Nome: Daniel Fernandes da Silva Pereira
Matrícula:21201910

Trabalho da disciplina Tópicos em Recuperação de Informação

Uma breve descrição sobre os objetivos do trabalho

O objeto do trabalho foi consolidar o conhecimento sobre o modelo vetorial. Modelo teórico que tem grande impacto na comunidade acadêmica e tem uma implementação relativamente fácil. Com a implementação é observável a teoria aplicada na prática e além do mais permite ver que simples atividades como tirar as stop words e uso do algoritmo de Porte realmente tem impacto na formação do índex e consequentemente na consulta em si.

Observações sobre a implementação

- Linguagem: Python na versão 2.7, o python foi escolhido por ser uma linguagem de script que geralmente apresentam vantagem em manipulação de texto e por poder usar com o paradigma orientado objeto.
- Bibliotecas:nltk, a biblioteca implementa o algoritmo de Porter
- Estrutura de dados: foi usado hash como estrutura padrão por ter busca constante e por ser nativo da linguagem

Descrição da implementação

A implementação desse trabalho teve um conjunto de classes que foram responsável por partes da execução do algoritmo.

Abaixo cito uma descrição de cada classe e suas principais funcionalidades:

- StopWord

O trabalho tem implementação própria de stop word. A classe lê um arquivo com as stop word e coloca dentro de um conjunto. Quando é usado a função clean é verificado se a lista de palavras enviada como parâmetro está dentro do conjunto de stop word se sim retira a palavra e senão o retorna a própria palavra.

- ArticleBase
  Classe responsável por fazer o parser de um arquivo e instanciar os objeto do tipo Article.

- Article
  Classe que descreve as informações que um artigo tem e tem as funções do parser de como cada campo do article irá ser tratado.

- QueryBase
  Classe responsável por fazer o parser de um arquivo e instanciar um objeto do tipo Article.
- Query
  Classe que descreve as informações que um query tem e as funções do parser de como cada campo da query irá ser tratado. Nesta classe é implementado as funções de idf e tf e cálculo das métricas.

- Vocabulary
  Classe que tem o índice. O índice é formado por palavras e cada palavra tem InvertedList.
- InvertedList
  É implementado um dicionario que associa a quantidade de ocorrência de determinada palavra em um artigo. Nesta classe é implementado as funções de idf e tf.

O link de sua implementação no GitHub:

https://github.com/daniel12fsp/trab_modelo_vetorial.git

Execução de sua implementação

É necessário que se crie o diretório chamado base, aonde ficaram os arquivos usados para base, e outro chamado query, que contém o arquivo de query(No git já foi adicionado os arquivos, logo depois do git clone será necessario somente executar conforme descrito abaixo). Não é necessário passar nenhum parâmento. O algoritmo vasculha esses diretório para entrada de dados.
Para executar você basta digitar: python2 vector_model-Daniel.py

Os resultados

O experimento teve como resultado: p@10 com 46,4% e map com 28,75% mostrando que apesar do uso de remoção de stop word e usando algoritmo de Porter é necessário o uso de outros passos intermediários para ter uma métrica de precisão melhor