Exercise 6

Web Mining

Time for completion: 2014-11-19 until 2014-12-03 (2 am at the latest)

1. Generate your own text-corpus, using the web-crawler implemented in exercise 3. Start with following link as initial seed:
   http://en.wikipedia.org/wiki/Data_mining.
   Use the IR-System from exercise 5 to transfer the web-sites into the internal representation. Don't forget to use an english stemmer.

   5 points

2. Extend your IR-System, that it provides three ranking algorithms for searching documents:
   a. Boolean Retrieval
   b. Vector Space retrieval
   c. Vector Space retrieval, regarding HTML markup
   d. Page Rank algorithm

   15 points

3. Implement a website, where querys can be entered and the ranking algorithm can be choosen. The results have to be ordered according to the used algorithm. Provide a choice method to specify the amount of returned result for the query at the web-site.

   5 points

4. Test your system. Formulate 20 querys, each with at least 3 words. Generate a test dataset. Evaluate the four algorithms according to the Precision, Recall and F1-Metric for different size of the result set (5,10…. 100)

   15 points

General Principles

Hand over your exercise in time via mail to harasic@inf.fu-berlin.de . Use the subject [WBI-1415]-exercise6-groupXY (replace XY with your group number).

If an exercise results pieces of software you have to hand over one **ZIP**-file containing:

- the *well-documented* sources
- a maven project containing its pom.xml
- a documentation how to compile and run the software

Software has to be implemented in Java conforming to version 1.7. The reference system where each application should run is the VM found under following link:

https://www.csw.inf.fu-berlin.de/teaching/ws1415/wbi/WBI.tar.gz

Username: ws1415
Password: wbiWS1415

If it is not possible to run your application with the predefined command line we will rate this exercise with 0 points. I will not debug your code.

Any documentation (not the documentation of source code) and written answers should be included as a PDF document within the above mentioned ZIP-file.

An exercise is passed if you reach 60% of the maximum number of points.

Contact person regarding the WBI exercises:
Marko Harasic
Mail: harasic@inf.fu-berlin.de