Universidade Federal de Uberlândia - Campus Monte Carmelo
Faculdade de Computação - Sistemas de Informação
Prof. Dr. Murillo G. Carneiro
GSI521 - Organização e Recuperação da Informação
Data de entrega: 22/03/2022

Trabalho 2: Modelo vetorial

O presente trabalho tem por objetivo a construção de uma estrutura de índice invertido e o desenvolvimento de um sistema de ORI baseado no modelo vetorial. O trabalho compreende uma atividade de codificação e outra de escrita do relatório.

A **atividade de codificação** deverá contemplar os seguintes passos:

- Receber uma coleção de documentos (com pelo menos 5 documentos). O script deverá receber o diretório de uma pasta na qual conste o arquivo correspondente a cada documento.
- Como preparação dos documentos, realizar pelo menos a conversão das palavras para minúsculo bem como a remoção de *stopwords* e pontuações em cada documento. Vide arquivos anexo para relação de *stopwords* e pontuações a serem consideradas.
- Construção do índice invertido a partir da relação de pares (termo, documento). Note que a repetição de termos em um mesmo documento deverá ser suprimida do índice.
- Construir a representação vetorial correspondente a cada documento a partir da ponderação TF-IDF.
- Desenvolver um modelo vetorial para consultas. Esse modelo deverá receber uma chave de consulta (com um ou mais termos), representá-la como um vetor a partir da ponderação TF-IDF, calcular a similaridade do cosseno entre o vetor de consulta e aqueles referentes à coleção de documentos, ranquear os documentos em relação à similaridade com a consulta e retornar apenas os top-2 melhores resultados ao usuário.

O **relatório** deverá ter entre 5 e 10 páginas e sua escrita atender aos seguintes pontos:

- Visão geral, com uma breve explicação sobre o funcionamento e relevância da ponderação TF-IDF e do modelo vetorial.
- Algoritmo, com uma explicação detalhada de toda a sequência do código fonte e análise crítica da solução codificada.
- Conclusões, ressaltar os resultados obtidos bem como os principais pontos que chamaram a sua atenção na realização do trabalho.

O algoritmo pode ser codificado na linguagem de preferência do aluno. Para entrega, crie uma pasta que contenha o relatório e os fontes necessários para execução em ambiente linux/Fedora 33+, além de um arquivo README explicando passo a passo a instalação, compilação e execução dos fontes (incluindo a realizaçãoz de novas consultas). A pasta deverá ser nomeada pelo código de matrícula do aluno, compactada ".zip" e enviada até a data de entrega estipulada no cabeçalho.