**598 Project 2**

**Weichuan Deng**

**112516608**

**Introduction**

The link information among them are stored in the file "project2\_data.csv" under /gpfs/projects/AMS598/projects/project2.

Each row in the file has two webpage numbers, indicating that the first webpage contains a link to the second webpage.

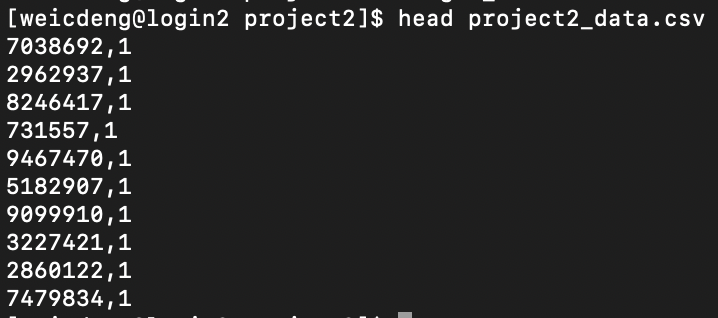
For example, (56800,678) means webpage 56800 has a link to webpage 678.

Use the taxation method (beta = 0.9) to calculate and report the top 10 webpages with the largest pagerank values.

You need to perform the map-reduce steps on SeaWulf but only need to run five (5) iterations.

**Dataset**

Firstly, I look over the dataset. The dataset is a csv file containing 2 columns the first one means the column of the element in matrix and the second one means the row of the element in matrix.

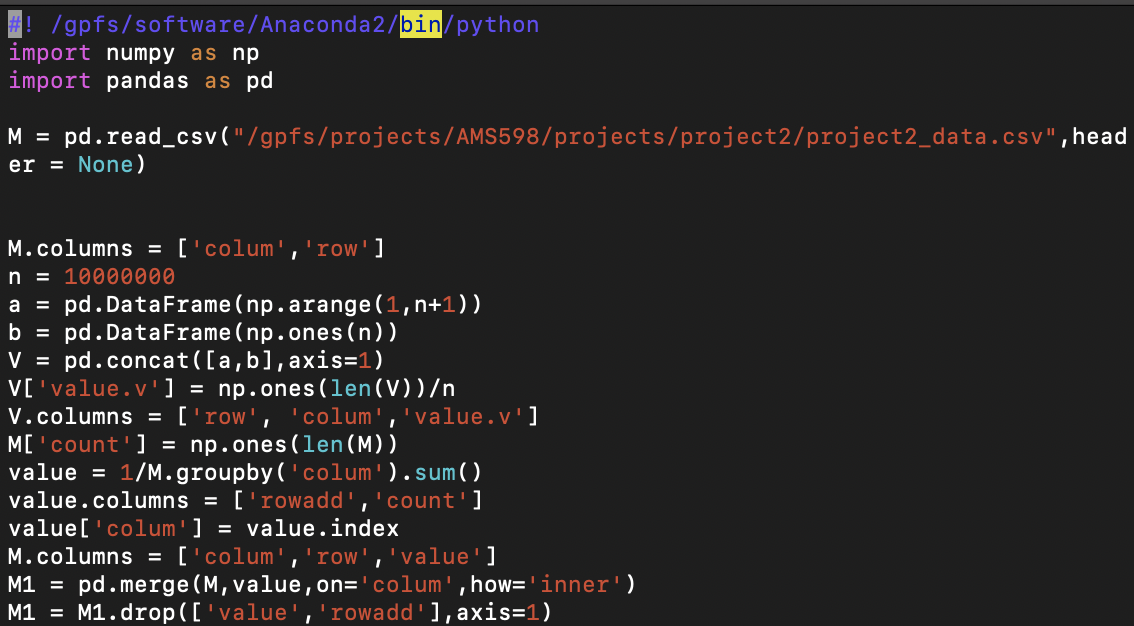


**Map-reduce**

I separate the transition matrix to 5 parts and each partition has 2000000 rows. Then I will get 5 partition of the page rank for 1 iteration. After combining those page rank, I move to next iteration.

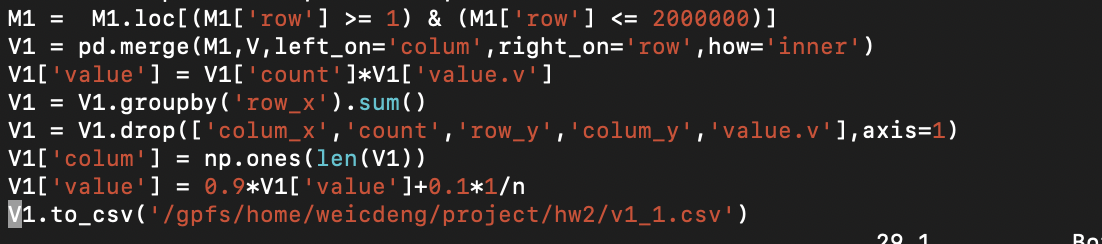
**Model**

The most important thing is the sum of each column in the transition matrix is 1. I add a ones column for the Dataframe at first. And calculate the number of non-zero row in this column. Then I use 1 divided by the number of rows in that columns to get the transition matrix. Also, I initialize the vector V.



Select 2000000 rows in the transition matrix to do the map-reduce algorithm with taxation:

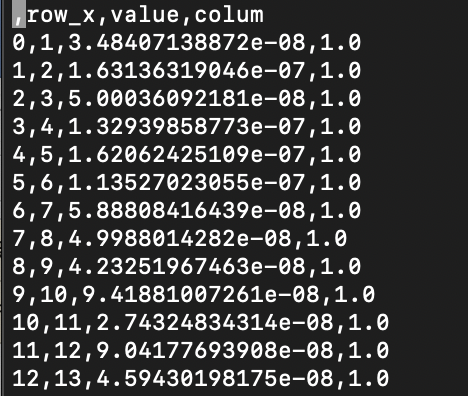
where , then save the first partition result of the first iteration as v1\_1.csv.



Then combine those result of the iteration. Save this file as v1.



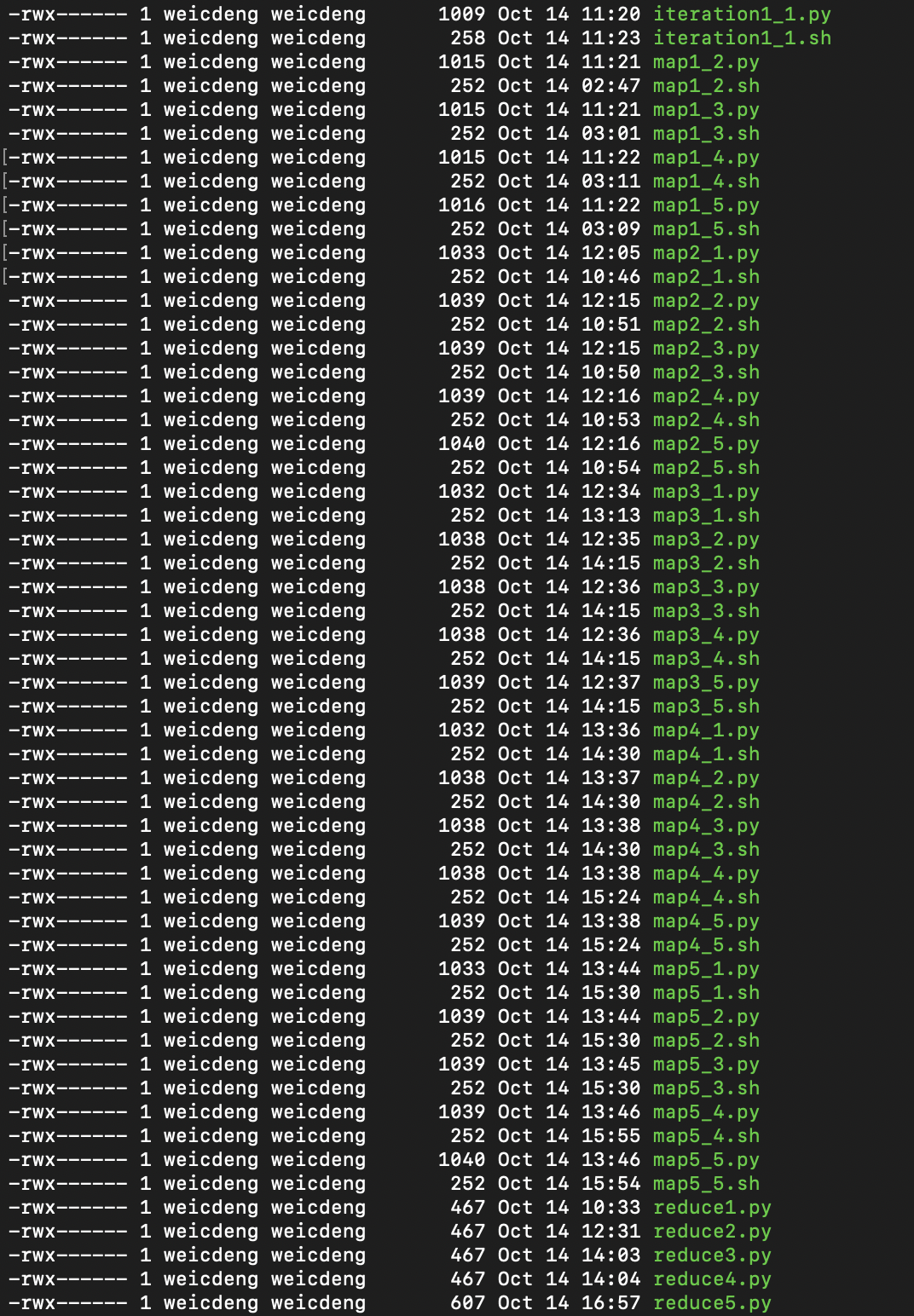
The result of the first iteration:

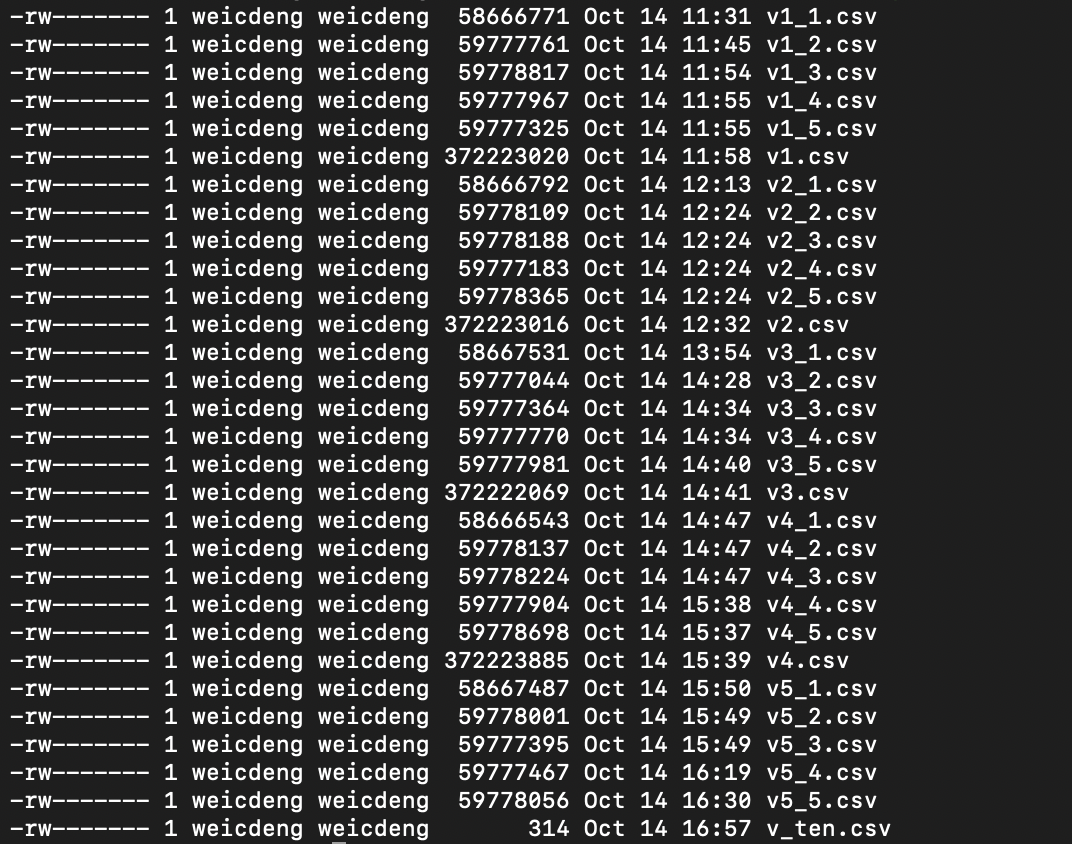


Then I just read the new V to do the next iteration.



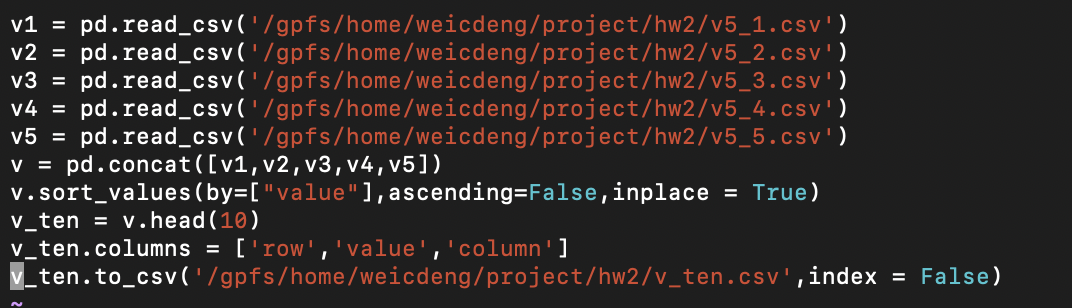
Five iterations and each iteration has five mappers.

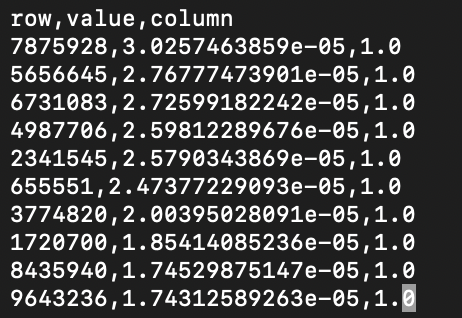




**Result**

Finally, after the fifth iteration, I get the final V. After sorting and ranking, Top 10 pages with the largest pagerank values are:

****

****

Download this file to my computer.

