Project 1
Yelp: Intro to SQL

Out: *September 24th, 2014*
Due: *October 1st, 2014, 11:59 P.M.*

1 Introduction

In this assignment, you will familiarize yourself with SQL queries and learn how to interact with a database through Java. You will compose five SQL queries on Yelp business data.

2 Overview of the Data

You will be using a small subset of Yelp's Academic Dataset, which provides data and reviews of businesses around 30 U.S universities, including Brown University. The TAs have already parsed the data into a SQLite database, which can be found in `/course/cs127/pub/yelp/yelp.db`.

For more information about Yelp's Academic Dataset, see `https://www.yelp.com/academic_dataset`.

2.1 Database Schema

Below is the database schema for three tables (`business`, `review` and `user`) along with the datatype for each column (field). For more information about datatypes in SQLite, you can refer to `http://www.sqlite.org/datatype3.html`.

2.1.1 business

```
‘id’: a unique identifier for this business (VARCHAR),
‘name’: the full business name (VARCHAR),
‘full_address’: localized address (VARCHAR),
‘city’: city (VARCHAR),
‘state’: state (VARCHAR),
‘latitude’: latitude (REAL),
‘longitude’: longitude (REAL),
‘stars’: star rating, rounded to half-stars (REAL),
‘review_count’: review count (INT),
‘open’: is the business still open for business? (INT),
‘photo_url’: photo url (VARCHAR)
```

2.1.2 review

```
'business_id': the identifier of the reviewed business (VARCHAR),
'user_id': the identifier of the authoring user (VARCHAR),
'stars': star rating, integer 1-5 (INT),
'text': review text (TEXT),
'useful_votes': count of useful votes (INT),
'funny_votes': count of funny votes (INT),
'cool_votes': count of cool votes (INT)
```

2.1.3 user

```
'id': unique user identifier (VARCHAR),
'name': first name, last initial, like 'Matt J.' (VARCHAR),
'review_count': review count (INT),
'useful_votes': count of useful votes across all reviews (INT),
'funny_votes': count of funny votes across all reviews (INT),
'cool_votes': count of cool votes across all reviews (INT)
```

3 Tools

3.1 SQLite

SQLite is installed on all Sunlab machines. It can be accessed from the command line using `sqlite3`. To load a pre-existing database (for example, the Yelp database), simply run

```
sqlite3 /course/cs127/pub/yelp/yelp.db
```

For more information on using SQLite from the command line, see `http://www.sqlite.org/sqlite.html`

3.2 SQLite Manager

SQLite Manager is a firefox add-on. It allows you to view the tables in the database and execute test queries. To get the SQLite Manager add-on, see `https://addons.mozilla.org/en-us/firefox/addon/sqlite-manager/`

To connect the database to the SQLite Manager:

1. Open firefox and navigate to Tools → SQLite Manager.
2. Navigate to Database → Connect Database.
3. Browse to `/course/cs127/pub/yelp`, change the file type selector from *.sqlite to All Files, and open `yelp.db`.

3.3 JDBC

We will be using SQLite database through a JDBC driver (Java Database Connectivity). The JDBC SQLite driver is included as a part of your stencil code.

There will not be an official help session on how to use JDBC, but TAs will be happy to answer questions on hours or via email. Students are highly encouraged to check out `http://web.archive.org/web/20100814175321/http://www.zentus.com/sqlitejdbc/`, which has a wonderful tutorial on working with JDBC and SQLite.

4 Working on the Project

4.1 Getting Started

To get started with the Java stencil, copy `/course/cs127/pub/yelp/stencil.tgz` into your course directory, and unpack it with `tar -xvzf stencil.tgz`.

4.2 Importing into Eclipse

1. Expand the stencil code inside your course directory. That should create a directory named “yelp”
2. Open Eclipse. From the top menu bar, navigate to File → New Java Project.
3. From there, uncheck “Use default location.”
4. Browse to the yelp directory inside your course directory. Click Finish.

5 The GUI Application

5.1 How to Run

To launch the GUI, either run App.java in Eclipse or run with the command `ant run` in the project directory.

5.2 How to Use

On the left pane, the application displays business information. Only seven businesses are shown. Note that the business names are clickable and once clicked, it will open your default browser to load an actual business page on Yelp.

On the right pane, the application displays the corresponding reviews for the selected business. Only seven reviews are shown. The “User Avg” button serves to provide the average rating for all reviews written by the user. The user names are clickable and once clicked, it will open your default browser to load an actual user page on Yelp.

In the bottom panel, the three buttons serve to provide different ranking algorithms for displaying seven businesses on the left pane.

5.3 Demo

Students are highly encouraged to check out the demo before starting the assignment. The demo can be found in `/course/cs127/pub/yelp/demo/demo.jar` and it can be run with the command `java -jar demo.jar`.

6 Your Assignment

Your task is to set up a connection to the database and write five queries to answer the following questions. All queries should be composed of a single SQL statement. All of your Java code should be written in DBStudentController.java, a stencil file provided for you.

For each query method, you will have to do two things. First, you will have to write an appropriate SQL query and execute it to return a ResultSet. **You must use PreparedStatement to execute the queries.** PreparedStatement is an efficient way to execute queries, especially for the ones that require inputs. More information can be found in `http://web.archive.org/web/20100814175321/http://www.zentus.com/sqlitejdbc/`. Second, you will need to extract the data from the ResultSet, create an appropriate data type to store them, and return it accordingly.

6.1 Queries

1. Get the businesses in Providence, RI that are still open. Results should be sorted by review counts in descending order. Return top 7 businesses.

   Input: N/A

   Output: Six columns - the business id, name, full address, review count, photo url, and stars of the business.

2. Get the reviews for a particular business, given the business ID. Results should be sorted by the review's useful vote counts in descending order. Return top 7 reviews.

   Input: Business ID

   Output: Four columns - the user id, name of the user, stars of the review, and text of the review.

3. Find the average star rating across all reviews written by a particular user.

   Input: User ID

Output: One column - the average star rating.

4. Get the businesses in Providence, RI that have been reviewed by more than 5 ‘elite’ users. Users who have written more than 10 reviews are called ‘elite’ users. Results should be ordered by the ‘elite’ user count in descending order. Return top 7 businesses.

   Input: N/A

   Output: Seven columns - the business id, business name, business full address, review count, photo url, stars, and the count of the ‘elite’ users for the particular business.

5. Get the businesses in Providence, RI that have the highest percentage of five star reviews, and have been reviewed at least 20 times. Results should be ordered by the percentage in descending order. Return top 7 businesses.

   Input: N/A

   Output: Seven columns - the business id, business name, business full address, review count, photo url, stars, and percentage of five star reviews

6.2 Hint

To return top n rows in SQLite, you can use the `LIMIT` clause (`http://www.sqlite.org/syntaxdiagrams.html#select-stmt`). For example, `SELECT * FROM business LIMIT 10;` will return the top 10 rows in the `business` table.

7 Handin

You can handin your project by running the following command from the directory containing all your files:

```
/course/cs127/bin/cs127_handin yelp
```