Exercise 1  Let a clustering for a set \( D \) of documents be defined as a function \( f : D \rightarrow \{1, \ldots, k\} \), for some integer \( k \), that is, \( f \) assigns each document to exactly one cluster.

(a) In the seminar, we have seen a number of algorithms for expressing documents in terms of some concepts instead of in the original terms. Say in one sentence what is the main difference between these algorithms and clustering, as defined above. You should make an effort to formulate your sentence well, and in a concise way; don’t just write a sloppy remark.

(b) There are many ways to use the output of a clustering algorithm for ranked retrieval, that is, for using \( f \) to compute similarity scores between queries and documents. Think of one such way (it may well be simple, as long as it does something reasonable), and give a short and concise description of it.

Exercise 2  There is an enormous literature on clustering algorithms. Pick one algorithm of your choice, and prepare a short presentation on it. From your presentation, everybody in our class should be able to understand how, in principle, the algorithm works. Name at least one advantage and one disadvantage of your algorithm. Does it have some provable properties? If so, mention them. Also name the source of your algorithm. Give a 1-slide description of the retrieval scheme that results from your clustering algorithm, according to what you worked out for Exercise 1(b) above.

One more constraint: no two of you should select the same algorithm! To realize this, as soon as you have made your choice, publish a (very) short description of it on the Wiki page for this exercise sheet. Before you start searching for an algorithm, you have to look at the Wiki to see which schemes have already been selected. If you then select a scheme, and when publishing it to the Wiki you find that someone else has selected it in the meantime, too, that’s ok, but only then.

Exercise 3  Write one, two sentences about how much time, on average, you invested for this seminar per week so far, and whether you found that reasonable. You don’t have to do this if you don’t want to. If you have other comments on the seminar so far, feel free to give them.