1 Distributed Indexing

Many parallel tasks, such as index construction, that run on distributed computer clusters are based on the MapReduce model.

1. Sketch a generic MapReduce schema and explain the input and output of the map as well as the reduce step.

2. Given is the following document collection:
   D1: to be or not to be
   D2: strive to be the best
   a) How can MapReduce be used to count the occurrences of all terms in this document collection? Describe the map/reduce input and output.
   b) Modify your described MapReduce schema so that it generates an inverted index.

3. Apply MapReduce to the problem of counting how often each term occurs in a set of files. Specify map and reduce operations for this task.

4. Assume that machines in MapReduce have 1 TB of disk space each. Assume further that the postings list of the term “the” has a size of 2 TB. Then the MapReduce algorithm as described cannot be run to construct the index. How would you modify MapReduce so that it can handle this case?