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 8

Submit until Tuesday, December 20 at **2:00pm**

Exercise 1 (10 points)

Copy your code from Exercise Sheet 2 (or from the master solution for that sheet if you prefer) to a new folder *sheet-08*. Extend your code by a method *preprocessing_vsm* that builds the term-document matrix from the inverted index (using BM25 scores as entries).

Then add a method *process_query_vsm* such that the result list is obtained via a multiplication of the term-document matrix with the query vector (and not via merging of the inverted lists, like in Exercise Sheet 2).

Note 1: You can use the code from the lecture as an orientation. The only difference is that you should use BM25 scores, whereas in the lecture we used simple tf scores.

Note 2: For debugging, it might be useful to keep your old *process_query* and *merge* methods around. However, please remove them in the final version of your submission: they will only make it harder for your tutor to understand your code and give you meaningful feedback.

Exercise 2 (10 points)

Re-run the evaluation of your system on the benchmark from Exercise Sheet 2, both with the original *process_query* (using the inverted index) and with the new *process_query_vsm* (using the term-document matrix). Make sure that the results are identical.

Re-run the evaluation with the columns of the term-document matrix normalized with respect to the L2-norm (that is, the sum of the squares of the entries of a column should sum to 1). Do this with BM25 scores, with ordinary tf.idf scores, and with tf scores. Which combination of score type and normalization or not gives the best result?

Add your code to a new sub-directory *sheet-08* of your folder in the course SVN, and commit it. Make sure that *compile*, *test*, and *checkstyle* run through without errors on Jenkins. Also commit the usual *experiences.txt* with your much appreciated brief and concise feedback.

What is more real and why: consciousness or matter?