Rio de Janeiro, 28 de Agosto de 2017.
Trabalho 1
PROFESSOR: EDUARDO LABER

1. Entre em um portal de notícias e colete 3.000 documentos (notícias).

2. Construa um vocabulário a partir dos tokens dos documentos coletados e indique o seu tamanho.

3. Remova acentos, transforme todos os tokens em lower case e jogue fora stop-words. Reconstrua o vocabulário com base nos tokens normalizados e indique seu tamanho.

4. Gere estatísticas e/ou visualizações para entender a distribuição de frequência dos tokens e a distribuição do tamanho dos documentos em tokens.

5. Com base no novo vocabulário crie uma representação *bag of words* para cada um dos documentos.

6. Calcule o quadrado da distância Euclideana entre cada par de pontos (documentos) através da força bruta e messa o tempo computacional deste procedimento. Armazene estes valores. Utilize dois loops para fazer isso e implemente o cálculo da distância

7. Seja $N$ o número de tokens do vocabulário. Para $n = 4, 16, 64, 256, 1024, 4096, 4^{\lfloor \log_4 N \rfloor}$, repita o procedimento abaixo 30 vezes

   - Obtenha uma matriz aleatória de $n$ linhas e $d$ colunas pelo método de Achiloptas e pelo método baseado na dist. Gaussiana, onde $d$ é o tamanho do vocaulário.
   - Messa o tempo computacional da geração das matrizes
   - Projete os 3000 documentos no espaço $R^n$ através das matrizes geradas. Messa o tempo da projeção
   - Messa o tempo para obter todas as distâncias ao quadrado entre os pontos projetados
   - Calcule a distorsão máxima em relação aos dados originais.
   - Calcule o limite superior da distorsão previsto pelo Lema de J.L.

8. Escreva um relatório descrevendo os experimentos e os resultados obtidos. Analise se os resultados obtidos estão de acordo com a teoria apresentada. Considere a média, o máximo e o mínimo dos 30 experimentos do item 4.

OBSERVAÇÕES IMPORTANTES: A criação das matrizes, a projeção e o cálculo das distâncias entre os pares devem ser calculados utilizando loops e não funções prontas.