**Processamento de dados textuais: aplicação da biblioteca NLTK como ferramenta analítica.**

Vinícius Andrade Lopes¹\*;João Vitor Matos2

1 Faculdade Rede de Ensino DOCTUM. Bacharel em Sistemas de Informação. Rua Uberlândia, 331 – Centro; 35160-024 Ipatinga, MG, Brasil.

2 Universidade Federal Fluminense (UFF). Doutor em Economia. Rua Miguel de Frias, 9 – Icaraí; 24220-900 Niterói, RJ, Brasil

\*autor correspondente: [vinicius.andlopes@gmail.com](mailto:vinicius.andlopes@gmail.com)

**Processamento de dados textuais: aplicação da biblioteca NLTK como ferramenta analítica.**

**Resumo**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares.

**Palavras-chave:** processamento; linguagem natural; análise; dados; automatização.

**Introdução**

Com o advento da tecnologia e o grande número de pessoas conectadas à internet, é inegável que a quantidade de informações trafegadas pelos principais meios de comunicação, tem se convertido em uma base de dados mundial não estruturada para empresas e profissionais, que possuem um grande potencial técnico e analítico, coletarem informações relacionadas a diversos seguimentos. No entanto, extrair informações robustas e conclusivas dessa base mundial de dados tem se tornado uma tarefa complexa, e alguns aspectos devem ser levados em consideração.

As dificuldades a serem encontradas durante a interpretação de dados textuais podem ser diversas, e elas ocorrem basicamente porque existem várias características textuais que precisam ser compreendidas. Os livros são classificados por gêneros literários, e a alteração no estilo e no conceito de escrita é bastante comum quando se compara um livro de fantasia com um romance, por exemplo. Alguns livros podem conter até mesmo gírias específicas, a variar da região de origem. Já as postagens feitas em redes sociais, são normalmente escritas de maneira informal e com a utilização de abreviações, caracteres especiais e emoticons.

Em complemento à dificuldade na compreensão dos dados, é possível se deparar com outro ponto que deve ser considerado durante o desenvolvimento de análises textuais: o idioma. Atualmente, existem mais de 6.500 idiomas falados em todo o mundo e, dentre eles, os mais populares são o Inglês, Chinês (Mandarim), Hindi, Espanhol e Francês (Gazeau, 2018). Dessa forma, identificar o idioma correto para atender as necessidades da análise textual, pode facilitar todo o processo de coleta de informações e otimizar a capacidade de processamento do “hardware”, visto que filtrar toda a base de dados a apenas um idioma, por exemplo, requer menos poder computacional.

Executar tarefas manuais para extração de informações em repositórios contendo um grande volume de dados não estruturados, ou estruturados, é uma tarefa praticamente impossível de ser realizada. Entretanto, não basta somente ter acesso a grandes volumes de dados. O ponto chave a ser realmente considerado, é como encontrar uma informação útil dentro de um arcabouço de dados diversificados (Yang et al., 2020). Desenvolver uma base de dados estruturada, sem erros de formatação e com os devidos parâmetros definidos, facilita a objetividade da visualização e análise das informações (Wickham, 2016).

Sendo assim, existe uma área da computação que tem por objetivo extrair representações e significados completos de textos escritos de forma livre em linguagem natural, denominada "Natural Language Processing" [NLP] ou Processamento de linguagem natural [PLN], em português (Indurkhya, 2010). Entende-se por linguagem natural os meios de comunicação mais comuns entre seres-humanos, como os próprios idiomas citados anteriormente. Seu crescimento e evolução acontece de geração em geração, e é difícil descrever linguagens naturais por completo seguindo preceitos explícitos, diferentemente de linguagens de programação e fórmulas matemáticas, que possuem princípios já estabelecidas de suas definições (Bird et al., 2009).

Basicamente, a PLN utiliza preceitos linguísticos como classe de palavras para realizar as análises, como por exemplo substantivos, verbos, adjetivos, pronomes, dentre outros, além de diversas estruturas gramaticais que têm por objetivo dar sentido às sentenças analisadas. Isso ocorre em função das várias representações de conhecimento, como um conjunto de palavras existentes em um idioma e seus significados, propriedades e regras gramaticais da linguagem, um grande vocabulário de palavras com relações semânticas, sinônimos e abreviações, e ontologias de entidade e ações (Indurkhya, 2010). O processamento de linguagem natural abrange qualquer tipo de desenvolvimento computacional de linguagens naturais, e consiste em compreender conteúdos descritos por humanos até o ponto de fornecer uma reposta válida sobre a informação analisada (Bird et al., 2009).

O processamento de linguagem natural pode ser definido, de forma simplificada, como uma modelo para identificar quem fez o quê, a quem, quando, onde, como e por que (Robertson, 1946). A PLN considera os textos como uma sequência de caracteres, respeitando a estrutura hierárquica da linguagem que está sendo analisada. Sendo assim, as técnicas de processamento de linguagem natural podem ser utilizadas em “softwares” de diversos seguimentos, como por exemplo corretores gramaticais, conversores de fala para textos, aplicações capazes de traduzir textos para outros idiomas, análise de sentimentos dos usuários mediante a um tema, dentre outros aspectos (Indurkhya, 2010).

**Material e Métodos**

A estratégia de pesquisa utilizada no desenvolvimento deste projeto será exploratória, visando apresentar de forma clara e concisa, didática e prática, a implementação das técnicas de processamento de linguagem natural disponibilizadas pela biblioteca “Natural Language Toolkit” [NLTK][[1]](#footnote-1). Descrever detalhadamente as principais etapas do processamento de textos de linguagens naturais é crucial para a compreensão do funcionamento analítico da ferramenta. As etapas de análises do PLN são decompostas em estágios, conforme apresentados na Figura 1.

Diagrama

Descrição gerada automaticamente

Figura 1. Estágios de análise do processamento de linguagem natural  
Fonte: Adaptado de Dale et al. (2000)

A execução da etapa de tokenização é um processo extremamente importante na análise de linguagem natural, pois permite manipular conteúdos textuais de maneira mais eficiente. O conceito que engloba essa técnica, segundo Castro (2017), é basicamente uma operação que é capaz de dividir cada sentença de um texto em uma sequência de tokens, onde cada token é uma unidade léxica como, por exemplo, uma palavra ou frase. Jurafsky (2000) complementa que a tokenização é o processo capaz de realizar um “split” de textos em palavras, frases, símbolos, ou qualquer outro conjunto de caracteres significativos, conhecido como tokens, que podem ser usados posteriormente para uma análise mais robusta.

Existe diversos métodos de executar essa técnica de tokenização, onde cada uma possui suas vantagens e desvantagens. Um método bastante comum na divisão de textos é a utilização de espaços vazios nas sequencias de caracteres, onde ocorre a geração desses tokens textuais a cada espaço vazio encontrado (JURAFSKY, 2000). No entanto, usar esse tipo de método para a tokenização de textos pode ser complicada em situações em que o espaço é utilizado de forma inconssistente, como por exemplo em frase escritas no idioma inglês que contenham algumas abreviações, como “Mr.” (Kilgarriff and Grefenstette 2003).

COLOCAR UMA IMAGEM QUE EXEMPLIFIQUE A TOKENIZAÇÃO

Outro método bastante comum é a tokenização baseada em expressões regulares. Basicamente, esse tipo de abordagem utiliza um conjunto de regras para identificar e separar os textos, tornando a etapa de tokenizaçao mais assertiva quando comparado ao método baseado em espaços vazios, porém aumentando o nível de complexidade durante a implementação algorítmica (Chang et al., 2014). Existe também um método de tokenizaçao baseada em aprendizado de máquina, onde um algoritmo é treinado com exemplos de textos já tokenizados e, a partir desse modelo já treinado, o sistema é capaz de identificar qual a melhor forma de realizar a tokenização do texto fornecido como “input”. Esse último método citado é mais eficiente e o mais complexo de ser implementado, pois requer um alto nível de conhecimento sobre o tema, sendo necessário desenvolver o conjunto de dados de treinamento ou possuir um conjunto com os parâmetros já treinados (Jiang et al., 2017).

A compreensão dos elementos linguísticos é uma etapa crucial no processamento de linguagem natural, onde é possível identificar e categorizar as palavras, símbolos e marcas de pontuações presentes em um texto. Essa tarefa é realizada por um processo chamado analisador léxico. De acordo com Jurafsky (2020), o analisador léxico é responsável por ler o texto, caractere por caractere, e gerar uma série de tokens, tais como palavras-chave, identificadores, números e símbolos, que serão utilizados posteriormente pelo sistema.

A etapa de análise léxica é fundamental para que o computador possa interpretar os dados textuais com alta precisão. Por exemplo, ao reconhecer uma palavra como um verbo, o sistema pode entender que essa palavra se relaciona a uma ação, e, portanto, pode ser utilizada para realizar certas tarefas. Além disso, a realização da etapa de análise léxica é essencial para o bom desempenho da etapa de análise sintática, pois os tokens gerados pelo analisador léxico serão usados como entrada para outros componentes do processamento de linguagem natural (JURAFSKY, 2000).

A análise sintática é uma fase da etapa de análise léxica, na qual fica responsável por identificar e analisar a estrutura gramatical de uma frase. Essa técnica é usada para determinar as funções sintáticas de cada palavra dentro de uma frase, buscando relações entre essas palavras (Chomsky, 2009). De forma similar as etapas descritas anteriormente, existem diversos métodos de realizar a análise sintática, dentre eles: gramáticas formais, análise estatística e análise de dependência.

As gramáticas formais, de forma resumida, são responsáveis pelo entendimento estrutural de cada frase fornecida como “input” para o algoritmo, identificando o contexto e a relevância de cada palavra (Chomsky, 2009). Já a análise estatística se baseia em modelos estatísticos para compreender a estrutura gramatical de uma frase a partir de dados de treinamento. Esses modelos têm capacidade de lidar com a ambiguidade e a variação na língua natural, se tornando mais robustos em relação aos erros e ao ruído (Charniak, 1997). A análise de dependência se concentra na relação de dependência entre as palavras de uma frase, ou seja, essa técnica tem como princípio de que as palavras de uma frase estão relacionadas entre si através da relação de dependência, como sujeito e verbo, objeto e verbo, dentro outros (Kübler, 2009).

Análise semântica é a etapa que tem por objetivo determinar o significado das palavras e frases, estabelecendo relações semânticas entre elas. Essa etapa é essencial para o entendimento do contexto e o conteúdo das palavras, sendo utilizada para tarefas de processamento de linguagem natural que evolve tradução e geração automática de textos (JURAFSKY, 2000).

Resumidamente, os estágios analíticos do processamento de linguagem natural são essenciais para que a máquina possa entender e processar dados textuais da melhor forma possível.

A PLN tem sua importância nos seguimentos científicos, econômicos, sociais e culturais. O ponto positivo é que essa tecnologia está em constante crescimento devido à implementação de suas técnicas em uma variedade de novas aplicações de linguagem, como descrito anteriormente. Por esta razão, é importante para uma ampla gama de pessoas, ter competências práticas para a utilização dessas técnicas englobadas pela PLN como, por exemplo, analistas de dados, engenheiro de dados, cientista de dados, desenvolvedores de “softwares”, cientistas da computação, acadêmicos, dentre outros (Bird et al., 2009).

NLTK é uma plataforma que trabalha com as diversas técnicas de processamento estatístico de linguagem natural, facilitando a implementação dessas ferramentas em diversos “softwares”. A plataforma possui recursos léxicos e disponibiliza bibliotecas de processamento de textos para tokenização, classificação, “stemming”, análise de raciocínio semântico, e um fórum ativo de discussão (Bird et al., 2009). Para desenvolver aplicações utilizando a biblioteca de código aberto NLTK, é necessário a utilização da linguagem de programação python[[2]](#footnote-2).

Python é uma linguagem de alto nível, orientada a objetos, interpretada e iterativa, de tipagem forte e dinâmica, com uma sintaxe relativamente simples e de fácil compreensão. A linguagem possui diversas estruturas como, por exemplo, dicionários, “arrays”, e uma gama de bibliotecas e módulos prontos para atender grandes demandas de processamento (Borges, 2014).

Uma das bibliotecas mais populares do Python é a NumPy, que manipula matrizes multidimensionais e realiza operações matemáticas de alto desempenho (Pedregosa et al., 2011). Pandas também é uma biblioteca muito importante para realizar tarefas de limpeza e preparação dos dados, pois oferece ao usuário ferramentas analíticas robustas para trabalhar com dados estruturados e fornece estrutura de dados de alto desempenho, tonando-se fundamental para a análise exploratória dos dados (McKinney, 2010).

De forma simplificada, bibliotecas em Python são um conjunto de ferramentas extremamente poderosas disponibilizadas ao usuário, no qual se torna possível realizar uma variedade de tarefas de forma eficiente (Borges, 2014). Também são fundamentais para o reaproveitamento de código.

No entanto, para ter acesso a todas as ferramentas de análise de dados disponibilizadas pela linguagem, é necessário instalar algumas dependências do python em um “hardware” com poder de processamento consideravelmente robusto. Visto que isso pode se tornar um limitante, existem algumas ferramentas de computação em nuvem que disponibiliza uma máquina virtual para realizar essa atividade, como por exemplo o Google Colab[[3]](#footnote-3).

O Google Colab é uma plataforma de computação em nuvem gratuita, baseada em Jupyter Notebooks, que permite aos usuários escrever e executar código em Python. Ele é mantido pelo Google Research e é uma ótima opção para quem procura uma plataforma de desenvolvimento colaborativo e acessível (Gunawan, 2020). A vantagem de usar uma plataforma de como o Google Colab é que ela permite acesso a recursos computacionais de alto desempenho, como Unidades de Processamento Gráfico “Graphics Processing Unit” [GPU] e Unidades de Processamento Tensorial “Tensor Processing Unit” [TPU], sem a necessidade de configurar e gerenciar infraestrutura própria. Isso torna o Colab especialmente útil para aplicações de aprendizado de máquina e processamento de grandes conjuntos de dados (Carneiro, 2018).

A extração de dados utilizados para exemplificar o funcionamento da biblioteca NLTK refere-se a arquivos textuais em formato txt. Esse formato foi escolhido por ser um documento de estrutura simples e de fácil manipulação, contendo sequencias de linhas com descrições de caracteres. Outro ponto positivo de arquivos com extensões .txt é a sua compatibilidade com todos os sistemas operacionais como Linux, Mac, Windows, Android e iOS, por exemplo.

AQUI VOU CITAR QUAIS LIVROS VOU USAR.

Após a etapa de extração dos dados, é essencial realizar a tarefa de tratar as informações coletadas, visando retirar itens indesejados que possam interferir nas análises futuras. Para tratar essa situação, a implementação de técnicas de expressão regular, ou “Regex”[[4]](#footnote-4), são frequentemente utilizadas, pois permitem identificar sequências de caracteres ou determinadas combinações considerando os parâmetros definidos pelo usuário. Pode-se obter resultados significativos utilizando essa ferramenta como, por exemplo, separar letras de números, identificar documentos seguindo padrões de caracteres (CPF, por exemplo), identificar extensões de documentos, dentre outros parâmetros. De forma simplificada, o desenvolvimento de funções com o objetivo de realizar o pré-processamento dos dados textuais contidos nos arquivos, é de grande importância para a parte analítica do projeto.

Com os dados obtidos após a implementação dos processos citados anteriormente, optou-se por desenvolver uma análise qualitativa das informações, onde o intuito é explorar e compreender as principais funcionalidades da biblioteca NLTK.

**Resultados Preliminares**

O desenvolvimento de um algoritmo de processamento de linguagem natural foi realizado seguindo etapas lineares para alcançar o objetivo final do projeto, que é, de forma resumida, gerar insights com informações extraídas de arquivos de texto. Nesse sentido, cada etapa do desenvolvimento da análise de dados textuais foi separada em blocos, para facilitar o entendimento de cada script.

**Conclusão(ões) ou Considerações Finais**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares.

**Agradecimentos** (opcional, 1 parágrafo sucinto)

O título da seção Agradecimentos deve ser alinhado à esquerda e grafado em negrito, primeira letra da palavra grafada em letra maiúscula. Trata-se de seção opcional, de no máximo três linhas, na qual o autor agradece aqueles que contribuíram de maneira relevante para o desenvolvimento do trabalho e elaboração do TCC, mas que não tiveram o envolvimento intelectual necessário à atribuição de coautoria do mesmo, abstendo-se totalmente da menção ou citação de nomes de empresas, instituições ou pessoas que permitiram ou contribuíram com o desenvolvimento do trabalho, a menos que esteja documentalmente autorizado a fazê-lo.

**Referências**

Borges, Luiz Eduardo. Python para desenvolvedores: aborda Python 3.3. Novatec Editora, 2014.

Indurkhya, Nitin, and Fred J. Damerau. Handbook of natural language processing. Chapman and Hall/CRC, 2010.

Robertson, D. W. "A note on the classical origin of" circumstances" in the medieval confessional." Studies in Philology 43.1 (1946): 6-14.

Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.

Gazeau, Valentin, and Cihan Varol. "Automatic spoken language recognition with neural networks." Int. J. Inf. Technol. Comput. Sci.(IJITCS) 10.8 (2018): 11-17.

Wickham, Hadley. "Data analysis." ggplot2. Springer, Cham, 2016. 189-201.

Yang, Jin, et al. "Brief introduction of medical database and data mining technology in big data era." Journal of Evidence‐Based Medicine 13.1 (2020): 57-69.

Dale, Robert; MOISL, Hermann; SOMERS, Harold (Ed.). Handbook of natural language processing. CRC press, 2000.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, v. 12, p. 2825-2830, 2011.

MCKINNEY, Wes et al. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 51-56.

GUNAWAN, Teddy Surya et al. Development of video-based emotion recognition using deep learning with Google Colab. TELKOMNIKA (Telecommunication Computing Electronics and Control), v. 18, n. 5, p. 2463-2471, 2020.

CARNEIRO, Tiago et al. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. IEEE Access, v. 6, p. 61677-61685, 2018.

CASTRO, Sergio M. et al. Automated annotation and classification of BI-RADS assessment from radiology reports. Journal of biomedical informatics, v. 69, p. 177-187, 2017.

JURAFSKY, Dan. Speech & language processing. Pearson Education India, 2000.

JURASFKY, Daniel; MARTIN, James H. An introduction to natural language processing, computational linguistics, and speech recognition. 2000.

KILGARRIFF, Adam; GREFENSTETTE, Gregory. Introduction to the special issue on the web as corpus. Computational linguistics, v. 29, n. 3, p. 333-347, 2003.

CHANG, Angel X.; MANNING, Christopher D. TokensRegex: Defining cascaded regular expressions over tokens. Stanford University Computer Science Technical Reports. CSTR, v. 2, p. 2014, 2014.

JIANG, Zhipeng et al. De-identification of medical records using conditional random fields and long short-term memory networks. Journal of biomedical informatics, v. 75, p. S43-S53, 2017.

CHOMSKY, Noam. Syntactic structures. In: Syntactic Structures. De Gruyter Mouton, 2009.

CHARNIAK, Eugene. Statistical parsing with a context-free grammar and word statistics. AAAI/IAAI, v. 2005, n. 598-603, p. 18, 1997.

KÜBLER, Sandra; MCDONALD, Ryan; NIVRE, Joakim. Dependency parsing. Synthesis lectures on human language technologies, v. 1, n. 1, p. 1-127, 2009.

**Apêndice ou Anexo** (opcional)

Apêndices são textos e/ou documentos que foram elaborados pelo autor e que são importantes para complementar a argumentação do trabalho. Anexos são textos ou documentos que ilustram, mas que não foram elaborados pelos autores. Apêndices deverão seguir as mesmas normas de formatação do restante do texto, inclusive para figuras e tabelas.

O TCC deverá conter no máximo 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s).

1. Página oficial: http://nltk.org/ [↑](#footnote-ref-1)
2. Página oficial: https://www.python.org/ [↑](#footnote-ref-2)
3. Disponível em: https://colab.research.google.com [↑](#footnote-ref-3)
4. Termo abreviado do inglês “regular expression”. [↑](#footnote-ref-4)