## Trabalho Prático de Recuperação de Informação

**Nome:** Roland Rodríguez Romero

**Matrícula:** 2150310

**Link no GitHub:** https://github.com/rolandr36/ri_project

## Objetivos do trabalho

O objetivo deste trabalho prático é a 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 100 consultas da coleção, guardando as respostas retornadas pela máquina de busca. O sistema será avaliado usando as métricas MAP e P@10.

## Descrição da implementação

O programa foi implementado em C++, foram usadas bibliotecas para o trabalho com arquivos e com containers como: unordered_map, vector e set. As estruturas de dados mais usadas foram os unordered_map, para guardar o vocabulário, as stopwords e as normas dos documentos (já que têm um tempo de aceso aos elementos de O(1)), o tipo vector, pra guardar as consultas e set pra guardar o conjunto dos documentos relevantes de cada consulta, utilizado no calculo das métricas.

No sistema foram implementados dois parser: um para o processamento da coleção de artigos e outro para o processamento das 100 consultas, o processamento é o reconhecimento do conteúdo dos arquivos e a importação pra o sistema.

Para a criação do vocabulário foram usados os campos: titulo, major subjects, minor subjects e abstract. Para diminuir o tamanho do vocabulário e ao mesmo tempo eliminar as palavras com pouco significado se usou um conjunto de stopwords, além disso se eliminaram as palavras de até duas letras.

## Compilação e execução

Para executar o programa primeiro é preciso compila-lo:

```
$ g++ main.cpp -o modvetorial
```

Logo pode ser executado,

```
$ ./modvetorial
```

Só é necessário rodar esse programa, já que ele processa a coleção, gera o arquivo invertido e depois processa as 100 consultas, dando uma saída resumida na tela e gerando um arquivo texto (results.txt) com as saídas mais detalhadas.

## Resultados obtidos

Os experimentos consistiram no processamento de 100 consultas na coleção e a avaliação foi feita usando o MAP e o P@10. Foram calculados o MAP e o P@10 para cada consulta e também o MAP e o P@10 geral (meia dos 100 valores individuais).

Nos resultados temos que o processamento de todas as 100 consultas (tendo em conta só o processamento das 100 consultas até obter o ranking resultado) é de 0,5 segundo. Enquanto a execução do programa completo, que inclui o processamento da coleção, geração do vocabulário e índice, processamento das 100 consultas, cálculo das métricas (MAP e P@10) e geração das saídas, tem um tempo de execução de aproximadamente 5 segundos.

As métricas usadas para a avaliação do sistema foram MAP e P@10. Analisando as consultas individualmente se aprecia que o valor do P@10 é muito variável, de 0 até 1, enquanto o P@10 meio pra as 100 consultas foi de 0,479. O MAP também teve resultados muito variáveis, de quase 0 até 0,72, mais o MAP meio para as consultas foi de 0,3539.

Observou-se que alguns documentos relevantes não foram retornados pela máquina de busca, foram analisados manualmente alguns desses casos, comprovando-se que embora os documentos fossem relevantes pra consulta, não compartilhavam palavras chaves com ela.