UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

3a lista de exercícios **22 de junho de 2020**

Instruções:

A lista deve ser respondida por grupos de até 3 pessoas (graduação) e individualmente (mestrado).
Resoluções idênticas de grupos distintos serão desconsideradas.
O código e demais dados devem ser anexados a cada questão.
Data limite para entrega: 03/07/2020.

Usando sua base de textos após os pré-processamento realizados na lista 2, realize as seguintes tarefas:

1. Escolha 10 documentos da base, e determine seu documento mais parecido (que não pode ser ele mesmo), usando:

a) Representação vetorial CountVectorizer com similaridade do cosseno;
b) Representação vetorial TF-IDF com similaridade do cosseno.

Mostre pelo menos um trecho de cada par de documentos parecidos que justifique essa semelhança.

2. Elabore um problema de classificação binária de textos coerente com sua base.

a) Rotule seus documentos manualmente (sugestão: separar em pastas distintas).
b) Extraia as representações vetoriais com CountVectorizer e TF-IDF.
c) Treine um classificador baseado em cada uma das duas representações vetoriais e Regressão Logística usando validação cruzada com 70% das amostras selecionadas para treino e 30% para teste. Exiba as matrizes de confusão, métricas de acurácia, precisão, recall e F1 score.
d) Faça o mesmo para o classificador Naive-Bayes.
e) Faça o mesmo para um outro classificador de sua preferência (pesquise na biblioteca Scikit-learn).
f) Compare os 6 resultados.