Documentation IR_P01

Team: “Marcel”

Our implementation utilizes the Apache Lucene Java Library for most of the IR related functions for the given task including analysing the given documents (tokenization) and ranking.

File input is mostly handled by the jSoup HTML manipulation library for Java.

According to the task we are using the Lucene English analyzer to tokenize the HTML body texts.

Major library functions and classes in use:

- **IndexWriterConfig, IndexWriter and IndexSearcher** : setSimilarity() to switch between ranking models ;

- **EnglishAnalyzer** : tokenStream(), stop word list and stemming information used for TokenStream and IndexWriterConfig

- **Directory & DirectoryReader** : IO support. File location management and sub folder support.

- **Document, TextField, Elements** : Process given HTML file and filter out information like title, date and body text with the help of TextFields.

- **TokenStream & StringReader** : Use EnglishAnalyzers tokenStream() function to stream tokens to tokenized String list

- **MultiFieldQueryParser** : Used to include all fields in index search (Titles, body, Date in our case)

- **ScoreDoc** : Contains documents with similarity scores based on user query

Our default search index weighting values are as follows:

Title: 0.5

Body: 2.0

Date: 0.5

Please consult the source code documentation inside the source files for further information.