Congresso Acadêmico UNIFESP 2020

Ciência e Universidade: Transformações para a Sociedade

Congresso Acadêmico Unifesp 2020

Um estudo comparativo entre *bag-of-words* e *word embedding* na classificação de textos

Celso Gabriel Vieira Ribeiro Lopes
Orientadora: Profa. Dra. Lilian Berton

13/07/2020

Congresso Acadêmico UNIFESP 2020

Inteligência artificial

- A Inteligência Artificial (IA) é uma área da ciência da computação que busca a automação do comportamento inteligente.

- Se divide em várias outras sub-áreas, dentre elas o Processamento de Linguagem Natural (PLN) e o Aprendizado de Máquina (AM).

Processamento de Linguagem Natural

• O Processamento de Linguagem Natural (PLN) busca fazer com que os computadores entendam as declarações ou palavras escritas em idiomas humanos.

• Surgiu para facilitar o trabalho do usuário e satisfazer o desejo de se comunicar com o computador em linguagem natural.

• É aplicada em diversas tarefas, como por exemplo **classificação de texto**, objeto de estudo desse trabalho.

Congresso Acadêmico UNIFESP 2020

Pré-processamento de texto

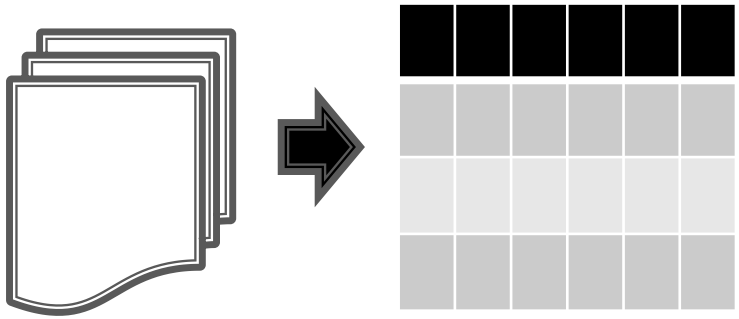
- *Tokenization*
- Normalização de palavras
- Remoção de *stopwords*
- *Stemming*

"O rato roeu a roupa do rei de Roma."

rato roer roupa rei roma

Extração de features

- Bag of Words (BOW)
- Word Embedding (Word2Vec)



Bag of Words

É um método muito utilizado na representação de documentos textuais. No caso de texto, este é representado como um vetor, onde cada posição possui o número de aparições de uma palavra ao longo do conteúdo do documento. Esse método utiliza duas maneiras para verificar a frequência das palavras.

- *Term Frequency* (TF)
- *Term Frequency - Inverse Document Frequency* (TF-IDF)

Word Embedding

- É um conjunto de técnicas que mapeiam a semântica e sintática de uma linguagem natural em um espaço usando estatísticas.

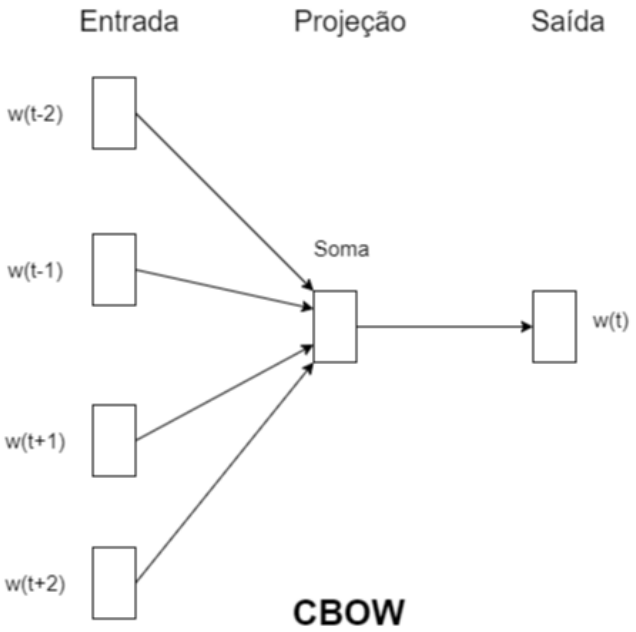
- Palavras de um conjunto de texto são mapeados para vetores. Esses vetores possibilitam um melhor desempenho nas tarefas do PLN, pois não tratam as palavras como únicas, em vez disso, refletem a similaridade e dissimilaridade entre elas.

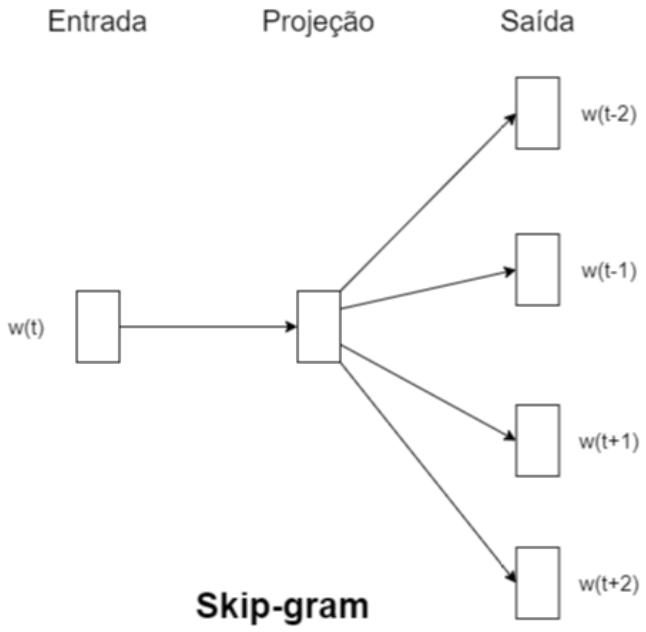
Word2vec

- É um modelo utilizado para calcular a representação de palavras como vetores via redes neurais.
- O word2vec tem se tornado um dos métodos mais usados no pré-processamento de textos.
- Ele oferece dois modelos de redes neurais em sua arquitetura: *continuous bag-of-words* (CBOW) e *skip-gram*.

- CBOW
- Skip-gram

Word2vec

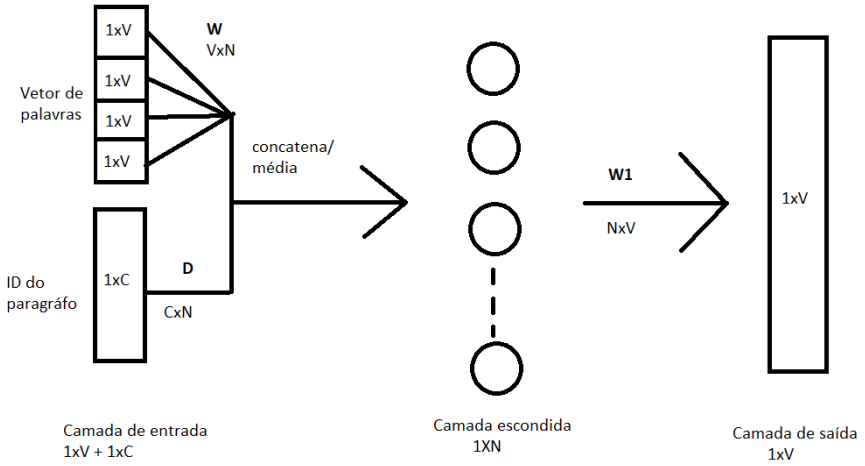

Entrada
Projeção
Saída
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Soma
w(t)
CBOW


Entrada
Projeção
Saída
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Skip-gram

Congresso Acadêmico UNIFESP 2020

Doc2vec

O conceito do Doc2vec é parecido com o modelo do Word2vec. A diferença é que o doc2vec utiliza o Word2vec e adiciona mais um vetor (ID do parágrafo) à entrada.



Aprendizado de máquina

- O Aprendizado de Máquina (AM) objetiva o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.
- Dentro do AM existem técnicas que podem ser divididas em aprendizado supervisionado e não-supervisionado.
- Os algoritmos supervisionados utilizados foram:
    - *k-Nearest Neighbors* (k-vizinhos mais próximos)
    - *Naive Bayes*
    - *Support-vector machine* (Máquina de suporte vetorial)
    - *Decision Tree* (Árvore de decisão)
    - *Multilayer Perceptron*

Congresso Acadêmico UNIFESP 2020

Materiais e métodos

- Para realizar a classificação de textos, todos os algoritmos foram e scripts foram feitos na linguagem de programação Python, junto com as **bibliotecas**: NLTK, Scikit-learn, Gensim, Pandas, Matplotlib e Numpy.

- Os ***datasets*** utilizados foram: Brown e Reuters

- As **medidas de validação**: acurácia, precision, recall, f1-score

Materiais e métodos

Brown

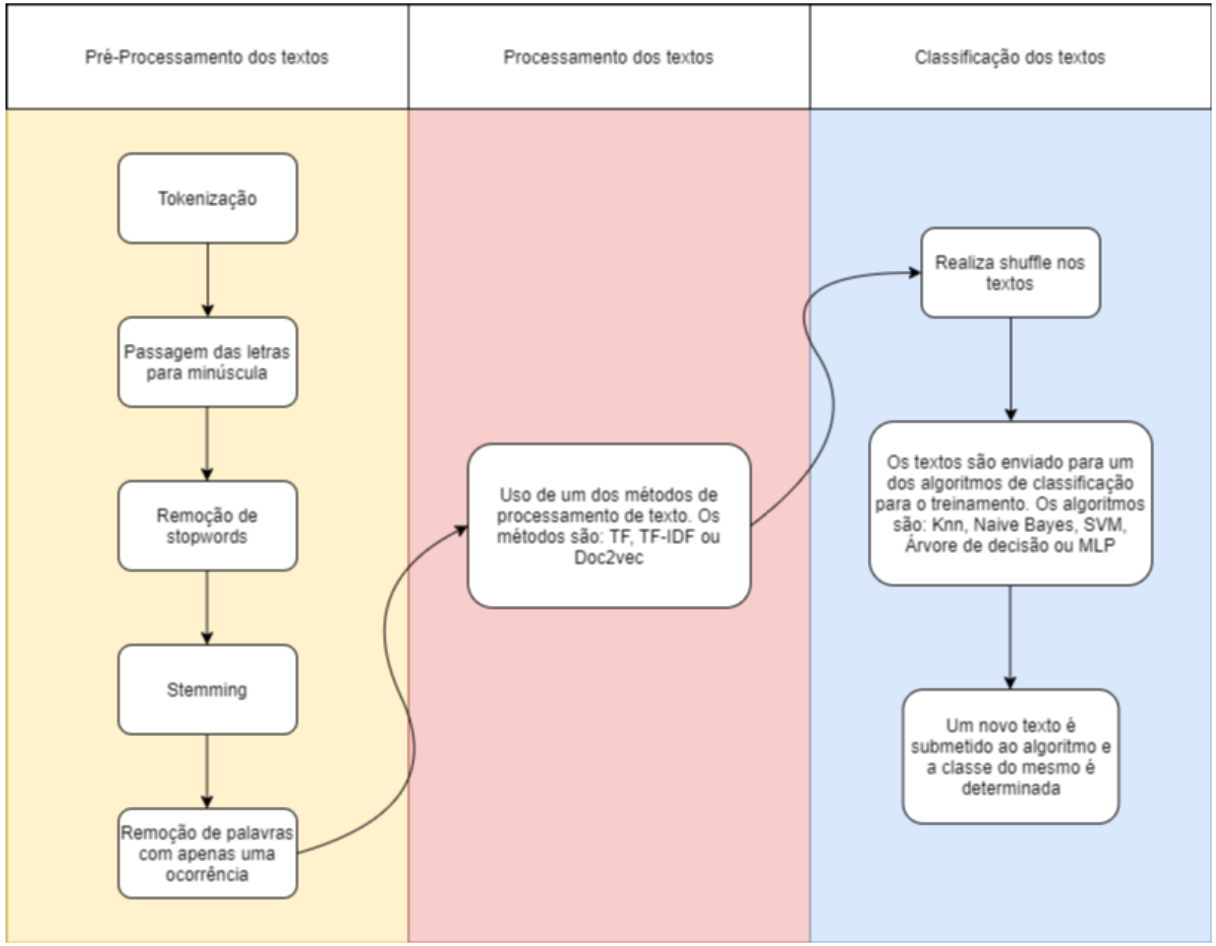
Classes	Quantidade de textos
learned	80
belles lettres	75
lore	48
news	44
hobbies	36
government	30
adventure	29
fiction	29
romance	29
editorial	27
mystery	24
religion	17
reviews	17
humor	9
science fiction	6

Reuters

Classes	Quantidade de textos	Classes	Quantidade de textos
earn	3964	wpi	29
acq	2369	orange	27
money fx	717	rapeseed	27
grain	582	strategic metal	27
crude	578	soy meal	26
trade	485	retail	25
interest	478	soy oil	25
ship	286	fuel	23
wheat	283	hog	22
corn	237	housing	20
dlr	175	heat	19
money supply	174	income	16
oilseed	171	lumber	16
sugar	162	sunseed	16
coffee	139	lei	15
gnp	136	dmk	14
gold	124	oat	14
veg oil	124	tea	13
soybean	111	platinum	12
bop	105	groundnut	9
nat gas	105	nickel	9
livestock	99	l cattle	8
cpi	97	rape oil	8
cocoa	73	coconut oil	7
reserves	73	sun oil	7
carcass	68	coconut	6
jobs	67	instal debt	6
copper	65	naphtha	6
cotton	59	potato	6
rice	59	propane	6
yen	59	jet	5
alum	58	cpu	4
gas	54	mdlr	4
iron steel	54	copra cake	3
ipi	53	cotton oil	3
barley	51	dfl	3
meal feed	49	nkr	3
rubber	49	palladium	3
palm oil	40	palmkernel	3
sorghum	34	rand	3
zinc	34	castor oil	2
pet chem	32	groundnut oil	2
tin	30	lin oil	2
lead	29	rye	2
silver	29	sun meal	2

Congresso Acadêmico UNIFESP 2020

Materiais e métodos



Congresso Acadêmico UNIFESP 2020

Resultados

- Resultados do *dataset* Brown utilizando os métodos TF e TF-IDF:

Algoritmos	Accuracy	F1 - Score	Precision	Recall
KNN	0.07133	0.03916	0.04185	0.05998
SVM	0.40199	0.27266	0.32017	0.30353
Naive Bayes	**0.52**	**0.44693**	**0.49632**	**0.48972**
Decision Tree	0.22599	0.15962	0.17440	0.17928
MLP	0.50600	0.43281	0.45905	0.47077

Accuracy	F1 - Score	Precision	Recall
0.438	0.33031	0.34857	0.39177
0.416	0.28555	0.30937	0.33542
0.51	**0.42641**	**0.46691**	**0.46584**
0.236	0.17474	0.19901	0.19133
0.478	0.37166	0.37166	0.39641

- Resultados do *dataset* Brown utilizando o método Doc2vec:

Algoritmos	Accuracy	F1 - Score	Precision	Recall
KNN	**0.402**	**0.29985**	**0.35027**	**0.33198**
SVM	0.358	0.18768	0.21376	0.24117
Naive Bayes	**0.366**	**0.29599**	**0.32917**	**0.32969**
Decision Tree	0.281	0.19438	0.21344	0.21141
MLP	0.444	0.33514	0.36913	0.36846

Congresso Acadêmico UNIFESP 2020

Resultados

- Resultados do *dataset* Reuters utilizando os métodos TF e TF-IDF:

Algoritmos	Accuracy	F1 - Score	Precision	Recall
KNN	0.63918	0.30760	0.329157	0.31766
SVM	0.30018	0.00727	0.00473	0.01577
Naive Bayes	0.61967	0.08884	0.12426	0.09617
Decision Tree	0.57940	0.22992	0.24574	0.23757
MLP	**0.65682**	**0.31762**	**0.33952**	**0.31916**

Accuracy	F1 - Score	Precision	Recall
0.55874	0.21127	0.24463	0.21576
0.44198	0.04472	0.06375	0.04523
0.65776	0.24026	0.26767	0.24382
0.55958	0.20235	0.22014	0.20579
0.63219	**0.27282**	**0.29794**	**0.27463**

- Resultados do *dataset* Reuters utilizando o método Doc2vec:

Algoritmos	Accuracy	F1 - Score	Precision	Recall
KNN	**0.809**	**0.33845**	**0.37119**	**0.34962**
SVM	0.744	0.17855	0.19488	0.19418
Naive Bayes	0.613	0.23456	0.23538	0.28406
Decision Tree	0.585	0.13691	0.14086	0.14459
MLP	**0.777**	**0.31128**	**0.32649**	**0.32869**

Resultados

• No *dataset* Brown, o algoritmo Naive Bayes obteve os melhores resultados considerando todas as *features*. O melhor resultado geral, foi obtido pelo **descritor TF e classificador Naive Bayes, com acurácia de 0,52 e f1-score de 0,44**.
• Esse dataset possuía 15 classes desbalanceadas.

• No *dataset* Reuters, o método Doc2vec trouxe um pequeno ganho de acurácia, no geral o método MLP obteve os melhores resultados. O KNN se destacou com o Word2vec.
• **O melhor resultado geral, foi obtido pelo descritor Word2vec e classificador KNN, com acurácia de 0,80 e f1-score de 0,33**.
• Esse *dataset* era mais desafiante pois possuía uma quantidade maior de classes desbalanceadas (45).

Congresso Acadêmico UNIFESP 2020

Referências

Russel, S.; Norvig, P. Inteligência Artificial. Tradução de Regina Célia Simille de Macedo. 3a ed. Rio de Janeiro: Elsevier, 2013.

Mitchell, T. M. Machine Learning: Portland: Book News, 1997.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation arXiv preprint arXiv:1309.4168, 2013

Agradecimentos

Obrigado pela atenção!

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Dúvidas

c.gabriel.vieira@hotmail.com

lberton@unifesp.br