Exercise 1  (14 points)

Copy your code from the last exercise sheet (or from the master solutions if you prefer) to a new folder sheet-08. Remove the methods concerning LSI and query processing (you don’t need that for this sheet), and keep only the methods needed to build the term-document matrix.

Then add the following methods, specified in the TIP file on the Wiki: initialize centroids (2 points), compute distances (4 points), compute assignment (4 points), compute centroids (4 points).

Make sure to write a unit test for each method (except initialize centroids if you use random numbers).

The methods for normalization are already fully coded, see the remarks on slide 24 of the lecture.

Exercise 2  (6 points)

Implement k-means using the code and methods from Exercise 1. Run it on the movie dataset from Exercise Sheet 2 (file movies2.txt) using $k = 50$. Write the top-10 terms from each of the final centroids to a file centroids.txt (format: one line per centroid, with the 10 words separated by spaces). Report on the Wiki: creation time for the term-document matrix, number of iterations, final RSS, and total running for your $k$-means method.

Commit your file centroids.txt to the SVN and briefly discuss it (along with the usual feedback) in your experiences.txt for this sheet.