Problem Set: Algorithms for MapReduce

Both problems are chosen exercises from Chapter 2 of the book *Mining of Massive Datasets*, available for free from `http://i.stanford.edu/~ullman/mmds.html` (linked to from the webpage of our PhD Open lecture).

You are allowed and encouraged to work on the problems in small groups, provided that:

- you acknowledge the names of the students you worked with, and
- you write up the solutions on your own.

Please send your solutions by e-mail to `klin@mimuw.[you know what]` by **Feb. 2, 2014**.

Problem 1 (Ex. 2.3.1 from the book): Design MapReduce algorithms to take a very large file of integers and produce as output:

(a) The largest integer,

(b) The average of all the integers,

(c) The same set of integers, but with each integer appearing only once,

(d) The count of the number of distinct integers in the input.

Problem 2 (Ex. 2.5.1 from the book): What is the communication cost of each of the following algorithms, as a function of the size of the relations, matrices, or vectors to which they are applied? (All section numbers refer to the book.)

(a) The matrix-vector multiplication algorithm of Section 2.3.2,

(b) The union algorithm of Section 2.3.6,

(c) The aggregation algorithm of Section 2.3.8,

(d) The matrix-multiplication algorithm of Section 2.3.10.