Talita Halboth Cunha Fernandes
GRR20165399
talitahalboth@gmail.com

Trabalho Prático 3
Inteliência Artificial 2020/2021

14 de fevereiro de 2021

1 Comparação da acurácia removendo palavras com determinada frequência

Foram realizados diversos teste removendo palavras do vocabulário de acordo com a frequência em que estes apareceram nos documentos utilizados. Foi usada apenas uma parte do dataset (os 1000 primeiros documentos) para que fosse viável realizar todos os testes em um tempo aceitável. Os testes realizados foram: com todas as palavras (sem remover stop words em inglês), removendo palavras que aparecem em 90% ou mais documentos, e da mesma forma, 80% e 70%. Também foram realizados testes removendo palavras que aparecem em 30%, 20% e 10% ou menos dos documentos. Cada teste foi repetido 30 vezes, para se encontrar um valor médio com significância estatística. Foram calculados os resultados de score, f1 score e perda (loss) médios. A seguir são demonstrados os resultados encontrados.

1.1 TF-IDF

Os testes descritos nessa seção foram os testes usando o modelo baseado em TFIDF.

1.1.1 Score

A imagem 1 apresenta o box-plot dos valores de score encontrado em cada teste realizado. Podemos ver que, removendo as palavras mais utilizadas, não houve tanta diferença no score em comparação com a remoção de plavras menos utilizadas.
Na tabela 1, encontram-se os valores de cáculo da estatística T student e do p-valor para teste de hipótese de igualdade entre a média do teste sem remoção de nenhuma palavra com os outros testes, considerando-se que as variâncias são diferentes. A média encontrada para o teste sem remoção de nenhuma palavra foi de 0.8166. Considerando um nível de significância de 1%, podemos assumir que a média para os testes removendo-se as palavras mais utilizadas tiveram, em média, o mesmo score que o teste sem remover nenhuma palavra. Já na questão de remover as palavras menos utilizadas, podemos afirmar com um alto grau de segurança que o score médio diminui se comparado com o teste em que não foram removidas palavras.

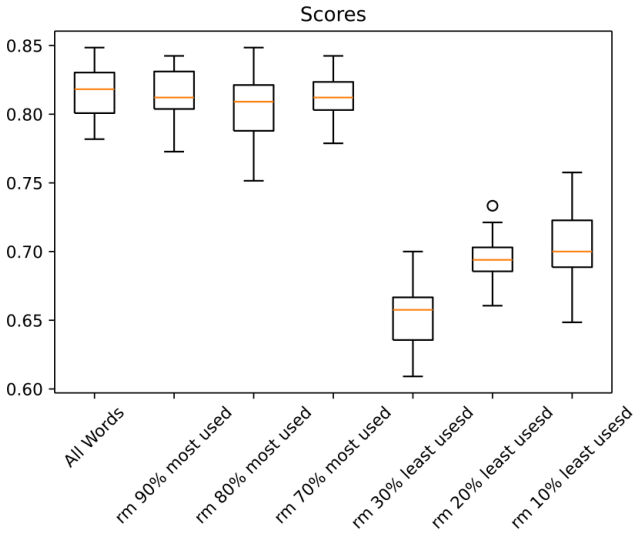

Figura 1: Boxplot de valores de score TF-IDF

1.1.2 F1 Score

A imagem 2 apresenta o box-plot dos valores de score encontrado em cada teste realizado, e a tabela 2 mostra os valores de média, estatística e p-valor como no caso do score. Neste caso, a média de F1 scores para o teste sem remover palavras foi de 0.8164. Os resultados são muito próximos dos valorse de score de cada teste, e, da mesma forma, podemos assumir o mesmo resultado.

Teste	Média	Estatística	p-valor
90% mais usadas	0.8149	0.3395	0.3677
80% mais usadas	0.8051	2.1099	0.0197
70% mais usadas	0.8112	1.1641	0.1246
30% menos usadas	0.6539	32.8763	0.0000
20% menos usadas	0.6949	26.2596	0.0000
10% menos usadas	0.7051	20.4905	0.0000

Tabela 1: Teste t-student para teste de hipótese do valor médio de score TF-IDF

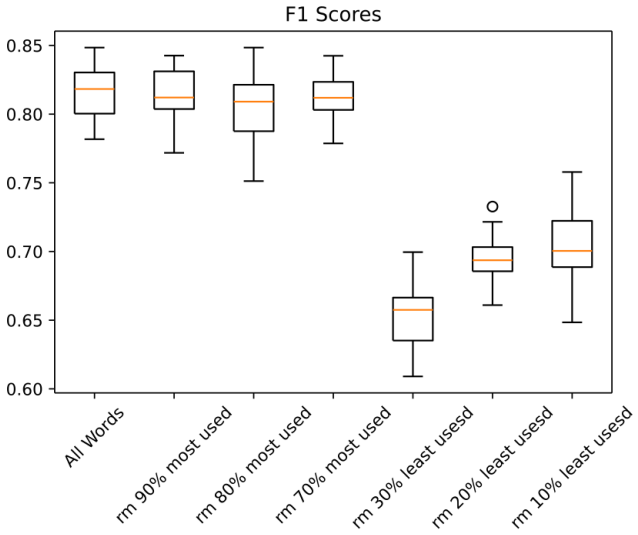

Figura 2: Boxplot de valores de F1 score TF-IDF

1.1.3 Loss

A figura 3 mostra os box-plots das médias encontradas para os valores de perda (loss) de cada teste, e a tabela 3 mostra as estatísticas e p-valores para teste de hipótese de igualdade de médias. A média encontrada no teste com todos os valores foi de 0.00472359. Podemos ver que em todos os casos, com significância de 1%, podemos rejeitar a hipótese de as médias serem iguais. Para os casos em que as palavras mais usadas foram removidas, o valor de perda foi menor que com todas as palavras, e já nos casos em que as palavras menos usadas foram removidas, a perda foi maior do que não removendo palavras.

1.2 Word2Vec

A seguir são mostrados os resultados paras os testes feitos usando o modelo Word2Vec.

1.2.1 Score

Da mesma forma que para o modelo TF-IDF, os valores de score removendo-se as palavas mais usadas foram muito próximos do valor sem remover palavra nenhuma (média de 0.7868), e são muito abaixo quando são removidas palavras pouco usadas. A figura 4 mostra os box-plots de cada teste e a tabela 4 mostra a estatística e os p-valores para o teste de hipótese correspondente.

1.2.2 F1 Score

A figura 5 mostra os box-plots do F1 score dos testes, e a tabela 5 mostra a estatística e os p-valores. A média encontrada para o F1 Score do modeo Word2Vec, sem remover palavras do vocabulário, foi de 0.7867.

1.2.3 Loss

A média encontrada para o valor de perda do modeo Word2Vec, sem remover palavras do vocabulário, foi de 0.01177075. Podem ser encontrados os box-plots do valor de perda para este modelo na figura 6, e a tabela 6 tem as estatísticas e p-valores.

Teste	Média	Estatística	p-valor
90% mais usadas	0.8149	0.3251	0.3731
80% mais usadas	0.8050	2.0981	0.0203
70% mais usadas	0.8112	1.1370	0.1301
30% menos usadas	0.6537	32.8594	0.0000
20% menos usadas	0.6949	26.2614	0.0000
10% menos usadas	0.7049	20.4662	0.0000

Tabela 2: Teste t-student para teste de hipótese do valor médio de F1 score TF-IDF

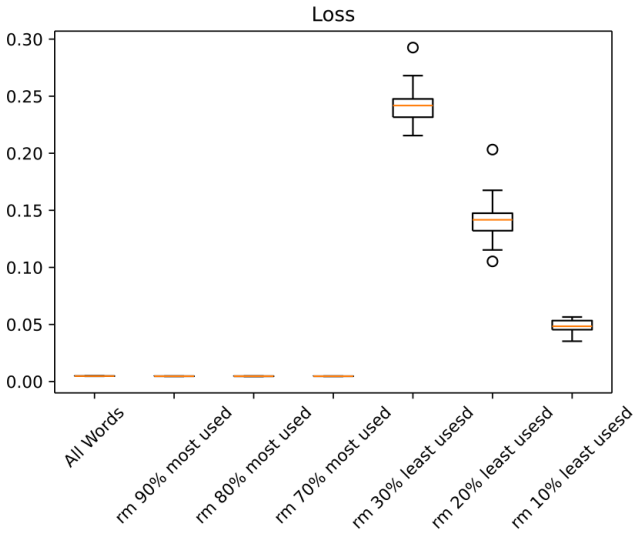

Figura 3: Boxplot de valores de loss TF-IDF

1.3 Conclusões

Com os resultados encontrados, podemos concluir que as palavras que aparecem com menos frequência são mais importantes para a detecção de sentimento em textos do que as palavras mais frequentes. Isto porque, na maioria das vezes, as palavras mais frequentes estão presentes em documentos com ambos os sentimentos, como “the“, “of“, “he“, etc. Estas palavras não são decisivas para a detecção do sentimento no texto. Mas palavras menos comuns, como “brutality“, “lonely“ e “hypnotizes“ são melhores em ajudar a descobrir o sentimento de uma frase.

2 Comparação da acurácia mudando parâmetros de min_count e window

Foram realizados testes mudando o valor do parâmetro min_count e window do modelo Word2Vec. Segundo a descrição da api da biblioteca Gensim, o parâmetro min_count ignora todas as palavras com frequência total menor que o valor passado. Já o parâmetro window é a maior distância entre a palavra atual e a predição de palavra em uma sentença[1].
Na figura 7 podemos ver os resultados do score obtido realizando testes com diferentes valores de min_count e window. O eixo X indica os valores de [min_count, window]. Uma coisa que podemos notar é que, no geral, valores window mais altos tendem a obter resultados melhores que os mais baixos. Já as médias dos valores de min_count ficaram menos organizadas, o que indica que este valor não foi tão significante quanto o valor de window para a acurácia do modelo. As figuras 10 e 11 mostram mais claramente a média dos valores de score separados entre window e min_count.
Este resultado nos mostra que, para os documentos em questão, é importante ter uma janela maior, pois com uma janela maior, é mais fácil para o modelo encontrar associações entre determinadas palavras. Para documentos com textos mais curtos, é provável que uma janela menor obtenha melhores resultados. Já no caso do min_count, em documentos mais curtos, o esperado é que quanto menor o valor, melhor o resultado. Isto porque para textos curtos, remover palavras no meio pode deixar quase impossível discernir o sentimento do texto. Já em textos maiores, podemos aumentar o valor de min_count com mais segurança, pode há mais contexto em cada documento para descobrir o sentimento.

Teste	Média	Estatística	p-valor
90% mais usadas	0.00460879	5.9162	0.0000
80% mais usadas	0.00460215	5.4046	0.0000
70% mais usadas	0.00460367	6.2090	0.0000
30% menos usadas	0.24115994	-84.3002	0.0000
20% menos usadas	0.14157398	-42.1991	0.0000
10% menos usadas	0.04847627	-43.7469	0.0000

Tabela 3: Teste t-student para teste de hipótese do valor médio de loss TF-IDF

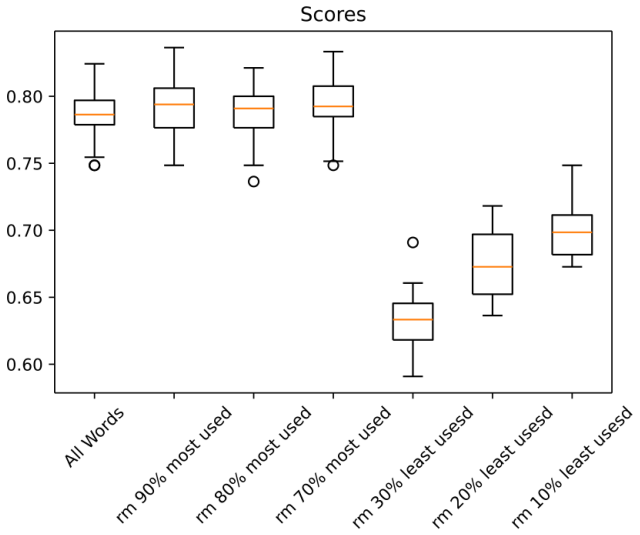

Figura 4: Boxplot de valores de score Word2Vec

Referências

[1] *Gensim Word2vec embeddings*. URL: https://radimrehurek.com/gensim/models/word2vec.html. (accessed: 13.02.2021).

Teste	Média	Estatística	p-valor
90% mais usadas	0.7921	-1.0081	0.1589
80% mais usadas	0.7887	-0.3966	0.3466
70% mais usadas	0.7935	-1.3286	0.0947
30% menos usadas	0.6322	30.3915	0.0000
20% menos usadas	0.6741	19.5935	0.0000
10% menos usadas	0.7017	16.3735	0.0000

Tabela 4: Teste t-student para teste de hipótese do valor médio de score Word2Vec

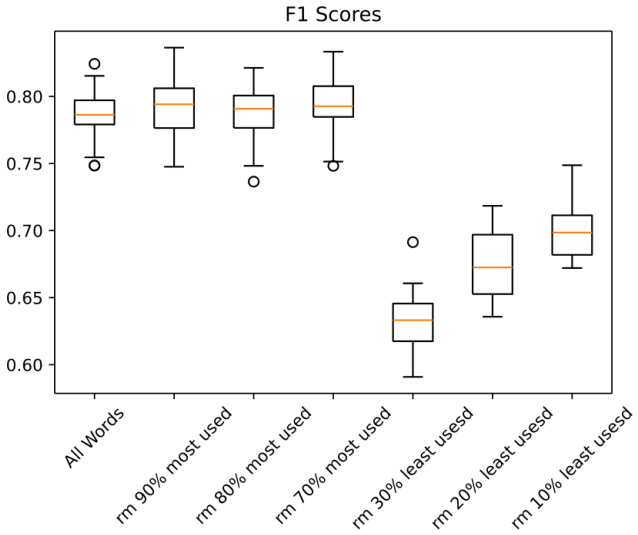

Figura 5: Boxplot de valores de F1 score Word2Vec

Teste	Média	Estatística	p-valor
90% mais usadas	0.7921	-1.0089	0.1587
80% mais usadas	0.7887	-0.4095	0.3419
70% mais usadas	0.7936	-1.3410	0.0926
30% menos usadas	0.6320	30.3058	0.0000
20% menos usadas	0.6740	19.5974	0.0000
10% menos usadas	0.7017	16.3293	0.0000

Tabela 5: Teste t-student para teste de hipótese do valor médio de f1 score Word2Vec

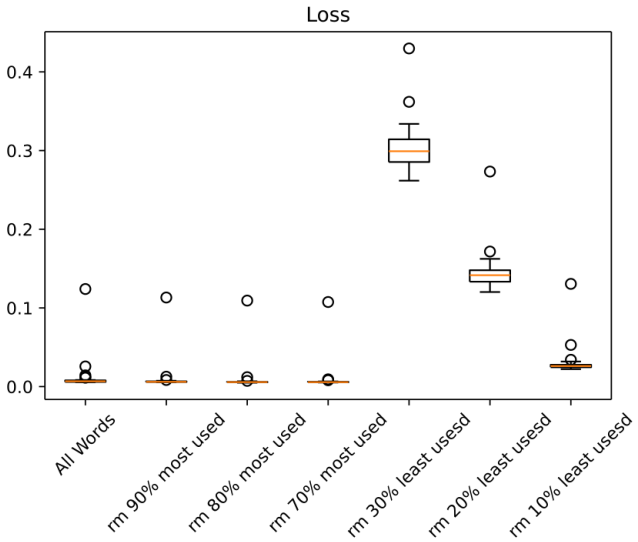

Figura 6: Boxplot de valores de loss Word2Vec

Teste	Média	Estatística	p-valor
90% mais usadas	0.00998256	0.3365	0.3688
80% mais usadas	0.00941490	0.4499	0.3272
70% mais usadas	0.00928824	0.4778	0.3173
30% menos usadas	0.30394078	-41.7365	0.0000
20% menos usadas	0.14581796	-21.3165	0.0000
10% menos usadas	0.03055162	-3.5220	0.0004

Tabela 6: Teste t-student para teste de hipótese do valor médio de f1 score Word2Vec

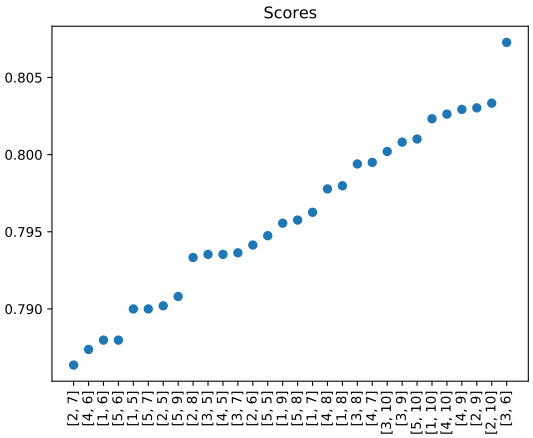

Figura 7: Sccater plot de valores de score variando min_count e window

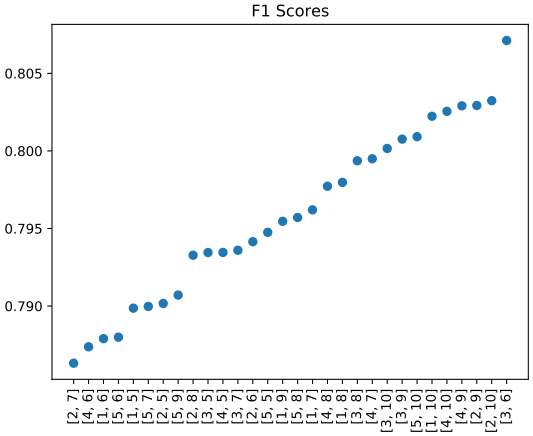

Figura 8: Sccater plot de valores de f1 Score variando min_count e window

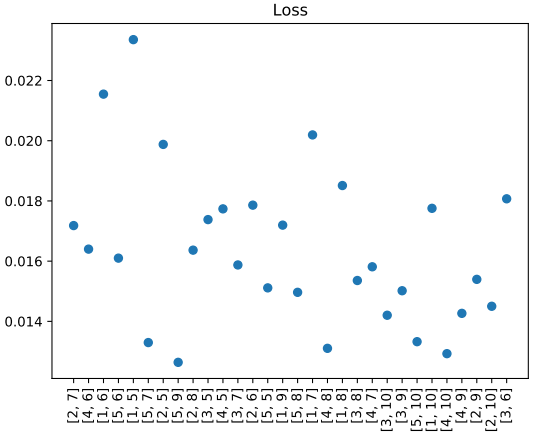

Figura 9: Sccater plot de valores de loss variando min_count e window

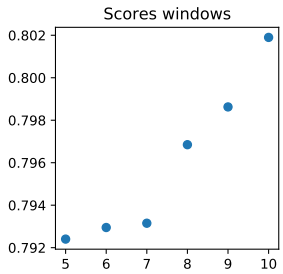

Figura 10: Sccater plot de valores de score variando window

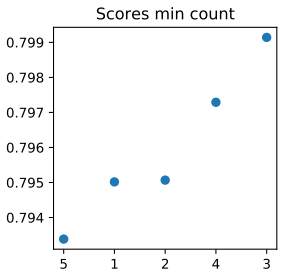

Figura 11: Sccater plot de valores de score variando min_count