UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

TÓPICOS ESPECIAIS EM RECUPERAÇÃO DE INFORMAÇÃO

Ministrada pelo professor DR. **David Braga Fernandes de Oliveira**

Aluno: Diego de Azevedo Barros – Matrícula: 2140176

Manaus – (2015) – AM

Objetivos principais na disciplina:

A ideia geral deste trabalho prático consiste em implementar o Modelo Vetorial para recuperação de informações. Para isso foi utilizado a linguagem de programação C++, que por ter uma rica biblioteca foi a escolhida para realização deste trabalho. Como estruturas para a implementação foi utilizado a STL (Standard Template Library) do C++ principalmente para armazenar coleções de dados, exemplos de containers utilizados: vector, map. O vocabulário e as listas invertidas da coleção forão armazenadas em um **map**<key, value>**,** sendo: <termo,<documento,frequencia>> hash_terms. O nome de todos os documentos da base de dados forão armazenadas em um **vector**<strings>.

Na implementação deste trabalho é carregado um arquivo contendo stop words, o qual é utilizado para remover da base palavras que não agregam valor em um documento, exemplo de stop words: a, e, com, portanto, conquanto, isso, esse, este etc. Também é carregado um arquivo contendo 50 consultas que serão avaliadas para esta base de dados. Estas consultas são passadas para a função ***calculate_similarity*** que calcula a similaridade entre os documentos para uma data consulta e mostra os 10 documentos com maior similaridade. Após isso é calculado a precisão para estes 10 documentos retornados comparando com uma lista de arquivos de documentos rotulados como relevantes. Ao final do processamento das 50 consultas juntamente com seus respectivos valores de precisão é calculado o **Map** (Média das Precisões Médias), afim de avaliar o comportamento dos algoritmos em relação à recuperação dos documentos rotulados como relevantes.

```
Query: Tênis Feminino
similaridades para consulta da query: 17
3464_tenis-floral-rosa_86036_301_1.jpg --- 0.00751926
3061_tenis-riana-branco-rosa_86001_301_1.jpg --- 0.00703431
3174_tenis-cassandra-branco_120962_301_1.jpg --- 0.00621304
5079_tenis-camila-branco-laranja_86023_301_2.jpg --- 0.00578911
3097_sapato-paula-floral_85566_301_1.jpg --- 0.00520195
3292_sapato-patricia-floral_85567_301_1.jpg --- 0.00516503
3648_tenis-feminino-branco-com-detalhes-pink_55426_301_1.jpg --- 0.00486406
3362_tenis-feminino-branco-com-detalhes-pink_55426_301_1.jpg --- 0.00486406
3078_sapato-tamires-floral_85560_301_1.jpg --- 0.00482525
2999_sapato-com-laco-azul_86008_301_1.jpg --- 0.00477451
Teste executado em: 8000 milisegundos
Precision: 0.6
```

Figura 1: Exemplo de 10 documentos retornados para uma dada consulta.

Neste exemplo, para a query: Tênis Femino forão retornados rotulados 6 documentos como relevantes, tendo uma precisão de 60%.

Especificação da Máquina utilizada nos experimentos:

Operating System	Linux Mint 17 Cinnamon 64-bit
Cinnamon Version	2.2.16
Linux Kernel	3.13.0-24-generic
Processor	Intel© Core™ i3 CPU M 370 @ 2.40GHz x 2
Memory	3.7 GiB
Hard Drive	476.7 GB
Graphics Card	Intel Corporation Core Processor Integrated Graphics Controller

Resultados:
A medida de avaliação utilizada foi a **Map** (Média das Precisões Médias), para uma quantidade de 50 consultas realizadas. O tempo de execução para cada consulta foi calculada em milisegundos.
O tempo de execução para cada consulta varia de 6000 a 9000 milisegundos.

Média das execuções das consultas = 7220.0 milisegundos
Map = (2.8 / 50) *100 = 5.6

Modo de Executar

Crie um diretório para o seu projeto, por exemplo: “Buscador”, coloque os arquivos dentro deste diretorio, todos na raíz: main.cpp, file_stop_words.txt, queries.txt, textDescDafitiPosthaus.txt e os arquivos numerados de 1 a 50.txt; estes arquivos são os documentos relevantes para cada consulta realizada. Certifique-se que tenha o compilador g++ instalado em sua máquina e execute o seguinte comando.

```
g++ main.cpp -o main
./main
```

Obs: caro queira dar saída em um arquivo, faça o seguinte:

```
g++ main.cpp -o main
./main >out.txt
```

Link do GitHub:
https://github.com/DiegodeAzevedoBarros/Vector-Model-in-Information-Retrieval