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

Exercise Sheet 9

Submit until Wednesday, January 9 at 4:00pm
postponed by one week due to lecture cancellation on January 9

This exercise sheet is about implementing k-means clustering for text documents and applying it to our example collection. Consider the explanations in the lecture and the code design suggestions on the Wiki. Note that you can resuse parts of your code from your *InvertedIndex* class.

Exercise 1 (5 points)

Write a class *KMeansClustering* with a method *readFromCsvFile* that constructs for each document a list of the ids of the terms it contains, together with their BM25 scores.

Exercise 2 (5 points)

Add a method *distance* that computes the distance between two given document vectors (using list intersection), and a method *average* that computes the average of a given number of document vectors (using simple component-wise addition). Also add a method *truncate* that for a given M truncates each document list to those M terms with the largest scores, and a method *normalize* that normalizes a given document vector such that the sum of the squares of its entries is 1.

Exercise 3 (5 points)

Add a method *computeClustering* that does k-means clustering using the methods from Exercise 2. Pick a random subset of the documents as the initial clusters. Truncate cluster centroids after each round. Terminate when the decrease in RSS falls below a given threshhold ε.

Exercise 4 (5 points)

Use your class to perform k-means clustering on our example collection, for $k = 50$ and $M = 1000$. Try to minimize both RSS and the running time of your *computeClustering* method, and report both numbers on the table linked on the Wiki. Write the 10 terms with the highest scores for each of your final cluster centroids into a file *clusters.txt* (one line per cluster). Briefly discuss in your experiences whether and how these results make sense.

Commit your code to our SVN, in a new sub-directory *exercise-sheet-09*, 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.