CptS -451 Introduction to Database Systems
Spring 2017

Project Milestone-1

Due Date: Thursday February 9th, 11:59pm

Summary:

In this milestone you will parse the Yelp JSON data and develop a simple database application. The goal of this exercise is to get you started in database programming early on. In Milestone3 you will develop a larger application with all required features.

Milestone Description:

1) Download the Yelp dataset from http://www.eecs.wsu.edu/~arslanay/CptS451/links.html. Look at each JSON file and understand what information the JSON objects provide. Pay attention to the data items in JSON objects that you will need for your application.

Download the sample JSON Parser program (C# or Python) from Blackboard (*Project\Sample JSON Parsing Code*). These programs provides example code for:

- reading JSON objects form a file and extracting certain key and value pairs from JSON objects,
- writing extracted data into a text file.

Please note that the sample code includes examples of extracting simple key values only. In a JSON object the key value can be an array or another JSON object (for example: hours), therefore you need to recursively parse those objects until you extract all data stored in JSON objects. You will write the code for parsing business, tips, user, and checkin JSON objects.

In `yelp_business.json` : Parse all keys except **review_count**, **attributes,** and **neighborhoods**.

In `yelp_tip.json`: Parse all keys.

In `yelp_user.json` : Parse all keys except **review_count, compliments** and **elite**.

In `yelp_checkin.json` : Parse all keys. (You need to aggregate the checkin information for the hours of the day. See below.)

Both versions (C# and Python) provide the same functionality; you may start with either. However, overall JSON parsing is easier in Python.

Parsing Check-in Data: The check-in objects include information about the number of check-ins for a particular business . The keys in the check-in JSON objects are in the form of:
"hour of the day – day of the week".
For example the key "9-3" corresponds to 9am-10am on Tuesday. (Day-0 is Sunday and day-6 is Saturday; time values are based on 24hour clock (i.e., military time))
The value of a key corresponds to the number of check-ins at the corresponding time-of-day and day-of-week. For simplicity, in your project you will aggregate the check-in information further and sum up the checkin values for morning hours (6am-12noon), afternoon hours (12noon-5pm), evening hours (5pm-11pm), and night hours (11pm-6am) (Assume start time of each interval is inclusive and end time is exclusive.) Therefore, for each day of the week, you will have 4 check-in values instead of 24: morning, afternoon, evening and night.

2) Download the "milestone1DB.csv" file from the link
http://www.eecs.wsu.edu/~arslanay/CptS451/links.html

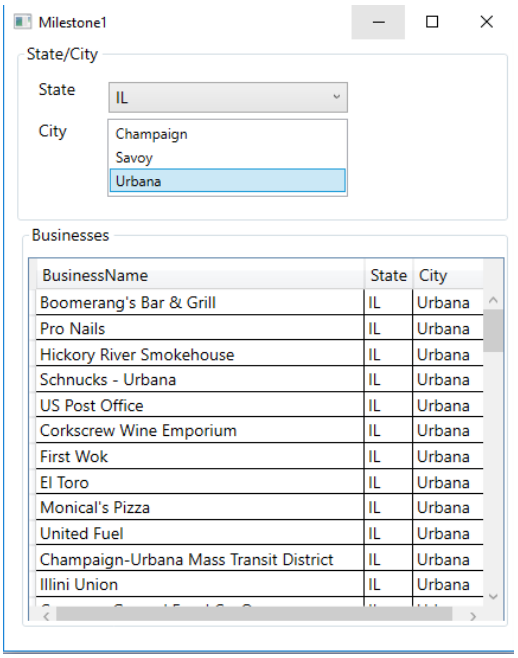
Create a database on MySQL with name "*Milestone1DB*" and create a table named "*business*". The schema of the *business* table should comply with the columns of the CSV file, i.e., there should be an attribute for each column of the CSV file. Please define the type and domain of each attribute based on the possible values that appear in the corresponding column.

The "milestone1DB.csv" file includes 3 columns: *name* (name of the business), *state*, and *city*.

Import the CSV file into this table by executing the following statement in the PostgreSQL command line. Please replace <path> with the directory path for the milestone1DB.csv file.

```
\copy business FROM '<path>/ milestone1DB.csv' DELIMITER ',' CSV
```

3) Write a simple application (either web or standalone) which connects to the *Milestone1DB* database and runs simple queries on the *business* table. A sample screenshot for your milestone1 application is shown below. The application will:
   - list the states that appear in business table and allow user to select a state;
   - when a state is selected, the zipcodes in that state will be listed;
   - when a zipcode is selected the list of the businesses will be listed.



A video tutorial on how to establish connectivity with the PostgreSQL in C# using Npgsql will be available on Blackboard.

You need to run the following queries on the *business* table:

```
SELECT DISTINCT state
FROM business
ORDER BY state;

SELECT city
FROM business
WHERE state= <selected state>
ORDER BY city;
```

```
SELECT name
FROM business
WHERE city= <selected city> AND state= <selected state>;
```

Milestone-1 Deliverables:

1. (40%) Source code for parsing all JSON data. Only submit your source code, **not the data files**.
2. (60%) Source code for your application. Only submit your source code, **not the data files**.

Create a zip archive *"<your-last-name>_milestone1.zip*" that includes your source code for JSON parsing and your sample application. Upload your milestone-1 submission on Blackboard until the deadline.

References:

1. Yelp Dataset Challenge, http://www.yelp.com/dataset_challenge/
2. Samples for users of the Yelp Academic Database, https://github.com/Yelp/dataset-examples
3. Yelp Challenge, University of Washington Student Paper 1 http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p08-fants.pdf
4. Yelp Challenge, University of Washington Student Paper 2, http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p10-michelmj.pdf