Instituto de Educação Superior de Brasília - IESB
Ciência de Dados e Inteligência Artificial

Sumarização de textos

por

Victor Augusto Souza Resende - 1922120027

Brasília - DF, 6 de Novembro de 2022

Conteúdo

1 Proposta e objetivos

A sumarização automatizada de textos pode ser feita por meio de técnicas de NLP ou Deep Learning (redes neurais). O resumo poderá ser extraído de duas formas, por meio abstrativa ou extrativo. As técnicas que serão utilizadas tem como foco extrativo, da qual coleta as partes principais do texto por meio de cálculos. A técnica abstrativa "lê o texto" e gera-se um texto novo, porém o desempenho talvez não seja tão escalável.

Assim, o projeto a seguir pretende a implementação de aplicação de uma interface que torne possível a sumarização de textos. Da mesma forma, a aplicação que será construída terá como objetivo a sumarização de artigos científicos por meio de upload do respectivo artigo em formato PDF, visando ajudar e auxiliar estudantes em uma leitura mais breve sobre determinado assunto.

2 Plano de implementação

O plano de atuação do projeto visa a organização e demonstração da arquitetura da qual será utilizada para o desenvolvimento e disponibilização do escopo apresentado anteriormente. Portanto, a seguir é apresentada a arquitetura e tecnologias das quais serão utilizadas.

2.1 Tecnologias e Algoritmos

As tecnologias das quais serão utilizadas visam a utilização de ferramentas denominadas open-sources, das quais possuem como vantagem a reprodução por qualquer pessoa da qual tenha interesse. Além disso, o código-fonte utilizado para a criação da interface para sumarização de textos referente a artigos científicos será disponibilizada no repositório GitHub do autor clicando aqui.

Portanto, a linguagem de programação utilizada será a linguagem Python, da qual se demonstra muito versátil para a implementação de atividades relacionadas ao processamento de linguagem natural (NLP). Da mesma forma, a implementação da interface será efetuada através das bibliotecas Streamlit, PDFplumber, re, NLTK, Unidecode, Spacy e Pandas, presentes na linguagem Python.

Por fim, referente aos algoritmos de processamento de linguagem natural, dos quais serão utilizadas para a sumarização dos textos, verificou-se a possibilidade dos algoritmos de Lun, distância de cossenos e td-idf. Entretanto, concluiu-se que modelos dos quais utilizam Deep Learning (redes neurais) foram mais eficientes na sumarização dos textos. Portanto, decidiu-se seguir com técnicas de Deep Learning, das quais já eram previamente treinadas e então disponibilizadas na comunidade de inteligência artificial Hugging Face.

2.2 Arquitetura

Essa seção visa demonstrar a arquitetura utilizada de maneira diagramada, a fim de tornar a percepção das tecnologias mais simples sobre o projeto em sua totalidade. Portanto, a seguir é apresentado um fluxograma da qual representa a idealização do projeto e suas características.

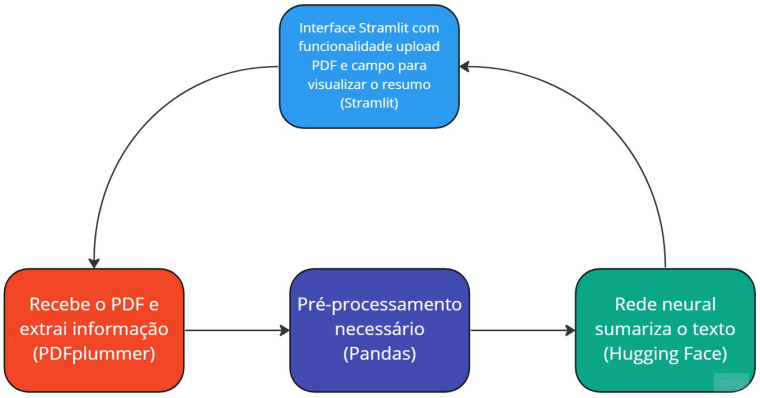

Figura 1: Arquitetura do projeto