UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

2a lista de exercícios **28 de julho de 2023**

Instruções:
A lista deve ser respondida por grupos de até 2 pessoas.
Resoluções idênticas de grupos distintos serão desconsideradas.
O código e demais dados devem ser anexados a cada questão.
Data limite para entrega: 13/08/2023.

Nas questões abaixo, você não está sendo avaliado pela implementação, mas pela sua capacidade de análise dos métodos aplicados em sua base (escolha dos hiperparâmetros, discussão sobre os resultados dentre os diferentes métodos, etc.). Respostas somente com implementação não serão avaliadas.

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

a) Determine o rótulo dos documentos (separando os documentos em classes bem definidas).
b) Extraia as representações vetoriais com CountVectorizer e TF-IDF, considerando os textos já processados como na primeira lista.
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 o classificador SVM com kernel linear.
f) Compare os 6 resultados.

2. Realize um comparativo entre os métodos LDA, SVD e NMF para realizar modelagem de tópicos. Para cada método:

a) empiricamente, tente identificar uma quantidade de tópicos adequada para seu problema. Exiba resultados que justifiquem a quantidade de tópicos escolhida;

b) identifique as 5 palavras mais relevantes de cada tópico;
c) identifique os 3 tópicos mais relevantes de 5 documentos quaisquer (você pode representar os tópicos por suas 5 palavras mais relevantes).
d) qual método apresentou melhores resultados, na sua opinião? Justifique com resultados/dados.

3. Realize um agrupamento dos dados usando a representação do CountVectorizer, seguindo os seguintes passos:

a) Aplique o algoritmo PCA, preservando 95% da variância nos dados. Qual a dimensão resultante dos dados projetados?
b) Aplique um algoritmo k-means nos dados projetados, tentando usar o método *elbow* para encontrar o valor de k ótimo.

4. Nesta questão você deve aplicar métodos de projeção multidimensional para visualizar os dados da segunda questão no espaço visual.

a) aplique os métodos de projeção multidimensional t-SNE e UMAP na representação CountVectorizer e plote os gráfico das projeções resultantes, colorindo os pontos de acordo com os grupos obtidos pelo k-means (questão 3b).
b) Experimente variar os hiperparâmetros **perplexity** do t-SNE e **n_neighbors** do UMAP. O que acontece com as projeções quando estes parâmetros são calibrados para valores menores ou maiores do que seus valores padrão?
c) Usando os melhores valores encontrados para **perplexity** e **n_neighbors**, compare o tempo de execução e a qualidade visual da projeção dos 2 métodos de projeção multidimensional (t-SNE e UMAP). Qual abordagem se saiu melhor em cada um desses aspectos?