EDUARDO MARCEL MAÇAN

GEOFIER: um Sistema de Anotação Geográfica de Textos com o Uso de Classificadores de Aprendizagem de Máquina

São Paulo

2015

EDUARDO MARCEL MAÇAN

GEOFIER: um Sistema de Anotação Geográfica de Textos com o Uso de Classificadores de Aprendizagem de Máquina

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

São Paulo

2015

EDUARDO MARCEL MAÇAN

GEOFIER: um Sistema de Anotação Geográfica de Textos com o Uso de Classificadores de Aprendizagem de Máquina

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

Área de concentração:
Engenharia de Computação

Orientador:
Prof. Dr. Edson Satoshi Gomi

São Paulo

2015

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Maçan, Eduardo Marcel

GEOFIER: um Sistema de Anotação Geográfica de Textos com o Uso de Classificadores de Aprendizagem de Máquina / E. M. Maçan -- versão corr. -- São Paulo, 2015.

66 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Aprendizado Computacional 2.Inteligência Artificial 3.Mineração de dados 4.Toponímia 5.Linguagem Natural I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Este trabalho é dedicado a meus pais,
Maria Neuza Cecchetti Maçan e Reinaldo Maçan.

AGRADECIMENTOS

Agradeço a minha esposa e companheira Andressa Juliane Martins, com quem dividi as angústias e as horas tardias, por me compreender, incentivar e manter nos trilhos. A meu orientador, Prof. Dr. Edson Satoshi Gomi, não apenas pela orientação na pesquisa mas também pelo acompanhamento ao longo da selva obscura, quando a estrada reta fora perdida. A meu filho Leonardo Malmegrim Maçan que desde sempre foi o incentivo para que eu nunca me acomodasse e, finalmente, agradeço a meus pais pelos ombros nos quais sempre me apoiei para que pudesse ver mais longe.

La cognizione del tempo preterito e del sito della terra è ornamento e cibo delle menti umane.

(Leonardo da Vinci)

RESUMO

A anotação geográfica de documentos consiste na adoção de metadados para a identificação de nomes de locais e a posição de suas ocorrências no texto. Esta informação é útil, por exemplo, para mecanismos de busca. A partir dos topônimos mencionados no texto é possível identificar o contexto espacial em que o assunto do texto está inserido, o que permite agrupar documentos que se refiram a um mesmo contexto, atribuindo ao documento um escopo geográfico. Esta Dissertação de Mestrado apresenta um novo método, batizado de Geofier, para determinação do escopo geográfico de documentos. A novidade apresentada pelo Geofier é a possibilidade da identificação do escopo geográfico de um documento por meio de classificadores de aprendizagem de máquina treinados sem o uso de um gazetteer e sem premissas quanto à língua dos textos analisados. A Wikipédia foi utilizada como fonte de um conjunto de documentos anotados geograficamente para o treinamento de uma hierarquia de Classificadores Naive Bayes e Support Vector Machines (SVMs). Uma comparação de desempenho entre o Geofier e uma reimplementação do sistema Web-a-Where foi realizada em relação à determinação do escopo geográfico dos textos da Wikipédia. A hierarquia do Geofier foi treinada e avaliada de duas formas: usando topônimos do mesmo gazetteer que o Web-a-Where e usando n-gramas extraídos dos documentos de treinamento. Como resultado, o Geofier manteve desempenho superior ao obtido pela reimplementação do Web-a-Where.

Palavras-chave: Anotação geográfica. Geotagging. Classificação automática de texto. Hierarquias de classificadores de texto. Gazetteers. Ambiguidade de topônimos.

ABSTRACT

Automatic text geotagging is the process by which mentions of place names and their positions in text are identified as metadata, allowing this information to be used by specialized applications, like Search Engines. It is possible to identify the geographic scope of a document by analysing the toponyms it mentions and then group documents by their geographic context, effectively adding a geographic scope to the documents. This dissertation presents a new method to identify the geographic scope of text, named Geofier. The novelty in Geofier is that it uses machine learning text classifiers, trained without the need of a gazetteer and without making assumptions regarding the language in which the documents are written. Wikipedia was used as the source for a geotagged text dataset in order to train a hierarchy of Naive Bayes and Support Vector Machine (SVM) classifiers. The Geofier hierarchy was then trained and evaluated, first using toponyms from the same gazetteer as Web-a-Where and then using n-grams extracted from the training samples as attributes. Geofier performed significantly better when compared to a Web-a-Where implementation.

Keywords: Geotagging. toponym ambiguity . Automatic text classifiers. Hierarchy of text classifiers. Gazetteers. Toponym ambiguity.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de consulta a um gazetteer. Detalhe do topônimo “Santa Mariana”. 15
Figura 2 – Sobreposição de áreas geográficas. 18
Figura 3 – Algoritmo do sistema Web-a-Where 19
Figura 4 – Web-a-Where: Algoritmo de desambiguação de topônimos 20
Figura 5 – Hierarquia Geográfica utilizada pelo Web-a-Where 21
Figura 6 – Arquitetura NewsStand 23
Figura 7 – Hiperplanos separando duas classes de exemplos 31
Figura 8 – Hiperplano ótimo separando duas classes de exemplos e seus support vectors 31
Figura 9 – Hierarquia de classificadores geográficos. Cada classificador corresponde a um nó, com classes que representam seus nós filhos 34
Figura 10 – Acurácia de Classificação Naive Bayes de artigos da Wikipédia em português para as cidades de São Paulo e do Rio de Janeiro, variando-se o número de documentos para treinamento 35
Figura 11 – Acurácia de Classificação com SVM de artigos da Wikipédia em português para as cidades de São Paulo e do Rio de Janeiro, variando-se o número de documentos para treinamento 36
Figura 12 – Algoritmo de Treinamento da Hierarquia Geofier 37
Figura 13 – Algoritmo de Classificação de Documentos do Geofier 37
Figura 14 – Artigos georreferenciados da Wikipédia em inglês. Nota-se a concentração dos textos em determinadas regiões 40
Figura 15 – Consulta ao banco de dados geográfico usando um polígono. Documentos de “San Francisco” 41
Figura 16 – Hierarquia geográfica dos Estados Unidos 43
Figura 17 – Artigos georreferenciados da Wikipédia em inglês, EUA Continental . . . 44
Figura 18 – Artigos georreferenciados da Wikipédia em inglês, Alaska 44
Figura 19 – Desempenho da hierarquia de classificadores com N-gramas, por região dos Estados Unidos 52
Figura 20 – Desempenho da hierarquia de classificadores com N-gramas, para estados dos Estados Unidos 53
Figura 21 – Desempenho da hierarquia de classificadores com N-gramas, para condados dos Estados Unidos 53
Figura 22 – Desempenho da hierarquia de classificadores com N-gramas combinados, para regiões dos Estados Unidos 54

Figura 23 – Desempenho da hierarquia de classificadores com N-gramas combinados, para estados dos Estados Unidos 55
Figura 24 – Desempenho da hierarquia de classificadores com N-gramas combinados, para condados dos Estados Unidos 56
Figura 25 – Proporção de n-gramas obtidos a partir dos nomes de locais do gazetteer 57

LISTA DE TABELAS

Tabela 1 – Nós XML de exportação da Wikipédia utilizados neste trabalho 39
Tabela 2 – Contagem de Ocorrências de nomes do gazetteer Web-a-Where 45
Tabela 3 – Contagem de Ocorrências de nomes do gazetteer, com iniciais minúsculas e maiúsculas, em uma coleção de páginas da Wikipédia. 45
Tabela 4 – Contagem de Ocorrências de nomes do gazetteer em uma coleção de páginas da Wikipédia versus população associada ao nome. 46
Tabela 5 – Quantidade de Atribuições de tags do Web-a-Where por nó da Taxonomia, Número de acertos por nó e global 47
Tabela 6 – Web-a-Where: Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para Estados e Condados dos Estados Unidos 48
Tabela 7 – As 15 cidades brasileiras com maior número de documentos georreferenciados na Wikipédia em português. 48
Tabela 8 – Os 15 condados dos EUA com maior número de documentos georreferenciados na Wikipédia em Inglês. 49
Tabela 9 – Os Estados dos EUA com maior e menor número de documentos georreferenciados na Wikipédia em Inglês. 49
Tabela 10 – Os estados brasileiros com maior e menor número de documentos georreferenciados na Wikipédia em Português. 50
Tabela 11 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para diferentes níveis da hierarquia geográfica dos Estados Unidos usando Naive Bayes 51
Tabela 12 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para diferentes níveis da hierarquia geográfica dos Estados Unidos usando SVM 51
Tabela 13 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios do Geofier usando Naive Bayes, sem um gazetteer. 55
Tabela 14 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para diferentes níveis da hierarquia geográfica dos Estados Unidos usando SVM 55

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Objetivos da Pesquisa 13
1.2 Organização da Dissertação 13

2 ANOTAÇÃO GEOGRÁFICA DE TEXTOS 14
2.1 Gazetteers Digitais 14
2.1.1 Ambiguidade de Topônimos 16
2.1.2 Escopo Geográfico 17
2.2 Exemplos de Sistemas de Anotação Geográfica 17
2.2.1 GIPSY 17
2.2.2 Web-a-Where 18
2.2.2.1 Desambiguação de Topônimos 19
2.2.2.2 Determinação do Escopo Geográfico 21
2.2.3 NewsStand 22
2.3 Desambiguação supervisionada de nomes de locais 24

3 CLASSIFICADORES DE TEXTOS 26
3.1 Vetorização de Documentos 26
3.2 Precisão, Revocação, Acurácia e F-score 27
3.3 TF-IDF 28
3.4 Classificadores Naive Bayes 29
3.5 Support Vector Machines (SVM) 30

4 O MÉTODO GEOFIER 33
4.1 Classificação para duas regiões 33
4.2 Método hierárquico para identificação de escopo geográfico de textos 35
4.3 Escolha dos classificadores do Geofier 37

5 AVALIAÇÃO DO GEOFIER 39
5.1 Extração de Documentos Georreferenciados da Wikipédia 39
5.2 Implementação do Web-a-Where 42
5.3 Hierarquia de Classificadores 48
5.4 Hierarquia de Classificadores com N-Gramas 51
5.5 Vocabulário Selecionado por tf-idf 55
5.6 Comparação entre Geofier e Web-a-Where 57

6 CONSIDERAÇÕES FINAIS 61

6.1 Trabalhos Futuros 61

Referências 63

1 INTRODUÇÃO

A busca por informações locais é facilitada se aos textos forem adicionados metadados que permitam a identificação do contexto geográfico ao qual o texto se refere. A importância do georreferenciamento da informação é justificada, por exemplo, pelos novos padrões de consumo de informação, principalmente nas consultas originadas de dispositivos móveis.

De acordo com o Google (FOX, 2011), em 2011 cerca de 20% de todas as consultas realizadas a partir de computadores pessoais foram efetuadas por usuários em busca de informação local. Este número chega a 40% para consultas realizadas a partir de dispositivos móveis.

Um estudo publicado pela empresa aferidora de audiência online Comscore em 2014 (LELLA ADAM; LIPSMAN, 2014) indicou que a maior parte do acesso à internet já não era realizado a partir de desktops, mas a partir de dispositivos móveis. Segundo este estudo aproximadamente 60% do tempo gasto online com mídias digitais era então realizado a partir destes dispositivos. Um novo estudo publicado em 2015 (LELLA ADAM; LIPSMAN, 2015) aponta o crescimento da categoria das pessoas que acessam a internet apenas através de seus telefones inteligentes.

A busca por informações sobre locais a partir de dispositivos móveis é uma consequência natural do fato de que pessoas em movimento geralmente buscam informações sobre pontos de interesse nas suas imediações. Anotar documentos da internet é um processo custoso e impraticável se realizado manualmente. Por esta razão, é fundamental encontrar um modo automático e eficiente para extrair informação geográfica de textos.

Vários projetos de pesquisa se dedicaram ao tema da extração de informação geográfica de textos, como o sistema GIPSY Woodruff e Plaunt (1994) e o NewsStand (SAMET et al., 2014; TEITLER et al., 2008). Entre estes sistemas o Web-a-Where (AMITAY et al., 2004; HAREL; AMITAY; SIVAN, 2006) se destaca por sua simplicidade e representatividade, graças ao uso de heurísticas populares aos demais sistemas de anotação geográfica de textos.

A determinação do escopo geográfico de um documento é feita a partir da identificação de nomes de locais, denominados topônimos, no texto. Os topônimos são obtidos por meio de consultas a um banco de dados de locais, denominado gazetteer.

A qualidade, abundância e atualidade das informações presentes no gazetteer são críticas para o desempenho dos métodos de anotação geográfica e sua manutenção é onerosa. A investigação de métodos que substituam ou automatizem a criação de um

gazetteer é portanto, fundamental.

Um dos principais problemas para a identificação de topônimos é que nomes de locais são frequentemente ambíguos, por exemplo, “São Paulo” pode se referir a uma cidade, um estado, um time de futebol ou uma figura religiosa. O nome precisa ser analisado dentro do contexto para que saibamos qual a interpretação correta. Heurísticas são utilizadas para identificar se os nomes encontrados no texto são de fato topônimos e a que locais se referem. Um levantamento das heurísticas mais utilizadas foi realizada por Leidner Leidner (2007).

Finalmente, os topônimos encontrados são analisados para determinar se o documento se refere a uma região específica e assim ser anotado com um rótulo que representa seu escopo, ou foco geográfico. A determinação do escopo geográfico de documentos é o problema estudado neste trabalho.

1.1 Objetivos da Pesquisa

O objetivo deste projeto de pesquisa é propor um método para determinação do escopo geográfico de documentos com o uso de classificadores de aprendizagem de máquina.

Este método não deve efetuar consultas a um Gazetteer contendo nomes de locais durante a determinação do escopo geográfico. O conhecimento geográfico necessário para a construção do método deve limitar-se aos nomes de regiões utilizadas para rotular os documentos da base de treinamento dos classificadores. Esta restrição possibilitará validar a hipótese de que os nomes de locais, tradicionalmente presentes no gazetteer, podem ser aprendidos automaticamente pelos classificadores a partir de uma coleção de textos georreferenciados.

A validação do método proposto será feita pela sua comparação com o método de anotação do Web-a-Where. Para isso será necessário implementar ambos, estabelecendo um patamar de comparação.

1.2 Organização da Dissertação

Esta dissertação está organizada da da seguinte maneira: Os capítulos 2 e 3 introduzem a anotação geográfica de textos e os classificadores de texto. O capítulo 4 apresenta o método Geofier. O capítulo 5 apresenta as fontes de dados e as transformações necessárias para sua utilização, bem como a avaliação de desempenho do Geofier em comparação com uma reimplementação do Web-a-Where. O capítulo 6 apresenta as considerações finais e os trabalhos futuros.

2 ANOTAÇÃO GEOGRÁFICA DE TEXTOS

A localização geográfica é parte frequente e importante da informação contida em textos mas sua identificação não é trivial, dada a variedade de formas e recursos que a linguagem escrita provê ao autor. A sintaxe da linguagem humana permite indireções, referências implícitas e ambiguidades de interpretação que são por vezes utilizadas intencionalmente, como recurso estilístico.

Documentos em formato eletrônico permitem que textos sejam anotados com metadados. Estes metadados em geral são específicos de aplicação e não é raro uma nova aplicação exigir a incorporação de novos metadados a um conjunto de documentos pré-existente.

A anotação geográfica de documentos consiste na adoção de metadados para a identificação de nomes de locais, chamados de topônimos, e a posição de suas ocorrências no texto como metadados. Estes metadados podem ser consultados por aplicações que se beneficiam deste tipo de informação, como é o caso de mecanismos de busca. A partir dos topônimos mencionados no texto é possível identificar o contexto espacial em que o assunto do texto está inserido, o que permite agrupar documentos que se refiram a um mesmo contexto, atribuindo ao documento um *escopo geográfico*.

Leidner (2007) define que o reconhecimento de topônimos “identifica um trecho de texto (i.e. posições inicial e final de caracteres) que constitui um topônimo, classificando-o como tal” e que a resolução de topônimos “[...]busca um conjunto de referentes candidatos (potenciais localidades) e aplica uma função de projeção que efetivamente escolhe o referente correto[...]” (LEIDNER, 2007, p. 29, tradução nossa). Para que um computador possa identificar nomes de locais em textos, distinguir precisamente a que local um nome se refere e determinar áreas mencionadas pelo texto em um mapa são necessários bancos de dados de nomes de locais. Estes bancos de dados são conhecidos como gazetteers.

2.1 Gazetteers Digitais

Um gazetteer digital é um banco de dados que armazena atributos associados a nomes de locais. Os atributos principais presentes em um gazetteer são **nome** (possivelmente incluindo variações de grafia), **localização** (coordenadas representando pontos, segmentos, ou áreas) e **tipo** (selecionado de um esquema de tipos de categorias para locais/acidentes geográficos) (HILL, 2000; GOODCHILD; HILL, 2008). No entanto, qualquer informação

associada a um local pode estar presente em um gazetteer digital, como a população do local, mapas, fotos e dados políticos e econômicos.

Alguns locais podem ser representados por uma única coordenada, como um monumento, ou o cume de uma montanha. Estes locais são frequentemente chamados de "Pontos de Interesse". Já Locais de grandes dimensões, como lagos, parques ou regiões administrativas podem ser representados por sequências de n coordenadas geográficas com $n > 2$, formando os vértices de um polígono fechado de n lados.

Regiões administrativas são áreas geográficas delimitadas a fim de auxiliar na organização de um território. São regiões administrativas, por exemplo, as 5 regiões brasileiras (Centro-Oeste, Norte, Nordeste, Sul e Sudeste), as unidades federativas do Brasil (estados) e os municípios brasileiros, os estados e condados dos Estados Unidos, além dos próprios países. As regiões administrativas possuem uma organização hierárquica com relações de pertinência, que convenciona-se chamar "hierarquia geográfica". O Estado de São Paulo "pertence" à região Sudeste do Brasil, por exemplo.

O Getty Thesaurus of Geographic Names Online (GETTY, 1987), o GeoNames (GEONAMES, 2005) e o OpenStreetMap Nominatim (OSM FOUNDATION, 2009) são exemplos notórios de gazetteers que podem ser consultados pela internet. A figura 1 mostra um exemplo de consulta a um gazetteer online. Nesta consulta estão evidentes o nome do local (Santa Mariana), um mapa com o polígono que representa suas fronteiras, rótulos com informações adicionais, como população (12437), sua hierarquia geográfica (Município, Microrregião, Mesorregião, Estado, Região e País) e outras informações específicas deste gazetteer.

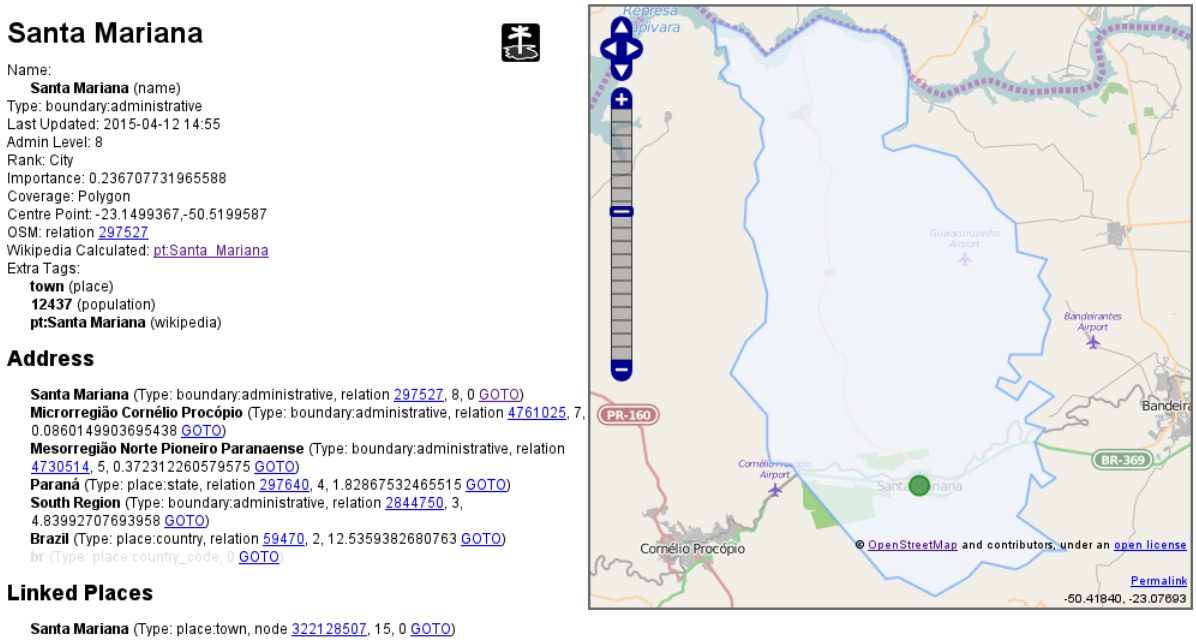
Os topônimos a serem buscados em um documento que será anotado geograficamente são tradicionalmente consultados a partir de um gazetteer, a exemplo dos sistemas NewsStand (SAMET et al., 2014; TEITLER et al., 2008), Web-a-Where (AMITAY et al., 2004; HAREL; AMITAY; SIVAN, 2006) e Gipsy (WOODRUFF; PLAUNT, 1994).

Uma vez encontrados nomes de locais no texto, o gazetteer é novamente consultado e as informações associadas aos topônimos são utilizadas para a desambiguação da interpretação dos topônimos identificados e a posterior determinação do escopo geográfico do documento.

2.1.1 Ambiguidade de Topônimos

A partir dos dados do censo demográfico brasileiro de 2010 disponibilizados pelo IBGE (IBGE, 2010a) é possível observar que aproximadamente 4% dos nomes de municípios então existentes no Brasil referiam-se a mais de uma localidade. "Bom Jesus" (PB, PI, RN, RS e SC) e "São Domingos" (BA, GO, PB, SC, e SE) são os nomes de municípios brasileiros com mais ocorrências repetidas, cada um presente em 5 estados diferentes.

Figura 1 – Exemplo de consulta a um gazetteer. Detalhe do topônimo “Santa Mariana”.



Fonte: Nominatim, OSM Foundation (2009)

O problema da ambiguidade de nomes de locais foi investigado desde os primeiros experimentos para o reconhecimento e resolução de topônimos. Woodruff e Plaunt (1994) são um bom exemplo de enumeração das diferentes maneiras com as quais locais falham em ser identificados corretamente por seus nomes. Entre os principais fatores que contribuem para a ambiguidade de topônimos, encontram-se:

- **Nomes de locais raramente são únicos**. Por exemplo, Cambridge pode se referir a Cambridge, Massachusetts ou Cambridge, Inglaterra;
- **Fronteiras políticas mudam com o tempo, como resultado de tratados ou guerras**. Nomes de locais mudam, o que faz com que seja difícil manter uma lista atualizada de nomes ao mesmo tempo em que se preserva sua utilidade para referências em documentos históricos;
- **Variação de grafia**. Nome de locais são frequentemente escritos de maneiras diversas em diferentes línguas.

Amitay et al. (2004) classificam os principais tipos de ambiguidade toponímica em **Geo/Geo** e **Geo/Non-Geo** pois locais podem ser confundidos com outros locais homônimos ou com palavras que não representem um lugar, respectivamente. Ao lidar com nomes de locais em uma coleção de textos históricos Smith e Crane (SMITH; CRANE, 2001) observaram que de todos os topônimos em seus documentos, 92% podiam ser traduzidos em mais de uma coordenada geográfica.

A ambiguidade do tipo Geo/Non-Geo, ou seja, nomes que não representem locais, mas são confundidos com nomes de locais (“Mariana” e “Mariana, MG”, por exemplo) depende da interpretação semântica do texto e da área geográfica sendo considerada. Por outro lado, a ambiguidade do tipo Geo/Geo, locais diferentes que possuem o mesmo nome (“California, Estados Unidos” e “Califórnia, Paraná”, por exemplo) pode ser quantificada, desde que um gazetteer suficientemente completo esteja disponível.

2.1.2 Escopo Geográfico

O escopo ou foco geográfico é um atributo de textos relacionados a locais. Andogah, Bouma e Nerbonne (2012) afirmam que todo documento possui um escopo geográfico. Ding, Gravano e Shivakumar (2000) definem o escopo geográfico de um recurso *w* da web como sendo “a área geográfica que o criador de *w* deseja atingir”, interpretação compartilhada por Buyukkokten et al. (1999) que também investigou o escopo geográfico de uma página web como sendo relacionado à “posição geográfica do web site em que a página reside”, Amitay et al. (2004) entendem o “foco geográfico” de um documento como sendo a região ou regiões nas quais se encontram a maior parte dos locais mencionados no texto.

Teitler et al. (2008) fazem referência a três tipos de escopo geográfico:

- **Escopo do Fornecedor** (*provider scope*) a localização geográfica do editor do documento;
- **Escopo do Conteúdo** (*content scope*) a geografia do conteúdo do documento; e
- **Escopo de Entrega** (*serving scope*) baseado na região do leitor ou consumidor do conteúdo.

Este trabalho de pesquisa adota a expressão **escopo geográfico** para referir-se ao **escopo do conteúdo** de documentos. Por escopo geográfico entende-se a região geográfica que abrange os topônimos mencionados no texto.

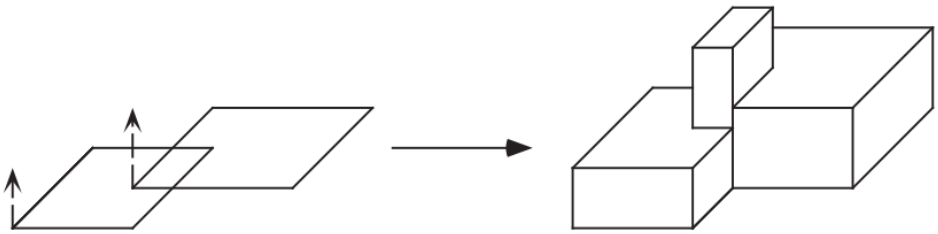
2.2 Exemplos de Sistemas de Anotação Geográfica

2.2.1 GIPSY

Um dos primeiros sistemas para anotação geográfica foi o GIPSY (Geographical Information Processing System, ou Sistema de Processamento de Informação Geográfica), desenvolvido por Woodruff e Plaunt (WOODRUFF; PLAUNT, 1994). O sistema GIPSY utiliza um algoritmo de três passos para a anotação geográfica. O primeiro passo identifica palavras e frases que possuem relevância geográfica, comparando-as com os nomes existentes num gazetteer. O segundo passo busca representações poligonais das áreas

mencionadas. O último passo combina os polígonos em uma representação tridimensional, tal como a apresentada na Figura 2. A sobreposição de polígonos das áreas mencionadas no texto cria um relevo tridimensional, o patamar mais alto deste relevo define um contorno que indica, graficamente, o escopo geográfico do documento analisado em um mapa.

Figura 2 – Sobreposição de áreas geográficas.



Fonte:Woodruff e Plaunt (1994)

Embora tenha apresentado bons resultados iniciais, as operações utilizando polígonos eram excessivamente custosas, fato observado pelos próprios autores.

2.2.2 Web-a-Where

O sistema Web-a-Where (AMITAY et al., 2004) foi um dos primeiros a propor a anotação geográfica automática de conteúdo da Web. A figura 3 exibe o algoritmo de anotação geográfica implementado pelo Web-a-Where.

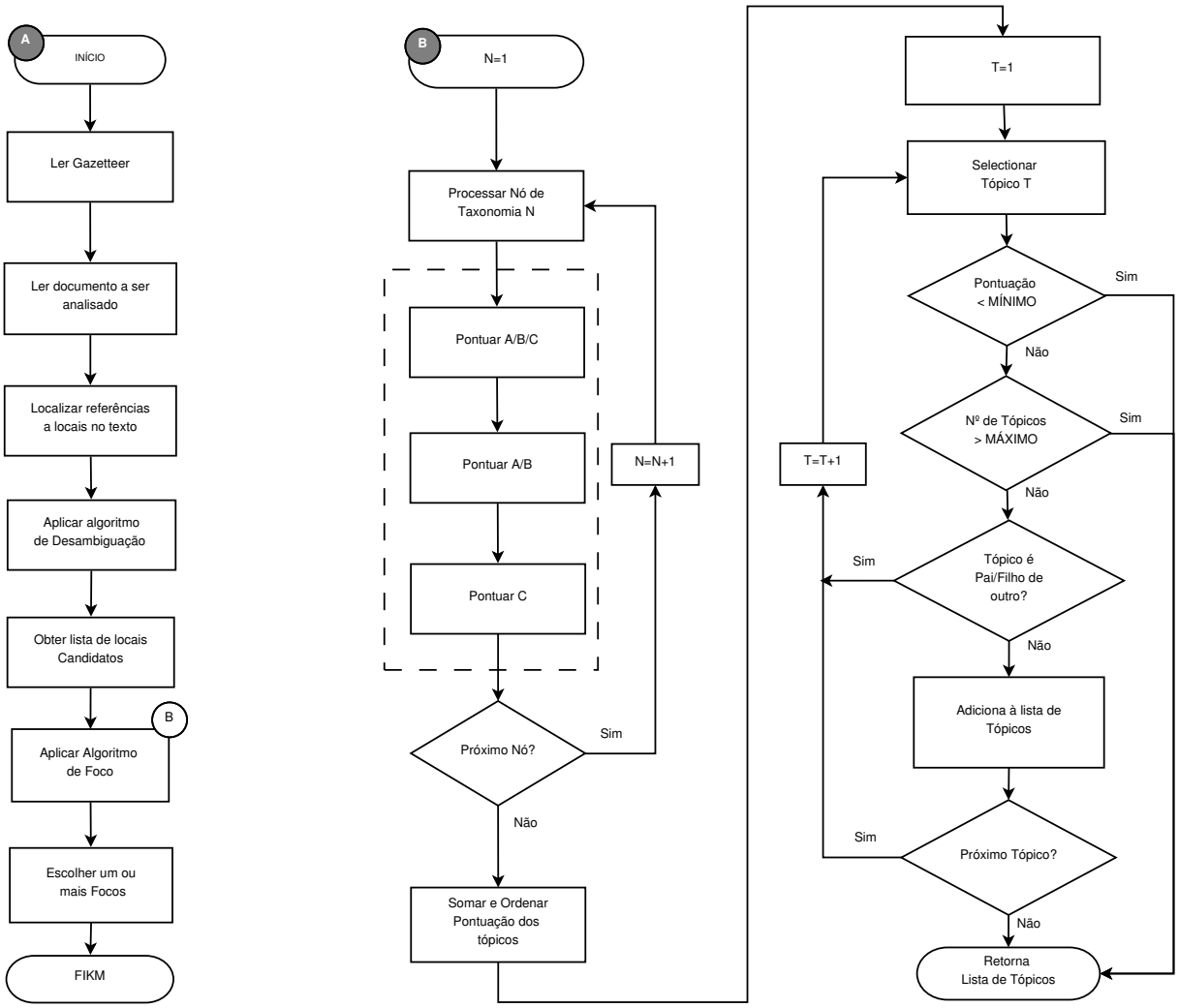
O primeiro passo executado pelo Web-a-Where é a busca nomes de cidades, estados e países presentes em seu gazetteer e a identificação de ocorrências destes topônimos nos textos. Todas as possíveis interpretações geográficas de cada nome são selecionadas e então o algoritmo irá decidir qual interpretação é a correta, atribuindo um valor real entre 0 e 1, como estimativa de confiança para aquela interpretação.

O gazetteer do Web-a-Where contém uma seção especial de termos que são nomes de locais, mas que tem significado não-geográfico mais frequente. As regras para resolver a ambiguidade do tipo Geo/Geo e Geo/Non-Geo utilizadas pelo Web-a-Where são:

Desambiguação Geo-Non-Geo

O Web-a-Where utiliza uma heurística baseada na contagem de frequência de termos (AMITAY et al., 2004, p. 275) para decidir se um nome identificado é realmente um nome de local ou se possui significado não geográfico. Para cada termo presente no gazetteer, é feita uma contagem do número de ocorrências do termo com letras iniciais maiúsculas, n_{mai}, e em letras minúsculas, n_{min}. Se um termo aparecer mais de

Figura 3 – Algoritmo do sistema Web-a-Where



Fonte: Adaptado de Harel, Amitay e Sivan (2006, fig. 5 e 6)

100 vezes ($n_{mai}+n_{min} \geq 100$) e na maioria das vezes grafado com iniciais minúsculas, ou seja, $\frac{n_{min}}{n_{mai}+n_{min}} > 0.5$, ele deve ser tratado como um termo não geográfico.

Se a contagem de ocorrências do termo em letras minúsculas for muito maior do que a população atribuída àquele local, então este termo é provavelmente não-geográfico e deve ser tratado de forma diferenciada. Esta contagem deve ser feita em uma grande coleção de páginas web. Os autores não especificam parâmetros para número ou tamanho mínimo de documentos, seus testes utilizaram uma coleção de 1.200.000 páginas que faziam parte da estrutura de domínios ".gov". Os autores notam que foi necessário analisar e tratar individualmente os falsos positivos e falsos negativos da lista de nomes selecionados por esta técnica.

Desambiguação Geo-Geo

Os autores do Web-a-Where assumem que a interpretação preferencial é sempre a do local com maior população (AMITAY et al., 2004, p. 276), na falta de outras evidências,

como um topônimo adjacente que torne a interpretação única, p. ex. “Los Angeles, California”.

2.2.2.1 Desambiguação de Topônimos

A Figura 4 apresenta o algoritmo de desambiguação de topônimos em pseudocódigo. Algumas funções foram introduzidas para torná-lo mais conciso e encapsular alguns conceitos do Web-a-Where. A interpretação padrão de um topônimo, para o Web-a-Where é o local mais populoso dentre os locais que compartilham um nome.

Figura 4 – Web-a-Where: Algoritmo de desambiguação de topônimos

Dados: Conjunto das menções a topônimos encontrados no texto, T.
Resultado: Lista de topônimos, ordenadas por peso.

```
1  Q ← ∅
2  para cada topônimo | topônimo ∈ T faça
3      se CompletamenteQualificado(topônimo) então
4          topônimo.peso ← 0.95
5          topônimo.significado ← SignificadoQualificado(topônimo)
6          Q ← Q ∪ {topônimo}
7  para cada topônimo | topônimo ∈ T \ Q faça
8      se ∃x ∈ Q | x.nome = topônimo.nome então
9          topônimo.significado = x.significado
10         se InterpretaçãoPadrão(topônimo) = InterpretaçãoPadrão(x) então
11             topônimo.peso← 0.90
12         senão
13             topônimo.peso← 0.85
14     senão
15         topônimo.significado ← InterpretaçãoPadrão(topônimo)
16         topônimo.peso ← 0.5
17 NQ = {x ∈ T | x.peso < 0.7}
18 contexto ← CalculaContexto(NQ)
19 se contexto != NULL então
20     para cada topônimo ∈ {x ∈ T | x.peso < 0.7} faça
21         significado ← InterpretaçãoPadrão(topônimo)
22         se FazParte(significado,contexto) então
23             topônimo.peso← 0.75
24         senão
25             topônimo.peso← 0.65
26 lista = OrdenaPorPeso({x | x ∈ T})
27 retorna lista
```

Fonte: elaborado pelo autor, a partir de Amitay et al. (2004)

O Web-a-Where atribui um peso para cada topônimo encontrado no texto, estes

pesos são valores reais entre 0.0 e 1.0 e representam uma medida da certeza de que aquele topônimo se refere a uma dada interpretação de local.

Inicialmente, identificam-se os topônimos completamente qualificados. Estes topônimos são caracterizados pela junção de dois topônimos que identifiquem um local unicamente, como “Chicago, IL”. Topônimos completamente qualificados recebem o peso 0.95, atribuído na linha 4.

Se um topônimo já foi completamente qualificado, mas aparece desacompanhado de seu qualificador, atribuir-se-á um valor entre 0.8 e 0.9 para esta ocorrência. Se a interpretação padrão do nome do local coincide com a do topônimo completamente qualificado, o valor atribuído será 0.9, caso contrário, 0.8 (linhas 8 a 13).

Todas as demais ocorrências de topônimos receberão o peso 0.5 e serão associadas às interpretações padrão de seus nomes: O local com a maior população.

Para todos os topônimos que não atingiram a pontuação mínima de 0.7, busca-se um contexto de desambiguação.

Para que tal contexto exista, ele deverá abranger a maior parte dos topônimos restantes. Na Figura 5 “Texas” é um contexto de desambiguação para os topônimos “Orlando”, “Garland”, “Forth Worth” e “Dallas”, pois a maioria dos topônimos analisados são subordinados a este nó da hierarquia geográfica.

Topônimos que fazem parte do contexto de desambiguação recebem um peso com valor 0.75 ou 0.65, dependendo se a interpretação do topônimo no contexto coincide com a interpretação padrão do topônimo ou não (linhas 19 a 25).

O algoritmo de desambiguação retorna a lista de topônimos em ordem decrescente de pesos atribuídos.

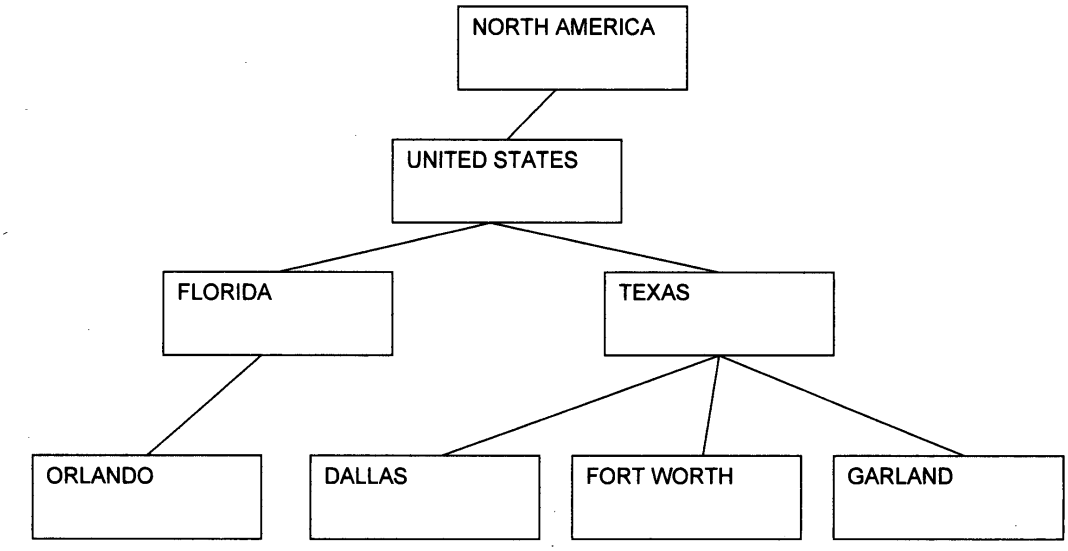
2.2.2.2 Determinação do Escopo Geográfico

O Web-a-Where modela o escopo geográfico como uma referência a um nó de uma hierarquia geográfica, composta por Continente, País, Estado e Local, conforme a figura 5.

O escopo geográfico é determinado através de um sistema de pontuação que utiliza o peso p atribuído a cada topônimo identificado no passo anterior. Ao encontrar um topônimo que representa uma cidade, soma-se o valor p^2 para o foco geográfico no formato Cidade/Estado/País.

Este peso é propagado para o escopo geográfico hierarquicamente superior, com um fator de atenuação $d = 0.7$. Após processar uma menção ao escopo geográfico Cidade/Estado/País e somar p^2 a sua pontuação, os valores p^2d e p^2d^2 são somados às pontuações dos escopos Estado/País e País, respectivamente. A propagação de pesos está assinalada com uma caixa tracejada na figura 3.

Figura 5 – Hierarquia Geográfica utilizada pelo Web-a-Where



Fonte:Harel, Amitay e Sivan (2006, fig. 3)

Os escopos geográficos são finalmente ordenados de acordo com sua pontuação e os quatro maiores entre aqueles com valor maior que 0.9 são selecionados para representar o(s) escopo(s) geográfico(s) do documento. Os valores de d e p utilizados pelo Web-a-Where foram determinados de forma empírica por seus autores.

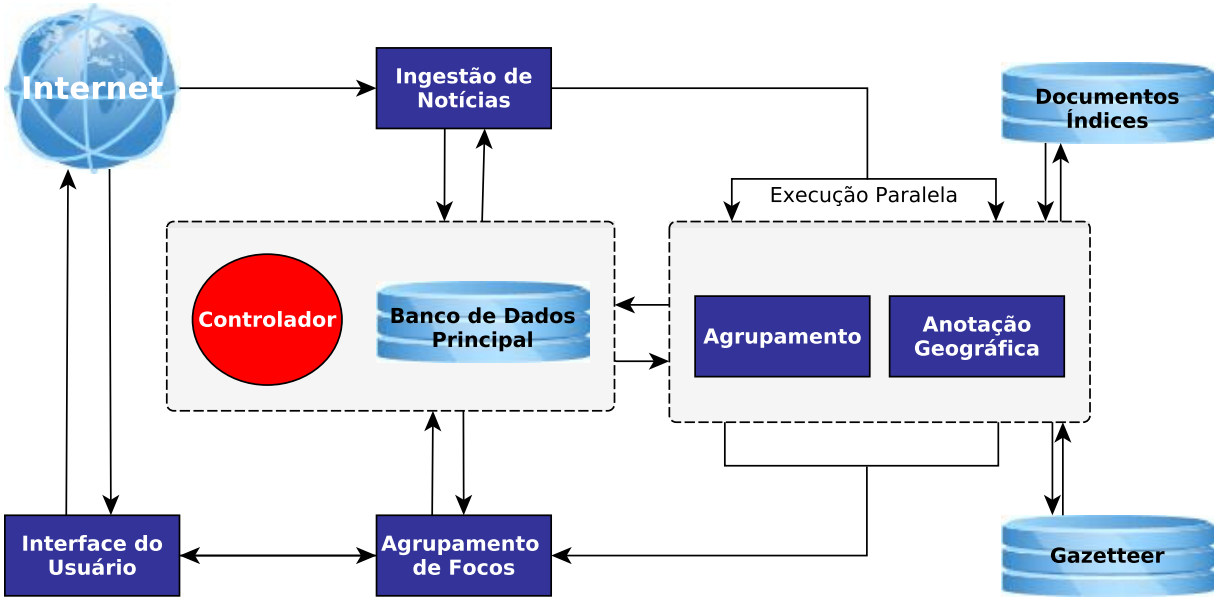
O Web-a-Where foi testado com um conjunto de páginas web classificadas manualmente (YAHOO, 1999) e foi capaz de identificar corretamente a que país um texto fazia referência em 92% dos casos, dos quais apenas 38% identificavam corretamente o estado ou cidade a que o texto se referia.

2.2.3 NewsStand

O NewsStand é um sistema que coleta, analisa e exibe notícias em uma interface de mapas (SAMET et al., 2014; TEITLER et al., 2008). Seus autores argumentam que embora mecanismos de busca de notícias existam, eles geralmente se concentram em palavras-chave ou tópicos e não lidam bem com dois tipos comuns de consulta espacial: “Onde o fato X aconteceu?” e “O que aconteceu no local X?”.

O NewsStand reconhece três tipos de escopo geográfico: do conteúdo, da fonte e do leitor (TEITLER et al., 2008) e utiliza elementos dos três contextos para a seleção e apresentação dos documentos para o usuário final por meio de uma interface baseada em mapas. Múltiplos subsistemas realizam as tarefas de coleta online de notícias, anotação geográfica, agrupamento por regiões e tópicos e exibição para o usuário final, conforme a figura 6.

Figura 6 – Arquitetura NewsStand



Fonte: Adaptado de Samet et al. (2014, Fig. 4)

A tarefas de anotação geográfica e determinação de escopo do NewsStand são derivadas de trabalhos anteriores de seus autores, principalmente do sistema STEWARD (LIEBERMAN et al., 2007) que foi estendido e especializado para o georreferenciamento de notícias online.

O sistema STEWARD combina o uso de um gazetteer com técnicas de processamento de linguagem natural para extrair frases do texto que potencialmente se refiram a locais, além do cálculo de TF-IDF para palavras individuais do texto (unigramas). A seleção de termos candidatos a nomes de locais encontrados no texto é chamado de vetor de características (*feature vector*) do documento. O vetor de características (*feature vector*) de um documento é um vetor contendo todos os locais mencionados no texto. O processo da obtenção do vetor de características é denominado extração do vetor de características (*feature vector extraction*).

Para cada termo do vetor de características encontrado no gazetteer são agregadas as suas intepretações geográficas. Estas potencialmente múltiplas interpretações são chamadas de *feature records* ou registros de características. A determinação a interpretação correta dentre os registros de características, a desambiguação dos topônimos propriamente dita, é realizada pelo subsistema especializado seguinte.

O sistema NewsStand verifica a relação entre pares de “topônimos encontrados no texto” e “atributos encontrados no gazetteer”. Esta heurística, batizada de *pair strength*, ou força de pares, compara pares de registros de características para verificar se eles são relacionados, baseados numa heurística que combina a distância em caracteres medida entre as menções dos pares dentro do texto, a distância geográfica entre os locais a que se

referem e suas respectivas populações.

Os pares com maior pontuação são escolhidos como verdadeiros e removidos da lista, removendo também interpretações que se tornem inválidas. Por exemplo, se “Springfield, IL” estiver em um par escolhido como verdadeiro, todos os pares que incluam “Springfield, MA” serão removidos.

O algoritmo de desambiguação termina quando todos os pares de registros de características forem removidos da lista, determinando a interpretação de cada registro, individualmente.

A Determinação do escopo geográfico é então realizada por um algoritmo denominado CARD ou Context-Aware Relevancy Determination (Determinação de Relevância Sensível ao Contexto) em que dois locais l_1 e l_2 são contextualmente relacionados em um documento d se l_1 e l_2 ocorrem com frequência dentro dos contextos um do outro em d. Por “contexto” entende-se proximidade dentro do texto, conforme descrito por(LIEBERMAN; SAMET, 2012).

Diz-se que um local *l* é importante para *d* se *l* está bem distribuído pelo documento *d* e ao mesmo tempo relacionado contextualmente a vários locais geograficamente próximos em *d*.

Estas informações são combinadas em uma pontuação que é utilizada para a determinação do escopo geográfico do texto.

2.3 Desambiguação supervisionada de nomes de locais

Overell (OVERELL, 2009) utiliza a WordNet (MILLER, 1995) para obter uma lista de termos que se referem a locais e identifica os verbetes correspondentes na Wikipédia. No passo seguinte, elencam-se pares de nomes de locais que ocorrem dentro de um mesmo verbete. A ideia principal é que locais geograficamente próximos são frequentemente mencionados no mesmo documento. A frequência com que pares de locais ocorrem juntos é utilizada para determinar o escopo geográfico correto.

Overell descreve e valida experimentalmente o que batizou de “Hipótese de Steinberg”, que afirma que locais geograficamente próximos ao assunto do texto e a seu autor possuem relevância muito maior do que locais distantes.

Estas abordagens representativas das técnicas de determinação de escopo geográfico e de extração de topônimos baseiam-se em consultas a listas de nomes de locais para a identificação de topônimos em um texto. A extensão e correção destas listas influencia diretamente na qualidade dos algoritmos, como observador por Amitay (AMITAY et al., 2004).

O estudo realizado por Overell indica que a relação entre os topônimos dentro do

texto e sua proximidade física são muito mais relevantes para a determinação da interpretação correta do local referenciado por um topônimo do que sua população. Parafraseando o Geógrafo Waldo Tobler e sua “Primeira Lei da Geografia”: “Tudo é relacionado a tudo mais, mas o que está próximo é mais relacionado do que o que está distante” (TOBLER, 1970, tradução nossa).

Um estudo da validade da primeira lei da geografia utilizando artigos georreferenciados da Wikipédia foi realizado por Hecht e Moxley (2009). Eles estudaram a estrutura de links dos verbetes georreferenciados da Wikipédia e concluíram que a probabilidade de existir um vínculo entre dois verbetes é 245 vezes maior para artigos distantes 50km do que a média de probabilidade entre dois verbetes aleatórios. Esta tendência é atenuada rapidamente e praticamente deixa de existir para artigos distantes pelo menos 650Km.

3 CLASSIFICADORES DE TEXTOS

Segundo Sebastiani (2002), a *categorização de textos* ou *classificação de textos* é a "atividade de rotular textos em linguagem natural com categorias temáticas obtidas a partir de um conjunto pré-definido". Ainda segundo Sebastiani, até os anos 80 esta tarefa era predominantemente realizada através de conjuntos de regras, obtidas a partir de conhecimento especializado em como classificar documentos nas categorias escolhidas.

A partir dos anos 90 esta abordagem baseada em regras deu lugar ao paradigma da *aprendizagem de máquina*, através do qual um processo indutivo constrói automaticamente um classificador de textos, "aprendendo" as características de cada categoria a partir de um conjunto de documentos previamente classificados.

Formalmente, a categorização de textos é a tarefa de designar um valor booleano para cada par $\langle d_j, c_i \rangle \in \mathcal{D} \times \mathcal{C}$, em que $\mathcal{D}$ representa o domínio de documentos e $\mathcal{C} = \{c_1, c_2, \ldots, c_{|C|}\}$ representa as classes, de acordo com as quais um documento será classificado. O valor Verdadeiro atribuído a $\langle d_j, c_i \rangle$ simboliza a decisão de atribuir d_j a c_i, o valor Falso simboliza a decisão de não atribuir d_j a c_i.

Mais formalmente, a tarefa de classificação automática de textos é a tarefa de aproximar a *função alvo* $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$ por meio de uma função $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$, chamada de classificador, hipótese ou modelo, tais que $\breve{\Phi}$ e Φ "coincidam tanto quanto possível". A medida desta coincidência é chamada "efetividade".

3.1 Vetorização de Documentos

Documentos em linguagem natural não são adequados para a análise por classificadores de aprendizagem de máquina. Por isso, buscam-se representações simplificadas que possam ser aplicadas aos documentos.

Um método possível é a Vetorização de Documentos, onde se transforma o documento em um vetor de pesos $\vec{d} = \langle w_{1j}, \ldots, w_{|\tau|j} \rangle$ em que τ é o conjunto dos termos $w_{i,j}$ que são encontrados ao menos uma vez em pelo menos um dos documentos do corpus.

A escolha do que configura um termo e como são calculados os pesos depende do domínio da aplicação. Se cada "termo" corresponder a uma palavra diferente do texto temos o modelo conhecido como *saco de palavras* (*bag of words*). Os termos podem também ser escolhidos através de um vocabulário controlado pré-existente.

Além da maneira como os termos do vocabulário são escolhidos, a vetorização dos documentos também difere na maneira como são calculados os pesos de cada termo.

O peso pode ser um valor booleano, indicando presença ou ausência de determinado termo do vocabulário no documento. Também é comum que os pesos representem as contagens do número de vezes que um termo aparece no documento.

O modelo do saco de palavras leva em conta cada palavra individualmente, ignorando as relações entre as palavras dentro do texto. Uma maneira simples de capturar a estrutura sintática da linguagem é estender o termo para conjuntos consecutivos de n palavras, conhecidos como *n-gramas*. Na prática se observa que n-gramas de tamanho maior que 2 pouco contribuem para as métricas estabelecidas de efetividade de classificação (MLADENIC; GROBELNIK, 1998; FURNKRANZ, 1998).

3.2 Precisão, Revocação, Acurácia e F-score

Precisão e Revocação (*precision* e *recall*) são métricas básicas definidas originalmente para a avaliação da efetividade de buscas feitas por um sistema de recuperação de informação simples (MANNING; RAGHAVAN; SCHUTZE, 2008).

A precisão P diz respeito à fração de documentos recuperados por uma busca que são relevantes para a consulta realizada. A revocação R mede a fração de documentos relevantes para uma consulta que são efetivamente recuperados pela busca.

Estas métricas também se aplicam a classificadores binários. Neste caso a precisão P é definida como a fração de itens atribuídos a uma determinada classe que realmente pertenciam a ela (equação 3.1) e a revocação R é a fração dos itens que pertencem a uma determinada classe e que foram corretamente atribuídos (equação 3.2).

$$P = \frac{\mid itens\ classificados\ corretamente \mid}{\mid total\ de\ itens\ classificados \mid} \tag{3.1}$$

$$R = \frac{\mid itens\ classificados\ corretamente \mid}{\mid total\ de\ itens\ pertencentes\ à\ classe \mid} \tag{3.2}$$

A métrica F-score F, combina precisão e revocação em um único valor, útil para a comparação entre classificadores. O F-score é a média harmônica entre precisão e revocação.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad = \quad \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{onde} \quad \beta^2 = \frac{1-\alpha}{\alpha} \tag{3.3}$$

O F-score permite dosar qual a influência de P e R na média, variando-se α. Para o caso comum em que P e R tenham a mesma relevância, temos $\alpha = \frac{1}{2}$ e $\beta = 1$, resultando

na medida F_1 ou, mais precisamente, $F_{\beta=1}$:

$$F_{\beta=1} = 2\frac{P.R}{P+R} \tag{3.4}$$

Para uma tarefa de classificação com múltiplas classes, a Precisão, Revocação e o F-score são calculados para cada classe individualmente. Uma alternativa é obter seu valor médio, através do cálculo da média aritmética do F-score de todas as classes.

A acurácia A é a fração das classificações corretas em relação ao total de classificações realizadas. Sejam N o total dos documentos usados pelo classificador C, $f(d) = 1$ se o documento d é classificado corretamente por C, e $f(d) = 0$ se o documento d não é classificado corretamente por C, a acurácia é definida como:

$$A = \frac{\sum_{k=1}^{N} f(d_k)}{N} \tag{3.5}$$

Estas medidas devem ser observadas em conjunto, na avaliação comparativa de desempenho entre classificadores distintos.

3.3 TF-IDF

Term Frequency (TF) (LUHN, 1957) e Inverse Document Frequency (IDF) (JONES, 1972) ou frequência do termo e frequência inversa de documentos é uma medida que busca avaliar o quanto um termo é importante para um documento dentro de uma coleção de documentos e carrega mais informação do que a mera frequência de ocorrência do termo nos documentos da coleção.

Duas importantes medidas em recuperação de informação são *precisão* e *revocação* (*precision* e *recall*). Em uma busca dentro de uma coleção de documentos, a precisão mede a porcentagem dos documentos retornados por uma busca, que são realmente relevantes. A Revocação mede a porcentagem de documentos relevantes que foram retornados pela busca.

Salton e Buckley (1987) observam que termos que aparecem com muita frequência em um documento são úteis para aumentar a revocação em uma busca. Estes termos devem, portanto fazer parte de uma medida de importância de um termo para uma coleção.

Termos com alta frequência sozinhos não são suficientes para assegurar um bom desempenho na recuperação de documentos, pois nas situações em que estes termos ocorram por toda a coleção, uma busca que privilegie estes termos irá recuperar muitos documentos, impactando na medida de precisão da busca.

Termos relevantes devem ter alta frequência em um documento, mas um baixo número de ocorrências dentro da coleção. Uma medida razoável da importância de um

termo para a coleção deve então ser composta do produto da frequência do termo pelo inverso da frequência na coleção.

Dados um termo t e um documento d pertencente a uma coleção D, definimos tf e idf:

$$tf(t,d) = \textit{número de ocorrências de t em d} \tag{3.6}$$

$$idf(t,D) = \log \frac{N}{n} \tag{3.7}$$

Onde n é o número de documentos que contém o termo t e N é o número total de documentos em D. O score tf-idf de um termo t pode então ser calculado da seguinte forma:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{3.8}$$

Onde t é um termo, d é um documento, n é o número de documentos em que o termo t aparece e N é o total de documentos na coleção.

Em modelos de linguagem aplicados à recuperação de informação é comum que se ajuste o modelo para se compensar termos ainda não vistos em documentos, evitando-se o denominador zero no cálculo de idf na equação 3.7, por exemplo. Esta técnica é conhecida como suavização (*smoothing*) e visa compensar os efeitos de um modelo baseado em características esparsas. Um estudo das diferentes técnicas de suavização em modelos de linguagem e seus efeitos na recuperação de informação foi realizado por Zhai e Lafferty (2004).

3.4 Classificadores Naive Bayes

O classificador Naive Bayes pode ser utilizado em tarefas de aprendizagem de máquina em que cada instância x de exemplo é descrita por uma conjunção de valores de atributos, e a função alvo $f(x)$ a ser aprendida pode assumir um número finito de valores de um conjunto V (MITCHELL, 1997).

Um conjunto de exemplos de treinamento é apresentado, descrito por uma tupla de valores de atributos $\langle a_1, a_2, \ldots, a_n \rangle$. Ao ser aplicado a uma nova instância desconhecida, o classificador Naive Bayes escolherá o valor mais provável da função alvo, dados os atributos que descrevem a nova instância, conforme a equação 3.9.

$$v_{MAP} = \underset{v_j \in V}{\arg\max}\, P(v_j \mid a_1, a_2, \ldots, a_n) \tag{3.9}$$

O teorema de Bayes expressa a relação entre uma probabilidade condicional e sua inversa, e é classicamente descrito da seguinte forma:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3.10}$$

A relação expressa pelo teorema de Bayes permite reescrever a equação 3.9:

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n \mid v_j)}{P(a_1, a_2, \ldots, a_n)} \tag{3.11}$$

Como o denominador é constante para todos valores de v_j , ele não terá influência no valor máximo e pode ser ignorado.

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n \mid v_j) \tag{3.12}$$

Os classificadores do tipo Naive Bayes assumem que os atributos são condicionalmente independentes, o que permite simplificar o cálculo da probabilidade da conjunção dos atributos $a_1 \ldots a_n$ como o produto das probabilidades individuais $P(a_1, a_2 \ldots a_n \mid v_j) = \prod_i P(a_i \mid v_j)$, o que leva à forma final do classificador Naive Bayes:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \tag{3.13}$$

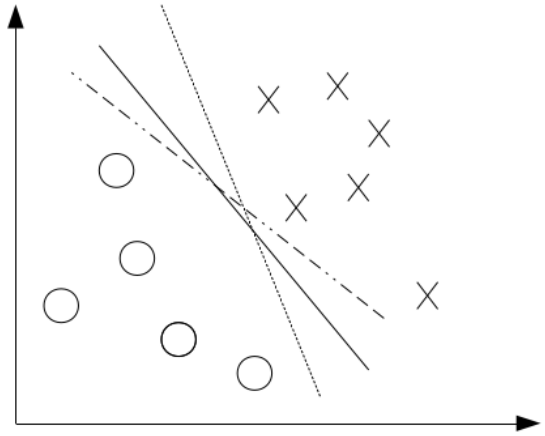
O modelo do saco de palavras representa um texto como o conjunto das palavras que o compõem, independentemente de ordem ou posição. A probabilidade de cada palavra em relação às classes é obtida através da contagem de ocorrências da palavra em uma determinada classe em relação ao número de ocorrências da palavra em todo o conjunto de treinamento. As probabilidades de cada classe são estimadas pelo número de exemplos da classe em relação ao total de exemplos disponíveis.

3.5 Support Vector Machines (SVM)

Um classificador SVM (Support Vector Machine) é um classificador definido por um hiperplano que separa instâncias de exemplos em duas classes. A figura 7 apresenta o problema em duas dimensões, para melhor ilustração da proposta. Em duas dimensões o hiperplano pode ser representado por uma reta.

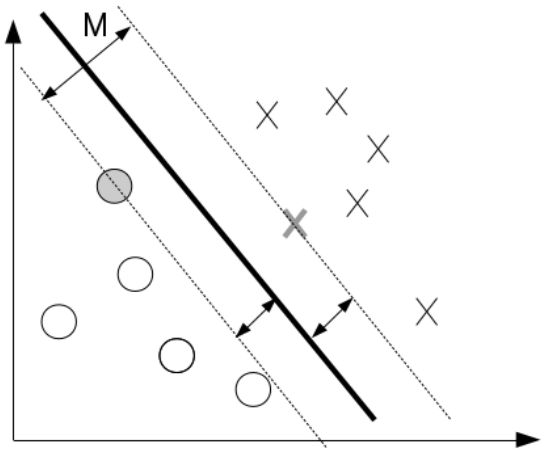
O classificador SVM busca encontrar o hiperplano que melhor separa as instâncias positivas e negativas, maximizando a distância entre os exemplos mais próximos do hiperplano ótimo. Estes exemplos são chamados de Support Vectors. A figura 8 ilustra um hiperplano que maximiza a margem (duas vezes a distância dos Support Vectors ao hiperplano) dos dados de treinamento.

Figura 7 – Hiperplanos separando duas classes de exemplos



Fonte: adaptado de Hastie, Tibshirani e Friedman (2009, p. 129)

Figura 8 – Hiperplano ótimo separando duas classes de exemplos e seus support vectors



Fonte: Fonte: adaptado de Hastie, Tibshirani e Friedman (2009, p. 134)

A fórmula do hiperplano canônico é dada por:

$$|\beta_0 + \beta^T x| = 1 \tag{3.14}$$

Em que x representa os support vectors, ou exemplos mais próximos do hiperplano. A distância entre um ponto e o hiperplano é dado pela equação 3.15.

$$\text{dist} = \frac{|\beta_0 + \beta^T x|}{||\beta||}. \tag{3.15}$$

De 3.14 e 3.15 temos que

$$\text{dist}_{\textit{support vectors}} = \frac{1}{||\beta||}. \tag{3.16}$$

A margem a ser maximizada equivale a duas vezes a distância entre os support vectors e o hiperplano.

$$M = \frac{2}{||\beta||} \tag{3.17}$$

A escolha de β e β_0 que maximiza a margem M é obtida através da minimização da lagrangiana, com respeito a β e β_0.

$$\min_{\beta,\beta_0} \mathcal{L}(\beta) = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(x_i^T\beta + \beta_0) - 1\right[\tag{3.18}$$

4 O MÉTODO GEOFIER

O método Geofier identifica o escopo geográfico de um texto por meio do uso de uma hierarquia de classificadores de texto. O Geofier utiliza a estratégia de divisão e conquista (*divide and conquer* sobre o espaço geográfico, aproveitando a forma hierárquica com que este espaço é organizado: O planeta é dividido em continentes, estes continentes são divididos em territórios e países que por sua vez adotam alguma forma de organização interna, também hierárquica.

A ideia principal é treinar um classificador de aprendizagem de máquina para cada nó da hierarquia geográfica, onde cada classe corresponda a um nó filho. Um classificador capaz de atribuir corretamente um texto a uma destas classes também determinará um escopo geográfico para o texto, pois cada classe corresponde diretamente a uma região no mapa.

Através de classificações sucessivas é possível percorrer a hierarquia até que se atinja uma das folhas, atribuindo ao documento o escopo geográfico mais específico correspondente. A escolha da raiz, do número de níveis da hierarquia a ser utilizada pelo Geofier dependerá exclusivamente dos objetivos específicos de implementação e da abrangência geográfica dos exemplos de treinamento disponíveis.

A Figura 9 exemplifica um possível “percurso” de um documento pela hierarquia até a atribuição do escopo geográfico mais específico. Neste exemplo, o nó “Brasil” da hierarquia geográfica foi escolhido como raiz, esta escolha simplifica a implementação do Geofier para o caso em que todos os documentos disponíveis pertencem à hierarquia geográfica brasileira.

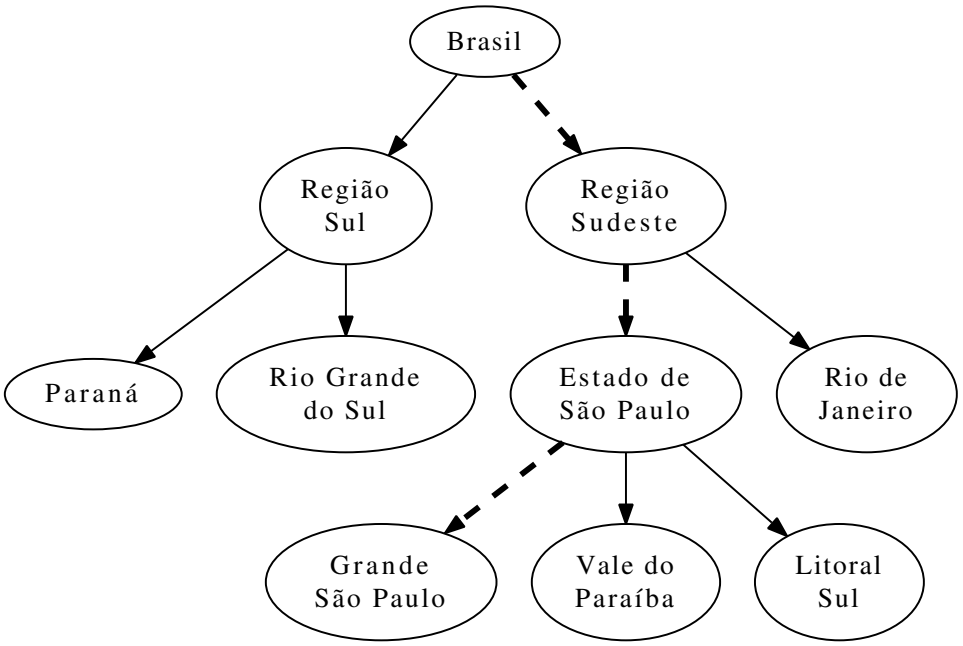
Para o treinamento desta hierarquia de classificadores é necessário um conjunto de documentos anotados com rótulos representando cada nível da hierarquia geográfica.

Foram realizados experimentos para avaliar a viabilidade da caracterização dos documentos com vetores TF-IDF com o intuito de usar este modelo para a identificação do escopo geográfico dos documentos, sem o uso de um gazetteer para consultas a nomes dos locais mencionados nos textos.

4.1 Classificação para duas regiões

Este experimento inicial serviu para avaliar a possibilidade de uso de classificadores de textos na identificação de um escopo geográfico. Os classificadores foram treinados sobre um vocabulário obtido a partir dos próprios exemplos, sem o uso de um gazetteer.

Figura 9 – Hierarquia de classificadores geográficos. Cada classificador corresponde a um nó, com classes que representam seus nós filhos



Fonte: preparado pelo autor

Documentos georreferenciados da Wikipédia em português cujas coordenadas estavam contidas dentro dos perímetros das cidades de São Paulo e do Rio de Janeiro foram extraídos e anotados com os rótulos “Rio de Janeiro” e “São Paulo”. Um classificador Bayesiano e uma SVM foram treinados usando este conjunto documentos, que continha 322 documentos na classe “São Paulo” e 114 na classe “Rio de Janeiro”, totalizando 436 documentos exemplo.

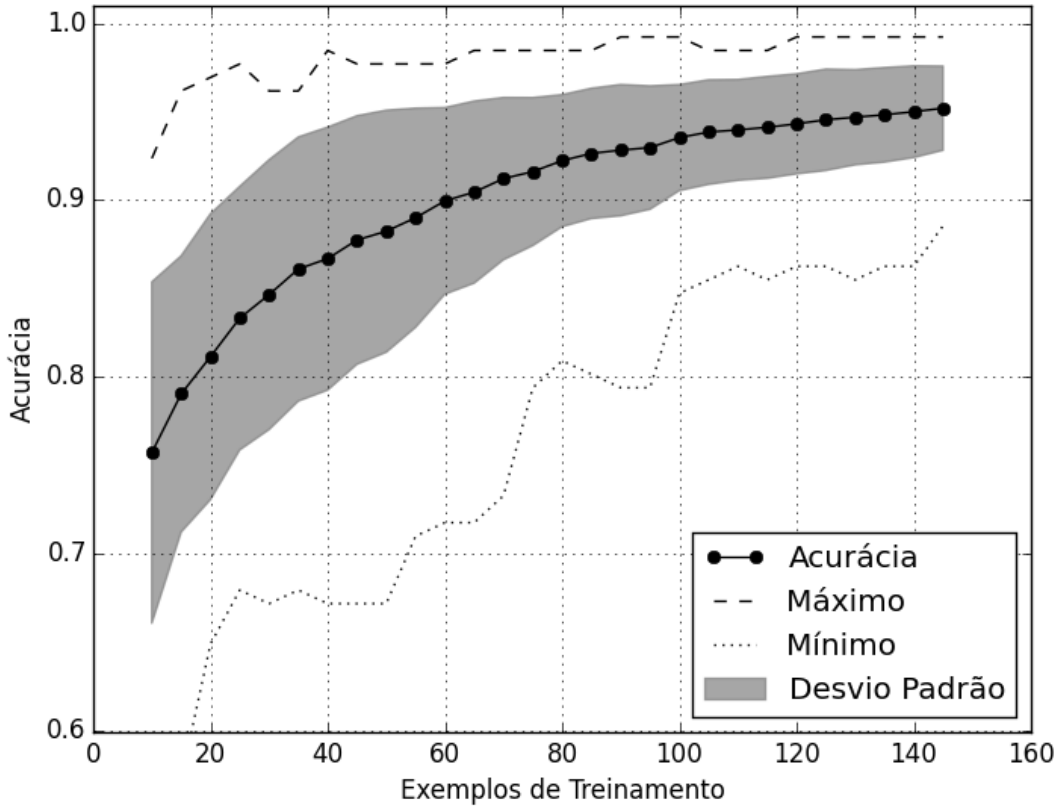
O vocabulário extraído destes documentos é composto de palavras isoladas (unigramas). As contagens de ocorrências dos termos deste vocabulário presentes em cada documento foram então utilizadas para criar uma representação vetorial de cada documento, adequada para o treinamento dos classificadores.

O experimento consistiu em realizar múltiplos testes de validação cruzada, variando o tamanho do conjunto de treinamento a cada teste. Os documentos restantes foram utilizados para validação e medição da acurácia do classificador.

Foram feitas 100 execuções para cada tamanho de conjunto de treinamento. Para cada execução de treinamento e validação foram registrados o maior e o menor valor da acurácia, além do cálculo da média e desvio padrão da acurácia, apresentados nas curvas de aprendizagem (*learning curves*) das figuras 10 e 11.

O classificador SVM apresentou acurácia superior ao Naive Bayes para a tarefa da classificação entre duas cidades, melhor média e menor desvio padrão para a acurácia.

Figura 10 – Acurácia de Classificação Naive Bayes de artigos da Wikipédia em português para as cidades de São Paulo e do Rio de Janeiro, variando-se o número de documentos para treinamento



Fonte: elaborado pelo autor

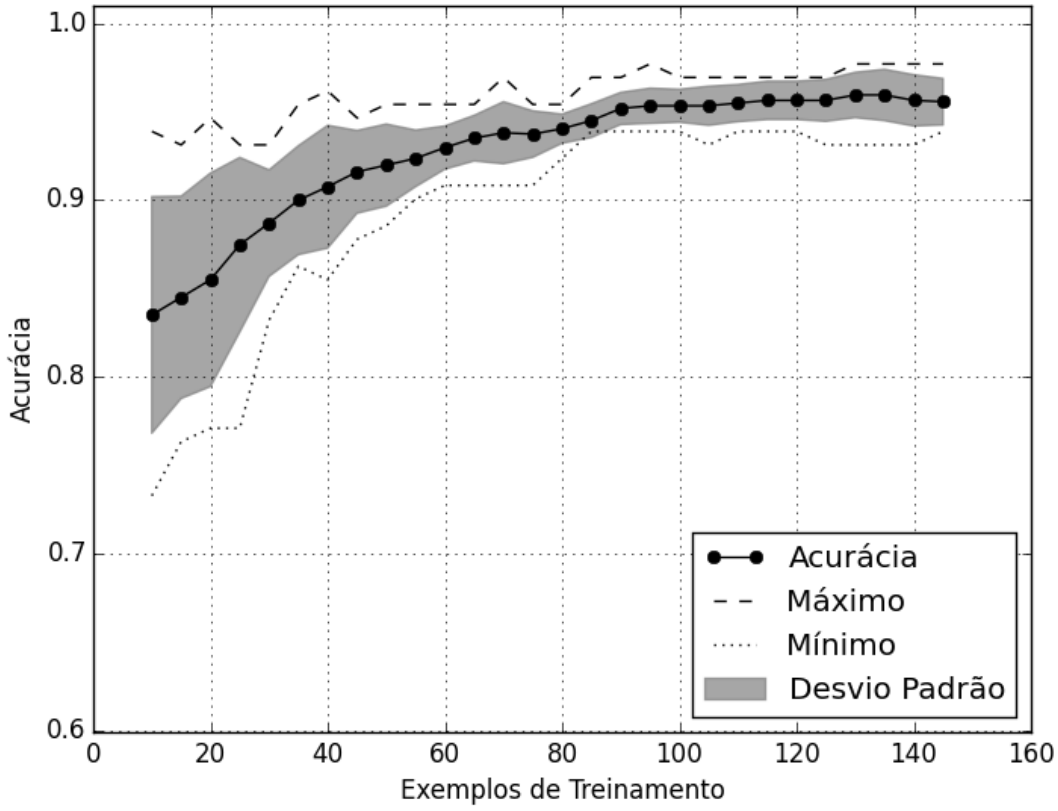
A execução deste experimento indica que é possível utilizar um classificador de aprendizagem de máquina de textos para atribuir regiões geográficas a documentos, pois foi possível atingir taxas acima de 90% de acerto de identificação do escopo geográfico correto para os documentos, com um número razoavelmente pequeno de exemplos de treinamento.

Observa-se que os classificadores começam a apresentar acurácia e margens de erro aceitáveis a partir de 60 exemplos para treinamento (aproximadamente 30 exemplos por classe). Esta observação é importante pois permite estabelecer um número mínimo de documentos que cada classe deve ter, para que se possam realizar os experimentos.

4.2 Método hierárquico para identificação de escopo geográfico de textos

O método Geofier consiste no treinamento de uma hierarquia de classificadores similar em estrutura à hierarquia da Região Geográfica em que se aplicará o método. Partindo

Figura 11 – Acurácia de Classificação com SVM de artigos da Wikipédia em português para as cidades de São Paulo e do Rio de Janeiro, variando-se o número de documentos para treinamento



Fonte: elaborado pelo autor

da raiz da hierarquia geográfica, selecionam-se os documentos do corpus relevantes para aquela região, e treina-se um classificador em que cada classe corresponderá a um dos nós filhos na hierarquia. Repete-se o treinamento para cada nó filho até que se chegue às folhas da hierarquia geográfica, conforme o algoritmo descrito na figura 12.

A função `ObtémExemplos` é responsável pela seleção dos documentos do corpus pertencentes à região de treinamento corrente. Sua implementação é trivial, pois todos os documentos do corpus de treinamento são anotados com os elementos da hierarquia geográfica, como descrito na seção anterior.

Para a classificação de um documento a hierarquia de classificadores deverá ser percorrida a partir de seu nó raiz. A cada visita a um nó, o nome da região geográfica em que o documento está sendo avaliado é acrescentado ao “caminho” percorrido na hierarquia. O classificador associado ao nó determinará a qual dos nós filhos o documento deve ser encaminhado. Repete-se o processo até que um nó folha seja alcançado, finalizando a classificação do documento, conforme algoritmo da figura 13.

Figura 12 – Algoritmo de Treinamento da Hierarquia Geofier

```
1  Função TreinaHierarquia(corpus, região)
2      hierarquia ← novo nó
3      hierarquia.nome = região
4      subordinados ← ListaFilhos (região)
5      se número de subordinados = 0 então
6          retorna null
7      senão
8          exemplos = ObtémExemplos(corpus,região)
9          hierarquia.classificador = TreinaClassificador(exemplos)
10         para i de 1 a número de subordinados faça
11             hierarquia.filho[i] = TreinaHierarquia(corpus, subordinados[i])
12     retorna hierarquia
```

Fonte: elaborado pelo autor

Figura 13 – Algoritmo de Classificação de Documentos do Geofier

```
1  Função ClassificaDocumento(hierarquia, documento)
2      n ← hierarquia ;
3      classificador ← n.classificador ;
4      caminho[0] = nó.nome ;
5      i ← 1 ;
6      enquanto classificador ≠ null faça
7          região = Classifica (classificador, documento) ;
8          nó ← nó.filho[região] ;
9          caminho[i] = nó.nome ;
10         classificador = nó.classificador ;
11         i = i + 1;
12     retorna caminho ;
```

Fonte: elaborado pelo autor

4.3 Escolha dos classificadores do Geofier

Os dois elementos principais do projeto do Geofier são a Hierarquia Geográfica que guia sua construção e a escolha dos classificadores que serão treinados.

Uma característica importante a ser observada para o treinamento do Geofier é o número de filhos nos nós da hierarquia geográfica. Cada nó possui um número de filhos n tal que $n_{min} \leq n \leq n_{max}$ em que n_{min} e n_{max} correspondem ao menor e maior número de filhos que um nó da hierarquia possui, respectivamente.

O valor de n_{max} é importante pois determinará o número máximo de classes com o qual um classificador da hierarquia será treinado. Na hierarquia brasileira, $n_{max} = 216$, pois a mesorregião “Noroeste Rio Grandense”, no Rio Grande do Sul, é composta por 216

microrregiões. Na hierarquia utilizada para os Estados Unidos, $n_{max} = 254$, pois o Estado do Texas é composto por 254 condados.

Embora n_{min} e n_{max} não imponham restrições ao método Geofier, é interessante observá-los para fins de implementação, em particular para a escolha dos classificadores que serão terinados.

Alguns classificadores, como o SVM, são essencialmente binários, isto é, classificam os elementos entre duas categorias. O classificador SVM é capaz de distinguir que elementos estão de qual dos dois "lados" de um hiperplano. Para que seja possível utilizar estes classificadores em problemas com múltiplas classes, múltiplos treinamentos serão efetuados. Há duas estratégias comuns para os treinamentos (MANNING; RAGHAVAN; SCHUTZE, 2008, p. 306 e 330).

A primeira, chamada "uma classe versus todas" (*One-versus-all - OVA Classification*) consiste em realizar n treinamentos, nos quais os exemplos de cada classe são separados dos demais para o treinamento de n classificadores binários. Um documento deve ser avaliado pelos n classificadores e a classe mais adequada deve ser escolhida. No caso específico das SVMs, a classe escolhida será aquela que apresentar a maior margem em relação ao documento classificado.

A segunda estratégia é treinar um conjunto (*ensemble*) de classificadores com cada par de classes. Segundo esta estratégia, durante a classificação de um novo documento a classe que obtiver o maior número de "votos" do conjunto deverá ser a escolhida. Embora esta estratégia exija $\frac{n(n-1)}{2}$ treinamentos, ela pode ser vantajosa, pois o número de exemplos a ser considerado para o treinamento de cada classificador do conjunto será menor.

O valor n_{max} (o maior número de classes presentes em um determinado nó da hierarquia geográfica selecionada) pode ter um grande impacto no custo do treinamento da hierarquia de classificadores do Geofier. O valor de n_{max} pode ser reduzido acrescentando níveis intermediários na hierarquia geográfica, se necessário.

Classificadores Naive Bayes avaliam a classe que provê a maior probabilidade de classificação dadas as evidências observadas. Sua formulação é capaz de tratar um problema com múltiplas classes naturalmente e são uma escolha popular para problemas de classificação de textos com várias classes mutuamente exclusivas.

Os classificadores Naive Bayes foram escolhidos pelo fato de não exigirem muita memória para armazenamento dos dados mesmo no treinamento de milhares de documentos, com centenas de classes e tamanhos de vocabulário superiores a 20.000 termos, nos testes realizados para esta pesquisa.

5 AVALIAÇÃO DO GEOFIER

O Geofier foi comparado com uma implementação do sistema Web-a-Where. O Web-a-Where (AMITAY et al., 2004) foi selecionado como base de comparação de desempenho por ser uma referência entre as diferentes técnicas de atribuição de escopo geográfico para textos retirados da web. Também apresenta acurácia satisfatória, segundo seus autores, a partir de um algoritmo razoavelmente simples de ser replicado.

Outra vantagem do Web-a-Where é utilizar heurísticas comuns a outros métodos de anotação geográfica. Implementando o Web-a-Where é possível analisar de forma comparativa a proposta da hierarquia de classificadores do Geofier com heurísticas comuns implementadas por outros métodos da literatura.

5.1 Extração de Documentos Georreferenciados da Wikipédia

A base de artigos da Wikipédia em inglês foi utilizada como fonte para um conjunto de documentos georreferenciados, a ser utilizada pelo Web-a-Where e pelo Geofier, na tarefa de identificação do escopo geográfico dos textos. A extração destes documentos é detalhada a seguir.

A Wikipédia é uma enciclopédia colaborativa de conteúdo aberto e, como tal, disponibiliza a íntegra de seu conteúdo para download, em diversos idiomas (WIKIPEDIA, 2014b). Diferentes arquivos de extração do conteúdo são disponibilizados em formato XML. Neste projeto de pesquisa foi utilizada a versão de 12 de janeiro de 2015 da Wikipédia em português e em inglês.

O arquivo de exportação XML da Wikipédia é composto de um preâmbulo dado pelos nós XML <siteinfo>. . . </siteinfo> e em seguida por uma sequência de nós <page>. . . </page>. O conteúdo de cada página e da metainformação, incluindo data de última modificação, usuário, texto completo e título estão organizados em nós do tipo <page>. Dos metadados, apenas o título de cada página será armazenado para eventual consulta.

Tabela 1 – Nós XML de exportação da Wikipédia utilizados neste trabalho

Nó XML	Descrição
<page>...</page>	Delimita o conteúdo de uma página
<title>...</title>	Título da página
<text>...</text>	Conteúdo em Linguagem de Marcação wikitext

Fonte: elaborado pelo autor

A URL de um verbete da Wikipédia é obtida adicionando-se à URL base o título do verbete, com espaços substituídos por sublinhados. Para a Wikipédia em inglês, a URL base é <http://en.wikipedia.org/wiki>.

A Wikipédia é composta por documentos escritos com uma linguagem de marcação própria chamada *wikitext*. Há algumas predefinições da linguagem para representar coordenadas geográficas no texto. Sempre que um verbete diz respeito a um local estas marcações são traduzidas para o usuário final como links para serviços de mapas online, através do Projeto Geohack (WIKIPEDIA, 2012).

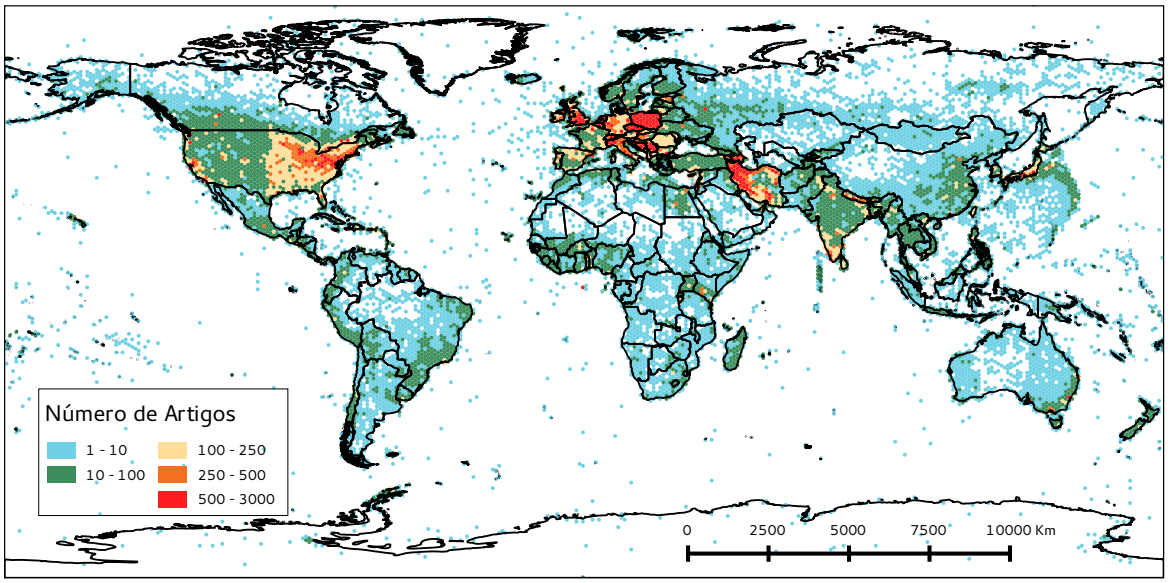
Para a construção de um dataset georreferenciado, todos os documentos das Wikipédias em português e inglês com marcações do tipo “geocoordenadas” (WIKIPEDIA, 2005b), “satélite” (WIKIPEDIA, 2005d), “coord” (WIKIPEDIA, 2005a), “Info/Município do Brasil” (WIKIPEDIA, 2005c) e “Infobox” (WIKIPEDIA, 2005e; WIKIPEDIA, 2005f) foram identificados através do uso de expressões regulares e tiveram suas respectivas coordenadas geográficas extraídas e armazenadas como metadados do documento. Especial atenção foi dada na exclusão de algumas marcações do tipo “Infobox” que diziam respeito a crateras em corpos celestes.

O conteúdo do verbete e sua metainformação como título, URL e coordenada de cada documento identificado, foi inserido em uma tabela de um banco de dados geográficos PostGIS (POSTGIS, 2012), resultando em uma base georreferenciada de documentos com 539.743 itens. A figura 14 exibe a distribuição geográfica destes documentos. Cada verbete georreferenciado da Wikipédia está representado com um ponto. Nota-se a distribuição irregular dos textos pelo espaço geográfico, concentrados em algumas regiões.

O IBGE (IBGE, 2010b) disponibiliza polígonos em alta resolução dos contornos de todas as cidades do Brasil. Para os Estados Unidos, dados cartográficos são disponibilizados pelo *United States Bureau of the Census* (CENSUS, 2014). Uma compilação das diferentes regiões administrativas de diversos países do mundo, curadas a partir de diversas fontes oficiais, é disponibilizada pelo projeto GADM (BERKELEY; INSTITUTE, 2012).

Estes polígonos, semelhantes aos descritos na página 15, foram armazenados em uma tabela do banco de dados geográfico Postgis. Desta forma é possível utilizá-los para realizar consultas dos documentos que pertencem a regiões administrativas arbitrárias. A Figura 15 mostra um exemplo de consulta que seleciona todos os documentos georreferenciados da Wikipédia cujas coordenadas estejam localizados dentro do perímetro do condado de San Francisco. Neste exemplo a tabela `USA_adm2` armazena os polígonos de todos os condados americanos na coluna `the_geom` e a tabela `enwikipedia` armazena todos os verbetes georreferenciados extraídos da Wikipédia, com coordenadas na coluna `xy`. A função booleana `ST_Contains()` retorna verdadeiro sempre que a geometria de seu segundo argumento está inteiramente contida na do primeiro.

Figura 14 – Artigos georreferenciados da Wikipédia em inglês. Nota-se a concentração dos textos em determinadas regiões



Fonte: elaborado pelo autor

As diferenças de organização da hierarquia geográfica entre países são frequentemente abstraídas através da utilização de nomes genéricos (nível 1, nível 2, etc...) e um dicionário que associe cada nível da hierarquia à sua interpretação local.

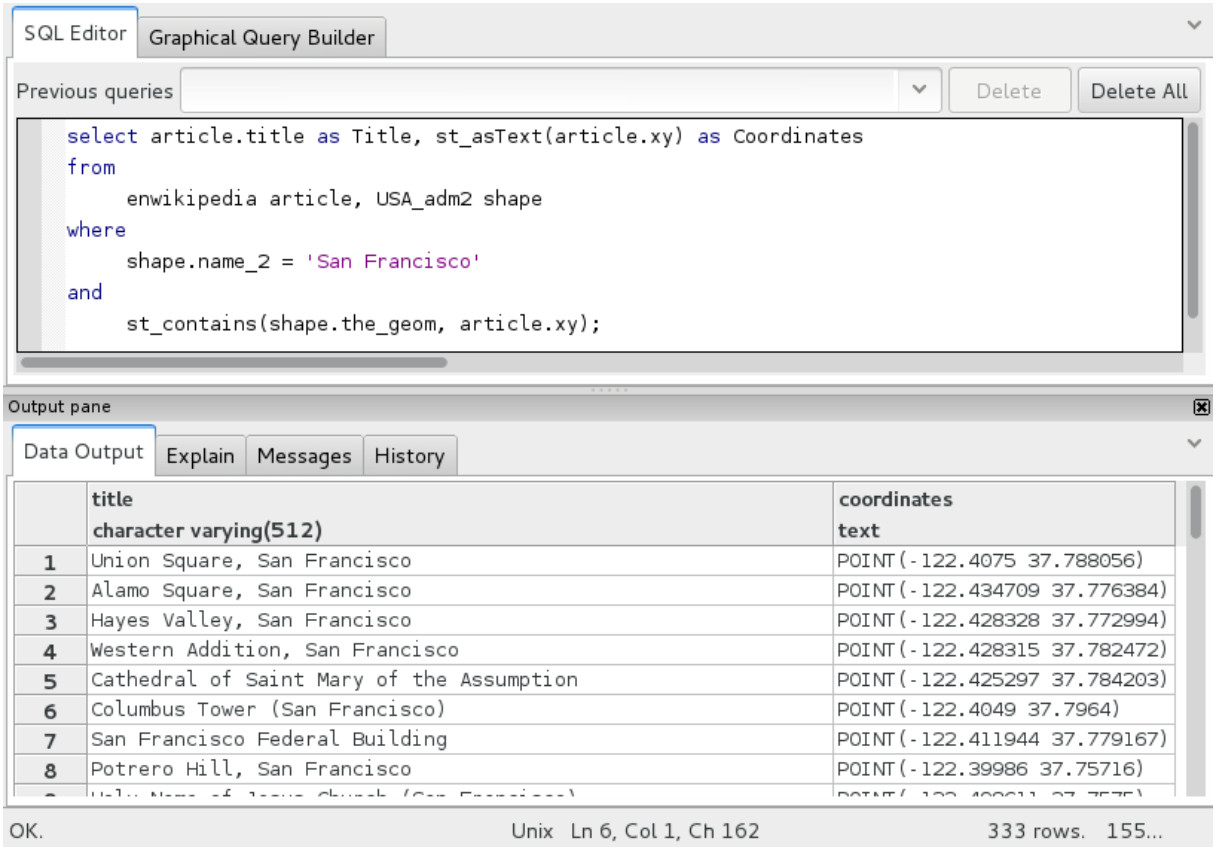
Para o Brasil, contornos das cidades brasileiras foram utilizados para selecionar os documentos georreferenciados pertencentes a cada município. Cada documento foi anotado com rótulos indicando a hierarquia geográfica completa à qual está subordinado. A hierarquia geográfica utilizada para experimentos com documentos brasileiros é composta por cinco níveis: Região, Estado, Mesorregião, Microrregião e Município (IBGE, 1993).

Ao contrário do Brasil em que a totalidade do território nacional está subordinada aos Municípios, nos Estados Unidos há áreas que não estão subordinadas à jurisdição de nenhuma cidade. A menor divisão administrativa que abrange todo o território dos Estados Unidos são os “Condados”. A hierarquia adotada para anotação geográfica de documentos pertencentes ao território americano neste trabalho é composta de quatro níveis: Região, Divisão, Estado e Condado.(CENSUS, 1994).

Estas regiões possuem entre si relacionamentos do tipo “contém” e “está contida em”. A hierarquia geográfica utilizada para o território dos Estados Unidos é apresentada pela Figura 16, com exceção dos condados pertencentes a cada estado.

Não foi possível encontrar uma ferramenta de tradução dos fontes dos artigos em wikitext para texto puro que satisfizesse as necessidades desse trabalho. Para resolver este problema foi necessária a construção de um espelho da Wikipédia em um servidor local,

Figura 15 – Consulta ao banco de dados geográfico usando um polígono. Documentos de “San Francisco”



Fonte: elaborado pelo autor

para a exibição do conteúdo dos verbetes em HTML sob demanda.

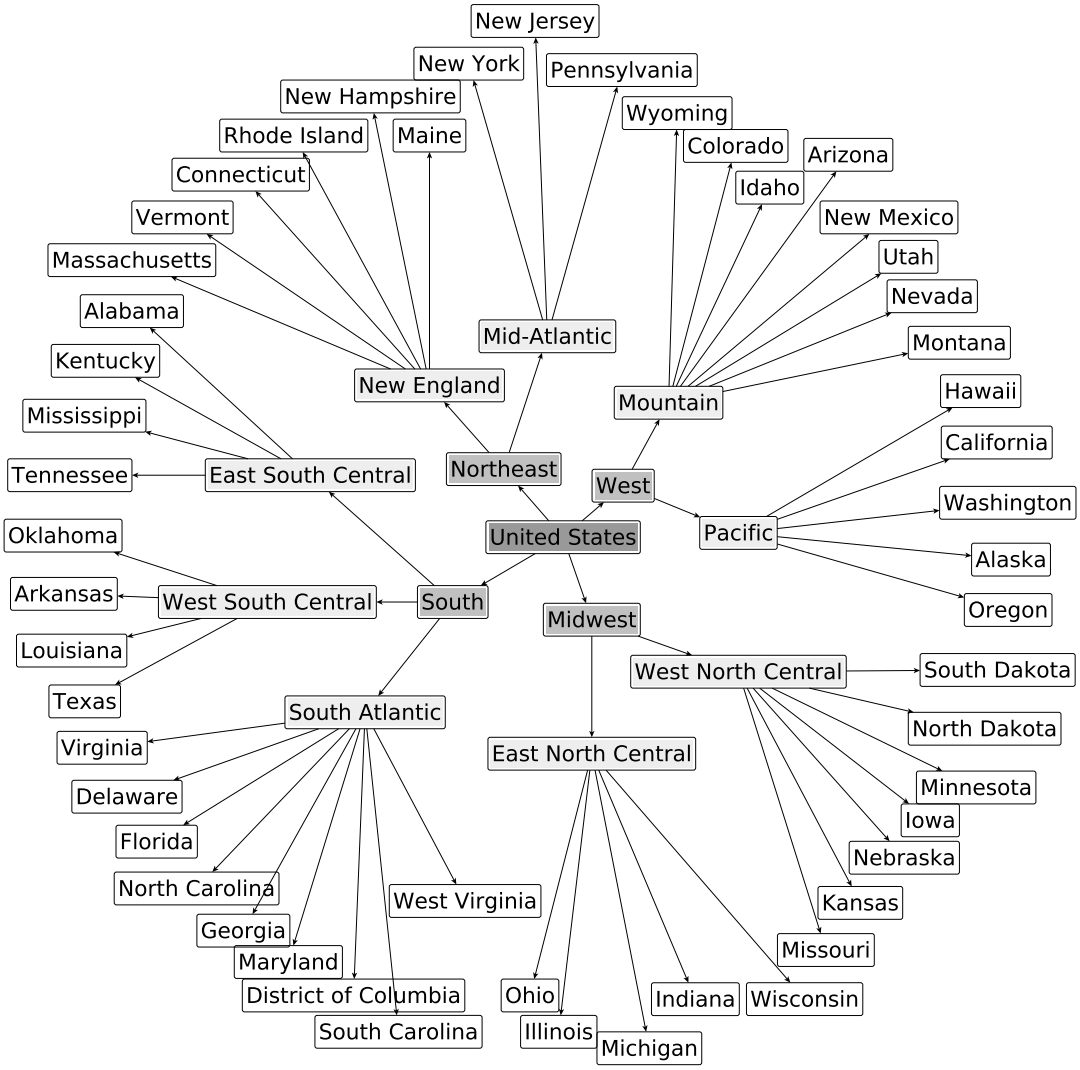
A representação HTML dos verbetes georreferenciados foi obtida através de requisições web para este espelho e convertida para texto puro, totalizando 7.863 documentos dentro do território brasileiro e 117.311 documentos georreferenciados e anotados com seus respectivos rótulos geográficos, dentro do território Americano.

5.2 Implementação do Web-a-Where

O Web-a-Where utilizava dados do site World-Gazetteer.com (HELDERS, 2005) para a construção de seu gazetteer. O site deixou de existir, porém foi possível recuperar o arquivo publicado do gazetteer através do Internet Archive (ARCHIVE, 1996). Os nomes de locais e dados populacionais utilizados neste trabalho para a implementação do Web-a-Where são provenientes do World-Gazetteer.com, para manter uma maior fidelidade de implementação em relação àquela utilizada pelo autor.

Para este experimento, limitamos o gazetteer e os documentos a serem analisados àqueles presentes no território americano, extraídos dos bancos de dados da Wikipédia (WI-

Figura 16 – Hierarquia geográfica dos Estados Unidos



Fonte: Preparado pelo autor

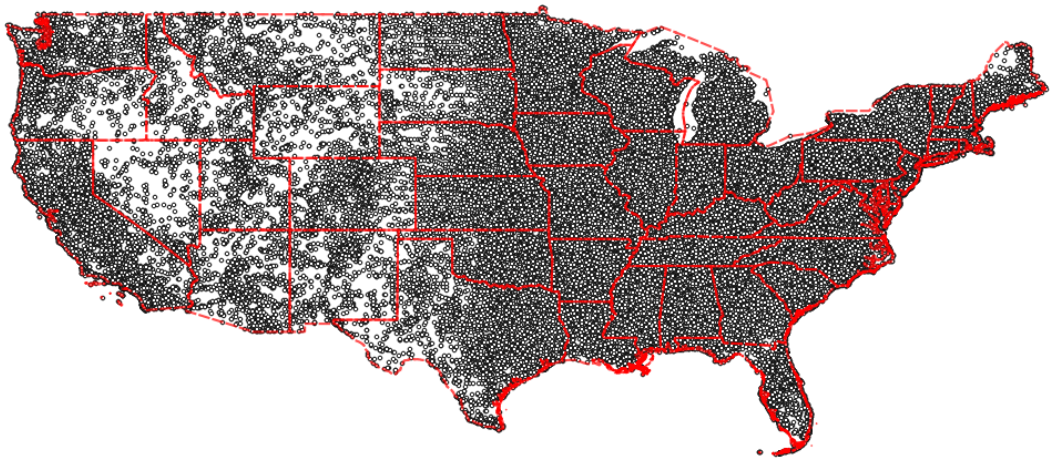
KIPEDIA, 2014b), conforme as figuras 17 e 18.

O Web-a-Where utiliza uma Taxonomia na forma *Continente/País/Estado/Local Povoado*, porém, os locais povoados (*populated places*) norte-americanos não cobrem toda a extensão territorial dos Estados Unidos, assim, nem todos os documentos georreferenciados da base de treinamento estão contidos em locais povoados.

Os condados e estados americanos, por sua vez, cobrem toda a extensão territorial americana (CENSUS, 1994, c.4, p.18), Para efeito de comparação com a técnica proposta neste trabalho, os documentos identificados pelo Web-a-Where também são anotados com o condado a que pertence o local atribuído ao documento pelo algoritmo.

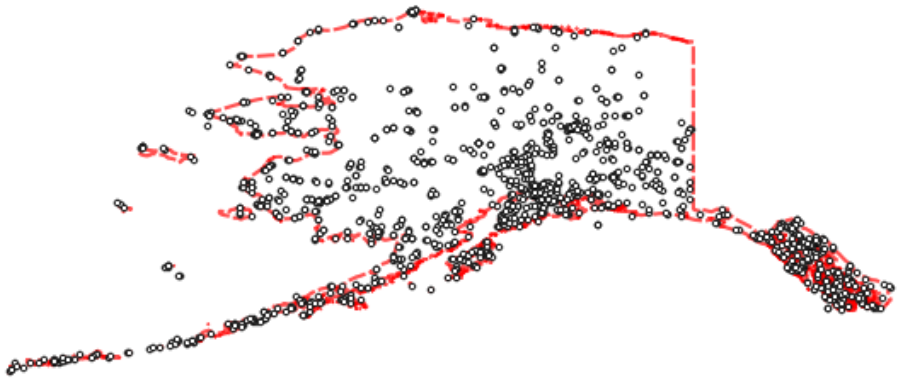
Para cada nome de local presente no gazetteer foi feita uma contagem, em todos os 15.205.271 artigos da Wikipédia em inglês. Esta contagem registrou o número de ocorrências de cada nome de local, tanto em minúsculas como com iniciais maiúsculas,

Figura 17 – Artigos georreferenciados da Wikipédia em inglês, EUA Continental



Fonte: elaborado pelo autor

Figura 18 – Artigos georreferenciados da Wikipédia em inglês, Alaska



Fonte: elaborado pelo autor

para determinar quais termos do gazetteer possuem caráter primariamente não geográfico e deveriam ter tratamento especial. A tabela 2 exibe os dez termos mais frequentes resultantes desta contagem.

Pelo método proposto pelos autores do Web-a-Where observa-se que os nomes das cidades de “Page”, Arizona e “Center”, Texas são ambíguos, ocorrendo com muito mais

Tabela 2 – Contagem de Ocorrências de nomes do gazetteer Web-a-Where

Nome	Ocorrências
page	43215790
center	10268952
August	4234890
University	4046191
York	3121374
New York	2930857
London	1773890
College	1589543
Washington	1125069
Union	992671

Fonte: elaborado pelo autor

frequência com significado não geográfico em textos da web. A tabela 3 mostra os primeiros 10 nomes da lista de locais do gazetteer que foram mais frequentemente encontrados escritos com iniciais minúsculas nos textos da Wikipédia.

Tabela 3 – Contagem de Ocorrências de nomes do gazetteer, com iniciais minúsculas e maiúsculas, em uma coleção de páginas da Wikipédia.

Nome	Ocorrências em Minúsculas	Ocorrências em Maiúsculas
Page	43215790	398440
Center	10268952	959318
Media	941121	593167
Reading	348446	177279
Gray	251358	224803
Mission	223142	172438
Hope	172665	141253
Economy	152933	116083
Price	139444	129797
Mobile	139022	133295

Fonte: elaborado pelo autor

A identificação destes nomes replica uma das etapas da construção da parte não geográfica do gazetteer do Web-a-Where, conforme descrito por seus autores: “Nomes de locais que apareceram mais de 100 vezes, mas que em sua maioria não estavam em maiúsculas como um nome próprio deveria, foram incluídos na seção não geográfica do gazetteer” (AMITAY et al., 2004, p.276)

Neste exemplo, as cidades de Page (Arizona), Center (Texas), Media (Pennsylvania), Reading (Pennsylvania, Massachusetts e Ohio), Gray (Maine e Louisiana), Mission (Texas e Kansas), Hope (Arkansas), Economy (Pennsylvania), Price (Utah) e Mobile (Alabama), entre outras, foram identificadas como tendo nomes com significado não geográfico por esta heurística. O segundo passo na identificação dos nomes da seção não geográfica

Tabela 4 – Contagem de Ocorrências de nomes do gazetteer em uma coleção de páginas da Wikipédia versus população associada ao nome.

Nome	Maior População	Ocorrências em Maiúsculas
August	8717	4234890
University	34953	4046191
York	39587	3121374
London	9279	1773890
College	11413	1589543
Union	56771	992671
Republic	10423	960862
Green	23894	754109
Martin	10209	699983
Mexico	10788	664841

Fonte: elaborado pelo autor

do gazetteer do Web-a-Where é identificar nomes que frequentemente aparecem em maiúsculas, mas que possuem significado eminentemente não geográfico. O Web-a-Where compara a população associada a um nome com o número de ocorrências do nome em coleção de textos de referência: “Nomes mencionados com muito maior frequência do que sua população garantiria também foram incluídos”. (AMITAY et al., 2004, p.276). Os autores não dão detalhes do que significa “muito maior frequência”, ou de como se daria esta garantia. Para estes experimentos serão feitas comparações simples entre a contagem de ocorrências de um nome e a maior população atribuída a este nome no gazetteer.

A tabela 4 mostra os 10 primeiros nomes do gazetteer que mais ocorreram escritos com letras iniciais maiúsculas na coleção de documentos e a população associada ao nome. Estes nomes, entre outros, foram adicionados à seção não geográfica do gazetteer por possuírem população menor do que o número de menções nos textos analisados. As cidades de August (California), University (Florida), York (Pennsylvania, Maine, Nebraska e South Carolina), London (Ohio e Kentucky), College (Alaska), Union (New Jersey, Missouri, South Carolina e Ohio), Republic (Missouri), Green (Ohio), Martin (Tenessee) e Mexico (Missouri), entre outras, foram adicionadas à seção não geográfica do gazetteer por obedecerem a esta regra.

O gazetteer utilizado na implementação deste trabalho contém apenas nomes de locais dos Estados Unidos, pois ele foi avaliado com documentos que descrevem locais e eventos deste país, por isso alguns nomes reconhecidamente geográficos como London e Mexico falham nesta regra. No gazetteer utilizado, o país México possui 103.872.328 habitantes e Londres 7.421.209. Se um gazetteer global fosse utilizado, estes termos não teriam entrado na seção não-geográfica, pois suas populações seriam maiores do que o número de menções na coleção de textos analisada.

O gazetteer resultante contém os nomes, populações e coordenadas de 51 estados (incluindo o distrito federal) e 5.095 nomes de cidades em sua seção "geográfica". A seção

"não geográfica"do gazetteer inclui 1.131 nomes de locais que foram identificados como tendo interpretação não geográfica na maioria dos casos.

Uma terceira seção foi criada para o gazetteer, contendo sequências que o Web-a-Where chama de "Qualificadas Completamente"(*Fully Qualified*). Estas sequências, dentro do escopo deste trabalho, referem-se a cidades acompanhadas de seus estados, como “New York, NY” , “Los Angeles, Cal.” ou “Houston, Texas”. Foram geradas 18.317 entradas com combinações de grafias de nomes de cidade e estados, com as variações de nomes de estados obtidas a partir de informações do World-Gazetteer.com.

Esta simplificação é possível porque o experimento foi realizado apenas dentro dos Estados Unidos. O único topônimo que pode qualificar outro nesta situação é uma menção a um estado. Numa implementação com um gazetteer e documentos globais, é necessário analisar outros tipos de topônimos qualificadores, por exemplo: “Cambridge”, seguido de “United States”.

A execução do Web-a-Where na totalidade dos 117.311 documentos exemplo resultou na atribuição de escopo geográfico a 116.917 documentos, conforme a tabela 5.

Tabela 5 – Quantidade de Atribuições de tags do Web-a-Where por nó da Taxonomia, Número de acertos por nó e global

Nível	Documentos Atribuídos	Acerto dentro do nível	Acerto global
País	45.271 (38.59%)	100%	99.66%
Estado	65.533 (55.86%)	96.8%	61.07%
Condado	6.113 (5.21%)	19.74%	1%
Nenhum	394 (0.34%)	-	-

Fonte: elaborado pelo autor

Apenas 5.21% do total de documentos avaliados pelo Web-a-Where foram atribuídos a uma localidade específica, dentre os quais 19% foram atribuídos corretamente. Do total, apenas 1% dos documentos foram atribuídos corretamente a um condado e 61.07% a um estado. Houve 0.34% de documentos que não traziam evidências suficientes para que o Web-a-Where pudesse atribuir qualquer escopo geográfico.

A média de Precisão para os documentos que foram atribuídos a algum estado pelo Web-a-Where é de 0.97 e sua taxa de revocação média é de 0.59, conforme a tabela 6. Isso significa que para um determinado estado americano 98% dos documentos que foram atribuídos àquele estado foram atribuídos corretamente (alta precisão), porém apenas 59% de todos os documentos que deveriam ser atribuídos ao estado o foram (menor revocação).

Em 117.311 documentos analisados, pouco mais de 6.000 documentos tiveram o escopo geográfico identificado com uma localidade e apenas 1161 documentos tiveram o escopo geográfico corretamente identificado neste nível hierárquico. As taxas de precisão e revocação são portanto, muito próximas de 0, para a maioria dos mais de 3.000 condados

americanos considerados, este resultado sugere a inadequação do método do Web-a-Where para este nível de precisão de atribuição do escopo geográfico.

Tabela 6 – Web-a-Where: Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para Estados e Condados dos Estados Unidos

Nível Hierárquico	Precisão	Revocação	$F_{\beta=1}$
State	0.97	0.59	0.72
County	0.00	0.00	0.00

Fonte: elaborado pelo autor

5.3 Hierarquia de Classificadores

O experimento anterior mostrou ser possível atribuir uma classe de documentos a uma coleção de textos georreferenciados, cujas coordenadas estivessem contidas em uma região geográfica. O classificador de textos foi capaz então de distinguir e separar estes documentos, atribuindo não apenas uma classe, mas uma área definida de território, em essência atribuindo um “escopo geográfico” ao documento.

É possível treinar um único classificador de textos para atribuir a um texto uma classe entre todas as cidades do território brasileiro, por exemplo. Tal classificador teria 5565 classes, uma para cada município brasileiro. Nos Estados Unidos, seriam 1837 condados.

Uma inspeção dos exemplos disponíveis para o treinamento mostra que a distribuição de exemplos no território é irregular. As tabelas 7 e 8 mostram as quinze cidades do Brasil e condados dos Estados Unidos com mais documentos no conjunto de verbetes georreferenciados da Wikipédia em português e inglês, respectivamente.

Há muitas cidades ou condados que não possuem exemplos suficientes para o treinamento e validação de um classificador, mas estas situações não ocorrem nos nós superiores da hierarquia geográfica, que agregam várias destas entidades, como estados, regiões de um país, ou suas regiões administrativas equivalentes. As tabelas 9 e 10 apresentam a contagem de documentos para os estados americanos e brasileiros presentes na Wikipédia em inglês e português, respectivamente.

O número de exemplos disponíveis para o estado de Roraima sugere que não será possível treinar um classificador capaz de distinguir documentos pertencentes a cada uma de suas 15 cidades, devido ao pequeno número de exemplos por classe e aos parâmetros coletados pela observação realizada no experimento anterior. No entanto, Roraima poderia ser, seguramente segundo estes parâmetros, uma das classes de um classificador que classifique documentos referentes à região norte do país, por estado.

Tabela 7 – As 15 cidades brasileiras com maior número de documentos georreferenciados na Wikipédia em português.

Índice	Cidade	Número de Documentos
1	Campinas, SP	409
2	São Paulo, SP	322
3	Rio de Janeiro, RJ	114
4	Brasília, DF	66
5	Belo Horizonte, MG	64
6	Curitiba, PR	61
7	Florianópolis, SC	41
8	Porto Alegre, RS	41
9	Recife, PE	41
10	São Carlos, SP	40
11	Cuiabá, MT	38
12	Salvador, BA	28
13	Fortaleza, CE	27
14	Coronel Fabriciano, MG	23
15	Belém, PA	21

Fonte: elaborado pelo autor

Tabela 8 – Os 15 condados dos EUA com maior número de documentos georreferenciados na Wikipédia em Inglês.

Índice	Cidade	Número de Documentos
1	Los Angeles, California	1137
2	New York, New York	793
3	King, Washington	752
4	Cook, Illinois	732
5	San Diego, California	541
6	Allegheny, Pennsylvania	522
7	Baltimore, Maryland	507
8	Kern, California	403
9	Miami-Dade, Florida	350
10	Riverside, California	350
11	Suffolk, New York	337
12	San Bernardino, California	337
13	San Francisco, California	333
14	Alameda, California	333
15	Middlesex, Massachusetts	329

Fonte: elaborado pelo autor

Este experimento consiste no treinamento de uma hierarquia de classificadores, seguindo a hierarquia geográfica dos Estados Unidos.

Dos documentos disponíveis para treinamento, separam-se aleatoriamente 30% para compor o conjunto de validação, antes de treinar a hierarquia.

Cada documento está anotado com região, divisão, estado e condado de que fazem parte, sendo possível selecionar dinamicamente quais documentos farão parte do conjunto de treinamento no momento do treinamento de cada classificador.

Tabela 9 – Os Estados dos EUA com maior e menor número de documentos georreferenciados na Wikipédia em Inglês.

Índice	Estado	Número de Documentos
1	California	9308
2	New York	5569
3	Pennsylvania	5285
4	Ohio	4884
5	Texas	4847
47	Hawaii	588
48	Vermont	516
49	Delaware	509
50	District of Columbia	327
51	Rhode Island	283

Fonte: elaborado pelo autor

Tabela 10 – Os estados brasileiros com maior e menor número de documentos georreferenciados na Wikipédia em Português.

Índice	Estado	Número de Documentos
1	São Paulo	1662
2	Minas Gerais	1059
3	Rio Grandedo Sul	646
4	Paraná	530
5	Bahia	464
23	Distrito Federal	66
24	Rondônia	62
25	Amapá	34
26	Acre	28
27	Roraima	21

Fonte: elaborado pelo autor

O Vocabulário utilizado para a vetorização dos documentos é fixo e dado pelo mesmo gazetteer utilizado pelo Web-a-Where, sem distinção prévia de termos que tenham maior ou menor significado geográfico. Cada documento é inicialmente transcrito em uma nova representação, dada por um conjunto de rótulos que indicam a presença de um nome de cidade ou estado no texto.

O seguinte fragmento de texto foi retirado do Verbete “Ada, Ohio” (WIKIPEDIA, 2014a) da Wikipédia: “Ada was founded in 1853 by S. M. Johnson. The town was named after the postmaster’s daughter, Ada. Ada has been noted for having one of the shortest place names in Ohio”. Na transcrição deste fragmento, os Termos “Ada” e “Ohio” serão substituídos por rótulos, os demais termos serão ignorados.

Cada ocorrência da cadeia “Ohio” corresponde a uma ocorrência do rótulo “uss-tate:Ohio”, na nova representação. Cada ocorrência da cadeia “Ada” é substituída por dois rótulos: “uscity:Ada/Ohio” e “uscity:Ada/Oklahoma”, pois há duas interpretações geográficas possíveis para a cadeia “Ada” no gazetteer.

Além da interpretação Geográfica do nome “Ada”, o fragmento exibe também uma interpretação não geográfica, já que a segunda ocorrência do nome refere-se a uma pessoa. Ao contrário do Web-a-Where e outros algoritmos que se utilizam de heurísticas para se antecipar à ambiguidade dos tipos Geo/Geo e Geo/Non-Geo, o Geofier confia esta tarefa aos mecanismos de aprendizagem dos classificadores.

Tabela 11 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para diferentes níveis da hierarquia geográfica dos Estados Unidos usando Naive Bayes

Nível Hierárquico	Precisão	Revocação	$F_{\beta=1}$
Region	0.96	0.96	0.96
Division	0.94	0.94	0.94
State	0.92	0.92	0.92
County	0.76	0.73	0.74

Fonte: elaborado pelo autor

Tabela 12 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para diferentes níveis da hierarquia geográfica dos Estados Unidos usando SVM

Nível Hierárquico	Precisão	Revocação	$F_{\beta=1}$
Region	0.98	0.98	0.98
Division	0.97	0.97	0.97
State	0.95	0.95	0.95
County	0.79	0.77	0.77

Fonte: elaborado pelo autor

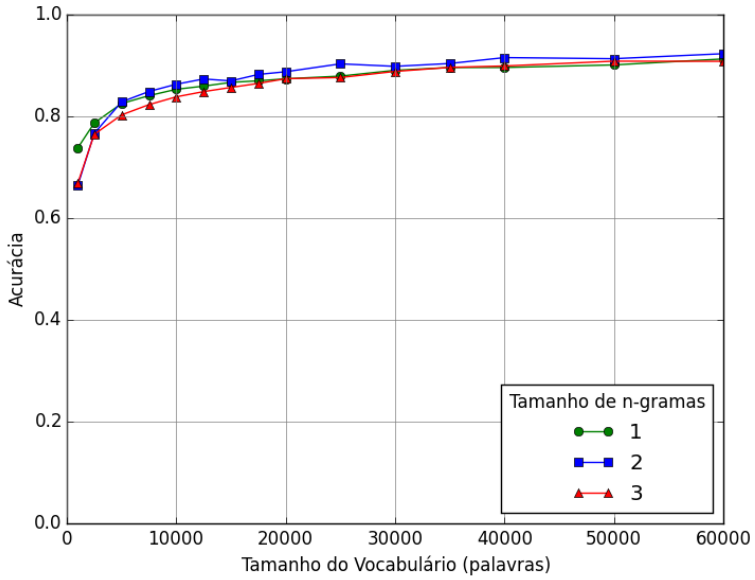
O uso de um vocabulário controlado, dado pelo mesmo gazetteer utilizado pelo Web-a-Where apresentou desempenho muito superior ao próprio Web-a-Where, tanto para classificadores Naive Bayes (tabela 11) quanto SVMs (tabela 12). Com as SVMs apresentando tanto melhor precisão quanto revocação em relação aos classificadores Naive Bayes, ainda que por uma pequena margem.

5.4 Hierarquia de Classificadores com N-Gramas

Um experimento inicial para avaliação dos efeitos do uso de diferentes tamanhos de vocabulário e combinações de n-gramas (descritos na seção 3.1) foi realizado para determinar a configuração ótima para avaliação de desempenho do Geofier. Os gráficos das figuras 19, 20 e 21 exibem a taxa de acertos de classificação da hierarquia de classificadores Naive Bayes, com tamanhos máximos de vocabulário diferentes e n-gramas de tamanho fixo 1, 2 e 3, separadamente.

O vocabulário é escolhido selecionando-se os n termos com maior pontuação tf-idf, o experimento de determinação dos parâmetros para o Geofier testou o desempenho de diferentes tamanhos de vocabulário entre 1.000 e 60.000 termos.

Figura 19 – Desempenho da hierarquia de classificadores com N-gramas, por região dos Estados Unidos



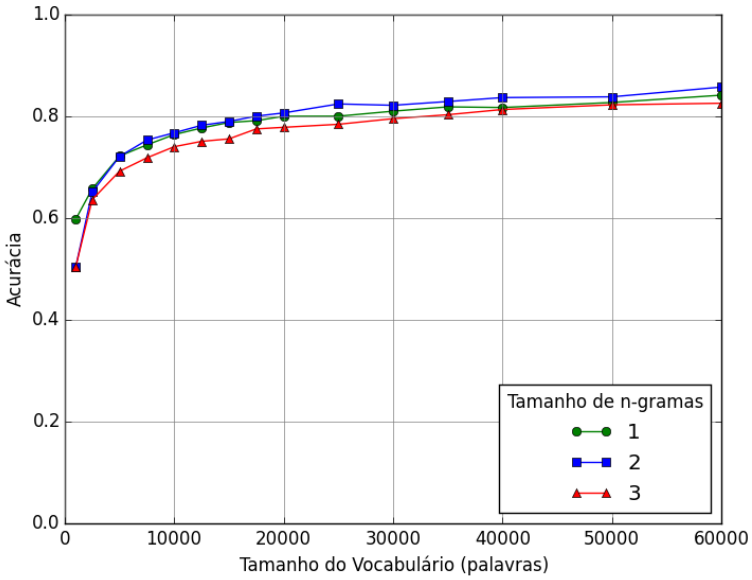
Fonte: elaborado pelo autor

Observa-se que para vocabulários de tamanho superior a 20.000 n-gramas os acréscimos em desempenho são cada vez menores, enquanto o custo de vetorização dos documentos aumenta para n-gramas de tamanho 2 ou 3. O vocabulário de n-gramas completo extraído dos documentos tem tamanhos de 708.952 unigramas, 12.191.567 termos para bigramas e 33.755.474 termos para trigramas, consequentemente o custo computacional para treinamento deverá ser levado em conta, comparado ao benefício proporcionado pela complexidade extra.

Verifica-se uma queda acentuada de desempenho para níveis mais específicos da hierarquia geográfica, observada já nos nível dos estados, conforme a figura 20. Esta queda se explica em parte pelo número menor de exemplos de treinamento e validação para regiões geográficas menores como os condados, exibidos na figura 21, mas também pela propagação dos erros de classificação cometidos nos níveis superiores da hierarquia durante o percurso. Este queda de desempenho é consistente com o observado em outros métodos na literatura e muito menos pronunciada do que a observada, por exemplo, na implementação do Web-a-Where utilizada neste projeto de pesquisa e apresentada na coluna “acerto global” da tabela 5.

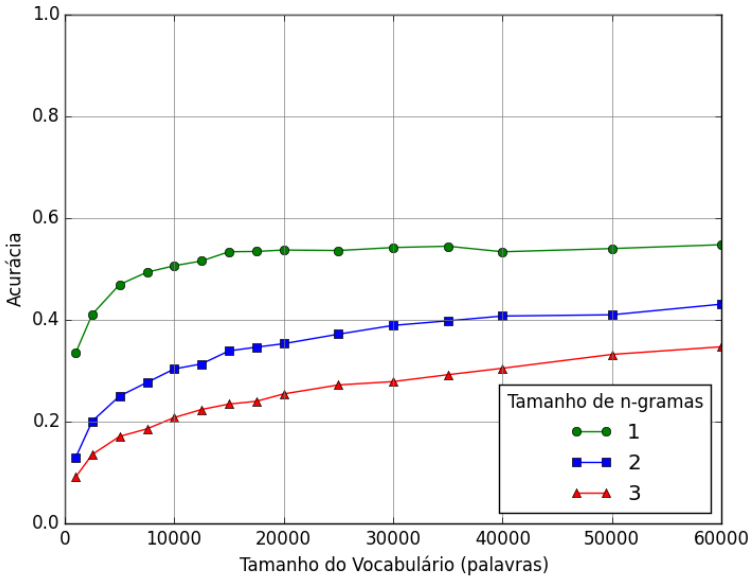
Para verificar qual o benefício de usar combinações de n-gramas de tamanhos

Figura 20 – Desempenho da hierarquia de classificadores com N-gramas, para estados dos Estados Unidos



Fonte: elaborado pelo autor

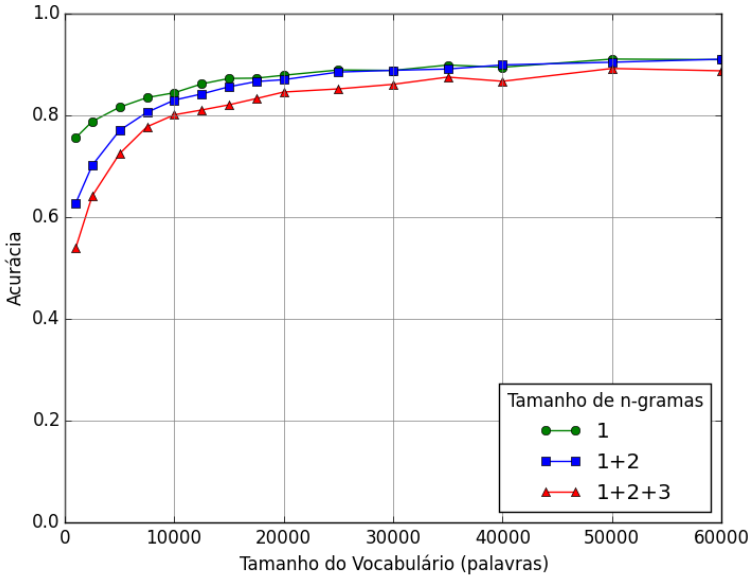
Figura 21 – Desempenho da hierarquia de classificadores com N-gramas, para condados dos Estados Unidos



Fonte: elaborado pelo autor

variados, o experimento foi repetido, desta vez, usando combinações de n-gramas de tamanho 1, 1+2 e 1+2+3. A vantagem de n-gramas de tamanho maior é que, espera-se, os n-gramas capturem um pouco da estrutura do texto. Esta é uma vantagem também para uma

Figura 22 – Desempenho da hierarquia de classificadores com N-gramas combinados, para regiões dos Estados Unidos



Fonte: elaborado pelo autor

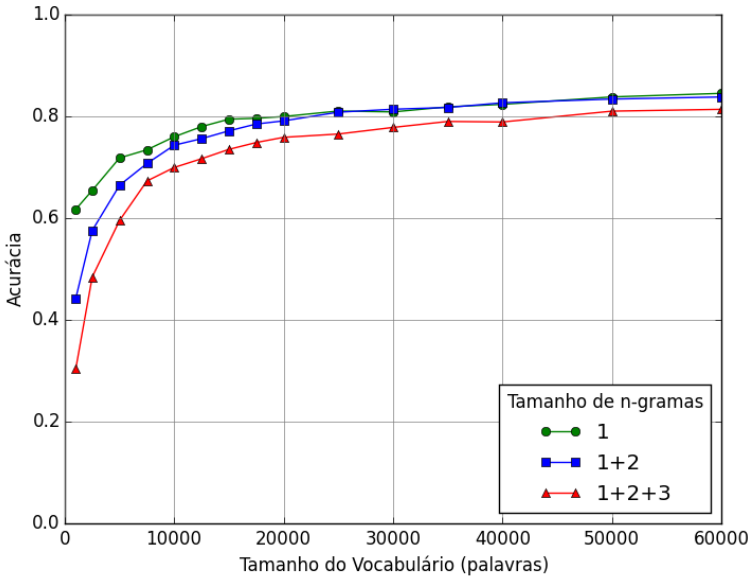
verificação manual do vocabulário selecionado automaticamente. Nomes compostos, como “San Diego” devem ser capturados pelo vocabulário de bigramas, enquanto no vocabulário composto exclusivamente de unigramas este nome de local estaria desmembrado em seus componentes “San” e “Diego”. A combinação de diferentes tamanhos de n-gramas deveria trazer o melhor dos dois mundos, pois apenas os termos com maior pontuação tf-idf são selecionados, independente do tamanho.

Observa-se experimentalmente que não existe vantagem, sob o ponto de vista de desempenho, em se usar n-gramas combinados para esta tarefa. O espaço dos n-gramas é cada vez mais esparso, quanto maior for n, ocorrências do mesmo n-grama são cada vez menos frequentes e há menos documentos exemplo nos níveis inferiores da hierarquia geográfica, daí a contribuição, por vezes negativa, do uso de n-gramas, conforme observado nas figuras 22, 23 e 24.

Dados os resultados deste experimento, o vocabulário dos 20.000 termos com maior valor TF-IDF e o uso combinado de unigramas e bigramas foi selecionado para a avaliação do desempenho do Geofier sem o gazetteer e uma breve análise do conteúdo do vocabulário selecionado automaticamente.

As tabelas 13 e 14 indicam desempenho equivalente ou melhor do Geofier, sem a necessidade de um gazetteer. A seção seguinte apresenta uma análise de evidências que podem indicar a razão deste desempenho em função do vocabulário selecionado.

Figura 23 – Desempenho da hierarquia de classificadores com N-gramas combinados, para estados dos Estados Unidos



Fonte: elaborado pelo autor

Tabela 13 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios do Geofier usando Naive Bayes, sem um gazetteer.

Nível Hierárquico	Precisão	Revocação	$F_{\beta=1}$
Region	0.98	0.98	0.98
Division	0.97	0.97	0.97
State	0.95	0.95	0.95
County	0.79	0.77	0.77

Fonte: elaborado pelo autor

Tabela 14 – Coeficientes de Precisão, Revocação e $F_{\beta=1}$ médios para diferentes níveis da hierarquia geográfica dos Estados Unidos usando SVM

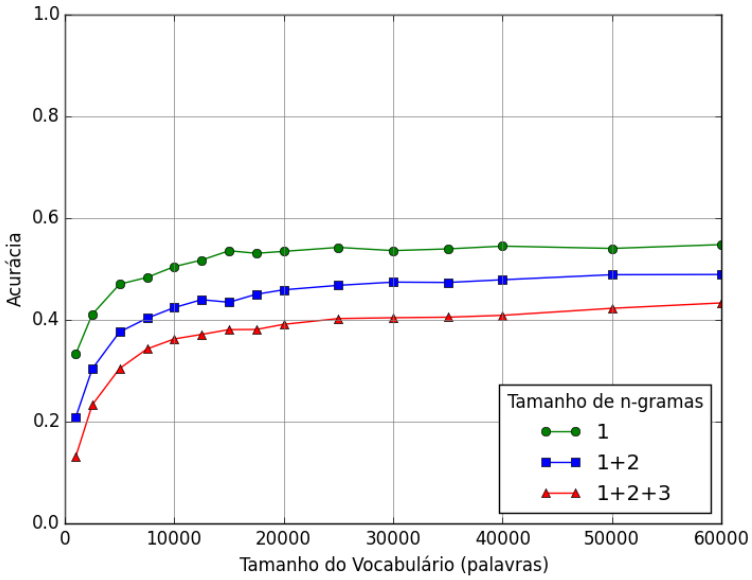
Nível Hierárquico	Precisão	Revocação	$F_{\beta=1}$
Region	0.99	0.99	0.99
Division	0.97	0.97	0.97
State	0.99	0.99	0.99
County	0.88	0.87	0.87

Fonte: elaborado pelo autor

5.5 Vocabulário Selecionado por tf-idf

Este experimento visa estimar quanto do gazetteer utilizado nos experimentos com o Web-a-Where e a Hierarquia de Classificadores foi coberto pelos unigramas e bigramas

Figura 24 – Desempenho da hierarquia de classificadores com N-gramas combinados, para condados dos Estados Unidos



Fonte: Elaborado pelo autor

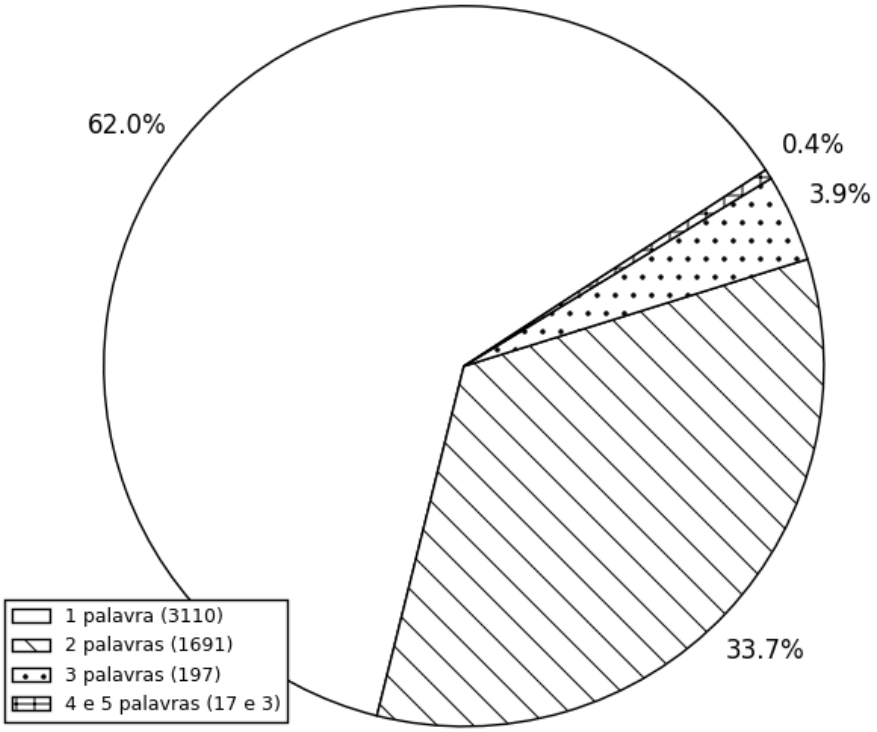
selecionados pelo cálculo de tf-idf para cada termo do vocabulário, pelo experimento 3.

O vocabulário selecionado foi limitado a 40.000 palavras, segundo indicativos de desempenho obtidos nas experiências anteriores. O conjunto de documentos utilizado contém 117.311 documentos. Estes documentos possuem 708952 unigramas, 12.191.567 bigramas e 33.755.474 trigramas diferentes.

Quanto maior o tamanho do n-grama, maior o número de n-gramas diferentes que poderão ser extraídos e menor a chance de que se repitam com frequência. O desempenho do uso de n-gramas de tamanhos variados em tarefas de classificação de textos foi estudado por Fürnkranz (FURNKRANZ, 1998), que observou que o uso de n-gramas de tamanho maior do que 2 pouco acrescenta ao desempenho da tarefa de classificação. As frequências de distribuição de n-gramas de diversos tamanhos em grandes corpora de textos foi estudada por Bardoel (BARDOEL, 2012), que observou que n-gramas maiores ocorrem sempre com menor frequência do que os n-gramas menores, pois os primeiros contém os segundos.

Os 5020 nomes de locais dos Estados Unidos usados pelo gazetteer são formados por nomes de 1 a 5 palavras, com predominância de nomes compostos de uma ou duas palavras. A figura 25 mostra a proporção dos comprimentos dos vários nomes de locais presentes no gazetteer. Observa-se que mais de 95% dos nomes presentes no gazetteer são compostos de uma ou duas palavras, portanto estes nomes deveriam estar presentes nos unigramas e bigramas selecionados.

Figura 25 – Proporção de n-gramas obtidos a partir dos nomes de locais do gazetteer



Fonte: elaborado pelo autor

5.6 Comparação entre Geofier e Web-a-Where

A divisão do território em regiões povoadas torna o gazetteer mais simples, mas limita a precisão que poderia ser alcançada pelo algoritmo, especialmente dentro dos Estados Unidos, onde uma grande área do território não está sob a jurisdição de uma municipalidade.

Uma das razões pelas quais o Web-a-Where falhou em atribuir documentos em regiões menores é dada pela interação entre o número de níveis na hierarquia geográfica, o número de menções a locais no texto e o fator de propagação dos pesos para os níveis hierárquicos superiores. Para textos com muitas menções a locais próximos, como é o caso dos textos da Wikipédia, um valor muito alto é propagado para os nós superiores da hierarquia, fazendo com que poucos documentos sejam classificados em seus condados, uma pequena parte em seus estados e a grande maioria seja classificada pertencentes ao país.

Os próprios autores do Web-a-Where argumentam que uma hierarquia com mais nós intermediários pode ser utilizada e que os valores de propagação de pesos propostos

no algoritmo publicado foram obtidos de maneira empírica. A otimização destes parâmetros, no entanto, não faziam parte do escopo deste trabalho, e distanciariam a implementação do Web-a-Where da que é frequentemente referenciada na literatura.

Um outro ponto fraco do Web-a-Where é sua dependência de um gazetteer com clara identificação de termos não geográficos. Ao identificar nomes de locais erroneamente, o Web-a-Where adiciona novos escopos geográficos possíveis para a tarefa de determinação do escopo. Estes escopos propagam seus pesos para os níveis superiores da hierarquia, aumentando a tendência de selecionar um escopo mais abrangente do que o realmente mencionado no texto.

A heurística utilizada para selecionar termos não geográficos sofre com as diferenças de escala entre a população do local e o número de menções ao termo em textos, sem nenhum tipo de normalização. Essa comparação funciona bem para exemplos extremos ou localidades pequenas, mas dependendo da população do local e do tamanho da base de textos onde é realizada a contagem, muitos falsos negativos e positivos ocorrem.

A correção destes falsos positivos é crucial para o desempenho do método e deve ser feita manualmente, como foi observado pelos autores, esta correção manual é certamente inviável para gazetteers muito grandes.

A heurística que considera o local de maior população como interpretação padrão para um topônimo também contribui para que locais pequenos tenham ainda menos relevância e escopos geográficos incorretos sejam selecionados, dando preferência a locais mais populosos, porém mais distantes do escopo do texto.

Ainda pior, a heurística da maior população incorpora ao algoritmo um viés presente na distribuição geográfica dos documentos exemplo, adequando os resultados do algoritmo à distribuição natural dos textos, sem efetivamente melhorar o desempenho do algoritmo em si. Uma análise envolvendo cada topônimo e os demais topônimos identificados no texto proporciona uma avaliação mais correta de seu significado, como observador por outros autores que publicaram trabalhos posteriores.

O Web-a-Where apresenta, no entanto, algumas características desejáveis: O Web-a-Where é facilmente capaz de distinguir entre um texto que contém evidências suficientes para seu algoritmo e um que não, através da identificação prévia dos locais presentes em seu gazetteer. O Web-a-Where também é capaz de determinar múltiplos escopos geográficos para um documento. Um texto comparando o custo de vida em São Paulo e no Rio de Janeiro poderia eventualmente ter estes dois escopos geográficos atribuídos. Esta característica torna sua utilidade mais geral do que textos com um foco principal, como é o caso das notícias e artigos enciclopédicos.

O Geofier apresentou acurácia muito superior ao Web-a-Where, partindo do mesmo gazetteer. Na primeira experiência com uma hierarquia de classificadores, os nomes de

locais do gazetteer do Web-a-Where foram utilizados para a criação do vetor de características que representariam os textos a serem classificados. Este experimento sugere que a análise de frequência dos termos e a aprendizagem automática da influência de cada termo na determinação da classe a que pertencem podem representar de maneira mais efetiva as sutilezas da ambiguidade geográfica dos topônimos do que um conjunto pré-estabelecido de regras, desenhadas pelo projetista do algoritmo seria capaz de capturar.

A substituição do gazetteer pela contagem de ocorrências de unigramas e bigramas, aliada à seleção de termos pela sua pontuação tf-idf ainda se mostrou mais eficiente do que o Web-a-Where e tão eficiente quanto a versão do Geofier que se apoiava em um vetor de características fixas, obtidos de um gazetteer. Uma análise dos termos selecionados por este método revelou que uma parte do gazetteer fora “aprendida” pelo método de seleção automática de atributos baseados em n-gramas, mas não todos.

O uso de classificadores SVM apresentou vantagem de desempenho de classificação em relação ao uso de Naive Bayes na maioria dos experimentos, porém o custo de treinamento do SVM tende a ser mais alto. As SVMs são classificadores essencialmente binários, em tarefas com múltiplas classes vários treinamentos precisam ser realizados para que cada par de categorias possa ser diferenciado. A hierarquia de classificadores contribui para manter o número de classes baixo, mas algumas regiões, ainda assim possuem classes (regiões geográficas) em número excessivamente grande. Por exemplo, o estado do Texas, nos Estados Unidos, possui 254 condados e provavelmente precisaria ser subdividido através da inclusão de uma divisão regional intermediária.

Os classificadores Naive Bayes apresentam vantagens quando o número de classes é excessivamente grande e podem ter vantagens práticas, mesmo que seu desempenho seja inferior ao da SVM.

Um vocabulário único e de tamanho fixo foi utilizado para o treinamento de todos os classificadores da hierarquia Geofier e a seleção automática dos termos com maior pontuação tf-idf foi feita levando-se em conta a relevância de cada termo para toda a coleção. É possível conseguir melhor desempenho se vocabulários específicos forem utilizados nos níveis superiores da hierarquia. Isso permitirá que termos com grande relevância para uma região menor sejam escolhidos para fazer parte do vocabulário de sua região. Com um vocabulário único tais termos precisariam ser muito repetidos em muitos documentos de uma região pequena, mas a quantidade de documentos disponíveis para regiões pequenas é menor, o que faz com que muitos termos relevantes não atinjam pontuação tf-idf suficiente para serem selecionados.

O uso de bigramas aparentemente acrescenta pouco para o desempenho do método, porém acrescenta muito à possibilidade de análise dos termos escolhidos para compor o vocabulário, já que uma parte significativa dos topônimos possui uma ou duas palavras. Foi possível observar que diversos topônimos importantes, como “Fifth Avenue” e “Central

Park” foram selecionados para compor o vocabulário do vetor de características, além de um grande número de topônimos, como “New York”, “Los Angeles” e “Dallas”.

Algumas construções da língua Inglesa foram reconhecidas. “Dallas”, por exemplo, aparecia no vocabulário de unigramas e também no vocabulário de bigramas com preposições de lugar, como “in Dallas” e “from Dallas”. Outros termos que apresentavam semântica geográfica, como “city center”, “residential neighborhood”, referências ao clima, como “subtropical climate” e celebridades como “George Bush”, “Bill Clinton” e “Barack Obama” também foram identificados como termos relevantes para a tarefa de identificação de escopo.

6 CONSIDERAÇÕES FINAIS

Este projeto de pesquisa especificou e implementou um método de identificação do escopo geográfico de um documento, que foi batizado Geofier. O método utiliza uma hierarquia de classificadores de aprendizagem de máquina, treinada a partir de documentos georreferenciados. Esta estratégia mostrou-se mais adequada, nos testes realizados, que o uso de consultas a um gazetteer, apresentando desempenho superior a outros métodos de determinação do escopo geográfico de documentos, em especial o Web-a-Where, com o qual foi diretamente comparado.

Embora o Web-a-Where tenha muitos passos e muitos detalhes, foi possível reproduzir a sua implementação. Esta reimplementação do algoritmo Web-a-Where ofereceu uma oportunidade única para compreender seus méritos e problemas, além de estabelecer uma base de comparação de desempenho para o Geofier.

O uso de múltiplos classificadores facilita a distribuição da tarefa de classificação entre vários computadores, permitindo escalar horizontalmente tanto as tarefas de treinamento quanto de atribuição de escopo geográfico a um volume potencialmente grande de documentos, obtidos da internet.

Outra vantagem prática da organização hierárquica dos classificadores é manter o número de classes relativamente baixo. Alguns tipos de classificadores, como os SVM precisam de múltiplos treinamentos para resolver problemas de classificação com múltiplas classes. São necessários n classificadores treinados para a estratégia “cada classe versus o resto” ou $n(n-1)$ classificadores treinados com cada par de classes individualmente. Um número grande de classes pode tornar o treinamento de um classificador nestas condições inviável na prática.

A técnica apresentada não depende de construções específicas de uma língua, tendo sido testada em português e em inglês. As palavras utilizadas foram separadas por espaços em branco. Existem tokenizadores eficientes para línguas orientais, o que sugere que a mesma técnica possa ser usada para estas línguas.

6.1 Trabalhos Futuros

O método Geofier pode ter seu desempenho melhorado pelo uso de vocabulários específicos para regiões menores, de forma que termos específicos de uma região pequena não tenham que competir por seu lugar no vocabulário com termos significativos por regiões maiores.

Outro ponto a ser tratado é que toda atribuição de escopo necessariamente percorre toda a profundidade da hierarquia geográfica, quando há situações em que o correto seria escolher um estado como escopo geográfico. Isto seria útil principalmente em regiões onde os classificadores não contam com exemplos de treinamento suficientes para atingir desempenho satisfatório. Métricas do desempenho obtidas durante o treinamento e validação podem ser utilizadas para que o algoritmo decida não prosseguir no percurso da árvore e atribua um escopo menos específico, porém com maiores chances de estar correto.

Outras técnicas de poda da árvore podem ser testadas, ao invés de treinar a árvore completa. Algumas regiões com poucos exemplos de treinamento disponíveis poderiam aguardar até que sejam apresentados exemplos suficientes para que o classificador seja treinado. Por exemplo, até que um número mínimo de exemplos de treinamento por classe estivesse disponível. Ainda outra possibilidade de otimização de desempenho seria paralelizar a implementação, por exemplo, com a execução de classificadores em máquinas diferentes.

As evidências colhidas da análise de termos frequentes na base de treinamento indicam um caminho promissor para a construção automática de um gazetteer, a partir de exemplos. Uma análise dos termos selecionados por frequência e por pontuação tf-idf realizadas por pares, como observado por outros autores pode ajudar a selecionar termos relevantes para cada classificador, através da informação mútua carregada pelos termos.

Adicionalmente, o estudo do efeito da evolução da grafia e variações nos topônimos ao longo do tempo propõe também a investigação da possibilidade de aplicação do Geofier na contextualização histórica de documentos, treinando-o com conjuntos de textos de diferentes períodos históricos.

REFERÊNCIAS

AMITAY, E.; HAR'EL, N.; SIVAN, R.; SOFFER, A. Web-a-where: geotagging web content. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2004. (SIGIR '04), p. 273–280.

ANDOGAH, G.; BOUMA, G.; NERBONNE, J. Every document has a geographical scope. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 81-82, p. 1–20, nov. 2012. ISSN 0169-023X.

ARCHIVE, I. *Internet Archive*. 1996. <http://archive.org/>. Acessado em: 19/08/2014.

BARDOEL, T. *Comparing n-gram frequency distributions: Explorative research on the discriminative power of n-gram frequencies in newswire corpora* . 31 p. Dissertação (Mestrado) — Tilburg University, Holanda, 5 2012.

BERKELEY, M. o. V. Z. University of; INSTITUTE the I. R. R. *Global administrative areas (boundaries)*. 2012. <http://www.gadm.org>. Acessado em 10/05/2014.

BUYUKKOKTEN, O.; CHO, J.; GARCIA-MOLINA, H.; GRAVANO, L.; SHIVAKUMAR, N. Exploiting Geographical Location Information of Web Pages. In: *WebDB (Informal Proceedings)*. [S.l.: s.n.], 1999. p. 91–96.

CENSUS, U. S. B. of the. *Geographic areas reference manual*. U.S. Dept. of Commerce, Economics and Statistics Administration, Bu reau of the Census, 1994. Disponível em: <https://books.google.com.br/books?id=fnQQp4N93QwC>.

CENSUS, U. S. B. of the. *Maps and Data - Geography - Us Census Bureau*. 2014. <https://www.census.gov/geo/maps-data/>. Acessado em 10/05/2014.

DING, J.; GRAVANO, L.; SHIVAKUMAR, N. Computing Geographical Scopes of Web Resources. p. 545–556, set. 2000.

FOX, N. *Official Google Blog: Ads are just answers*. 2011. <http://googleblog.blogspot.com.br/2011/10/ads-are-just-answers.html>. Acessado em 10/12/2013.

FURNKRANZ, J. A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, n. 1998, p. 1–10, 1998.

GEONAMES. *Geonames Geographical Database*. 2005. <http://www.geonames.org/>. Acessado em 17/06/2014.

GETTY. *Getty Thesaurus of Geographic Names*. 1987. <http://www.getty.edu/research/tools/vocabularies/tgn/>. Acessado em 17/06/2014.

GOODCHILD, M. F.; HILL, L. L. Introduction to digital gazetteer research. *Int. J. Geogr. Inf. Sci.*, Taylor & Francis, Inc., Bristol, PA, USA, v. 22, n. 10, p. 1039–1044, jan. 2008. ISSN 1365-8816.

HAREL, N.; AMITAY, E.; SIVAN, R. *Method and system for determining the focus of a document*. Google Patents, 2006. US Patent App. 11/165,527. Disponível em: <http://www.google.com/patents/US20060004752>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. Springer, 2009. Disponível em: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.

HECHT, B.; MOXLEY, E. Terabytes of tobler: Evaluating the first law in a massive, domain-neutral representation of world knowledge. In: *Proceedings of the 9th International Conference on Spatial Information Theory*. Berlin, Heidelberg: Springer-Verlag, 2009. (COSIT'09), p. 88–105.

HELDERS, S. *World Gazetteer Data (Zip File)*. 2005. <https://web.archive.org/web/20051229193956/http://www.world-gazetteer.com/dataen.zip>. Acessado em: 19/07/2014.

HILL, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. *Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000*, LNCS 1923, p. 280–290, 2000.

IBGE. *Normas de apresentação tabular*. 3. ed. Rio de Janeiro: Centro de Documentação e Disseminação de Informações. Fundação Intituto Brasileiro de Geografia e Estatística, 1993. Acesso em: 21 ago 2013.

IBGE. *Cadastro Nacional de Endereços para Fins Estatísticos*. 2010. <ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Cadastro_Nacional_de_Enderecos_Fins_Estatisticos>. Accessado em 19/07/2012.

IBGE. *Malhas digitais Brasileiras*. 2010. <ftp://geoftp.ibge.gov.br/malhas_digitais/municipio_2010/>. Acessado em: 19/07/2012.

JONES, K. S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, v. 28, n. 1, p. 11–21, 1972. ISSN 0022-0418.

LEIDNER, J. L. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Tese (Doutorado) — School of Informatics, University of Edinburgh, 2007. Disponível em: <http://www.era.lib.ed.ac.uk/bitstream/1842/1849/1/leidner-2007-phd.pdf>.

LELLA ADAM; LIPSMAN, A. *The US Mobile App Report*. 2014. Disponível em: <http://www.comscore.com/Insights/Presentations-and-Whitepapers/2014/The-US-Mobile-App-Report>.

LELLA ADAM; LIPSMAN, A. *2015 US Digital Future in Focus*. 2015. Disponível em: <http://www.comscore.com/Insights/Presentations-and-Whitepapers/2015/2015-US-Digital-Future-in-Focus>.

LIEBERMAN, M. D.; SAMET, H. Adaptive context features for toponym resolution in streaming news. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2012. (SIGIR '12), p. 731–740.

LIEBERMAN, M. D.; SAMET, H.; SANKARANARAYANAN, J.; SPERLING, J. STEWARD: architecture of a spatio-textual search engine. In: *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems - GIS '07*. New York, New York, USA: ACM Press, 2007. p. 1.

LUHN, H. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, v. 1, n. 4, p. 309–317, Oct 1957.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

MILLER, G. A. Wordnet: A lexical database for english. *Commun. ACM*, ACM, New York, NY, USA, v. 38, n. 11, p. 39–41, nov. 1995. ISSN 0001-0782.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

MLADENIC, D.; GROBELNIK, M. Word sequences as features in text-learning. *Proceedings of the 17th Electrotechnical and Computer Science Conference*, p. 145–148, 1998.

OSM FOUNDATION. *Nominatim*. 2009. <http://nominatim.openstreetmap.org/>. Acessado em 27/05/2014.

OVERELL, S. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. 1–181 p. Tese (Doutorado) — Department of Computing, Imperial College London, 2009.

POSTGIS, P. S. C. *PostGIS*. 2012. <http://postgis.net>. Acessado em: 19/07/2012.

SALTON, G.; BUCKLEY, C. *Term Weighting Approaches in Automatic Text Retrieval*. Ithaca, NY, USA, 1987.

SAMET, H.; SANKARANARAYANAN, J.; LIEBERMAN, M. D.; ADELFIO, M. D.; FRUIN, B. C.; LOTKOWSKI, J. M.; PANOZZO, D.; SPERLING, J.; TEITLER, B. E. Reading news with maps by exploiting spatial synonyms. *Commun. ACM*, ACM, New York, NY, USA, v. 57, n. 10, p. 64–77, set. 2014. ISSN 0001-0782.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys*, v. 34, n. 1, p. 1–47, mar. 2002.

SMITH, D.; CRANE, G. Disambiguating geographic names in a historical digital library. *. . . and Advanced Technology for Digital Libraries*, 2001.

TEITLER, B. E.; LIEBERMAN, M. D.; PANOZZO, D.; SANKARANARAYANAN, J.; SAMET, H.; SPERLING, J. NewsStand. In: *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08*. New York, New York, USA: ACM Press, 2008. p. 1.

TOBLER, W. A computer movie simulating urban growth in the detroit region. *Economic Geography*, v. 46, n. 2, p. 234–240, 1970.

WIKIPEDIA. *Predefinição: Coord*. 2005. <http://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Coord/doc>. Acessado em 10/12/2013.

WIKIPEDIA. *Predefinição: Geocoordenadas*. 2005. <https://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Geocoordenadas>. Acessado em 10/12/2013.

WIKIPEDIA. *Predefinição: Info/Município do Brasil*. 2005. <https://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Info/Munic%C3%ADpio_do_Brasil>. Acessado em 10/12/2013.

WIKIPEDIA. *Predefinição: Satélite*. 2005. <http://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Sat%C3%A9lite>. Acessado em 10/12/2013.

WIKIPEDIA. *Template: Infobox*. 2005. <https://en.wikipedia.org/wiki/Template:Infobox>. Acessado em 10/12/2014.

WIKIPEDIA. *Wikipedia: Lista de Infoboxes com conteúdo geográfico*. 2005. <https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes#Geography_and_place>. Acessado em 10/12/2014.

WIKIPEDIA. *GeoHack*. 2012. <https://wiki.toolserver.org/view/GeoHack>. Accessado em 19/07/2013.

WIKIPEDIA. *Ada, Ohio*. 2014. <https://en.wikipedia.org/wiki/Ada,_Ohio>. Acessado em 17/09/2014.

WIKIPEDIA. *Wikimedia Database Dumps*. 2014. <http://dumps.wikimedia.org/backup-index.html>. Acessado em 10/05/2014.

WOODRUFF, A. G.; PLAUNT, C. GIPSY : Automated Geographic Indexing of Text Documents Previous Work in Georeferencing of Text Documents. p. 1–21, 1994.

YAHOO. *DMOZ.org - The Open Directory Project*. 1999. <http://www.dmoz.org>. Acessado em: 10/01/2014.

ZHAI, C.; LAFFERTY, J. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, ACM, New York, NY, USA, v. 22, n. 2, p. 179–214, abr. 2004. ISSN 1046-8188.