Documentation Programming Task IR

Team Badass Mofos

Members: Michael Thane M.Nr.:

Theodora Kyriakou M.Nr.:

Sebastian Hennig M.Nr.: 221088

We will first describe the general Workflow and then go into detail for each step.

# General Workflow

First we read the all relevant files from the Folder that is parsed via the command line. Then these files are parsed into **Doc objects.** These Objects contain all information to create the Index. We then check if there is already an index stored in the folder which was also passed via command line. If there is we load that index and if not we create a new index out of these Doc objects. After that it is decided which ranking model is used also dependent on the user input. When all these parameters are set we query the index with the last parameter that is parsed via the command line and print the results.

## Getting Relevant Files

We first get all relevant files (txt and html) from the folder provided in the command line args. We do this using a recursive folder traversal function which reads in all files in a folder when they have the right format and then proceeds to go to any subfolders and do the same thing again until there are no more subfolders left and returns all files found.

## Extracting relevant information

In the next step we extract the important information of the files into Objects. These are named **Docs.** A Doc has the following attributes: title, content and path. We create these Objects using a Factory. The getInstance() method of our factory takes a file path as input and returns a Doc object of that file. To successfully manage this step the files first need to parsed depending on their format.

## Parsing

Text:

@TODO Michael

HTML:

@TODO Michael

Jsoup Version:

Once we parsed all the files into our Docs we can use these to create the index.

## Indexing

@TODO Theodora

Lucene Version: 6.6.0

Lucene Analyzer Commons: 6.6.0

Lucene Query Parser: 6.6.0

## Ranking Model

@TODO Theodora

## Query

@TODO Theodora