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

# Search Engines WS 09/10

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

## Exercise Sheet 13

complete until Thursday, February 11th

### Exercise 1

In the lecture we have written code for hierarchical clustering using the single-link heuristic that runs in $O(n^3)$ time, where $n$ is the number of points. Modify the code so that it uses the *complete-link* heuristic for merging clusters and so that it runs in $O(n^2 \cdot \log n)$ time, using a priority queue as briefly discussed in the lecture. Run your code for $n = 10^1, 10^2, 10^3, \ldots$ (go as high as you can on your machine) and output $T/n^2$, where $T$ is the running time. Does $T/n^2$ indeed look logarithmic in $n$, as it should? (Note that for $n = 10^i$, $\log n$ grows linearly with $i$.)

### Exercise 2

Modify the code from the lecture so that it runs in $O(n^2)$ time using a so-called NBM-array (NBM = next best merge). The NBM-array stores for each cluster representative (= the smallest index of a point in the cluster) the representative of the cluster with the highest single-link similarity. To update the NBM-array after each iteration, make use of the fact that single-link is best-merge persistent, as discussed in the lecture. Check $T/n^2$ analogously to the previous exercise.

### Exercise 3

Show by a counterexample that the complete-link heuristic is not best-merge persistent.

### Exercise 4

Consider the flat clustering $\mathcal{C}$ that results when stopping a hierarchical clustering of $n$ points after $k$ iterations (that is, we have $n - k$ clusters now). Let $s_k$ be the similarity of the last cluster pair merged before we stopped. Let $G$ be the graph with $n$ vertices (one for each point), and with an edge between two vertices if and only if the similarity between the respective points is at least $s_k$. Then prove the following two statements. (1) If the merging heuristic was single-link, the clusters of $\mathcal{C}$ are exactly the connected components of $G$. (2) If the merging heuristic was complete-link, the clusters of $\mathcal{C}$ are exactly the maximal cliques of $G$. You can assume that the similarities between all pairs of points are different.