DSCI 551 – Spring 2021

Homework 5 (Hadoop MapReduce), 100 points

Due: 4/25 Sunday

1. [40 points] Use the provided roster file (CSV format, the same as one used for hw1), write a Hadoop MapReduce program Part.java to find the number of people from different participation countries. Example output:

```
China	3
United States of America		5
…
```

Execution format: hadoop jar part.jar Part input output

Assume roster file is stored under the input directory.

Submission: Part.java part.jar and your output file (part-r-00000).

2. [60 points] Use the world database provided (3 JSON files: country.json, city.json and countrylanguage.json), write a PySpark program using Spark Dataframe to answer the following SQL questions. **For questions a and b, also write a PySpark program using RDD API to answer the question.**

   a. [15 points (5 points for RDD)] Select name
   From country
   Where continent = "North America";

   b. [15 points (5 points for RDD)] select country.name, city.name from country join city on country.Capital
   = city.ID;

   c. [5 points] Select distinct continent
   From country;

   d. [10 points] select language from countrylanguage where countrycode = 'CAN';

   e. [15 points] select continent, avg(LifeExpectancy) as avg_le
   from country
   group by continent
   having count(*) >= 20
   order by count(*) desc
   limit 1;

Submission: dataframe-a.py … dataframe-e.py rdd-a.py rdd-b.py

All the results should be printed out directly in terminal when execute:

python dataframe-a.py

Also submit a pdf file that contains both script and result for each question.

Use Python3.8 for this homework, all your scripts will be tested on ec2.