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

Vinícius Andrade Lopes¹\*;João Vitor Matos2

1 Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País

2 Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País

\*autor correspondente: nome@email.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 NLTK. 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 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.

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).

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.

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”[[1]](#footnote-1), 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 desenvolvido 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 título da seção Resultados Preliminares deve ser alinhado à esquerda, grafado em negrito com as primeiras letras das palavras em letras maiúsculas. É permitido que a seção seja dividida em subtópicos com formatação de acordo com a descrição no item 1.1 Formato e margens, apresentados na mesma ordem da seção Material e Métodos. Nesta seção devem ser apresentados os resultados parciais da pesquisa, ou seja, os resultados obtidos até o momento.

**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.

**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. Termo abreviado do inglês “regular expression”. [↑](#footnote-ref-1)