INF 551 – Spring 2020

Homework #5

Due: May 3, Sunday
100 points

In this homework, we will consider again the country data set, country.csv, city.csv, and countrylanguage.csv, similar to that you have seen in homework 1. But note that all NULL values in the data set are replaced with empty string '', and header lines (1st line) have been removed. Use the data set attached with this handout.

1. [Hadoop MapReduce, 55 points] Write a MapReduce program AvgExp.java that implements the following SQL query:

```
SELECT Continent, avg(LifeExpectancy)
FROM country
where GNP > 10000
group by Continent
having count(*) >= 5;
```

   You can take WordCount.java as the template. But note the following:

   - You may want to use split function of Java instead of StringTokenizer:
     - https://docs.oracle.com/javase/7/docs/api/java/lang/String.html#split(java.lang.String)
   - Replace IntWritable with FloatWritable
   - It is OK that your implementation does not utilize a combiner.
   - Name your jar file "ae.jar".

   Execution format: hadoop jar ae.jar AvgExp.java <input-hw5> <output-hw5>

   Ignore the angle brackets.

   Where the <input-hw5> directory stores a single file "country.csv".

   Submission: AvgExp.java ae.jar

2. [Apache Spark, 45 points] For each of the following questions, write a Spark program in Python. You can assume that all three csv files, country.csv, city.csv, and countrylanguage.csv (note files names are all lowercase letters), are available in the same directory where you execute the code. Note that you should NOT use Spark DataFrames or Spark SQL for this homework.
   a. [10 points] Find the 10 most populated countries in a given continent. Return the names of countries and their populations in the descending order of populations. Name your script "pop10.py".
      - Execution format: spark-submit pop10.py "Asia"
        - This find the top-10 most populated countries in Asia.
      - Sample output:

        China, 1277558000.0
        India, 1013662000.0

...

b. [10 points] Find names of countries which do not have any languages recorded (in the country language table). Output one line per country. Name your script “no-lang.py”.
c. [15 points] Find names of countries which have at least 10 unofficial languages. Order the countries by the descending order of the number of unofficial languages. Name your script “unofficial10”.py.
d. [10 points] Implement the SQL query in Question 1. Name your script “avgexp.py”.

Submission:

1. Compress your code and jar package into a compressed file with naming convention: <Firstname>_<Lastname>_hw5.zip

2. For q1, AvgExp.java and ae.jar should be submitted. In q2, pop10.py, no-lang.py, unofficial10.py and avgexp.py should be submitted. There are 6 files to be included in the compressed folder in total.

3. Naming error in submitted files will result in 5 points deduction.