CENTRO UNIVERSITÁRIO FACENS

REPORT 4

Documento de Projeto

IDENTIFICAÇÃO

Nº	NOME	e-mail	Telefone
160068	Guilherme Proença Cravo da Costa	cravo-guilherme1@hotmail.com	(15)996143888
121058	Renato Druzian	rendruzian@gmail.com	(11)983524314

TÍTULO: Sumarização de notícias

LÍDER DO GRUPO: Renato Druzian

ORIENTADOR: Johannes von Lochter

Data da Entrega: 09 / 06 /2020

_____

Visto do Orientador

SUMÁRIO

1 INTRODUÇÃO DO TRABALHO

O mundo está cada vez mais repleto de informações não-estruturadas, principalmente texto. Mídias sociais, como Twitter e Facebook, tiveram alto crescimento nos últimos anos e influenciam diariamente com opiniões e notícias.

Sites de notícias são fontes provedoras de informações muitas vezes confiáveis, mas o volume de notícias nem sempre é possível de ser acompanhado por uma pessoa ocupada. (Rino & Pardo, 2003) "...viajar pelas páginas de notícias a fim de apreender o que é essencial exige tempo, capacidade de identificar o que é relevante, no grande volume de informações disponível, e capacidade de mentalizar, de forma coerente, o conteúdo essencial...".

Máquinas começaram a ser empregadas para realizar tarefas que antes eram das pessoas, como secretarias que resumem notícias financeiras para os patrões ou agentes de *home brokers* geram *insights* para investidores, que possibilitou a diminuição do tempo de muitos processos.

Em IA, uma das técnicas mais recentes para tratar de sumarização de texto são redes neurais recorrentes. Este método possui aplicações para solucionar algumas análises de sentimento, entidades nomeadas e sumarização de texto.

Primeiramente, para a sumarização de texto, há a coleta das notícias que para utilização como entrada na rede neural, nessa coleta são captados os títulos e os textos das notícias onde o título fica sendo nosso parâmetro de comparação para a saída da rede neural.

Posteriormente, na etapa de teste, que consiste em dividir os dados coletados em treino e teste, esses dois grupos podem ter tamanho que for necessário. A divisão mais comum é 70% para treino e 30% para teste, a rede aprenderá com o grupo de treino e o resultado obtido pela mesma será analisado com o grupo de teste.

A rede neural aprenderá lendo o texto da notícia e fará uma ligação com o título, já na etapa de teste lerá as notícias do grupo que ela não conhece os dados e tentará gerar um título, ao término será realizado uma análise dessa saída com o título original, com isso será feito a análise, de acordo com (Ferneda, 2006) "As redes neurais artificiais se diferenciam pela sua arquitetura e pela forma como os pesos associados às conexões são ajustados durante o processo de aprendizado".

Entretanto, deve-se observar que se a maioria das notícias na parte de treinamento possuírem títulos sensacionalistas ou que não condizem com as reais informações apresentadas no texto, será gerado resultados não confiáveis, podendo com isso gerar um algoritmo enviesado, ou seja, que pensa de forma muito parecida com a fonte das notícias.

Para evitar esse problema e garantir melhor assertividade, é necessário um grande volume de dados e de várias fontes, evitando assim criar algum viés na rede neural.

Por fim, a área científica poderá beneficiar-se para melhor inserção de títulos em artigos, assim como resenhas ou textos gerais em instituições acadêmicas, pois a sumarização possibilita que mais textos sejam lidos em menor tempo, sem que haja perda no sentido para que a mensagem seja transmitida.

2 TRABALHOS CORRELATOS

A sumarização de textos possui métodos úteis para diminuir o grande volume de informações e, a combinação deles pode trazer bons resultados, já que todos tem uma desvantagem na sua utilização.

Estudos feitos por (Brownlee, 2017) apontam o funcionamento do método *Bag of Words*, comumente usado para auxiliar um computador a compreender textos atribuindo um peso para cada palavra, descartando palavras com pesos menores sem muita informação como, *stop words*, e mais relevância as com maior peso.

O uso de redes neurais para sumarização de textos realizado por Kryscinski et al. (2019), aponta que existem partes críticas na análise da documentação que são fragmentos com informações mais significativas como, palavras contidas no sumário.

Para Chen, K., Corrado, G., Dean, J., Tomas, M., & Sutskever, com o uso distribuído de alta performance de representação de vetores do *Skip-Gram*, o processo de palavras similares é melhorado em linguagem natural.

3 O QUE SERÁ FEITO

Para a apresentação do TCC1 será apresentado uma implementação de sumarização extrativa, onde realizamos a sumarização apenas com termos contidos no texto da notícia principal.

Nesta primeira etapa foi utilizado um *dataset* com 287 notícias, e realizado a sumarização delas, esta etapa foi desenvolvida para entender da melhor maneira como é realizada a sumarização para que ao utilizar a rede neural para isso fique mais claro quais são as possíveis entradas e possíveis problemas que podemos encontrar

4 O QUE NÃO SERÁ FEITO

Neste trabalho não será entregue:

- Um produto comercial;
- De maneira escalável;
- Sumarização de qualquer linguagem, idioma, apenas português;
- A capacidade de sumarização em qualquer domínio, apenas notícias;

5 BENEFÍCIOS

Os benefícios deste trabalho são auxiliar na criação de resumos de notícias para uma leitura rápida e caso o leitor tenha interesse na notícia, pode entrar na página do site da mesma para obter mais detalhes.

Os meios acadêmico e científico são beneficiados devido a criação de textos diários que podem ter melhorias significativas na elaboração de artigos, resenhas, anotações, entre outros.

Por fim, melhorar a compreensão do leitor em relação ao texto, de qual maneira que, haja mais coesão e coerência.

6 METAS PARA O TCC 2

As metas para o TCC2 são melhorar o algoritmo ou mudar a forma de análise dos dados:

- Desenvolver uma rede neural para sumarização da notícias
- Aumentar a quantidade de notícias a ser analisada
- Finalizar a monografia e apresentação para banca.

7 RECURSOS UTILIZADOS

Para o desenvolvimento até o momento foi utilizado o Google Colaboratory, Anaconda versão 1.9.12 para criação de ambiente de desenvolvimento e simplificar a instalação das bibliotecas utilizadas no desenvolvimento.

Com o Anaconda foi montado um ambiente com Python versão 3.7, além das bibliotecas BeautifulSoup4, Pandas, Numpy, Tensorflow.

REFERÊNCIAS

Brownlee, J. (2017). *Machine Learning Mastery*. Fonte: Machine Learning Mastery: https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Chen, K., Corrado, G., Dean, J., Tomas, M., & Sutskever, I. (s.d.). Distribuited representations of words and phrases and their compositionality., (p. 9).

Ferneda, E. (2006). *Redes neurais e sua aplicação em sistemas de recuperação de.* Ribeirão Preto.

Kryscinski, W., Keshar, N. S., McCAnn, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation., (p. 13).

Luo, Q., Xu, W., & Guo, J. (2014). A study on the CBOW model's overfitting and stability., (p. 4). Shangai.

Rino, L., & Pardo, T. (2003). A sumarização automática de textos principais características.