Students have to solve the programming tasks and get at least 50% of all points in order to pass this lecture.

Programming Task P01

Program your own (command-line based) Information Retrieval system using *Apache Lucene*[1] (at least version 3.6, currently the newest version is 6.2.1). *Lucene* is an open source search library that provides standard functionality for analyzing, indexing, and searching text-based documents. The following criteria have to be met by your Information Retrieval system. Your program shoud ...

- Using Lucene, parse and index *Plain Text* and *HTML* documents that a given folder and its subfolders contain. List all parsed files.
- Consider the English language and use a stemmer for it (e.g. Porter Stemmer).
- Select an available search index or create a new one.
- Print a ranked list of relevant articles given a search query. The output should contain the most relevant documents, their rank, path, last modification time, relevance score and in addition for *HTML* documents title and summary.
- Search multiple fields concurrently: not only search the document's text (body tag), but also its title and date (for *HTML* documents).

The program should be written in a way that it is runnable without taking any look in the source code or even adapting the source code. Create a jar-File named IR_P01.jar. It should process the input:

```
java -jar IR_P01.jar [path_to_document_folder]
```

Please send your solutions (jar-File AND source code) via e-mail until 01. Dec. 2016, 08:00 a.m. to afraa.ahmad-alyosef@ovgu.de.

(10 points)

[1] http://lucene.apache.org/java/