Trabalho Prático da disciplina de Recuperação da Informação

Aluno: Marcio Palheta Piedade

O trabalho foi desenvolvido em C++, utilizando apenas bibliotecas nativas, como string.h e math.h

O projeto foi organizado em arquivos “.h“ e suas respectivas implementações “.cpp”.

A leitura do arquivo XML contendo todos os documentos a serem processados é feita a partir do arquivo “parser.cpp”, que recebe um arquivo XML e o converte para um arquivo .DAT, formatado de acordo com o projeto, removendo convertendo todos os caracteres para maiúsculo e removendo caracteres acentuados.

Quando do processamento dos documentos, é criado o vocabulário do projeto e armazenado em uma Arvore binária de busca e cada termo possui uma lista encadeada contendo todas as ocorrências dos termos em documentos da coleção. Nesse momento, o sistema remove as stopwords de cada documento.

Precisei me dedicar ao fechamento do trabalho de PPCC e não consegui concluir a parte de avaliação da máquina de buscas.

A seguir, os passos necessários para compilar e executar o projeto

1. Parser do XML

$ g++ core.cpp parser.cpp -o parser

$ ./parser origem.xml > destino.dat

2. Indexacao e processamento de consultas:

$ g++ core.cpp termo.cpp vocabulario.cpp documento.cpp indexador.cpp processadorconsulta.cpp -o processador

$ ./processador destino.dat