Brandeis University
Department of Computer Science
COSI 132a - Networked Information Systems - Fall 2020

Homework #4: Programming with MapReduce

Due date March 17th

A-30 Points Provide the pseudo-code for the map and reduce functions that sort a list of names. Your solution should eliminate duplicates (words that appear more than once).

Hint: First think how you can get partial ordering of the words. Then think about an alternative key partitioning function that can help achieve a total ordering.

B-30 Points Assume you are given a list of strings [`filename`, `md5hash`] pairs where `filename` represents the name of some file and `md5hash` is a hash value of that string when we apply the MD5 hash function. The `md5hash` is also a string. Some file names may appear in the list more that once. You can assume that the MD5 hash functions generated unique hash values for its input, i.e., no two different file names are given the same hash value. Please provide the pseudo-code for the map and reduce functions that find the names of **duplicate** files, i.e., the file names that appear more than once (Note that two files are duplicates if their md5hash values are equal). Your code should report **only distinct** file names.

C-40 Points Facebook has a feature which lists common friends with a person when you visit his or her profile page. 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. In the context of this exercise let's assume that common friends of a person will be listed for pair of persons (who are also friend themselves).

Assume that you are given a file which consists of millions of lines in the following format:

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

...

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 friends listed in the file.

For example, you should report a file with each line's format is:

(Person A, Person B) : Person C, Person D, where A and B are friends and C, D are their common friends