Recuperação de Informação
Trabalho Prático

Instituto Tecnológico Educacional da Amazônia (ITEAM) 09.03.2020

1 Visão geral do trabalho prático

Neste trabalho, o aluno deverá implementar uma máquina de busca para uma coleção de documentos. O modelo de RI a ser usado nesta implementação é o Modelo Vetorial. Abaixo, apresentamos uma breve descrição sobre a coleção de documentos a ser usada, bem como detalhes sobre o trabalho de implementação.

2 Coleção de documentos

A coleção de documentos a ser adotada neste trabalho será a CFC (Cystic Fibrosis Collection). Essa coleção consiste de 1.239 documentos publicados entre 1974 e 1979 sobre a doença genética Fibrose Cística, que é causada por um gene defeituoso que faz com que o corpo produza um líquido conhecido popularmente como muco, que se acumula nas passagens respiratórias dos pulmões e também no pâncreas. Essa coleção pode ser obtida através do seguinte link:

Download

cfc.zip

2.1 Descrição

A coleção CFC contém 6 arquivos de documentos, cada um contendo os artigos publicados em um dado ano (1974 a 1979). Cada artigo contém os seguintes campos:

- **Paper Number:** Número onde os dois primeiros dígitos contém o ano de publicação do artigo, e os três dígitos restantes contém um índice que varia de 1 até o número de artigos publicados naquele ano.
- **Record Number:** Número ID do artigo, variando de 1 até 1239.
- **Medline Acession Number:** A CFC é um subconjunto do banco de dados MEDLINE.
- **Author(s):** Autores do artigo.
- **Title:** Título do artigo.
- **Source:** Fontes de citações bibliográficas.

- **Major Subjects:** Contém os tópicos principais do artigo.
- **Minor Subjects:** Contém os tópicos secundários do artigo.
- **Abstract/Extract:** Contém o abstract do artigo (ou, as vezes, o texto completo).
- **References:** A lista completa de referências do artigo.
- **Citations:** Lista de artigos que citaram o presente artigo.

Além dos arquivos de artigos, a CFC contém um arquivo com 50 consultas e seus respectivos relevantes. As avaliações de relevância foram feitas por 4 pesquisadores diferentes. Desta forma, para cada par artigo/consulta, existem 4 avaliações diferentes. Cada avaliação foi feita através da atribuição de uma nota de relevância, que varia de 0 a 2, com os seguintes significados:

- **2:** Altamente relevante
- **1:** Pouco relevante
- **0:** Não relevante

Exemplo de resposta de um documento: 513 0010. Nesse exemplo, o record number de artigo é 513.

O primeiro, o segundo e o quarto pesquisadores deram nota 0 de relevância, indicando que o documento 513 é não relevante para a consulta avaliada. O terceiro pesquisador deu nota 1 de relevância, expressando a opinião de que o documento é mediamente relevante.

3 Descrição do trabalho de implementação

Este trabalho prático consiste na implementação de um sistema de recuperação de informação (RI), usando o modelo vetorial, para possibilitar buscas na coleção CFC. Esse sistema de RI deverá processar as 50 consultas da coleção, guardando as respostas retornadas pela máquina de busca. As respostas serão guardadas em função dos Record Numbers dos artigos de resposta.

4 Avaliando os resultados

Você deverá comparar os resultados retornados pelo seu sistema com a base de relevantes da coleção. As métricas MAP e P@10 deverão ser usados para comparar os resultados. Para este trabalho de implementação, deve-se considerar como relevante todos os documentos citados no arquivo de consultas, sem considerar a nota de relevância dos avaliadores.

5 Disponibilizando sua implementação para correção

Sua implementação deverá ser compartilhada através do GitHub (`https://github.com/`). Para corrigir seu trabalho, o professor irá clonar sua implementação através do comando git clone. Não é preciso disponibilizar os arquivos da coleção e nem arquivos de índice, uma vez que eles serão gerados durante a correção do trabalho. Desta forma, use o .gitignore para não versionar os índices e os arquivos da coleção.

6 Relatório do trabalho prático

Você deverá escrever um relatório sobre os resultados obtidos em seu trabalho prático. No entanto, esse relatório deverá possuir apenas 2 (duas) páginas, contendo as seguintes informações: seu nome e matrícula; uma breve descrição sobre os objetivos do trabalho; uma breve descrição sobre sua implementação (linguagem, bibliotecas, estruturas de dados); o link de sua implementação no GitHub; uma breve tutorial de compilação e execução de sua implementação; e uma breve descrição sobre os resultados obtidos (tempo de processsamento das 50 consultas e o MAP/P@10 obtidos nos experimentos).

7 Data de entrega

O trabalho deverá ser entregue até o dia 16 de Março de 2020.

8 Observações

1. Trabalho em dupla;
2. Os alunos deverão usar as linguagens C ou C++ durante suas implementações.