CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Luciene Debortoli Dueli

TERMOS GEORREFERENCIADOS COMO MEIO DE RECUPERAÇÃO DE INFORMAÇÃO DE NOTÍCIAS NA WEB

Timóteo

2017

Luciene Debortoli Dueli

TERMOS GEORREFERENCIADOS COMO MEIO DE RECUPERAÇÃO DE INFORMAÇÃO DE NOTÍCIAS NA WEB

Trabalho de conclusão de curso apresentado ao curso de engenharia de computação do Centro Federal de Educação Tecnológico de Minas Gerais como pré-requisito para obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Leonardo Lacerda Alves

Timóteo

2017

Luciene Debortoli Dueli

Termos georreferenciados como meio de recuperação de informação de notícias na Web

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Trabalho Aprovado. Timóteo, 18 de agosto de 2017.

Prof. Leonardo Lacerda Alves
Orientador

Prof. Marcelo de Sousa Balbino
Prof. Convidado

Prof. Odilon Correa da Silva
Prof. Convidado

[!h] A Deus, que me deu força nesta caminhada.
A todos que estiveram ao meu lado
me apoiando e incentivando.

Agradecimentos

A Deus por ter me dado saúde e força para superar as dificuldades encontradas ao longo deste caminho.

Aos meus pais João Bosco e Maria do Carmo pela capacidade de acreditarem e investirem em mim.

À minha irmã Silvany pelo forte incentivo e aos meus irmão Marco Antônio e Luciano pelo carinho. A todos que compartilharam minha ausência, mas que hoje colho o fruto de todo o esforço, o apoio e compreensão foi fundamental para chegar até aqui.

Ao meu noivo, Douglas, obrigada pelo carinho, a paciência e por sua capacidade de me trazer paz na correria de cada semestre.

Ao meu orientador Leonardo Lacerda, pelo suporte e incentivos durante este trabalho.

Agradeço a todos os professores por me proporcionar conhecimento não apenas racional, mas manifestação do caráter e afetividade da educação no processo de formação profissional.

Resumo

Com tantas informações na internet muitas das vezes os usuários tem dificuldades de encontrar o que desejam. No caso de noticias da web existem vários métodos para a recuperação da informação, mas muitas vezes os termos escolhidos são feitos sem padrão e estão ligados aos sentimentos momentâneos do redator ao indexar conteúdos. No processo de recuperação é comum que os usuários levem em consideração o contexto geográfico em função de sua localização. O trabalho objetiva identificar características comuns em webjornais que deem suporte às escolhas de termos relevantes e georreferenciados, para isso é importante a identificação do local em que ocorreu a notícia. Desta forma, foram coletadas diversas notícias de 23 veículos da Região Sudeste, toda a coleção de notícias foram tokenizados e armazenadas em um banco de dados para análises e identificação de possíveis indicadores de localidade. Foram analisadas as palavras em diferentes modos, considerando quantidades diferentes de palavras por notícias, apenas palavras com as iniciais maiúsculas e apenas o final do texto. Com os dados obtidos, foi possível concluir que palavras com a primeira letra maiúscula abrange um número maior de indicadores de localidade e ainda se for considerado apenas o lead do jornal já seria suficiente para identificação de indicadores de localidade tendo como ponto positivo a queda do custo de processamento.

Palavras-chave: Recuperação da informação, indicadores de localidade, notícias, tags

Abstract

With a lot of information on internet many times users have difficulties finding what they want. In case of web news there are several methods for information retrieval, but often the terms chosen are made without a pattern and linked to the momentary feelings of the editor when indexing content. In recovery process, it's common that users take into account the geographical context according to their location. The work aims to identify common characteristics in web news that support the choice of relevant and georeferenced terms, for this, it's important to identify the place where the news occurred. In this way, several news were collected from 23 vehicles of the Southeast Region, the entire collection of news were tokenized and stored in a database for analysis and identification of possible locality indicators. Words were analyzed in different ways: considering different amounts of words in each news, only words with the initial capital letters and only the end of the text. With the data obtained, it was possible to conclude that words with the first capital letter cover a greater number of locality indicators and even if it is considered only the lead of the newspaper would already be enough to identify locality indicators having as a positive point the fall of processing cost.

Keywords: Information retrieval, location indicators, news, tags

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A Internet é uma das principais fontes de informação para muitos usuários e os sistemas de buscas são os métodos mais utilizados para a recuperação da informação na Internet (DETERS, 2003). No entanto, recuperar páginas na Web não é uma tarefa tão simples. Por mais que seja uma necessidade comum para os usuários, a busca costuma ser uma tarefa frustrante, sendo que o usuário pode não conseguir encontrar o que deseja diante de tantas informações apresentadas (LOH; WIVES; FRAINER, 1999). No caso especíco de notícias, são inúmeros veículos de informação em diversos países e em vários idiomas e as notícias não são indexadas com sua referência geográfica, desta forma a recuperação de informação para o usuário pode não ser tão interessante por apresentar informações que sejam muito distantes do seu espaço geográfico.

No caso dos sistemas de recomendação, se tenta prever conteúdos e informações, principalmente jornalísticos, nos quais o usuário possa se interessar, baseando-se no perfil do usuário, seu comportamento e localização (JUNIOR; BRANCO; BARBOSA, 2010). Outro mecanismo é a busca direta, pela qual o usuário informa algumas palavras que indicam o tema de seu interesse (MORAIS; AMBRÓSIO, 2007). A busca direta utiliza métodos de busca de texto para localizar documentos relevantes, onde as buscas normalmente acontecem diretamente nos textos originais. Por fim, os mecanismos baseados em visualização, este permite que o usuário encontre informação por meio de interfaces gráficas, reduzindo algumas dificuldades na recuperação de informação, permitindo a interação com formas alternativas de representação e facilitando não só a recuperação, mas também a navegação e o compartilhamento de informação com outros usuários (GALDO; VIEIRA; RODRIGUES, 2009). O mecanismo de nuvens de tags é um bom exemplo de mecanismo baseado em visualização.

Em qualquer dos métodos, as escolhas dos termos relevantes não é uma tarefa tão simples. De acordo com (GOUVÊA; LOH, 2007), o processo de geração ou seleção de termos geralmente é feito por autores de notícias e até mesmo por usuários sem critérios bem definidos. Ainda, (NASCIMENTO; NEVES, 2012) diz que o ato de indexar ou etiquetar conteúdos está ligado diretamente aos sentimentos momentâneos de quem está indexando, e, pela falta de conhecimento, acaba criando polissemia, sinonímia, inflexões e erros ortográficos, prejudicando a recuperação da informação.

A maioria dos métodos baseiam-se em uma heurística que faz uso da frequência de palavras no texto e na coleção de textos, o que, segundo (REYNAR, 1999), é um bom indicador do conteúdo de um documento. Métodos baseadas na frequência relativa e na frequência inversa ajudam a definir quais termos podem ou devem ser usados para recuperar determinado texto (LOH; WIVES; FRAINER, 1999).

Por outro lado, outra heurística comum é considerar que a localização da notícia seja um indicador adequado de relevância e do tema para o usuário, em função de sua própria localização. É comum que os usuários levem em consideração o contexto geográfico dos do-

cumentos no processo de recuperação de informação, (VARGAS, 2012). Isso torna os resultados mais úteis para eles. No entanto, os sistemas de buscas tradicionais ainda não tratam da contextualização geográfica.

Desta forma, o objetivo deste trabalho é identificar características comuns em textos jornalísticos produzidos por veículos de alguns estados brasileiros e que dêem suporte à escolha de termos relevantes para a recuperação de notícias georreferenciadas.

1.1 Método

O presente trabalho utiliza textos jornalísticos como bases de dados. Dessa forma, são extraídos textos de diversos sites de noticias da região Sudeste do Brasil, esta escolha foi devido ao maior conhecimento de nomes de locais da região.

Vale ressaltar que nem todos os sites selecionados fazem o uso de tags em suas páginas, podemos citar como exemplos os sites que utilizam tags: Estadão e Uai, essas tags podem ser encontradas logo abaixo da narração da noticia. Como nem todos os sites possuem tags definidas será feito um processo de tokenização a partir do texto de cada notícia.

1.2 Problema

O presente trabalho estudará notícias da web para verificar se é possível sua recuperação georreferenciada, identificando o local ou uma aproximação que ocorreu a notícia. A maior dificuldade será encontrar a localização ou pelo menos uma aproximação do local que se refere a notícia.

2 Revisão da literatura

Tag, em inglês, significa etiqueta ou rótulo, que segundo (VIEIRA; CARVALHO; LAZZARIN, 2013) corresponde uma palavra escolhida pelo autor de modo informal e pessoal, utilizada para descrever o conteúdo de informação e permite uma indexação baseada em palavras-chaves, facilitando a busca e a recuperação das publicações.

O estudo de tags pode ser um ponto de partida para reconhecer os termos que os leitores de notícias consideram mais importantes e se existe algum padrão na escolha de termos para melhor indexar e recuperar informações pelos usuários de forma que sua navegação seja mais produtiva e o conteúdo seja relevante para o usuário. Neste capítulo são apresentados trabalhos relacionados às técnicas de processamento e mineração de dados trazendo o conceito de tags.

2.1 MTCIR: A multi-term tag cloud information retrieval system

Devido à popularização do uso de termos e palavras-chave em recuperação de informação, muitos deles não representam de fato o contexto do objeto ou podem conter imprecisões, pois quem escolhe esses termos são pessoas comuns e não especialistas. Assim, a recuperação de informação pode-se tornar falha.

Desta forma, (TORRES-PAREJO et al., 2013) apresentaram MTCIR, uma nuvem de tags multi-termo de sistema de recuperação de informação.

> "O sistema utiliza uma interface visual com base em nuvens de tag multi-prazo, que apresenta o conteúdo da base de dados e podem ser usados como ajuda no processo de pesquisa" [1] (TORRES-PAREJO et al., 2013, p.5448, tradução nossa).

O problema foi concentrado ao acesso à informação em bases de dados de textos e levado em consideração a dificuldade do usuário em fazer uma busca, principalmente quando não tem conhecimento do conteúdo do banco de dados. Como solução para esse problema, dar ao usuário um conjunto de sugestão de consulta auxilia-o no processo de busca com resultados mais apropriados. Com esta possibilidade o uso de tags como meio de categorização se torna essencial e a nuvem de tags facilita a procura, exploração e representação do conteúdo do banco de dados, uma vez que a nuvem de tag tem como característica as tags mais relevantes. Pelo fato da ausência de um sistema de tagging, os termos foram gerados a partir do texto no banco de dados.

> "O uso de etiquetas de multi-prazo fornece determinado contexto e fornece orientação para os usuários, aumentando assim a satisfação no processo de recuperação" [2](TORRES-PAREJO et al., 2013, p. 5449, tradução nossa).

[1] The system uses a visual interface based on multi-term tag clouds, which presents the content of the database and can be used as assistance in the search process.

[2] The use of multi-term tags provides certain context and orientation for the users, thus increasing the satisfaction in the retrieval process.

Conforme (RIVADENEIRA et al., 2007), as nuvens de tags também podem dar suporte aos usuários nos seguintes itens:

- Pesquisa: localização de termos específicos;
- Navegação: navegação em diversas páginas por meio das nuvens de tags sem propósito específico;
- Impressão: visão rápida dos conteúdos da página, inclusive dos mais populares;
- Reconhecimento: distinção de palavras iguais, mas com significados diferentes, como por exemplo, manga que pode ser fruta ou parte de uma peça de roupa.

(TORRES-PAREJO et al., 2013) propuseram uma metodologia geral para geração de nuvens de tags de texto, utilizando técnicas de mineração de texto e ferramentas externas. Essas estruturas são:

- Pré-processamento sintático: limpeza de dados no sentido sintático, tokenização, remoção de stopwords e etc. Algumas ferramentas para auxiliar nesta fase: ANNIE, Lucene, PrePro2010 e DB2DS. Para este pré-processamento os autores optaram pelo DB2DS
- Pré-processamento semântico: detecta e agrupa sinônimos, cada grupo de sinônimos terá um representante canônico. Neste processo é identificado o sentido de um conjunto de termos. Para detectar os sinônimos são necessários os seguintes passo:
  - Parte do discurso de tagging (POS): O termo é marcado como parte do discurso e calculado o POS do texto de origem de acordo com o contexto. Algumas sugestões de software são: Tree Tagger, TNT, SVMToll e Stanford Tagger. Os autores escolheram Stanford para executar esta tarefa.
  - Senso desambiguação da palavra: determinação do senso de cada palavra. WordNet, Lesk algorithm, Simplified Lesk algorithm e Adapted Lesk Algorithm podem ser usados para esta tarefa, os autores optaram por usar Adapter Lesk Algorithm por oferecer melhores resultados.
  - Geração conjunto de sinônimos: É atribuído a um conjunto de sinônimos todos os termos com o mesmo sentido.
  - Seleção representativa canonical: É determinado um termo para representar cada conjunto de sinônimo.

  Nesta etapa pode ser utilizadas ferramentas como dicionários eletrônicos, tesauros e WordNet.
- Representação forma intermediaria: é uma forma de armazenar os textos pré-processados, para a geração da nuvem de tags, com os temos mais frequentes. Os autores usam a forma de representação AP-Sets com uma versão estendida do Ap-Sets, denominada

AP-Seqs. “O Ap-Sets são organizados em estruturas ordenadas ponderadas chamadas wapo-estrutura”[3] (TORRES-PAREJO et al., 2013, p. 5451, tradução nossa).

- Geração da nuvem de tags: A nuvem de tags é gerada a partir dos termos mais frequentes utilizando Wordle.

(TORRES-PAREJO et al., 2013) geraram nuvens de tags de textos de temas diferentes. Foram calculadas métricas, que retornam valores no intervalo [0,1] a partir de dados recuperados diante de uma consulta. Estas métricas são para calcular cobertura, sobreposição e equilíbrio.

> Cobertura nos dá a fração dos dados originais cobertos pelos termos da nuvem de tags. [...] Sobreposição determina em que medida os diferentes termos na nuvem de tags podem recuperar os mesmos objetos nos dados originais [...]. Equilíbrio é uma métrica para medir a quantidade de resultados obtidos pelos termos da nuvem de tags [...]. (TORRES-PAREJO et al., 2013, p. 5453-5454, tradução nossa).

De acordo com (TORRES-PAREJO et al., 2013), os resultados obtidos foram semelhantes aos resultados feitos por experiências não computacionais.

2.2 Métodos para seleção automática de tags para descrever noticias na Web

De acordo com (ZIESEMER et al., 2012), quando se há uma má escolha de tags os resultados das buscas podem se tornar invisíveis ou insuficientes para o usuário. Como meio de minimizar estes problemas vem surgindo muitos estudos que buscam métodos para sugestão de tags automáticas para o autor do conteúdo, portanto, com as sugestões, as tags seriam mais qualificadas e economizaria tempo na escolha delas. Bertocchi relata que:

> “Em boa parte dos cibermeios, a prática do tagueamento no jornalismo ainda está sendo feito de forma subjetiva e descontrolada. No processo de etiquetamento, leva-se em conta o repertório individual de cada jornalista da redação. As redações abrem mão, assim, de uma técnica jornalística objetiva para etiquetar narrativas”. (BERTOCCHI, 2009, p.15)

Gouvêa, Loh e Garcia (2008) trabalharam com notícias da web e utilizaram técnicas já existentes de seleção automática de tags e foram comparadas considerando a natureza do texto jornalístico e assim sugerir métodos de identificação de tags mais relevantes para descrição da noticia. Considerando o fato de que os textos jornalísticos são escritos utilizando a técnica de pirâmide invertida, sendo a parte superior da notícia contendo os dados mais importantes e a inferior as informações menos relevantes, a pesquisa de (GOUVÊA; LOH; GARCIA, 2008) só é valida para texto acerco desta característica.

Uma análise preliminar de padrões de ocorrências de tags em uma coleção de notícias cadastradas no site Delicious, (GOUVÊA; LOH, 2007) perceberam que a porcentagem de tag cadastradas que estavam presente nas notícias foi baixa, ou seja, muitas tags cadastradas

[3] The AP-Seqs are organized in weighted ordered structures called WAPO-Structures.

estavam relacionadas somente com quem as cadastrou. As tags cadastradas que estavam presentes no texto estavam mais relacionadas com o título e o primeiro parágrafo. Em números, 27% das tags escolhidas estavam presentes na noticia e 74,2% das noticias possuíam ao menos uma das tags presente no texto. Também foi comprovado que a escolha de substantivos como tag são bem eficazes.

Logo, o objetivo deste trabalho foi comparar diferentes técnicas de seleção automática de tags aplicados sobre notícias. O experimento foi feito com 1000 textos jornalísticos de dois sites de noticias. Foi utiliza o calculo TFIDF (term frequency–inverse document frequency) e devido este calculo depender dos textos analisados foram feitas outras analises divididas em 10 técnicas:

> Título: esta técnica seleciona como tags os substantivos presentes no título da notícia;
>
> 1ª Frase: esta técnica seleciona os substantivos presentes na 1ª frase do texto da notícia;
>
> Top 3: esta técnica seleciona os 3 substantivos mais frequentes no texto da notícia;
>
> Título e 1ª Frase: esta técnica seleciona os substantivos presentes tanto no título da notícia quanto na 1ª frase do texto;
>
> Título e TOP 3: esta técnica seleciona os substantivos presentes tanto no título da notícia quanto entre os 3 substantivos mais frequentes no texto;
>
> 1ª Frase e TOP 3: esta técnica seleciona os substantivos presentes tanto na 1ª frase do texto da notícia quanto entre os 3 substantivos mais frequentes neste texto;
>
> Top 3 TFIDF_local 1000: esta técnica seleciona as 3 palavras mais relevantes da notícia utilizando o método TFIDF, considerando a frequência da palavra no texto da notícia e a frequência na coleção de teste composta de 1000 notícias;
>
> Top 3 TFIDF_old 1000: esta técnica é semelhante à anterior, mas utiliza uma coleção de treino com 1000 notícias publicadas 1 ano antes das notícias da coleção de teste (a idéia é testar se uma coleção de treino mais antiga pode gerar resultados semelhantes ou não);
>
> Top 3 TFIDF_new 1000: esta técnica é semelhante às anteriores mas utiliza como treino uma coleção de 1000 notícias do mesmo período de tempo das notícias de teste (mas notícias diferentes das de teste);
>
> Top 3 TFIDF_new 4000: idem à anterior mas utilizando uma coleção de treino com 4000 notícias (do mesmo período da coleção de teste), a ideia é testar se um número maior de notícias de treino pode melhorar os resultados. (GOUVÊA; LOH; GARCIA, 2008, p.2)

Depois de aplicadas as técnicas (GOUVÊA; LOH; GARCIA, 2008) separaram os textos por tags para avaliar as qualidades das tags usando as métricas de coesão e acoplamento. Em seguida foi utilizada a função cosseno para medir a similaridade. Com os dados obtidos foi calculado o quociente da coesão pelo acoplamento, pois os resultados isolados não são suficientes. Para validar as similaridades obtidas foram adicionadas as notícias em cluster aleatórias assim que foi calculada a média da similaridade do cosseno.

Como resultado, a utilização do TFIDF apresentou os melhores resultados, no entanto o custo é o maior visto que é necessária sua análise para cada nova notícia inserida na coleção. O método de TFIDFnew aparece como alternativa, demonstrando resultados um pouco

inferiores ao anterior mas tendo como vantagem o custo mais baixo para sua execução, visto que não é necessária sua análise a cada notícia inserida

Desconsiderando o método TFIDF, as intersecções aumentaram a qualidade das análises de substantivo no Título e 1ª Frase (destaque para Título com TOP 3 e 1ª Frase com TOP 3) os quais são mais simples de executar e obtiveram resultados um pouco inferires ao TFIDF, mas com um custo bem menor.

O trabalho apresentado de (GOUVÊA; LOH; GARCIA, 2008) não identifica as tags que co-ocorrem nas noticias, apenas estuda o caso de geração automáticas de forma eficaz, um caso a parte para geração da uma nuvem de tags com eficiência para recuperação de informação, além de não associar as tags com outros critérios como temporal e geoespacial por exemplo.

2.3 Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing

(GOUVÊA et al., 2008) analisaram notícias da web na língua portuguesa (onde a localidade não estava explícita no texto) para identificar indicadores de localização. Li, citado por Gouvêa et al. (2008) diz que uma alternativa para resolver o problema de ambiguidade é utilizando técnica do Word Sense Disambiguation, criando estrutura chamadas de Gazetteers.

> "Gazetteer digital ou simplesmente gazetteer, é um dicionário geoespacial de nomes geográficos. Também conhecido como dicionário toponímico ou dicionário de lugares, é uma versão digital dos índices de topônimos, geralmente encontrados em Atlas." (MACHADO, 2011, p.33)

> "Um gazetteer digital (ou simplesmente gazetteer) é um dicionário geo-espacial de nomes geográficos, ou seja, uma coleção de nomes de lugares associados à sua localização e a outras informações descritivas"(VASCONCELOS, 2006, p.22)

"A função básica de um gazetteer é informar a localização de um lugar dado seu nome, sendo essa a forma mais natural de se referenciar."(VASCONCELOS, 2006)

Segundo (GOUVÊA, 2009), os Gazetteers "têm sido utilizados para armazenar os variados tipos de referências às localidades e sua posição espacial (utilizando para isso coordenadas geográficas)". Por outro lado apresentam pontos negativos, que de acordo com (GOUVÊA, 2009), por serem criados manualmente, os gazetteers podem apresentar falta de informações que identifiquem alguns tipos de lugares que são definidos como indicadores de localidade. No Brasil, os gazetteers não possuem cobertura suficiente para reconhecer os lugares do país com granularidade espacial mais fina.(VASCONCELOS, 2006)

Alguns gazetteers tem suas limitações, como relata (MACHADO, 2011), gazetteers disponibilizados na web possuem várias limitações tornando difícil o uso para recuperação de informação, eles apresentam estruturas simples composto por apenas o nome do lugar, o tipo e o footprint (par de coordenadas de localização) e não inclui nomes de lugares intra-urbanos, problema também citado por (VASCONCELOS, 2006), além de não existir recursos

para registrar e utilizar o relacionamento espacial entre os elementos, salve as estimativas de proximidades baseadas nas coordenadas. Pelas limitações e importância desta ferramenta, (MACHADO, 2011) deixa claro que há a necessidade de uma geração nova de gazetteers que seja efetivo na recuperação de informação geográfica.

Alguns trabalhos citados pelos autores utilizaram algumas técnicas para a criação de gazetteer, como aprendizado supervisionado, utilização da Wikipédia combinado com Wordnet e Flickr, cada um com seus processos diferenciados. Mas o problema em usar a Wikipédia, Wordnet ou Flickr segundo (GOUVÊA et al., 2008), é que eles são atualizados manualmente, acarretando falta de informações necessárias para a criação do gazetteer.

O objetivo do trabalho de (GOUVÊA et al., 2008) foi identificar indicadores de localização geográficas de textos jornalísticos da Web, supondo que a maioria dos textos tenham um indicador de localização e acreditando que, com análise estatística, possa recuperar notícias de acordo com a localização.

O primeiro passo foi a coleta de notícias aleatórias na Internet, posteriormente foi feita a identificação de indicadores de localidades, desta forma foram selecionadas todas as palavras com a primeira letra maiúscula que correspondem aos nomes próprios, considerando também os nomes compostos e expressões que incluem preposições. A relação entre nomes de cidades e indicadores de localização possuem um peso que representa a probabilidade da relação. A ideia foi calcular um peso local dentro de uma notícia e um peso global (que representa a importância da relação da cidade quanto ao indicador). Com os pesos foram criados vários gazetteers, cada um possuindo um conjunto de cidades e cada cidade foi associada a uma lista de indicadores de localização sendo que cada associação tem um peso global. Os gazetteers criados por (GOUVÊA et al., 2008) foram:

- 3000 NP x NP antiga: composto por 3000 mil notícias entre os anos 2001 e 2006, considerando apenas nomes próprios.
- 3000 NP x NP nova: igual ao gazetteer anterior, porém as notícias são de 2007 a 2008.
- 6000 NP x NP (antigo e nova): junção dos dois gazetteers anteriores, o objetivo é testar com maior número de notícias para gerarem melhores resultados.
- 3000 NP x NP (SA): formado por 3000 notícias recentes e que estão relacionados com apenas uma cidade.
- Linha de Base: criado a partir de um banco de dados que contém todas as cidades e bairros brasileiros

Com as análises, (GOUVÊA et al., 2008) observaram que os quatro gazetteers geram melhores resultados se comparada com a Linha de Base e chegaram a uma conclusão:

> "Concluímos então que as notícias são úteis para a criação de gazetteers e também melhora os processos de georreferenciamento. Notícias podem ajudar a identificação de indicadores de localização que não estão relacionados com ruas e bairros" (GOUVÊA et al., 2008, p.9, tradução nossa)

Com as comparações entre os gazetteers, também concluíram que grande coleção de notícias e notícias mais recentes têm melhores resultados para a construção de gazetteers e textos que possuem apenas uma cidade geram poucos indicadores de localidade.

Com o trabalho, Gouvêa et al. (2008) demonstraram que a construção de gazetteers com indicadores de localização são mais úteis para melhorar os processos de georreferenciamento do que usar nomes de cidades, bairros e ruas e que estes indicadores podem ser descobertos pela análise de notícias da web. As notícias também podem trazer diferentes indicadores de localização como nome de pessoas muito importantes relacionadas a locais geográficos.

Os autores ressaltam ainda que com a proposta abordada, a criação de Gazetteers pode ser feita de forma automática, através de captura de notícias na web, cobrindo um grande números de locais e mantendo informações atualizadas sobre cada local com menos esforço. Além disso as notícias são mais acessíveis do que nomes de bairros e ruas, pois banco de dados com estes nomes são difíceis de serem encontrados ou são pagos.

3 Criação do banco de palavras a partir das notícias

Para as análises dos conteúdos das notícias foi preciso criar um dicionário de palavras, ou seja, a tokenização das notícias, que passaram por vários processos até a criação do banco de dados. Este capítulo tratará das etapas dos processos realizados desde a coleta das notícias até a popularização do banco de dados.

3.1 Coleta de dados

Primeiramente foi feita a coleta de notícias de 23 veículos da região Sudeste do Brasil através do software HTTrack (2015). Foram coletadas no total 22.959 notícias. Na tabela 1 são listados os veículos de notícias e as respectivas quantidades de notícias.

Tabela 1: Relação dos veículos escolhidos e as respectivas quantidades de notícias coletadas

Veículo	Quantidade de notícias
A Gazeta	3028
Correio do Sul	30
Diário de Notícias	1133
Diário de Petrópolis	42
Diário do Aço	383
Diário Rio Doce	2558
Diário São Paulo	172
Espírito Santo Hoje	321
Extra Rio de Janeiro	369
Folha Contagem	1823
Folha São Paulo	248
Folha Vitória	1035
Hoje Em Dia	401
Jornal da Manhã Online	332
Jornal Vale do Aço	237
Net Diário Rio de Janeiro	4320
O Dia Rio de Janeiro	122
O Estadão de São Paulo	2038
O Tempo	325
Portal O Rio	164
Sete Dias	147
Tribuna de Minas	3682
Uai	49

Fonte: Elaborada pela autora.

Das próximas vezes que forem citados os jornais Diário do Aço, Diário Rio Doce, Espírito Santo Hoje, Jornal da Manhã Online, Diário Petrópolis e Jornal Vale do Aço, serão usadas as respectivas siglas DA, DRD, ESH, JMO, DP e JVA.

Após a coleta de todos os veículos, foi preciso selecionar apenas os arquivos referentes às notícias, pois o software utilizado faz a coleta de todos os diretórios e arquivos do site, como imagens e outros arquivos do computador remoto. Por isso, os arquivos foram analisados e aqueles relacionados às notícias foram selecionados e armazenados em diretórios, sendo um para cada veículo.

3.2 O Banco de dados

O banco de dados foi constituído com apenas uma tabela com quatro atributos:

- nomepalavra: cada palavra, ou seja, os tokens
- posição: cada palavra de uma notícia é associada a uma posição, assim podemos identificar o antecedente e o sucessor de cada uma, isto é feito de forma que toda vez que começar uma nova notícia a posição reinicie em zero. Com a última posição é possível identificar a quantidade de palavras de cada notícia.
- notícia: nome da notícia referente à palavra, ou seja, nome do arquivo, já que a maioria dos arquivos foram salvos com o nome do titulo da noticia.
- jornal: nome do veículo referente à notícia e à palavra.

3.3 Tokenização

Pelo fato das notícias estarem em formato HTML, foi preciso remover todas as tags HTML e informações que não pertenciam ao corpo da notícia e que poderiam interferir no resultado dos processos subsequentes. Alguns processos foram realizado antes da tokenização, na figura 1 representa de forma esquematizada os processos descritos abaixo:

No primeiro passo, foi feito o reconhecimento manual dos arquivos em HTML de cada veículo, identificando padrões de tags que limitavam o corpo da notícia para que fosse copiada e armazenada em um arquivo em formato txt, ainda com as tags HTML. Este processo foi realizado com um algoritmo desenvolvido na linguagem java.

O segundo passo foi a padronização de caracteres de todos os documentos HTML, uniformizando todos os arquivos para o padrão UTF-8, padrão já estabelecido na maioria dos arquivos.

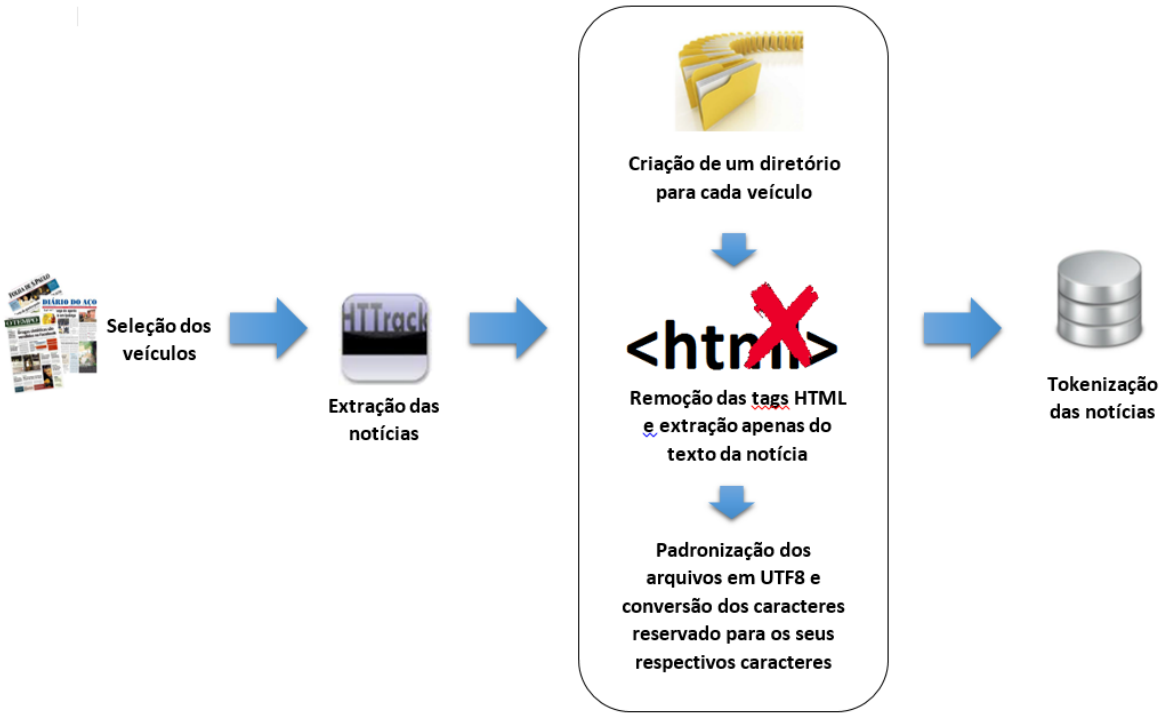
Após a padronização foi preciso normalizar a acentuação dos textos devido aos caracteres reservados (entities) do HTML. Para este terceiro passo foi utilizado um código em PHP adotando a função html_entity_decode que converte todas as entidades HTML para seus respectivos caracteres.

O quarto passo foi a retirada das linhas vazias com o uso da função preg_replace e expressão regular em PHP. Este processo não é obrigatório uma vez que na fase de tokenização bastaria acrescentar na expressão regular uma notação para ignorar as linhas vazias.

A última etapa foi a tokenização, onde foi criada uma expressão regular para que separassem os tokens com o uso da função spliti em PHP, ela separa as strings em array utilizando expressões regulares, a separação foi realizada onde encontrassem os seguintes caracteres: interrogação (?), exclamação (!), ponto (.), aspas simples esquerda e direita (‘,’), aspas duplas("), aspas duplas direita e esquerda (“,”), crase (‘), barra invertida (/), dois ponto (:), número três sobrescrito (3), número dois sobrescrito (2), barra vertical (|), indicadores ordinais (º, ª), grau (°), travessão (—), asterisco (*), reticências (. . .), apostrofo (’), parentes e espaços. Após a tokenização as strings foram salvas no banco de dados com suas respetivas informações.

Algumas observações devem ser consideradas, pela opção de não quebrar as palavras com hífen, ele não foi incluído na expressão regular, dessa forma no banco de dados haverá palavras concatenadas com hífen (isso se não foi colocado espaço entre a palavra e o hífen), por exemplo, vende-se será considerado como palavra única. Também foi considerado o hífen como palavra por existir espaços entre ele, outro caso que será encontrado são palavras concatenadas a um hífen, por exemplo, Urbanos-, podendo ser ou não um erro do redator por colocar ou deixar de colocar um espaço na ligação entre as palavras.

Figura 1: Esquema dos processos de seleção dos veículos até a tokenização



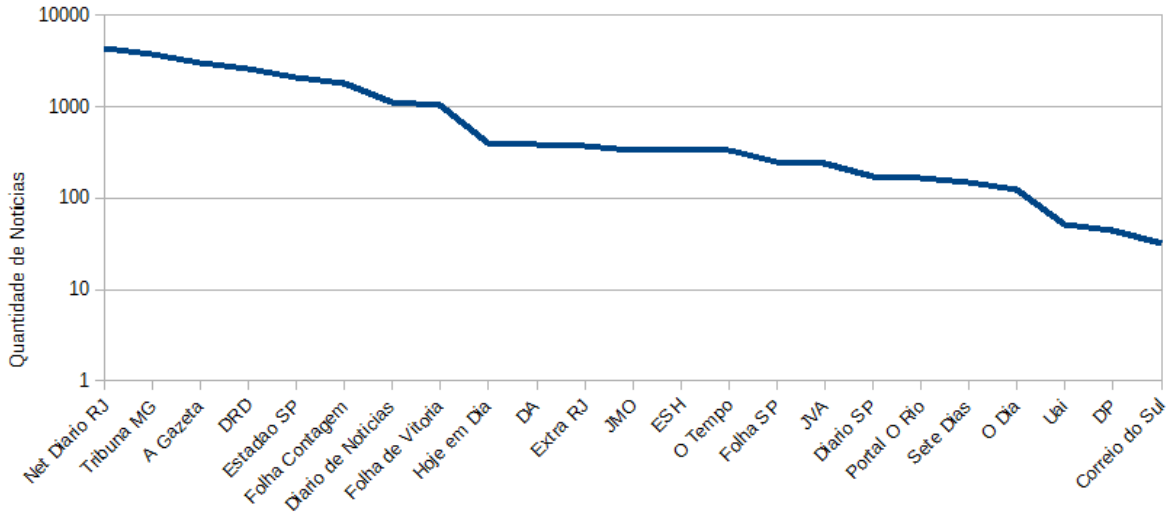
Fonte: Reproduzida pela autora

4 Análises a partir do banco de dados

Após a finalização do dicionário de palavras foram executadas diversas buscas para análises e reconhecimento de todo o conteúdo coletado.

Primeiro apuramos a quantidade de palavras, na figura 2 é possível verificar a variação de palavras entre os veículos, podemos notar que existem jornais com grandes diferenças de tamanhos, mas isto pode ser pelo fato do veiculo tem uma maior quantidade de notícias ou por serem redações muito grandes.

Figura 2: Quantidade de palavras por veículo

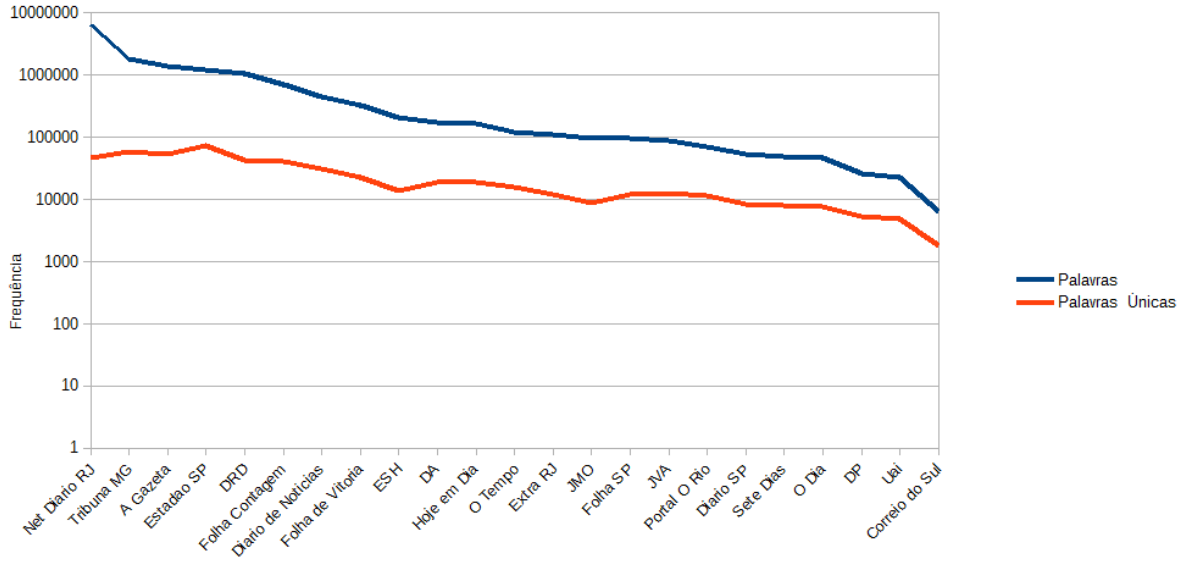


Fonte: Resultado de pesquisa

Após a criação do banco de dados, utilizamos duas maneiras de se referir aos tokens, palavras e palavras únicas. Como o próprio nome já diz, palavras, que correspondem a cada token individualmente e palavras únicas que representa o token uma única vez mesmo que ele se repita posteriormente, por exemplo, “Ele concerta peça por peça”, nesta frase temos 5 palavras e 4 palavras únicas. Das 22.959 notícias, temos no total 14.714.665 palavras e 164.830 palavras únicas, podemos observar essas frequências por veículos na figura 3.

A alta frequência de palavras do jornal Net Diário RJ é o resultado de um número muito grande de notícias que são constituídas de vários textos jornalísticos de assuntos semelhantes, aumentando consideravelmente a quantidade de palavras deste veículo. Mesmo o número de palavras sendo bem alta, a quantidade de palavras únicas é bem inferior. Esse fenômeno é ocasionado pela similaridades dos diversos textos, logo as palavras se repetem ocasionando um menor número de palavras únicas.

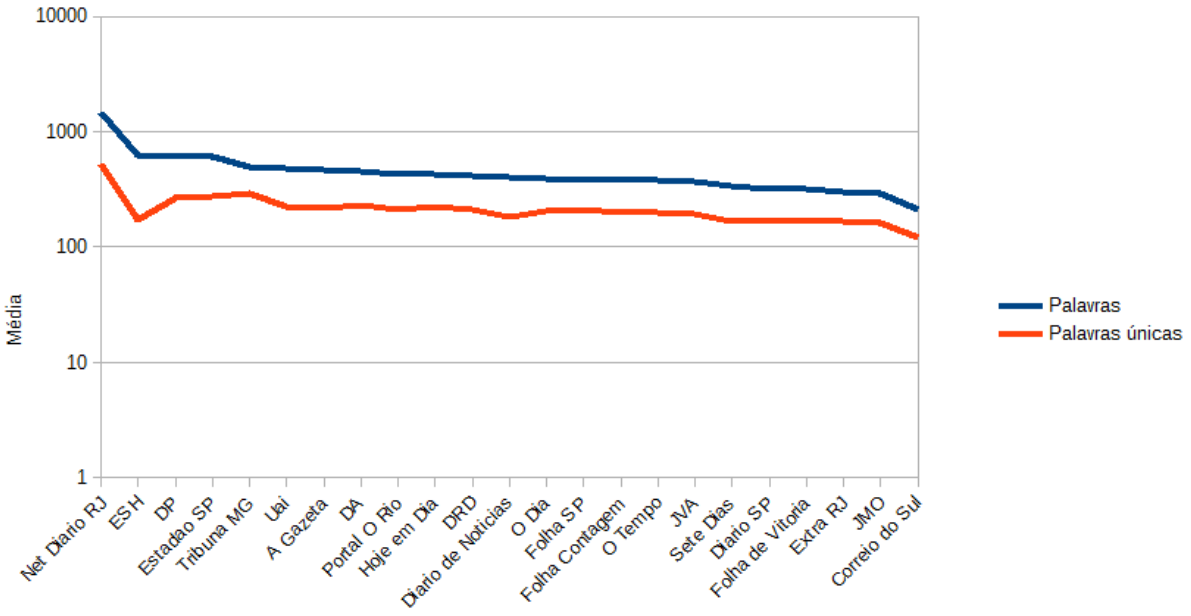
Figura 3: Frequência de palavras e palavras únicas por veículo



Fonte: Resultado de pesquisa

Na figura 4 é apresentada a média de palavras e palavras únicas por veículo. A média de palavras está entre 212 e 1484. Se desconsiderarmos o jornal Net Diário RJ devido ao relato anterior, o intervalo cai para 212 a 631 palavras. Logo pode-se dizer que é possível um jornalista redigir uma matéria para o jornal dentro desta média de palavras. Por outro lado para um leitor ser capaz de compreender uma notícia, ele precisa conhecer aproximadamente entre 121 e 531 palavras diferentes da língua portuguesa.

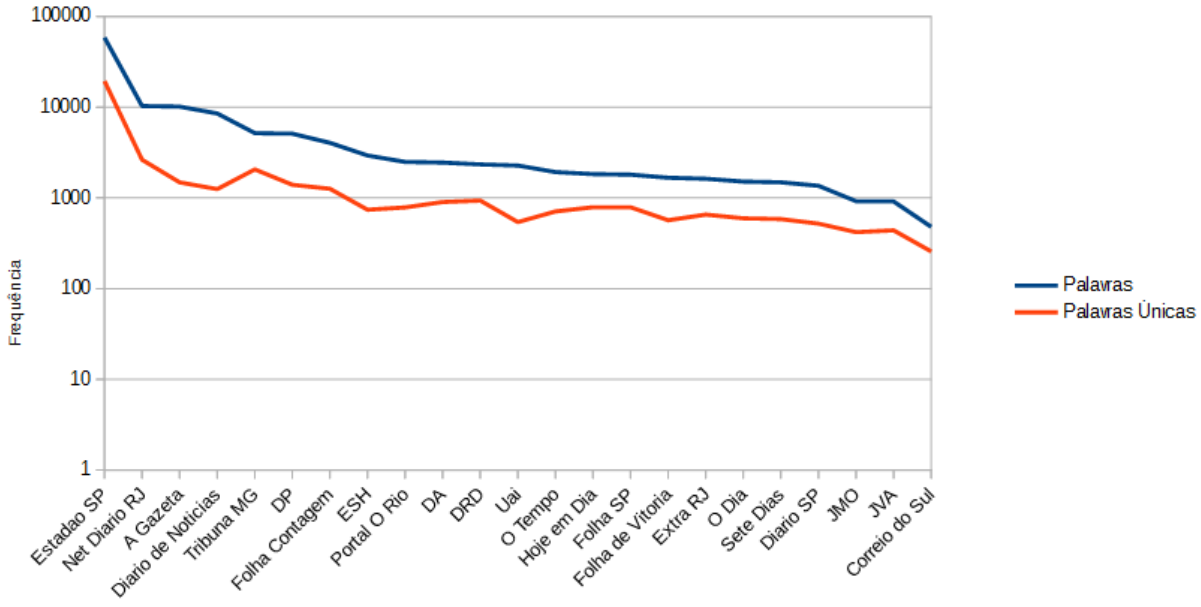
Figura 4: Média de palavras e palavras únicas por veículo



Fonte: Resultado de pesquisa

Quanto ao tamanho das notícias, nem sempre quando a quantidade de palavras é grande a de palavras únicas também será, isso é um fator que depende muito do conteúdo do texto. Na figura 5 mostra a quantidade máxima de palavras e de palavras únicas da maior notícia de cada jornal.

Figura 5: Notícia com maior número de palavra e palavras únicas por veículo



Fonte: Resultado de pesquisa

O Estadão SP é o veículo que tem a maior notícia. Esse alto número se dá pelo fato de ser uma lista de aprovados do vestibular da Fuvest, com a presença de grande quantidade de nomes, sobrenomes e o respectivo número de inscrição dos aprovados acarretando em uma elevada coleção de 58.645 palavras.

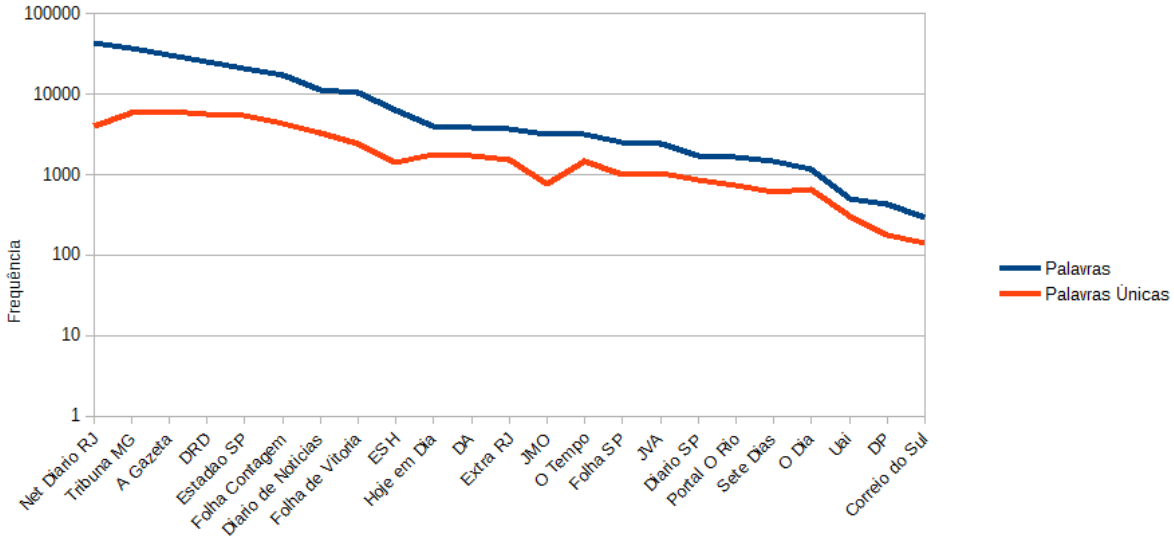
O Net Diário RJ é o veículo que tem o maior números de notícias coletadas e tem as maiores notícias em quantidade de palavras, no entanto, a quantidade de palavras únicas é bem menor. A figura 5 mostra que o intervalo entre as linhas dos outros veículos possuem uma diferença menor, a causa dessa diferença pode ser pela existência de um conjunto de notícias de um mesmo assunto, sendo assim, resultando em uma extensa reportagem contendo um número muito grande de palavras e pelo fato dos textos serem de mesmo assunto a queda de palavras únicas pode ter se dado por esse motivo.

A Gazeta tem a terceira maior notícia. Seus arquivos que possuem os maiores números de palavras são editais de casamentos e notícias com a mesma natureza daquelas do veículo Net diario RJ, tendo o mesmo efeito sobre as frequências de palavras. As notícias com o maiores números de palavras do jornal A Gazeta são conteúdos sobre vagas de empregos e concursos que também são matérias longas e com pequena variedade de palavras.

Considerando que as notícias coletadas foram escritas seguindo o modelo de pirâmide invertida, selecionamos as 10, 20, 50 e 100 primeiras palavras de cada notícia para analisar o comportamento entre as figuras.

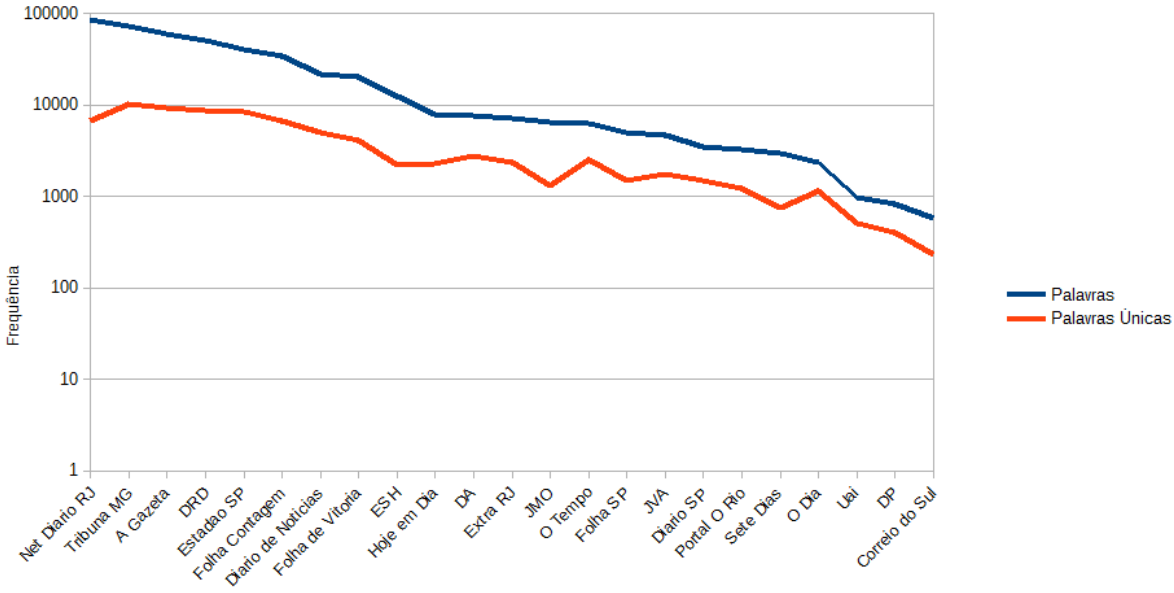
Na figura 6, o intervalo entre as linhas dos jornais O Dia e Uai estão bem próximas, logo a variedade de palavras entre as 10 primeiras posições são bem grandes. Esse mesmo comportamento pode ser observado na figura 7 com relação as 20 primeiras posições.

Figura 6: As 10 primeiras posições de cada notícia por veículo



Fonte: Resultado de pesquisa

Figura 7: As 20 primeiras posições de cada notícia por veículo



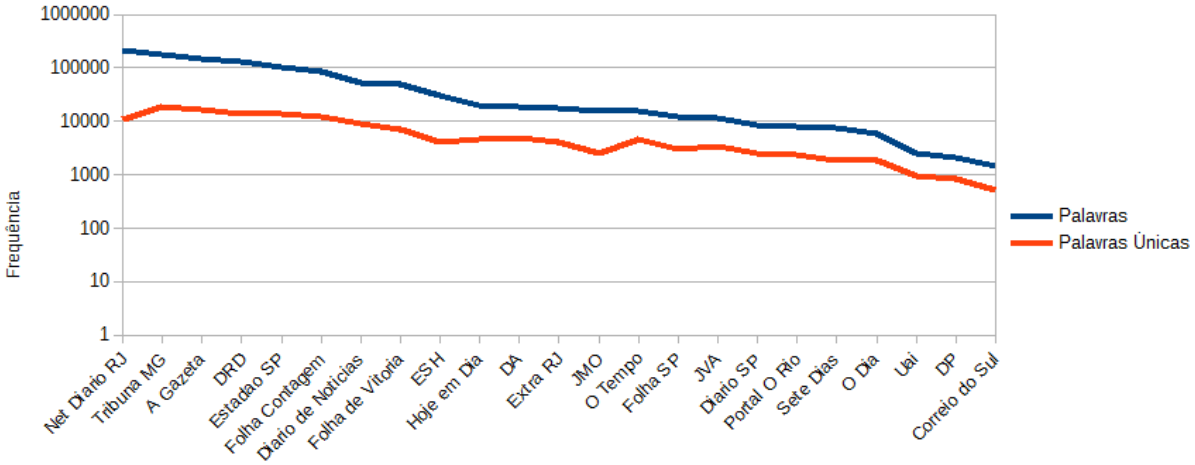
Fonte: Resultado de pesquisa

Se compararmos as figuras 6 e 7, elas tem o comportamento bem parecidos, principalmente a linha azul referente às palavras, algumas mudanças foram os realces dos picos pertencentes aos veículos O Tempo e O Dia.

Já a figura 8, que reúne as 50 primeiras palavras, apresenta uma grande diferença

sobre os outros, suas linhas são mais suaves, além dos intervalos entre as linhas serem mais uniformes e menores, não tirando os pequenos destaques do pico do jornal O Tempo e os leves vales dos jornais Net Diário RJ, ESH, JMO e Uai.

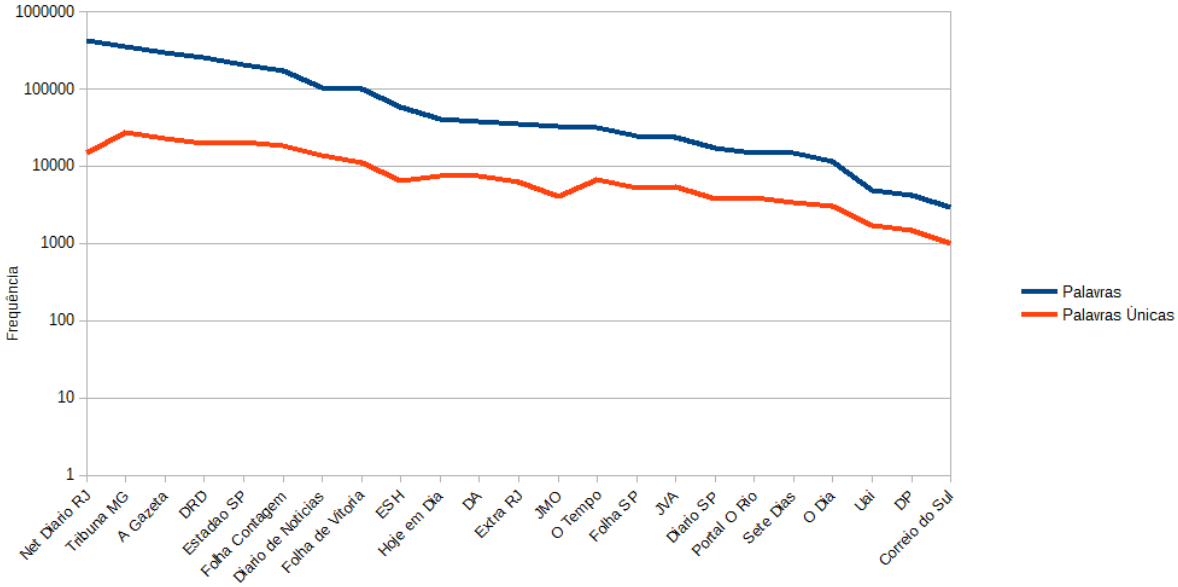
Figura 8: As 50 primeiras posições de cada notícia por veículo



Fonte: Resultado de pesquisa

A figura 9 representa as 100 primeiras palavras que também têm as linhas mais suaves, mas em contrapartida o intervalo entre elas são um pouco maiores, se comparado a imagem anterior.

Figura 9: As 100 primeiras posições de cada notícia por veículo



Fonte: Resultado de pesquisa

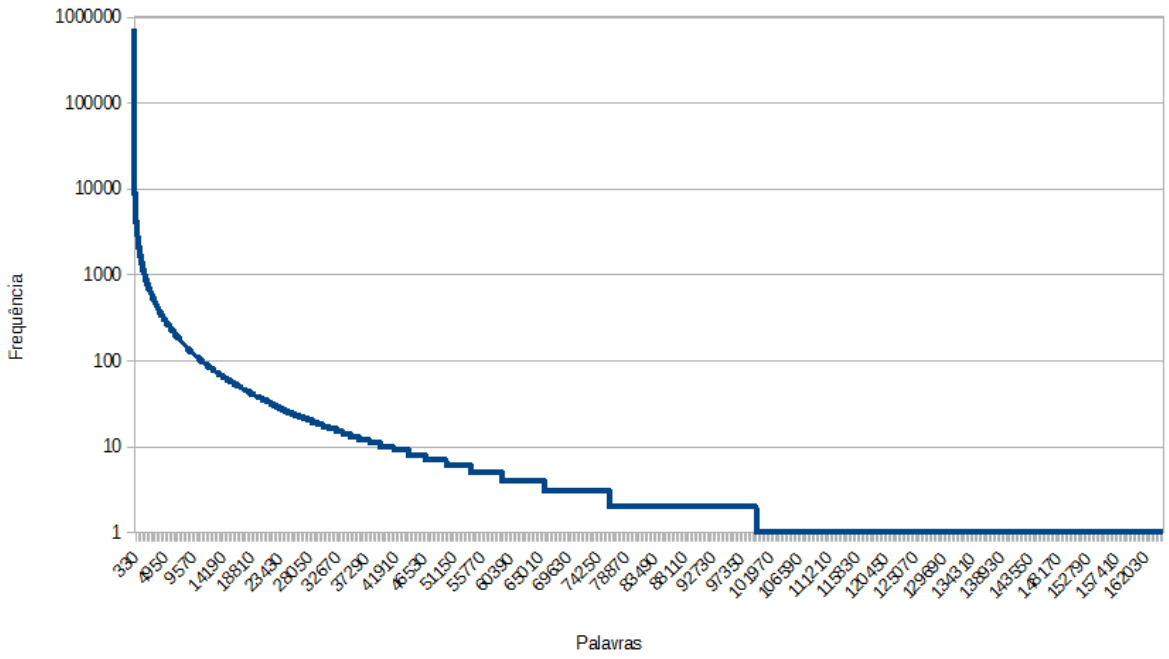
Com uma pesquisa mais aprofundada poderíamos avaliar se entre 50 e 100 primeiras palavras é possível escrever um texto jornalístico com precisão. Além da possibilidade de verificar se com as primeiras linhas de uma notícia da web é suficiente para a criação de gazetteers,

pois se o padrão seguido pelo redator for de pirâmide invertida logo nos primeiros parágrafos poderá ser identificado fortes indicadores de localização, que segundo (CANAVILHAS, 2006), os dados mais importantes ficam na parte superior da redação, que são informações que respondem às seguintes perguntas: o quê, quem, onde, como, quando e por quê. Note que a localização, onde, é uma das primeiras perguntas a serem tratadas, logo se for comprovado que é possível, o custo e esforço para a criação e manutenção de gazetteers serão menores.

Agora analisando em um campo mais amplo, ou seja, vamos explorar os números em relação às palavras existentes em nosso banco de dados.

Pela linha do gráfico da imagem 10 podemos observar que mais da metade das palavras representam uma frequência muito baixa, totalizando 123.191 palavras que se repetem menos de 10 vezes, são aproximadamente 74,74%, mais do que a metade do nosso dicionário de palavras. Se for considerado os dados das 10, 20, 50 e 100 primeiras palavras de cada notícia e gerando os gráficos nesta mesma configuração obtemos semelhanças no seu comportamento em quase sua totalidade.

Figura 10: Frequência das palavras



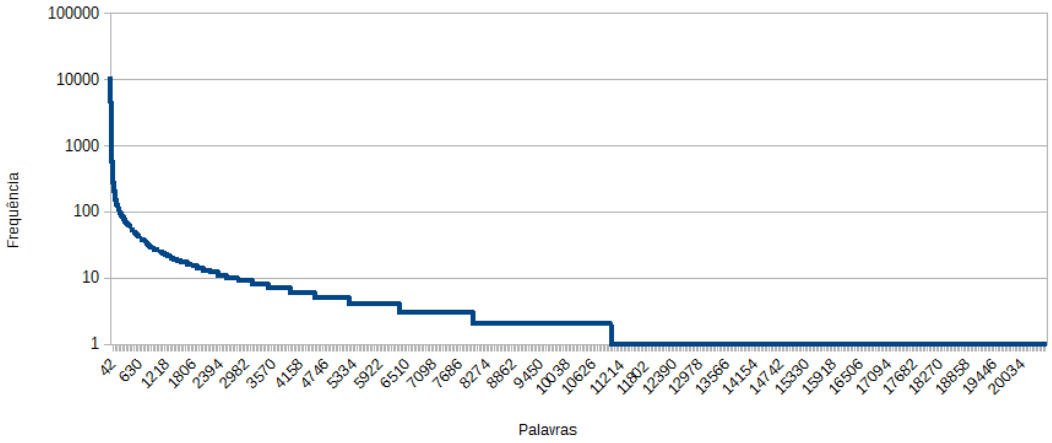
Fonte: Resultado de pesquisa

Na figura 11, aproximadamente 86,33% das palavras se repetem menos de 10 vezes.

Na figura 12, com as 20 primeiras palavras, aproximadamente 83,64% das delas ocorrem no máximo 9 vezes.

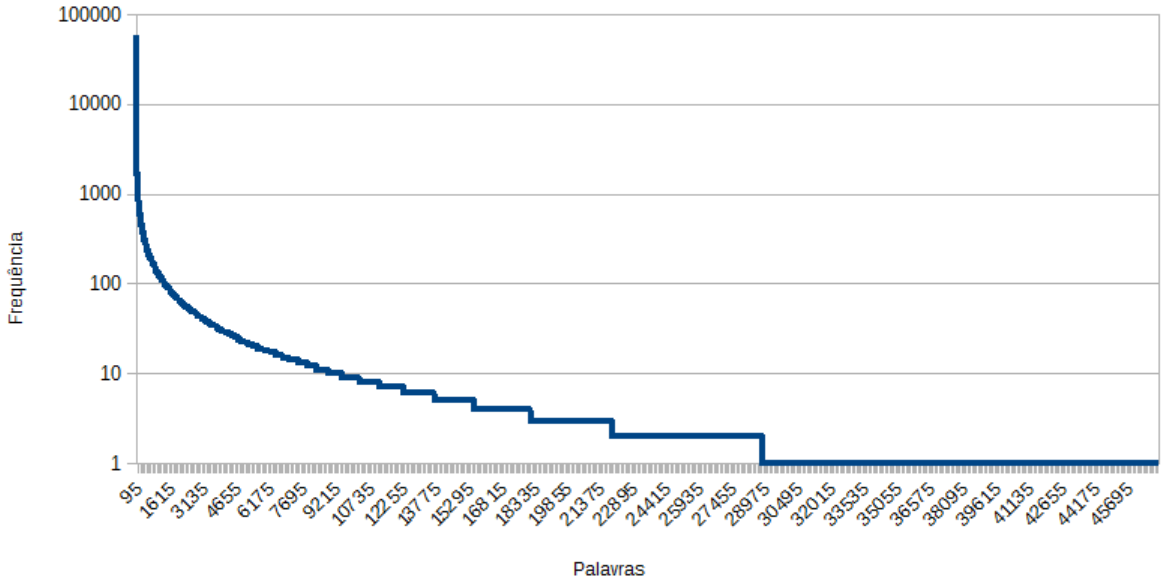
Já com as 50 primeiras palavras, aproximadamente 79,78% delas ocorrem menos de

Figura 11: Frequência das 10 primeiras palavras de cada notícia



Fonte: Resultado de pesquisa

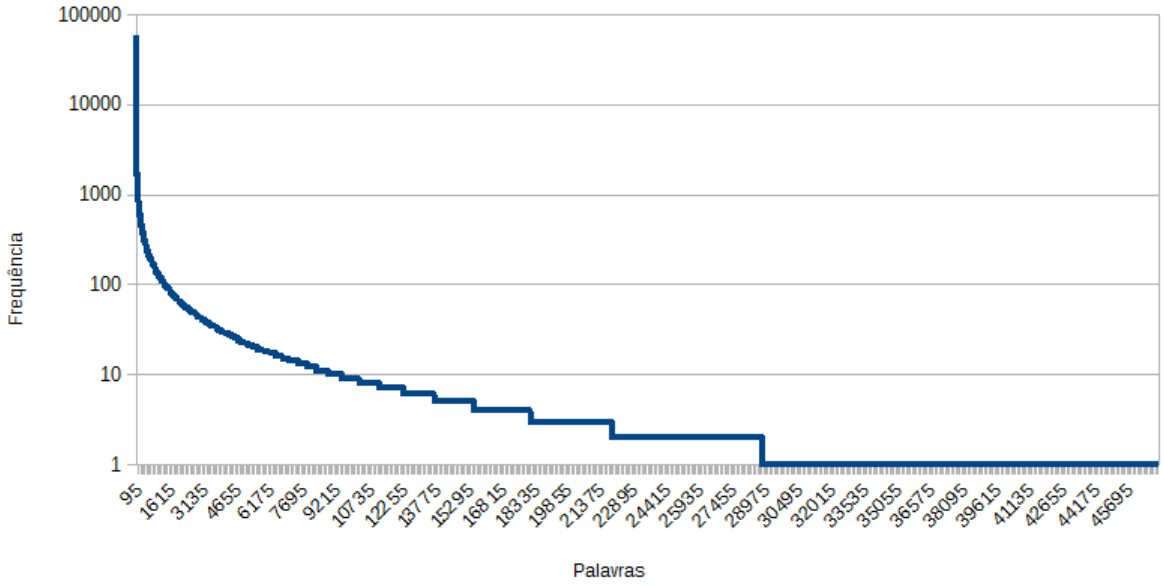
Figura 12: Frequência das 20 primeiras palavras de cada notícia



Fonte: Resultado de pesquisa

10 vezes, conforma na figura 13.

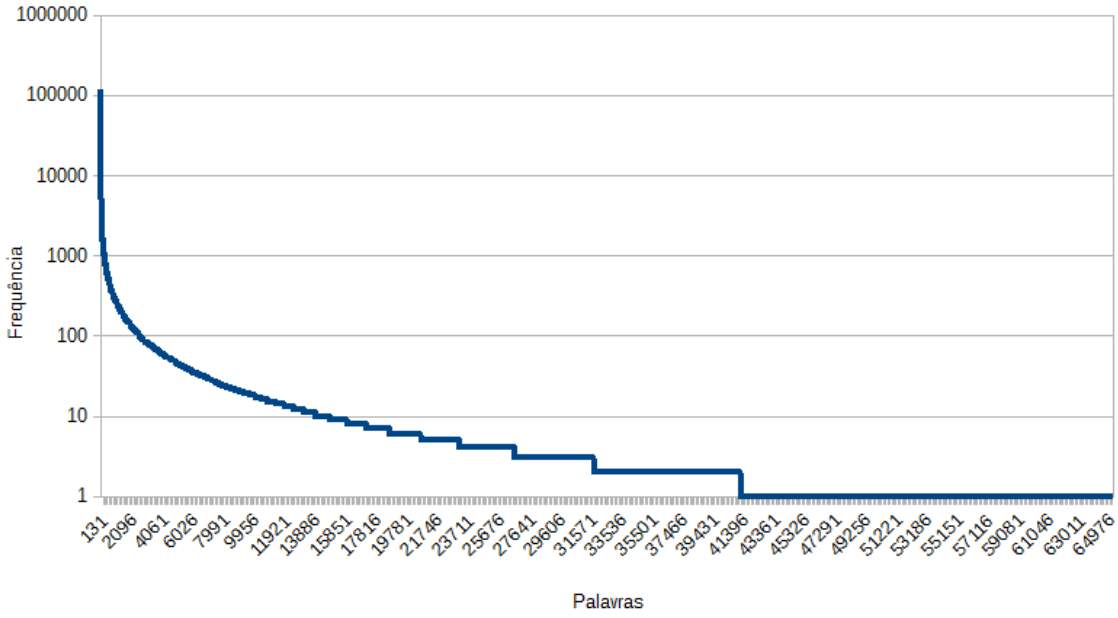
Figura 13: Frequência das 50 primeiras palavras de cada notícia



Fonte: Resultado de pesquisa

Com as 100 primeiras palavras aproximadamente 77,33% delas aparecem menos de 10 vezes conforme a figura 14.

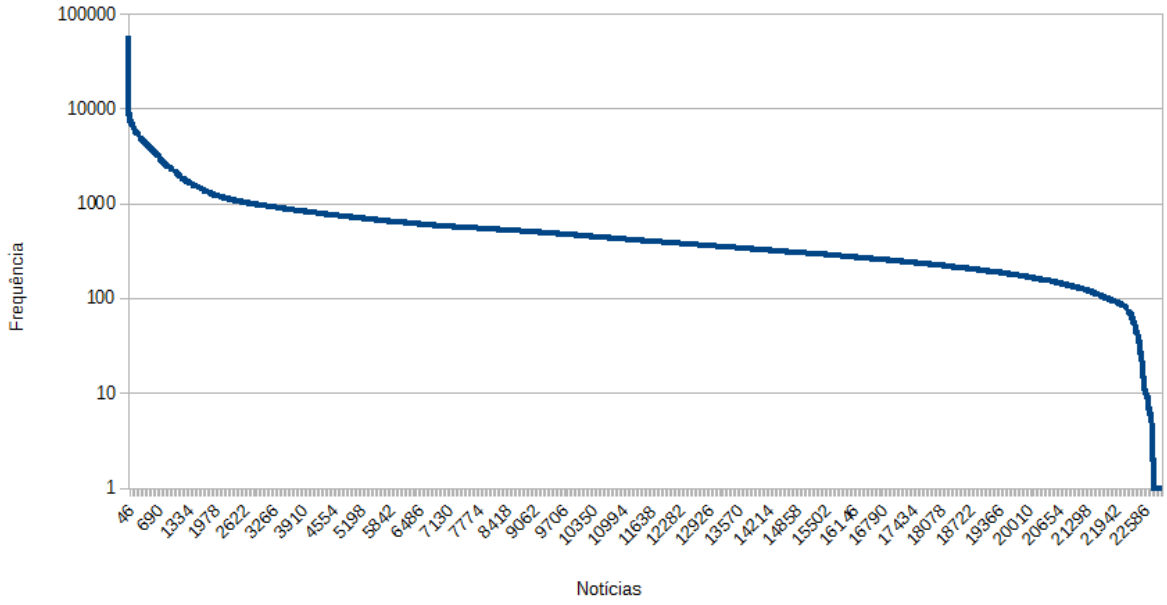
Figura 14: Frequência das 100 primeiras palavras de cada notícia



Fonte: Resultado de pesquisa

Pela figura 15 podemos observar que a grande maioria das notícias possuem em média de 100 a 1000 palavras.

Figura 15: Média das palavras por notícias



Fonte: Resultado de pesquisa

4.1 Análise dos termos para possível identificação de indicadores de localidade e termos de desambiguação

Com o surgimento da Internet não podemos afirmar que os web jornais sigam ou não o padrão da pirâmide invertida, há uma discussão quanto a isso. Segundo (MÓDOLO, 2007), no lead clássico que geralmente corresponde ao primeiro parágrafo, contém as respostas das seguintes perguntas sobre o assunto da notícia: quem, o que, quando, onde, como e por quê. Se não há espaço para responder todas essas perguntas o lead se estende a um novo parágrafo, o sublead, respondendo as perguntas: como e por quê. Dessa forma no segundo nível da pirâmide é apresentado os dados secundário e respectivamente o final, mas o interessante para esta pesquisa são as respostas das perguntas onde e quem, a primeira nos leva provavelmente com exatidão a localidade e a segunda pode servir como parâmetro de desambiguação, desta forma apenas a lead já é suficiente para extrair os possíveis indicadores de localidades.

De acordo com (CANAVILHAS, 2006), nos jornais impressos o espaço para redigir a notícia é limitado e faz com que os jornalistas tenham que seguir a técnica da pirâmide invertida, assim é possível escrever de forma que se for preciso diminuir o tamanho da redação o editor pode subtrair um dos últimos parágrafos sem comprometer a notícia. Já nas notícias na web os cortes por motivos espaciais não serão um problema e o jornalista pode oferecer mais recursos para o leitor, como links de outros textos e até mesmo elementos multimédia.

A figura 16 é uma notícia do jornal Net Diário, pode-se notar que neste texto a maioria dos nomes de localidades estão concentrados no topo, os retângulos em vermelho evidenciam os nomes de localidades em vários níveis, como região Serrana, cidade Teresópolis e bairro do Alto, como também há outros nomes das proximidades da principal cidade da notícia que é

Teresópolis.

Figura 16: Notícia do jornal Net Diário

Ministro do Turismo visita Teresópolis

Postado em 26 agosto 2014. Tags: cervejaria, Polo Cervejeiro, Região Serrana, St. Gallen, Turismo

– Vinicius Lages conheceu atrativos do Polo Cervejeiro da Região Serrana

O ministro do Turismo Vinicius Lages esteve em Teresópolis no último sábado, 23, para visitar as instalações da Vila St. Gallen, localizada no bairro do Alto. Lages, que estava acompanhado do secretário de estado de Turismo do Rio, Claudio Magnavita, foi recebido pelo vice-prefeito Marcio Catão e pelo secretário municipal de Turismo, Ronaldo Fialho. Na ocasião, o ministro também passou por Nova Friburgo e Petrópolis. O objetivo foi conhecer o Polo Cervejeiro da Região Serrana a fim de que o Ministério apoie o projeto 'Rota Cervejeira do Rio de Janeiro – Cervejas das Montanhas', que está em fase de elaboração.

Ministro Vinicius Lages conversa com vice-prefeito Marcio Catão e com secretário de Turismo, Ronaldo Fialho, sobre o circuito cervejeiro e gastronômico da Região Serrana (Foto: Jeferson Hermida)

"Recebi o convite da secretaria de estado de Turismo para conhecer as instalações das cervejarias que integram o Polo Cervejeiro da região. Estou impressionado com a Vila St. Gallen, um equipamento turístico muito bonito e bem estruturado", avaliou o ministro Lages, que estava acompanhado do diretor do Departamento de Financiamento e Promoção de Investimento no Turismo, Eduardo Golin.

A Vila St. Gallen é um empreendimento da Cervejaria Therezópolis destinado à diversão e à experiência dos amantes da cerveja com os múltiplos estilos da bebida e as diversas formas de harmonizá-la. O espaço temático completa três anos de existência em outubro deste ano e possui diversos ambientes, como o Bistrô 1912, onde o cliente conhece o que há de mais sofisticado na relação gastronomia e cervejas especiais; e a Vila St. Gallen, na qual o visitante conhece a história, os ingredientes, os principais estilos e o processo de produção da bebida.

O vice-prefeito Marcio Catão ressaltou que a iniciativa poderá contribuir para o desenvolvimento dos municípios integrantes do projeto. "O ministro veio para chancelar uma ideia que está sendo discutida há mais de um ano, que é a Rota Cervejeira. Esse circuito poderá se tornar um produto espetacular não só para o turismo da região e de Teresópolis em especial, mas também para o desenvolvimento econômico dessas cidades", pontuou Catão. O secretário municipal de Trabalho, Emprego e Economia Solidária, Lucas Bonifácio, e o subsecretário de Estudos e Projetos, Carlos Tucunduva, também participaram da visita.

Fonte: Reproduzida pela autora

Supondo que todas as notícias dos veículos coletados seguem o modelo da pirâmide invertida, foram criadas várias tabelas divididas em duas classificações, as de letras minusculas e maiúsculas e outra que são apenas palavras que apresentam a primeira letra maiúscula:

- Top_2000_das_100_primeiras: formada pelas 2 mil palavras mais frequentes das 100 primeiras palavras de cada notícia de cada veículo.
- Top_2000_das_100_ultimas: criada a partir das 2 mil palavras mais recorrentes das 100 últimas palavras de cada notícia. Neste processo foi encontrado algumas dificuldades que será mencionado no capítulo 6 com as soluções encontradas para contornar os problemas.
- Top_2000_geral: de forma geral foram selecionadas as 2 mil palavras mais frequentes considerando todas as palavras de todas as notícias.

- Top_2000_das_100_primeiras_maiusculas: foram selecionadas as 2 mil palavras com a letra inicial maiúscula das 100 primeiras palavras de cada notícia.
- Top_2000_das_ultimas_maiusculas: das 100 últimas palavras de cada noticia foram selecionadas apenas as com letra inicial maiúscula, extraindo as 2 mil com maiores frequências.
- Top_2000_maiuscula_geral: formada pelas 2 mil palavras com a primeira letra maiúscula mais recorrentes de todas as notícias, considerando todas posições do texto.

O motivo por considerar apenas as palavras com a inicial maiúscula é pela grande possibilidade de indicar o maior número de nomes de lugares, rua, avenida, bairro, cidade, estado, pessoas importantes e tantas outras que podem ser indicadores de localidades e até mesmo termos que podem servir para distinguir o significado de uma palavra (desambiguação), assim diminuindo o número de tokens que se tornaria lixo.

Não devemos esquecer que palavras escritas com a primeira letra maiúscula não são apenas os nomes próprios, mas também indicam início de período ou citação. Existem várias outras situações que este recurso é utilizado. No entanto, não vamos considerar casos específicos, pois nossos textos jornalísticos já foram tokenizados, ou seja, não dá mais para identificar ponto final e início de parágrafos/trechos, sendo assim foram selecionadas todas as palavras com a letra inicial maiúscula, no total foram 79.908 palavras, sendo um número grande para analisar uma a uma e validar quais são possíveis indicadores de localidade e termos de desambiguação, dessa forma foi limitado à um número de 2000 palavras por tabela.

Com as tabelas criadas foram analisadas palavra por palavra e retirada aquelas que não interessavam, grande parte dos substantivos e os nomes próprios foram considerados. Para as palavras que geraram dúvidas quanto ser um indicativo de localidade ou não, foi pesquisado no site dos Correios [1] se existe alguma rua, bairro ou cidade que contenha a palavra em seu nome. Na tabela 2 é possível visualizar o resultado desta análise.

A classe A é representada por palavras no geral, ou seja, palavras que as iniciais podem ser tanto maiúsculas quanto minusculas, já a classe B são apenas palavras com a inicial maiúscula.

Tabela 2: Resultado da análise das tabelas criadas com as 2000 primeiras palavras

Classe	Tabela	Indicadores	Palavras	Percentual
A	Top_2000_das_100_primeiras	360	195591	18
	Top_2000_das_100_ultimas	317	150709	15,85
	Top_2000_geral	332	997584	16,1
B	Top_2000_das_100_primeiras_maiusculas	988	165468	49,4
	Top_2000_das_ultimas_maiusculas	972	149957	48,6
	Top_2000_maisucula_geral	976	839832	48,8

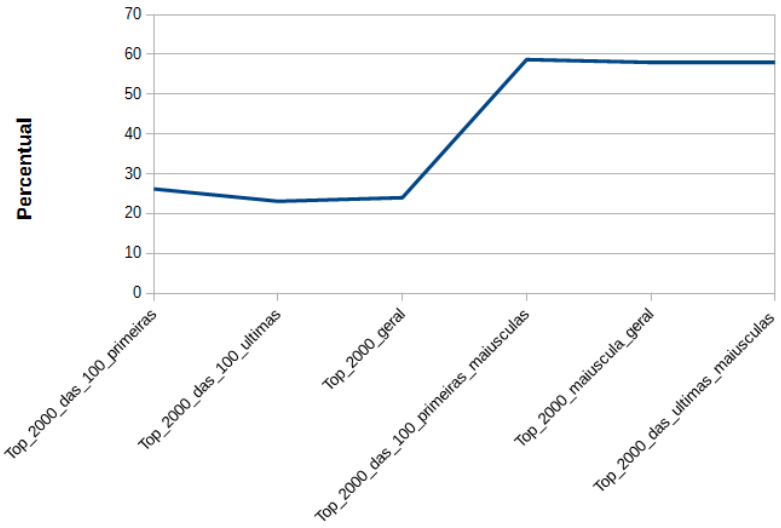
Fonte: Elaborada pela autora.

[1] http://www.buscacep.correios.com.br/sistemas/buscacep/

De acordo com a tabela 2 podemos observar que nas duas classes os piores resultados foram apresentados pelas tabelas que consideraram apenas as últimas palavras. O que torna válida a ideia de que as notícias seguem o modelo da pirâmide invertida, visto que os melhores resultados são os formados pelas 100 primeiras palavras em cada uma das classes.

Na tabela 2 e na imagem 17 é possível analisar de maneira mais clara que a tabela Top_2000_geral tem um número maior de palavras se comparada com as demais. Entretanto, seu percentual de indicadores é bem mais baixo em relação à tabela Top_2000_maiusculas_geral, sendo este um resultado esperado, visto que as palavras com a primeira letra maiúsculas garantem um maior número de nomes próprios, contribuindo para o aumento dos indicadores de localidades.

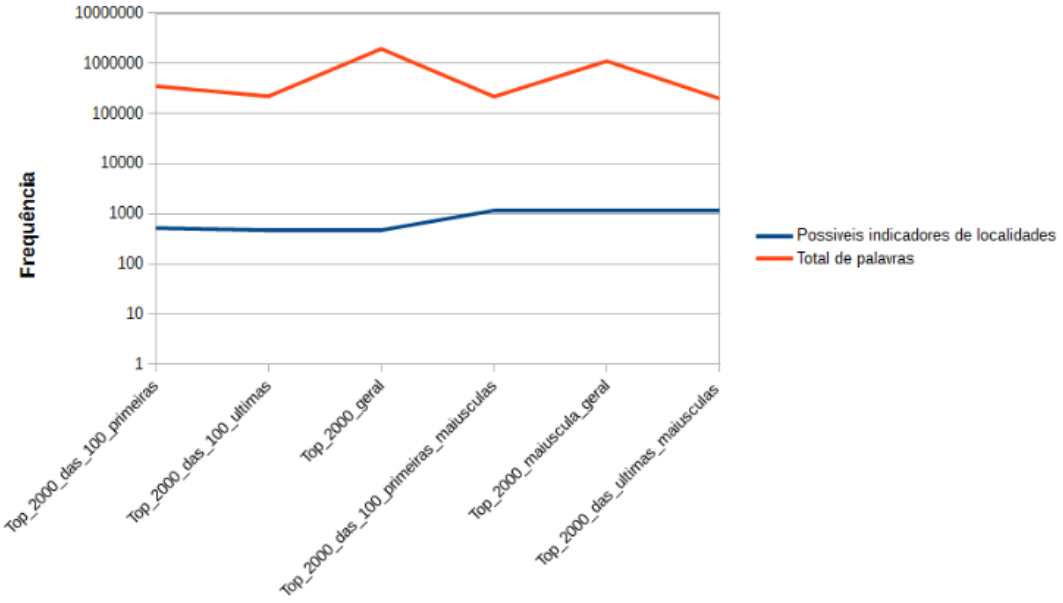
Figura 17: Resultados obtidos das tabelas criadas



Fonte: Resultado de pesquisa

Também é nítido que a tabela Top_2000_das_100_primeiras_maiusculas tem um total de palavras mais baixo e ao mesmo tempo um número de possíveis indicadores bem elevado se comparados às outras tabelas, ou seja, a quantidade de palavras é menor mas possuem mais variedade de indicadores de localidades. Outros testes podem ser feitos aumentando as posições de 100 para 200 ou 300 para verificar se os resultados podem ser melhores e o quanto melhores.

Figura 18: Resultados obtidos das tabelas criadas



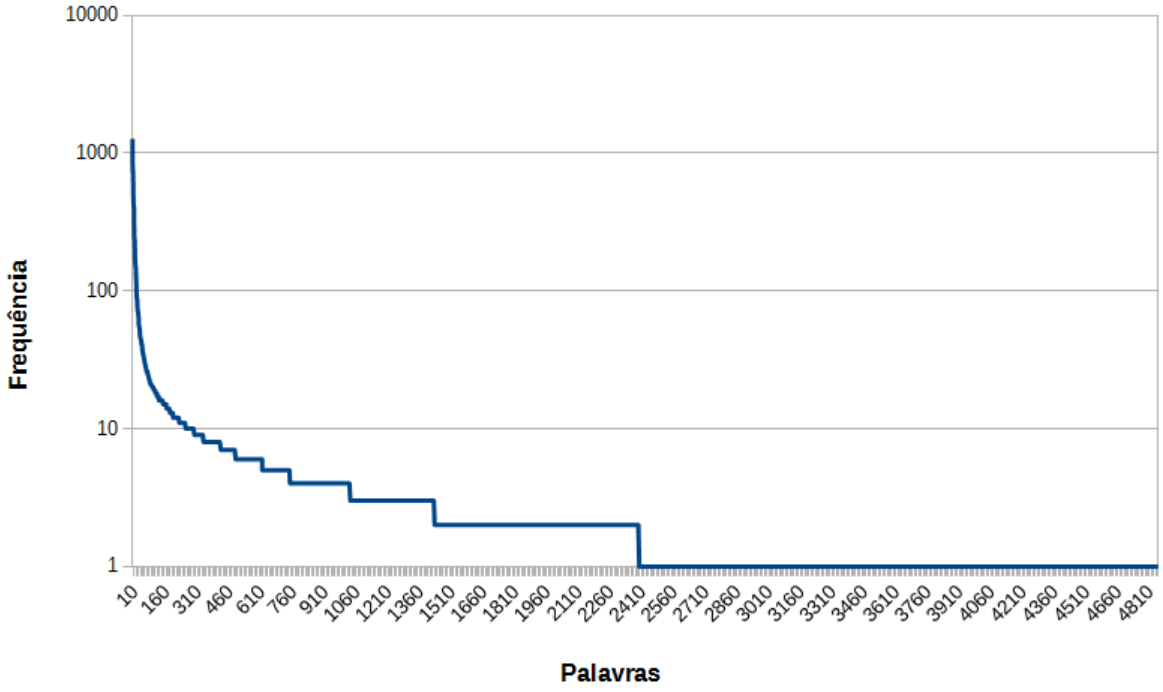
Fonte: Resultado de pesquisa

5 Exploração das notícias do jornal Uai

Foi escolhido como amostra o jornal Uai, tendo em vista o conhecimento da área geográfica que o veículo cobre e por ter uma menor coleção de notícias. No jornal em questão existem 22.640 palavras e 4.853 palavras únicas.

Na figura 19 podemos observar que o comportamento é o mesmo da figura 10 que se refere à frequência das palavras, a maioria das palavras se repetem menos que 10 vezes, o que difere os gráficos são as quantidades de palavras, no jornal Uai são 12.558 contra 123.191 palavras de toda coleção de jornais.

Figura 19: Frequência de palavras do jornal Uai

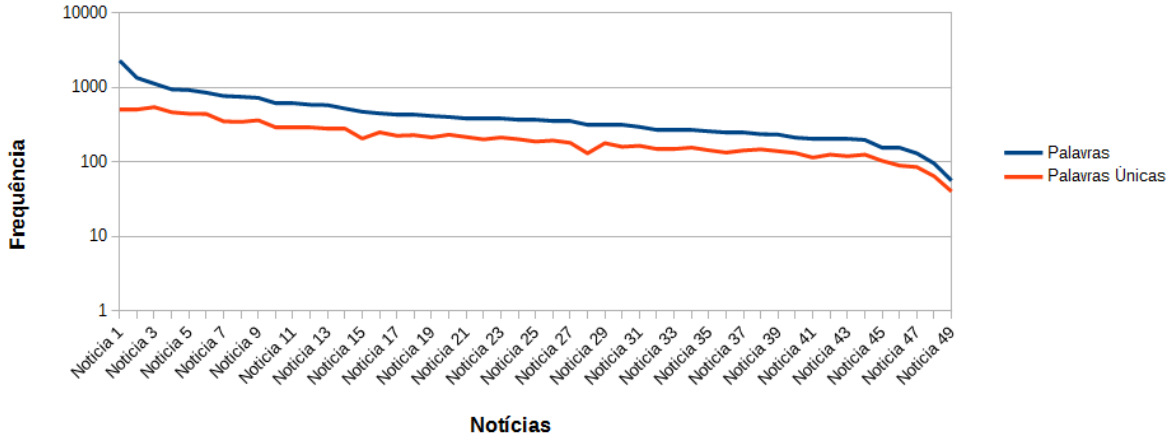


Fonte: Resultado de pesquisa

A figura 20 e 3 tem uma pequena semelhança, mesmo a figura 20 representando todas as notícias do jornal Uai e a figura 3 sendo separado por veículo, podemos notar que existe uma aproximação entre as linhas, elas começam opostas uma da outra, ou seja, uma possui um grande número de palavras e outra uma menor frequência de palavras únicas e terminam bem mais próximas, caindo o número de palavras e palavras únicas.

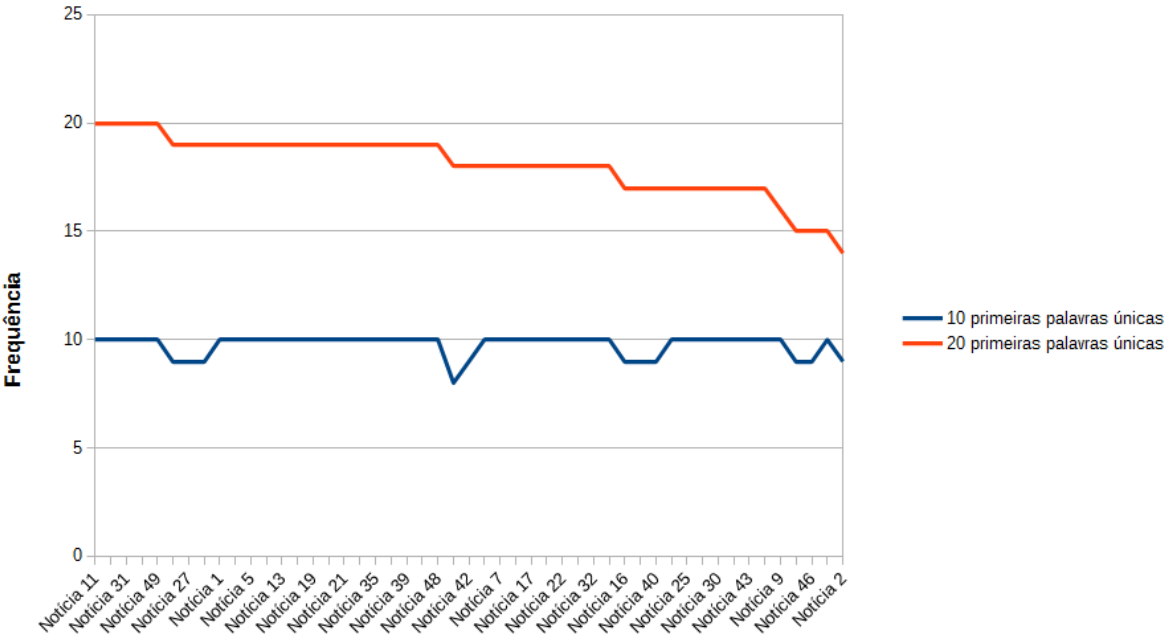
Na figura 21 foi contabilizado o número de palavras e palavras únicas das 10 e 20 primeiras palavras de cada notícia. Nas 10 primeiras posições, são poucas notícias em que as palavras se repetem, mais precisamente 11 notícias que pelo menos uma palavra se repete, a grande maioria possui um conjunto variado de palavras. Já nas 20 primeiras posições é o contrário, apenas 5 das notícias são formadas por 20 palavras diferentes. Pode-se observar

Figura 20: Frequência de palavras e palavras únicas do jornal Uai



Fonte: Resultado de pesquisa

Figura 21: Frequência de palavras e palavras únicas das 10 e 20 primeiras posições das notícias do jornal Uai



Fonte: Resultado de pesquisa

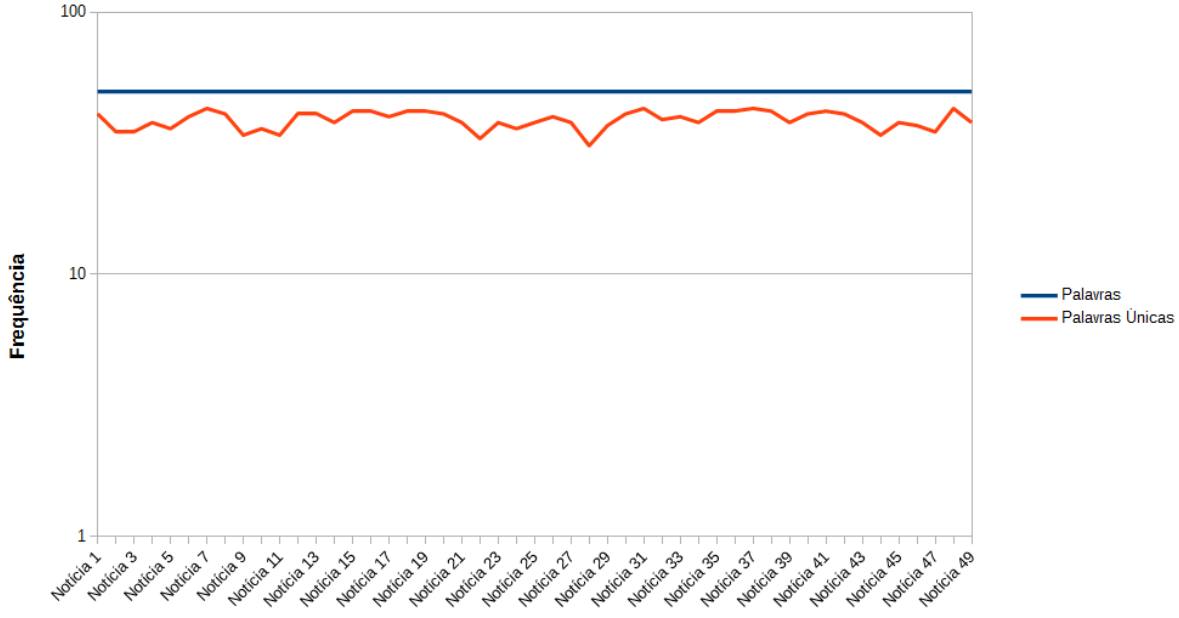
que frases a partir de 8 palavras começam a se repetir.

Na figura 22 são contabilizadas as 50 primeiras palavras de cada notícia, podemos observar que a quantidade de palavras únicas na maior parte são bem próximas da quantidade de palavras. Se compararmos com a figura 8, que contempla todos os veículos, esta proximidade das linhas é algo em comum para as 50 primeiras palavras.

Na figura 23 são contempladas com as 100 primeiras posições das notícias, podemos observar que nem todas as notícias são escritas com até 100 palavras. Se comparado com

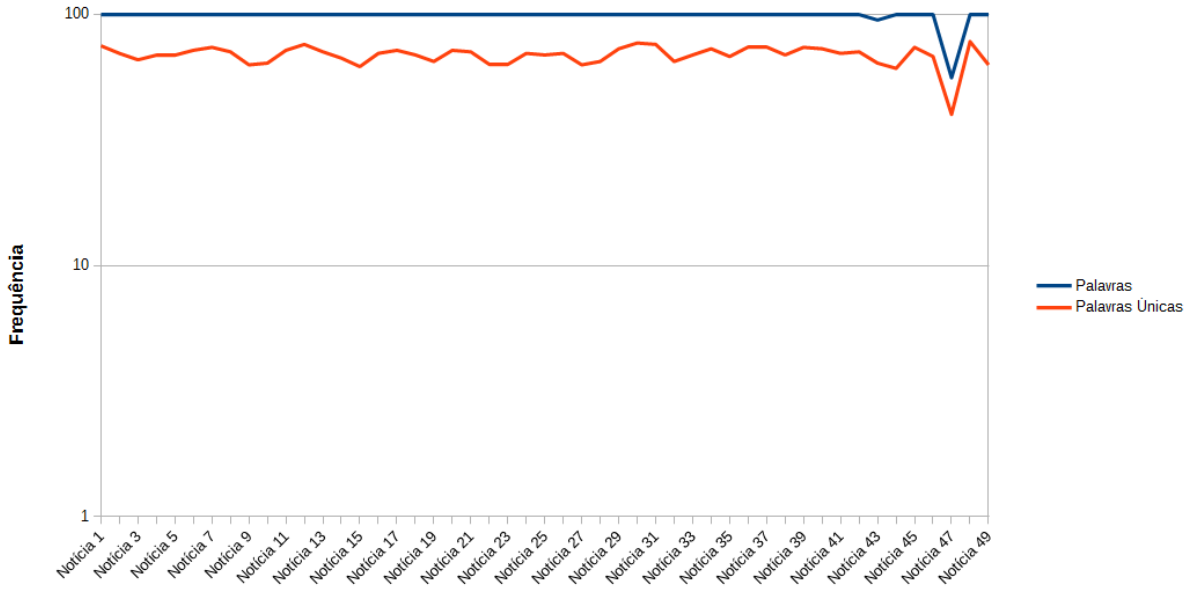
o gráfico anterior a distância entre as linhas são maiores, caso que também ocorreu com os gráfico 8 e 9, que foram representados com as 50 e 100 primeiras palavras respectivamente, considerando todos os veículos.

Figura 22: Frequência de palavras e palavras únicas das 50 primeiras posições das notícias do jornal Uai



Fonte: Resultado de pesquisa

Figura 23: Frequência de palavras e palavras únicas das 100 primeiras posições das notícias do jornal Uai



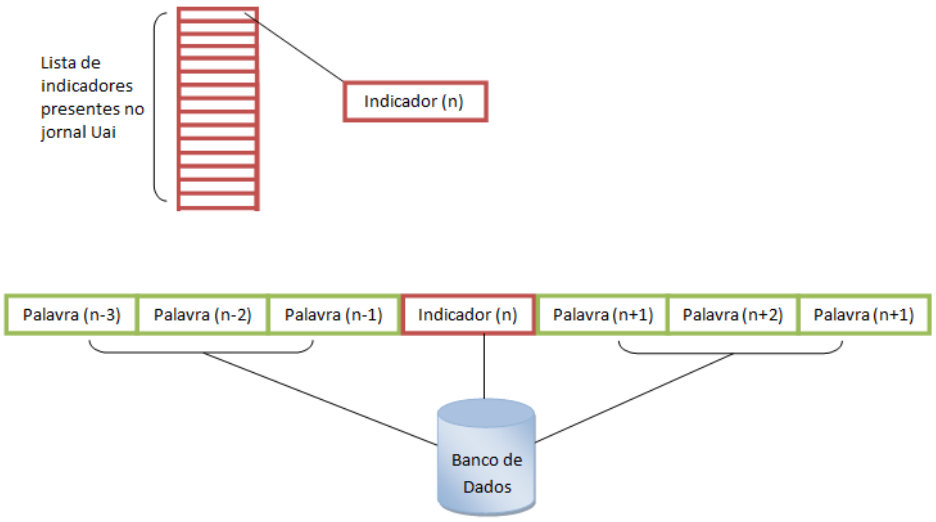
Fonte: Resultado de pesquisa

5.1 Análise dos termos do jornal Uai com os resultados identificados de localidade e termos de desambiguação

Com base no resultado das tabelas anteriormente analisadas, foi destacado o melhor resultado, Top_2000_das_100_primeiras_maiusculas, sendo adotada as mesmas característica para o objeto de estudo do presente tópico, onde passaremos a analisar as palavras próximas de cada indicador de localidade da tabela e se, somente se, estiver presente no jornal em destaque.

Anteriormente, das 2 mil palavras selecionadas no início do processo, rendeu um saldo de 988 palavras que são possíveis indicadores de localidades e das 988 palavras 260 estão presentes nas notícias do jornal Uai. Se dividirmos pela coleção do jornal Uai são 4 indicadores para cada notícia. Para análise foi criado uma nova tabela, uai_palavras_vizinhas, selecionando as três palavras que antecedem e sucedem, como mostra o esquema na figura 24.

Figura 24: Esquema de seleção das palavras vizinhas



Fonte: Produzido pela autora.

Abaixo os atributos da tabela criada e na figura 25 um exemplo da tabela:

- Palavra_vizinha: são as palavras antecessores e sucessores de cada indicador de localidade.
- Posicao: número referente a posição da palavra na notícia.
- Noticia: nome da notícia em que a palavra se encontra.
- Palavra_origem: o possível indicador de localidade que originou do processo anterior.

- Posicao_origem: posição da palavra de origem na notícia, ou seja posição do possível indicador de localidade.

Figura 25: Exemplo de comparação de indicadores e as palavras vizinhas

	A	B	C	D	E
1	palavra_vizinhas	posição	noticia	palavra_origem	posicao_origem
912	Avenida	7	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
913	BH	4	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
914	de	1	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
915	de	3	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
916	na	6	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
917	protestam	5	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
918	saúde	2	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	BH	4
919	Afonso	8	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
920	Agentes	10	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
921	Avenida	7	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
922	BH	4	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
923	na	6	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
924	Pena	9	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
925	protestam	5	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Avenida	7
926	Afonso	8	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
927	Agentes	10	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
928	Avenida	7	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
929	comunitários	11	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
930	na	6	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
931	Pena	9	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
932	protestam	5	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Afonso	8
933	Afonso	8	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9
934	Agentes	10	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9
935	Avenida	7	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9
936	comunitários	11	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9
937	de	12	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9
938	na	6	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9
939	Pena	9	agentes-de-saude-de-bh-protestam-na-avenida-afonso-pena.html	Pena	9

uai_3_palavras_vizinhas | Plan1 | Plan2 | Plan3 | Plan4 | Plan5 | Plan6 | Plan7 | Plan8 | Plan9 | Plan10 | Plan11 | ...

Fonte: Produzido pela autora.

Com a tabela populada, foi analisada palavra por palavra, desta forma foram excluídas palavras que não tinham relevância, permanecendo apenas aquelas que podem ser indicativos de localidades. Como forma de classificação foi criado uma tabela com 10 tipos de categorias para os indicadores de localidade sendo:

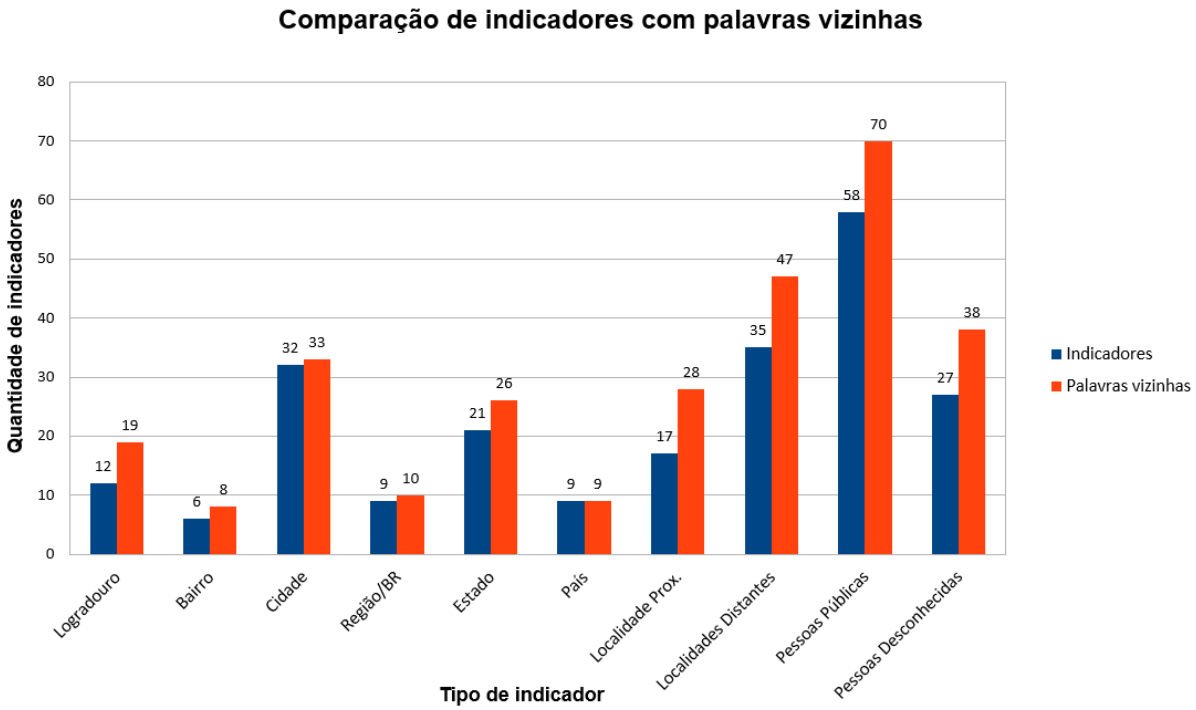
- Logradouro: nome de ruas e avenidas
- Bairro: nome de bairros
- Cidade: nome de cidades e municípios
- Região\BR: mesorregiões, rodovias e regiões como por exemplo Triângulo Mineiro
- Estado: nomes de estados
- País: nomes de países
- Localidades Próximas: são locais próximos de onde ocorreu o fato da notícia
- Localidades Distantes: são nomes de locais distantes da ocorrência dos fatos

- Pessoas Públicas: nomes de pessoas conhecidas nacional e internacionalmente, como por exemplo nome de políticos, estes nomes podem servir para ajudar a descobrir de que local a notícia se refere.
- Pessoas Desconhecidas: nomes de pessoas incomuns, que são pouco conhecidas.

Foi feita uma leitura de cada notícia para reconhecer cada palavra e classificá-la da forma correta. Algumas notícias apresentaram alguns nomes em mais de uma forma, como Belo Horizonte e BH, assim como Minas Gerais sendo também representada de forma abreviada, MG ou apenas Minas. Neste caso foi contabilizada o nome apenas uma vez por se tratar da mesma localidade. Outro aspecto notado neste processo foi que nomes de ruas, avenidas e bairros quase sempre vem com o indicativo na frente, isso é perceptível ao comparar com as palavras vizinhas e até mesmo nomes de localidades compostas, uma vez que na tokenização elas foram desmembradas.

Pode-se observar na figura 26 que a maioria dos indicadores são nomes de pessoas públicas sendo a grande maioria nomes de políticos, precisamente são 58 nomes de pessoas para indicadores e 70 nomes para palavras vizinhas, logo vai existir jornais com mais de um nome de pessoa pública. O indicador do tipo país tem a mesma ocorrência nas duas circunstâncias, mas neste caso de estudo se trata de jornais da região Sudeste e pela alta granularidade não se torna de extrema importância, mas já impediria o processo de analisar as notícias que não ocorreram no Brasil.

Figura 26: Comparação de indicadores com palavras vizinhas



Fonte: Produzido pela autora.

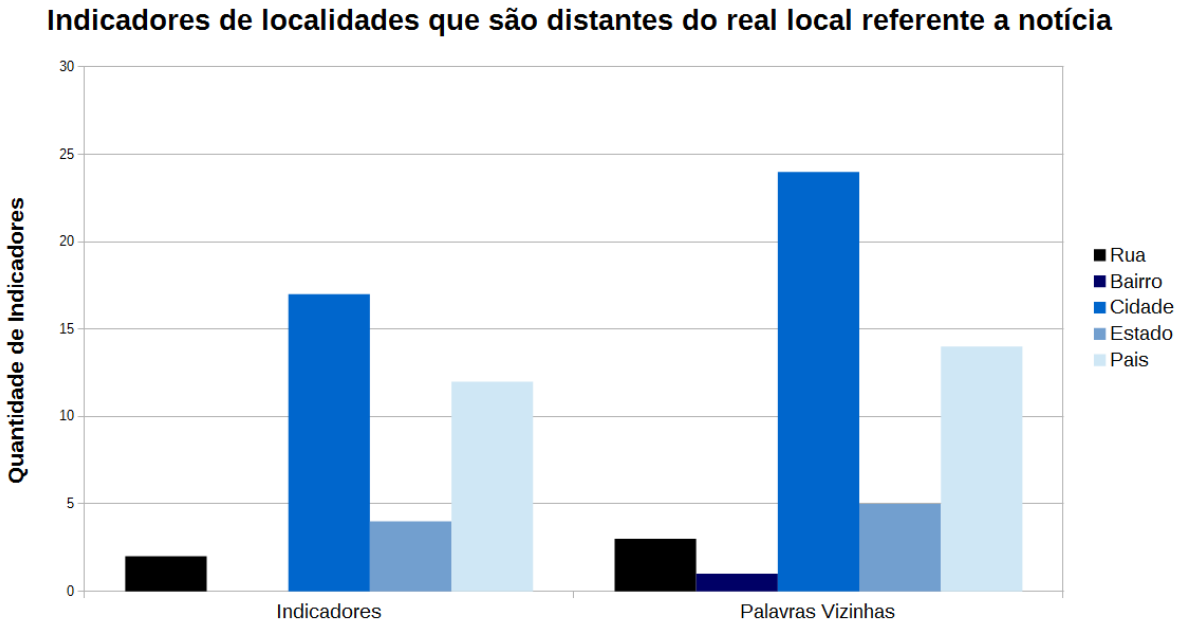
Nomes compostos foram tratados como apenas um termo. Como por exemplo, na lista

de indicadores existem as palavras Afonso e Pena, já no momento de análise foi considerado como logradouro por se trata do nome de uma avenida de Belo Horizonte.

Outro tipo de indicador que também chama a atenção é de localidades distantes, existe uma quantidade grande de nomes de localidades que são afastados do real local que aconteceu o fato da notícia, se tornando irrelevante para o índice de indicadores, mas que por outro lado pode ser confundido se não existir conhecimento de todo o noticiário e todos os indicadores de localidade, tornando um falso indicador.

Na figura 27 pode ser analisado as classificações destas localidades, a grande parte são de nomes de cidades e países, esses nomes se não forem bem analisados podem ser confundidos e levar a desvirtualização do real espaço geográfico.

Figura 27: Análise dos indicadores de cidades distantes do real local da notícia



Fonte: Produzido pela autora.

6 Dificuldades encontradas para criar a tabela com as 100 últimas palavras de cada notícia

Para buscar as últimas 100 palavras de cada notícia deparamos com certa dificuldade, pois o número de palavras no banco de dados é razoavelmente grande, excedendo o tempo de execução do script, foram várias tentativas até chegar a solução, que deveria ser feito a seleção por veículo.

A principal lógica para selecionar as últimas posições, foi calculando o intervalo entre a última e as 99 palavras anteriores, o ponto chave foi identificar a última posição utilizando a função MAX() e depois subtrair 100 da posição máxima (última posição) e assim descobrir o intervalo de cada uma das notícias para reunir todas elas e construir um ranking das palavras mais recorrentes da parte inferior da redação.

Utilizar apenas um select para fazer toda essa manobra tem um alto custo temporal, na tentativa de executar o script foram mais de 24 horas que ainda não foram suficientes para concluir a seleção. A alternativa encontrada foi executar o script por veículo, assim diminuiria o tempo de execução já que as comparações feitas entre as tuplas seriam reduzidas consideravelmente, mas para isso foi criada uma tabela auxiliar com as colunas:

- ultima_pos: última posição da palavra da notícia
- marcador: marcação da posição que deve-se começar a selecionar a palavra até a última posição
- noticia: nome da notícia que as posições pertencem
- veiculo: nome do veículo que as posições pertencem

Como existem pequenas notícias com menos de 100 palavras, o cálculo do intervalo foi atribuído um valor negativo, para que não ocorressem divergência de valores, pois no banco de dados não existe posição negativa, substituímos todos os valores negativos por zero. Para agilizar o processo e ganhar tempo, após o veículo processado ele foi excluído da tabela para diminuir o seu tamanho e consequentemente o tempo de execução do script.

Abaixo segue o script como exemplo para a seleção das 100 ultimas posições das noticias de um jornal específico.

create table net_diario_rj as SELECT p.‘nomepalavra‘, p.‘posicao‘, p.‘noticia‘, p.‘jornal‘ FROM ‘copia_palavra‘p, ‘aux_posicao‘a WHERE p.‘noticia‘ = a.‘noticia‘ and a.‘jornal‘ = ’net diario rj’ and p.‘jornal‘ = a.‘jornal‘ and p.‘posicao‘ <= a.‘maxi‘ and p.‘posicao‘ >= a.‘min‘

7 Conclusão

Com o trabalho podemos concluir que os indicadores de localidade se encontram nos primeiros parágrafos das notícias, ou seja no lead, o que confirma que jornais eletrônicos ainda seguem o formato da piramide invertida. O interessante de se trabalhar apenas com o lead é que as informações de localidade se encontram logo nos primeiros parágrafos reduzindo assim o tamanho do texto e consequente um menor custo de processamento. Considerar palavras com a primeira letra maiúscula abrange um maior número de possíveis indicadores. Ao analisar as palavras vizinhas elas podem facilitar no georreferenciamento, pois podem ajudar na identificação do tipo de indicador e na identificação de nomes compostos das localidades, antes do nome de logradouros vem escrito se é o nome de uma avenida ou rua por exemplo, sendo assim é um auxílio na identificação e classificação do indicador de localidade e até mesmo de identificar nomes compostos.

7.1 Resultados

Com o trabalho desenvolvido e o estudo dos jornais podemos concluir que pelas médias de palavras por notícias que foram encontrados, uma notícia pode ser escrita com 212 a 631 palavras. Se considerar apenas o lead esta média possivelmente pode cair, uma vez que o tamanho do texto irá diminuir.

Os indicadores de localidade estão mais concentrados no primeiro parágrafo, ou seja, no lead, onde ficam as principais informações sobre a notícia, desta forma ao considerar apenas as 100 primeiras palavras do texto, teve um maior número de indicadores de localidades, mas o benefício maior foi com a consideração das palavras com a primeira letra maiúscula (não podemos considerar sendo apenas substantivos próprios, pois no processo de tokenização foram retirados os pontos, perdendo a identificação de início de orações). Desta forma podemos concluir que os webjornais ainda são escritos com a estrutura de pirâmide invertida e é possível encontrar a localidade no lead do jornal.

Com a análise feita do jornal Uai constatamos que a maior parte de indicadores de localidade encontrados são nomes de cidades, países e estados e ao considerar as palavras vizinhas dos possíveis indicadores mais recorrentes tem um melhor aproveitamento, pois ao lado de um indicador pode existir outros indicadores e até mesmo um complemento do nome de alguma localidade, se tratando de nomes compostos, que ao serem tokenizados tenham sidos separados.

Nem sempre ter um conjunto muito grande de palavras remete uma grande quantidade de possíveis indicadores de localidade, a consideração das 2 mil palavras mais recorrente obteve o maior montante de termos, em compensação um dos piores índices de indicadores de localidades, isso de dar pelo fato da repetição muito grande de palavras que são desconsideradas no processo subsequente, sendo assim, selecionar de forma geral as palavras mais

repetidas acaba incluindo muitos termos sem valores.

Determinados tipos de notícias não são boas bases para análises, como foi o caso do jornal Estadão SP, que possui um texto com nomes dos aprovados do vestibular da Fuvest, desta forma ocasionando um vício para o dicionário de palavras, pois com a grande variedade de nomes e sobrenomes de pessoas comuns não serviriam como palavras de desambiguação para indicadores de localidades. Outro ponto, levando em consideração apenas as palavras com a letra inicial maiúscula este tipo de texto consideraria todas as palavras, podendo causar um desvio de localidades, pois existem muitos lugares com nomes de pessoas, podendo gerar uma confusão ao selecionar os possíveis indicadores de localidade que na verdade não se referem a nomes de lugares mas sim de cidadãos comuns.

7.2 Considerações e trabalhos futuros

Durante o trabalho alguns pontos foram percebidos e que possivelmente poderiam ter ajudado a encontrar melhores resultados. Levando em consideração os resultados encontrados, as palavras com a primeira letra maiúscula são as que cobriram um maior número de possíveis indicadores de localidade, se ao tokenizar as notícias não tivessem extraídos os pontos, seria fácil de identificar se a palavra com a letra inicial maiúscula é de fato um nome próprio ou apenas o inicio de frase/paragrafo. Se não tivesse desconsiderado os pontos a lista selecionada teria um maior número de nomes próprios o que levaria um maior aproveitamento das palavras como indicadores de localidade e consequentemente reduziria as palavras comuns que no contexto não representaram nada.

O trabalho ficou limitado à veículos de língua portuguesa de alguns estados brasileiros. O interessante seria fazer o estudo de veículos de outros países, de idiomas diferentes e até mesmo de veículos que sejam de língua portuguesa mas que sejam de outros países. Desta forma poderia ser analisado o comportamento e diferenças entre os resultados que aqui foram encontrados.

O presente trabalho considerou todos as categorias de notícias, sendo assim, existem algumas notícias que não são de relevância, como por exemplo, textos que sejam de receitas culinárias, a não ser que seja de algum prato típico de alguma região, caso contrário se torna dispensável pois não agrega valor por não apresentar algum indicador de localidade. Logo se for consideradas apenas categorias que apresentem indicadores de localidade pode trazer resultado diferentes.

Uma abertura deste trabalho é a criação de um gazetteer à partir das notícias de jornais, pois assim seria um dicionário mais completo, incluindo componentes de localização desde nomes de ruas a nomes de países e termos para desambiguação.

Referências

BERTOCCHI, D. Considerações sobre o uso de tags em narrativas jornalísticas digitais. *7o. SBPJor–Encontro Nacional de Pesquisadores em Jornalismo*, São Paulo, v. 40, p. 19, 2009. Citado na página 14.

CANAVILHAS, J. Webjornalismo: Da pirâmide invertida à pirâmide deitada. *BOCC–Biblioteca Online de Ciências de Comunicação*, [S.l.], 2006. Citado nas páginas 27 e 30.

DETERS, J. I. *Método de Ordenação de Documentos na Web Baseado no Tempo de Permanência. Florianópolis, 2003. 88 f.* 2003. Tese (Doutorado) — Dissertação (Mestrado em Ciências da Computação)–Universidade Federal de Santa Catarina, Florianópolis, Florianópolis, SC, 2003. Citado na página 10.

GALDO, A.; VIEIRA, A. F. G.; RODRIGUES, R. S. Classificação social da informação na web: tecnologia, informação e gente. *Datagramazero, Rio de Janeiro*, v. 6, n. 9, p. 25–34, 2009. Citado na página 10.

GOUVÊA, C. *Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web*. 2009. 88 p. Dissertação (Mestrado em Ciência da Computação) — Universidade Católica de Pelotas, Pelotas, RS, Brasil, 2009. Citado na página 16.

GOUVÊA, C.; LOH, S. Folksonomias: identificação de padrões na seleção de tags para descrever conteúdos. *Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.21529/RESI*, [S.l.], v. 6, n. 2, 2007. Citado nas páginas 10 e 14.

GOUVÊA, C.; LOH, S.; GARCIA, L. F. F. Métodos para seleção automática de tags para descrição de notícias na web. In: ACM, 14., 2008, Vila Velha, Espírito Santo, Brasil. *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. New York, NY, USA, 2008. p. 81–84. Citado nas páginas 14, 15 e 16.

GOUVÊA, C. et al. Discovering location indicators of toponyms from news to improve gazetteer-based geo-referencing. In: INPE, 10., 2008, Rio de Janeiro. *GeoInfo*. [S.l.], 2008. p. 51–62. Citado nas páginas 16, 17 e 18.

HTTRACK. *HTTrack Website Copier*. 2015. Disponível em: <https://www.httrack.com>. Citado na página 19.

JUNIOR, W. T. L.; BRANCO, C. F. C.; BARBOSA, P. Sistemas de recomendação de notícias nas mídias sociais buscam substituir o gatekeeping dos meios de comunicação de massa. *Comunicação & Inovação*, v. 10, n. 19, 2010. Citado na página 10.

LOH, S.; WIVES, L. K.; FRAINER, A. S. Recuperação semântica de documentos textuais na internet. In: *Conferência Latino-Americana de Informática (CLEY)*. Porto Alegre: [s.n.], 1999. p. 827–836. Citado na página 10.

MACHADO, I. M. R. *Um gazetteer ontológico para recuperação de informação geográfica*. 2011. 90 p. Tese (Mestre em Ciência da Computação) — Master's thesis, Departamento de Ciência da Computação da Universidade Federal de Minas Gerais, 2011. Citado nas páginas 16 e 17.

MÓDOLO, C. M. Infográficos: características, conceitos e princípios básicos. In: *XII Congresso Brasileiro de Ciências da Comunicação da Região Sudeste*. S.l], organization=Intercom, conference-number=12, conference-year=2007, conference-location=Juiz de Fora, MG, Brasil: [s.n.], 2007. Citado na página 30.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineraçao de textos. *Relatório Técnico–Instituto de Informática (UFG)*, Goias, 2007. Citado na página 10.

NASCIMENTO, G. F. C. d. L.; NEVES, D. A. d. B. Folksonomia como estratégia de indexação dos bibliotecários no del. icio. us. 2012. Citado na página 10.

REYNAR, J. C. Statistical models for topic segmentation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. [S.l.], 1999. p. 357–364. Citado na página 10.

RIVADENEIRA, A. W. et al. Getting our head in the clouds: toward evaluation studies of tagclouds. In: ACM, 2007, New York, NY, USA. *Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l], 2007. p. 995–998. Citado na página 13.

TORRES-PAREJO, U. et al. Mtcir: A multi-term tag cloud information retrieval system. *Expert Systems with Applications*, Elsevier, [S.l.], v. 40, n. 14, p. 5448–5455, 2013. Citado nas páginas 12, 13 e 14.

VARGAS, R. N. P. *Identificação da cobertura espacial de documentos usando mineração de textos*. 2012. Tese (Mestrado em Ciência da Computação e Matematica) — Universidade de São Paulo, 2012. Citado na página 11.

VASCONCELOS, K. A. Borges de. *Uso de uma ontologia de lugar urbano para reconhecimento e extração de evidências geo-espaciais na Web*. 2006. 181 p. Tese (Doutor em Ciência da Computação) — Universidade Federal de Minas Gerais, 2006. Citado na página 16.

VIEIRA, D. V.; CARVALHO, E. B. d.; LAZZARIN, F. A. Uma proposta de modelo baseado na web 2.0 para as bibliotecas das universidades federais. *Encontro Nacional de Pesquisa em Ciências da Informção*, São Paulo, n. 9, 2013. Citado na página 12.

ZIESEMER, A. d. C. A. et al. Recomendação de tags para mídia social colaborativa: da generalização à personalização. Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012. Citado na página 14.