UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

2a lista de exercícios **8 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: 17/06/2020.

Pesquise e escolha uma base contendo vários textos, em português ou inglês. Sugestão: procurar no Kaggle. Em seguida:

1. Determine a distribuição de comprimentos dos textos (em quantidade de caracteres), listando estas quantidades e plotando um histograma.

2. Aplique os seguintes passos de pré-processamento aos textos:

- Remova todas as palavras que contêm números;
- Converta as palavras para minúsculas;
- Remova pontuação;
- Tokenize os textos em palavras, gerando um dicionário único com n tokens e convertendo cada texto em um vetor de dimensão n com a respectiva contagem de palavras.

Em seguida, encontre as palavras mais frequentes.

3. Aplique os seguintes passos de pré-processamento aos textos processados na questão anterior:

- Remova *stopwords*;
- Realize rotulação de POS;

- Realize stemização;

Exiba os resultados em alguns textos. Verifique quais são as palavras mais frequentes e compare com as palavras mais frequentes da questão anterior.