**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 a internet, é inegável que a quantidade de informações trafegadas pelos principais meios de comunicação, tem se tornado uma base de dados não estruturados mundial 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 de dados mundial 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 isso ocorre 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 especificas de cada país. 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 a dificuldade na compreensão de dados textuais, é possível se deparar com outro ponto que deve ser levado em consideração no desenvolvimento de processamentos e análises e textuais: o idioma. Atualmente, existem mais de 6500 idiomas falados em todo o mundo e, dentre eles, os mais populares são inglês, chines (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 reduz 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. Ademais, não é 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).

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 extrais uma base de informações robustas 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 serão 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; utilização das técnicas de processamento de linguagem natural para aquisição de informações e plotagens de gráficos; identificação de *outliers* na base de dados; construção e estruturação de 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)*

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

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

WICKHAM, Hadley. Data analysis. In: SPRINGER **ggplot2**. Second Edition. ed. [S.l.]: [s.n.], 2016. p. 189-201.

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. ISSN 1.