Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Jennifer Olivia Leiria Albanaz

Reconhecimento de Entidades Nomeadas em resultados de licitações publicados em Diários Oficiais

Curitiba
2020

Jennifer Olivia Leiria Albanaz

Reconhecimento de Entidades Nomeadas em resultados de licitações publicados em Diários Oficiais

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Walmes Marques Zeviani

Curitiba

2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Reconhecimento de Entidades Nomeadas em resultados de licitações publicados em Diários Oficiais

Jennifer Olivia Leiria Albanaz[1]
Walmes Marques Zeviani[2]

Resumo

O Processamento de Linguagem Natural é uma área dedicada a desenvolver a capacidade tecnológica de compreensão da linguagem do homem pela máquina. Entretanto, um dos principais desafios neste campo ainda é a dificuldade de automatizar o entendimento de contexto e semântica de determinadas informações para capturalas corretamente. Neste artigo buscou-se construir um modelo que fosse capaz de identificar empresas vencedoras de licitações em resultados divulgados em Diários Oficiais e para isso, foram utilizadas técnicas de Reconhecimento de Entidades Mencionadas, considerado um importante recurso para o processamento de textos em grande escala. Utilizando o spaCy, uma biblioteca de código aberto para Processamento de Linguagem Natural, foi treinado um algoritmo de redes neurais para reconhecer entidades nos textos de Diários Oficiais coletados. A partir disso, foi possível aplicar o modelo a uma base de resultados desconhecidos e encontrar novas oportunidades de negócios em 90% dos casos.
Palavras-chave: Reconhecimento de Entidade Mencionada, Processamento de Linguagem Natural, Processamento de Textos, Aprendizado de Máquina

Abstract

Natural Language Processing is dedicated to develop technology ability for the understanding of human language, by machines. However, the main challenge in this field is the need of meaning in context and semantics of certain information in order to capture it accurately. The purpose of this is article was to build a model able to identify bidding winners in published texts by Official Journals. Techniques of Named Entity Recognition, considered as an important resource for large-scale text processing, were used. By means of using spaCy, an open source library for Natural Language Processing, a neural network algorithm was trained to recognize entities at texts collected from Official Journal. There after it became possible to apply the model to a database of unknown results and find new business opportunities in 90% of cases.

Keywrods: *Named Entity Recognition, Natural Language Process, Text Process, Machine Learning.*

1 Introdução

De forma cada vez mais veloz cresce a quantidade de textos e informações armazenados pela humanidade. David Lewis afirmou em seu livro *Information Overload* [1], que em 2019 o mundo geraria 20 vezes mais informação do comparado ao ano de 2005. O crescimento de dispositivos inteligentes e conectados tem sido exponencial e o acesso facilitado à tecnologias de alto nível permite a geração de novos dados com muito mais velocidade.

A necessidade de informação é considerada vital para o homem e viabiliza a adaptação às condições externas da existência [2]. Porém, é percebido que o ser humano já atingiu o limite físico e biológico para consumir, processar e analisar a quantidade de informações à disposição.

A extração da informação pode ocorrer de forma natural ou através de consultas a um conjunto de dados armazenados e, tais consultas podem representar um Sistema de Informação, composto por entrada, processamento e saída. Uma das várias subtarefas dessa extração é o Reconhecimento de Entidades Mencionadas (REM) que tem por objetivo localizar e categorizar elementos textuais tais como nomes de lugares, pessoas e organizações, por exemplo. O conhecimento obtido a partir da classificação pode auxiliar a análise semântica de texto, mostrando-se assim um importante recurso para gerenciamento de documentos, mineração de dados e processamento de textos em grande escala.

Dentre os diversos tipos de publicações textuais existentes, sejam elas do meio físico ou digital, destacam-se para fim deste artigo especificamente os Diários Oficiais.

Os Diários Oficiais são jornais criados, mantidos e administrados por órgãos públicos para publicar atos normativos e administrativos de interesse geral da população, como por exemplo leis, atas de plenários e licitações.

Uma licitação é um procedimento obrigatório durante um processo aquisitivo dentro de uma gestão pública, que busca selecionar a melhor proposta para o órgão. Deve ser acessível a qualquer cidadão e de tratamento igualitário a todos os interessados em fornecer produtos ou serviços. Os critérios de decisão são estabelecidos em

[1]Aluno do programa de Especialização em Data Science & Big Data, jennyalbanaz@gmail.com.
[2]Professor do Departamento de Estatística - DEST/UFPR.

edital e vence o participante que oferecer as opções mais vantajosas, com a melhor qualidade e o menor preço.

Segundo a Lei de Licitações e Contratos 8.666/93, a administração do órgão pode exigir que a empresa vencedora garanta a capacidade de cumprir todas as obrigações do contrato a ser firmado entre as partes. Essa lei também prevê a possibilidade da apresentação de uma apólice de seguro-garantia, que tem por objetivo assegurar o cumprimento de obrigações contratuais. Desta forma, vencedores de licitações de obras ou serviços para órgãos públicos, são clientes latentes para uma empresa de seguro garantia, que pode ofertar o seu produto com antecipação.

Contudo, os resultados de licitações são divulgados pelo Diário Oficial do órgão responsável pelo edital, por meio de textos em linguagem natural, não havendo qualquer definição de padrão estrutural ou de categorização para tais publicações. Ou seja, a captura desses resultados depende de uma análise de semântica e contexto, o que ainda se mostra como grande desafio para os sistemas computacionais.

O Processamento de Linguagem Natural é um ramo da Ciência da Computação que utiliza técnicas de extração com o objetivo de desenvolver uma comunicação mais simplificada na relação homem-máquina. Entretanto, um dos principais desafios do processamento de textos de forma inteligente se dá principalmente pelas diversas características das palavras, como por exemplo, polissemia (mesma palavra com muitos significados), sinonímia (palavras diferentes com o mesmo significado) e raridade (palavras pouco conhecidas ou utilizadas). E ainda no âmbito deste artigo, também encontrou-se carência de pesquisa ou literatura relacionada a REM em língua portuguesa.

Diante disso, o presente trabalho teve por objetivo construir um sistema de REM em língua portuguesa, utilizando técnicas de aprendizado em máquina para processar publicações de Diários Oficiais e identificar os vencedores de licitações.

2 Desenvolvimento

A lei 12.527/2011, denominada lei de acesso à informação pública, estabelece a obrigatoriedade de publicação dos editais de licitação na Internet. Cabe a cada órgão publicar informações concernentes a procedimentos licitatórios, inclusive os respectivos editais e resultados, bem como todos os contratos celebrados, em seu *website*.

Para o presente artigo, reuniu-se uma base contendo 119.885 resultados de licitações, os quais serão referidos como registros, divulgados em um período de 108 dias, em diferentes *sites* de prefeituras ou governos, com objeto de contrato exclusivo para obras e prestação de serviços.

2.1 Análises Iniciais

Na análise dos registros foi possível identificar que em alguns casos o nome da empresa vencedora era acompanhado por seu respectivo CNPJ. Então, por meio de de uma expressão regular definida para o padrão deste identificador, extraiu-se o CNPJ da empresa e cruzando esse dado com uma base de pessoas jurídicas da Receita Federal recuperou-se a Razão Social destes vencedores. Com isso, foi selecionado 16% dos registros, reduzindo a média diária de 1.109 oportunidades da base inicial, para 178.

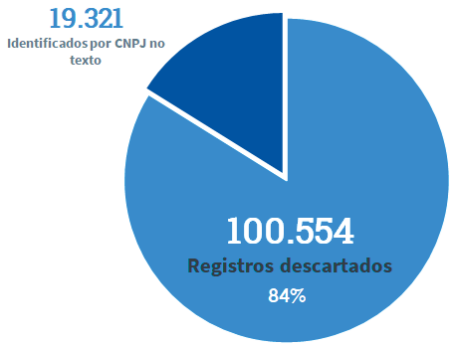

Figura 1: Análise dos Registros

Com a constatação do baixo aproveitamento utilizando técnica simplificada, buscou-se, então, soluções relacionadas a modelos de Processamento de Linguagem Natural (NLP).

2.2 Modelos Disponíveis

Foram pesquisadas as principais bibliotecas livres disponíveis em Python, tais quais NLTK, spaCy, Stanford e OpeNLP. Além da investigação própria e análise da relevância de cada biblioteca para a necessidade da pesquisa, a decisão também foi fundamentada em outras pesquisas científicas realizadas, embora a maioria não em língua portuguesa. Destas, destacou-se principalmente os autores Omran e Treude com seu artigo *Choosing an NLP Library for Analyzing Software Documentation* [3], no qual realizaram uma revisão completa da literatura e uma série de experimentos comparativos entre as bibliotecas. Em seus resultados, spaCy obteve a melhor experiência comparada a todas as demais.

spaCy é uma biblioteca de código aberto para Processamento de Linguagem Natural avançado em Python e foi projetado especificamente para processar e entender grandes volumes textuais [4]. Dispõe de vários recursos de conceitos linguísticos, dentre eles o REM. Uma entidade mencionada é um "objeto do mundo real" que recebe um nome, como por exemplo pessoas, locais, moedas e organizações.

A biblioteca possui modelos pré-treinados, em diferentes idiomas, com várias palavras categorizadas. Essas categorias são chamadas de anotações e tratam-se de marcações de classe (em inglês *part-of-speech tagging* ou POS-Tagging), identificando as palavras com base na sua gramática (substantivos, verbos, adjetivos, advérbios, etc.). O modelo codifica as anotações, mantendo-nas alinhadas, e cria as estruturas de dados necessárias para

um acesso eficiente, chamadas de *GoldParse*. Deste modo, pode aprender, por exemplo, uma frase com várias palavras e identificar a semântica entre elas. À medida que o modelo é aplicado a diferentes textos, esse é atualizado com os novos dados e estruturas, gerando um ciclo de treinamento.

Embora possua anotações em vários idiomas, no tocante a entidades em português ainda é pouco desenvolvido, principalmente quando se trata da categoria Organizações (ORG). A Figura 2 apresenta um exemplo de texto simplificado, utilizando o *displaCy*, uma ferramenta online de visualização dos modelos pré-treinados disponíveis para REM.

Figura 2: Consulta de REM em texto simplificado

É possível perceber que embora o modelo reconheça a palavra Brasil como um Local, o nome como uma Pessoa e o símbolo como uma Moeda, o modelo pré-treinado não conseguiu associar Universidade Federal do Paraná a uma Organização.

Experimentando um texto de Diário Oficial o modelo atendeu menos ainda a expectativa da identificação, conforme Figura 3.

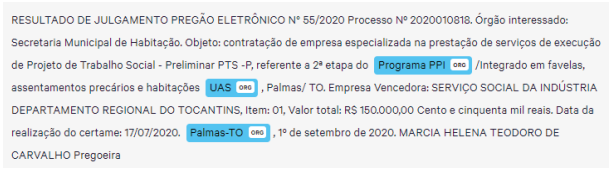

Figura 3: Consulta de REM em publicação de D.O.

Neste caso esperava-se encontrar *Secretaria Municipal de Habitação* e *Serviço Social da Indústria* como organizações.

Sendo assim, optou-se por ignorar os modelos existentes e treinar o algoritmo com os dados da base selecionada a partir da identificação do CNPJ, criando anotações e estruturas exclusivas para esse problema particular [5].

O spaCy utiliza redes neurais convolucionais para processar o conjunto de dados, predizendo a entidade de cada registro. A predição é comparada com o resultado conhecido, ou seja o texto na posição indicada, e então a diferença entre eles é computada (loss). Com o aprendizado da comparação o modelo se ajusta e a cada iteração essa diferença tende a diminuir, tornando-o mais preciso.

2.3 Pré-processamento

A base selecionada para treino e teste resultou em 19.321 registros com 2 colunas: o texto do Diário Oficial, chamada de síntese, e o nome da empresa vencedora, chamada razão social. O tamanho do texto da síntese é bem variado, sendo entre 92 e 32.759 caracteres, e em média 1.048 caracteres. O total de empresas únicas na base é de 4.321, sendo que a mais participativa aparece em 95 registros diferentes. Embora várias empresas tenham uma frequência alta, a mediana do grupo é de 2 participações por empresa (Figura 4).

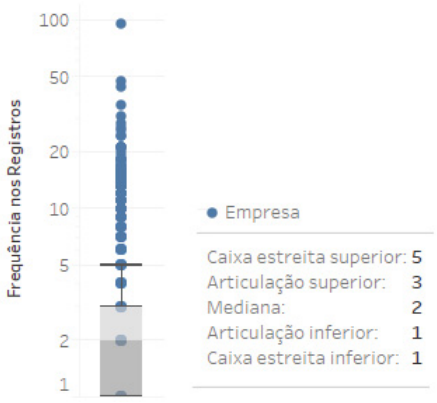

Figura 4: Frequência das empresas nos registros

Notou-se que em alguns registros as empresas apareciam destacadas em caixa alta, desta forma, optou-se por transformar todo o texto da síntese também caixa alta, para evitar uma tendência no modelo. Foram ainda removidos todos os caracteres especiais para evitar erros no código relacionados a pontuação e acentuação, principalmente porque parte das empresas possuem caracteres tais como *ampersand* (&) e hífen (-).

Na sequência, a razão social foi localizada dentro da síntese para identificar sua posição inicial e final em relação ao número de caracteres do texto e a partir desse trecho criar a anotação Organização (ORG). Com isso, foi criado um vetor de treino no qual é passado o texto e as posições iniciais e finais da entidade de cada registro (Figura 5).

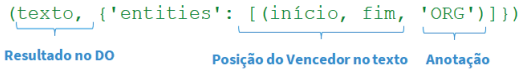

Figura 5: Vetor de Treino

2.4 Treino e Teste

Inicialmente a base foi dividida aleatoriamente em 75% para o treino e aplicando o modelo treinado na base de testes foram identificadas 94% das empresas. Porém, percebeu-se que poderia haver memorização dos nomes, já que é possível a mesma empresa estar nos dois conjuntos devido à sua frequência de participações.

Desta forma, optou-se por fazer um segundo teste, removendo as duplicatas de razão social e mantendo assim apenas empresas distintas. Também dividida em 75% para treino, o teste retornou 90% de acerto. Entretanto, julgou-se que a base ficou muito menor do que a original, e foi ignorado o fato de que por mais que a empresa se repetisse, o texto na qual ela estava inserida é diferente, dado que os processos não se repetem.

Por fim, manteve-se a base completa, porém garantindo que todos os registros da mesma empresa fossem concentrados na mesma porção, ou seja, as bases de treino e teste são conjuntos disjuntos.

Para encontrar o número de iterações necessárias foi definido que o modelo parasse o treino a partir do momento em que a diferença computada (loss) não fosse menor que a anterior por 3 vezes consecutivas.

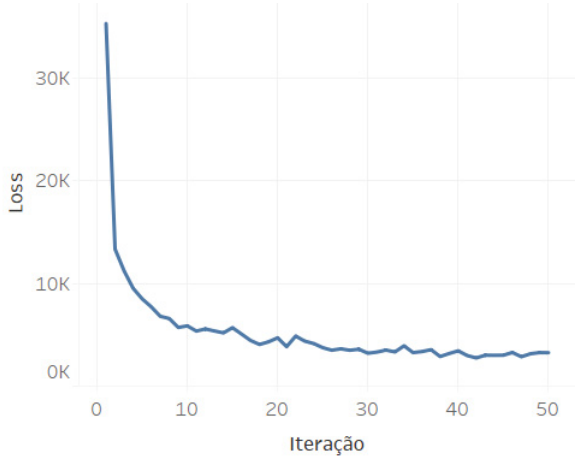

Figura 6: Iteraçãos x Loss

A precisão, sensibilidade e f-score deste último treino foram calculados para avaliação do modelo, buscando verdadeiro positivo, falso positivo e falso negativo, respectivamente.

Scorer	Results
Precision	97.95176
Recall	97.75799
F-Score	97.56498

2.5 Resultados

O modelo treinado foi aplicado à base na qual os vencedores não foram identificados através do CNPJ. O resultado foi de 90.633 registros retornados. Deste resultando extraiu-se uma amostra para analisar manualmente cada registro, e foram identificadas 90% das empresas corretamente como vencedoras. Desta forma, pode-se afirmar que com o trabalho desenvolvido será possível automatizar diariamente cerca de 1.000 oportunidades contra 178 do início da pesquisa, um efeito bastante relevante, já que inicialmente havia uma perda de 84% das oportunidade.

As empresas reconhecidas pelo algoritmo foram relacionadas com a base da Receita Federal e importadas em um novo banco de dados, sendo consideradas como novas oportunidades para ofertas de seguro garantia.

3 Conclusão

O propósito deste trabalho foi de construir um modelo que pudesse identificar automaticamente o máximo possível de vencedores de licitações em Diários Oficiais. Mesmo que processamento textual, principalmente em linguagem natural, seja ainda um ramo muito desafiador, o resultado obtido na pesquisa foi bastante satisfatório e pode alavancar a vantagem competitiva de uma empresa de seguro garantia. Entende-se ainda que a medida que a base for crescendo, o modelo terá mais dados para aprendizado e obterá resultados cada vez mais assertivos.

Agradecimentos

A Junto Seguros pela disponibilização dos dados iniciais para a pesquisa e por me subsidiar nesta especialização. Ao Prof. Walmes Zeviani pela orientação e tempo dedicados e a todo o corpo docente da Especialização em Data Science e Big Data da Universidade Federal do Paraná pelo conhecimento disseminado e dedicação ao ensino.

Referências

[1] D. Lewis. *Information Overload: Practical Strategies for Surviving in Today's Workplace*. Penguin, 1999.

[2] J. Shapiro V. Frantz and V. Voiskunskii. *Automated Information Retrieval: Theory and Methods*. Academic Press, 1997.

[3] F. N. A. Omran and C. Treude. *Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments*. 2017.

[4] spaCy.io. *Online Documentation*. Disponível em https://spacy.io. Acessado em julho, 2020.

[5] K. Jaiswal. *Custom Named Entity Recognition Using spaCy*. Disponível em: https://towardsdatascience.com/custom-named-entity-recognition-using-spacy-7140ebbb3718. Acessado em julho, 2020.