Relatório de Recuperação da Informação
Implementação do Modelo Vetorial

Nome: Brandell Cássio Corrêa Ferreira
Matrícula: 21453372

1 Objetivo do Trabalho

O objetivo deste trabalho prático é utilizar os conceitos aprendidos na matéria para implementar uma máquina de busca usando como *baseline* o modelo vetorial. A coleção de documentos utilizada foi a CFC (*Cystic Fibrosis Collection*), referente a relatórios médicos sobre uma doença genética chamada Fibróse Cística, causada por um gene defeituoso. Esta coleção possui 1.239 documentos, que foram publicados entre 1974 e 1979. Outro objetivo, foi aplicar métricas de avaliação aprendidas em sala de aula, como a **MAP**(Mean Average Precision) e **P@10**(Precision at 10), e visualizar o quão bom é o modelo vetorial aplicada a esta base.

2 Imlplementação

Para a implementação foi utilizada as lingugagens *C/C++*. Utilizei *C++* apenas para deixar o código organizado em classes, mas as funções, estão escritas basicamente em *C*. Como estrutura de dados para o vocabulário utilizei um *Hash*, já que o mesmo possui um acesso quase constante.

No vocabulário, cada palavra é acompanhada por um *id*, que é usado para a construção da lista invertida. A lista invertida é basciamente um vetor de listas encadeadas e em cada posição do vetor há :

- Lista invertida de Documentos (Com documento e TF (*Term Frequency*))
- Número de Documentos em que o termo aparece

- IDF *(Inverse Document Frequency)*

Criei uma estrutura chamada **Espaço Vetorial**, onde calculo o vetor para cada documento, onde a coordenada i é referente ao peso do termo i do vocabulário no documento.

Para fazer as consultas, criei uma estutura especifica que possui:

- As palavras da consulta
- Os documentos relavantes

Utilizei como método de ordenação o *Quick Sort* e para fazer a busca de documentos, utilizei a *busca Binária*.

Utilizei a busca para verificar as métricas. Já a ordenção, para gerar o rank.

3 Tutorial de Compilação e Execução

Para baixar o projeto basta dar clone no repositório que está no link abaixo: **https://github.com/brandellcassio/vetorial.git**.

Após isso, modifique o acesso ao arquivo **configure.sh**, e o execute :
$ chmod a+x configure.sh
$./configure.sh

Este script é responsável por fazer o build do projeto. Depois de executar, será criado dois scripts dentro da pasta build/: o *Indexer* (que indexa e cria os índices) e o *Reader* (Responsável por rodar as consultas). Para rodá-los corretamente você precisa indicar o caminho da coleção, acrescentando "cf" ao final do caminho. Por exemplo:
$ ./Indexer /caminho/para/cfc/cf

4 Resultados Obtidos

Após rodar o executável *./Reader* verifiquei que o tempo de execução varia entre *0.04s* e *0.11s* para as 100 consultas, e o resultados das métricas foram:

- **MAP Geral**: 0.269 ou 26,9%
- **P@10 Geral**: 0.456 ou 45,6%