Slide 1: Logo Facens

Slide 2: Capa do trabalho

Este trabalho sobre sumarização de textos e a apresentação foi desenvolvida pelos alunos Guilherme Proença Cravo da Costa e Renato Druzian com o auxílio do orientador Prof. Johannes von Lochter.

Slide 3: Agenda

A apresentação será baseada no seguinte cronograma: Introdução, Trabalhos correlatos, Escopo do projeto, Recursos utilizados, Demonstração do Protótipo e, para o encerramento, as Metas para o TCC2.

Slide 4: Introdução

Com sumarização de textos, é possível realizar mais leitura em menos tempo devido ao grande volume de informação e a falta de tempo de hoje em dia.

Com isso, dados de sites de notícias foram coletados com o método webscrapping, que consiste em estruturar os dados para posterior análise.

Também é necessária a implementação de um algoritmo apto para a sumarização, e coletar datasets prontos, que representam os dados de maneira tabular com linhas e colunas.

Slide 5: Trabalhos correlatos

O estudo feito por Chen, K., Corrado, G., Dean, J., Tomas, M., & Sutskever, I., relata que ao representar palavras similares em um vetor, o aprendizado em linguaguem natural possui melhor desempenho.

Ao utilizar o modelo Skip-Gram, o treinamento é feito para identificar palavras semelhantes e, então, relacioná-las ao texto para que, por exemplo, cidades sejam associadas aos seus respectivos países.

No artigo de Kryscinski, W., Keshar, N. S., McCAnn, B., Xiong, C., & Socher, R., é proposto simplificar documentos extensos sem perder o sentido com linguagem natural.

Os datasets necessitam ser em larga escala para melhor análise. Assim, o experimento realizado com dois testes comparativos com humanos, um irrestrito, para escrever resumos com o que acharam mais relevante do texto, e o restrito, baseando o resumo em três perguntas associadas ao artigo.

O estudo notou que os escritores com restrições, dissertaram mais detalhadamente resumos sem informações adicionais.

Já no artigo de Luo, Q., Xu, W., & Guo, J., o estudo é realizado com método CBOW, tem o objetivo de torná-lo menos instável e melhorar o vetor de palavras.

Os resultados apontaram melhora na precisão de 30,80% para 38,43% na relação entre pares de palavras em relação a 52 milhões de palavras com um vetor de duzentas dimensões.

Os dois métodos usados são a propor codificação de frequência de palavra inversa com estrutura regular e estrutura regularizada à função objetivo do modelo CBOW.

Slide 6: Escopo do projeto

Com a implementação de um webscrapping para coleta e um algoritmo para a sumarização das notícias, quanto maior possível for o dataset, o resultado será mais preciso.

Por enquanto, aproximadamente, 10 páginas de sites de notícias foram analisadas.

Slide 7: Recursos utilizados

Os recursos utilizados, até o momento, são o

Collaboratory para centralizar as pesquisas sem a necessidade de um espaço físico

Anaconda Navigator versão 1.9.12 para facilitar o gerenciamento e a implementação de pacotes.

E o Python versão 3.7 como linguagem de programação, com o auxílio das bibliotecas BeautifulSoup4, Pandas, Numpy, Tensorflow.

Slide 8: Demonstração do protótipo

(a definir)

Slide 9: Metas para o TCC2

Para encerrar a apresentação, as metas para o TCC2 são: analisar todos os dados obtidos com o TCC1 para evoluir o protótipo.

Ampliar a quantidade de datasets de notícias, visto que, para o TCC1 o número foi reduzido para a análise ser mais efetiva.