**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: Vinícius Andrade Lopes | | **Data início curso**: \_\_27/10/2021\_\_\_\_ |
| **Orientador**: João Vitor Matos | | **Defesa em:**  05/2023 |
| **Curso**: MBA Data Science e Analytics | **Modalidade**: Distância | Turma: 212 |

1. **Título do projeto** *(Inicial)*

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

1. **Introdução**

*(Contextualizar e apresentar a problemática do tema geral, ou seja, a importância do tema proposto e sua relevância. O texto deverá ser escrito de forma impessoal e toda informação utilizada deverá ser embasada por meio de trabalhos de fontes confiáveis com as devidas citações dos autores)*

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, normalmente são 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 (VALENTIN 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, 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* (Processamento de Linguagem Natural), comumente chamada de *NLP* (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 (STEVEN BIRD, 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).

1. **Objetivo**

*(Qual o objetivo principal do trabalho, ou seja, qual pergunta deve ser respondida ao final da sua pesquisa)*

O objetivo geral deste trabalho é descrever detalhadamente as principais técnicas utilizadas para o processamento de dados textuais usando a biblioteca NLTK e linguagem de programação python, exemplificando alguns cenários onde essas técnicas podem ser aplicadas para extrair uma base de informações consistentes e significantes.

Para isso, será desenvolvido um documento onde é possível constatar as aplicações práticas das funções de análise de dados textuais disponibilizadas pela biblioteca NLTK. A base de dados utilizada para a aplicação dessas técnicas, será composta por livros no formato txt.

Os objetivos específicos deste trabalho são: especificar as funcionalidades da biblioteca NLTK que serão utilizadas para o desenvolvimento analítico, descrever - de forma simplificada - o funcionamento de cada uma das técnicas implementadas no decorrer deste projeto, utilizar técnicas de processamento de linguagem natural para aquisição de informações e plotagens de gráficos, identificar outliers na base de dados, construir e estruturar uma base de dados consistente antes da aplicação de técnicas analíticas.

1. **Material e Métodos**

*(Descrever o(s) método(s) de coleta de dados e a(s) ferramenta(s) de análise a ser(em) utilizada(s) no trabalho de conclusão de curso, ou seja, como será a condução da pesquisa e a forma de obtenção dos resultados, por exemplo, fontes de dados, técnicas, procedimentos, índices, entre outros)*

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). Dessa forma, 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 utilizados 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).

A PLN tem sua importância nos seguimentos científicos, econômicos, sociais e culturais. O ponto positivo é que a PLN está em constante crescimento devido a implementação de suas técnicas em uma variedade de novas tecnologias de linguagem, como descrito anteriormente. Por esta razão, é importante, para uma ampla gama de pessoas, ter competências praticas par a utilização dessas técnicas abordadas pela PLN, como por exemplo, analistas de dados, engenheiro de dado, cientista de dados, desenvolvedores de softwares, cientistas da computação, acadêmicos, dentre outros (STEVEN BIRD, 2009).

A NLTK – Natural Language Toolkit, é uma plataforma que trabalha com as diversas técnicas de processamento estatístico de linguagem natural, facilitando a implementação dessas técnicas 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 (STEVEN BIRD, 2009).

1. **Resultados Esperados**

*(Descrever os resultados que são esperados após a realização da coleta e análise dos dados, ou seja, quais resultados são esperados ao final da pesquisa)*

Após descrever e realizar todas as tarefas propostas durante o desenvolvimento deste trabalho, o resultado esperado é ter uma documentação que ofereça uma base de conhecimento consistente para a implementação de técnicas de processamento de dados textuais utilizando a biblioteca NLTK, de forma a exemplificar detalhadamente suas principais funções e demonstrar alguns cenários onde a aplicabilidade dessas técnicas textuais podem ser utilizadas para coletar informações consistentes.

1. **Cronograma de Atividades**

*(Adicionar as “Atividades planejadas”, assim como o período (tempo para desenvolver cada atividade) planejado para a realização de cada atividade, sendo que deverá ser adequado ao calendário de entregas das etapas do trabalho de conclusão de curso definido pela Coordenação. Marcar com um “x” a coluna que corresponde ao período planejado para desenvolver cada atividade planeja)*

Base a partir de agosto.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Pesquisa sobre o tema | **X** |  |  |  |  |  |  |  |  |  |
| Principais materiais teóricos sobre o tema abordado no trabalho |  | **X** | **X** | **X** |  |  |  |  |  |  |
| Entrega do projeto de pesquisa |  |  |  | **X** |  |  |  |  |  |  |
| Introdução ao desenvolvimento prático e definições de bibliotecas |  |  | **X** |  |  |  |  |  |  |  |
| Implementação da biblioteca NLTK |  |  | **X** |  |  |  |  |  |  |  |
| Aplicação das técnicas de NLP em uma base de dados |  |  | **X** | **X** | **X** |  |  |  |  |  |
| Estatística |  |  |  |  | **X** |  |  |  |  |  |
| Revisão dos resultados obtidos através do algoritmo |  |  |  |  |  | **X** |  |  |  |  |
| Resultados preliminares |  |  |  |  |  |  | **X** |  |  |  |
| Desenvolvimento da versão final do TCC |  |  |  |  | **X** | **X** |  |  |  |  |
| Validações estruturais e regras de formatação |  |  |  |  |  | **X** |  | **X** |  |  |
| Revisão final |  |  |  |  |  |  | **X** | **X** |  |  |
| Entrega do trabalho de conclusão de curso |  |  |  |  |  |  |  |  | **X** |  |
| Elaboração da apresentação da defesa do TCC |  |  |  |  |  |  |  |  | **X** |  |
| Entrega da apresentação final da defesa |  |  |  |  |  |  |  |  |  | **X** |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

1. **Referências Bibliográficas**

*(Listagem das bibliografias citadas no projeto de pesquisa, seguindo rigorosamente as Normas do MBA USP ESALQ – Consulte o manual de “Normas para Elaboração do Trabalho de Conclusão de Curso” disponível no Sistema TCC)*

# Bibliografia

INDURKHYA, Nitin A. D. F. J. **Handbook of natural language processing**. 2nd Edition. ed.

ROBERTSON, D. W. A Note on the Classical Origin of "Circumstances" in the Medieval Confessional. **Studies in Philology**, v. 43, p. 6–14, 1946. ISSN 00393738. Published for University of North Carolina Press.

STEVEN BIRD, Ewan K. A. E. L. **Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit**.

VALENTIN GAZEAU, Cihan V. Automatic Spoken Language Recognition with Neural Networks. **I.J. Information Technology and Computer Science**, v. 10, p. 11-17, 08 ago. 2018.

WICKHAM, Hadley. Data analysis. In: SPRINGER **ggplot2**. 2nd Edition. ed. [S.l.]: [s.n.], 2016. p. 189-201. ISBN 978-3-319-24277-4. https://doi.org/10.1007/978-3-319-24277-4\_9.

YANG, Jin A. L. Y. A. L. Q. A. L. L. A. F. A. A. W. T. A. Z. S. A. X. A. A. L. J. Brief introduction of medical database and data mining technology in big data era. **Journal of Evidence - Based Medicine**, v. 13, p. 57-69, 2020. Published for Wiley Online Library.