

**COORDENADORIA DE ENGENHARIA COMPUTAÇÃO**

**GUILHERME PROENÇA CRAVO DA COSTA**

**RENATO DRUZIAN**

**SUMARIZAÇÃO DE NOTÍCIAS**

**Sorocaba/SP**

**2020**

**Guilherme Proença Cravo da Costa**

**Renato Druzian**

**SUMARIZAÇÃO DE NOTÍCIAS**

Trabalho de conclusão de curso apresentado ao Centro Universitário Facens como exigência parcial para obtenção do diploma de graduação em Engenharia da Computação.

Orientador: Prof. Johannes von Lotcher

**Sorocaba/SP**

**2020**

FICHA CATALOGRAFICA

ELABORADA PELA “BIBLIOTECA FACENS”

SXXXs

Sobrenome, Nome do Aluno 1; Sobrenome, Nome do Aluno 2; Sobrenome,

Título do trabalho: subtítulo se houver/ por Nome do Aluno.- Sorocaba, SP: [s.n.], 2020.

XX f.; 29cm.

Trabalho de Conclusão de Curso (Graduação) – Centro Universitário Facens, Coordenadoria de Engenharia de Computação – Curso de Engenharia de Computação, 2020.

Orientador: Prof.

1. Palavra-chave. 2. Palavra-chave. 3. Palavra-chave. I. Faculdade de Engenharia de Sorocaba. II. Título.

CDD XXX

**Guilherme Proença Cravo da Costa**

**Renato Druzian**

**SUMARIZAÇÃO DE NOTÍCIAS**

Trabalho de conclusão de curso apresentado ao Centro Universitário Facens como exigência parcial para obtenção do diploma de graduação em Engenharia da Computação.

Orientador: Prof. Johannes von Lotcher

Sorocaba, XX de XXXXXXXXXX de 2020

BANCA EXAMINADORA

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Esp/Me/Dr e Nome

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Esp/Me/Dr e Nome

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Esp/Me/Dr e Nome

**AGRADECIMENTOS**

Frase de um autor que representa o espírito desse trabalho. Frase entre aspas duplas.

Nome do autor

**RESUMO**

**Palavras-Chave**:

**ABSTRACT**

**Key-words**:

**LISTA DE ILUSTRAÇÕES**

[Figura 1 - CBoW 16](#_Toc39943334)

[Figura 2 – Skip Gram 17](#_Toc39943335)

**LISTA DE EQUAÇÕES**

Equação 1 – Skip Gram 16

Equação 2 – Probabilidade Skip Gram 17

Equação 3 – GloVe 18

Equação 4 – Probabilidade GloVe 19

Equação 5 – Vetor resultante GloVe 19

Equação 6 – Simplificação da Equação 5 19

Equação 7 – Função de custo do modelo GloVe 19

**LISTA DE SIGLAS**

IA Inteligência Artificial

BoW *Bag of Words*

S2S *Sequence to sequence*

GloVe *Global Vectors for Word Representations*

CBoW *Continous Bag of Word*

**SUMÁRIO**

[1 INTRODUÇÃO 13](#_Toc39933661)

[2 REPRESENTAÇÃO DE TEXTO 15](#_Toc39933662)

[2.1 Bag of Words 15](#_Toc39933663)

[2.2 Word Embeddings 15](#_Toc39933664)

[2.3 Skip Gram 16](#_Toc39933665)

[2.4 Cbow 17](#_Toc39933666)

[2.5 Glove 17](#_Toc39933667)

[3 MÉTODOS DE SUMARIZAÇÃO 18](#_Toc39933668)

[3.1 Sequence to Sequence 18](#_Toc39933669)

[4 METODOLOGIA 18](#_Toc39933670)

[5 RESULTADOS 18](#_Toc39933671)

[6 CONCLUSÃO 18](#_Toc39933672)

# INTRODUÇÃO

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.

# REPRESENTAÇÃO DE TEXTO

O texto da forma como é conhecida não pode ser interpretado pelo computador, por isso, é necessário alterar a forma de representar texto para que o computador possa interpretá-lo e analisar o resultado.

Existem duas formas amplamente utilizadas no meio acadêmico que trazem resultados satisfatórios, esses métodos são o *Bag of Words* e *Word Embeddings*, que através de cálculos matemáticos uma máquina compreenda uma linguagem humana.

## Bag of Words

Esse método consiste em analisar todo o *dataset* e montar um dicionário com todas as palavras contidas, ou seja, criar um “saco de palavras” com todas as amostras de palavras, sem que haja repetições da mesma.

Após coletar todas as palavras e analisar uma frase, é criado um *array* com 1 e 0 para simbolizar se aquela palavra existe ou não no dicionário criado.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dicionário | A | FACENS | é | nota | máxima | no | Enade |
| Frase | FACENS | conseguiu | nota | máxima | no | Enade |  |
| *Array* | 1 | 0 | 1 | 1 | 1 | 1 |  |

E com esse *array*, é possível realizar cálculos e análises para fazer comparações, em outras palavras, nesse formato o computador consegue “entender” a frase.

## Word Embeddings

Segundo estudo de (Bolukbasi, Chang, Zour, Saligrama, & Kalai, 2016), este método representa cada palavra ou alguma frase, em vetores de palavras de n-dimensões, e esse vetor serve como um dicionário.

Com esse dicionário criado, o computador criará relações semânticas entre as palavras para medir a distância entre os vetores que estão sendo analisadas, e com isso, pode-se criar analogias tal como:

Com este exemplo simples pode-se verificar o quão poderoso pode ser o método *word embeddings.*

## Skip Gram

Conforme estudo de (Chen, Corrado, Dean, Tomas, & Sutskever) o *Skip gram* é um método baseado em vetores, ou seja, cada palavra é transformada em um vetor e tem como finalidade prever quais são as combinações de palavras disponíveis para uso antes e depois da palavra, dado a mesma como entrada.

Assim, é possível analisar o contexto em que a palavra está inserida para poder encontrar essas combinações. Outro uso deste método é, dado dois vetores ao realizar a soma deles, gerar na saída um terceiro vetor com relevância em relação a soma dos dois vetores, como no exemplo abaixo:

Nesse exemplo, realiza-se a conexão das palavras Brasil e Capital pode-se notar que a saída dele é Brasília.

Outro ponto que pode ser notado, é que ao treinar um modelo com *skip gram*, encontra-se uma combinação de palavras que tenha sentido no contexto.

Dado uma sequência de palavras de treino, w1, w2, ... wn, o skip gram tem o objetivo de aumentar a probabilidade de uma palavra fazer sentido no contexto.

Equação – Skip Gram

Onde tem-se que c é o tamanho do contexto de treino, quanto maior o c maior são os exemplos de treino, maior o tempo de treino e consequentemente melhor será o resultado do modelo.

A base da fórmula do *skip gram* é definida por , que também podemos escrever como:

Equação – Probabilidade Skip Gram

Já e são as representações dos vetores de entradas e saídas da palavra w, e W é o número de palavras no vocabulário, ou seja, o número total de palavras presentes no conjunto de treino.

## Cbow

Como (Luo, Xu, & Guo) disse em seu trabalho, *CBoW* é mais uma técnica de representação de texto que utiliza vetores para simbolizar palavras, esta técnica também faz uso de redes neurais recorrentes para avaliar e prever uma palavra em um determinado contexto.

De uma forma bem simples pode-se dizer que o *CBoW* faz o caminho inverso do *Skip Gram,* onde o último quer, através de uma palavra prever toda a frase, enquanto o CBoW possui a frase mas com uma palavra faltando, e ele pretende prever essa palavra faltante.

O uso de redes neurais recorrentes se enquadra, pois as palavra antes são representadas como sendo o passado e as palavra após a lacuna como sendo o futuro, e entende-se como lacuna a palavra que se deseja prever.

Abaixo podemos ver duas imagens que mostram de maneira mais clara a diferença entre eles.

Uma imagem contendo objeto, relógio

Descrição gerada automaticamente

Figura - CBoW

Uma imagem contendo objeto, relógio

Descrição gerada automaticamente

Figura – Skip Gram

## Glove

Conforme estudos realizados por (Pennington, Socher, & Manning, 2014) os Vetores Globais, ou *Global Vectors,* são amplamente utilizados para criar vetores semânticos em representação da linguagem. Esses vetores são utilizados em aplicações de classificação de documentos, respostas automáticas de questões, recuperação de informação e análises textuais.

Existem dois modelos conhecidos para aprendizagem de vetores de palavras, que são:

* Fatorização de matrizes, utilizados para análise semântica;
* Janela de contexto local, como o skip gram por exemplo

Para continuar no estudo do GloVe vamos definir algumas coisas, X representa uma matriz que conta a ocorrência de uma palavra do conjunto de teste por uma palavra do conjunto de treino e representa o número de vezes que uma palavra j apareceu no contexto da palavra i. Com isso temos:

Equação – GloVe

E por último temos a probabilidade de uma palavra j ocorrer no contexto de i:

Equação – Probabilidade GloVe

Por fim temos a equação resultante do vetor, F, que irá apontar para a palavra que mais se adequa ao contexto:

Equação – Vetor resultante GloVe

Onde temos que e são vetores de palavras e e são vetores de palavras livres de contexto. Simplificando a Equação 5, chegamos em:

Equação – Simplificação da Equação 5

O índice *k* é referente a palavra que se está comparando com *i* e *j*, para com isso poder dizer se *k* está mais próximo de *i* ou de *j*. Se o resultado estiver mais próximo de 0 entre as palavras *k* e *i*, então existe algum tipo de ligação entre *k* e *i*.

Agora analisando a Equação 5, se o valor estiver próximo de 1 então as duas palavras, *i* e *j*, tem ligação com a palavra *k.*

Por fim (Pennington, Socher, & Manning, 2014) diz que a função de custo para o modelo é:

Equação – Função de custo do modelo GloVe

Onde temos que *V,* é todo o vocabulário disponível para treino.

# MÉTODOS DE SUMARIZAÇÃO

## Sequence to Sequence

# METODOLOGIA

# RESULTADOS

# CONCLUSÃO

**REFERÊNCIAS**

Adelson, P., Arora, S., & Hara, J. (s.d.). Abstractive text summarization with Quasi-Recurrent neural netwaorks. p. 7.

Bolukbasi, T., Chang, K.-W., Zour, J., Saligrama, v., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing embeddings. *30th Conference on Neural Information Processing Systems (NIPS 2016)*, (p. 9). Barcelona.

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.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. *50th Annual Meeting of the Association for Computational Linguistics.* Jeju: 10.

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. (s.d.). A study on the CBOW model's overfitting and stability., (p. 4). Beijing.

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization., (p. 12). Palo Alto.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (p. 12). Doha.

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