Universidade Federal de Uberlândia

Faculdade de Computação – Sistemas de Informação

FACOM32605 – Organização e Recuperação de Informação

Profa. Fernanda Maria – 2025/ 1º Sem

2ª TP – Tarefa de Programação 2

Objetivo: Construção e comparação de um modelo vetorial de RI

Data de entrega: 16/07/2025 até às 23:50hs.

Membros: Individual ou em dupla.

Faça os seguintes passos:

1) Crie uma coleção de documentos com as letras das músicas do top 100 da Billboard de 2023 (https://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_2024). Trabalhe com somente as 20 primeiras músicas. Cada uma das músicas deverá ser transformada um arquivo .txt.
2) Crie o vocabulário (termos de indexação) dos arquivos de texto removendo qualquer tipo de pontuação e considerando somente letras minúsculas. Qualquer caractere que não seja uma letra deve ser removido.
3) Usando uma linguagem de programação Python, implemente o modelo vetorial para encontrar os pesos TF-IDF para cada documento e também o grau de similaridade com relação a uma consulta que vocês irão testar.
4) Responda:
    a. Qual o tamanho do vocabulário da coleção (quantos termos)?
    b. Qual termo, em qual documento, possui o maior peso TF-IDF?
    c. Quanto tempo o programa demorou para criar o vocabulário da coleção de músicas?
    d. Quanto tempo o programa demorou para calcular o TF-IDF da coleção inteira?

Submissão: envie um .zip com as músicas, o código, um print do console com os valores TF-IDF da coleção (o print não precisa mostrar tudo!) e um PDF com as respostas do item 4).

Universidade Federal de Uberlândia

Faculdade de Computação – Sistemas de Informação

FACOM32605 – Organização e Recuperação de Informação

Profa. Fernanda Maria – 2025/ 1º Sem