INF 551 – Fall 2019

Homework #5

Due: December 6, Friday
100 points

1. [60 points] Write a Hadoop MapReduce program, count.java, that takes the Sells table (stored as a comma-separated-value or CSV file) in the beers database, and computes the same results as the following SQL query. You can assume that there are no NULL prices in the table.

```
Select bar, count(*)
From Sells
Where beer like 'bud%'
Group by bar
Having max(price) <= 5
```

Example input file:

```
Joe's bar,bud,3
Mary's bar,bud light,4,
…
```

Sample Output format: **(Please use "\t" as delimiter between bar name and number)**

```
Bob's bar	3
Joe's bar	3
…
```

Execution format:

hadoop jar count.jar count input/sells.txt output
(where sells.txt is the file storing the content of the sells table)

2. [40 points] For each of the following queries, write a Spark program in Python to implement the query. Assume that all tables in the beers database are stored in the CSV files.

   a. Implement the same query as Question 1.

   Execution format: spark-submit q1.py input/sells.txt q1.txt
   Output format for q1.txt is same as above.

   b. For each bar, compute the average price of beers sold at the bar.

   Execution format: spark-submit q2.py input/sells.txt q2.txt
   Output format for q2.txt: **(Please use "\t" as delimiter between bar name and number)**

```
Bar	Average_price
Bob's bar	3
Joe's bar	3
```

c. Find all drinkers that frequent some bars but do not like any beers.

Execution format: spark-submit q3.py input/frequents.txt input/likes.txt q3.txt

Output format for q3.txt:

Drinker
Steve

d. Find all drinker-beer pairs such that the drinker likes the beer and frequents a bar that sells the beer.

Execution format: spark-submit q4.py input/likes.txt input/frequents.txt input/sells.txt q4.txt

Output format for q4.txt: **(Please use "\t" as delimiter between drinker name and beer name)**

Drinker Beer
Steve Bud

Submissions:

For q1: count.jar, count.java

For q2: q1.py, q2.py, q3.py, and q4.py

Submission Criteria:

1. Please **submit all the files in 1 folder.**
2. Please **append all files with your names**. Eg. firstname_lastname_filename.py/.java/.jar
3. For q1, **please do not use any library other than org.apache.hadoop.*, java.***
4. For **q1**, you should implement **Hadoop MapReduce** for the task.
5. For q2, **please use python 3** and do not use any library other than Python Standard Library and **pyspark**.
6. For q2, **you should implement the query in spark operation., no for loops allowed.**