

IS622 - Homework 2

Brian Chu / Sept 6, 2015

Exercise 2.3.1: Design MapReduce algorithms to take a very large file of integers and produce as output:

(a) The largest integer

Map: For each tuple in R, produce key-value pair (m,1) where m is the maximum integer in the chunk.

Reduce: From the combined values of m, select and output the maximum value(s) M.

(b) The average of all the integers

Map: For each tuple in R, produce key-value pair (a,1) where a is the average value of all the integers in the chunk.

Reduce: From the combined values of a, output the average value A.

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

Map: For each tuple in R, produce key-value pair (t,n) where t is the integer and n is the number of occurrences.

Reduce: Each unique key t represents a group. Output the subset $R[t_0, t_1, \dots, t_n]$.

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

Map: For each tuple in R, produce key-value pair (t,n) where t is the integer and n is the number of occurrences.

Reduce: Each unique key t represents a group (t,N). Output θ where θ is the sum of rows in the dataset.

Exercise 2.3.2 : Our formulation of matrix-vector multiplication assumed that the matrix M was square. Generalize the algorithm to the case where M is an r-by-c matrix for some number of rows r and columns c.

- M is now an R x C matrix. Vector V, therefore, has to be length C elements.
- i is the row index on the range 1 to R.
- j is the column index on the range 1 to C.

The calculation for the resulting product vector X with length R is:

$$X_j = \sum_{i=1}^R M_{ij} V_j, j \in [1, 2, \dots, C]$$

Map: Matrix element M_{ij} produces the key-value pair (i, $M_{ij} V_j$).

Reduce: Sums all the values with the same key i. Outputs (i, x_i).