**UNIVERSIDADE FEDERAL DE ALAGOAS**
**INSTITUTO DE COMPUTAÇÃO**

**Processamento de Linguagem Natural**
**Professor: Thales Vieira**

**5a lista de exercícios** **29 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: 09/08/2020.

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

**1.** O objetivo dessa questão é desenvolver um buscador de documentos.

a) escolha e aplique um modelo do tipo word2vec a seus textos, compatível com o idioma de seus textos (inglês ou português).
b) escolha 5 palavras de consulta que não estão em nenhum dos textos. Para cada palavra de consulta, encontre as 3 palavras **de seu conjunto de textos** mais parecidas com cada uma das palavras de consulta e exiba os documentos onde estas palavras aparecem.
c) usando as mesmas palavras do item acima, recupere os 3 documentos cujo word vector médio é mais próximo de cada palavra de consulta.
d) realize o procedimento acima usando três modelos com dimensão distinta.

**2.** Aplicando a representação Doc2Vec:

a) Resolva a segunda questão da 3a lista com esta nova representação e compare com os resultados obtidos anteriormente.
b) Resolva a segunda e terceira questão da 4a lista com esta nova representação e compare com os resultados obtidos anteriormente.