Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Björn Buchhold

Information Retrieval
WS 2013/2014

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

UNI FREIBURG

Exercise Sheet 8

Submit until Tuesday, December 17 at 4:00pm

Exercise 1 (5 points)

Copy a suitable version of your class *InvertedIndex* from one of the first exercise sheets (or from the master solution if you prefer) to a new folder *exercise-sheet-08*. Extend the class by a method *writeTermDocumentMatrix(String name, int m)* that writes the term-document matrix for the m most frequent terms (that is, those with the m longest inverted lists) to a file *<name>.matrix*. For the entries of the matrix, use BM25 scores with the default settings for b and k. Write in sparse-matrix format, that is, for each non-zero entry in the matrix, write a line of the form: *<row-index> <column-index> <value>*, with the three numbers separated by spaces. Also write to a separate file *<name>.terms* the words corresponding to the m most frequent terms.

Exercise 2 (5 points)

Write an Octave script *readTermDocumentMatrix.m* that reads files *<name>.matrix* and *<name>.terms*, as produced by your code of Exercise 1 above, into Octave.

Exercise 3 (5 points)

Write an Octave script *computeMostRelatedTerms.m* that computes the ids of the 100 most related terms pairs as follows. For a given term-document matrix A (as produced by the script from Exercise 2) and a given k, compute the term-term association matrix $T = U_k \cdot U_k^T$, where U_k is the matrix consisting of the first k columns from the U of the singular value decomposition $A = U \cdot S \cdot V^T$. Then determine the 100 largest entries of T, ignoring the entries on the diagonal.

Exercise 4 (5 points)

Write an Octave script *writeRelatedTermsToFile.m* that writes the terms (not the ids) from what was computed in Exercise 3 to a file *<name>.<k>.term-pairs*. Run all your scripts for $m = 1000$ and different values of k, namely 5, 10, and 50. The three resulting *.term-pairs* files should also be committed to the SVN. Briefly discuss their contents in your *experiences.txt* for this exercise sheet. In particular, explain why you think you obtained some of the term pairs you did, and argue which value of k you think gives you the most meaningful results for this data set.

[please turn over]

Add your code and your Octave scripts to a new sub-directory *exercise-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. As usual, also commit a text file *experiences.txt* where you briefly describe your experiences with this exercise sheet and the corresponding lecture. As a minimum, say how much time you invested and if you had major problems, and if yes, where. Don't forget the brief(!) discussion asked for in Exercise 4 above.