

CECS 521 - Database Architecture

Spring 2016

Homework Assignment 3

Due: Friday, April 1
@ 11:59pm

In your course project you would develop a data analysis application for Yelp.com's business review data. The emphasis would be on the database infrastructure of the application.

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 this project you would 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 42,153 businesses, 252,898 users, and 1,125,458 reviews from Phoenix (AZ), Las Vegas (NV), Madison (WI) in United States and Waterloo (ON) and Edinburgh (ON) in Canada. (http://www.yelp.com/dataset_challenge/). In your project you will use a smaller and simplified dataset. This simplified dataset includes only **20,544** businesses, the reviews that are written for those businesses only, and the users that wrote those reviews.

The Yelp JSON files that you will use in this project are available on beachboard.

(Note: Please make sure to use the dataset available on beachboard, not the one from the Yelp.com website)

See Appendix-A for an overview of the Yelp Academic Dataset.

Overview & Requirements:

You would develop a target application which runs queries on the Yelp data and extracts useful information. The primary users for this application will be potential customers seeking for businesses that match their search criteria. Your application will have a user interface that provides the user the available business categories (main and sub-categories) and the attributes associated with each business category. Using this application the user will search for the businesses from various business categories that have the properties (attributes) the user is looking for. The user can filter the search results by state, city or zipcode. In addition, the user may choose to view only the businesses that are open or closed on certain days of the week. The application should also allow the user to view the reviews provided for each business.

You will be designing your application a standalone Java application.

Example screenshots of a possible application are available in Appendix-B. 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 and pull out the information. However your GUI should provide the basic functionality for easy browsing of the business categories and attributes (as illustrated in Appendix-B). Creativity is encouraged! Additional functionality will be considered for extra credit.

Project Details:

I. Part 1

- Download the Yelp dataset from Camino. 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 (For example, categories, attributes,...etc.)
- You may have to modify your database design from Homework 2 to model the database for the described application scenario on page-1. 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.
- 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.
- Populate your database with the Yelp data. Generate INSERT statements for your tables and run those to insert data into your DB.

II. Part 2

Implement the application for searching local businesses as explained in section "Overview & Requirements". In this milestone you would:

- Write the SQL queries to search your database.
- Establish connectivity with the DBMS.
- Embed/execute queries in/from the code. Retrieve query results and parse the returned results to generate the output that will be displayed on the GUI.
- Implement a GUI where the user can,
 - o browse through main and sub-categories for the businesses; select the business categories that user wants to search for; (*note: The list of the main categories is given in Appendix-C. All other categories that appear in the business objects are sub-categories. Such a distinction is made for easier browsing of the business categories.*)
 - o browse through the attributes of the selected categories; specify the attributes she/he is interested in;
 - o search for the businesses that belong to the main and sub-categories that user specifies and have the attributes that user selects, (The user should be able to search for the businesses that have either all the attributes (AND condition) or that have any of the attributes specified (OR condition))
 - o filter out the search results for a specific state, city, or zipcode,
 - o filter out the businesses which are open or closed on specific days of the week,
 - o select a certain business in the search results and list all the reviews for that business. (*note: The review list should also include the names of the users who provided those reviews*)

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

Required .sql files:

You are required to create two .sql files:

1. createdb.sql: This file should create all required tables. In addition, it should include constraints, indexes, and any other DDL statements you might need for your application.
2. dropdb.sql: This file should drop all tables and the other objects once created by your createdb.sql file.

Required Java Programs:

You are required to implement two Java programs:

1. populate.java: This program should get the names of the input files as command line parameters and populate them into your database. It should be executed as:
"> java populate yelp_business.json yelp_review.json yelp_checkin.json yelp_user.json".
Note that every time you run this program, it should remove the previous data in your tables; otherwise the tables will have redundant data.
2. hw3.java: This program should provide a GUI, similar to figure 1, to query your database. The GUI should include:

- a. List of main business categories.
- b. List of sub-categories associated with the selected main category(ies).
- c. List of the attributes associated with the selected sub-categories.
- d. A dropdown menu to filter results based on days and hours the business is open/closed.
- e. List of results.
- f. List of the reviews provided for a specific business.

Grading guideline:

Points	
5	Creating/Dropping database tables
10	Populating database
10	JSON parsing
20	GUI containing all of requirements mentioned for user interfaces.
15	Listing of Category/Subcategory/Attributes
20	Filtering Search/All/Any attributes, city/state/zip and days of the week
10	Listing of reviews for a business
10	List of results

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>

Appendix-A

Yelp's Academic Dataset

Yelp has made available a dataset which contains user reviews for 42,153 businesses in Phoenix (AZ), Las Vegas (NV), Madison (WI) in United States and Waterloo (ON) and Edinburgh (ON) in Canada. The purpose was to provide a real-world data set to promote research in various areas of research. The dataset includes 5 types of data objects: *business*, *review*, *user*, *tip*, and *check-in*. 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*.

The fields of objects are given below:

Business Objects

Business objects contain basic information about local businesses.

```
{
  'business_id': (encrypted business id),
  'full_address': (localized address),
  'hours': (the days of the week when business is open; the opening and closing times on those days)
  'open': True / False (corresponds to closed, not business hours),
  'categories': (categories associated with the business)
  'city': (city),
  'state': (state),
  'latitude': latitude,
  'longitude': longitude,
  'review_count': review count,
  'name': (business name),
  'neighborhoods': [(hood names)],
  'stars': (star rating, rounded to half-stars),
  'attributes': (business properties),
  'type': 'business'
}
```

Review Objects

Review objects contain the review text, the star rating, and information on votes Yelp users have cast on the review. Use *user_id* to associate this review with others by the same user. Use *business_id* to associate this review with others of the same business.

```
{
  'votes': {
    'useful': (count of useful votes),
    'funny': (count of funny votes),
    'cool': (count of cool votes)
  }
  'user_id': (the identifier of the authoring user),
  'review_id': (the identifier of the reviewed business),
  'stars': (star rating, integer 1-5),
  'date': (date, formatted like '2011-04-19'),
  'text': (review text),
  'type': 'review',
  'business_id': (the identifier of the reviewed business)
}
```

User Objects

User objects contain aggregate information about a single user across all of Yelp (including businesses and reviews not in this dataset).

```
{
  'yelping_since': (the date when user account was created)
  'votes': {
    'useful': (count of useful votes across all reviews),
    'funny': (count of funny votes across all reviews),
    'cool': (count of cool votes across all reviews)
  }
}
```

```

    }
    'review_count': (review count),
    'name': (first name, last initial, like 'Matt J.'),
    'user_id': (unique user identifier),
    'friends': (friends of the user),
    'fans': (number fans of the user),
    'average_stars': (floating point average, like 4.31),
    'type': 'user',
    'compliments': (comments from other users),
    'elite': ()
}

```

Checkin

```

{
  'type': 'checkin',
  'business_id': (encrypted business id),
  'checkin_info': {
    '0-0': (number of checkins from 00:00 to 01:00 on all Sundays),
    '1-0': (number of checkins from 01:00 to 02:00 on all Sundays),
    ...
    '14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),
    ...
    '23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)
  } # if there was no checkin for a hour-day block it will not be in the list
}

```

Tip

```

{
  'user_id': (encrypted user id),
  'text': (),
  'business_id': (encrypted user id),
  'likes': (),
  'date': (),
  'type': 'tip'
}

```

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

Appendix-B

Sample Application

The screenshot shows the 'Yelp Tool' window. On the left, under the 'Start' tab, there is a list of main categories with checkboxes. The categories are: Active Life, Arts & Entertainment, Automotive, Car Rental, Cafes, Beauty & Spas, Convenience Stores, Dentists, Doctors, Drugstores, Department Stores, Education, Event Planning & Services, Flowers & Gifts, Food, Health & Medical, Home Services, Home & Garden, Hospitals, Hotels & Travel, Hardware Stores, Grocery, Medical Centers, Nurseries & Gardening, Nightlife, Restaurants, Shopping, and Transportation. The 'Restaurants' category is selected. To the right of the categories list are two empty columns. At the bottom, there are input fields for 'Day of the week', 'From:', 'To:', and 'Search for:', along with 'Search' and 'Close' buttons.

Figure 1- List the main categories that appear in "business" data.

The screenshot shows the 'Yelp Tool' window with the 'Restaurants' category selected. The left pane now displays sub-categories and attributes. The sub-categories listed are: Kosher, Landmarks & Historical Bui, Laotian, Latin American, Lebanese, Leisure Centers, Live / Raw Food, Local Flavor, Local Services, Lounges, Malaysian, Mass Media, Meat Shops, Medical Spas, Mediterranean, Mexican (selected), Middle Eastern, Modern European, Mongolian, Moroccan, Music Venues, Outlier Stores, Pakistani, Party & Event Planning, Patisserie / Cake Shop, Performing Arts, Persian / Iranian, Personal Chefs, Personal Shopping, and Peruvian. The attributes listed are: Accepts Credit Cards, Accepts Credit Cards, false, Accepts Insurance, false, Ages Allowed_18plus, Ages Allowed_21plus, Ages Allowed_allages, Alcohol_beer_and_wine, Alcohol_full_bar, Alcohol_none, Ambience_casual, Ambience_classy, Ambience_divey, Ambience_hipster, Ambience_intimate, Ambience_romantic, Ambience_touristy, Ambience_trendy, Ambience_upscale, Attire_casual, Attire_dressy, Attire_formal, By Appointment Only, By Appointment Only, false, BYOB, BYOB/Corkage_no, BYOB/Corkage_yes_corkage, BYOB/Corkage_yes_free, BYOB_false, Caters, and Caters_false. The right pane shows a table with columns: Business, City, Sta..., and Stars. The bottom of the window has the same input fields and buttons as Figure 1.

Figure 2 - List the sub-categories associated with the selected main category(ies). List the attributes associated with the selected sub-categories.

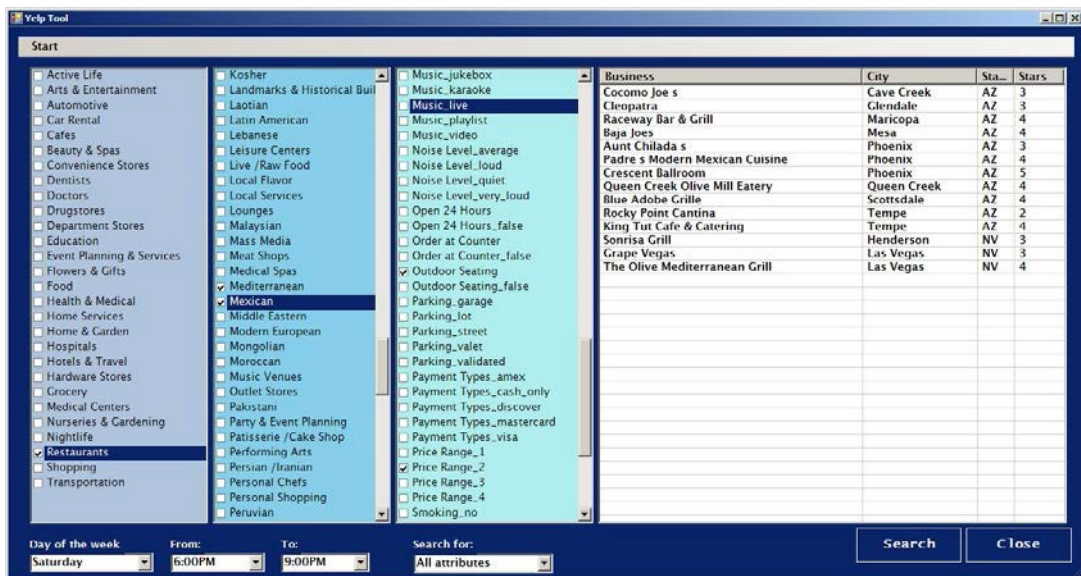


Figure 3 - Search for businesses for the selected sub-categories which have "all" the selected attributes and which are open on Saturday between 6PM and 9PM.

ReviewDate	Stars	Review Text	UserID	Useful Votes
10/3/2011 ...	5	Great sound, nicely laid out space, good service, solid beers on tap... This...	Andrew	7
10/4/2011 ...	5	Finally downtown has a great music venue and Crescent Ballroom could n...	Eric	1
10/4/2011 ...	5	Me: How long have you guys been open? Bartender: About... four hours. ...	Ana	2
10/4/2011 ...	5	This is the new mecca of Phoenix music. Charlie Levy who booked for th...	Dave	11
10/4/2011 ...	5	Frankly, I think it's going to be hard to see anyone anywhere else after yo...	Steve	0
10/5/2011 ...	5	All I can say is LOVE, LOVE, LOVE that Phoenix finally has something to b...	Will	0
10/6/2011 ...	5	I haven't had a chance to see a show here, but I can't wait to. Talking wit...	B	2
10/7/2011 ...	5	What fantastic people! Met the general manager and she couldn't be more...	Holly	0
10/8/2011 ...	5	... so you have been frustrated with the lack of any good venues in Phoe...	Matthew	1
10/8/2011 ...	5	Great space, great sound, great owner. Best place in Phoenix to see a ban...	Jim	0
10/11/2011 ...	5	I can't say I am an expert critic at music venues by any means, but what I ...	Lindsey	4
10/16/2011 ...	5	A great new venue! We really needed a new music venue here, and Cresce...	Aaron	0
10/21/2011 ...	5	Went with my daughter to the St. Vincent /Cate Le Bon show at this venue...	Julie	2
10/22/2011 ...	5	I can't believe what a great venue the Crescent Ballroom is. It's as if the o...	Becky	3
10/24/2011 ...	5	Love it! Great atmosphere and drink selection. Also, the menu looked am...	courtney	0
10/26/2011 ...	5	Not a single complaint about this place! We decided to bar hop a bit bef...	Anne	0
11/3/2011 ...	5	This place is going to put PHX on the map. Wide variety of music every ni...	ROYAL	0
11/8/2011 ...	5	Music venue: I believe this is the best one in town. Perfect combination of...	Jennifer	0
11/10/2011 ...	5	Only been here once, but I absolutely love this place. It almost feels like a...	Dace	3
11/18/2011 ...	5	Finally, a decent room for a mid-sized club acts. My new favorite venue. ...	Shaq	0
11/21/2011 ...	5	I am so grateful this place has opened! I went to see Phantogram here a c...	Liz	0
11/29/2011 ...	5	Fun venue, saw phantogram here and really enjoyed it, spacious and soun...	Becca	0
12/2/2011 ...	5	This is the venue downtown phoenix has been waiting for! Go for a show...	Adrian	0
12/3/2011 ...	5	Let me preface this review on my background. I own a audio and lighting ...	Travis	0
12/18/2011 ...	5	What a great addition to downtown Phoenix!! Now, I have yet to see a sho...	Kat	4
12/22/2011 ...	5	I love this place! Just what downtown Phx needed. Great food and a great ...	Lesley	1
12/22/2011 ...	5	67 Tempeville, simply amazing! 62, 60, great sound, service, and show...	Chloe	0

Figure 4 - When clicked on a business, the reviews provided for that business and the users who wrote those reviews are listed.

Appendix-C

Main Business Categories

1. Active Life
2. Arts & Entertainment
3. Automotive
4. Car Rental
5. Cafes
6. Beauty & Spas
7. Convenience Stores
8. Dentists
9. Doctors
10. Drugstores
11. Department Stores
12. Education
13. Event Planning & Services
14. Flowers & Gifts
15. Food
16. Health & Medical
17. Home Services
18. Home & Garden
19. Hospitals
20. Hotels & Travel
21. Hardware Stores
22. Grocery
23. Medical Centers
24. Nurseries & Gardening
25. Nightlife
26. Restaurants
27. Shopping
28. Transportation