INF 551 – Fall 2019

Homework #3: SQL

Due: November 3, Sunday (end of day)
100 points

Consider again the LA Restaurants & Market Health data set available at Kaggle: https://www.kaggle.com/cityofLA/la-restaurant-market-health-data. In particular, we consider the two CSV files: one for inspections; the other for violations.

1. [20 points] Write an SQL script "load.sql" that does the following:
    a. Creates a table "inspections" for the inspection data set; and a table "violations" for the violation data set. Your tables should be stored in a database called "inf551" with both user & password being "inf551".
    b. Loads the data in the csv files into the respective tables. You may refer to: https://dev.mysql.com/doc/refman/5.7/en/load-data.html for details on "load data" statement in MySQL.

    Note that load.sql will assume the two data sets are located at the same directory in the name of "violations.csv" and "inspections.csv".

    Submission: <firstname>_<lastname>_load.sql

2. [50 points] Write an SQL query for each of the following questions:
    a. Find out names of facilities whose name contains "cafe" (case insensitive) and had a violation with code "F030".
    b. Find out names of facilities that have the highest inspection scores.
    c. Find out which facility (by id) has the largest number of violations. Output the names of such facilities (ascending order).
    d. Find out which facilities that had inspections done but do not have any violations (as recorded in the violations data set). Output names of such facilities (ascending order).
    e. For each different letter grade in inspections, output the average score of facilities receiving the letter grade.

    Submission: Name files as <firstname>_<lastname>_a.sql, <firstname>_<lastname>_b.sql …

3. [30 points] Write a Python script "good.py" that answers the question 2.d above.
    Note that your script should use Python **mysql-connector** to connect to the "inf551" database mentioned above. Output the results to a file whose name is specified in command line.
    Submission: <firstname>_<lastname>_good.py
    Execution: python good.py output_file_name.txt

INF 551 – Fall 2019

Requirements

1. Python Environment : Python3.6
2. Packages : The Python Standard Library and mysql-connector
3. Submission :

   For question 1, **<firstname>_<lastname>_load.sql**

   For question 2**, <firstname>_<lastname>_a.sql**, **<firstname>_<lastname>_b.sql**, and so on.

   For question 3, **<firstname>_<lastname>_good.py**

   Then submit all files in a zip file named as <firstname>_<lastname>_hw3.zip
4. Command to Execute Your Code :

```
# for question 3
$ python <firstname>_<last_name>_good.py output_file_name.txt

(all arguments are the paths to the files, please not hard-code)
```

5. Output Format :

For question 3, please strictly follow the output format: one restaurant name per line.

(output file path refers to **output_file_name.txt** above)

```
ALL INDIA CAFÉ
ANDY'S DONUTS
BIOBAR
.
.
.
```

Grading Criteria

1. If your programs can not be executed with the command specified above, there will be 40% penalty.
2. If your programs can not be executed with the required Python version, there will be 30% penalty.
3. If you use non-standard python packages (except for mysql-connector package) then 30% penalty.
4. If your .py takes more than 5 minutes for each to complete, there will be 20% penalty.
5. Please do not keep any "print" statements, they will lead to 10% penalty.
6. Please do not hard-code file names for Q3, else 10% penalty.
7. Please submit all files under 1 zip file in the format mentioned in the requirement.
8. Late homework will be deducted by 10% for every 24 hours that it is late. (no credit after 72 hours)