Trabalho Prático Wikipédia – Busca e Avaliação do Resultado

Relatório
O relatório deverá ser obrigatóriamente e Jupyter e deverá constar:

- Analise a ocorrência de termos no documento fazendo, no mínimo:
  - Quais são os 10 termos com maior e menor IDF da coleção? Com base nos termos de menor IDF, será que podiamos propor stopwords novas? Com base nesses termos, existe algo que você poderia melhorar no processamento?

  - Apresente o gráfico de frequência das palavras. Tais palavras devem ser ordenadas decrescentemente de acordo com a sua frequência. Note que a frequencia da palavra é a frequencia total do termo na coleção (Definido por F(termo) na aula sobre TF-IDF). De forma similar, faça o gráfico do IDF de cada palavra (veja o grafico similar na aula sobre TF-IDF). Mostre também, quais que tipos de palavras (específicas? Erros de processamento?) possuem TF alto, mediano e baixo. Faça o mesmo para IDF.

  - [opcional] Qual é o tamanho do vocabulario (número de palavras) quando usamos stemming e remoção de stopwords? e quando não usamos?
- Faça algumas consultas exemplos, mostrando o tempo de execução e o número de documentos retornados para cada consulta para cada modelo implementado.
- Apresentação e discussão dos gráficos de avaliação (precisão e revocação @5, @10, @25, @50 revocação @5, @10, @25, @50) do modelo vetorial.

Para plotar os graficos, você pode usar as bibliotecas seaborn ou matplotlib.