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 9

Submit until Tuesday, Januar **14** at 4:00pm (deadline extended by one week on January 6)

This exercise sheet is about clustering documents using k-means. See the Wiki for a code skeleton of the class *Clustering* for further specifications and **a lot of useful implementation advice**. In particular, for the sake of a simple and efficient implementation, it is explained there how you should wirk with two different representations of the documents: DENSE and SPARSE.

Exercise 1 (4 points)

Write a method *buildDocumentsFromInvertedIndex* that builds documents in SPARSE representation from a given inverted index with BM25 scores.

Exercise 2 ($3 \times 2 = 6$ points)

Write a method *normalize* that normalizes a given document in SPARSE representation such that the sum of the squares of the scores is 1.

Write a method *truncate* that for a given M, truncates a given document in DENSE representation to those entries with the (at most) M largest scores and returns a corresponding document in SPARSE representation.

Write a method *distance* that computes the distance $1 - x \cdot y$ between two normalized documents x and y, given in SPARSE representation, where $x \cdot y$ is the dot product.

Exercise 3 (5 points)

Write a method *cluster* that performs k-means clustering for the documents built using the method from Exercise 1. Pick a random subset of size k of these documents as the initial centroids. Implement a suitable termination condition that achieves a good compromise between small RSS and small running time.

Exercise 4 (5 points)

Write a method *writeCentroidsToFile* that writes the top-10 terms from each centroid to a file named *clusters.txt*. The format should be: one line per centroid, with the 10 terms separated by spaces. Run your whole algorithm for $k = 50$ and $M = 1000$ on the collection linked on the Wiki (it's a subset of the collection from Exercise Sheet 1). Report on the Wiki: document creation time, final RSS, and total running time for k-means. Commit your file *clusters.txt* to the SVN and briefly discuss your results in your *experiences.txt* for this sheet.

[please turn over]

Add your code, as well your result file *clusters.txt*, to a new sub-directory *exercise-sheet-09* 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.