Homework 4: Query Execution

DSCI 551 – Spring 2025

Due: 11:59pm, 4/9, 2025, **Wednesday**

Points: 100

In this homework assignment, you will work on a data set on LA restaurant inspection, named "inspetions.csv". The data lists the health inspection reports for the restaurants in LA in 2015 to 2018. Note that there may be multiple inspections for a restaurant.

You can find more information about the data set here: https://www.kaggle.com/datasets/cityofLA/la-restaurant-market-health-data?resource=download. But note that the dataset used in this homework is slightly different from the one on the above Web site. Please use the inspections.csv provided to you, attached with this handout.

Consider the following SQL query (assume data are stored in a table called "inspections):

```
Select facility_name, avg(score)
From inspections
Where facility_name like '%COFFEE%' and program_status = 'ACTIVE'
Group by facility_name
Having count(*) >= 10
order by facility_name;
```

Write a Python program sql2py.py to find the answer to the above query using the provided csv file (note your code should work on other csv files in the same format).

Requirements:

- Your code should NOT load the entire inspections.csv file into memory. Instead, it should assume that the csv file may contain a huge amount of data that might not fit into the memory. This corresponds to the setting in database system where tables might store terabytes of data on a machine with gigabytes of memory. Hint: you may use read_csv with chunksize specified in pandas.
- Execution format:
    - python3 sql2py.py <inspections.csv> <chunk_size>
      where the chunk_size specifies how many rows of data are to be read at the same time from the csv file.
    - Note also that inspections.csv used to test your code might contain a different subset of inspection records that one provided to you.

Permitted libraries:

pandas, sys, collection, traceback

SUBMISSION DETAILS:

- Students are required to submit two files:
    1. YourName_sql2py.py — your completed Python script (replace "YourName" with your full name using underscores, e.g., John_Smith_sql2py.py)
    2. The output of your script on the provided inspections.csv file, saved as a .txt or .csv file (e.g., John_Smith_output.txt)
- Do not load the entire CSV file into memory. Use chunked reading as demonstrated in the template.
- Do not modify the structure of the template file. Just complete the code where indicated. You may add helper functions if needed.
- The test script will evaluate your script using multiple CSV files with the same format. Points will only be awarded if the script returns the expected result.
- Make sure your code runs with the following command format:

  python3 YourName_sql2py.py <inspections.csv> <chunk_size>

- Note: The data file (e.g., inspections.csv) should be in the same folder as YourName_sql2py.py.
- You will get 0 points if the code breaks due to syntax errors or any other issues. Please test your script thoroughly before submission.