Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Patrick Brosi

Information Retrieval
WS 2016/2017

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 10

Submit until Tuesday, January 17 at **2:00pm**

Exercise 1 (20 points)

Implement Latent Semantic Indexing (LSI) using the following functionality and indexing:

1. Copy your code from ES8 (or from the master solutions for that sheet if you prefer) to a new folder *sheet-10*.

2. (8 points) Extend your code by a method *process_query_lsi* that computes results by projecting both the query vector and the term-document matrix into the latent space (Variant 2 from the lecture). The necessary preprocessing (in particular, the computation of the term-document matrix and the computation of the SVD) should be done in two methods *preprocess_vsm* (re-use from ES 8) and *preprocess_lsi*. For your term-document matrix only use the m most frequent terms to keep the matrix small. The dimension k of the latent space and the number m of terms should be second and third command-line arguments (after the file name).

3. (4 points) Repeat the evaluation of your system on the benchmark from ES2, without LSI (using your code from Exercise 1), with LSI (using your code from Exercise 2), and using a linear combination of the two (with parameter λ, as explained in the lecture). Play around with the parameters m, k and λ and post your best results in the result table on the Wiki. Leave the BM25 parameters constant for the three runs.

4. (8 points) Add a method *related_term_pairs* that computes the term-term association matrix T explained in the lecture (based on the SVD), and returns the first 100 term pairs in value-sorted order (term pairs with highest values first, ignoring pairs of equal terms). Write these term pairs to a file *term_pairs.txt* (one term pair + value in T per line) and commit that file to the SVN as part of your solution. Briefly(!) discuss in your *experiences.txt* why you think that LSI "found" some of these term pairs.

[please turn over with left eigenvector]

Add your code and your file *term_pairs.txt* from Exercise 4 to a new sub-directory *sheet-10* of your folder in the course SVN, and commit it. As usual, make sure that *compile*, *test*, and *checkstyle* run through without errors on Jenkins. And let us partake in your adventures for this sheet in the usual *experiences.txt*.

How to go to bed early?