Relatório

Alunos: Jackson Lima e Mikael Souza

Componentes:

- extractor.py: Extrai a base do as arquivos cf74 - cf79
- cleardocs.py: Remove stopwords, pontuações e aplica stemming na base de dados
- clearqueries.py: Remove stopwords, pontuações e aplica stemming nas consultas
- metrics: Apresenta as métricas de avaliação MAP e NDCG
- heap.py: Implementa estrutura heap para seleção dos k primeiros resultados
- similarity.py: Implementa similaridade para pesquisa termo a termo
- tfidf.py: Implementa tf, idf e índice invertido

Como executar:

1. make build
2. make up

Se falhar certifique-se que o python 2.7 é o padrão no seu sistema operacional, se não será necessário substituir `python` pela váriavel de ambiente correspondente no seu sistema operacional para a versão 2.7 no arquivo `Makefile`.

Resultados:

Baseline:

- NDCG
  - Média: 0.45312 (45%)
  - Erro: 0.0785 (7%)
- MAP
  - Média: 0.2515 (25%)

Aplicando peso nas consultas:

- NDCG
  - Média: 0.4628 (46%)
  - Erro: 0.0746 (7%)
- MAP

- Média: 0.2732 (27%)

Entendimento

Observamos que as consultas tinham termos em comum, então nossa intuição foi que a importância desses termos é maior que as outras. Para dar maior importância utilizamos 80% da base de consultas e calculamos os termos mais frequentes, desses termos utilizamos 20% dos termos mais frequentes (por ordem de maior frequência). Para realizar as avaliações utilizamos 20% da base de consultas restantes.