# 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).

1. [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.**
   1. [15 points (5 points for RDD)] Select name

From country

Where continent = "North America";

* 1. [15 points (5 points for RDD)] select country.name, city.name from country join city on country.Capital

= city.ID;

country.join(city, country.Capital== city.ID)[[country.name,city.name]]

* 1. [5 points] Select distinct continent

From country;

country[['Continent', 'Region']].distinct()

* 1. [10 points] select language from countrylanguage where countrycode = 'CAN';
  2. [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

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