CptS - 451 Introduction to Database Systems
Online

Project Description

In your semester long CptS 451course project you will develop a data search application for Yelp.com's business review data. The emphasis will be on the database infrastructure of the application.

Learner Objectives:

At the conclusion of this assignment you will gain experience in:

- ✓ Database modeling and design
- ✓ Populating the database with large datasets
- ✓ Querying large databases
- ✓ Optimizing query performance through indexes
- ✓ JSON parsing
- ✓ Database Application Development

Overview:

In 2013, Yelp.com has announced the "Yelp Dataset Challenge" and invited students to use this data in an innovative way and break ground in research. In your project you will query this dataset to extract useful information for local businesses and individual users.
The Yelp data is available in JSON format. The original Yelp dataset includes and **5.9M** reviews by **1.5M** users for **188K** businesses from United States, Canada, UK, and Germany. (https://www.yelp.com/dataset) In your project you will use a smaller dataset that your instructor created. This simplified dataset includes only **11,481** businesses, **192,999** users, and **416,490** reviews written for those businesses.

You will be given sample code (Python) to parse some of the Yelp JSON files (available on Blackboard). The Yelp dataset that you will use in this project is available as an attachment to the Project Description course item on the Learning Management Service website.
(Note: Please do not use the database from the Yelp.com website)
See Appendix-B for an overview of the Yelp Academic Dataset.

In addition, you will use a US Census zipcode dataset which provides the population and average income for each US zipcode area in the US. This dataset is available as an attachment to the Project Description course item on the Learning Management Service website. (You need to run the insert statement in the zipData.sql file to insert the zipcode data.)

Requirements:

You will develop a target application which runs queries on the Yelp data and extracts useful information. The primary users for this application will be customers/investors that would like to get information about businesses in a given location.
Using this application, the users can gather information about:

- the businesses in a given state, city, and/or zipcode,
- the businesses that belong to a certain category,
- the overpriced, popular, and successful businesses,
- etc.

You will implement this project as a standalone Python application.

A detailed description of the application and example screenshots are available in Appendix-A. In evaluating your work instructor's primary focus will be primarily on how you design your database and how efficiently you can search the database. However, your GUI should provide the basic functionality for easy search of the business. Creativity is encouraged! Additional functionality will be considered for extra credit.

You will be given more detailed milestone descriptions when they are assigned.

Submission Instructions:

You will submit the deliverables for milestones on **Blackboard (learn.wsu.edu)**. For each milestone you will create a .zip files that contains all deliverables for that milestone, name the .zip files as *<yourteamname>_milestoneX.zip*, and submit it to the corresponding milestone dropbox on Blackboard. Specific submission details for each milestone will be provided under milestone descriptions.

Project Milestones:

I. Milestone-0: (no submission required)

Download and install PostgreSQL Database Server. You may download the latest version from the link https://www.postgresql.org/

II. Milestone-1:

1) Parse JSON Data:

   Download the Yelp dataset from the Project Description course item on the Learning Management Service website. 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. The milestone-1 description will specify which data items you shouldn't parse in the *business, review, check-in, user* JSON objects.

   Download the sample program from Blackboard (*Project/ Sample JSON Parsing Code*). The sample code:
   - reads JSON objects form a file and extracts certain key and value pairs from JSON objects, and
   - writes the 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: business categories and 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, user, review, and checkin JSON objects.

2) i) Design a database schema that models the database for the described application scenario in Appendix-A and provide the ER diagram for your database design. Your database schema doesn't necessarily need to include all the data items provided in the JSON files. Your schema should be precise but yet complete. It should be designed in such a way that all queries/data retrievals on/from the database run efficiently and effectively. In Milestone2 you will revise your ER model.

   ii) Translate your ER model into relations and produce DDL SQL statements for creating the corresponding tables in a relational DBMS. Note the constraints, including key constraints, referential integrity constraints, not NULL constraints, etc. needed for the relational schema to capture and enforce the semantics of your ER design.

3) Build a very simple database application:

Download the "Milestone1DB.csv" file from the Milestone-1 folder, available in the Project Resources section of the online course space. Create a database on PostgreSQL with name "milestone1DB" and create a table named "business". You will import the CSV file into this table. Detailed instructions are available in Milestone-1 specification. (Note that the schema of this table should comply with the columns of the CSV file.)

Write a simple application (either web or standalone) which connects to the milestone1DB database and runs simple queries on the business table. The goal of this exercise is to get you started in database programming early on. In Milestones 2 and 3, you will develop a larger application with all required features.

The instructor will provide a video which explains how to establish connectivity with PostgreSQL in Python using `psycopg2` driver. Instructor will provide the queries you need to run on your table (see Milestone 1 specification).

Milestone-1 Deliverables:

1. (25%) Source code for parsing all JSON data. Only submit your source code, **not the data files**.
2. (40%) The E-R diagram for your database design. To create your ER diagram, I suggest you to use Edraw Max (https://www.edrawsoft.com/download-edrawmax.php) . You may also use your favorite drawing tool (e.g., Visio, Word, PowerPoint). Should be submitted in .pdf format. Name this file "*<your-team-name>_ER_v1.pdf*"
3. (35%) Source code for your application. Only submit your source code, **not the data files**.

Create a zip archive *"<your-team-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.

You will demonstrate your Milestone1 to the instructor and the TA.

III. Milestone-2: (Deadline TBA)

1) Revise your database schema (ER model and relations).
2) Populate your database with the Yelp data. Generate INSERT statements for your tables and run those to insert data into your DB. You will also write and additional scripts to update the information stored in your database.

   Write triggers and assertions to ensure the validity and consistency of the information stored in your database. Details will be available in Milestone2 specification.
3) Write a 1- to- 2 page paper describing the metrics you proposed for classing businesses as popular or successful.
4) Build the alpha-prototype of your application.

Milestone-2 Deliverables:
(Weights of the deliverables are TBA)

1. The revised E-R diagram. **Should be submitted in .pdf format.** Name this file *"<your-team-name>_ER_v2.pdf"*
2. SQL script file containing all SQL statements (i.e., CEATE TABLE statements, UPDATE statements, and TRIGGERS) . Name this file *"<your-team-name>_SQL.sql"*
3. The code of your application.

Create a zip archive *"<your-team-name>_milestone2.zip*" that includes your ER diagram and SQL script files. Upload your milestone-2 submission on Blackboard until the deadline.
You will demonstrate your Milestone2 to the instructor and the TA.

IV. Milestone-3: (Deadline: TBA)
In this milestone you will complete the implementation of your application. A detailed description of the application requirements is provided in Appendix-A.

Milestone-3 Deliverables:
(Weights of the deliverables are TBA.)

1. The source code of your application. **Please only upload your source code, not your DB files.s**

Create a zip archive *"<your-team-name>_milestone3.zip*" that includes your source code. Upload your milestone-3 submission on Blackboard until the deadline.

You will demonstrate your final project to the instructor and the TA. The demonstration schedule will be announced in mid-April.

References:

1. Yelp Dataset Challenge, https://www.yelp.com/dataset
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

Appendix-A

Application Specification

The primary users for this application will be potential customers seeking for businesses. Using this application, the users can gather information about:

- the businesses in a given state, city, and/or zipcode,
- the businesses that belong to a certain category,
- the population, avg income per person in a zipcode,
- the overpriced, popular, and successful businesses in a zipcode, etc.

You should implement this project as a standalone Python application. Below you will find screenshots to help you visualize the required functionality.

Business Search:

Using this application, user can search for the businesses which are within a certain state, city, zip and which belong to a certain category. In addition, the application analyzes the properties of each business in the zipcode and estimates which of the businesses are considered successful or popular. See below for more details.

Use Cases:

1. User selects a state, city, and zipcode. When search button is pressed, the businesses in that state/city/zipcode are displayed (see Figure-2). The following information is provided for each business returned in the search result.
   - Business name
   - Address and city
   - Business rating (stars)
   - # of reviews provided for the business
   - Average rating (stars) of the reviews provided for the business
   - Total number of check-ins

     (Note: You should (i) *query* t*he review table to calculate the number of reviews and avg review rating* and (ii) *query the check-in table to calculate the number of check-ins for each business* and (iii) *update those attribute values in the business table*.)

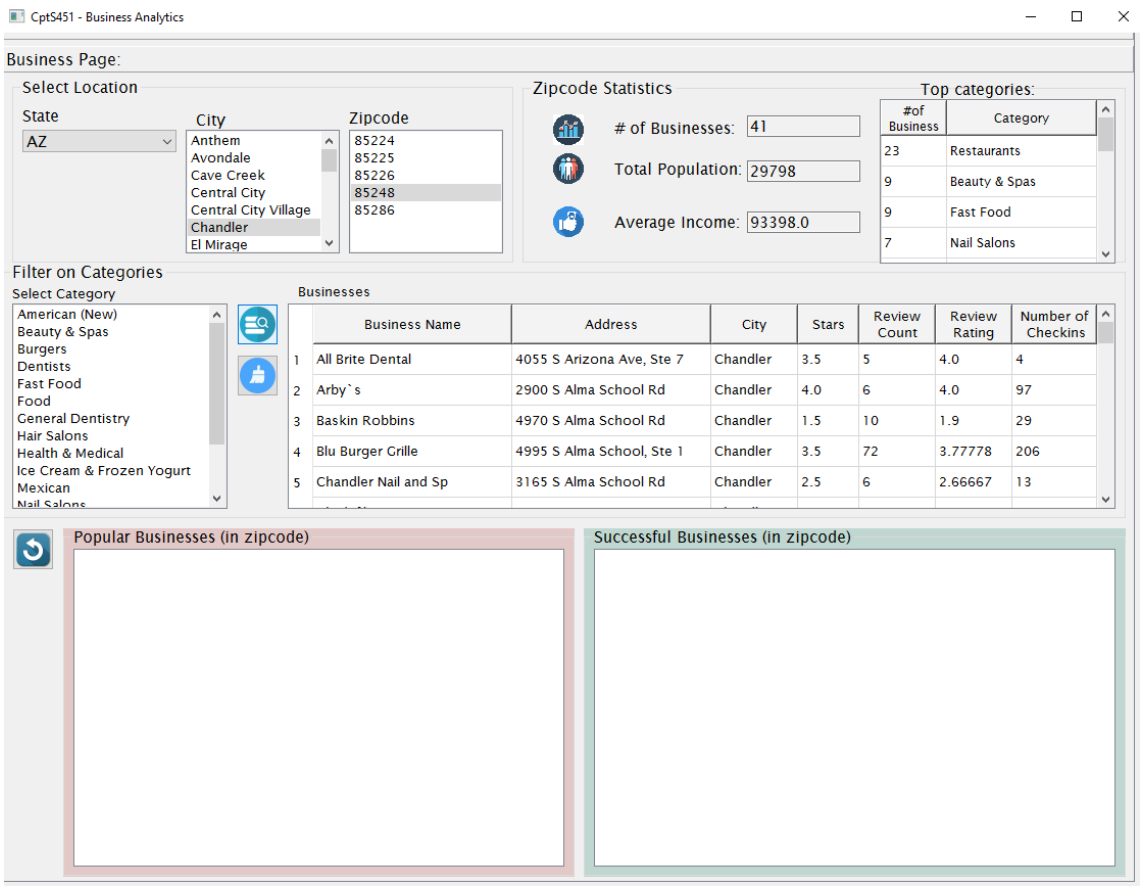

Figure 2 – Searching for the businesses in a Chandler, AZ, 85248

2. When the user selects a zipcode (in use case-1), the following information for the selected zipcode will be displayed (see Figure-2 above):
   - Total number of businesses in the zipcode,
   - The total population of the zipcode
   - The average income of the employees in the zipcode.

   The population and average income information are provided in the US Census dataset.

3. The user might refine the results by specifying a business category. The search will return the businesses which belong to the selected category (see Figure-3).

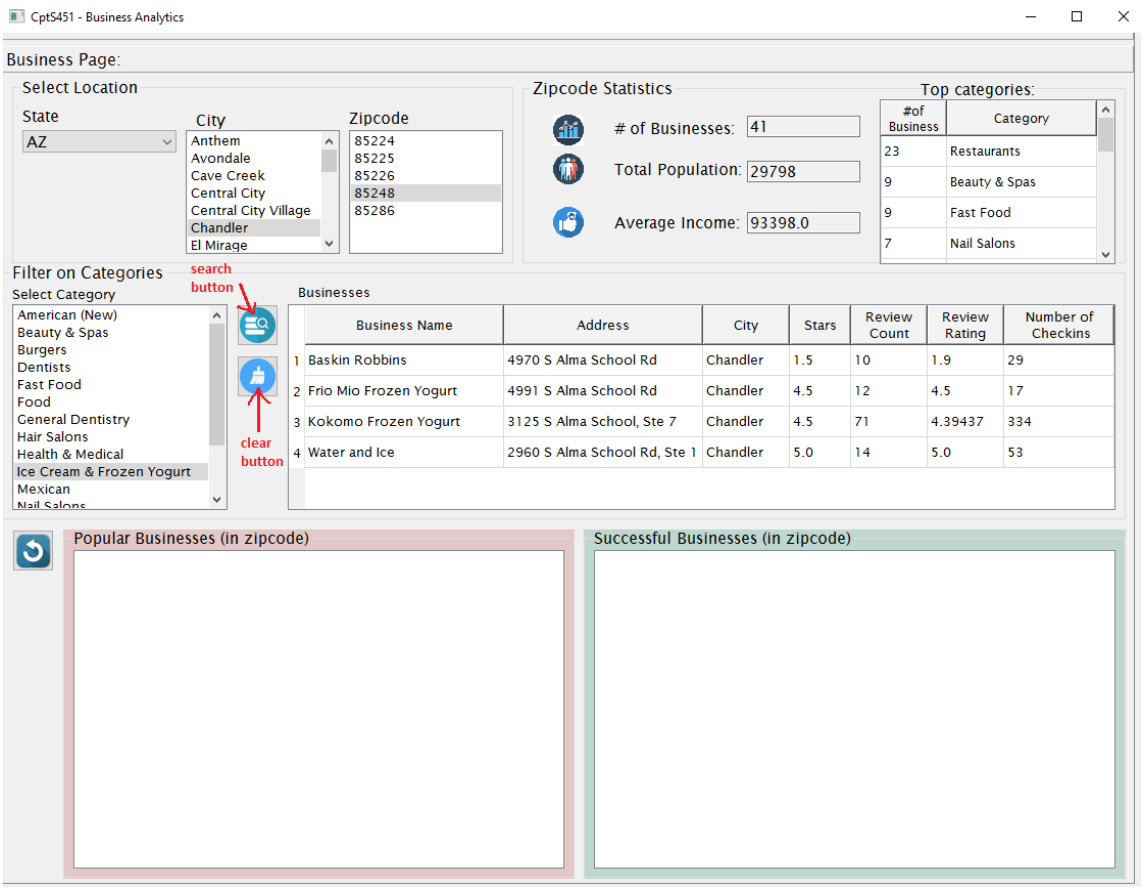

Figure 3 – Searching for the businesses with category 'Ice Cream & Frozen Yogurt' in Chandler, AZ, 85248.

4. Another major use case for this application is analyzing the business properties and identifying the businesses that may be classified in one of the following groups.
   - Popular businesses that seem to attract more customers compared to other businesses in the same category.
   - Successful businesses which have been serving the community for a long time and which have loyal customers.

   In the above description, the terms "popular" and "successful" are vaguely defined. You need to propose and formulate your own metrics for classifying the businesses into these three groups and specify which information you will use in your analysis. In milestone 2, you will write a 1 to 2 page long paper where you describe your proposed metrics in detail. And in your application, you will implement these metrics (in terms of SQL queries) and query the business data to identify the successful, and popular businesses.
   (*Note that you need pre-analyze the Yelp data and extract some additional features/details about the businesses (by running some SQL queries), store those extracted information in the database, and use them in your queries for finding the popular and successful businesses.*)

   Figure-4 shows an example classification of businesses in the selected zipcode. The sample implementation doesn't apply any specific metrics but classify the businesses based on the values of certain business attributes. Your results for the same zipcode will be different in your application. Also, you may display different set of attributes in your results.

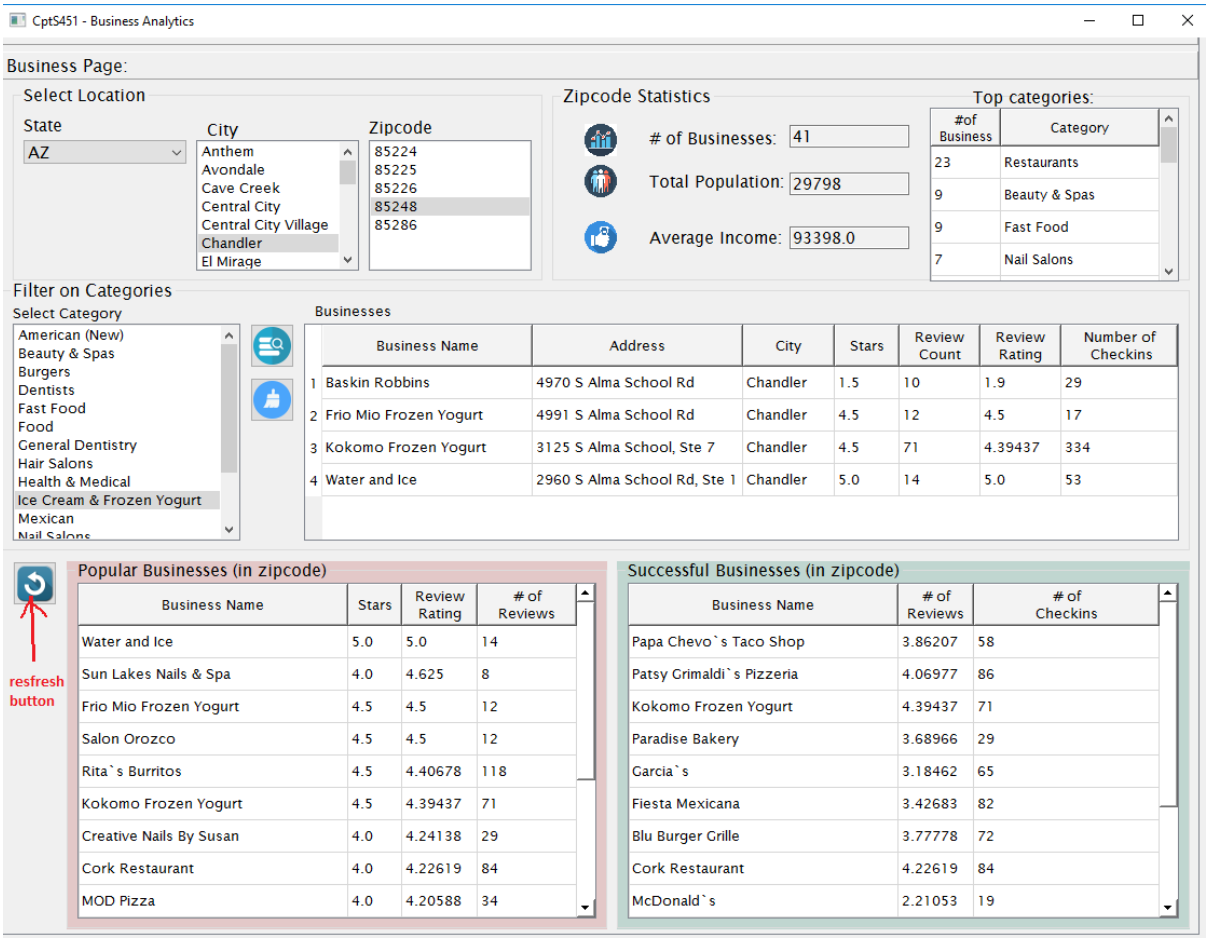

Figure 4 – When the "refresh" button is pressed, the overpriced, popular and successful businesses in the selected zipcode are displayed.

Please note that all data should be kept in the database and should be retrieved from the database when needed. You are not allowed to create internal data structures to store the data.

Appendix-B

Yelp's Academic Dataset

Yelp has made available a dataset which contains user reviews for **188K** businesses from United States, Canada, UK, and Germany. The purpose was to provide a real-world data set to promote research in various areas of research. The dataset includes 6 types of data objects: *business*, *review*, *user*, *tip*, *check-in, and photos*. Every object contains a 'type' field, which tells whether it is a *business*, a *user*, or a *review*. *Business* objects contain basic information about local businesses. *Review* objects contain the details of the reviews by users for the businesses. *Review*'s user_id associates the reviews with the *user* objects. Similarly, *review*'s business_id associates each review with the *businesses*.

Detailed description of the data objects is available at:
https://www.yelp.com/dataset/documentation/json

In your project, you will only parse *business*, *user*, *review,* and *check-in objects.*
Review data: You don't need to store the 'elite' and 'compliments' information in your database.
Business data: You don't need to store business open/close times, neighborhood, and lat/long coordinates. Also you should only store the business attribute information that may be helpful in identifying successful, popular, and overprices businesses (see usecase B.4). You can exclude the attributes that you don't need.

Usage of this dataset is governed by the Academic Dataset Terms of Use.