Terrier

Recuperação da Informação 2019.1

Thiago Coelho

Introdução e Objetivo

Terrier é um sistema de Recuperação da Informação *open source* escrito em Java. É desenvolvido pela *School of Computing Science, University of Glasgow*.

O objetivo do trabalho foi utilizar essa ferramenta para a indexação e recuperação de documentos de uma coleção de teste.

Foi utilizada a versão v5.1 do Terrier e o sistema operacional Windows 10.

Pré-processamento

Foi aplicado um pré-processamento para a remoção de caracteres especiais de todos os arquivos.

O script utilizado para a remoção está em https://github.com/Thiagohnc/rec-info/blob/master/remove_char_especial.py

Indexação e Configuração

- *bin\trec_setup.bat <caminho-para-a-pasta-com-os-documentos>*

- *collection.spec*
  - É possível escolher o que não deve ser indexado

- *terrier.properties*
  - É possível customizar algumas propriedades
  - *Stemmer* (*termpipelines*)
  - *Stopwords* (*stopwords.filename*)
  - Método de Ponderação (*trec.model*)

- *bin\terrier batchindexing*

Consultas

Pode-se realizar consultas através de um arquivo *xml*.

Um exemplo de consulta é o seguinte:

<top>
<num> 1 </num>
<PT-title> consulta </PT-title>
</top>

bin\terrier batchretrieval -t <caminho-para-o-arquivo-com-a-consulta>

Os resultados estarão em var/results, dentro da pasta do Terrier.

Consultas - Sintaxe

- termo1 termo2
  - Documentos com termo1 ou termo2
- +termo1 +termo2
  - Documentos com termo1 e termo2
- +termo1 -termo2
  - Documentos com termo1, mas sem termo2
- termo1^2.3
  - Multiplica o peso de termo1 por 2.3

Consultas - Sintaxe

- “termo1 termo2” ou +”termo1 termo2”
  - Documentos em que termo1 e termo2 apareçam em uma frase
  - Útil para a consulta 3 (documentos sobre filmes cuja história se passa na Escócia)
  - A consulta foi +filme +”passa Escocia”
  - A intenção foi pegar documentos que falassem algo como “O filme se passa na Escocia”
  - De fato, o primeiro resultado possui a frase *“O diretor começou a rodar o filme na Escócia, onde se passa a história, e atualmente comanda a sua fase final na Irlanda.”*

Consultas Interativas

bin\terrier interactive

Com esse comando é possível realizar consultas rápidas direto do terminal.

```
terrier query> opio^3 papoila^3 papoula^3 cultivo producao
```

```
        Displaying 1-1000 results
0 FSP941225-037 18.556451475326156
1 FSP940515-058 18.15219188610657
2 FSP940227-120 16.46837716203282
3 FSP950319-071 12.018673920764773
4 FSP940227-119 11.352938388183675
5 FSP951224-031 11.238244452485201
6 FSP940904-101 11.189106892486905
7 FSP940920-078 9.61184101682344
8 FSP941023-075 9.538661199336518
```

Integração com Aplicação Java

Um dos casos de uso comuns para a ferramenta é como um componente de busca dentro de uma aplicação grande.

É possível importar bibliotecas do Terrier para fazer indexação e recuperação de documentos dentro de uma aplicação java.

Essa funcionalidade não foi explorada para o trabalho.

Referências

- http://terrier.org