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 2

Submit until Wednesday, November 7 at 4:00pm

Exercise 1 (5 points)

Copy your code from the last exercise sheet to a new folder *exercise-sheet-02* and modify the *buildFromCsvFile* method (or add an option) so that it takes all sentences with the same URL (in *wikipedia-sentence.csv*) as one document.

Reason: with sentences as documents, the tf.idf scoring schemes that you should implement and evaluate in the tasks below do not make much sense.

Please also convert all words to lower case before adding them to the index (that is, make your search case-insensitive).

Exercise 2 (5 points)

Modify your *InvertedIndex* class so that your *buildFromCsvFile* method computes inverted lists with BM25 scores, for arbitrary given parameters k and b. Modify your *QueryProcessor* class to consider the scores (use boolean retrieval, as explained in the lecture). Consider the explanations and the implementation advice (slide before references) given in the lecture.

Exercise 3 (5 points)

Inspect the top-10 documents returned by your *SearchMain* program for the query *relativity theory*. Try three different BM25 parameter settings: $b = 0$ and $k = 0$ (binary), $b = 0$ and $k = 1000$ (standard tf.idf), and $b = 0.75$ and $k = 1.75$ (BM25 default). For each of these (manually) determine the Precision@10 (see page 2 for what is supposed to be relevant to this query). Report the results in your *experiences.txt* for this exercise sheet (see below), and briefly discuss them there.

If you like, also try some other queries.

Exercise 4 (5 points)

Commit your code to our SVN, in a new sub-directory *exercise-sheet-02*. Make sure that everything runs through without errors on Jenkins (compiler, unit tests, checkstyle).

Also commit, in that sub-directory, 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.

Definition of relevance for the query *relativity theory*:

Documents which say something substantial about relativity theory. For example, documents about a scientist who has done substantial research on relativity theory. Or documents that explain a substantial aspect of relativity theory.

Documents which only mention relativity theory in passing but are really about something else are not relevant. Same for documents which mention the two words only separately, in unrelated passages.