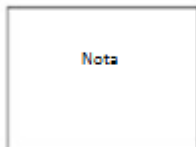




Centro Federal de Educação Tecnológica –
Trabalho final: Mineração de Textos
Prof. Gustavo Guedes



Aluno(a): _____

Turma: _____

Data: _____

Este teste é para ser codificado em Python, Java ou R.
LEIA AS QUESTÕES ATÉ O FINAL ANTES DE COMEÇAR.

O conjunto de dados rotulado **Sentence Polarity Dataset v1.0** (<https://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz>) possui 5331 sentenças positivas e 5331 sentenças negativas. Carregue os textos do conjunto em um DataFrame. Em seguida, você deve realizar duas tarefas:

- agrupar os textos utilizando o kMeans ($K=2$) e apresentar a nuvem de palavras de cada um dos grupos. Antes de agrupar, remova as stopwords (lembre-se que o conjunto é em português). Use WordCount, ou seja, TF.
- realizar a tarefa de classificação utilizando pelo menos 3 algoritmos e apresentando o F1 de cada um deles. Obrigatórios: Naive Bayes e kNN. O terceiro vocês podem escolher, pode ser o SVM. Utilize k-fold validation com 5 folds. Imprima na saída a média dos F1's dos 5 folds. Use o TFIDF.
- divida o conjunto de dados em treino e teste (80% para treino e 20% para teste). Treine o algoritmo kNN com os 80% de treino e teste com os 20% restantes. Imprima a matriz de confusão (tem função pronta para isso). Além disso, imprima a revocação, precisão e F1. Use WordCount, ou seja, TF.

Para submeter seu trabalho: faça download do arquivo ipynb no colab e suba no Teams. Caso seja em R ou Java, suba todos os arquivos necessários para a execução.

*obs: são disponibilizados no arquivo rt-polaritydata.tar.gz dois arquivos txt. Um .neg e um .pos, contendo sentenças negativas e positivas, respectivamente. Dentro do arquivo compactado existe um readme caso julgue relevante ler.