TDT4117 Information Retrieval - Autumn 2016
Assignment 4

Deadline for delivery in ITS learning is 13.11.2016

Task 1 : Page rank and HITS

- Compare page rank and HITS and briefly describe the main ideas of both approaches and point out their differences.
- Given the graph below, compute hub and authority scores for webpages labeled as A, B, C and D using HITS algorithm. Perform at least 3 iterations of the algorithm and illustrate your computations by providing formulas filled with values for at least one iteration.

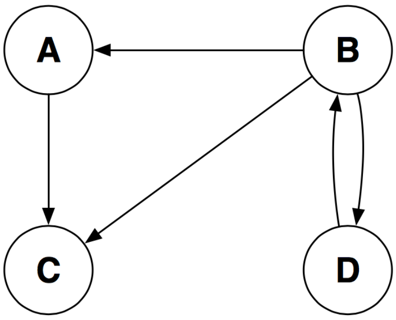

Figure 1: Graph of websites connected by links.

Task 2 : Structured Indexing and Retrieval in Lucene

Throughout this assignment, you will work with a subset of the 20 news collection (1907 out of around 18.500 e-mails posted online), which you can download from its learning ('20news-part.zip').

The java files and libraries ('ir4-sample.zip') come with a class, which you should complete, that is 'MyDocument'. There is an empty method that you are supposed to implement in the course of this assignment.

Make sure to import the following libraries to your project:

- lucene-core-4.10.1.jar
- lucene-demo-4.10.1.jar
- lucene-analyzers-common-4.10.1.jar

Subtask A

Lucene offers the possibility of indexing a document in several fields (i.e. subject, body, from). If a collection is indexed in this way it becomes easy to search across fields or in specific fields only.

The task is to update the given ‘MyDocument’ class and implement the ‘Document(File f)’ method to index the following fields per document:

- id: the name of the file.
- from: whatever is stored in the from field of the given message.
- subject: the subject of the e-mail.
- contents: the actual e-mail contents.

All fields ‘from’, ‘subject’, and ‘contents’ should be searchable, i.e. store their respective term vectors. Look at the given ‘NewsDocument’ class that reads a document and returns the texts for ‘from’, ‘subject’, and ‘contents’, and ‘id’.

Subtask B

One tool to manipulate Lucene indices is the Lucene Index Toolbox or Luke, see `https://github.com/DmitryKey/luke/releases/`. In this task, you have to use Luke by downloading the jar from:

```
https://github.com/DmitryKey/luke/releases/tag/luke-4.10.1
```

and opening it by typing ‘java -XX:MaxPermSize=512m -jar luke-with-deps.jar’. Once you start the application, choose the path to the index directory of subtask A. Use Luke to search for the term ‘Vancouver’ in different fields. Use Luke’s search tab and set the analyzer to ‘org.apache.lucene.analysis.standard.StandardAnalyzer’ and select different fields. Show screenshots of the results and explain the behavior of the system.

Notes

Possibility to work in group.
Submit the source code of ‘MyDocument’ class together with your report.
Deliveries must be typed and in PDF format.