Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Patrick Brosi

Information Retrieval
WS 2016/2017

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

UNI FREIBURG

Exercise Sheet 9

Submit until Tuesday, January 10 at **2:00pm**

Exercise 1 (20 points)

Write a program for k-means clustering with the following components and functionality.

1. Copy your code from the last exercise sheet (or from the master solution if you prefer) to a new folder *sheet-09*. Remove the methods concerning query processing (you don't need them for this sheet), and keep only the methods needed to build the term-document matrix.

2. Implement the k-means algorithm following the specifications in the TIP file on the Wiki. You must implement the test cases provided there. As usual, insignificant variations or extensions of the test cases are no problem. You don't have to test the *initialize_centroids* method, if you implement it using random numbers.

Note 1: you need L_2-normalization at two places in the algorithm. You already needed this normalization for ES8. Just copy your code from there or take the code from the master solution.

Note 2: you don't need any other normalization for this sheet. If you think you do, think again.

3. Implement a function that outputs the top-10 words of each centroid vector according to the following order: for each centroid vector, sort the words by $x_i \cdot idf_i$, where x_i is the score of term i in the centroid vector and idf_i is the inverse document frequency of term i (which you can easily compute from the inverted index).

4. Implement a main function such that your program can be called as follows: *python3 kmeans <file> <number of clusters>*. The program should output the time needed to build the term-document matrix, the total time for the k-means clustering, the final RSS, and the number of iterations. The program should also output the top-10 words from each centroid, according to 3.

5. Run your algorithm on the movie dataset from ES1 (which was also used in ES2 and ES8) with 20 clusters. Report the results in the table on the Wiki. If your results deviate starkly from the others or the baseline provided, join a Zen monastery. Briefly discuss your centroids (along with the usual feedback) in your *experiences.txt* for this sheet.

[turn over at christmas]

Add your code to a new sub-directory *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. Also commit the usual *experiences.txt* with your much appreciated brief and concise feedback.

Is reincarnation compatible with our current understanding of the world?