UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

2a lista de exercícios **31 de julho de 2025**

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: 21/08/2025.

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 de textos coerente com sua base, com pelo menos 3 classes.

a) Determine qual será 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 préprocessados como na primeira lista.
c) Treine um classificador baseado em cada uma das duas representações vetoriais, e escolhendo três dos seguintes classificadores: Regressão Logística, Naive Bayes, SVM (com kernel linear), k-NN ou Random Forest. Use 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 micro e macro.
d) Compare os resultados.

2. Você deverá aplicar dois métodos de modelagem de tópicos sobre uma base textual, utilizando a representação TF-IDF dos textos:

- um método clássico, à sua escolha entre LDA, NMF ou SVD truncado;
- e o método BERTopic.

a) Defina um número de tópicos apropriado para sua base e justifique a escolha com base em interpretação qualitativa ou métricas de coerência.
b) identifique as 5 palavras mais relevantes de cada tópico;
c) identifique o tópico mais relevante 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 agrupamento dos dados usando o algoritmo k-means, tentando encontrar um valor adequado para k. Use as seguintes representações vetoriais:

a) TF-IDF
b) Distribuição de tópicos obtida com o método clássico escolhido na questão 2.

4. Nesta questão você deve aplicar métodos de projeção multidimensional para visualizar os dados no espaço visual, usando os resultados da segunda e terceira questões.

a) aplique os métodos de projeção multidimensional t-SNE e UMAP na representação TF-IDF e plote os gráfico das projeções resultantes, colorindo os pontos de acordo com os grupos obtidos em 3a.
b) experimente variar os hiperparâmetros **perplexity** do t-SNE e **n_neighbors** do UMAP no experimento da questão 4a. O que acontece com as projeções quando estes parâmetros são calibrados para valores menores ou maiores do que seus valores padrão?
c) aplique os métodos de projeção multidimensional t-SNE e UMAP nos vetores usados na questão 3b. Encontre bons valores para os hiperparâmetros **perplexity** do t-SNE e **n_neighbors** do UMAP, e plote os gráfico das melhores projeções obtidas por cada método de projeção, colorindo os pontos de acordo com os grupos obtidos em 3b.