

Ferramentas de Coleta e Análise de Dados de Licitações Públicas

Matheus Y. Yamashiro

Prof. Dr. Kelton A. Pontara Prof. Me. Miguel J. das Neves

Universidade Estadual Paulista “Júlio de Mesquita Filho”
UNESP Câmpus de Bauru

Agosto/2022

Licitações – O que são?

- Licitações são os métodos de compra e venda (alienação) de bens feitos pelo Estado.
- Zaffari (2021) diferencia as negociações públicas das negociações particulares:
 - “O particular não tem o dever de prestar contas dos negócios que faz”
 - “O Estado deve selecionar a melhor proposta e a mais vantajosa [...]”
- A liberdade do particular no âmbito público “daria margem a escolhas impróprias” (CARVALHO FILHO, 2017).
- Elaborados para obedecer “aos princípios da legalidade, impessoalidade, moralidade, publicidade e eficiência”, conforme rege a Constituição Federal (BRASIL, 1988).

Licitações – Como analisar esses documentos?

- Os Editais e os Diários Oficiais são escritos utilizando a linguagem natural, utilizada naturalmente pelos humanos em geral.
- Também chamada de linguagem “não estruturada”, possui ambiguidades que dificultam a organização e leitura automatizada.
- Não é um processo trivial. Por isso desenvolveu-se a área de Processamento de Linguagem Natural (PLN, ou NLP, do inglês *Natural Language Processing*).

- Jurafsky e Martin (2021) resumem as competências necessárias:
 - fonética (o som);
 - morfologia (os componentes);
 - sintaxe (a estrutura);
 - semântica (os significados); e
 - discurso (o todo).
- Rodrigues (2020) dá um breve resumo da área:
 - historicamente: implementações de regras e manipulações;
 - atualmente: aprendizagem de máquina (ML, do inglês *Machine Learning*) e aprendizagem profunda (do inglês *Deep Learning*)

- Cadeia de caracteres
 - C, I, Ê, N, C, I, A, _, D, A, _, C, O, M, P, U, T, A, Ç, Ã, O
- *Bag-of-words* (BoW)
 - $[1, 0, 0]^t = \text{CIÊNCIA};$
 $[0, 1, 0]^t = \text{DA};$
 $[0, 0, 1]^t = \text{COMPUTAÇÃO}$
 - $[1, 1, 1]^t$
- *n-grams*
 - (CIÊNCIA, DA); (DA, COMPUTAÇÃO)
 - (CIÊNCIA DA COMPUTAÇÃO)
- *word-embeddings*
 - $\text{CIÊNCIA} = \mathbf{v}^1 \in \mathbb{R}^n$
 - $\text{DA} = \mathbf{v}^2 \in \mathbb{R}^n$
 - $\text{COMPUTAÇÃO} = \mathbf{v}^3 \in \mathbb{R}^n$

Sejam **A** e **B** representações de documentos ou palavras.

- Similaridade de Coincidência
 - *Strings*, *sacolas*, *word-embeddings*...
 - $\mathbf{A} = \mathbf{B} \Rightarrow \forall i \in [0, n], a_i = b_i?$
- Similaridade de Cosseno
 - *Sacolas*, *word-embeddings*, vetores
 - $\mathbf{A} \approx \mathbf{B} \Rightarrow \cos(\mathbf{A} \angle \mathbf{B}) \approx 1$
- Similaridade de Jaccard
 - *Sacolas*, *word-embeddings*, conjuntos
 - $$\frac{\mathbf{A} \cap \mathbf{B}}{\mathbf{A} \cup \mathbf{B}} = \frac{\mathbf{A} \cap \mathbf{B}}{\mathbf{A} + \mathbf{B} - (\mathbf{A} \cap \mathbf{B})}$$

- Determinar as informações mais importantes
- Reduzir de modo que o conteúdo seja o menos afetado possível

Rino e Pardo (2003) colocam que podem ser:

indicativo que “transmite somente uma ideia vaga”; e
informativo que “contém todos os seus aspectos principais”.

Seja n_i um documento i num *corpus* N , f_{ij} a frequência do termo j no documento i , e q_i a quantidade de documentos de N em que o termo j aparece, a importância TF.IDF do termo j para o documento i no *corpus* N é dada por

$$TF \times IDF = \frac{f_{ij}}{|n_i|} \times \log_2 \left(\frac{|N|}{q_i} \right)$$

Metodologia – Ferramentas Utilizadas

- *Python* e bibliotecas padrão
- *Jupyter*
- *spaCy*
- *Apache TikaTM*
- *BeautifulSoup*
- *MySQL*
- *Django*
- *Podman*

Metodologia – Obtenção dos Dados (Diários Oficiais)

- Obtenção da página e dos *links* de download
- Obtenção dos DOs em .pdf
- Raspagem dos DOs pelo *Apache TikaTM* (saída em .xhtml)
- Remoção de *tags* não utilizadas (*div*, *meta*...) utilizando *BeautifulSoup*
- Inserção dos parágrafos no *MySQL*

Metodologia – Obtenção dos Dados (Tabelas)

- Obtenção da página da tabela (*link* direto)
 - Licitações Abertas
 - Licitações Suspensas
 - Licitações Encerradas
- Extração dos dados da tabela pelas *tags* de tabela utilizando *BeautifulSoup*
- Obtenção da página com os detalhes da licitação utilizando os *links* da tabela
- Extração dos dados utilizando *BeautifulSoup*
- Inserção dos dados no *MySQL*

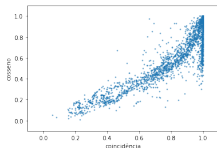
- Obtenção dos dados no *MySQL*
 - Textos
 - Sacolas
 - Valores pré-calculados
- *Tokenização* em *n-grams*
- Aplicação das métricas
 - Similaridade Cosseno
 - Similaridade Jaccard
 - Similaridade de Coincidência
- Exibição na página *web*

Demonstração

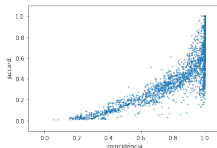
- Página inicial
- Busca de termos
- *Download* de licitações
- *Download* de DOs
- “Leitura” de DO
 - Termos principais
- Licitações semelhantes

- Comparação das Métricas

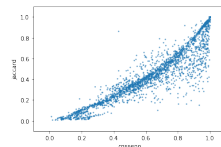
Figura: Comparação da performance dos métodos de comparação de texto



(a) Coincidência \times Cosseno



(b) Coincidência \times Jaccard

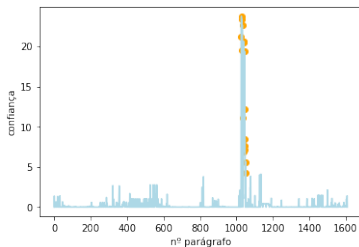


(c) Cosseno \times Jaccard

Fonte: Próprio autor

- Performance da Métrica de Confiança


Figura: Exemplo dos níveis de confiança de um DO





Fonte: Próprio autor


Conclusão e Trabalhos Futuros


- Breve estudo e/ou implementação de técnicas de:
 - PLN
 - Mineração de Dados
 - *Web-development*/Desenvolvimento *fullstack*
- Aprimoramento da Métrica de Confiança/Comparação de Vocabulários
- Aprimoramento da interface *web*
- Melhor uso da técnica de *word-embeddings* para busca de termos semelhantes, por exemplo
- Detecção de erros de digitação


 BRASIL. Constituição da república federativa do brasil de 1988. Brasília, DF, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 05 mai. 2022.

 CARVALHO FILHO, J. d. S. *Manual do direito administrativo*. 31. ed. São Paulo: Atlas, 2017.

 JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 3. ed. Pearson Prentice Hall, 2021. Disponível em: <<https://web.stanford.edu/~jurafsky/slp3/>>. Acesso em: 20 jul. 2022.

 RINO, L. H. M.; PARDO, T. A. S. A sumarização automática de textos: principais características e metodologias. In: *Anais do XXIII Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2003. v. 8, p. 203–245.

 RODRIGUES, S. M. A. F. Introdução ao processamento de linguagem natural. In: SAGAH. *Processamentos de Linguagem Natural*. 1. ed. Porto Alegre: Grupo A, 2020. cap. 1, p. 13–34. ISBN 978-6-55-690057-5. Disponível em: [〈https://online.vitalsource.com/books/9786556900575〉](https://online.vitalsource.com/books/9786556900575). Acesso em: 05 mai. 2022.

 ZAFFARI, E. Licitações públicas: aspectos introdutórios e legais. In: SAGAH. *Licitações e Contratos*. 1. ed. Porto Alegre: Grupo A, 2021. cap. 1, p. 13–23. ISBN 978-65-5690-218-0. Disponível em: [〈https://online.vitalsource.com/books/9786556902180〉](https://online.vitalsource.com/books/9786556902180). Acesso em: 09 mai. 2022.