Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Marjan Celikik

Search Engines WS 09/10

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

UNI FREIBURG

Exercise Sheet 3

complete until Tuesday, November 10th

Exercise 1

Consider a query and a set of documents in the vector space model, with all vectors normalized. (1) Prove that ranking the documents by euclidean distance to the query vector gives the same result as ranking the documents by cosine similarity with the query vector. Provide both a mathematically rigorous proof, as well as an intuitive explanation (one or two sentences). (2) Does this statement remain true when the query vector is not normalized? Why or why not? (3) Show, by a counterexample, that the statement is not true in general if the document vectors are not normalized.

Exercise 2

Extend your parsing, index construction, and query processing code from Exercise Sheet 1 to incorporate a score for each word occurrence, as explained in the lecture. The query processor should rank hits by these scores. Pick one of the scoring schemes from the lecture or your own invention, as long as it is meaningful and non-trivial.

Exercise 3

For your document collection from Exercise Sheet 1 (or another one if you wish), pick a non-trivial query of your choice, and find (by whatever means) as many documents in your collection which are relevant for that query as you can. Then run your query processor from Exercise 2 for that query. Determine the various precision / recall measures presented in the lecture. Pick three highly-ranked non-relevant documents returned and discuss why they were returned despite their non-relevance. Find three relevant documents that were not returned, and discuss why that happened.

Exercise 4

Consider top-k retrieval where the score for a document is the *maximum* (and not the sum) of the scores of the individual query words for that document. (1) Prove that in that case, it suffices to look at the top-k documents of each of the involved inverted lists, provided they are appropriately ordered, and say how they have to be ordered for this statement to be true. (2) Give a counterexample for k = 3, and a three-word query, showing that this statement is not true when scores are aggregated by sum (as discussed in the lecture) instead of by maximum (as for (1) above).