

An Approach to Attribute Selection Using Zipf's Law and the TF-IDF Measure in the Patent Classification Process

Uma abordagem na seleção de atributos usando a Lei de Zipf e a Medida TF-IDF no processo de classificação de patentes

Carlos Gabriel S. Rodrigues, Claudia A. Martins

¹Instituto de Computação – Universidade Federal de Mato Grosso (UFMT) – Cuiabá – MT – Brasil

carlos.rodrigues@sou.ufmt.br, claudia@ic.ufmt.br

Abstract. *Natural language processing aids in understanding data through linguistic methods combined with machine learning and statistical techniques. In this work, algorithms related to word frequency are being investigated with the aim of analyzing the relevance of words to a dataset. Zipf's law combined with Luhn cuts and the TF-IDF measure are used in selecting the most relevant attributes for the classification process in the domain of patent data.*

Keywords: *Zipf's law, Luhn cuts, TF-IDF measure, natural language processing.*

Resumo. *O processamento de linguagem natural auxilia a compreensão dos dados por meio de uma métodos linguísticos aliados a técnicas e algoritmos de aprendizado de máquina e estatística. Neste trabalho estão sendo investigados algoritmos relacionados com a frequência de palavras, com o objetivo analisar a relevância das palavras para um conjunto de dados. A lei de Zipf combinadas com os cortes de Luhn e a medida TF-IDF são utilizadas na seleção dos atributos mais relevantes para o processo de classificação no domínio de dados de patentes.*

Palavras-chave: *Lei de Zipf, Cortes de Luhn, medida TF-IDF, processamento de linguagem natural.*

1. Introdução

O processamento de dados textuais não consiste apenas ao armazenamento, mas também no desenvolvimento de algoritmos direcionadas para ferramentas de busca, classificação, tradução entre outros, aplicadas em diversos domínios como redes sociais e tratamento de crimes digitais envolvendo, *bullying*, difamação e indicação de conteúdos impróprios, (Correa, 1999).

Patentes são documentos textuais com registros de propriedade temporária de uma invenção dando ao seu inventor o direito legal sobre ela¹. O processamento de dados em documentos de patentes consiste em garantir o desenvolvimento tecnológico e intelectual

¹ INPI (2022). Instituto Nacional da Propriedade Industrial. Disponível em: < <https://www.gov.br/inpi/pt-br> >. Acesso em: 20 jun. 2022.

de dezenas de pessoas. Esses repositórios de patentes se tornam uma fonte rica de conteúdo e inovação, sendo necessárias ferramentas computacionais para ajudar a manter e recuperar a informação armazenada e identificar possíveis erros ou fraudes relacionadas.

Este trabalho está inserido na investigação de técnicas para processamento e classificação automática de patentes. A classificação neste domínio é naturalmente complexa devido à natureza intrínseca dos dados, a alta dimensionalidade e a sobreposição das classes. Nesse contexto, este trabalho se propõe a investigar e analisar os atributos (palavras) do conjunto de dados usando técnicas baseadas na relação da lei de Zipf e da medida TF-IDF na redução da dimensionalidade dos dados. Isso consiste em identificar as palavras mais frequentes e investigar a relação lógica com a classe a que pertence. O objetivo é selecionar os atributos mais relevantes na discriminação de cada classe visando melhorar o desempenho dos classificadores e a criação de um dicionário. Trabalhos relacionados já utilizam a lei de Zipf como calibração de representação vetorial dos documentos (Cao et al., 2019) e a abordagem TF-IDF para melhorar a seleção de atributos (Jing et al., 2002).

2. Metodologia proposta

O objetivo desse trabalho foi utilizar uma combinação da lei de Zipf e a medida TF-IDF, junto com os cortes de Luhn, na identificação das palavras mais relevantes na discriminação de cada classe no domínio de patentes. Dessa forma, a metodologia utilizada segue os seguintes passos: i) conta-se a frequência das palavras; ii) gera-se a curva de Zipf com as frequências obtidas; iii) aplica-se os cortes de Luhn usando como parâmetro TF-IDF; iv) seleciona os atributos.

O conjunto de dados utilizado foi obtido da *World Intellectual Property Organization* (WIPO) e possui 75.239 patentes organizadas de forma hierárquica do sistema de classificação *International Patent Classification* (IPC) (WIPO, 2019) cuja organização se baseia numa hierarquia de níveis e subníveis que varia de A até H (Fall et al., 2003). Segundo Zipf (1949), existe uma relação entre a frequência e a posição em um processo de classificação das palavras de um conjunto de dados. Para Zipf, um texto segue o padrão na qual poucas palavras possuem uma alta frequência, palavras de frequência média aparecem significativamente e muitas palavras de frequência baixa.

A Lei de Zipf aplicada a um conjunto de documentos é um procedimento que consiste em pegar todos os termos no conjunto de textos e contar o número de vezes que cada termo aparece. Se o histograma resultante for ordenado de forma decrescente, ou seja, o termo que ocorre mais frequentemente aparece primeiro, então a forma da curva é a “curva de Zipf” para aquele conjunto de documentos. Para mostrar o comportamento da sua lei, Zipf utilizou o texto de um livro da literatura inglesa para apresentar a curva gerada graficamente. O eixo “x” do gráfico representa a frequência e o eixo “y” as palavras. Notou-se um padrão, no qual a palavra mais frequente ocorre aproximadamente duas vezes mais do que a segunda palavra mais frequente, três vezes mais que a terceira palavra mais frequente e, assim por diante, mostrando que esse fenômeno ocorre em qualquer texto.

Dessa forma, em um texto sem o pré-processamento, as palavras mais frequentes são denominadas de *stopwords* e aquelas com frequências menores podem ser erros

ortográficos, números e palavras muito específicas. Ou seja, nem as palavras com alta frequência e nem as palavras com baixa frequência são interessantes na discriminação do conteúdo (Allahverdyan et al., 2013).

Para buscar as palavras mais relevantes que se concentram no espaço médio da curva de Zipf, Luhn (1958) usou a lei como uma hipótese nula para especificar dois pontos de corte, os quais denominou de superior e inferior, para excluir termos não relevantes. Os termos que excedem o corte superior são os mais frequentes e são considerados comuns por aparecer em qualquer tipo de documento, como as preposições, conjunções e artigos. Já os termos abaixo do corte inferior são considerados raros e, portanto, não contribuem significativamente na discriminação dos documentos.

Diferente da lei de Zipf, que usa a frequência simples de ocorrência da palavra, a medida *Term Frequency – Inverse Document Frequency* (TF-IDF) utiliza a frequência da palavra no texto, porém inversamente proporcional à frequência da palavra em todo o conjunto de dados. Assim, busca-se identificar as palavras mais relevantes na discriminação de um documento específico.

3. Experimentos e resultados

Os documentos de patentes são organizados em seções de A-H. A classe C possui 21,6% dos documentos e possui a maior quantidade de documentos no conjunto de dados já a classe D possui a menor, apenas 2,3%. Como as classes estão desbalanceadas, e considerando a intersecção das classes, o processo de classificação obteve o melhor desempenho para a classe C (Química e Metalurgia) e o pior desempenho para a classe D (Têxtil e Papel). As classes mais difíceis de buscar o limite de separação durante o aprendizado ocorreram entre: G (Física) versus H (Eletricidade); A (Necessidades Humanas) versus B (Transporte e Operações); B (Transporte e Operações) versus E (Construção Física) (Martins et al, 2021), como mostrado na Tabela 1.

Tabela 1. Frequência de palavras em cada classe

<i>Classe</i>	A	B	C	D	E	F	G	H
<i>Palavra</i>	36.925	33.630	55.968	11.783	13.605	20.035	28.309	24.580

Esse trabalho consistiu em selecionar as palavras que possam ser mais representativas de cada classe da coleção para a criação de um dicionário de palavras e sinônimos, ou um *thesaurus* no qual palavras com significados similares são agrupadas juntas, usando a Lei de Zipf e os cortes de Luhn.

A lei de Zipf foi aplicada separadamente em cada uma das oito seções. O comportamento obtido foi similar à curva padrão de Zipf em cada uma das seções. Está ilustrada na Figura 1 a curva de Zipf para a seção A e os cortes superior e inferior de Luhn, representados nas linhas tracejadas nas cores laranja e verde, respectivamente. Esta seção contém diversos documentos de propriedades registrados pelo IPC relacionados à área de “Necessidades Humanas” como agricultura, alimentos, artigos pessoais ou domésticos, saúde. Assim, a palavra mais frequente é o substantivo inglês 'device' na seção A 8450 vezes, seguido de 'method' (7780), 'relates' (7194), 'material' (5074), 'product' (4885).

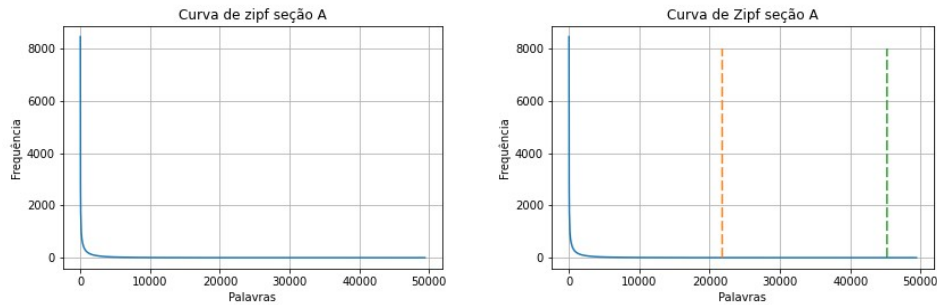


Figura 1. A curva de Zipf para a seção A (esquerda) e os cortes de Luhn (direita)

Após, foi calculado o valor TF-IDF das palavras para esta seção A. Os maiores valores TF-IDF das palavras do conjunto dados foi 'resilienty' com valor 9.0, seguidos por 'mccuster' (9.60), 'disadvantaged' (9.60), 'humphries' (9.60) e 'chiavaro' (9.60), considerando que esta seção A está relacionada com artigos pessoais ou domésticos. Na Figura 2 é mostrado a quantidade de palavras com o mesmo valor TF-IDF.

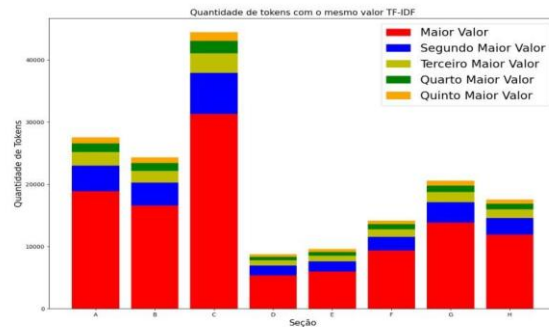


Figura 2. Palavras com o mesmo valor TF-IDF para cada seção.

Algumas das principais palavras de cada seção são mostradas na Tabela 2, com suas respectivas frequências: Frequência Simples (FS) e Frequência TF-IDF. Comparativamente, como já discutido, palavras com FS provavelmente não irão coincidir com as com as mesmas TF-IDF. Como é possível observar, a mesma palavra 'Device' foi uma das mais frequentes em três seções sendo, portanto, uma *stopword*.

Tabela 2. Principais palavras de cada classe e suas frequências

<i>Seção</i>	<i>FS</i>	<i>TF-IDF</i>
A	Device (8450)	Resilienty (9.60)
B	Material (14079)	Unsatisfying (9.60)
C	Process (17121)	Withdistribution (10.00)
D	Fabric (3173)	Voile (7.75)
E	Device (12761)	Rationalised (8.32)
F	Engine (9668)	Penstock (9.66)
G	Method (1398)	Marnar (9.14)
H	Device (12761)	Theic (9.66)

Observa-se que as seções D, E e F são as que possuem um menor número de documentos. Isso pode impactar na frequência já que há menor variabilidade de

documentos disponíveis e, neste caso, o valor de frequência mais alta de E e F se sobressai à frequência da seção como a seção A que possui quase oito mil patentes a mais e, portanto, menos palavras. Um estudo está sendo realizado para analisar o desempenho na seleção das palavras com maior valor TF-IDF e a seleção de acordo com a curva de Zipf e os cortes de Luhn. Nesse último caso, a seleção ocorre baseado nos valores obtidos com o TF-IDF, cujo corte superior considera o limite das palavras com os maiores TF-IDF e uma porcentagem proporcional ao número de palavras para definir o limite do corte inferior.

6. Conclusão

Análise de dados de patentes são naturalmente complexos e apresentam diversas regiões de sobreposição, visto que, uma determinada invenção pode conter características de mais de uma classe (Seção) ou assunto ao mesmo tempo. A linguagem extremamente técnica e rebuscada de textos de patentes torna o processo de classificação difícil devido a grande quantidade de documentos e a alta dimensionalidade. Assim, foram investigados métodos que descrevem fenômenos linguísticos da frequência de palavras em um texto se segue algum padrão, baseado nos cortes de Luhn e na lei de Zipf. Foram analisados a busca por palavras discriminantes a partir das informações de cada classe dos documentos.

Apesar de serem técnicas distintas e com usos diferentes, o TF-IDF e as leis de Zipf podem ser utilizadas de forma complementar visando a análise de dados textuais baseados em frequência e na relevância de palavras. Porém, devido a alta dimensionalidade e a complexidade do domínio, estão sendo investigados métodos para seleção das palavras mais relevantes na discriminação das classes, visando melhorar o desempenho e acurácia dos classificadores e na criação de um dicionário de termos.

Referências

- Allahverdyan, A. E., Deng, W., and Wang, Q. A. 2013. "Explaining Zipf's law via a mental lexicon". *Physical Review E*, v. 88, n. 6, pp. 062804.
- Correa, L. M. S. (1999). "Aquisição da linguagem: uma retrospectiva dos últimos trinta anos". *Revista DELTA: Documentação de estudos em linguística teórica e aplicada*. DOI <https://doi.org/10.1590/S0102-44501999000300014>.
- Fall, C. J., Tórcsvári, A., Benzineb, K., & Karetka, G. (2003). "Automated categorization in the international patent classification". *ACM SIGIRForum* (37:1), pp. 10–25. URL <http://portal.acm.org/citation.cfm?doid=945546.945547>.
- Jing, L., Huang, H., and Shi, H. 2002. "Improved feature selection approach TFIDF in text mining", in *Proceedings. International Conference on Machine Learning and Cybernetics*. IEEE. pp. 944-946.
- Luhn, H. P. (1957). "A Statistical Approach to Mechanized Encoding and Searching of Literary Information". *IBM Journal of Research and Development*. 1 (4): 309-317. ISSN 0018-8646. doi:10.1147/rd.14.0309
- Wipo (2019). *Guide to the International Patent Classification*. Tech. rep. URL <http://www.wipo.int/classifications/ipc/>
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.