INF 551 – Fall 2018

Homework #5

Due: November 28, Wednesday
100 points

1. [70 points] Write a Hadoop MapReduce program, Max.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, max(price)
From Sells
Where beer != 'Summerbrew'
Group by bar
Having count(*) > 1
```

Example input file:

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

Sample Output format:

```
Bob's bar	3

Joe's bar	3

…
```

Execution format:

```
hadoop jar max.jar Max input/sells.txt output
```

where sells.txt is the file storing the content of the Sells table.

2. [30 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 q2_a.txt

q2_a.txt example:

Bar Max_Prize

Bob's bar 3

Joe's bar 3

…

(ps: use "\t" as delimiter)

b. Find all drinkers that like some beers but never frequent any bars.
   Execution format: spark-submit q2.py input/likes.txt input/frequents.txt q2_b.txt
   Q2_b.txt example:
   Drinker
   Steve

c. Find all drinker-beer pairs such that the drinker likes the beer and frequents a bar that sells the beer.
   Execution format: spark-submit q3.py input/likes.txt input/frequents.txt input/sells.txt q2_c.txt
   Q2_c.txt example:
   Drinker Beer
   Steve Bud

Submissions:

For q1:

Source codes: Max.java, max.jar

Output files: part-r-00000

For q2:

Source code: q2_a.py, q2_b.py, and q2_c.py

Output files: q2_a.txt, q2_b.txt, and q2_c.txt

Important Notes:

1. Please prepend your name to all the submission files as before to facilitate the grading. e.g. firstname_lastname_Max.java, firstname_lastname_max.jar firstname_lastname_q2_a.py … **DO NOT** place them in a folder or zip file.
2. For q1, **please do not use any library other than org.apache.hadoop.*, java.***
3. For q1, you should implement hadoop MapReduce for the task.
4. For q2, **please use python 2.7** and do not use any library other than Python Standard Library.
5. For q2, **you should implement the query in spark operation.**