, criamos uma nova base com tweets não geolocalizados publicados pelos 601 mil usuários. A base de tweets não-geolocalizados tem volume de mais de 125 milhões de tweets. Temos, portanto, duas bases distintas: a base **Geo** contendo apenas os tweets geolocalizados e a base **NotGeo** que contém os tweets sem nenhuma marcação de GPS, ambas contendo tweets dos 601 mil usuários coletados até Março de 2014. Coletamos tweets não-geolocalizados para, assim como Han et al. [2014], podermos avaliar como se comparam os métodos de inferência de localização baseados em texto quando aplicados a tweets geolocalizados e tweets não geolocalizados.

Com a finalidade de estudar como esses usuários interagem com outros usuários na rede de seguidos-seguidores do Twitter, recuperamos os **ids** de até 5.000 seguidos e 5.000 seguidores de cada usuário. Não identificamos, intencionalmente, todos os seguidos/seguidores para não esbarrar nas limitações impostas pelas respectivas interfaces e por sabermos de trabalhos anteriores desenvolvidos por Rout et al. [2013] e Crandall et al. [2010] que determinam que usuários que apresentam um grande número de seguidos/seguidores são, em geral, celebridades, *spammers* ou contas falsas, que se distinguem do usuário comum do Twitter.

3.2 A base em números

Como não é possível coletar todos os tweets publicados através da API Streaming devido a limitações explicadas na seção 3.1, o número de tweets geolocalizados é mais volumoso nos meses de 2014 devido a utilização da API REST que permite a coleta de todos os tweets recentes do usuário até o limite de 3.200 tweets por usuário. Essa desproporção pode ser vista na Figura 3.1a.



(a) Distribuição de tweets geolocalizados por mês

(b) Distribuição de tweets não geolocalizados por mês

Figura 3.1: Distribuição dos tweets das bases Geo e NotGeo por mês

Como pode ser visto na Figura 3.2, aproximadamente 60% dos usuários têm menos de 100 tweets geolocalizados publicados na nossa base.

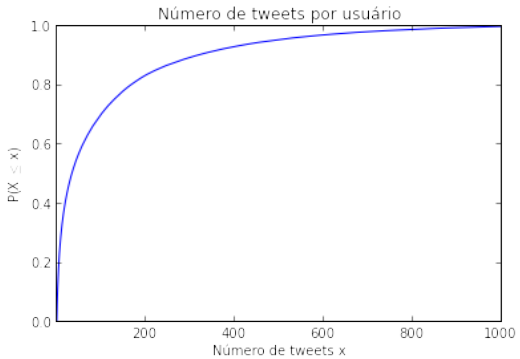


Figura 3.2: Número de usuários por quantidade de tweets postados

Descartamos usuários com menos de 3 tweets geolocalizados e para os usuários restantes, designamos o município brasileiro onde o usuário possui mais tweets geolocalizados como aquele onde ele reside ou passa maior parte do tempo, ou seja, a **localização do usuário**. Na Figura 3.3, mostramos a quantidade de usuários por

proporção da cidade mais frequente. Note que para a maior parte dos usuários, mais de 90% dos seus tweets se originam da mesma cidade.

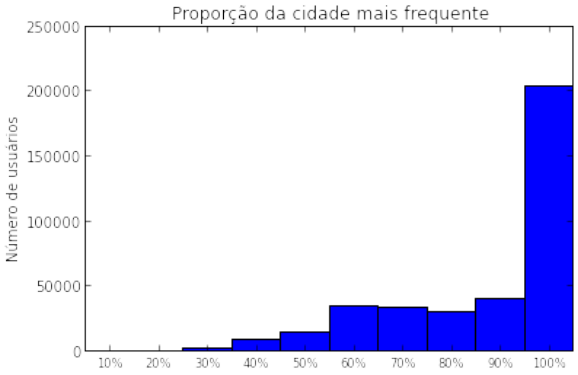


Figura 3.3: Para a maior parte dos usuários, a sua cidade principal aparece em mais de 90% dos seus tweets

O número de usuários não é uniforme para todas as cidades do Brasil, como pode ser visto na Figura 3.4. Ao compararmos as listas das 30 cidades mais populosas do país (Figura 3.5) e as 30 cidades mais frequentes na base (Figura 3.4), notamos que as cidades mais frequentes estão também entre as mais populosas. Isso sugere que, dentre outros fatores, o número de habitantes de uma cidade tem relação proporcional ao número de tweets geolocalizados publicados na mesma. No entanto, as posições de algumas cidades nos *rankings* das Figuras 3.4 e 3.5 não coincidem, como é o exemplo de Salvador que é a terceira cidade mais populosa, porém a décima terceira cidade com maior número de usuários. Enquanto Porto Alegre é a terceira cidade com maior número de usuários, mas é a décima em população. Isso se deve a diversos fatores como a penetração e acesso da população a tecnologia. Áreas mais ricas podem oferecer à população uma possibilidade maior de obter aparelhos dotados de GPS. Se usarmos o PIB per capita como uma medida grosseira de riqueza de uma cidade, a cidade de Porto Alegre tem o PIB per capita superior a 30 mil reais, enquanto em Salvador o mesmo não ultrapassa 15 mil reais de acordo com pesquisa publicada pelo IBGE em 2010[3]. **Essa caracterização dos dados sugere que existe um viés que deve ser levado em consideração em trabalhos de monitoramento do Twitter: alguns grupos sociais e áreas geográficas podem estar subrepresentados nas amostras que um pesquisador coleta.**

Os usuários apresentam um comportamento homogêneo quanto ao número de cidades de onde postam. Quando saem de suas cidades, eles tendem a visitar cidades próximas - localizadas até 200km de distância - como pode ser visto na Figura 3.6a.

[3]ftp://ftp.ibge.gov.br/Pib_Municipios/2010/pdf/tab01.pdf

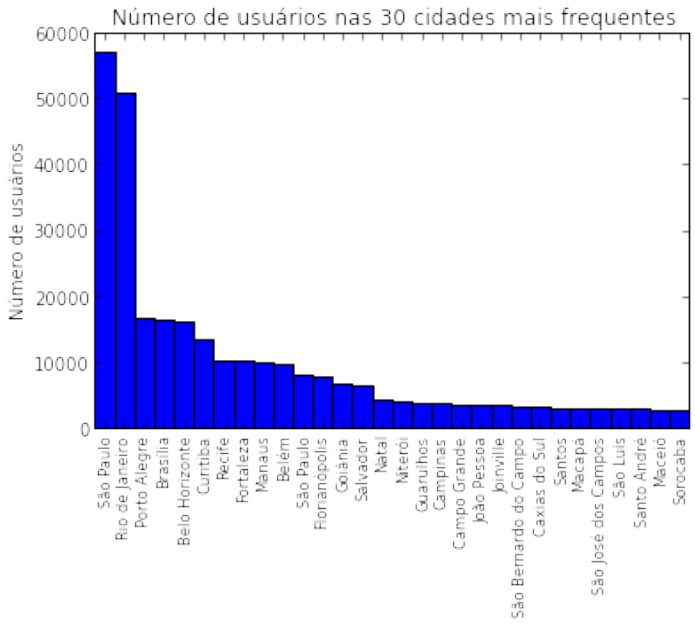


Figura 3.4: As 30 cidades brasileiras com maior número de usuários geolocalizados

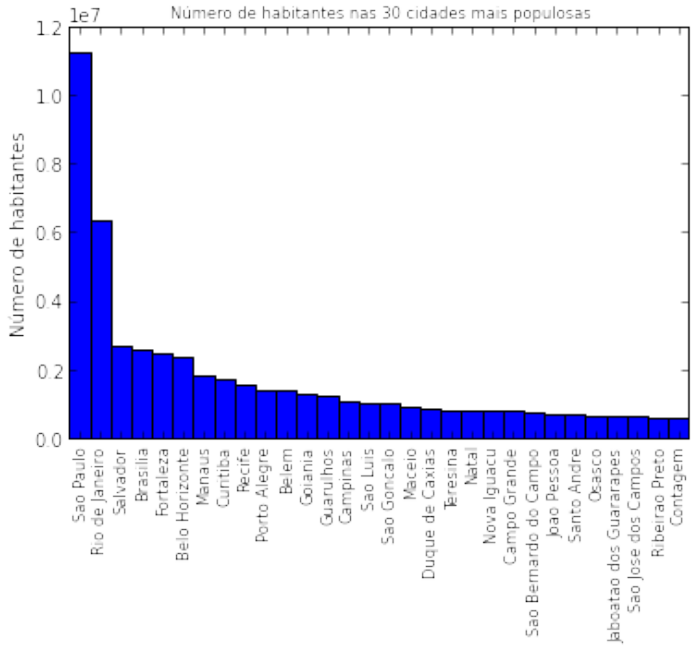
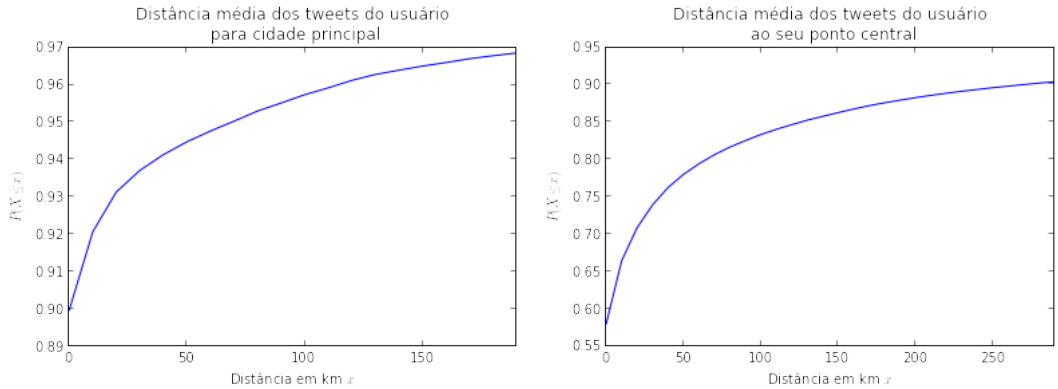


Figura 3.5: As 30 cidades brasileiras com maior número de habitantes

Além de definirmos a cidade principal do usuário, definimos o ponto central como o centro de massa dos tweets produzidos pelo usuário, isto é, definimos a latitude (lat_c) e longitude (lon_c) do ponto central como:

$(lat_c, lon_c) = (\frac{1}{M}\sum_{i=1}^{M} lat_i, \frac{1}{M}\sum_{i=1}^{M} lon_i)$ onde M é o número de tweets do usuário e lat_i e lon_i são a latitude e longitude do tweet i. Como mostrado na Figura 3.6b, cerca

de 80% dos usuários posta a uma distância média inferior a 100km do ponto central, o que, juntamente com as Figuras 3.3 e 3.6a, nos leva a concluir que os tweets da base Geo possuem alta localidade espacial para usuários.



(a) 90% dos usuários não posta a partir de mais de uma cidade

(b) 80% dos usuários posta em média dentro de um raio menor do que 100km

Figura 3.6: Comportamento dos usuários de acordo com seu padrão de movimentação

3.3 Rede de usuários

A hipótese de que a amizade no Twitter é um indício de relação fora do mundo virtual e, portanto, de proximidade geográfica foi explorada em alguns trabalhos, como por Rout et al. [2013] e Davis Jr et al. [2011], entre outros listados na seção 2.1.3.

Podemos definir diversas relações entre os usuários do Twitter, organizando-os como uma rede ou grafo. Existem quatro redes já utilizadas em trabalhos anteriores e que servem de apoio para inferir a localização do usuário: rede de seguidores, rede de seguidos, rede de amigos e rede de menções.

Na Figura 3.7 vemos a distribuição de seguidos (ou *followees*) para os usuários da nossa base no último mês do processo de coleta explicado na Seção 3.1.1. Cerca de 80% dos usuários possui até 80 seguidos. Podemos ver uma curva semelhante para o número de seguidores (*followers*), também na Figura 3.7. Cerca de 80% dos usuários têm até 60 seguidores. O número de amigos de cada usuário é menor do que o número de seguidos e seguidores. Em geral, nem todas as pessoas que um usuário segue o seguem de volta. Na nossa base, 80% dos usuários possui no máximo 40 amigos. Por outro lado, menos de 10% dos usuários não possui nenhum amigo, seguido ou seguidor.

A rede de seguidos, seguidores e, consequentemente, de amigos muda com o passar do tempo. Usuários podem deixar de seguir certos usuários e seguir outros por uma série de motivos que podem indicar mudança de interesse. Durante o período avaliado,

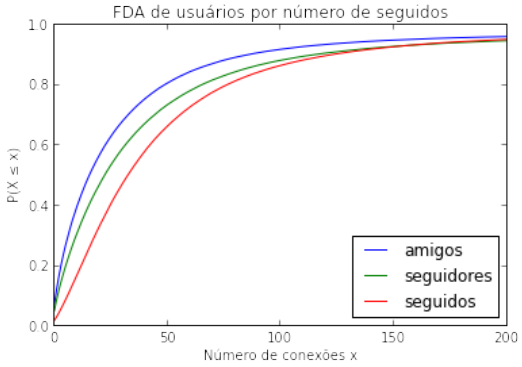
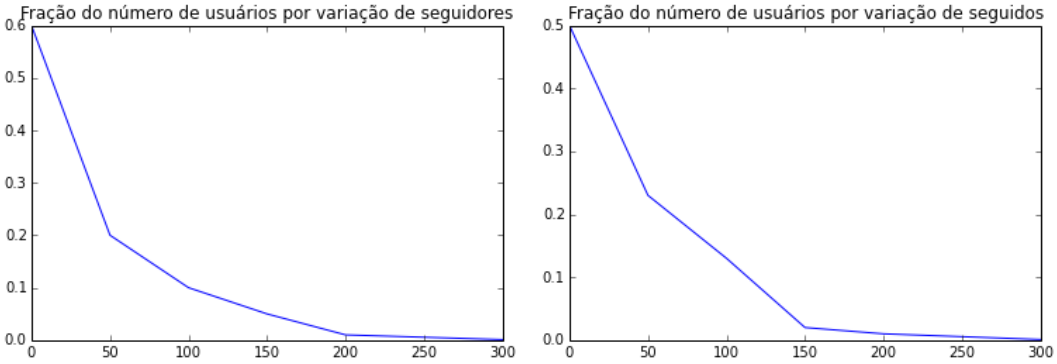


Figura 3.7: Função de distribuição acumulada para número de amigos, seguidos e seguidores

analisamos as mudanças que ocorreram nos números de seguidos e seguidores dos nossos usuários. Por limitações impostas pelo número de requisições que é possível fazer à API rest do Twitter, não conseguimos analisar o histórico de quais usuários deixaram de ser seguidos.

Nas Figuras 3.8a e 3.8b, podemos observar que a maior parte dos usuários (cerca de 60%) não apresenta variação no número de seguidores, enquanto cerca de 50% dos usuários não apresenta variação no número de seguidos. A variação é calculada entre o primeiro e último tweet coletado de cada usuário entre janeiro de 2013 e junho de 2014.



(a) Variação do número de seguidores (b) Variação do número de seguidos

Figura 3.8: Dinâmica da rede de seguidos/seguidores no Twitter

Finalmente, como podemos ver nas Figura 3.9, a maior parte dos usuários se relaciona com outros usuários que estão, em média, a no máximo 1.000km de distância da sua cidade principal. Para fins de comparação, essa distância equivale à aproximadamente a distância entre Belo Horizonte e Joinville (SC). Esse número, no entanto, é enviesada por amigos/seguidos/seguidores que estão muito distantes do usuário

analisado. Na Figura 3.10, vemos que aproximadamente 35% das conexões de seguidos/seguidores estão na mesma cidade (0km de distância), enquanto mais de 50% das conexões de amizade estão a menos de 100km de distância. Para fins de comparação, a distância de 100km equivale a, aproximadamente, a distância entre a cidade de São Paulo e São José dos Campos.

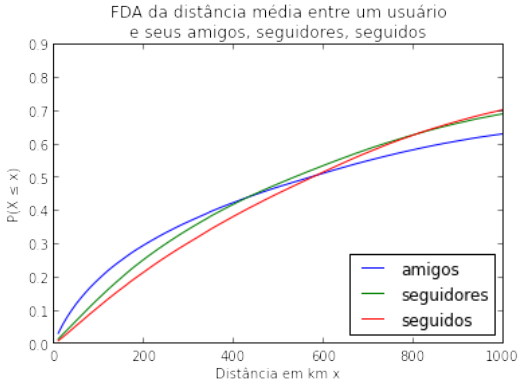


Figura 3.9: A distância média entre um usuário e seus amigos é menor que 1,000km para 63% dos usuários

Continuando a análise na Figura 3.10, concluímos que a rede de amizades possui maior localidade espacial em relação ao usuário do que a rede de seguidos e seguidores, e é ligeiramente menos densa, como mostrado na figura 3.7. Por esses motivos, ela aparece como uma boa alternativa para inferência de localização geográfica de usuários.

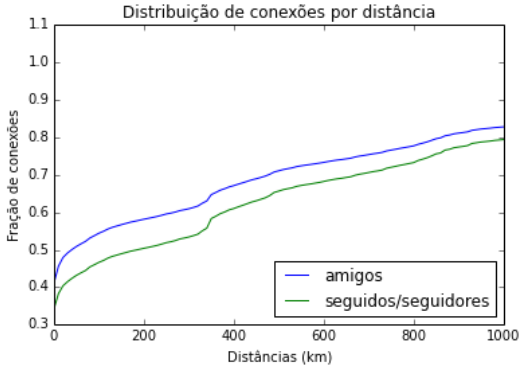


Figura 3.10: Função de distribuição acumulada da distância por tipo de conexão

Relacionar os usuários através de **menções** é outra forma de organizá-los como uma rede. Menções são referências que usuários fazem a outros usuários da comunidade através de seus nomes de usuários. Essas referências servem para indicar repostagem do conteúdo referido (quando o nome do usuário é antecedido pelos caracteres **RT**, indicando a ação de *retweetar*) ou para dirigir-se diretamente ao usuário referido.

Há diversas formas de indicar uma conexão entre usuários através de suas menções, mas, como Jurgens et al. [2015], neste trabalho consideramos que há uma conexão entre dois usuários **A** e **B** se ambos fazem uma menção ao outro pelo menos **uma vez** durante o período observado. Considere, por exemplo, uma janela de tempo igual a 15 dias. Um usuário **A** se conectará a um usuário **C** se e somente se o intervalo máximo entre uma menção do usuário A a C é menor ou igual a 15 dias.

Um dos fatores estudados nesse trabalho é a influência do tempo em características que podem ser utilizadas para inferência da localização. Assim, para avaliar como o tamanho da janela de tempo impacta a rede de menções, consideramos somente os usuários contidos na nossa base de dados. Dessa forma, se o usuário **A** menciona o usuário **B**, mas esse não faz parte da nossa base de dados, essa referência não é considerada.

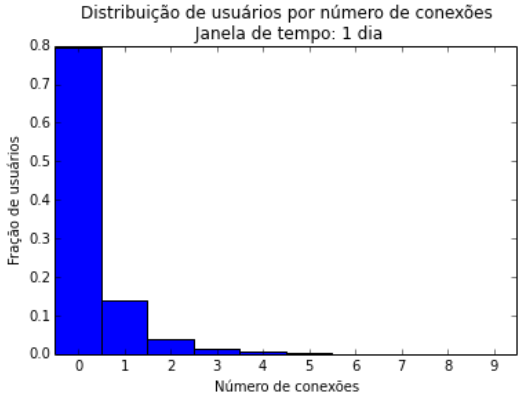
É de se esperar que quanto maior a janela de tempo observada, maior é a quantidade de mensagens disponíveis e, portanto, maior é o número de menções feitas entre usuários. No entanto, como podemos observar nas Figuras 3.11a, 3.11b e 3.11c, o aumento do tamanho da janela de tempo pouco influencia no número de conexões criadas. Para a janela de tempo de 1 dia, temos apenas 5% menos conexões que o período de tempo para o período que envolve a base inteira. Isso sugere que usuários se respondem grande rapidez, uma vez iniciada a interação - ou no caso oposto, nunca se respondem, não importa quanto tempo passe.

Nas figuras 3.12a, 3.12b e 3.12c, analisamos como a distância entre os usuários se comporta com o aumento da janela de tempo. O aumento de 5% no número de conexões observado na janela de tempo igual ao período inteiro da base pouco afeta a distância entre os usuários conectados. Observe que um número superior a 70% das conexões existe entre usuários que estão a menos de 10km de distância um do outro, mostrando uma localidade geográfica ainda maior do que a da rede de amizades e, portanto, pode levar a resultados mais precisos.

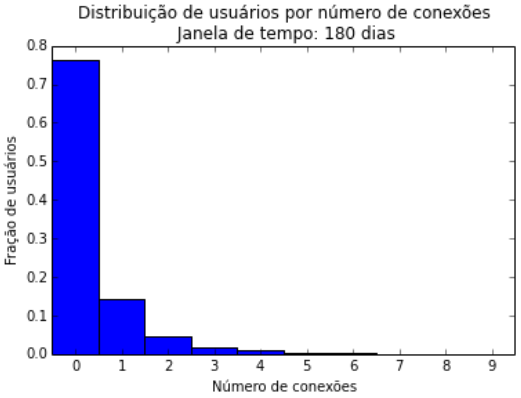
3.4 Atributos do perfil do usuário

Ao preencher o seu perfil, algumas informações são requisitadas ao usuário pelo Twitter, dentre elas:

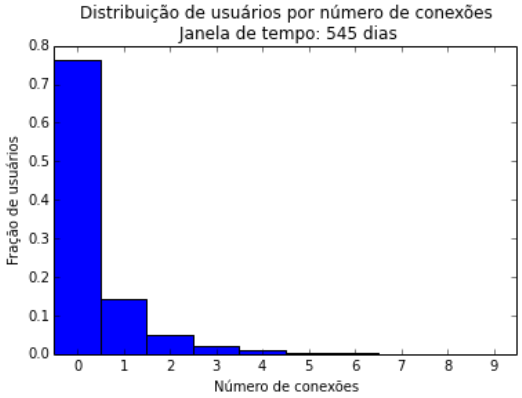
1. **Descrição do usuário dono do perfil**: contém um texto curto (140 caracteres) e pode ser utilizado tanto de forma jocosa quanto para descrever a profissão, interesses e outros aspectos da vida do usuário.



(a) Janela de tempo de 1 dia: aproximadamente 80% dos usuários não possui nenhuma conexão com outros usuários da base



(b) Janela de tempo de 180 dias: pouco mais de 75% dos usuários não possui nenhuma conexão com outros usuários da base



(c) Janela de tempo de 545 dias: pouco mais de 75% dos usuários não possui nenhuma conexão com outros usuários da base

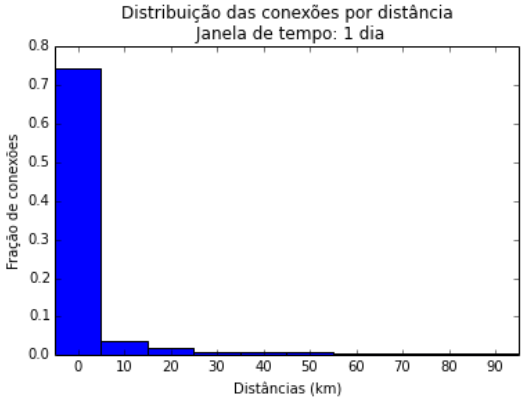
Figura 3.11: A distribuição de conexões por usuário pouco muda com a alteração da janela de tempo

2. **Localização declarada** que é um texto livre que, por isso mesmo, nem sempre pode ser utilizado de forma direta, já que é passível de não ser bem estruturada ou até mesmo falsa e jocosa.

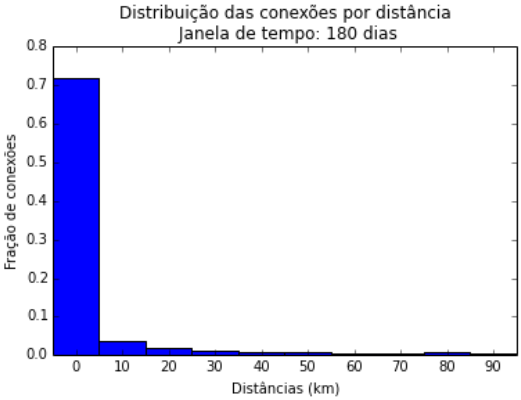
3. **Fuso horário utilizado pelo usuário** que indica o fuso horário preferencial do dono da conta.

Esses atributos já foram utilizados previamente para inferir a localização dos usuários sem levar em consideração sua dinamicidade com o passar do tempo.

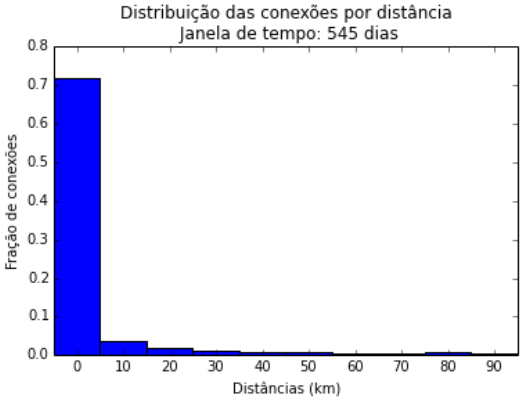
Verificamos algumas alterações temporais desses atributos nos usuários da nossa base de dados e temos que entender se essas modificações têm impacto direto na locali-



(a) Aproximadamente 75% das conexões estão a menos de 10km de distância



(b) Pouco mais de 70% das conexões estão a menos de 10km de distância



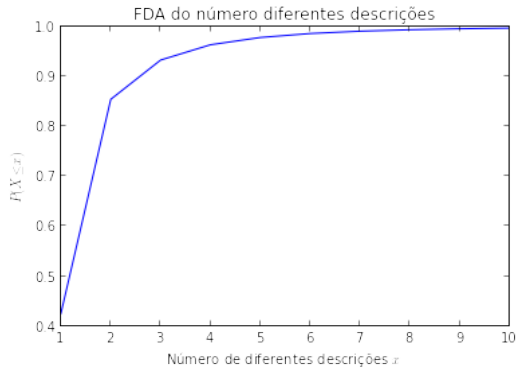
(c) Pouco mais de 70% das conexões estão a menos de 10km de distância

Figura 3.12: A distribuição da distância das conexões entre usuários pouco muda com a alteração da janela de tempo

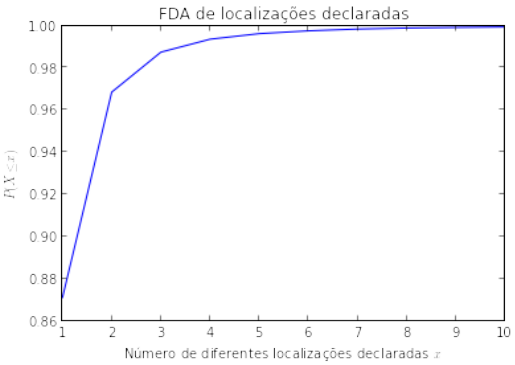
zação do usuário. Durante o período analisado (2013-2014), houve mudanças tanto na descrição de perfis quanto na localização declarada e fuso horário de onde publicaram seus tweets.

Para o caso da descrição, a maioria dos usuários mudou a descrição pelo menos uma vez durante o período analisado, como mostramos na Figura 3.13a. Por outro lado, menos de 15% dos usuários muda a localização declarada no seu perfil do Twitter, como podemos ver na Figura 3.13b, o que pode ser correlacionado com a pouca variação de locais encontrados na Figura 3.6a.

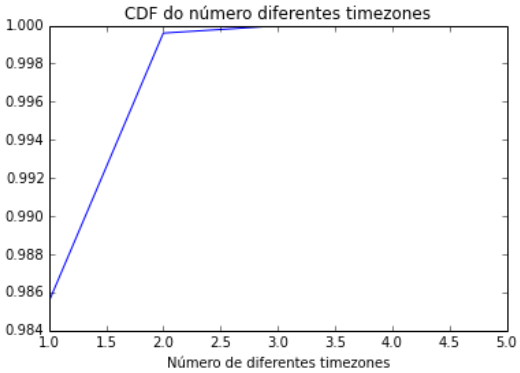
O Brasil, pela sua extensão, possui quatro diferentes fusos horários. Portanto, ao identificar de qual fuso horário alguém publica suas mensagens no Twitter, é possível reduzir o número de possíveis localizações de um usuário. De acordo com a Figura 3.13c, menos de 2% dos usuários tweeta de fusos horários diferentes - o que está de



(a) Menos de 45% dos usuários possui a mesma descrição durante todo período



(b) Menos de 15% dos usuários muda a sua localização declarada durante o período



(c) Menos de 2% dos usuários esteve em mais de um fuso horário no período

Figura 3.13: Número de diferentes descrições, localizações declaradas e fusos horários dos usuários

acordo com a observação feita sobre a pouca mobilidade de usuários com base na Figura 3.6a, uma vez que um usuário só mudaria de fuso horário caso mude de cidade.

Sendo **a descrição do usuário** o atributo com maior variação, cabe investigar se a descrição pode ser considerada como fonte para inferência da localização do usuário, uma vez que ela varia muito mais do que o número de cidades que o usuário visita. Além disso, investigaremos se esse caráter dinâmico implica na atualização dos modelos de inferência criados a partir desse campo.

3.5 Texto dos tweets

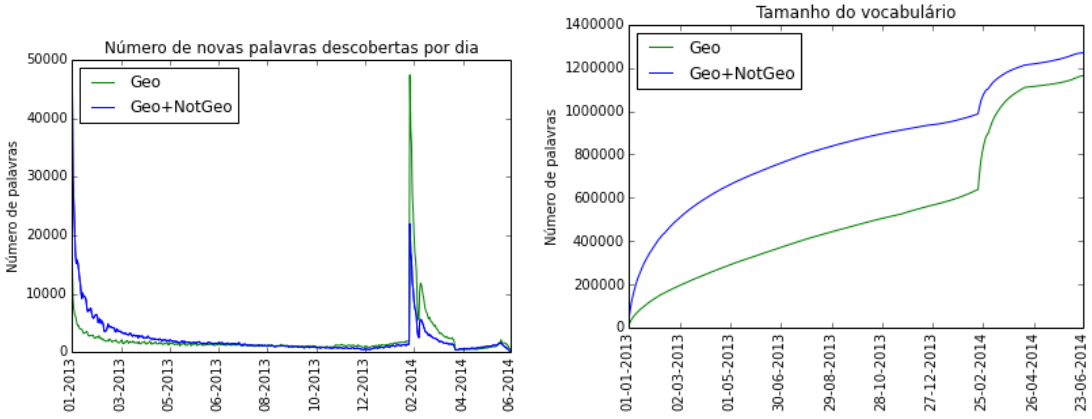
O texto contido nos tweets é uma das principais fontes para inferência da localização dos usuários, uma vez que os mesmos podem discutir assuntos de interesse e/ou predominantes em uma região. Perguntamo-nos, portanto, se esses assuntos e interesses

são homogêneos com o passar do tempo.

O vocabulário do Twitter, isto é, o conjunto de palavras únicas usadas pelos usuários, varia com o período analisado, seja por adição de novas palavras e/ou *hashtags* que caem no uso comum ou pela ausência de expressões que caíram em desuso.

Na Figura 3.14a, vemos que inicialmente há um grande número de palavras encontradas por dia tanto na nossa base de tweets geolocalizados quanto não-geolocalizados. A tendência de que o número de palavras novas encontradas por dia diminua segue até o início do ano de 2014, quando a taxa de tweets coletados por dia sobe significativamente, como explicamos na Seção 3.2. Note que enquanto a base de tweets geolocalizados tem um pico de quase 50 mil palavras novas encontradas em uma única data, a base de tweets geolocalizados e não-geolocalizados tem um pico de aproximadamente 20 mil novas palavras. Isso se deve ao fato de que o vocabulário que é novo para a base de tweets geolocalizados já havia sido visto anteriormente na base de tweets não geolocalizados, uma vez que ela é maior que a primeira.

O tamanho do vocabulário tende a se estabilizar, como vemos na Figura 3.14b, com uma possível convergência dos tamanhos do vocabulário das bases Geo e Geo+NotGeo. Com a finalidade de evitar a contagem de palavras que foram digitadas erroneamente e/ou são pouco utilizadas, para ambas as análises foram consideradas somente as palavras utilizadas por no mínimo 10 usuários distintos.



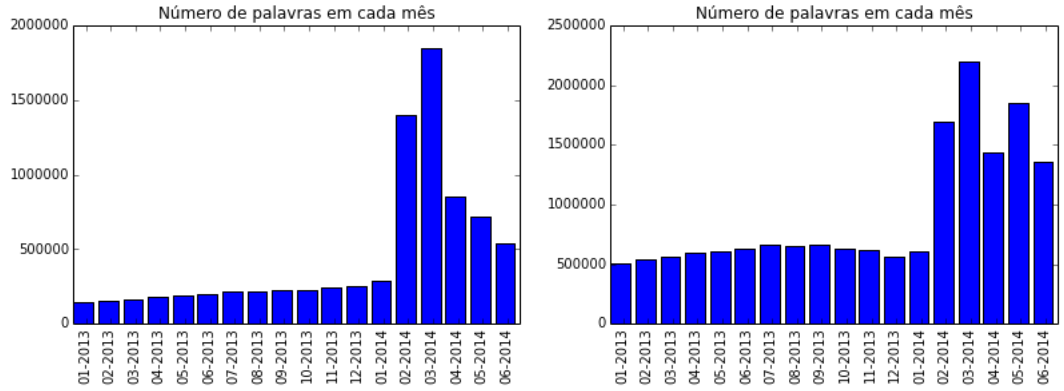
(a) Número de palavras novas encontradas por dia

(b) O número de palavras únicas encontradas sempre cresce.

Figura 3.14: Evolução do vocabulário considerando as bases de Tweets geolocalizados e não-geolocalizados

O tamanho do vocabulário tem impacto direto no número de atributos textuais que podem ser utilizados para criação de modelos linguísticos que descrevam nossa base de dados. Considere, por exemplo, um modelo de unigrama onde cada palavra é

considerada como um atributo do modelo. Nas Figuras 3.15a e 3.15b vemos o tamanho do vocabulário para cada mês, isto é, o número de palavras únicas com no mínimo 10 ocorrências no período. Não é surpresa que o número de palavras únicas na base de tweets Geo+NotGeo seja maior, uma vez que contém os tweets da base Geo acrescidos com os tweets não geolocalizados.



(a) Número de palavras que podem ser utilizadas como atributos em cada mês na base Geo

(b) Número de palavras que podem ser utilizadas como atributos em cada mês na base Geo+NotGeo.

Figura 3.15: O número de palavras disponíveis muda em cada mês, independente da base utilizada

Nos modelos baseados em unigrama, para cada palavra do vocabulário é atribuída uma probabilidade independente das outras de que ela apareça nos documentos. Queremos saber o quão bem um modelo de unigrama criado em um período de tempo T_i consegue generalizar para um novo período de tempo T_{i+1}, a fim de avaliar como o uso do vocabulário muda com o tempo.

Para responder essa pergunta, utilizaremos o conceito de **perplexidade**, comum na área de processamento de linguagem natural [Martin & Jurafsky, 2000]. Dado um modelo linguístico, a perplexidade do mesmo se relaciona de forma inversa com a probabilidade que ele atribui a ocorrência de uma amostra desconhecida. Probabilidades maiores são associadas a maior capacidade do modelo linguístico de gerar a amostra testada e, consequentemente, temos uma menor perplexidade. Modelos que não generalizam para novos documentos resultam em valores de perplexidade mais altos.

Na Figura 3.16, analisamos a perplexidade do modelo criado a partir dos unigramas encontrados no mês de janeiro de 2013. Para criação do modelo, consideramos de forma simplificada que a probabilidade de ocorrência de uma palavra $p(w_i)$ é dada por:

$$p(w_i) = \frac{o_i + 1}{W + V},$$

onde o_i é o número de ocorrências da palavra w_i, W é o número total de palavras do corpus referente ao período de janeiro de 2013 e V é o tamanho do vocabulário considerado. Utilizamos a suavização de Laplace para que nenhuma probabilidade seja zerada.

Na Figura 3.16, mostramos a evolução da perplexidade em cada um dos meses tanto para a base Geo quanto para Geo+NotGeo. Observe que com o passar do tempo a perplexidade do modelo tende a aumentar frente a novos dados. Isso corrobora a nossa hipótese de que, por seu caráter dinâmico, modelos linguísticos criados a partir de tweets precisam ser, no mínimo, atualizados com o passar do tempo. A fim de descobrir o quão útil é a nossa base de tweets não geolocalizados, criamos também um modelo somente com os tweets de janeiro de 2013 da base NotGeo e medimos a perplexidade na amostra de tweets geolocalizados. Observe que a perplexidade desse modelo a partir de março de 2013 é sempre inferior ao modelo criado na própria base Geo.

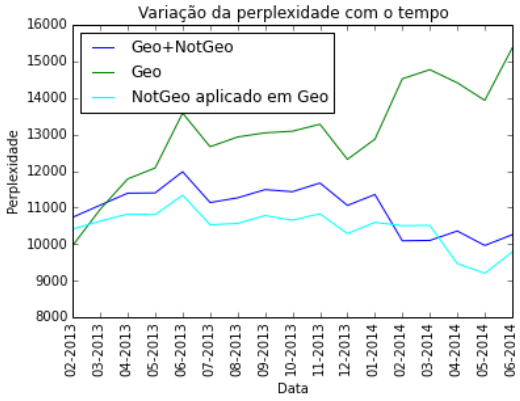


Figura 3.16: Perplexidade do modelo de unigrama criado em Jan-2013 aplicado nos meses seguintes

Finalmente, analisamos como se comporta o conjunto de palavras que melhor caracterizam cada localização. Primeiramente, agrupamos as cidades como descrito na Seção 4.5 e aplicamos o teste χ^2 para enumerar os percentis mensais das palavras com maior dependência da localização.

O teste χ^2 é comumente utilizado para examinar o grau de independência entre duas variáveis aleatórias. A tabela de contingência 3.1 representa os valores das variáveis. A equação da estatística é dada por:

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	na cidade c	em cidade diferente de c
w	$O_{w,c}$	$O_{w,C}$
W	$O_{W,c}$	$O_{W,C}$

Tabela 3.1: Tabela de contingência para palavra e co-ocorrência na cidade c

Onde O_i representa uma observação (i.e., co-ocorrência de uma cidade (c) e a palavra (w)) e n é o número de células na tabela. $O_{w,c}$ e $O_{W,C}$ denotam a ocorrência da palavra w na cidade c e palavras diferentes de w em cidades que não sejam c, respectivamente. $E_{w,c}$ denota a frequência esperada para w em c, calculada a partir das probabilidades marginais e número total N:

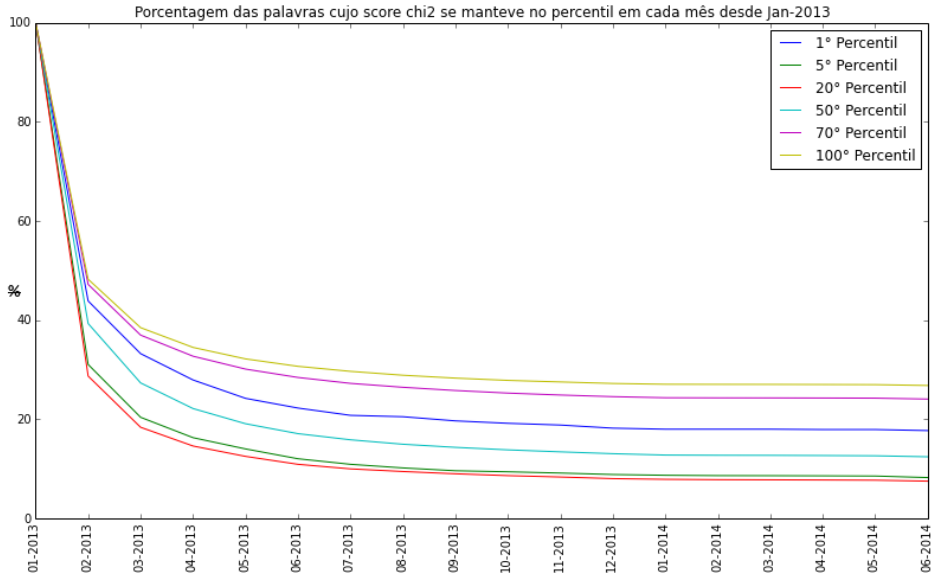
$$E_{w,c} = P(w)P(c)N = \frac{O_{w,c} + O_{w,C}}{N} \frac{O_{w,c} + O_{W,c}}{N} N$$

$$N = O_{w,c} + O_{W,c} + O_{w,C} + O_{W,C}$$

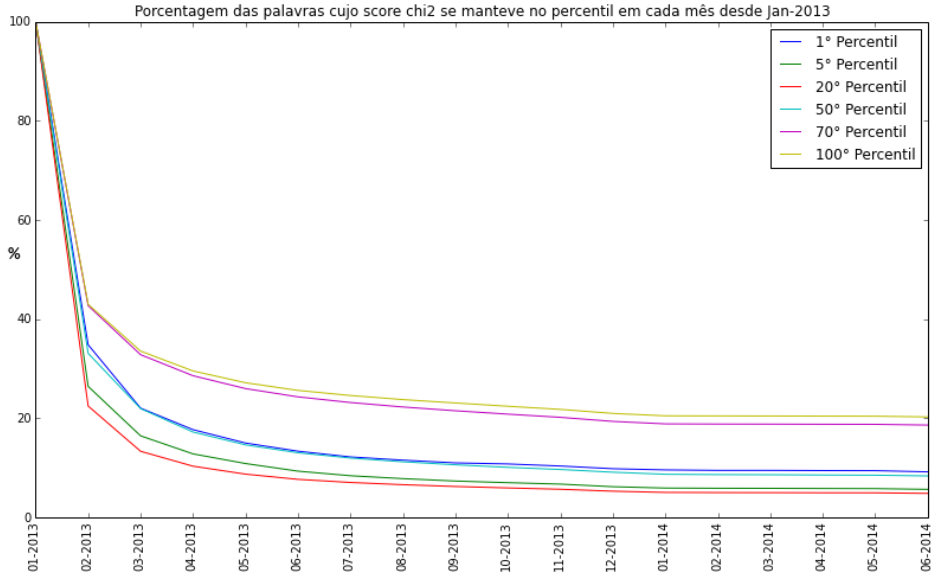
Altos valores de χ^2 indicam que a hipótese nula H_0 de independência deve ser rejeitada e, portanto, a ocorrência de um termo e a cidade (classe) são dependentes. Para cada percentil, os termos mais significativos são considerados como aqueles cujos valores do χ^2 são maiores do que determinado limiar. Para o primeiro percentil, por exemplo, temos que o valor mínimo dentre os termos listados é maior do que 99% dos outros termos da base inteira.

Nas Figuras 3.17a e 3.17b, para cada mês atribuímos a porcentagem dos termos que se mantiveram em relação ao mês anterior. Ao consideramos todo o conjunto de palavras, no percentil 100, a porcentagem de palavras mantidas aumenta em decorrência do aumento de palavras consideradas em cada mês. Note que a partir do mês de maio, esse valor tende a se estabilizar, não importando qual o percentil analisado. Isso sugere que existem palavras que são sempre mantidas como significativas e importantes para a inferência da localização dos usuários enquanto outras são renovadas e atualizadas.

Para demonstrar de forma qualitativa esse fenômeno, na Tabela 3.2, listamos as 10 palavras de maior média de χ^2 que se mantiveram durante todos os meses no primeiro percentil. Essas palavras são nomes de cidades e estados que, obviamente, não possuem caráter temporal. Enquanto isso, a lista das 10 palavras mais descritivas de agosto de 2013 **que não se mantiveram** no primeiro percentil nos demais meses contém não só nomes de lugares, mas se relacionam também a eventos como o **zfestival**, a vinda da banda **emblem3** para o Brasil e a um show do cantor Luan Santana (vide o termo



(a) Na base Geo, menos de 40% das palavras são mantidas durante um ano como as mais descritivas de acordo com o teste χ^2



(b) Na base Geo+NotGeo, menos de 30% das palavras são mantidas durante um ano como as mais descritivas de acordo com o teste χ^2 para maior parte dos percentis

Figura 3.17: Evolução do conjunto de termos com maior χ^2 em cada mês

luansantanahojebarretos).

3.6 Ground truth

A cidade onde o usuário vive é uma informação desconhecida que desejamos inferir. No entanto, precisamos de definir uma forma de associar os usuários a cidades para

Agosto de 2013	**Sempre mantidas**
dourados	macapa
cachoeirinha	belem
zfestival	manaus
mtvhottest	recife
araci	maceio
bieber	fortaleza
emblem3	natal
luansantanahojebarretos	rn
realizem	teresina
mcduduzinholindobrabogostosoetesao	lajeado
gurupi	Curitiba

Tabela 3.2: Palavras mais descritivas em Agosto de 2013 e palavras que sempre se mantiveram como descritivas durante o período

termos um conjunto para treino e avaliação de método.

Em todos os trabalhos anteriores, os autores de alguma forma relacionam os usuários à cidade mais comum entre os seus tweets como o ponto principal ou de origem do usuário - sendo esse o *ground truth* do experimento.

Por termos uma base de dados que se expande pelo período superior a um ano, podemos observar como os usuários mudam de acordo com o *ground truth*. Para tanto, selecionamos os usuários que pubicaram no mínimo três tweets geolocalizados na nossa base. Determinamos sua cidade mais frequente durante o ano inteiro e mostramos na Tabela 3.3 a concordância de cada mês em relação ao *ground truth*. Podemos observar que coerentemente com os resultados obtidos na seções anteriores, por volta de 80-92% dos usuários tem como cidade mais frequente em cada mês a cidade também descrita como *ground truth*. Isso corrobora a hipótese de que usuários tendem a se movimentar de forma limitada durante o ano e que as cidades onde residem continuam sendo as mais frequentes na base de dados, mesmo que eles possam viajar e publicar tweets geolocalizados a partir das cidades que visitam.

Jan	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
89%	85%	88%	86%	94%	87%	85%	90%	91%	92%	84%	82%

Tabela 3.3: Concordância entre o local mais frequente entre os tweets de cada usuário por mês e mais frequente durante o ano

Capítulo 4

Métodos para inferência da localização do usuário

Neste trabalho, definimos como a localização do usuário aquela onde o usuário reside e/ou passa a maior parte do seu tempo e de onde há, portanto, a maior probabilidade de que ele publique novos tweets. A localização real do usuário é definida como nosso *ground truth* é aquela mais comum entre os tweets geolocalizados do usuário na base inteira.

Dados um usuário e suas informações públicas, exceto sua localização exata no momento da publicação de seus tweets, queremos que um método de inferência lhe atribua uma localização que se aproxime tanto quanto possível de sua localização real. Este capítulo discute diversos métodos já propostos na literatura para inferência e localização baseados em diferentes atributos. Nosso objetivo é avaliar se, ao utilizar esses métodos considerando variações temporais, os resultados podem ser melhorados. Além disso, propomos maneiras de considerar conjuntamente as inferências feitas pelos diferentes métodos através de técnicas de meta-aprendizado. Por último, discutimos também um dos desafios enfrentados por esses métodos: a escassez de dados em algumas localidades e propomos uma nova forma de lidar com o problema.

4.1 Inferência a partir das relações entre usuários

Existem na literatura métodos que trabalham tanto com a rede de menções [Jurgens et al., 2015] quanto com a rede de amizades [Davis Jr et al., 2011], como mostrado na Seção 2.1.3. Entendemos que a rede de menções é mais fácil de ser obtida. No entanto, gostaríamos de descobrir as vantagens de uma rede sobre a outra, sendo consumidas pelos mesmos métodos. Escolhemos quatro algoritmos de inferência de localização

baseada em rede, por utilizarem técnicas distintas. A descrição de todos pode ser encontrada resumidamente na seção 3.3.

1. Davis Jr et al. [2011], por se tratar de um dos algoritmos de menor custo computacional, mais simples mas que apresentou resultados robustos no trabalho original (**VotoVizinhança**);

2. Rout et al. [2013], que utiliza uma estratégia de classificação de usuários utilizando SVM e atributos da rede e, de agora em diante denominado (**Wheres**);

3. Backstrom et al. [2010], que utiliza um modelo probabilístico para determinar a localização do usuário (**FindmeMethod**);

4. Kong et al. [2014], que foi um trabalho inicialmente desenvolvido utilizando a rede de menções que atribui pesos diferentes para as localizações de cada usuário na rede de amigos baseado no conceito de proximidade social (**Spot**).

Nos trabalhos citados que utilizam a rede de mencões, um usuário se conecta ao outro se ambos se mencionam um número de vezes acima de determinado limiar em qualquer momento do período de coleta das bases. No nosso trabalho, analisamos se o modelo criado muda caso sejamos mais ou menos estritos quanto ao tempo mínimo entre uma menção de um usuário ao outro. Queremos saber, por exemplo, se ao restringirmos que deve haver interação entre dois usuários no período de 1 dia para haver uma conexão entre eles obtemos resultados diferentes dos modelos em que não há restrição de tempo entre uma interação e outra para que a conexão seja considerada.

4.2 Inferência textual

Muitos temas discutidos por usuários no Twitter possuem caráter local, envolvendo assuntos como esporte, política e entretenimento. Usuários estão interessados naquilo que lhes afeta, como condições climáticas, decisões políticas, eleições municipais, estaduais etc. Da mesma forma, usuários podem fazer referência a locais de seu interesse ou descrever eventos que aconteceram em determinados locais conhecidos por outros membros da comunidade que conhecem a região.

Dessa forma, não seria surpreendente que pudéssemos identificar a localização de um usuário de acordo com conteúdo de suas postagens. Um ser humano que conheça assuntos relacionados a determinado local e alguns de seus pontos de interesse - como praças, shopping centers etc - poderia utilizar essa informação para inferir a localização de um usuário somente analisando seus tweets. Como exemplo, temos o tweet ilustrado

na Figura 4.1, em que a usuária referencia um local que leva no nome a denominação de uma região bastante conhecida de Belo Horizonte: a região da Pampulha. Além disso, "Promove Pampulha é um ponto de referência na região.

Figura 4.1: Locais de referência são citados mais frequentemente em tweets na própria cidade

Algumas palavras, expressões idiomáticas e gírias são mais comuns em determinadas regiões e são indicativas da origem dos usuários. Obviamente alguns usuários podem mover-se de cidade e mudar de região levando consigo as especificidades de fala do seu lugar de origem, mas considerando a massa de usuários que utilizam o Twitter, esses usuários seriam a exceção e não a regra. Na Figura 4.2, podemos ver um usuário que se utiliza de um expressão tipicamente mineira: "Uai"e, não surpreendentemente, seu tweet foi publicado de uma cidade no estado de Minas Gerais.

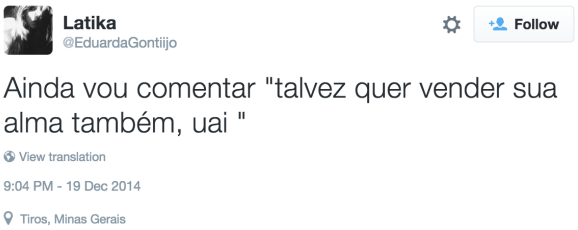


Figura 4.2: Expressões idiomáticas podem ser utilizadas para identificar a localização do usuário

Para nos utilizarmos do texto presente nos tweets, podemos criar modelos que levam em consideração as diferenças na forma de se expressar e nos assuntos de cada região. Com a abundância e crescimento cada vez maior de dados disponíveis, a utilização de modelos criados por técnicas e algoritmos de aprendizado de máquina são cada vez mais frequentes para solucionar problemas conhecidos de linguagem natural, como classificação de textos em assunto [Lee et al., 2011], análise de sentimentos [Ribeiro Jr et al., 2012] etc. Uma maneira de enxergar nosso problema de forma que possamos tirar vantagem da massa de dados que possuímos é como um problema de classificação multi-classes [Manning & Schütze, 1999].

Um problema de classificação multi-classes pode ser definido da seguinte maneira: dada uma instância e seus atributos, essa instância participa de uma e somente uma das n possíveis classes, sendo $n > 2$.

Analogamente, definimos o nosso problema de inferência como:

1. Cada classe corresponde a uma região ou localização dentre as possíveis para o universo de usuários. Temos, portanto, um conjunto de classes possíveis L que representam os locais possíveis onde se localizam os usuários

2. Cada usuário U_i é uma instância do problema retirado de um universo de usuários U. Cada usuário U_i participa somente de uma classe definida, ou seja, está associado a um único local.

3. O conteúdo textual do conjunto de tweets produzidos pelo usuário U_i é fornecido como atributo para nosso algoritmo de classificação. Caso a localização do usuário seja conhecida, esse usuário e seus atributos são incorporados no conjunto de treino a ser fornecido para o classificador. Um usuário com localização desconhecida tem seus atributos fornecidos ao classificador para inferência do possível local de onde ele publica suas mensagens.

4.2.1 Modelo linguístico estático e Modelo linguístico temporal

Em trabalhos anteriores, a criação de modelos linguísticos para inferência de geolocalização é feita através de uma massa de tweets coletados a partir da localização informada pelo usuário ou tweets geolocalizados. Modelos linguísticos locais feitos a partir de um conjunto de tweets - sem levar em consideração o momento em que foram gerados - tendem a captar assuntos que são comuns na região em que os tweets foram publicados. Uma das hipóteses que motiva esse trabalho é a de que um modelo linguístico de valor temporal pode ter maior poder de predição.

No tweet mostrado na Figura 4.3a, o nome do prefeito da cidade de Belo Horizonte é citado. Intuitivamente, espera-se que, a menos que ele seja notícia estadual ou nacional, seu nome seja mais frequente em tweets de usuários que moram/visitam a cidade, uma vez que ele tem impacto direto na mesma. No mesmo tweet, o usuário cita um evento: O FIQ (Festival Internacional de Quadrinhos), que aconteceu no mês de novembro de 2013 em Belo Horizonte (http://www.fiqbh.com.br/) - e portanto, um evento de caráter temporal e local. Nossa hipótese é de que eventos locais e temporais tenham uma repercussão mais forte nos tweets dos usuários que moram/frequentam o local ou a região onde o evento acontece.

Exemplos de eventos que podem repercutir nas mídias sociais, incluem shows, eventos climáticos, desportivos, exposições de arte que acontecem na cidade, eleições (seguidas de nomes dos candidatos locais) etc. Na Figura 4.3 vemos tweets sobre um

show que acontecem em Salvador, a comemoração do título cruzeirense (time de Belo Horizonte) sob chuva e a exposição de obras de Escher que seguiu para outra cidade após ser exposta em Belo Horizonte.

(a) Prefeito da cidade é citado em um evento que acontece na cidade de Belo Horizonte

(b) tweets sobre artistas podem ser mais frequentes quando os mesmos visitam a cidade

(c) tweet sobre eventos climáticos e eventos desportivos de impacto local

(d) tweet sobre uma das exposições que aconteciam na cidade de Belo Horizonte

Figura 4.3: Tweets com informações geográficas

Tais tweets - que tratam de eventos de caráter temporal - contêm informações que podem auxiliar na criação de um modelo linguístico que tem um tempo limitado de vida. Afinal, não chove sempre em Belo Horizonte, títulos podem ser comemorados em outros eventos desportivos e exposições e artistas viajam por diferentes cidades. O prefeito da cidade, no entanto, será Márcio Lacerda por alguns anos e a região da Pampulha continuará a ser citada frequentemente.

Dessa forma, temos que palavras que são mais frequentes durante determinado tempo em uma cidade do que em outra, mas que, com o passar do tempo, não se tornam tão significativas para predizer a cidade da qual o tweet se origina. Durante nosso projeto, trabalharemos com essa hipótese para criação de modelos linguísticos temporais para inferência de localização do usuário e compararemos seu desempenho com um modelo linguístico que não leva em conta a dimensão de tempo.

Modelo Estático: Como em trabalhos anteriores, trabalhamos com a premissa de que um conjunto de tweets pode ser utilizado para gerar um modelo para inferir corretamente a localização de um novo usuário a partir do conteúdo de seus tweets, independente da data de publicação dos mesmos. O funcionamento desse método segue como:

- **Fase de treino**: Dada uma base de treinamento contendo os usuários U, cada instância de treino fornecida para o algoritmo de aprendizado de máquina consiste no texto composto por todos os tweets T_i do usuário U_i associados com a classe ou localização determinado para o usuário L_i.

- **Fase de inferência**: Dado um usuário u_i de localização desconhecida e o seu conjunto de tweets t_i, a instância fornecida ao modelo gerado na fase de treino consiste no texto composto por todos os tweets t_i e a localização inferida é aquela sugerida pelo modelo.

Modelo temporal Neste método, trabalhamos com a hipótese de que modelos gerados com os tweets publicados no período de tempo $[t_{inicial}, t_{final}]$ possuem maior poder de predição que os tweets de localização desconhecida também criados dentro dessa janela de tempo. Como mostramos na tabela 3.2, a importância de termos descritivos para cada região muda a cada mês. O funcionamento desse método segue como:

- **Fase de treino:** A nossa base de usuários para treino contém o conjunto de usuários U e um conjunto de tweets T gerados no intervalo de tempo $[t_0, t_n]$. Cada tweet t é associado a um usuário u e cada usuário associado a um local. O período de tempo $[t_0, t_n]$ é dividido em d *partições* de tamanho igual a $\Delta dias$ e denominados $t_0, t_1, ..., t_d$. Cada tweet é atribuído a uma *partição* de acordo com sua data de criação. Dessa forma, temos nossa massa de tweets distribuída em d conjuntos. Para cada partição, agrupamos os tweets disponíveis de acordo com seu autor e utilizamos tais instâncias como treino associadas às localizações dos usuários aos quais eles pertencem para criação de d modelos: um para cada partição de tempo.

- **Fase de inferência:** Temos d classificadores, cada qual para um intervalo de tempo diferente. Dado um novo usuário de localização indefinida, seus tweets são distribuídos para cada uma das d partições de tempo de acordo com sua data de publicação. Cada modelo sugere uma localização para o usuário e essa é definida como aquela votada pela maioria dos modelos.

Note que para esse método, assim como para o método estático, supomos que a localização definida para o usuário durante o intervalo de tempo $[t_0, t_n]$ é a mesma. Isso vai de encontro ao que observamos na Tabela 3.3, que mostra que no período de um ano a localização mais frequente de cada usuário em cada um dos meses é a mesma que a localização mais frequente do usuário durante o ano em mais de 80% dos casos.

4.2.2 Criação de modelos linguísticos: seleção de atributos e treino

Nossa tarefa de criação dos nossos modelos linguísticos, sejam temporais ou estáticos, está divida em três fases: **limpeza ou pré-processamento do texto, seleção de**

atributos/termos e escolha do classificador. A seguir, descrevemos cada uma dessas fases.

1. **Pré-processamento do texto**

   Uma das principais características do Twitter é a sua limitação quanto ao número máximo de 140 caracteres que cada publicação pode conter. Dessa forma, usuários utilizam-se da criatividade para conseguir transmitir a mensagem desejada frente a essa limitação: utilizando-se de abreviações, omissão de pontuação e até palavras inteiras da sentença. Usuários expressam-se na plataforma das mais variadas maneiras: utilizando-se desde a forma padrão da língua até o uso intencional ou não de palavras e sentenças que fogem da norma culta. Portanto, o primeiro passo para criação de um conjunto de atributos a ser fornecido para o classificador que gerará nosso modelo é remover termos irrelevantes, seja por sua pouca utilização ou por irrelevância para a tarefa de classificação, através de:

   - Remoção de ***stopwords***. Cada língua possui um conjunto de palavras que possuem alta frequência em qualquer tipo de texto. Essas palavras geralmente são preposições, pronomes, mas também podem incluir substantivos, adjetivos e verbos que são amplamente utilizados. Essas palavras são conhecidas por aumentarem inutilmente o número de atributos que são fornecidos para classificadores, uma vez que não possuem qualquer poder preditivo. No nosso método, utilizamos uma lista de ***stopwords*** da língua portuguesa e retiramos tais palavras dos tweets da nossa base.
   - Normalização da forma como as palavras são escritas através da retirada de acentos e *cedilhas* de qualquer palavra. Como usuários podem escrever a mesma palavra de formas diferentes e não têm a obrigação de seguir a norma culta da língua, normalizamos todas as palavras. Aqui podemos potencialmente incorrer em erros, dado que algumas palavras de significados diferentes podem ser escritas da mesma forma com a possível diferença de um sinal gráfico.
   - Remoção de termos que ocorram abaixo de um limiar de frequência nos textos da base. Aqui, definiremos um limiar experimental para o qual uma palavra será removida. Dessa forma, tentamos retirar termos que aparecem simplesmente por não seguirem a norma ortográfica comum no Twitter, intencionalmente ou não, e termos irrelevantes.

2. **Seleção de atributos**

Nem todas as palavras provenientes da primeira etapa são igualmente importantes para discriminar as localizações. É conhecido que, para executar tarefas de classificação de texto, um grande número de palavras disponíveis pode degradar a eficácia do algoritmo de aprendizagem [Manning & Schütze, 1999]. Para melhorar a performance dos nossos algoritmos, podemos, primeiramente, efetuar uma seleção das palavras que melhor discriminam cada uma das localizações candidatas.

No nosso trabalho, escolhemos o teste χ^2 para seleção de atributos. No teste χ^2, como explicado na seção 3.5, utilizamos a estatística χ^2 para ordenar as palavras de acordo com a sua dependência com o atributo de classe que, no nosso caso, é a localização do usuário. Utilizamos as N palavras com maior valor. O valor de N é estudado empiricamente.

3. **Escolha do Algoritmo para classificação**

   Existem vários algoritmos de aprendizado de máquina que podem ser aplicados para a tarefa de classificação multi-classe de textos. No entanto, muitos algoritmos de aprendizagem não são apropriados para essa tarefa em particular por razões de escalabilidade. O algoritmo Support Vector Machine [Cortes & Vapnik, 1995], por exemplo, não é indicado para tarefas que possuem um número massivo de classes [Han et al., 2014]. Além disso, gostaríamos de ter um algoritmo que possa ser facilmente retreinado para incorporar novos dados provenientes do *stream* do Twitter.

   Com isso em mente, escolhemos estudar a aplicabilidade dos seguintes algoritmos ao nosso problema:

   - **Naive Bayes Multinomial**, que implementa o algoritmo Bayes para dados multinomialmente distribuídos. É uma das variações clássicas do algoritmo de Naive Bayes utilizado na classificação de texto e que, apesar de considerar a independência dos atributos, traz bons resultados em tarefas de classificação de texto [McCallum et al., 1998] . A distribuição é parametrizada pelo vetor $\theta_y = (\theta_{y1}, ..., \theta_{yn})$ para cada classe y onde n é o número de atributos (que no nosso caso é o número de palavras selecionadas pela seleção de atributos do item anterior) e θ_{yi} é a probabilidade $P(x_i|y)$ do termo i aparecer na amostra que pertence a classe y.
   - **Regressão logística**, que é um modelo de classificação em que as probabilidades geradas são modeladas de acordo com uma função logística [Fan

et al., 2008]. A regressão logística mede a relação de dependência entre uma classe (cidade) e uma variável geralmente contínua.

4.3 Utilizando atributos do perfil do usuário

A meta-informação que acompanha os tweets é uma importante fonte de informação geográfica a ser considerada. Nesse método, escolhemos dois dados fornecidos pelo próprio usuário:

1. Localização declarada
2. Descrição do perfil do usuário

Todas essas informações fornecidas pelo usuário podem ser imprecisas ou errôneas, uma vez que este tem liberdade de fornecer qualquer informação como forma de texto livre (para localização declarada e descrição) ou que achar coerente ou útil.

Essas informações também diferem com respeito à sua utilidade. A localização declarada pode indicar corretamente e de forma explícita à localização do usuário, enquanto a descrição pode simplesmente fazer referência à localização do usuário ou a fatos relacionados ao local onde o usuário mora.

O usuário da Figura 4.4c faz referência estruturada à cidade de Belo Horizonte. Uma das formas utilizadas para referir-se à cidade é através da sigla **BH** ou **B.H.**. Neste caso, o usuário também fornece o estado em que está, através da sigla do estado **MG**. Essa informação pode ser importante em casos em que há uma cidade com o mesmo nome em outros estados do país.

Na Figura 4.4b, vemos um usuário que inclui na sua descrição de perfil o nome da cidade onde trabalha como vereador. Neste caso, a descrição faz uma referência direta à cidade. Pode-se pensar que, dado que o usuário está se descrevendo, ele deve referir-se diretamente a uma cidade somente se possui ligação estreita a ela, como no caso em que more ou tenha nascido na referida cidade.

Na Figura 4.4a, vemos um usuário se definindo como "atleticano", denominação dada ao torcedor que torce para o time de futebol chamado "Atlético". Existem diversos times de futebol com esse nome no Brasil, sendo que os de maior torcida são, pela ordem: Atlético Mineiro, Atlético Paranaense e Atlético Goianiense. Sabendo desse fato, podemos inferir com alguma margem de erro que o usuário é de Minas Gerais, Paraná ou Goiás.

Mas como utilizar tais informações fornecidas pelo usuário de forma voluntária? Abaixo são descritos os métodos utilizados nos nossos experimentos:

(a) Usuário que faz referência a um time de futebol da cidade de Belo Horizonte

(b) Usuário que faz referência ao nome da cidade na sua descrição

(c) Usuário usa a abreviação do nome da cidade onde mora no campo de Localização

1. **Casamento exato de nome de cidades na localização fornecida pelo usuário**: a partir de um *gazetteer* - dicionário de nomes e informações geográficas - procuramos encontrar na localização fornecida pelo usuário o nome (e suas variações) de cidades existentes na área de interesse, ou seja, de onde julgamos que usuários possam estar publicando mensagens.

2. **Utilização de modelo criado a partir do texto do campo de localização dos usuários**: neste método, utilizamos o conjunto de usuários com localização conhecida como treinamento para um algoritmo de classificação supervisionada, para que este aprenda os padrões encontrados nas descrições dos usuários em cada região. Essa pode ser considerada uma variação dos métodos de inferência textual, aplicada no contexto das localizações dos usuários. Nossa hipótese é de que algoritmos de classificação possam aprender palavras que designam cidades de forma não padronizada, como *belzonte* para Belo Horizonte.

3. **Casamento exato de nome de cidades no texto de descrição do usuário**: como no método **1**, procuramos por nomes de cidades e suas variações no texto de descrição do usuário.

4. **Utilização de modelo criado a partir do texto da descrição dos usuários**: neste método, utilizamos o conjunto de usuários com localização conhecida como treinamento para um algoritmo de classificação supervisionada para que este aprenda quais os padrões encontrados nas descrições dos usuários em cada região. Essa é uma variação do método **2**, aplicada no contexto das descrições dos usuários ao invés de seus tweets.

Como mostrado na Seção 3.4, essas informações são dinâmicas e podem ser modificadas pelo usuário com o passar do tempo. Portanto, investigaremos como a atualização dos modelos em intervalos de tempo influencia na eficácia dos modelos treinados.

4.4 Combinação de modelos

Todos os métodos mencionados anteriormente possuem um grau maior ou menor de confiança para a inferência da localização do usuário.

Nesta seção, propomos a utilização de um combinação dos resultados obtidos pelos métodos descritos anteriormente no capítulo de forma a fornecer uma predição mais robusta da localização do usuário.

Combinaremos os resultados de três formas:

1. **Votação**: um das formas mais simples de combinar os resultados obtidos de modelos base. Neste caso, cada modelo contribui para decisão com seu voto e a localização geográfica com maior número de votos é a escolhida. No caso de empate, a localização do método que apresenta maior precisão é a escolhida durante a fase de validação após o treino é o escolhida.

2. **Votação ponderada**: no caso da votação simples, cada classificador tem direito a um voto e o peso de todos os votos é o mesmo. Na votação ponderada, cada modelo tem seu voto ajustado. No nosso trabalho, avaliamos duas formas de ajustar os pesos dos votos: utilizando a acurácia obtida pelo modelo (**Voto-acc)** em uma partição de validação e através de um algoritmo genético (**Voto-GA)**. Para o ajuste de pesos usando o algoritmo genético, seja N o número de modelos base considerados (C). Cada indivíduo da nossa população no algoritmo genético é codificado como um vetor de números reais de tamanho N, onde a posição i do vetor ($1 \leq i \leq N$) é o peso do voto do classificador $c_i \in C$. Nosso objetivo é minimizar a taxa de erros das classes inferidas a partir do voto. Algoritmos genéticos [Holland, 1975] têm se mostrado um método robusto e prático de otimização. Um algoritmo inicia com um conjunto de indivíduos que representam possíveis soluções para o problema em questão. Essa população inicial pode ser criada aleatoriamente ou não. O valor da função objetivo (*fitness function*) de cada indivíduo é avaliado e o algoritmo manipula um subconjunto da população na busca de melhores resultados através de operações genéticas como mutação e cruzamento criando uma nova geração da população. Uma das vantagens do uso de pesos ajustados para otimizar a votação, é a possibilidade de interpretar os pesos dados a cada algoritmo como sua confiança, por exemplo.

3. **Meta-árvore de decisão**: induzimos uma meta-árvore de decisão com atributos utilizados para medir a confiança de cada um dos modelos - chamados atributos ordinários - além da localização inferida por cada um dos modelos - chamado

de atributos de classe. A meta-árvore de decisão, ou **MDT**, foi proposta por Todorovski & Džeroski [2000] como uma nova forma de combinar o resultado de diferentes classificadores. As **MDTs** possuem estrutura idêntica às das árvores de decisão (**ODT**): cada um de seus nós internos especifica um teste a ser realizado sobre um atributo do conjunto de itens a serem classificados. Cada resultado do teste separa esse conjunto inicial em sua própria sub-árvore.

A MDT possui dois tipos de atributos: os **atributos ordinários** e os **atributos de classe**. Os atributos ordinários são aqueles relacionados a descrever as saídas dos classificadores e/ou os atributos dos itens a serem classificados e os atributos de classe são aqueles preditos pelos classificadores. Cada teste dos nós internos busca dividir a base de forma que as sub-árvores possuam a maior acurácia por classificador predominante possível, dada por:

$$info(S) = 1 - \max_{c \in C} acuracia(c, S), \tag{4.1}$$

onde S é o conjunto de exemplos e C o conjunto de classificadores base utilizados. Diferentemente das ODTs que atribuem uma classe às suas folhas, as MDTs atribuem classificadores às folhas de sua árvore e, portanto, não inferem diretamente a classe de um novo item a ser testado.

Na Figura 4.5, exemplificamos uma MDT que deve escolher entre o resultado de dois classificadores: o KNN e o Naive Bayes. Os atributos ordinários utilizados pela árvore são a probabilidade atribuída a classe escolhida pelo KNN (knn_prob) e a probabilidade atribuída pelo algoritmo Naive Bayes (naive_prob). As folhas indicam qual classificador deve ser utilizado para cada instância do problema.

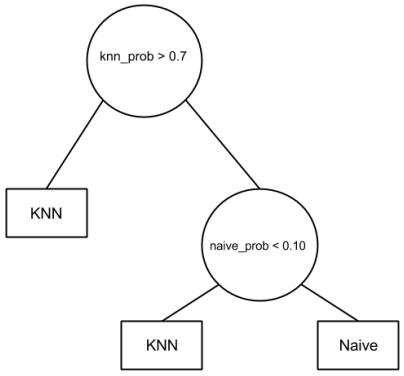
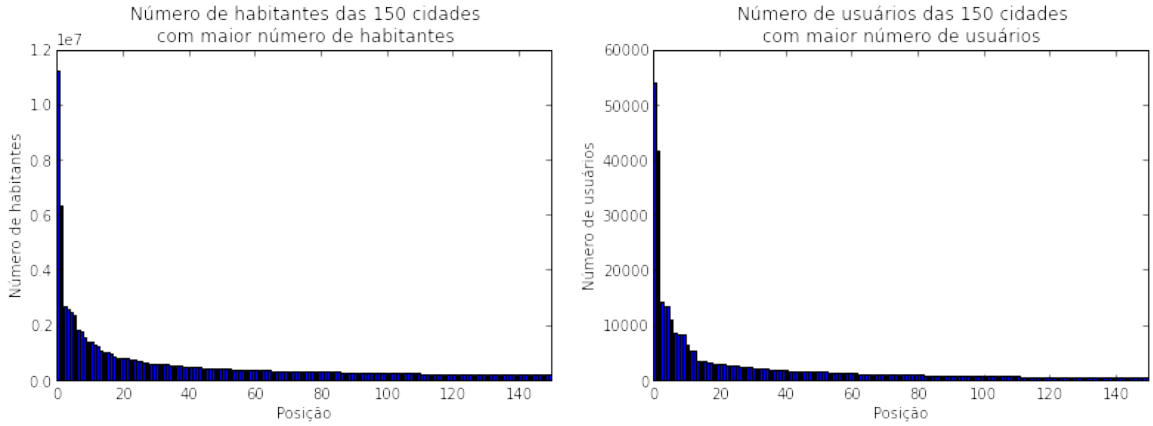


Figura 4.5: Exemplo de MDT: a decisão entre o classificador KNN e Naive Bayes é feita levando em consideração a probabilidade atribuída à classe por cada um

4.5 Desafio: escassez de dados

Um dos desafios enfrentados ao inferir a localização de um usuário é a esparsidade de tweets em regiões pouco populosas. Como observamos nas Figuras 3.4 e 3.5, parece existir uma relação entre a população e o número de usuários que publicam a partir da cidade. Dessa forma, cidades pouco populosas tendem a ter poucos ou nenhum tweet geolocalizado.

Para o Brasil, por exemplo, temos 5.507 cidades, com números de habitantes e usuários variados. O número total de habitantes do país, segundo o censo de 2010, é de mais de 190 milhões de pessoas e o número de usuários na nossa base soma pouco mais que 600 mil. Como podemos observar nas Figura 4.6, as distribuições de habitantes e usuários seguem ambas um decaimento exponencial. O número de usuários é bastante reduzido para para algumas cidades. A 150^a cidade com maior número de usuários possui apenas 453 usuários, enquanto a 150^a cidade mais populosa possui mais de 182 mil habitantes.



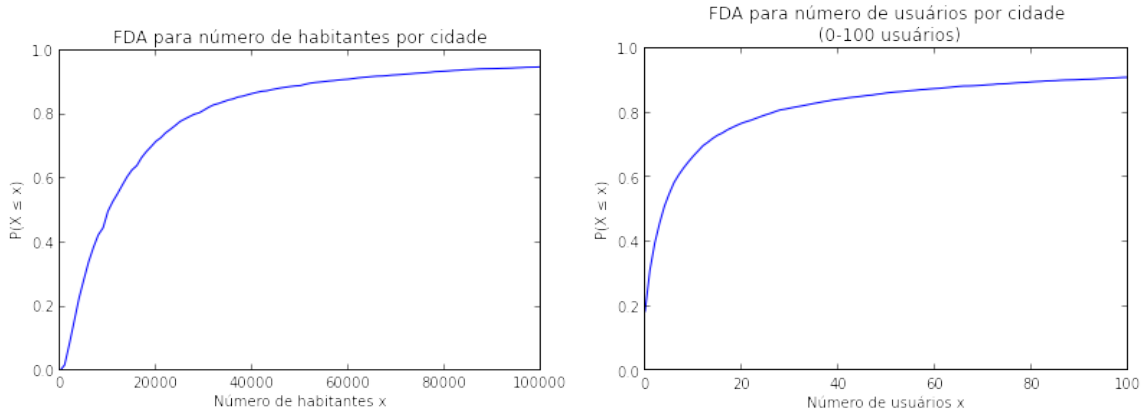
(a) Mesmo com decaimento rápido, a 150^a cidade mais populosa possui 182mil habitantes

(b) Enquanto a 150^a cidade com mais número de usuários geolocalizados possui apenas 453 usuários geolocalizados

Figura 4.6: Distribuição de usuários e habitantes por cidade

Das 5.507 cidades, mais de 60% possui menos de 20 usuários, como observado na Figura 4.7b. Da mesma forma, mais de 60% das cidades possuem no máximo 20 mil habitantes, de acordo com a função de distribuição acumulada (FDA) traçada na Figura 4.7a.

Dessa forma, não é surpreendente que a maior parte dos usuários geolocalizados no Brasil estejam concentrados em poucas cidades. Como podemos ver na Figura 4.8, as 100 cidades com maior número de usuários geolocalizados abrigam mais de 60% dos usuários da nossa base.



(a) Mais de 60% das cidades possui no máximo 20mil habitantes

(b) Mais de 80% das cidades possui no máximo 100 usuários geolocalizados

Figura 4.7: Distribuição de usuários e habitantes por cidade

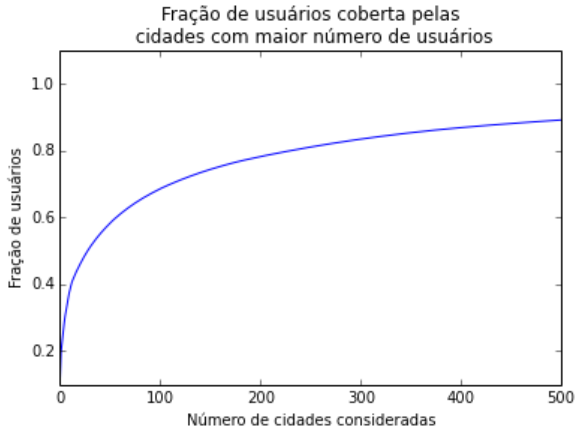


Figura 4.8: As 500 cidades com maior número de usuários geolocalizados abriga mais de 80% dos mesmos

Uma forma de compensar a pequena quantidade de usuários existente em várias cidades consiste em agrupá-las de forma que elas formem um conjunto maior - que chamaremos daqui em diante de ***agrupamento*** - considerando-as como uma cidade. A partir de então, qualquer usuário que tenha sua localização em uma das cidades pertencentes a determinado agrupamento terá suas coordenadas redefinidas sendo a daquele agrupamento. Dado um conjunto de agrupamentos G , o agrupamento g_j contém um conjunto de cidades C_j, cada cidade c_i pertencente a C_j tem sua representação geográfica dada por um ponto com coordenadas $latitude_i, longitude_i$. As coordenadas do agrupamento coincidem com o centro do agrupamento g_j que é dado pelo centro de massa do conjunto de coordenadas correspondente ao seu conjunto de cidades.

Antes de detalhar os métodos que consideramos para agrupar as cidades, defini-

mos algumas características para os agrupamentos:

1. **Distância máxima das cidade do agrupamento até o centro do agrupamento a que pertencem**. Dadas as cidades de um agrupamento, as distâncias das suas cidades até o seu centro são determinantes para expressar o erro na localização de um indivíduo. Ao definirmos que um usuário posta mensagens a partir do agrupamento g_j, estamos na verdade dizendo que ele pode estar em qualquer uma das cidades C_j do agrupamento, mas para a finalidade de localização o posicionaremos no centro do agrupamento. Dessa forma, uma vez que o usuário foi atribuído ao agrupamento que contém sua cidade, o maior erro esperado quanto a distância é igual a maior distância (d_{jmax}) entre o centro de g_j e alguma das suas cidades. Se $d_{jmax} = 0$, temos um agrupamento contendo apenas uma cidade. Valores pequenos d_{jmax} diminuem o erro de distância entre a localização definida para o usuário (como agrupamento) da sua localização real.

2. **Distância média ponderada entre as cidades do agrupamento e o seu centro**. Cidades com maior número de usuários devem estar mais próximas dos centro do agrupamento do que as cidades com menor número de usuários. Para entender melhor, suponha uma situação em que a cidade A tenha 100 usuários e a cidade B tenha 10 usuários. Se a cidade A está a 100km de distância do centro do agrupamento e a cidade B a 10km do centro do agrupamento, o erro de distância total é $100 \times 100km + 10 \times 10km = 10.100km$. No entanto, se temos uma situação contrária, o erro de distância seria $100 \times 10km + 10 \times 100km = 2000km$. Dessa forma, a distância média ponderada pelo número de usuários é um outro indicativo da qualidade do conjunto de agrupamentos criados.

3. **Tamanho do agrupamento**. Agrupamentos muito grandes ou muito pequenos podem atrapalhar na inferência da localização de futuros usuários. Um agrupamento com poucos usuários pode não ter exemplos suficiente para generalizar um modelo de inferência. Usaremos o coeficiente de Gini para medir o quão bem distribuídos estão usuários entre os agrupamentos. O coeficiente de Gini mede a desigualdade entre os valores de uma distribuição e é bastante utilizada em economia para medir níveis de desigualdade, como os de salário, por exemplo. O coeficiente de Gini varia no espaço real $[0, 1]$, onde 0 representa uma situação em que a distribuição de usuários é igual a distribuição uniforme e 1 representa a situação extrema em que um agrupamento possui todos os usuários e os demais agrupamentos não possuem nenhum.

No nosso trabalho, estudamos duas formas de agrupar as cidades. A primeira é encontrada no artigo de Han et al. [2014] e a segunda foi concebida durante o nosso trabalho. Para ambas, consideramos a localização da cidade como aquela que representa a sua sede urbana.

- **Agrupamento por número mínimo de usuários/GroupMinUsers.** Este método é divido em duas etapas.

  1. Dado um conjunto de cidades C e um limite inferior mínimo de habitantes, Min_u, cada cidade com o número de habitantes maior do que Min_u é considerada um agrupamento. Temos um conjunto de agrupamentos G formado por essas cidades.
  2. As cidades restantes possuem menos usuários do que o valor Min_u. Elas são agrupadas com as cidades selecionadas no passo 1 cuja distância seja mínima.

- **Agrupamento por número mínimo de usuários e distância/GroupMinDistUsers.** Enquanto o método anterior favorece a formação de agrupamentos ao redor de cidades populosas, nosso método favorece a fusão de cidades menores, porém próximas, para formar agrupamentos mais homogêneos. Nosso método consiste em:

  1. Dado um conjunto de cidades C, considere cada cidade como um agrupamento.
  2. Agrupe todos os agrupamentos cujas distâncias de seus centros seja menor do que Min_d. Atualize os centros dos agrupamentos de acordo com seu novo centro de massa.
  3. Para cada agrupamento cujo número de usuários for menor do que Min_u e que não tenha sido fundido a outro durante o passo 3, funda-o com o agrupamento mais próximo. Atualize os centros dos agrupamentos.
  4. Repita os passos 2 e 3 até que todos os agrupamentos tenham um número de usuários igual ou maior a Min_u.

Experimentamos com várias configurações dos métodos descritos para entendermos como as características dos agrupamentos gerados são influenciados pelos parâmetros iniciais.

Inicialmente analisamos o comportamento de **GroupMinUsers**. Na Figura 4.9a vemos que a maior distância do centro de um agrupamento aumenta quando o parâmetro Min_u aumenta. Isso se deve ao fato de que um número muito pequeno de cidades

tem o número mínimo de usuários para iniciar um agrupamento. Com o crescimento do parâmetro Min_u, o número de agrupamentos diminui, como mostrado na figura 4.9e até o ponto em que temos apenas um agrupamento. Isso se reflete no menor coeficiente de Gini possível que é igual a zero, como mostrado na figura 4.9b, que diminui a medida que o número mínimo de usuários necessário para formar um agrupamento aumenta. Inicialmente, com um número pequeno, toda e qualquer cidade era seu próprio agrupamento, com um número variado de usuários. Enquanto queremos agrupamentos mais balanceados, queremos diminuir também a distância das cidades ao centro de seus respectivos agrupamentos. Vemos aqui um *trade-off* entre o coeficiente de Gini e a distância entre as cidades e o centro do agrupamento, já que pela Figura 4.9c e 4.9f vemos que a distância para o centro aumenta à medida que mais cidades são agrupadas. O número de usuários por agrupamento cresce, obviamente, à medida que o número mínimo de usuários aumenta.

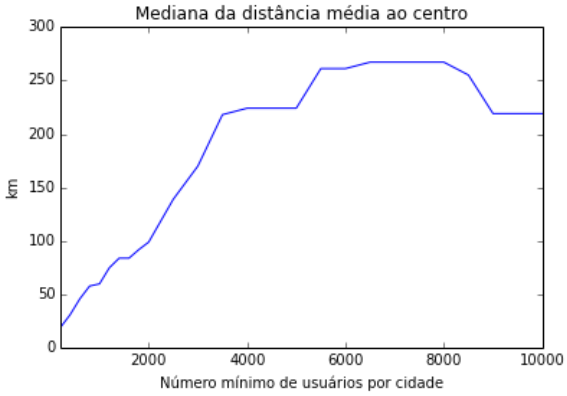
Para o segundo método, **GroupMinDistUsers**, temos duas variáveis a serem analisadas: o número mínimo de usuários e a distância mínima dos agrupamentos. Variamos o número mínimo de usuários de 200 a 10.000 e a distância mínima de zero até 2.000km. Pela Figura 4.10a, podemos ver claramente um retângulo onde as distâncias máximas são menores até o valor de até 2.000 usuários e 180km de distância. Ao contrário, os menores coeficientes de Gini estão nos extremos, onde o número mínimo de usuários no agrupamento é grande. Focamos nossas análises no primeiro retângulo para melhor visualização dos dados. Novamente, como podemos observar, quando a distância entre as cidades e o centro dos agrupamentos cresce (Figuras 4.11a e 4.11b), o coeficiente de Gini (Figura 4.11) e número de agrupamentos diminui (Figura 4.11c), assim como nos agrupamentos anteriores. No entanto, note que as medianas de distâncias máximas (4.11a) são menores do que o método anterior para coeficientes de Gini semelhante. Isso se deve ao fato de que agrupamos cidades próximas para formar agrupamentos com mais usuários, ao invés de simplesmente fundi-las com agrupamentos com um número de usuários grande. O número de usuários representando nas Figuras 4.11d e 4.11c, mostram que pequenas distâncias e pequenos valores para o número mínimo de usuários favorece a formação de agrupamentos com menor número de usuários o que não é desejável para se manter o balanceamento entre os agrupamentos.

Para avaliar os agrupamentos formados, levamos em consideração os seguintes critérios, já descritos anteriormente:

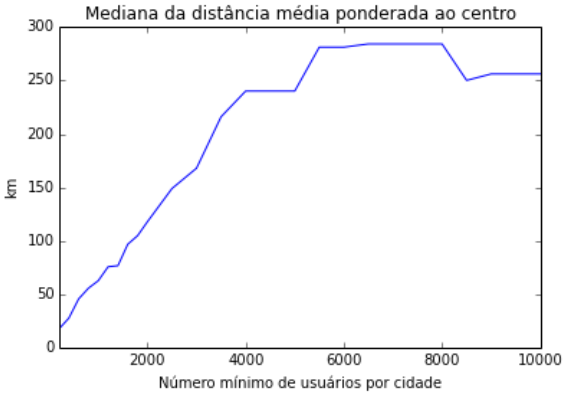
- Coeficiente de Gini (C_g): quanto mais próximo o coeficiente está de zero, melhor.
- Distância média entre as cidades e os centros dos seus agrupamentos (D_m). Aqui, normalizamos a distância média igualando-a a razão entre a mesma e a distância

(a) Desejamos valores pequenos para as distâncias máximas entre as cidades e o centro do seu agrupamento

(b) Coeficientes de Gini menores representam conjuntos de agrupamentos melhor balanceados.

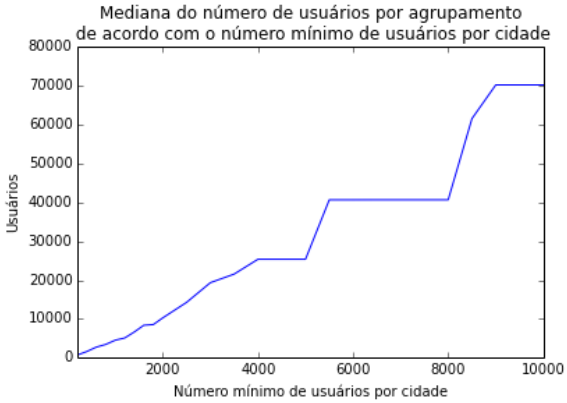


(c) Coeficientes de Gini menores representam conjuntos de agrupamentos melhor balanceados.



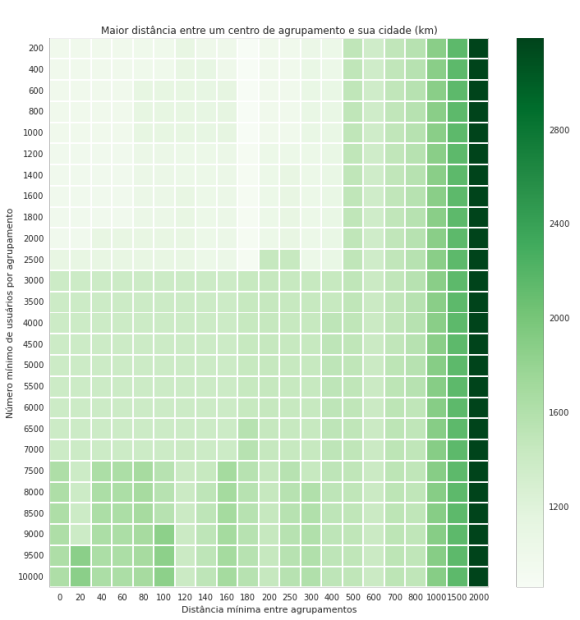
(d) Coeficientes de Gini menores representam conjuntos de agrupamentos melhor balanceados.

(e) Podemos controlar o número de agrupamentos calibrando o valor do número mínimo de usuários por cidade.

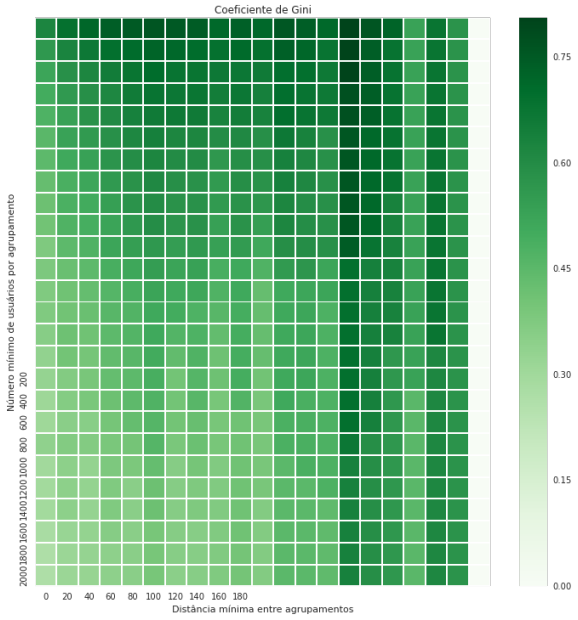


(f) Um menor número de agrupamentos significa um maior número de usuários por agrupamento.

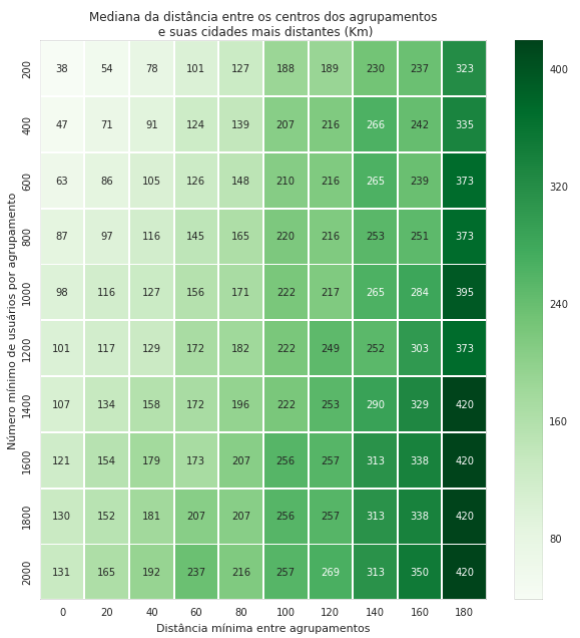
Figura 4.9: Comparação dos vários conjuntos de agrupamentos gerados



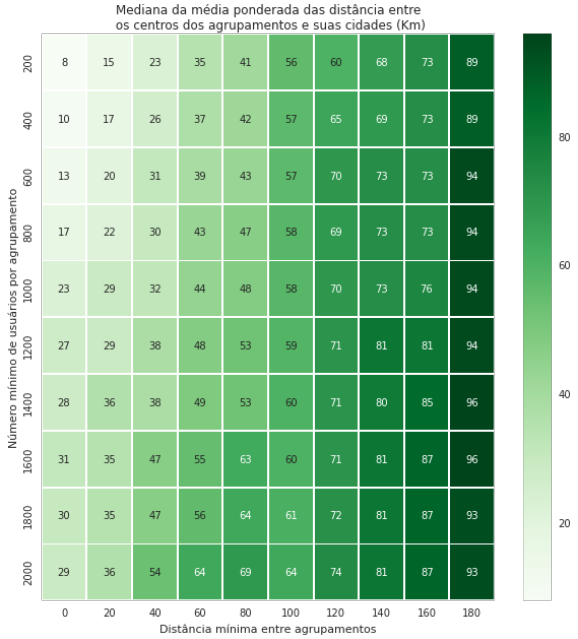
(a) Desejamos valores pequenos para as distâncias máximas entre as cidades e o centro do seu agrupamento



(b) Coeficientes de Gini menores representam conjuntos de agrupamentos melhor balanceados.



(c) Uma média não ponderada não descreve o erro real que ocorre, já que cidades com maior número de usuários terão erros mais frequentes

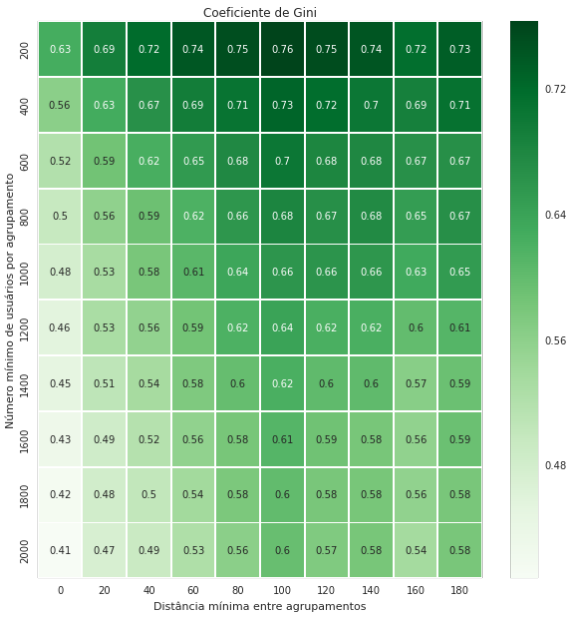


(d) Para privilegiar cidades com maior número de usuários, ponderamos as distâncias de acordo com o número de usuários das cidades

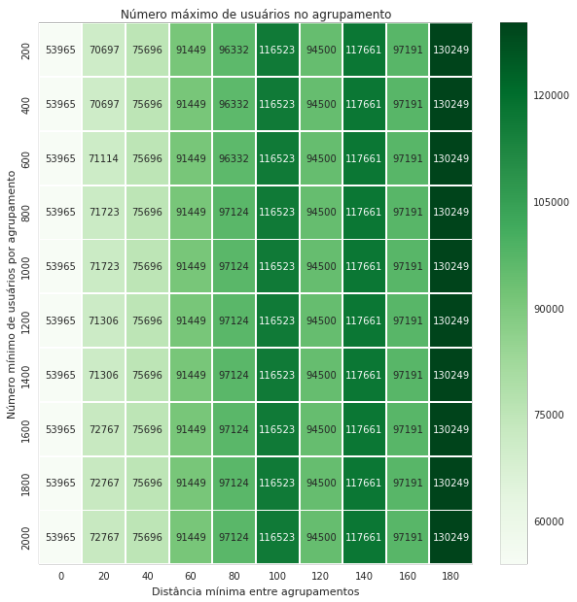
Figura 4.10: Comparação dos vários conjuntos de agrupamentos gerados

média máxima possível para a área analisada, isto é, quando temos apenas um agrupamento.

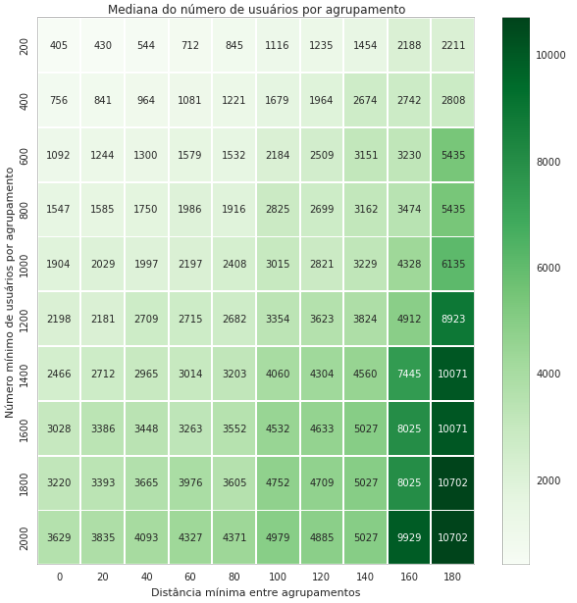
(a) Um número alto de agrupamentos indica com maior precisão onde os usuários se encontram, mas podem ser muito desbalanceados



(b) Coeficientes menores representam conjuntos de agrupamentos com distribuição de usuários mais próxima da uniforme



(c) Agrupamentos com um grande número de usuários são resultado da distribuição não uniforme de usuários entre as cidades



(d) Agrupamentos com um número pequeno de usuários são difíceis de serem bem discriminados

Figura 4.11: Comparação dos vários conjuntos de agrupamentos gerados

- Distância média ponderada entre as cidades e os centros dos seus agrupamentos (D_p). Novamente, normalizamos a distância média ponderada igualando-a a razão entre a mesma e a distância média ponderada máxima possível para a

área analisada, isto é, quando temos apenas um agrupamento.

Definimos uma medida de qualidade do conjunto de agrupamentos $Q(G)$ da seguinte forma:

$$Q(G) = \frac{aC_g + bD_m + cD_p}{a + b + c}$$

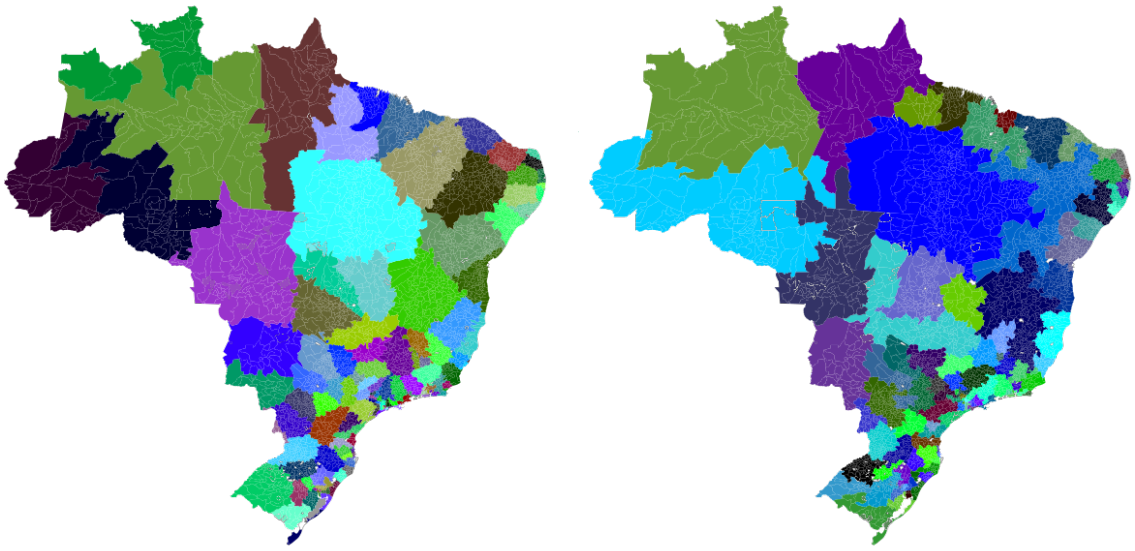
Onde a, b e c são coeficientes inteiros que representam a importância de cada um dos três fatores em relação aos outros. $Q(G)$ varia de zero até um, sendo que quando mais próximo de 0, melhor a qualidade dos agrupamentos gerados segundos os critérios estabelecidos através de a, b e c.

Variando o parâmetro do método 1 entre $[200, 10.000]$, e os parâmetros do método 2 entre $[200, 1000]$ usuários mínimos e $[0, 2000]$ km de distância, encontramos o valor que minimiza nossa métrica de qualidade para ambos os métodos, para os valores $a = 1, b = 2, c = 1$:

- **GroupMinUsers:** mínimo de usuários deve ser igual a 611, atingindo um score igual a 0.16. Foram criados 139 agrupamentos para esse caso.
- **GroupMinDistUsers:** mínimo de 1800 usuários e 3km, atingindo um score igual a 0.13. Foram criados 100 agrupamentos para esse caso.

Na Figura 4.12, podemos visualizar os agrupamentos formados pelos dois métodos. Note que no agrupamento **GroupMinDistUsers**, notamos no sudeste do Brasil agrupamentos maiores devido ao agrupamento de cidades menores ao invés de em torno dos grandes centros urbanos. Por outro lado, como observamos no nordeste e em parte do centro-oeste brasileiro, em **GroupMinDistUser** alguns agrupamentos tem formato irregular devido a estratégia de escolha e mudança do centro de agrupamentos.

Uma vez explicado como tentamos superar o desafio de esparsidade de tweets entre as regiões, explicaremos nas próximas seções as técnicas utilizadas para inferir a localização dos usuários.



(a) Agrupamento GroupMinUsers (b) Agrupamento GroupMinDistUsers

Figura 4.12: Comparação dos agrupamentos gerados pelas duas técnicas apresentadas

Capítulo 5

Experimentos e resultados

Este capítulo apresenta os resultados experimentais obtidos pelos métodos de inferência de localização apresentados no capítulo anterior sobre a base apresentada no Capítulo 3. Na Seção 5.1, apresentamos as métricas utilizadas para avaliar a eficácia dos métodos tanto em precisão quanto em relação à cobertura dos usuários da base de dados. Na Seção 5.2 apresentamos os resultados individuais de cada método base, e finalmente realizamos uma análise comparativa dos mesmos na Seção 5.3. Por fim, na Seção 5.4 combinamos os métodos base através de sistemas de votação e uma árvore de meta decisão.

5.1 Métricas de sucesso

Publicações anteriores obtiveram pouca concordância em métricas para avaliar a eficácia dos algoritmos comparados, reduzindo as possibilidades de comparar diretamente os resultados obtidos em diferentes trabalhos. Além do mais, as métricas existentes frequentemente testam diferentes capacidades dos algoritmos.

A métrica mais óbvia é a **acurácia**, que é dada pela fração de usuários cujas localizações foram corretamente inferidas. A acurácia não leva em consideração a distância entre a localização inferida e a localização real do usuário, o que a torna limitada para mensurar os erros cometidos pelos métodos. Além da acurácia, existem outras métricas bem conhecidas na literatura para avaliação de métodos de inferência de localização em redes sociais, levando em consideração principalmente o Twitter e Facebook.

Rout et al. [2013] produzem um *ranking* das cidades de acordo com a probabilidade de cada uma ser aquela de onde se originam os tweets de um usuário. A métrica **Precision@k** reporta a percentagem de localizações inferidas em que a loca-

lização correta está entre as ***k*** com maior probabilidade em um *ranking*. Quando k=1, a métrica é equivalente a acurácia - que é utilizada por Davis Jr et al. [2011]. Essa métrica não leva em consideração a distância entre a localização predita e a real do usuário.

McGee et al. [2013], por exemplo, utilizam a métrica **Acc@k** que mede a porcentagem de inferências feitas em até **k** unidades de distância. Essa métrica, diferentemente da anterior, possui a vantagem de levar em consideração a distância entre a localização predita e a real do usuário. No entanto, a métrica é sensível à escolha do **k**: valores baixos podem tornar a métrica semelhante a uma simples utilização do conceito de acurácia, enquanto valores altos de **k** tornam os resultados de dois algoritmos indistinguíveis, não importando o quão distantes as posições inferidas estão das localizações corretas dos usuários.

Kong et al. [2014] usam uma abordagem em que o erro é estimado a partir da média de distâncias entre a localização predita e a real do usuário. No entanto, essa métrica não é robusta o suficiente, uma vez que pode ser enviesada, por exemplo, por uma quantidade de erros (*outliers*) para usuários cuja distância é muito grande.

Priedhorsky et al. [2014] apresentam uma métrica baseada na confiança da inferência feita pelos algoritmos que considera a distribuição de probabilidade em uma área geográfica. No entanto, nem todos os métodos de inferência oferecem uma distribuição de probabilidades para suas predições o que, portanto, torna a métrica de difícil utilização para comparar vários dos métodos da literatura.

Em **Jurgens [2013]**, o autor reporta a performance dos algoritmos utilizando a função de distribuição acumulada (FDA) que mostra a porcentagem de inferências com erro menor do que x unidades de distância. A **FDA** pode ser vista como uma extensão da métrica **Accuracy@k**, uma vez que representa todos os valores de k.

No nosso trabalho, adotaremos a recente métrica publicada por Jurgens et al. [2015] que estende Jurgens [2013], atribuindo um valor estatístico para a FDA. A FDA atribui a probabilidade de um usuário ser classificado com um erro de distância menor ou igual a x por um método:

$$FDA(x) = P(distancia \leq x)$$

A área sob a curva (ASC) de uma FDA nos fornece uma forma de quantificar a performance geral dos métodos, onde aqueles que melhor aproximam os usuários de sua localizações reais possuem valores altos para $F(x)$ para valores pequenos de x. No entanto, ajustes sobre essas medidas devem ser feitas.

Primeiro, a penalidade sobre predições erradas deve escalar de formar inversa a

distância, uma vez que assumimos que a diferença de 50km, por exemplo, é mais discrepante quando comparamos 150km com 100km do que quando comparamos 1.000km com 1.050km. Para tanto, computamos a curva da FDA seguindo a escala logarítmica de x, e portanto:

$$FDA(x) = P(distancia \leq \log x)$$

Por outro lado, sabemos que a distância máxima de erro que um algoritmo pode cometer é limitada pela maior distância entres as cidades analisadas. Nós podemos, portanto, normalizar a media da ASC para que ela esteja no intervalo $[0, 1]$. No caso em que área normalizada tem o valor máximo de 1, o método inferiu corretamente todas as localizações do usuário. No caso em que a área tem valor normalizado igual a 0, temos um método que inferiu localizações o mais distantes possíveis da localização real do usuário.

No nosso trabalho, estamos tanto interessados na capacidades dos algoritmos que utilizam agrupamentos de inferir o agrupamento correto do usuário quanto na capacidade do método de agrupamento e método de inferência de posicionar o usuário tão perto quanto possível da sua localização real. Ao mesmo tempo, estamos interessados em descobrir qual a cobertura da base que cada método consegue atingir. Para tanto, utilizaremos as cinco métricas de avaliação, sendo as quatro primeiras relacionadas à precisão dos métodos e a última relacionada à cobertura de usuários conseguida por eles:

1. **ASC-g**: área sob a curva normalizada da FDA da distância entre o centro do agrupamento inferido e o agrupamento da localização real do usuário. O valor da **ASC-g** possibilita comparar a eficácia de dois métodos quanto às suas capacidades de inferir o agrupamento correto do usuário. Essa medida não é utilizada para os métodos de casamento exato do nome de cidades no campo de descrição e localização declarada, uma vez que ambos fazem a previsão com a granularidade de cidade e não de agrupamento.

2. **ASC-c**: área sob a curva normalizada da FDA da distância entre a localização inferida e a localização real do usuário. Para o caso dos métodos que utilizam agrupamentos de cidades, a localização inferida é o centro do agrupamento atribuído ao usuário. Agrupamos as cidades a fim de superar a escassez de dados para caracterizar áreas pouco populosas, mas idealmente queremos que a localização inferida esteja tão próxima quanto possível da sua localização real com granularidade de cidade. No entanto, **ASC-c** nos permite avaliar dois métodos e

seus agrupamentos utilizados quanto a distância da sua localização real e o centro do agrupamento para o qual a posição foi inferida.

3. **Acc@100**: fração dos usuários cuja localização inferida está a, no máximo, 100km de distância da sua localização real com granularidade de cidade. Para esta medida, não estamos interessados nos erros aproximados como em **ASC-c** e **ASC-g**, somente nas localizações corretamente inferidas. O valor de 100km, escolhido em trabalhos anteriores, permite que inferências feitas na mesma área metropolitana ou regiões semelhantes não sejam penalizadas.

4. **Mediana**: a mediana das distâncias entre as localizações inferidas e a localização do usuário. Embora pouco confiável por estar sujeita a valores extremos, a mediana nos oferece uma ideia intuitiva quanto à distância entre o local inferido e as localizações reais dos usuários.

5. **Revocação**: fração dos usuários da base cujas localizações foram inferidas. Enquanto alguns métodos podem alcançar bons resultados nas métricas relacionadas a precisão, eles podem não inferir a localização da maioria dos usuários.

5.2 Experimentos

Nas próximas seções, apresentaremos os experimentos feitos para cada uma das técnicas apresentadas e uma análise comparativa dos resultados obtidos em cada um deles individualmente e suas possíveis combinações. Para todos os experimentos, utilizamos a mesma configuração de validação cruzada utilizando 10 partições.

Utilizamos a base descrita no Capítulo 3 e avaliamos a eficácia dos métodos em duas formas de agrupamento de cidades: o GroupMinDistUsers e o GroupMinUsers. O agrupamento gerado por GroupMinDistUsers é composto de 100 regiões que agregam as cidades brasileiras, enquanto o agrupamento de GroupMinUsers é composto de 149 regiões. A escolha dos parâmetros para os métodos de agrupamentos foi descrita na Seção 4.5. Dos aproximadamente 601mil usuários iniciais, removemos aqueles que possuem menos de três tweets ou cujas velocidades de movimentação entre um ponto e outro tenham sido superiores a 800km/h (velocidade aproximada de um Boeing 737), restando aproximadamente **525** mil usuários na nossa base. Essa último passo é feito com o intuito de retirar usuários que são prováveis *bots* ou que compartilham contas, uma vez que a mudança entre localizações muito distantes em um espaço curto de tempo implica que eles se movimentaram a velocidades acima das possíveis por meios de transporte usuais.

Os resultados dispostos nas tabelas deste capítulo foram comparados usando o Teste-t de *Student* com 95% de confiança [Jain, 1991]. Em cada tabela, os resultados marcados com ▲ são estatisticamente superiores aos demais valores daquela tabela para sua métrica. Caso mais de um valor esteja marcado com ▲, esses resultados são estatisticamente semelhantes e superiores aos demais.

5.2.1 Inferência através da rede de relacionamentos

Nesta seção, analisaremos como se comportam os algoritmos de inferência de localização através da rede de relacionamentos dos usuários. Aplicamos os métodos em dois tipos de relacionamentos já mencionados na Seção 3.3: a relação de amizade e a relação através de menções.

A rede de amizades pode ser vista como a interseção da rede de seguidos e seguidores de um usuário. Como dizem Davis Jr et al. [2011], os amigos de um usuário tendem a estar mais próximos do mesmo, enquanto aqueles usuários que são somente seguidos ou seguidores possuem uma probabilidade maior de estarem mais distantes. Mostramos que isso é verdade para nossa base na Figura 3.10.

A rede de menções também já foi utilizada em trabalhos anteriores para identificar a localização do usuário. Um dos seus atrativos, para autores que trabalharam com a inferência de usuários do Twitter, é o fato de ela ter menor custo de coleta do que a rede de seguidos e seguidores, de acordo com a as restrições impostas pela política de uso da API do Twitter, explicada na Seção 3.1.

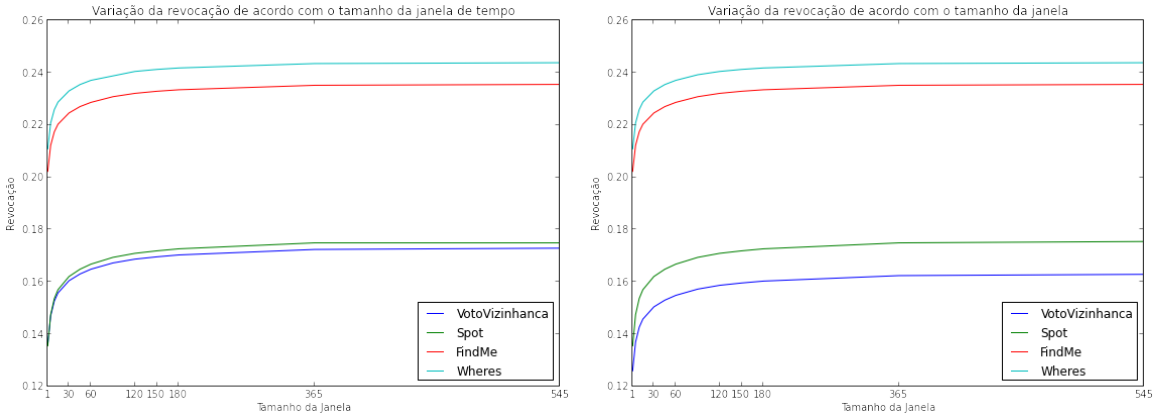
Em Jurgens et al. [2015], os autores avaliaram a eficácia dos principais métodos de inferência de localização da literatura que utilizam a rede de relacionamentos dos usuários. Para tanto, os autores utilizaram a rede de menções, mesmo que os trabalhos originais tenham utilizado a rede de amizades, como Davis Jr et al. [2011]. Aproveitamos a oportunidade, portanto, para comparar a eficácia dos métodos nos dois tipos diferentes de rede, além de analisarmos se o tamanho da janela de tempo entre menções recíprocas dos usuários pode influenciar nos resultados da rede criada.

5.2.2 Rede de menções

Como explicado na seção 3.3, uma aresta entre dois usuários é criada na rede de menções caso ambos mencionem um ao outro durante o intervalo de tempo da coleta. Nos experimentos anteriores da literatura, os autores não se preocuparam em analisar se o intervalo de tempo entre a menção do usuário A ao usuário B e a do usuário B ao usuário A importa.

Variamos o tamanho da janela de tempo entre os valores de 1 a 545 dias - tempo total da coleta. Como podemos observar na Figura 5.1, a **revocação** para cada um dos métodos aumenta até o valor em que o tamanho da janela é igual a 120 dias e, a partir daí, se estabiliza. Isso ocorre porque o número de menções recíprocas em uma pequena janela de tempo é menor do que em um período de tempo igual a 120 dias, tornando o grafo menos denso. A partir de então, aumentar o tamanho da janela de tempo considerada parece não afetar a revocação obtida.

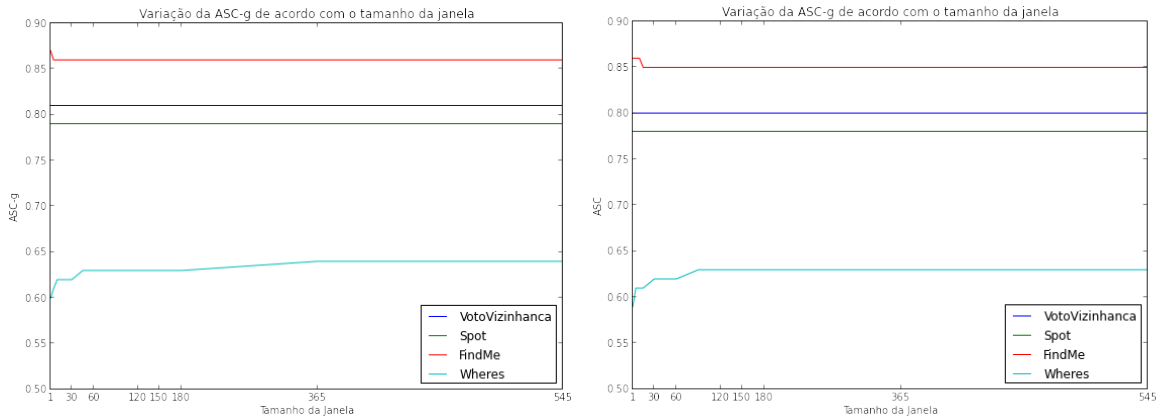
Observamos que a revocação do método **Wheres** foi superior nas duas formas de agrupamento, enquanto **VotoVizinhança** foi o que obteve os piores resultados em ambos os casos. Isso se deve ao fato de que o algoritmo de **VotoVizinhança** conta somente com os usuários diretamente conectados a um nó para inferir sua localização - e que um dos critério para atribuição de uma localização desconhecida é o número mínimo de conexões com localização previamente conhecida e o número mínimo de votos que um local deve receber para ser considerado o local predito. O método **Wheres**, no entanto, consegue propagar a localização de usuários através da rede, aumentando o número de usuários com localizações recuperadas em quase 8% em relação ao método de **VotoVizinhança**.



(a) GroupMinDistUsers (b) GroupMinUsers

Figura 5.1: A revocação de todos os métodos atinge seu máximo e se estabiliza a partir do tamanho de 120 dias

Quanto à distância dos agrupamentos preditos para os agrupamentos reais dos usuários, voltamos a ter um comportamento idêntico entre as duas formas de agrupar os usuários como podemos ver na Figura 5.2. Enquanto **Wheres** possui a maior **revocação** comparado aos demais algoritmos, sua **ASC-g** é inferior ao de todos, com o valor de 0.63, enquanto os demais algoritmos apresentam pontuação superior a 0.78. A eficácia dos algoritmos é praticamente a mesma, não importando a janela de tempo.



(a) GroupMinDistUsers (b) GroupMinUsers

Figura 5.2: A ASC-g é a mesma para os dois tipos de agrupamento

Na Tabela 5.1, sumarizamos os resultados obtidos para todos os métodos com seus melhores parâmetros para a janela de tempo igual a 120 dias, já que os resultados para janelas maiores é semelhante.

Embora os valores de **ASC-g** para os mesmos métodos sejam semelhantes nas duas formas de agrupamento, o acerto levando em consideração a cidade do usuário e a distância do centro dos agrupamentos preditos tem resultados diferentes. Enquanto o GroupMinDistUsers apresenta **ASC-c** acima de 0.60 o agrupamento gerado por GroupMinUsers exibe performance inferior a 0.60 para todos os algoritmos de rede testados. Tais resultados apoiam nossa premissa de que a escolha da forma de agrupamento é fundamental para minimizar os erros das localizações inferidas. O método com melhores resultados, o **FindMe**, tem resultados pouco melhores do que **Spot** e **VotoVizinhança** para **ASC-c, ASC-g e Mediana** enquanto **Wheres** possui os piores resultados para essas métricas. A revocação de todos os métodos não ultrapassa 25%, resultado de a maioria dos usuários da rede de menções não ser conectado a outro usuário da nossa base.

	GroupMinUsers				**GroupMinDistUsers**			
	VotoVizinhança	**Spot**	**FindMe**	**Wheres**	**David**	**Spot**	**FindMe**	**Wheres**
ASC-g	0.79	0.78	0.85	0.63	0.80	0.79	0.86▲	0.63
ASC-c	0.58	0.58	0.61	0.49	0.63	0.62	0.66▲	0.52
Acc@100	0.66	0.65	0.70	0.49	0.73	0.72	0.78▲	0.53
Mediana(km)	51.31	53.00	48.90	115.3	20.41	21.63	17.88▲	75.35
Revocação	0.16	0.16	0.23	0.24▲	0.16	0.17	0.23	0.24▲

Tabela 5.1: Avaliação dos métodos sobre a rede de menções

5.2.3 Rede de amizades

A rede de amizade possui maior número de arestas, já que, como mostrado na Figura 3.10, menos de 10% dos usuários não possui nenhum amigo em nossa base. Esperamos, portanto, que a **revocação** seja, pelo menos, maior do que a obtida na rede de menções ao aplicarmos os mesmos métodos e formas de agrupamento.

Na tabela 5.2, mostramos o desempenho dos diferentes métodos aplicados na rede de amizades. Novamente, cada um dos métodos apresenta resultados semelhantes para os diferentes agrupamentos com relação a ASC-g, com exceção de **FindMe**, que obteve o melhor resultado em GroupMinDistUsers (0.86). Para ASC-c, cada método obteve um resultado estatisticamente diferente para os diferentes agrupamentos em que foi aplicado, sendo que GroupMinDistUsers conseguiu o melhor resultado para todos os casos. O método **Spot** foi o que sofreu maior perda tanto para **Acc@100** e **Mediana**, em relação aos resultados obtidos para rede de menções, ao contrário de **Wheres** para o qual as mesmas métricas foram melhores na rede de amizades do que na de menções.

Note que a **revocação** de todos os algoritmos é próxima a 90% - um valor bem maior do que a **revocação** conseguida com a rede de menções.

	GroupMinUsers				**GroupMinDistUsers**			
	VotoVizinhança	**Spot**	**FindMe**	**Wheres**	**David**	**Spot**	**FindMe**	**Wheres**
ASC-g	0.81	0.61	0.70	0.80	0.82▲	0.62	0.72	0.78
ASC-c	0.58	0.47	0.52	0.57	0.63	0.51	0.56	0.60
Acc@100	0.68	0.48	0.57	0.66	0.74▲	0.53	0.63	0.70
Mediana(km)	50.78	115.32	66.83	53.06	20.60▲	77.63	37.92	27.75
Revocação	0.90	0.92▲	0.90	0.89	0.91	0.89	0.91	0.91

Tabela 5.2: Avaliação dos métodos de inferência aplicados na rede de amizades

O aumento das arestas entre os usuários foi essencial para conseguirmos uma melhor cobertura dos usuários com localização inferida. Embora a rede de menções tenha um custo inferior para ser obtido que a rede de amizades, ela não teve uma cobertura de usuários tão grande quanto a obtida pela rede de amizades, explicando, potencialmente, alguns dos resultados ruins obtidos por Jurgens et al. [2015].

5.2.4 Inferência a partir do texto dos tweets

Nesta seção avaliamos a eficácia dos métodos aplicados aos tweets de acordo com:

1. Classificadores: **Naive Bayes** e **Regressão Logística**.
2. Duas base, sendo que uma contém apenas tweets **geolocalizados** (Geo) e a outra contém tweets **geolocalizados e não-geolocalizados** (Geo+NotGeo)

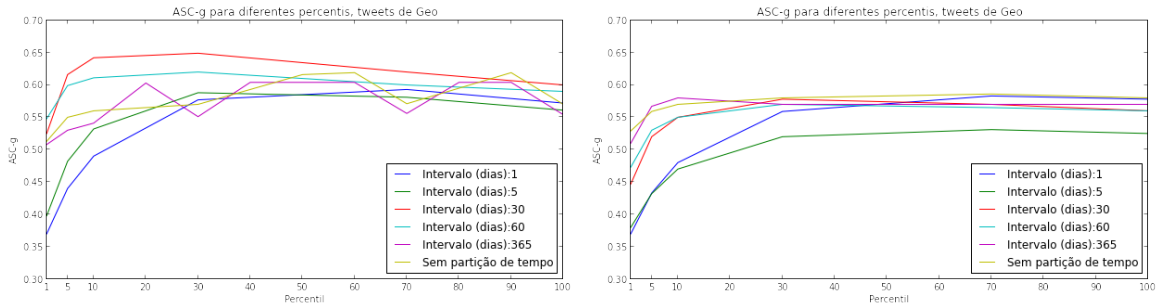
3. Duas configurações de agrupamento: **GroupMinDistUsers** e **GroupMinUsers**

Para melhorarmos a eficácia dos classificadores, fizemos seleção de atributos utilizando o teste χ^2, selecionando diferentes percentis dos atributos com maior pontuação para cada localização. A variação dos percentis utilizados foi feita em passos de 10, exceto para o quinto percentil, que é um valor intermediário entre o primeiro e décimo percentil.

Para ambas configurações de agrupamento, testamos diferentes janelas de tempo de acordo com o que foi explicado na seção 4.2.2, contando votos para cada agrupamento em cada janela de tempo classificada. Além de utilizarmos diferentes janelas de tempo, usamos uma partição sem divisão de tempo para contrastarmos a eficácia de dividirmos nossa base em períodos.

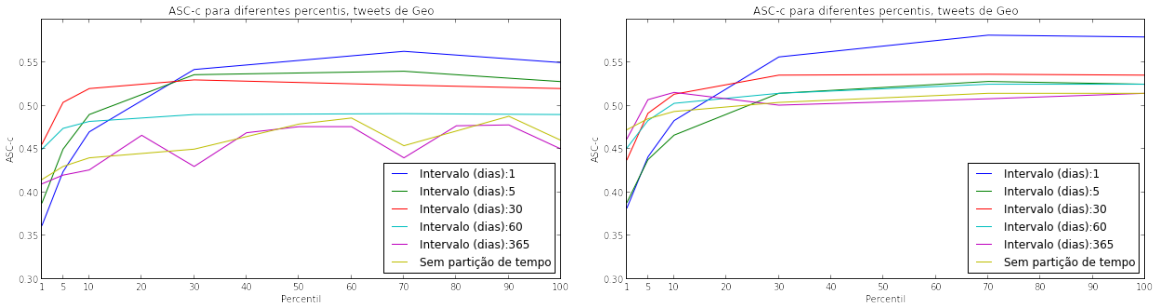
Na Figura 5.3 avaliamos os resultados do algoritmo Naive Bayes para as duas formas de agrupamento. O GroupMinDistUsers obteve resultados sistematicamente inferiores ao GroupMinUsers para a métrica ASC-g, com pouca distinção entre o modelo linguístico sem partição de tempo e os modelos com janela de tempo. Para o GroupMinUsers, no entanto, o modelo com janela de tempo igual a 30 dias obteve resultado superior aos demais no trigésimo percentil.

Embora o GroupMinUsers demonstre ter maior eficácia para inferir o agrupamento do qual o usuário faz parte, a Figura 5.4 apresenta resultados bastante semelhantes, principalmente para a janela de tempo igual a um dia e o modelo estático de classificação dos tweets, mostrando que o GroupMinDistUsers compensa as distâncias dos erros ao colocar os usuários mais próximos das suas cidades quando o agrupamento correto é inferido.



(a) GroupMinUsers (b) GroupMinDistUsers

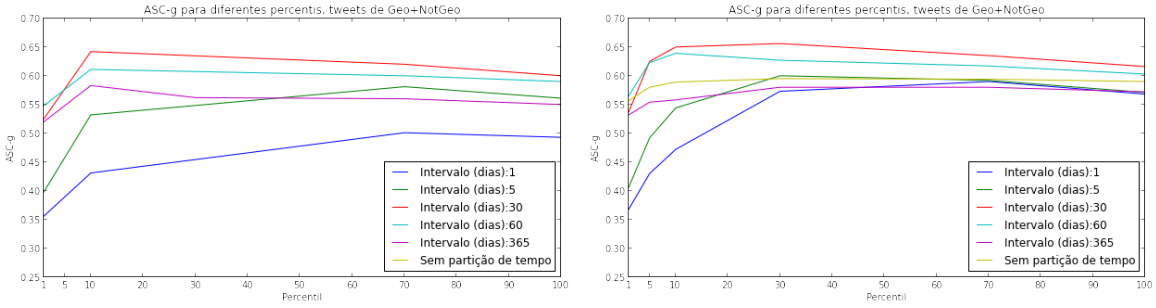
Figura 5.3: Avaliação da ASC-g para classificação utilizando Naive Bayes e tweets geolocalizados.



(a) GroupMinUsers (b) GroupMinDistUsers

Figura 5.4: Avaliação da ASC-c para classificação utilizando Naive Bayes e tweets geolocalizados.

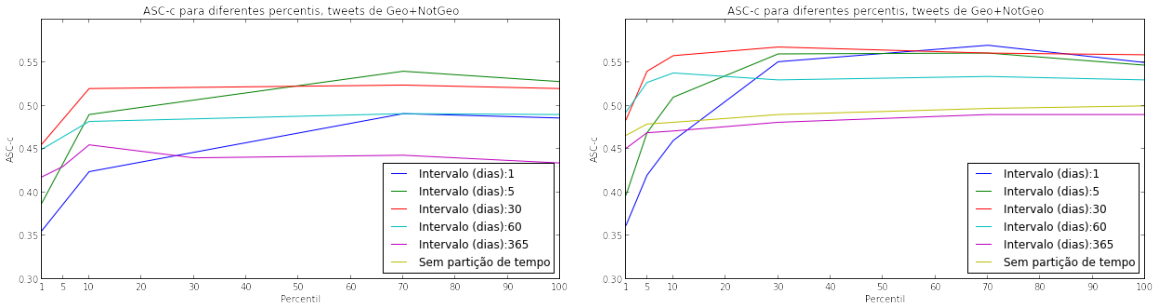
Ao utilizarmos as bases de tweets localizados e não localizados como treino e teste para nossos algoritmos de classificação, o GroupMinDistUsers consegue um desempenho superior na inferência da localização em relação ao GroupMinUsers, como vemos na Figura 5.5. O algoritmo que utiliza o modelo com atualização no intervalo de 30 dias teve melhor resultado tanto em uma forma de agrupamento como na outra, e o intervalo de tempo igual a um dia é o mais sensível à variação do percentil utilizado para seleção de atributos. Nas Figuras 5.6, novamente vemos que o GroupMinDistUsers fornece uma predição mais próxima da cidade real do usuário do que o GroupMinUsers.



(a) GroupMinUsers (b) GroupMinDistUsers

Figura 5.5: Avaliação da ASC-g para classificação utilizando Naive Bayes e tweets geolocalizados e não-geolocalizados.

Após testar as demais variações de classificador e configurações de agrupamento em diversos percentis na **base Geo**, sumarizamos os resultados na Tabela 5.3. Nessa tabela, os valores correspondem ao resultado obtido na melhor variação do percentil para cada método. Os valores acompanhados de ▲, são estatisticamente superiores aos demais valores de **toda tabela** com 95% de confiança pelo Teste T de *Student* para sua métrica. Os valores em negrito indicam que eles são superiores somente aos demais valores de toda **sua linha** com 95% de confiança.



(a) GroupMinUsers (b) GroupMinDistUsers

Figura 5.6: Avaliação da ASC-c para classificação utilizando Naive Bayes e tweets geolocalizados e não-geolocalizados.

O algoritmo Naive Bayes aplicado em GroupMinUsers obteve os melhores resultados em **ASC-g, ASC-c, Acc@100 e Mediana**. Por outro lado, o algoritmo Regressão Logística obteve seus melhores resultados no agrupamento **GroupMinDistUsers**, mas esse são estatisticamente inferiores aos resultados de Naive em GroupMinUsers. Os métodos que não utilizam partição de tempo obtiveram a maior revocação e empataram estatisticamente com os melhores resultados dos métodos que consideram a partição de tempo com o algoritmo de Regressão Logística. Foram considerados aproximadamente **120 mil termos únicos** nessa base.

	Naive Bayes na base Geo											
	GroupMinUsers						**GroupMinDistUsers**					
Intervalo(dias)	**1**	**5**	**30**	**60**	**365**	**S/Tempo**	**1**	**5**	**30**	**60**	**365**	**S/Tempo**
ASC-g	0.59	0.59	0.65 ▲	0.62	0.60	0.62	0.58	0.53	0.58	0.57	0.58	0.59
ASC-c	0.56▲	0.54	0.53	0.49	0.48	0.49	0.55	0.50	0.51	0.50	0.49	0.49
Acc@100	0.55▲	0.51	0.53	0.50	0.48	0.49	0.55 ▲	0.47	0.51	0.50	0.50	0.51
Mediana (km)	66.3 ▲	86.6	79.5	102.6	116.7	105.1	71.0	132.0	86.2	101.8	98.8	86.1
Revocação	0.75	0.67	0.65	0.74	0.80	0.80	0.81	0.81	0.77	0.81	0.90▲	0.90▲
	Regressão Logística na base Geo											
	GroupMinUsers						**GroupMinDistUsers**					
Intervalo(dias)	**1**	**5**	**30**	**60**	**365**	**S/Tempo**	**1**	**5**	**30**	**60**	**365**	**S/Tempo**
ASC-g	0.39	0.39	0.56	0.55	0.52	0.55	0.42	0.40	0.54	**0.58**	**0.58**	**0.57**
ASC-c	0.37	0.38	0.48	0.44	0.40	0.42	0.39	0.39	**0.50**	0.47	0.46	0.47
Acc@100	0.30	0.30	0.46	0.43	0.38	0.43	0.30	0.31	**0.49**	0.46	0.46	**0.49**
Mediana (km)	391.7	359.6	98.2	124.7	147.9	121.0	382.0	374.3	**77.1**	**79.4**	112.6	**78.7**
Revocação	0.85	0.83	.72	0.63	0.90▲	0.90▲	0.87	0.86	0.70	0.62	0.90▲	0.90▲

Tabela 5.3: O GroupMinUsers apresenta melhores resultados para ASC-g que o GroupMinDistUsers, quando Naive Bayes é aplicado a base de somente tweets geolocalizados

Na Tabela 5.4, utilizamo-nos do mesmo esquema da tabela anterior para base **Geo + NotGeo**, avaliando os ganhos ao usarmos também os tweets não-geolocalizados. Dessa vez, a configuração de agrupamento **GroupMinDistUsers** obteve melhores resultados tanto para o algoritmo Naive Bayes quanto para o de Regressão Logística. O algoritmo de Naive Bayes aplicado nas bases Geo+NotGeo e GroupMinDistUsers

obteve resultados relacionados a precisão melhores que o algoritmo de Regressão Linear aplicado às mesmas configurações. Além disso, para esse método e configuração, a **Mediana** da distância representa quase a metade da obtida na melhor configuração da base **Geo**, além de valores superiores de **Acc@100, ASC-g e ASC-c**.

O método que não utiliza partição de tempo na base Geo+NotGeo possui valores sistematicamente inferiores ao método que utiliza intervalo de tempo de 30 dias, exceto para revocação. Notamos, portanto, vantagem na utilização do método que leva em conta as partições de tempo se desejamos melhorar nossas medidas de precisão sem nos importar com perder em revocação.

	Naive Bayes nas bases Geo+NotGeo											
	GroupMinUsers						**GroupMinDistUsers**					
Intervalo(dias)	**1**	**5**	**30**	**60**	**365**	**S/Tempo**	**1**	**5**	**30**	**60**	**365**	**S/Tempo**
ASC-g	0.50	0.58	0.64	0.61	0.58	0.61	0.59	0.60	0.66▲	0.64	0.58	0.59
ASC-c	0.49	0.54	0.52	0.49	0.45	0.50	0.57	0.56▲	0.57	0.54	0.49	0.50
Acc@100	0.44	0.51	0.53	0.50	0.44	0.51	0.56	0.56	0.61▲	0.58	0.51	0.53
Mediana (km)	201.5	86.6	83.4	102.7	127.7	107.3	58.7	58.7	37.5▲	54.6	90.2	79.1
Revocação	0.86	0.67	0.65	0.74	**0.90**	**0.90**	0.79	0.66	0.63	0.72	**0.90**	**0.90**
	Regressão Logística nas bases Geo + NotGeo											
	GroupMinUsers						**GroupMinDistUsers**					
Intervalo(dias)	**1**	**5**	**30**	**60**	**365**	**S/Tempo**	**1**	**5**	**30**	**60**	**365**	**S/Tempo**
ASC-g	0.42	0.43	**0.61**	0.60	0.56	0.60	0.42	0.43	0.59	0.60	0.58	**0.61**
ASC-c	0.41	0.42	0.53	0.49	0.45	0.47	0.42	0.42	**0.54**	0.52	0.49	0.51
Acc@100	0.33	0.33	0.51	0.47	0.42	0.47	0.33	0.34	**0.53**	**0.53**	0.49	**0.53**
Mediana (km)	376.6	359.6	88.2	117.7	143.6	119.0	370.9	359.9	**74.2**	**76.4**	108.3	**78.7**
Revocação	0.89	0.87	0.0.72	0.64	0.99▲	0.99▲	0.88	0.88	0.71	0.64	0.99▲	0.99▲

Tabela 5.4: Os métodos que utilizam partição de tempo possuem resultados melhores para **ASC-g, ASC-c, Acc@100 e Mediana** do que o método sem partição de tempo. No entanto, esse último apresenta maior **revocação**

5.2.5 Inferência a partir dos atributos

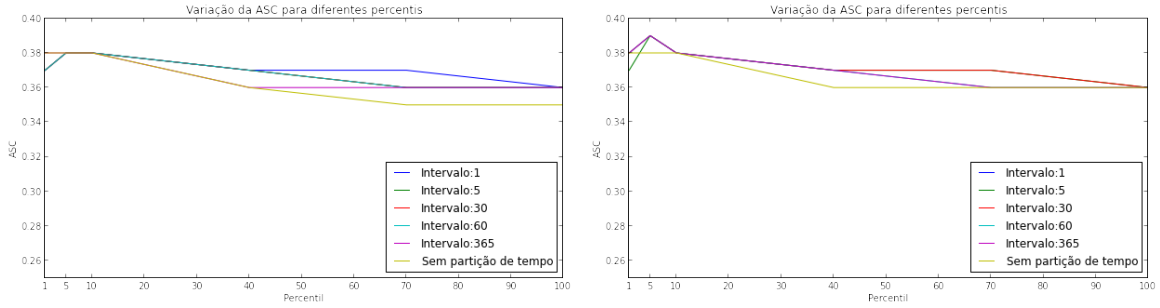
Nesta seção, relatamos os resultados obtidos a partir dos algoritmos de inferências com base nos atributos do perfil do usuário: a **localização declarada** e **descrição** fornecida pelo usuário.

Como explicado na Seção 4.3, utilizaremos duas abordagens para trabalhar com esses campos:

1. Utilização de um algoritmo de classificação sobre um modelo unigrama para inferir a localização do usuário

2. Utilização de um casamento exato de nomes oficiais de cidades e seus sinônimos no texto dos campos

Para o campo de descrição, aplicamos o algoritmo Naive Bayes com atualização do modelo de classificação a cada intervalo de tempo e sem atualização do modelo de classificação. Como mostrado na Figura 5.7. Os resultados para cada métrica são estatisticamente indistinguíveis, como sumarizamos na Tabela 5.5,

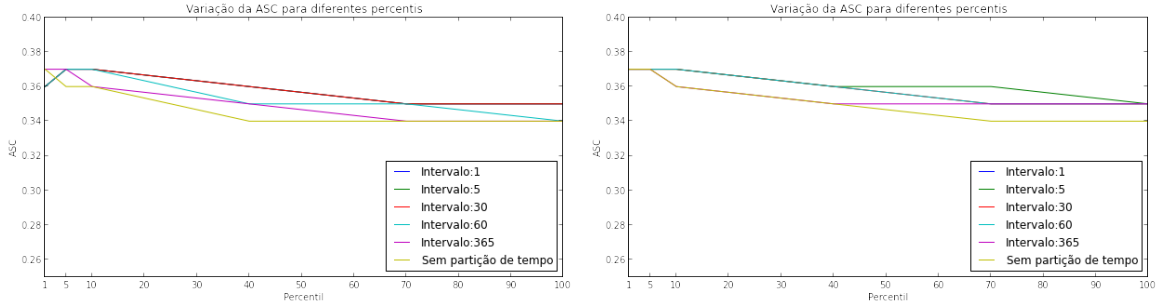


(a) Classificação com Naive Bayes no campo de Descrição - **GroupMinUsers**

(b) Classificação utilizando Naive Bayes no campo de Descrição - **GroupMinDistUsers**

Figura 5.7: Avaliação da distância entre o agrupamento predito e o agrupamento do usuário

Enquanto isso, a escolha do tipo de agrupamento de cidades escolhido pouco importa para a distância entre as cidades reais do usuário e a distância ao centro do agrupamento inferido, como vemos na Figura 5.8. Novamente, seguindo a tendência para distância dos agrupamentos, o algoritmo sem partição de tempo apresenta resultados inferiores. A **revocação** para os classificadores aplicados na descrição é igual a **0.65**.



(a) Classificação com Naive Bayes no campo de Descrição - **GroupMinUsers**

(b) Classificação utilizando Naive Bayes no campo de Descrição - **GroupMinDistUsers**

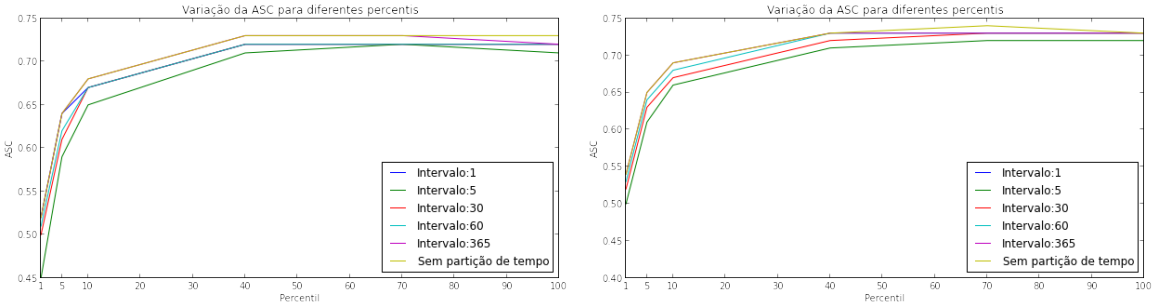
Figura 5.8: Avaliação da distância entre o a cidade do usuário e o centro do agrupamento inferido

Ao aplicarmos o algoritmo de classificação no campo de localização, a despeito de encontrarmos o campo preenchido por textos jocosos, conseguimos maiores valores

	Campo Descrição											
	Naive - GroupMinUsers						Naive - GroupMinDistUsers					
Intervalo	1	5	30	60	365	545	1	5	30	60	365	545
ASC-g	0.38	0.38	0.38	0.38	0.38	0.38	0.39▲	0.39▲	0.39▲	0.39▲	0.39▲	0.38▲
ASC-c	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37
Acc@100	0.26	0.26	0.26	0.26	0.26	0.26	0.27▲	0.27▲	0.27▲	0.25	0.25	0.27▲
Mediana	402.59▲	403.75▲	404.39▲	405.60▲	408.73▲	413.88	406.62▲	404.66▲	405.94▲	405.34▲	414.65	413.44
Revoc.	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65

Tabela 5.5: Os resultados não são sensíveis quanto janela de atualização de tempo. Os resultados são apresentados com os melhores conjuntos de atributos para cada partição de tempo.

de **ASC-g** que vários métodos aplicados no texto de tweets (Seção 5.2.4) e do campo de descrição, como vemos na Figura 5.9.



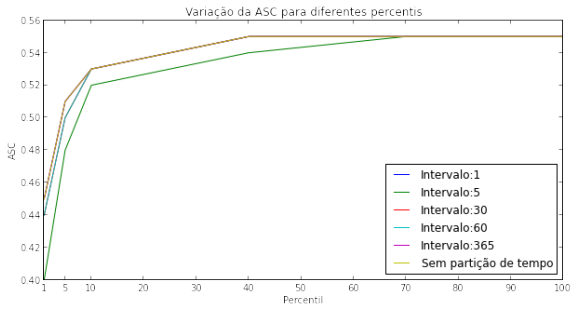
(a) Classificação com Naive Bayes no campo de Localização - **GroupMinUsers**

(b) Classificação com Naive Bayes no campo de Localização - **GroupMinDistUsers**

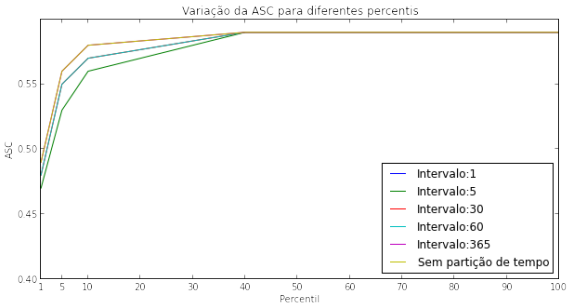
Figura 5.9: Avaliação quanto a distância entre o agrupamento inferido e o agrupamento do usuário

Novamente, para distância real entre a cidade e o centro do agrupamento inferido, representada pela métrica **ASC-c**, a escolha do método de agrupamento pouco importou para os resultados obtidos, exceto para as janelas de tempo igual a cinco e número de atributos abaixo do quinto percentil, onde os **GroupMinDistUsers** conseguiu uma pontuação superior. A revocação para os classificadores aplicados ao campo de localização é **0.55**.

A Tabela 5.6 sumariza os resultados dos nossos experimentos utilizando o classificador no campo de localização. Para cada partição de tempo, apresentamos o melhor resultado obtido ao variamos o subconjunto de termos utilizado, de acordo com o teste χ^2. Os resultados, para cada métrica, obtidos em cada tipo de agrupamento, não são sensíveis quanto ao tamanho da janela de tempo, mostrando que o tamanho da janela de tempo para atualização do modelo não contribuiu nos resultados. Por outro lado, embora o **ASC-g** de ambos os agrupamentos sejam indistintos, a **ASC-c, Acc@100**



(a) Classificação com Naive Bayes no campo de Localização - **GroupMinUsers**



(b) Classificação utilizando Naive Bayes no campo de Localização - **GroupMinDistUsers**

Figura 5.10: Avaliação quanto a distância entre a cidade do usuário e o centro do agrupamento inferido

e Mediana do GroupMinDistUsers apresenta resultados melhores, justamente por privilegiar agrupamentos mais compactos e com menor espalhamento de usuários.

	Campo Localização											
	Naive - GroupMinUsers						Naive - GroupMinDistUsers					
Intervalo(dias)	1	5	30	60	365	545	1	5	30	60	365	545
ASC-g	0.71	0.71	0.72	0.72	0.73	0.73	0.70	0.72	0.73▲	0.73▲	0.73▲	0.74▲
ASC-c	0.54	0.54	0.55	0.55	0.55	0.55	0.59▲	0.59▲	0.59▲	0.59▲	0.59▲	0.59▲
Acc@100	0.58	0.58	0.58	0.58	0.60	0.60	0.65	0.65	0.66▲	0.66▲	0.66▲	0.66▲
Mediana (km)	58.96	58.91	58.95	58.85	58.84	58.83	29.85▲	29.55▲	29.60▲	29.52▲	29.42▲	29.34▲
Revocação	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55

Tabela 5.6: A atualização do campo de descrição periodicamente não afeta a inferência

Finalmente, utilizamos a técnica de casamento exato entre os nomes das cidades encontrados em GeoNames[1] e os campos de descrição e localização do usuário. Embora seja uma técnica simples, ela obteve precisão melhor do que a dos classificadores, embora a custo de menor **revocação**.

Utilizamos duas formas para escolher a cidade inferida: utilizando o nome encontrado no campo de descrição/localização do último tweet ou a mais frequente em todos os tweets - uma vez que o usuário pode mudar tais campos. Ambas as estratégias, no entanto, produziram resultados semelhantes

Uma vez que com o casamento exato obtemos o nome de uma cidade, é possível inferir a localização do usuário ao nível de cidade e, portanto, a métrica **ASC-g** não faz sentido para avaliação deste método. Embora a revocação tenha sido baixo tanto para descrição quanto para localização, conseguimos uma **Acc@100** igual a 0,84 para o campo de localização declarada. Para o campo de descrição, porém, a aproximação

[1]http://www.geonames.org/

da cidade inferida com a real tem resultados piores do que através da utilização de um classificador, sugerindo que a citação da cidade onde o usuário reside no campo de descrição não é tão comum quanto esperado .

	Localização	**Descrição**
ASC-c	0.84▲	0.30
Acc@100	0.84▲	0.25
Mediana(km)	0.0▲	972.4
Revocação	0.37▲	0.11

Tabela 5.7: Casamento exato de nomes de cidade com os campos de descrição e localização

5.3 Análise comparativa dos métodos de inferência

Cada um dos métodos analisados nas seções anteriores possui suas vantagens e limitações para inferir a localização de um usuário. A classificação dos algoritmos de rede é tão boa quanto o número de conexões que ele possui com usuários que contenham informações geolocalizadas, enquanto a classificação de usuários baseada nos textos dos tweets depende do número de tweets que ele publicou e a existência de palavras com alta relevância geográfica.

Na Tabela 5.8, apresentamos os resultados de cada método com seus melhores parâmetros segundo o **ASC-c**. Mostramos em negrito os resultados do método de **Casamento Exato - Loc**, uma vez que ele apresenta de forma isolada os melhores resultados para **ASC-g, ASC-c, Acc@100 e Mediana**. No entanto, sua revocação é abaixo do 40% e, portanto, se mostra inviável para inferir a localização de todos os usuários da base. Seguido do **Casamento Exato - Loc**, temos o método **FindMe** na rede de menções, que possui bom desempenho tanto em **ASC-g, ASC-c, Acc@100 e Mediana**, mas revocação abaixo de 30%. Por fim, em terceiro lugar, temos o método **VotoVizinhança** na rede de amizade, que apresenta desempenho de menos de 10% abaixo do FindMe, mas revocação de 91%.

A seguir, calculamos a concordância entre esses métodos considerando a o agrupamento inferido na configuração **GroupMinDistUsers**, que foi a que apresentou melhores resultados para **ASC-c, Acc@100 e Mediana**. Dependendo do grau de discordância, podemos ter indicações de que combinar os resultados desses métodos pode ser uma abordagem vantajosa para tarefa de inferência.

O coeficiente de kappa de Cohen é uma medida estatística que mede a concordância entre dois anotadores, descontando os acertos ocorridos por acaso. O coeficiente de Kappa de Cohen é dado pela equação:

		GroupMinUsers					GroupMinDistUsers				
		ASC-g	ASC-c	Acc@100	Mediana	Revoc.	ASC-g	ASC-c	Acc@100	Mediana	Revoc.
Rede Menções	VotoViz.	0.79	0.58	0.66	51.31	0.16	0.80	0.63	0.73	20.41	0.16
	FindMe	0.85	0.61	0.70	48.90	0.23	0.86▲	0.66▲	0.78▲	17.29▲	0.23
	Wheres	0.63	0.49	0.49	115.3	0.24	0.63	0.52	0.53	75.35	0.24
	Spot	0.78	0.58	0.65	53.0	0.16	0.79	0.62	0.72	21.63	0.17
Rede Amizades	VotoViz.	0.81	0.58	0.68	50.78	0.90	0.82	0.63	0.74	20.60	0.91
	FindMe	0.70	0.52	0.57	66.83	0.90	0.72	0.56	0.63	37.92	0.91
	Wheres	0.80	0.57	0.66	53.06	0.89	0.78	0.60	0.70	27.75	0.91
	Spot	0.61	0.47	0.48	115.32	0.92▲	0.62	0.51	0.53	77.63	0.89
Casamento Exato	Descrição	0.30	0.25	0.16	972.4	0.11	0.30	0.26	0.16	972.4	0.11
	Loc.	**0.84**	**0.84**	**0.84**	**0.0**	0.37	**0.84**	**0.84**	**0.84**	**0.0**	0.37
Classificação de Campos	Descrição	0.38	0.37	0.26	404.39	0.55	0.39	0.39	0.37	405.94	0.55
	Loc.	0.72	0.55	0.58	58.95	0.65	0.73	0.59	0.66	29.60	0.65
Classificação Tweets Temporal	Naive	0.64	0.52	0.53	83.4	0.65	0.66	0.57	0.61	37.5	0.63
	Regressão	0.61	0.53	0.51	88.2	0.69	0.59	0.54	0.53	74.2	0.71
Classificação Tweets	Naive	0.58	0.45	0.44	0.51	0.91	0.59	0.49	0.51	90.2	0.90
	Regressão	0.60	0.47	0.47	119.00	0.91	0.58	0.49	0.49	108.30	0.90

Tabela 5.8: Comparação da melhor configuração de cada método. Mediana dada em km.

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}, \tag{5.1}$$

onde $P(a)$ é a concordância relativa observada e $P(e)$ é a probabilidade hipotética de que a concordância tenha ocorrido por acaso. Se os anotadores estão em completo acordo, temos $\kappa = 1$ e no outro caso extremo, $\kappa = 0$.

Na Tabela 5.9, mostramos a concordância entre dois métodos de inferência de acordo com o coeficiente de Kappa de Cohen na diagonal inferior e a fração da base de usuários cuja localização foi inferida por ambos os métodos na diagonal superior.

O casamento exato de padrões de nomes de cidades para o campo de descrição possui concordância inferior a 0,18 com relação a todos os outros métodos, inclusive com o modelo criado pelo algoritmo Naive Bayes sobre as palavras do próprio campo (0,10). Ao mesmo tempo, a porcentagem de usuários inferidos pelo método de casamento no campo de descrição e outros métodos (primeira linha da tabela) é de no máximo 10%, o que pode ser explicado pela baixa **revocação** do método (vide Tabela 5.8). Por outro lado, o casamento exato de nomes no campo de localização declarada do usuário e os métodos de localização através da rede de amigos inferem cerca de 35% dos usuários - praticamente a mesma porcentagem de usuários inferida pelo algoritmo de casamento exato sozinho (37%), sugerindo que há uma redundância no uso dos dois métodos. Porém, note que os coeficientes de Kappa de Cohen inferiores a 0.60 entre os métodos de rede e de casamento exato sobre o campo de localização que contradizem essa informação. O coeficiente de Kappa igual a 0.84 e porcentagem da base classificada pelos métodos Casamento-loc e Naive-loc a 37% sugere que ambos são redundantes, o

que pode ser explicado por ambos serem aplicados no mesmo campo.

Dentre os métodos utilizados para inferência, **VotoVizinhança** é o que possui maior concordância com os demais. Mesmo os pares desses métodos inferindo mais de 90% dos usuários da base, κ é inferior a 0.70 para todos os casos, exceto entre **VotoVizinhança** e **Wheres**, onde há a maior concordância (0.87), sugerindo que ambos são redundantes. Ao compararmos os métodos de rede aplicados tanto na rede de amigos e menções, primeiro notamos que o número de usuários inferidos por ambos é limitado pelo baixa revocação da rede de menções. Os coeficientes de concordância entre os mesmos métodos aplicados em redes diferentes não ultrapassa 0.60, exceto para **VotoVizinhança**. A concordância entre métodos diferentes aplicados na rede de menções é alta, acima de 0.87, exceto para **FindMe** e **Wheres** que, no entanto, inferem a maior fração de usuários para modalidade **Rede de Menções**.

Os métodos de classificação aplicados aos tweets possuem concordâncias próximas a 0.50. Uma exceção é Naive-Geo+NotGeo e Naive-geo, que possuem uma concordância de 0.70. A fração de usuários rotulada por qualquer par de métodos aplicados a tweets é inferior a 0.60. A discordância é ainda maior entre os modelos de classificação aplicados a tweets e os aplicados a redes de amizades ou menções. Essa alta discordância é uma oportunidade a ser explorada na combinação dos algoritmos, discutida na próxima Seção.

		Casamento		Rede Amigos				Rede Menções				Tweets, 30 dias				Naive	
		Desc	**Loc**	**VotoV.**	**Spot**	**Wheres**	**FindMe**	**VotoV.**	**Spot**	**Wheres**	**FindMe**	**Reg. Geo+NGeo**	**Reg. Geo**	**Naive Geo+NGeo**	**Naive Geo**	**Desc**	**Loc**
Casamento	**Desc**		0.05	0.1	0.1	0.11	0.11	0.02	0.02	0.03	0.03	0.05	0.08	0.05	0.07	0.1	0.08
	Loc	0.15		0.34	0.35	0.37	0.37	0.06	0.07	0.1	0.09	0.18	0.28	0.19	0.24	0.3	0.36
Rede Amigos	**VotoV.**	0.11	0.63		0.9	0.9	0.9	0.16	0.17	0.23	0.22	0.43	0.67	0.47	0.59	0.61	0.5
	Spot	0.07	0.4	0.59		0.92	0.92	0.16	0.17	0.24	0.23	0.44	0.69	0.48	0.61	0.62	0.52
	Wheres	0.11	0.59	0.87	0.55		0.97	0.16	0.17	0.24	0.23	0.47	0.73	0.5	0.64	0.64	0.54
	FindMe	0.09	0.5	0.66	0.45	0.61		0.16	0.17	0.24	0.23	0.47	0.73	0.5	0.64	0.64	0.54
Rede Menções	**VotoV.**	0.12	0.58	0.74	0.51	0.72	0.59		0.16	0.16	0.16	0.08	0.12	0.09	0.11	0.11	0.09
	Spot	0.11	0.57	0.73	0.5	0.71	0.58	0.93		0.17	0.17	0.08	0.13	0.1	0.12	0.12	0.1
	Wheres	0.08	0.42	0.54	0.38	0.52	0.43	0.95	0.87		0.24	0.11	0.18	0.13	0.16	0.16	0.14
	FindMe	0.13	0.63	0.71	0.49	0.68	0.58	0.92	0.88	0.67		0.11	0.18	0.13	0.15	0.16	0.13
Tweets Temporal 30 dias	**Reg. Geo+NGeo**	0.07	0.34	0.35	0.26	0.34	0.25	0.41	0.4	0.31	0.38		0.43	0.33	0.38	0.32	0.27
	Reg. Geo	0.04	0.18	0.18	0.13	0.18	0.13	0.19	0.18	0.15	0.18	0.55		0.41	0.56	0.49	0.41
	Naive Geo+NGeo	0.1	0.46	0.49	0.37	0.48	0.38	0.53	0.52	0.41	0.51	0.52	0.33		0.42	0.34	0.29
	Naive Geo	0.07	0.37	0.37	0.28	0.36	0.29	0.4	0.39	0.31	0.38	0.46	0.36	0.7		0.44	0.36
Naive	**Desc**	0.11	0.11	0.11	0.09	0.11	0.08	0.1	0.1	0.13	0.11	0.15	0.09	0.16	0.14		0.47
	Loc	0.1	0.84	0.51	0.34	0.48	0.4	0.48	0.47	0.35	0.51	0.34	0.18	0.41	0.34	0.11	

Tabela 5.9: A diagonal inferior mostra a concordância segundo Kappa de Cohen, enquanto a diagonal superior mostra a porcentagem dos usuários cujas localizações puderam ser inferidas pelos dois métodos.

5.4 Combinação de resultados

Nesta seção, avaliamos a combinação dos resultados dos métodos base através de sistemas de votação e meta-árvore de decisão. Para combinar os métodos, usamos validação cruzada com dez partições. Nos próximos experimentos, utilizamos a configuração de agrupamento de **GroupMinDistUsers**, uma vez que ela apresentou os melhores resultados para **ASC-c, Acc@100 e Mediana.**

Para o método de votação, seja C o nosso conjunto de métodos base e w_i o peso do voto do classificador $c_i \in C$. A localização escolhida como a inferida para o usuário é aquela que possui a maior soma de votos ponderados pelo peso w_i dos métodos de inferência base. Como explicado na seção 4.4, utilizamos os seguintes sistemas de votação:

1. **Voto majoritário simples**: os votos de todos os métodos possuem o mesmo peso. Cada uma das 10 partições é utilizada uma vez como partição de teste e os algoritmos base são treinados nas 9 partições restantes. Nenhum ajuste é feito no peso dos votos.

2. **Voto Acc**: cada uma das 10 partições é utilizada como partição teste uma única vez. Das 9 partições restantes, 8 são utilizadas no treino dos métodos base e 1 partição é utilizada como validação. Para cada método $c_i \in C$, o seu peso $w_i = ASG - c_i$, onde $ASG - g_i$ é a **ASG-g** obtida após treinar o método na base separada para treino e validá-lo na partição de validação.

3. **Voto - GA**: os pesos dos votos de cada método são ajustados usando um algoritmo genético [Holland, 1975] que otimiza a acurácia obtida pela votação. Cada indivíduo do algoritmo genético é composto por N números reais, onde N é o número de métodos de inferência C. O objetivo do algoritmo genético é maximizar a acurácia dos votos ponderados. Para ajustarmos os pesos dos votos de cada método através do algoritmo genético, utilizamos a configuração de probabilidade de mutação igual a 0.05 e de cruzamento igual a 0.95. O número de indivíduos em cada população é igual a 50 e o número de gerações igual a 150. Para cada uma das 10 partições utilizadas como teste, as 9 restantes são utilizadas com treino para os métodos base e ajuste de pesos através do algoritmo genético.

Para combinarmos resultados de diferentes métodos de inferência base através da meta árvore de decisão, precisamos fornecer-lhe, além das classes inferidas por cada método, atributos que representem a confiança com a qual fizeram a inferência. Como

exemplo, suponhamos que um dos classificadores base seja o Naive Bayes: neste caso, fornecemos a classe predita e sua probabilidade segundo o algoritmo.

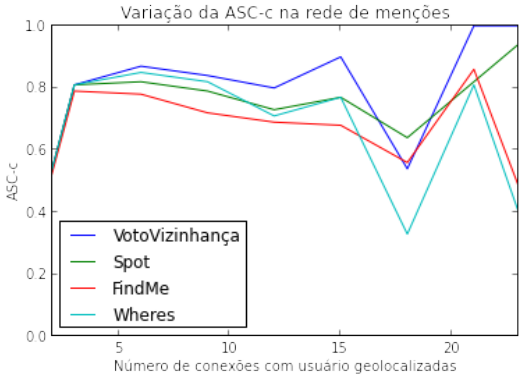
Para nossos métodos, no entanto, não possuímos essa probabilidade fornecida diretamente pelos algoritmos e, portanto, temos a tarefa de extraí-la através da análise dos resultados, tentando identificar quais fatores se relacionam com a eficácia da inferência.

A seguir, descrevemos os atributos fornecidos à meta árvore de decisão e a justificativa de como eles se relacionam com a eficácia de cada método. Para os métodos de rede, analisamos como a qualidade da inferência se relaciona à quantidade de conexões geolocalizadas que um usuário possui. Usaremos a **ASC-c** para medir a qualidade da inferência, uma vez que ela leva em consideração a proximidade das inferências feitas para as localizações reais dos usuários.

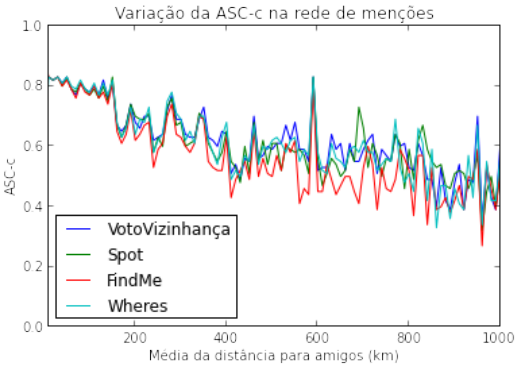
Na Figura 5.11a, as tendências do algoritmo de **VotoVizinhança** e **Spot** parecem ser diferentes das do **FindMe** e **Wheres** quanto ao número de conexões na rede de menções: à medida que o número de conexões aumenta, **Spot e VotoVizinhança** parecem se beneficiar, ao contrário de **FindMe e Wheres**. No entanto, como a variação do número de conexões por usuário na nossa rede de menções é pequena, é difícil dizer com certeza se essa tendência é real. Para a rede de amizades (Figura 5.11c), com um intervalo maior de número de conexões, vemos que todos os algoritmos apresentam melhores resultados para a faixa de número de amigos inferior a 100, e para **VotoVizinhança** o ASC-c oscila bastante à medida que usuários têm um maior número de amigos. Tanto na rede de menções quanto na rede de amizades (Figuras 5.11b e 5.11d), a distância média entre a localização inferida para o usuário e a localização dos usuários com os quais ele se conecta é um indicativo da qualidade da inferência. Quanto maior a distância, menores as chances de termos a localização correta.

No método de inferência temporal aplicado sobre os tweets, o usuário pode ter tweets presentes em diferentes períodos de tempo (tweets em todos os meses, por exemplo). Para cada período de tempo em que ele possui tweets, o classificador (Naive Bayes ou Regressão Logística), indica a probabilidade de que cada localização seja a do usuário. A localização inferida é aquela com maior probabilidade no maior número de janelas de tempo - ou seja, com maior número de votos. As análises a seguir foram feitas usando a seleção de atributos com corte no trigésimo percentil e atualização do classificador a cada 30 dias.

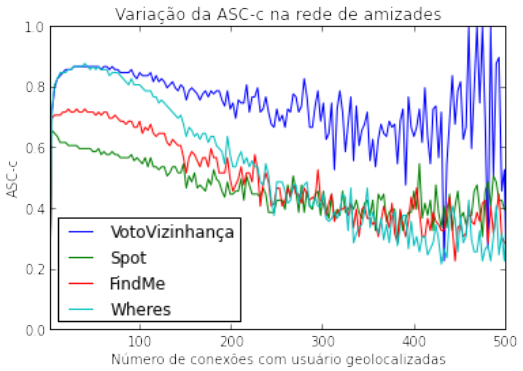
Nas Figuras 5.12a e 5.12c, vemos que o número de períodos de tempo avaliados pelo método influencia na qualidade da inferência gerada. Em geral, ao aumentarmos o número de votos apurados pelo método tanto utilizando o Naive Bayes quanto a Regressão Logística temos resultados melhores. Enquanto isso, a média da proba-



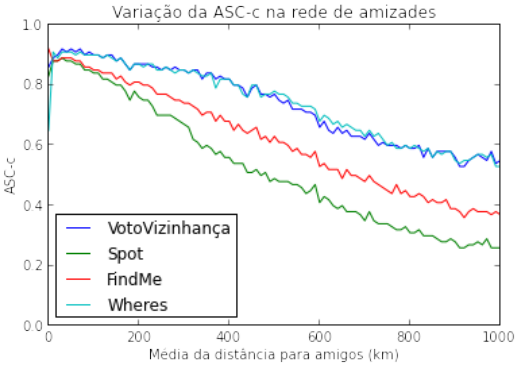
(a) **VotoVizinhança** e **Spot** parecem seguir tendências diferentes de **Wheres** e **FindMe** a medida que cresce o número de conexões



(b) A medida que a distância média entre o usuário e suas conexões aumenta, a variação dos valores de ASC-c também aumenta



(c) **VotoVizinhança** apresenta a maior variação de resultados quando o usuário possui muitos amigos geolocalizados



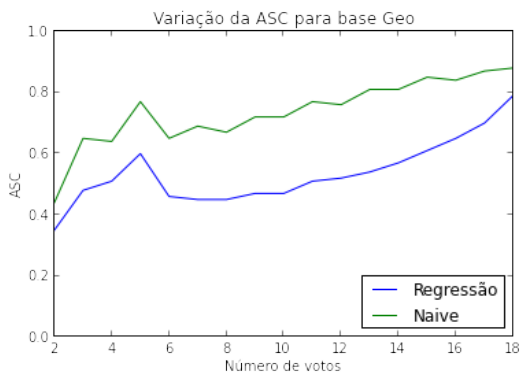
(d) A média de distância entre o usuário e seus amigos é um indicativo de inferência incorreta

Figura 5.11: Variação de **ASC-c** de acordo com o número de conexões geolocalizadas do usuário e a distância média entre a localização inferida para o usuário e as localizações dos usuários com os quais conecta

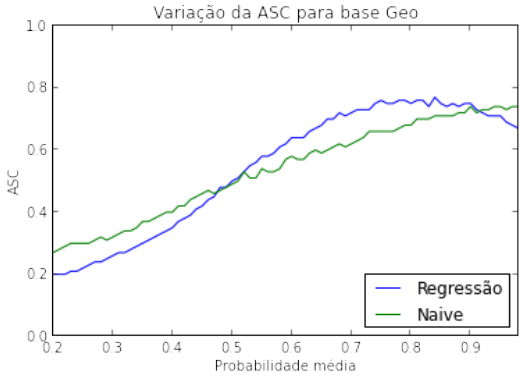
bilidade de cada localização tem influências diferentes para Naive Bayes e Regressão Logística. Nas Figuras 5.12b e 5.12d, vemos que, para probabilidades inferiores a 0.5, Naive Bayes possui maior ASC-c, enquanto para valores superiores a 0.5 a Regressão Logística possui maior eficácia.

Avaliando o comportamento do Naive Bayes nos campos de descrição e localização declarada do usuário, para média de probabilidades inferiores a 0.6 temos vales bruscos na ASC-c para o campo de localização declarada, enquanto para o campo de descrição temos uma curva mais suave de crescimento, com maior inclinação a partir da probabilidade 0.70.

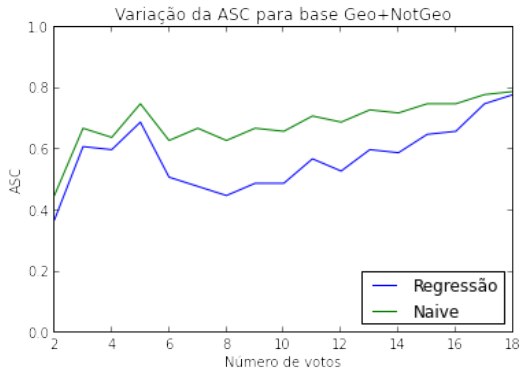
Finalmente, para o casamento exato de cidades no campo de localização e des-



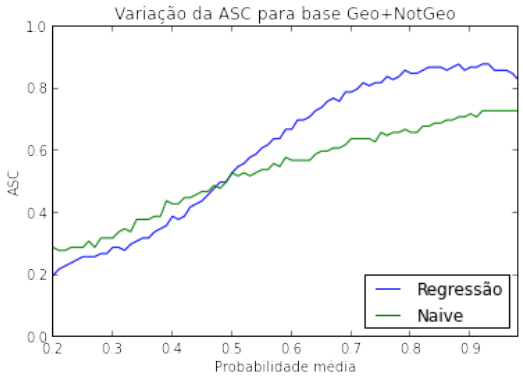
(a) A ASC-c é maior para usuários que possuem tweets em um número maior de diferentes períodos de tempo



(b) Os algoritmos apresentam diferentes faixas de probabilidade em que obtêm melhores ASC-c



(c) O número de votos influencia da mesma forma ambos algoritmos



(d) A ASC-c é mais sensível ao aumento da probabilidade da regressão logística

Figura 5.12: Variação da ASC-c de acordo com a probabilidade média dos algoritmos e o número de períodos em que os algoritmos classificaram tweets para o usuário

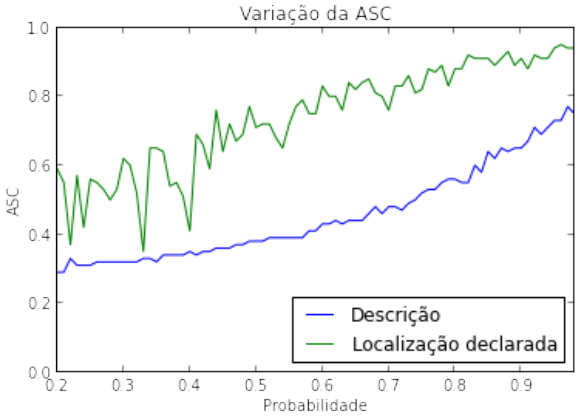


Figura 5.13: Para probabilidades abaixo de 0.6, o algoritmo de Naive Bayes apresenta bruscas variações quanto ao ASC-c para o campo de localização

crição, o próprio fato de haver um casamento exato indica uma alta confiança para o caso do campo de localização declarada com uma média de **Acc@100** igual a 84% (Tabela 5.8), embora o mesmo não seja verdade para o casamento exato no campo de descrição. Para cada um dos métodos de casamento, fornecemos à meta árvore de decisão um atributo discreto, indicando a ocorrência de um casamento exato com o nome oficial de alguma cidade brasileira.

Na Tabela 5.10, sumarizamos os atributos ordinários fornecidos à meta-árvore de decisão. Lembre-se que, como explicado na seção 4.4, diferente da árvore de decisão comum, buscamos uma árvore cujas folhas determinem qual classificador deverá ser utilizado.

	Atributos Ordinários
Rede de Menções **Rede de Amizades**	- Número de Conexões com usuários geolocalizados - Média da distância da localização inferida entre o usuário e os usuários com os quais se conecta
Casamento Exato	- Ocorrência do casamento
Classificação aplicada no campo de descrição/localização	- Probabilidade atribuída à localização inferida
Classificação aplicada em tweets em diferentes períodos de tempo	- Probabilidade média atribuída à localização inferida - Número de votos
Classificação aplicada em tweets **Classificação aplicada em tweets**	Probabilidade atribuída à localização inferida

Tabela 5.10: Atributos ordinários fornecidos para meta árvore de decisão

Os métodos baseados na rede de amizades obtiveram os melhores resultados para as métricas **ASC-g, ASC-c, Acc@100 e mediana** , como vimos na tabela 5.8. No entanto, como explicamos na Seção 3.1, a rede de amigos é o atributo do usuário com maior custo de ser recuperado, ao passo que as menções feitas a outros usuários, descrição do perfil e localização declarada são atributos que já vêm incorporados em cada tweet coletado. Ao avaliarmos a utilização dos sistemas de Voto-GA e a meta árvore de decisão, para cada método utilizamos dois conjuntos de testes: um que inclui os resultados e atributos ordinários provindos de todos os métodos de inferência de localização de usuários e outro que exclui os resultados e atributos ordinários dos métodos de inferência que utilizam a rede de amizades. Nosso objetivo, com o último teste, é descobrir se conseguimos uma eficácia satisfatória comparável ao que conseguimos com os métodos aplicados na rede de amizades, uma vez que o custo de obtenção desses dados é menor.

Os resultados das diferentes formas de combinação dos resultados estão sumarizados na Tabela 5.11. A **revocação** de todas as formas de combinação apresentadas na tabela 5.11 são iguais ou superior a **0.97**, e portanto significativamente melhores que o valor máximo encontrado de **0.91** na Tabela 5.8 com a utilização de um único

	ASC-g	ASC-c	Acc@100	Mediana	Revocação
Voto Simples	0.78	0.61	0.71	24.82	0.99▲
Voto Acc	0.79	0.63	0.72	21.62	0.99▲
Voto - GA	0.85	0.65	0.77	18.27	0.99▲
Voto - GA S/Amigos	0.71	0.58	0.63	33.80	0.98
Árvore de Decisão	0.86▲	0.65▲	0.78▲	17.76▲	0.97
Árvore de Decisão S/Amigos	0.64	0.57	0.55	56.16	0.77

Tabela 5.11: Combinando todos os métodos de inferência, 99% da base é coberta sem perda na precisão

modelo. Esses resultados indicam que nossa expectativa de cobrir um maior número de usuário ao combinar os algoritmos foi atendida.

As combinações **Voto Simples** e **Voto Acc** tiveram eficácia inferior quando comparadas com os métodos individuais de rede de amizade, de menções e casamento de localização em relação às métricas de precisão. Porém, a **revocação** dos métodos individuais aplicados à rede de menções e casamento no campo de localização não chegam a 50% da base de usuários. Os resultados próximos de **Voto Simples** e **Voto Acc** mostram que ganha-se pouco ao ajustarmos os pesos de acordo com as acurácias dos métodos utilizados neste trabalho.

Na Tabela 5.12, mostramos os pesos atribuídos pelo algoritmo genético a cada método após treinamento na base inteira de usuários. O método de **VotoVizinhança** aplicado à rede de amigos recebe o maior dos pesos, em contraste com o peso igual a 0,10 quando aplicado a Rede de Menções também do mesmo algoritmo. Três métodos aplicados a rede de amigos estão entre os 5 com maior peso, em conformidade com seus altos valores de **revocação** e **ASC-c e Acc@100** vistos na Tabela 5.8. Dentre os métodos aplicados aos tweets, os que receberam maior peso foram aqueles aplicados às bases Geo+NotGeo, enquanto para os votos do campos do perfil do usuário os métodos aplicados ao campo de localização tiveram maior peso do que aqueles dos métodos aplicados ao campo de descrição. O peso do voto do modelo criado a partir do campo de localização utilizando o Naive Bayes foi maior do que o peso atribuído ao casamento exato que, embora tenha maiores **Acc@100, ASC-c, ASC-g**, tem uma revocação baixa igual, a 37%.

O fato de os métodos baseados na rede de amizades conseguirem o maior peso era esperado, uma vez que eles, individualmente, conseguem bons resultados, como mostramos na Tabela 5.8. Averiguamos, então, o desempenho da combinação dos algoritmos, exceto os aplicados na rede de amizades, com seus pesos gerados pelo algoritmo genético. Podemos ver os pesos gerados na Tabela 5.13. Os votos com maior peso correspondem a algoritmos aplicados na rede de menções, campo de localização e texto na base Geo e Geo+NotGeo. Novamente, o campo de localização declarada esteve entre os

Método	Peso	Método	Peso
Rede Amigos - VotoVizinhança	0.98	**Rede Amigos - FindMe**	0.91
Naive - Loc	0.89	**Naive Tweets - Geo+NotGeo**	0.86
Rede Amigos - Where	0.85	**Casamento Localização**	0.72
Regressão Tweets - Geo+NotGeo	0.70	**Rede Menções - FindMe**	0.63
Rede Menções - Wheres	0.41	**Rede Menções - Spot**	0.42
Casamento Descrição	0.33	**Naive Desc**	0.36
Naive Tweets - Geo	0.12	**Rede Amigos - Spot**	0.20
Regressão - Geo	0.10	**Rede Menções - VotoVizinhança**	0.10

Tabela 5.12: Pesos dos votos em **Votos-GA**. Somente os pesos diferentes de zero são mostrados

cinco métodos com maior peso, tanto aplicando o casamento exato quanto utilizando o algoritmo Naive. Esse resultado sugere que o atributo de localização declarada carrega tanta informação sobre a localização do usuário quanto a rede de menções. Ao mesmo tempo, para os métodos aplicados nos textos dos tweets, o algoritmo Naive Bayes aplicado tanto na base Geo+NotGeo e Geo obteve pesos semelhantes, enquanto o casamento exato no campo de descrição tem um peso igual a 0.07, mostrando-se o menos expressivo. Embora alguns métodos tenham obtido pesos pequenos, eles garantem alta revocação na ausência de inferência de localização por outros métodos.

Método	Peso	Método	Peso
Rede Menções - FindMe	1.12	**Casamento Localização**	1.0
Naive - Loc	0.53	**Naive - Geo+NotGeo**	0.5
Naive - Geo	0.49	**Rede Menções - Spot**	0.29
Reg. - Geo+NotGeo	0.08	**Casamento Descrição**	0.07
Rede Menções - Wheres	0.06	**Rede Menções - VotoVizinhança**	-0.01
Naive - Descrição	-0.05	**Reg. - Geo**	-0.05

Tabela 5.13: Pesos dos votos em **Votos-GA S/Amigos**, sem a rede de amigos. Somente os pesos diferentes de zero

As inferências mais precisas, no entanto, foram feitas pela meta árvore de decisão que, no entanto, obteve menor revocação que os outros métodos de combinação. Na Tabela 5.14, mostramos a fração de usuários inferida por cada um dos métodos de acordo com a meta árvore de decisão treinada sobre os usuários e as saídas dos classificadores. Quando utilizamos a rede de amigos, o algoritmo de **VotoVizinhança** classifica a maior parte do usuários. Logo em seguida, temos o algoritmo Naive Bayes aplicado ao campo de localização. Os métodos baseados em texto são utilizados para inferir a localização de uma parcela menor que 1% dos usuários nessa configuração.

Ao removermos a rede de amizades, porém, ainda na tabela 5.14, os métodos baseados em texto são escolhidos para inferir a maioria das localizações dos usuários, seguido pela rede de menções. O algoritmo de casamento exato no campo de localiza-

ção, embora tenha bons resultados para precisão das suas inferências, é desprestigiado pela árvore por ter uma revocação abaixo dos demais métodos. Esse comportamento é diferente do observado em Votos-GA, que dá um voto para os métodos aplicados sobre o campo de localização. A revocação também diminui bastante ao retirarmos os resultados da rede de amizades do conjunto de treinamento da meta árvore de decisão.

Embora tanto os resultados de **Voto-GA S/Amigos e Meta Árvore de decisão S/Amigos** tenham sido aquém dos obtidos pelos outros métodos de combinação, ambos apresentam resultados competitivos em relação aos algoritmos de Voto-Vizinhança e FindMe aplicados individualmente na rede de amizades, especialmente **Voto-GA S/Amigos** para as métricas **ASC-c e Acc@100**. Isso demonstra que, na falta da rede de amizades, a combinação dos resultados dos outros métodos pode compensar essa ausência.

Todos os métodos		**Sem Rede de Amizades**	
Método	**Fração inferida**	**Método**	**Fração inferida**
Rede Amigos - VotoVizinhança	0.47	Reg. - Geo+NotGeo	0.59
Naive loc	0.27	Naive - Geo	0.19
Rede Amigos - Wheres	0.13	Rede Menções - FindMe	0.08
Rede Amigos - FindMe	0.08	Casamento - Loc	0.07
Casamento - Loc	0.04	Naive - Loc	0.04
Rede Amigos - Spot	0.01	Naive - Geo+NotGeo	0.03

Tabela 5.14: Fração de usuários cujas localizações foram inferidas por cada método de acordo com a meta árvore de decisão

Portanto, se a rede de amizade estiver disponível, a combinação com a qual obtemos melhor resultados é a que utiliza a meta árvore de decisão e, portanto, seu uso é aconselhável, embora com menor revocação. Por outro lado, caso não utilizemos a rede de amizades, o sistema de votação com pesos ajustados pelo algoritmo genético tem melhores resultados para as métricas relacionadas a precisão, além de maior revocação.

Capítulo 6

Conclusão e trabalhos futuros

Existem diversas abordagens para inferir a localização geográfica de usuários do Twitter quando não temos disponíveis tweets suficientes com localização derivada de GPS. Neste trabalho, avaliamos métodos que se baseiam na rede de amizades, rede de menções, tweets públicos, descrição e localização declarada do perfil dos usuários, cada uma com suas vantagens e desvantagens específicas. Comparando os métodos entre si, aqueles baseados na rede de amizades obtiveram os melhores resultados tanto em precisão de localização de usuários quanto em revocação, seguidos dos métodos que utilizam a localização declarada do usuário e finalmente dos modelos linguísticos treinados sobre os tweets públicos dos usuários.

Embora os métodos derivados dos tweets públicos dos usuários tenham mostrado precisão inferior às dos métodos de inferência através da rede, os primeiros utilizam atributos do usuário, cujo custo de recuperação é menor. Mostramos que o modelo de inferência baseado no texto precisa ser atualizado periodicamente para apreender mudanças no uso de termos localizados e propomos um novo método para inferência da localização dos usuários levando em consideração vários períodos de tempo. O algoritmo de Naive Bayes aplicado sobre tweets geolocalizados e não-geolocalizados obteve os melhores resultados quando atualizado a cada 30 dias. Com essa configuração, nosso método obteve precisão até 10% maior do que o modelo de texto estático.

Embora a precisão do método de casamento exato no campo de localização declarada seja alto, a sua revocação só não é mais baixa do que a revocação do método de casamento exato no campo de descrição. O modelo criado pelo algoritmo Naive Bayes sobre o campo de localização declarada produz resultados melhores do que os modelos criados sobre os tweets, por exemplo, mas os usuários classificados pelo mesmo têm uma grande sobreposição com os usuários classificado pelo casamento exato no campo de localização, mostrando redundância entre os dois métodos. Os métodos que uti-

lizam o campo de descrição obtiveram os piores resultados tanto em precisão quanto revocação, sugerindo que o campo não seja tão útil para inferência da localização dos usuários quanto os demais atributos disponíveis.

Portanto, para os campos preenchidos voluntariamente pelo usuário, o campo de localização declarada se mostrou o mais eficaz e o método que melhor o utiliza é o casamento exato com o nome oficial de cidades. No nosso trabalho, utilizamos uma lista simples de nomes de cidades, deixando de aproveitar informações outras informações geográficas como o nome de estados, por exemplo. Além do mais, a baixa precisão encontrada no casamento exato do nome de cidades no campo de descrição sugere o uso de técnicas mais sofisticadas para diferenciar quando uma palavra é utilizada para designar um local ou refere-se a outro tipo de entidade. Nossos resultados têm potencial de serem melhorados, no futuro, com a utilização de diferentes gazetteers e outras técnicas de reconhecimento de entidades aplicadas aos campos de descrição e localização declarada do usuário, apontando para trabalhos futuros.

Mostramos que os usuários tendem a se relacionar com outros usuários próximos de si no mundo real tanto através de menções quanto através da relação de amizade. Os métodos aplicados à rede de relacionamentos dos usuários apresentaram, em geral, alta precisão. A rede de amizades, devido a restrições atuais da API do Twitter, é mais difícil de ser obtida do que a rede de menções que, por isso mesmo, vem sendo utilizada como alternativa à rede de usuários em alguns trabalhos como o de Jurgens et al. [2015]. Mostramos, no entanto, que a substituição de uma pela outra não leva aos mesmos resultados, principalmente quanto à cobertura, que é menor quando os mesmos algoritmos foram aplicados na rede de menções. O método de Rout et al. [2013] obteve menor precisão na rede de menções, enquanto o método de Kong et al. [2014] foi mais prejudicado na rede de amizades. Tanto os métodos de Davis Jr et al. [2011] e Backstrom et al. [2010] obtiveram resultados de precisão próximos tanto na rede de menções quanto na rede de amizades, sendo os melhores métodos de inferência na categoria de métodos aplicados à rede de relacionamentos. O método de Davis Jr et al. [2011], no entanto, é mais atraente por apresentar resultados que se aproximam do de Backstrom et al. [2010] e ser computacionalmente mais barato, já que sua inferência é feita baseada na localização mais frequente dos seus amigos, sem necessidade de treinamento prévio.

Por terem comportamentos diversos, alguns usuários podem revelar - mesmo que implicitamente - sua localização através de um ou mais atributos que estudamos no nosso trabalho. Mostramos que ao utilizar somente um deles nunca conseguimos 100% de revocação. A combinação dos métodos-base através de sistemas de votação resultaram em uma cobertura de quase 100% da base de usuários. No sistema de

votação através de votos ponderados cujos pesos foram atribuídos por um algoritmo genético obtivemos resultados superiores, inclusive, em precisão. O uso de uma meta árvore de decisão para escolha do método cuja localização deve ser atribuída ao usuário obteve os melhores resultados de precisão tanto sobre os métodos base quanto sobre as outras formas de combinação dos mesmos.

Avaliamos a combinação dos métodos que não utilizam a rede de amizade - uma vez que ela é a mais custosa de ser obtida. A revocação foi superior a de qualquer um desses métodos base utilizados individualmente, mas os resultados relacionados às métricas de precisão foram inferiores tanto na combinação no sistema de **Votos-GA** quanto **Meta árvore de decisão**. Isso nos leva a conclusão de que os métodos baseados na rede de amizade obtêm os melhores resultados enquanto os outros métodos base podem ser utilizados como métodos complementares.

A esparsidade de dados em cidades pouco populosas é um desafio para os métodos de inferência de localização. Em trabalhos anteriores, essas cidades foram ignoradas ou agrupadas com cidades maiores. A escolha do tipo de agrupamento de cidades feito mostrou forte influência nos resultados obtidos. Mostramos que **ASC-g** semelhantes para tipos de agrupamentos diferentes gera erros diferentes quanto à distância da cidade real do usuário.

Analisamos duas formas de agrupamento, mas existe espaço para avaliação de outras formas de agrupamento de cidades, como *k-d tree*, que analisaremos nos próximos trabalhos. No nosso trabalho, não filtramos quais usuários são possíveis *bots* e *spammers*, deixando essa análise para trabalhos futuros. Além disso, não utilizamos a informação de fuso horário preferencial do usuário, que é uma fonte de informação importante principalmente quando a tarefa de inferir a localização do usuário leva em conta múltiplos países. Por fim, pretendemos validar nossos resultados em outras bases, estendendo nosso trabalho para levar em consideração a localização de usuários em nível mundial.

Referências Bibliográficas

Backstrom, L.; Sun, E. & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. Em *Proceedings of the 19th international conference on World wide web*, pp. 61--70. ACM.

Blei, D. M.; Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993--1022.

Bouillot, F.; Poncelet, P.; Roche, M. et al. (2012). How and why exploit tweet's location information? Em *International Conference on Geographic Information Science (AGILE)*.

Chandra, S.; Khan, L. & Muhaya, F. B. (2011). Estimating twitter user location using social interactions–a content based approach. Em *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 838--843. IEEE.

Cheng, Z.; Caverlee, J. & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. Em *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759--768. ACM.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273--297.

Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D. & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436--22441.

Davis Jr, C. A.; Pappa, G. L.; de Oliveira, D. R. R. & de L Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735--751.

Eisenstein, J.; O'Connor, B.; Smith, N. A. & Xing, E. P. (2010). A latent variable model for geographic lexical variation. Em *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277--1287. Association for Computational Linguistics.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R. & Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871--1874.

Finkel, J. R.; Grenager, T. & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. Em *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 363--370. Association for Computational Linguistics.

Gelernter, J. & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753--773.

Gomide, J.; Veloso, A.; Meira Jr, W.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. Em *ACM Web Science Conference (WebSci)*, pp. 1--8.

Graham, M.; Hale, S. A. & Gaffney, D. (2013). Where in the world are you? geolocation and language identification in twitter. *CoRR*, abs/1308.0683, abs/1308.0683.

Han, B.; Cook, P. & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pp. 451--500.

Hecht, B.; Hong, L.; Suh, B. & Chi, E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 237--246. ACM.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press.

Ikawa, Y.; Enoki, M. & Tatsubori, M. (2012). Location inference using microblog messages. Em *Proceedings of the 21st international conference companion on World Wide Web*, pp. 687--690. ACM.

Jain, R. (1991). *The Art of Computer System Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling.* John Wiley & Sons.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. Em *ICWSM*.

Jurgens, D.; McCorriston, J.; Xu, Y. T. & Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. Em *ICWSM*.

Kinsella, S.; Murdock, V. & O'Hare, N. (2011). I'm eating a sandwich in glasgow: modeling locations with tweets. Em *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 61--68. ACM.

Kong, L.; Liu, Z. & Huang, Y. (2014). Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, 7(13).

Kwak, H.; Lee, C.; Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? Em *Proceedings of the 19th international conference on World wide web*, pp. 591--600. ACM.

Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M. M. A.; Agrawal, A. & Choudhary, A. (2011). Twitter trending topic classification. Em *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 251--258. IEEE.

Lee, R. & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. Em *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 1--10. ACM.

Lieberman, M. D.; Samet, H. & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. Em *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pp. 201--212. IEEE.

Lotan, G.; Graeff, E.; Ananny, M.; Gaffney, D.; Pearce, I. & Boyd, D. (2011). The revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5:1375--1405.

Mahmud, J.; Nichols, J. & Drews, C. (2012). Where is this tweet from? inferring home locations of twitter users. Em *ICWSM*.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Martin, J. H. & Jurafsky, D. (2000). Speech and language processing. *International Edition*.

McCallum, A.; Nigam, K. et al. (1998). A comparison of event models for naive bayes text classification. Em *AAAI-98 workshop on learning for text categorization*, volume 752, pp. 41--48. Citeseer.

McGee, J.; Caverlee, J. & Cheng, Z. (2013). Location prediction in social media based on tie strength. Em *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 459--468. ACM.

Paradesi, S. M. (2011). Geotagging tweets using their content. Em *FLAIRS Conference*.

Popescu, A. & Grefenstette, G. (2010). Mining user home location and gender from flickr tags. Em *In proceedings of the International AAAI Conference on Weblogs and Social Media*.

Priedhorsky, R.; Culotta, A. & Del Valle, S. Y. (2014). Inferring the origin locations of tweets with quantitative confidence. Em *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 1523--1536. ACM.

Ren, K.; Zhang, S. & Lin, H. (2012). Where are you settling down: Geo-locating twitter users based on tweets and social networks. Em *Information Retrieval Technology*, pp. 150--161. Springer.

Ribeiro Jr, S. S.; Junior, Z.; Meira Jr, W. & Pappa, G. L. (2012). Positive or negative? using blogs to assess vehicles features. *Encontro Nacional de Inteligência Artificial*, pp. 1--12.

Rodrigues; Assuncao & Pappa (2013). Uncovering the location of twitter users. Em *Brazilian Conference on Intelligent Systems,*.

Roller, S.; Speriosu, M.; Rallapalli, S.; Wing, B. & Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. Em *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1500--1510. Association for Computational Linguistics.

Rout, D.; Bontcheva, K.; Preoţiuc-Pietro, D. & Cohn, T. (2013). Where's@ wally?: a classification approach to geolocating users based on their social ties. Em *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pp. 11--20. ACM.

Sakaki, T.; Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. Em *Proceedings of the 19th international conference on World wide web*, pp. 851--860. ACM.

Schulz, A.; Hadjakos, A.; Paulheim, H.; Nachtwey, J. & Mühlhäuser, M. (2013). A multi-indicator approach for geolocalization of tweets. Em *Proceedings of the Seventh International Conference on Weblogs and Social Media, International AAAI Conference on Weblogs and Social Media*.

Smith, D. A. & Crane, G. (2001). Disambiguating geographic names in a historical digital library. Em *Research and Advanced Technology for Digital Libraries*, pp. 127--136. Springer.

Sultanik, E. A. & Fink, C. (2012). Rapid geotagging and disambiguation of social media text via an indexed gazetteer. Em *ISCRAM, 2012*, pp. 1–10.

Takhteyev, Y.; Gruzd, A. & Wellman, B. (2012). Geography of twitter networks. *Social Networks*, 34(1):73--81.

Todorovski, L. & Džeroski, S. (2000). *Combining multiple models with meta decision trees*. Springer.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G. & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178--185.

Weng, J. & Lee, B.-S. (2011). Event detection in twitter. Em *International AAAI Conference on Weblogs and Social Media*.

Wing, B. & Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. Em *ACL*, volume 11, pp. 955--964.

Yardi, S. & Boyd, D. (2010). Tweeting from the town square: Measuring geographic local networks. Em *International AAAI Conference on Weblogs and Social Media*.

Zong, W.; Wu, D.; Sun, A.; Lim, E.-P. & Goh, D. H.-L. (2005). On assigning place names to geography related web pages. Em *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pp. 354--362. ACM.