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 as Metas para o TCC2.

Slide 4: Introdução

Sumarização de textos, consiste em sintetizar um texto grande, como notícias, em textos mais curtos que possua a idéia principal da notícia, e caso o leitor queira ter mais informações sobre a notícia, ele pode ir até o site e ler o texto na integra, logo com sumarização, é possível ler mais textos em menos tempo.

Existem duas formas de realizar a sumarização, a extrativa ou a abstrata.

Sumarização extrativa, consiste em criar o texto sem adicionar palavras ou termos que não estejam no texto completo.

Já a sumarização abstrata, cria um texto pode criar um texto “com suas palavras”, ou seja, ela pode adicionar palavras ou termos que não estejam no texto completo

Slide 5: Trabalhos correlatos

O estudo feito por Chen, 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, 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, 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

O trabalho consiste em criar um sumarizador de notícias, para esta primeira parte desenvolvemos um algoritmo para a sumarização extrativa, onde usa apenas palavra contidas no texto principal e sem o uso de redes neurais.

Primeiramente, foi criado um scrapping para coletar notícias e títulos, esse scrapping foi desenvolvido com python e a biblioteca BeautifulSoup, gerando como saída um json com as todas as informações necessárias.

Já o algoritmo de sumarização foi criado com Python e as bibliotecas ScikitLearn, NLTK e NetworkX.

Esse algoritmo usa a vetorização de palavras e frases para que os textos possam ser processados sem maiores problemas, além disso utilizamos a similaridade de cosseno para poder sumarizar a notícia

Slide 7: Recursos utilizados

Os recursos utilizados, até o momento, são o

Google Collaboratory para compartilhar códigos de testes

Anaconda Navigator versão 1.9.12 para criar e padronizar os ambientes, minimizando alguns possíveis problemas, como tempo de seção e internet instável.

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: desenvolver uma rede neural para realizar a sumarização de notícias, aumentar o dataset de notícias e verificar se com o aumento teremos uma diminuição no erro do algoritmo