DSCI 551 – HW5

(Implement SQL Using MapReduce)

(Fall 2020)

100 points, Due 11/22, Sunday

In this homework, consider the region table in the Covid-19 dataset. You are provided with a CSV file "region.csv" containing only the *province*, *city, elderly_population_ratio*, and *nursing_home_count* of rows in the table. Your task is to write a Hadoop MapReduce program Sum.java to implement the following query. Assume only one reduce task is used.

```
select province, sum(nursing_home_count)
from covid19.region
where elderly_population_ratio >= 20
group by province
order by province
```

Sample execution format:

hadoop jar sum.jar Sum input

where the "input" is the directory that stores the csv file.

Submission:

Sum.java, sum.jar (jar file containing the class files), and your output file part-r-00000.

Hint using Java:

- Using replace function to remove quotes in strings.
- Using Float.parserFloat(s) to convert string s into float.
- Using Integer.parseInt(s) to convert string s into integer.