Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Florian Bäurle

Information Retrieval
WS 2012/2013

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

UNI FREIBURG

Exercise Sheet 8

Submit until Wednesday, December 19 at 4:00pm

Exercise 1 (6 points)

Extend your class *InvertedIndex* (copy your best version from a previous exercise sheet or the master solution) by a method *writeTermDocumentMatrix(String name, int m)* that writes the term-document matrix for the m most frequent terms to a file *<name>.matrix*.

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, in a separate file *<name>.terms*, the words corresponding to the m most frequent terms.

Exercise 2 (8 points)

Write an Octave script that, for a given k, computes the truncated singular value decomposition $U_k \cdot \Sigma_k \cdot V_k^T$ of your term-document matrix from Exercise 1. Use *svds(A, k)*, which is for sparse matrices, not *svd*, which is for dense matrices; see slide 21 from the lecture. Then compute the term-term association matrix $T = U_k \cdot U_k^T$, and output those 100 pairs of (different) terms with the largest entry in T. Output the names of the terms, not just the term indices (that's what the file *<name>.terms* from above is for).

Exercise 3 (6 points)

Run your script from Exercise 2 with $m = 1000$ and $k = 10, 50, 100, 500$. Write the output (100 lines per file) into text files *term-pairs.k<k>.txt*, which you should also commit to our SVN. In your *experiences.txt*, briefly discuss these results. In particular, say for which value of k you get the most meaningful results, how meaningful those term-pair associations are, and why you think they received a high score in T.

Commit your code to our SVN, in a new sub-directory *exercise-sheet-08*, and as usually make sure that Checkstyle and Tests run through without errors on Jenkins. Also commit a text file *experiences.txt* with your feedback. As a minimum, say how much time you invested and if you had major problems, and if yes, where.