Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Marjan Celikik

Search Engines WS 09/10

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

UNI FREIBURG

Exercise Sheet 12

complete until Thursday, February 4th

Exercise 1

Implement the k-means algorithm from the lecture when the "points" are strings of text. As distance between two texts take the Jaccard distance $|A \cap B|/|A \cup B|$, where A and B are the sets of distinct words in the two texts. As the average of m texts take the text consisting of the M words with the largest total number of occurrences in the m texts, where M is a parameter. Use a reasonable stopping criterion of your choice.

Exercise 2

Run your k-means algorithm from Exericse 1 on the DBLP dataset from the last exercise sheet. Pick $k = 3$ and find a reasonable value for M. (Understand that if M is too large, your program will run very slow, and if M is too small the centroids will be of poor quality.)

Exercise 3

Extend your program to compute the RSS of your clustering as well as the precision and recall of your clustering with respect to the "perfect" clustering given by the class labels SIGGRAPH, SIGIR, and STOC. Provide these numbers in a nice table, as usual, for three different runs of your k-means algorithm (each with a different random initialization of the three centroids).

Exercise 4

Prove that k-means does not satisfay the consistency property defined in the last part of the lecture. This holds for arbitrary $k \geq 2$, but you only have to prove it for $k = 2$.

Hint: Construct a point set with three subsets X_1, X_2 and Y. Then carefully define distances between all these points such that 2-means will pick one centroid in $X_1 \cup X_2$ and one in Y, but when the points within X_1 and within X_2 are moved closer together (with all other distances remaining the same) then 2-means will pick one centroid in X_1 and one in X_2, thus yielding a different clustering.