Universidade Federal do Ceará – UFC
Centro de Ciências – CC
Mestrado e Doutorado em Ciências da Computação - MDCC
Estruturas de Dados

Exercício: Índice Invertido
Objetivos: Exercitar os conceitos de índice invertido.
Data da Entrega: 10/02/2025

OBS 1: Exercício Individual.
OBS 2: A entrega da lista deverá ser executada utilizando-se o SIGAA.

NOME: _____MATRÍCULA: _____

Questão 1

Crie um arquivo Jupyter Notebook e realize as seguintes operações:

a) Realizar o "restore" do arquivo (dump) denominado fd_whatsapp_0911_2023.zip no PostgreSQL. Esse arquivo está disponível no link a seguir: https://drive.google.com/drive/folders/1kEEnmZUVJEgYTynZjU6qMICbEVd8wKca?usp=sharing
b) Remova os trava-zaps.
c) Remover textos com menos de 5 palavras.

d) Implemente uma estrutura de Índice Invertido formado por um arquivo contendo o vocabulário e outro contendo a lista de ocorrências. Considere cada mensagem presente no conjunto de dados (ou seja, postada no WhatsApp) como um documento. Para a geração do vocabulário aplique os processos de remoção de "stop words" e "Stemming". Utilize o modelo TF-IDF.

e) Implemente um mecanismo de consulta que receba um conjunto de termos separados por ";" e retorna os "Top K" documentos que possuem esses termos.

Exemplo de Consultas:

- Dino
- Dino; Emendas
- Dino; Emendas; Lira
- Dino; Emendas; Lira; Centrão

OBS: lembre-se que será necessário aplicar os processos de "stop words" e "Stemming" aos termos recebidos como entrada antes de realizar a busca no arquivo de vocabulário.

f) Implemente uma interface gráfica simples que possibilite a um usuário executar consultas.

g) Para avaliar o desempenho do seu sistema de recuperação de textos utilize as seguintes métricas:
    - Precisão
    - Revocação
    - Precisão R
    - Medida-E

A avaliação deste trabalho se dará em duas etapas:

1ª. Vídeo de Apresentação: Cada estudante irá disponibilizar um vídeo (no Youtube) apresentando o todo o código gerado, bem como as ferramentas utilizadas. O estudante pode utilizar slides e notebooks na produção do vídeo.

2ª. Avaliação do Código: O professor da disciplina irá avaliar a qualidade dos códigos (notebooks) gerados pelo estudante, bem como a utilização das ferramentas utilizadas e as análises realizadas.

A avaliação do trabalho irá envolver os seguintes quesitos:

- Qualidade e organização do código (Notebook);
- Clareza da descrição das atividades realizadas e dos resultados obtidos;
- Domínio do Tema;

PS. Não serão aceitos trabalhos que não forem apresentados (por meio de vídeo disponibilizado no Youtube).

PS. Cada estudante será responsável pela disponibilização do ambiente (software e hardware) necessário para a gravação da apresentação do seu trabalho.

Os Notebooks, Arquivos CSVs, Dumps e URL do vídeo deverão ser disponibilizados, em formato .ZIP, no SIGAA. Caso o tamanho do arquivo .ZIP ultrapasse o limite máximo permitido pelo SIGAA, o estudante pode disponibilizar um link para um repositório no Google Drive.

"Se não posso estimular sonhos impossíveis, não devo negar o direito de sonhar com quem sonha.".

Paulo Freire