Organização e Recuperação da Informação

Trabalho prático 01

O algoritmo deve ser entregue até o dia às 23:59 do dia 05 de Dezembro. No início do código-fonte deve constar um comentário com o nome e a matrícula de todos os integrantes do grupo. Compactar os arquivos de código-fonte necessários em formato .zip. **Não utilizar .rar, .7z ou outros.** O nome do arquivo deve seguir o padrão **Trabalho_pratico_01-GrupoX.zip**, onde x corresponde ao número do grupo informado na tabela disponível no Microsoft Teams.

O objetivo deste trabalho é implementar, **em Python**, um sistema completo de indexação e recuperação de informação capaz de manipular dinamicamente uma coleção de documentos. O sistema deve construir e atualizar o vocabulário, a matriz TF-IDF e o índice invertido sempre que documentos forem **inseridos** ou **removidos**. As buscas devem contemplar *consultas booleanas*, *consultas por similaridade* e *consultas por frases*. Todo o código deve ser desenvolvido pelo grupo, com exceção da remoção de *stopwords* e da radicalização, que podem ser realizadas por meio de bibliotecas prontas..

O programa final deve exibir um menu que permita ao usuário inserir e remover documentos, consultar a coleção e visualizar tanto o índice invertido quanto a matriz TF-IDF. Todas as estruturas internas devem ser atualizadas automaticamente após qualquer alteração na coleção, garantindo que os resultados apresentados reflitam fielmente seu estado atual.

Funcionalidades obrigatórias

1) Leitura dos documentos pelo arquivo ‘colecao - trabalho 01.json’.
2) Remoção de *stopwords* e radicalização (bibliotecas podem ser utilizadas).
3) Construção e atualização do vocabulário da coleção.
4) Construção e atualização da matriz TF-IDF.
5) Construção e atualização do índice invertido.
6) Execução de consultas booleanas usando a matriz TF-IDF e exibição dos documentos que satisfazem a consulta.
7) Consultas por similaridade de cosseno utilizando o índice invertido, com exibição do ranqueamento dos documentos.
8) Buscas por frases utilizando o índice invertido, com exibição do ranqueamento dos documentos retornados.
9) Apresentação dos resultados de forma organizada (vocabulário, TF-IDF, índice invertido e resultados das consultas).

Interações do menu

O menu deve exibir opções para:

1) Adicionar **um documento por vez** à coleção (seguindo a ordem do JSON).
2) Adicionar **todos os documentos** da lista.
3) Remover um documento da coleção pelo seu identificador.
4) Exibir o vocabulário atualizado.
5) Exibir a matriz TF-IDF atual.
6) Exibir o índice invertido completo por **posição de palavras**.
7) Realizar consultas booleanas.
8) Realizar consultas por similaridade.
9) Realizar consultas por frase.
10) Executar quaisquer outras operações que o grupo considerar necessárias para o funcionamento do sistema.