Uni Freiburg, Web Science Group
Prof. Peter Fischer
Systems Infrastructure for Data Science - Winter 2014/15

Exercise Sheet #10: Map/Reduce

January 15, 2015

A. How would you sort a list of names using MapReduce? Please provide the pseudo-code for the map and reduce functions. Does your solution eliminate duplicates? Describe modifications for providing duplicate eliminated results.

B. Assume you are given a list of [filename : string, md5hash : string] pairs. How would you find the names of duplicate files where you should report only distinct file names? Two files are duplicate if their md5hash values are equal. Please provide the pseudo-code for the map and reduce functions. An example for this elimination would look as follows: [Name1, 123], [Name2, 123], [Name1, 123], [Name1, 456] $\rightarrow$ [123,[Name1, Name2]]: File [Name1, 456] does not have any duplicates, [Name1, 123] is only shown once.

C. Facebook has a feature which lists common friends with a person when you visit his or her profile page. In the context of this exercise let's assume that common friends will be listed for pair of persons who are already friends. We would like to implement this feature using MapReduce. The friendship is a bi-directional relationship, if A is a friend of B then B is a friend of A too. Assume that you are given a file which consists of millions of lines in the following format:

   PersonA [PersonB, PersonC, PersonD, ...]
   PersonB [PersonA, PersonD, PersonE, ...]...

   where each line lists a person's name followed by list of his/her friends. Given this friendship list file, please provide the pseudo-code for map and reduce functions for finding common friends among all pairs of persons listed in the file.