UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

4a lista de exercícios **13 de julho 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: 24/07/2020.

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

1. 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;
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?

2. Realize um agrupamento dos dados seguindo os seguintes passos:

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

3. Aplique projeção multidimensional para visualizar os dados no espaço visual comparando os métodos t-SNE e UMAP da seguinte maneira:

a) plote os gráfico das projeções resultantes, codificando as cores dos pontos de acordo com o grupo obtido na questão 2;
b) compare o tempo de execução dos 2 métodos.