Homework 1 – MapReduce Programming Assignment (100 points)
Submission Deadline: September 27, 2018 (12:29pm)

Dataset (rel.csv):

Homework Requirements:
Please download the dataset named "rel.csv" from the link above. In this dataset, each line is a pair of strings separated with comma. Formally, each line can be represented as (c_1, c_2).

Part I. (50 points) Given the pairs in the "rel.csv" dataset, please write a MapReduce program according to the following requirements.

Map phase:
Order the pair of strings alphabetically.
For example, given a pair (c_1, c_2),
if $c_1 <= c_2$, then the ordered pair is still (c_1, c_2);
if $c_2 < c_1$, then the ordered pair is (c_2, c_1);

Reduce phase:
Output all ordered pairs in the dataset as well as the count of each ordered pair.
In addition, please use counter to compute the total number of **unique** ordered pairs.

Part II. (50 points) Given the pairs in the "rel.csv" dataset, assume *L* is the set of all first components of the pairs and *R* is the set of all second components of the pairs. Please write a MapReduce program to compute the set difference *R – L*, that is, the set of strings appearing in the second components of the pairs but not appearing in the first components of the pairs.

Submission Details:
Please submit a compressed file containing the following files to Canvas:

(1) A Word or PDF document containing the following:
   a. Code for the two MapReduce programs, as well as comments indicating the above requirements;
   b. For the output results obtained by the two programs, please upload them to a public accessible site (e.g., google drive) and make them downloadable. Please provide the sharable web links to these output results in this document.
(2) Original MapReduce programs in Java.
(3) Compiled jar files for the two MapReduce programs that can be directly executable on Hadoop.

NOTE:
For those who are using a programming language other than Java (e.g., Python), please submit your code in the corresponding language, and include specific instructions and/or commands to run your code in the document.