COEN 280 - Database Systems Winter 2022

Homework Assignment 3

Due: Friday, Mar 8 @11:59pm Demo: TBD

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 Camino. (Note: Please make sure to use the dataset available on Camin, 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 and users that match their search criteria. Your application will have a user interface that provides the user the available business categories (main, sub-categories), the attributes associated with each business category along with business review and yelp user information 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.

Faceted search has become a popular technique in commercial search applications, particularly for online retailers and libraries. It is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple **filters**. Faceted search is the dynamic clustering of items or search results into categories that let users drill into search results (or even skip searching entirely) by any value in any field. Users can then "drill down" by applying specific constraints to the search results. Look at https://react.rocks/tag/Faceted Search for some examples.

In this application, the user can filter the search results using available business attributes (i.e. facets) such as category, sub-category, attributes, reviews, stars and votes. Each time the user clicks on a facet value; the set of results is reduced to only the items that have that value. Additional clicks continue to narrow down the search—the previous facet values are remembered and applied again.

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!

Project Details:

1. Part 0

Install Oracle Database 11gR2 or later. Consult the instructions provided on Camino under Assignment 3. You will be installing a docker container with an Oracle Database install on this environment that you will be using for your assignment.

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.
- After you populated your database, created indexes on frequently accessed columns of its tables using CREATE INDEX statement. This will help speed up query execution times. You have some flexibility about which indexes to choose.

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.
- Business Search: Implement a GUI where the user can search for movies that match the criteria given.
 - Browse through main categories for the businesses (See Appendix C); select the business attributes 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.
 - The usage flow of the GUI is as follows:
 - 1) Once the application is loaded, main categories are loaded from the backend database. Note that, selection of business main categories (single or multiple) is required. For instance, assume that use selects *Restaurants* as the main category.
 - 2) The subcategories matching the main category selection will be listed under subcategories column. Since user selected *Restaurants* in previous step, only sub-categories values that its main category is *Restaurants* should appear in the sub-categories panel. Note how faceted search work here. After step 1, the set of results is reduced to only the businesses that belong to *Restaurants* category. The user can select desired sub-categories values. This attribute is optional in building the query. User might not select a sub-category at all. Assume that use selects *Mediterranean* as the sub-category value.
 - 3) Attribute column is the next selection. Similar to No 2, the set of results is reduced to only the subcategories **AND** main category selection. The user can select desired attribute values (single or multiple selections). This attribute is optional in building the query. User might not select a attribute values at all. Since user selected *Restaurants*, and *Mediterranean* in previous steps, only attribute values that appeared in business with main-category = Restaurants **AND** sub-

- category = *Mediterranean*, should appear in the attribute selection panel. Assume that user selects *Outdoor Sitting* as the desired attribute.
- 4) Review column is the next selection. You can specify review duration (from/to) and enter the star and vote values into the text box. The attributes under the Review column are also optional.
- 5) City textfield is the next selection. You can enter any city name in the city textfield and all businesses with specified categories, subcategories and attributes located in that city can be selected. Note that city textfield is independent of category, subcategory, and attribute selection mentioned above, which means that if you do not pick any category, subcategory and attributes, then all businesses located in that city will be selected. Note that city textfield can have no text.
- The application should be able to search for the businesses that have either all the specified values (AND condition) or that have any of the values specified (OR condition). For example:
 - if user selected AND condition, and selected *Restaurants* and *Cafes* as main categories, sub-categories of businesses that have *Restaurants* **AND** *Cafes* as main categories, should be listed in the next panel.
 - If user selected OR condition, and selected *Restaurants* and *Cafes* as main categories, subcategories of businesses that have *Restaurants* **OR** *Cafes* as main categories, should be listed in the next panel.

Note that the relation between facets (or business characteristics) is always **AND**. However, the relation between values of one facet can be set to be OR or AND.

Example:

Consider the below example on the AND/OR selection. Assume the following example:

BusinessID	Category	Sub-category
1	restaurant	Mediterranean
2	restaurant	Mexican
3	restaurant	Mediterranean

Suppose User selects Restaurant as main category and both Mediterranean and Mexican as subcategory. Also user selects AND from the "Search for" drop down menu. This means that attributes of businesses that are (Restaurant, Mediterranean) AND (Restaurant, Mexican) should appear in the attribute column.

Look for the conjunction of attributes between business 1, 2, 3 that follow the above rule. Per above example the following attributes should show in the attribute panel since they are common between all three businesses: (remember that user selected AND from the "Search for" drop down menu)

```
Ambience_Good_True
Price_Range_1_False
```

Suppose User selects Restaurant as main category and both Mediterranean and Mexican as subcategory.

Also user selects OR from the "Search for" drop down menu.

This means that attributes of business that are (Restaurant, Mediterranean) OR (Restaurant, Mexican) should appear in the attribute column. So, you have to look for disjunction of attributes between business 1, 2, 3 that follow the above rule.

Per above example, what shows in attribute panel is:

```
Music_Loud_True
Ambience_Good_True
Parking_Street_False
Price_Range_1_False
Music_Loud_False
```

Select a certain business in the search results and list the following for that business(s):

- Business, City, State, Stars
- 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)
- User Search: Implement a GUI where the user can perform a search for users that match the criteria given
 - o The usage flow of the GUI is as follows:
 - 1) User can specify user search using attributes such as member_since, review_count, number of friends, average stars and number of friends.
 - 2) Clicking on "Execute User Query" will show user matches, yelping since and average stars
 - 3) select a certain user in the search results and list all the reviews given by that user.
 - 4) Users can be selected based on vote_funny, vote_cool and vote_useful.
 - 5) Review count, number of friends, average stars, vote_funny, vote_cool and vote_useful should have 5 different comparators, which are >, >=, =, <=, <, followed by the edit text indicating a number.

Note: The user can do either a Business search or User search (not both) at any given time through the GUI. **Note:** 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.

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 businessesin 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-\overline{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 0:00 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 business id),
'likes': {},
     'date': {},
     'type': 'tip'
}
```

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

Appendix-B Active Life Arts & Entertainment Adult Education Bars
Butcher
Chicken Wings Beer, Wine & Spirits Ambience_classy_false Automotive Beauty & Spas ✓ Cafes Car Rental
Dentists CSA Chinese Ambience_hipster_true
Ambience_trendy_false Ambience_intimate_false Conven Review from 2007-01-01 -Cuban
Electronics Caters_false
Dogs Allowed_true Department Stores Doctors Cultural Center Caters_true ☐ Good For_brunch_false

☐ Good For_dinner_true Event Planning & Services Flowers & Gifts ✓ Food Stands Food Trucks Good For_dessert_false Halal Hardware Stores Health & Medica Irish Good for Kids true Has TV false Home Services
Hotels & Travel Local Flavor Noise Level_quiet
Parking_lot_false Outdoor Seating_false
Parking_lot_true Home & Garden Hospitals Pakistani Medical Centers

Nurseries & Gard ☐ Nightlife Real Estate
Specialty Sch Russian
Sports Bars Parking_valet_false Parking_validated_false
Takes Reservations_false Shopping ☐ Transportation Tires ☐ Tobacco Shops Waiter Service_true Wheelchair Accessible_fals 1 getSubCategory Las Vegas ------Figure 1- Yelp Application Main UI (Business Search) Arts & Entertainment

Active Life Beauty & Spas Cafes Car Rental -Department Stores Doctors Flowers & Gifts Event Planning & Services Food
Hardware Stores Grocery
Health & Medical Home & Garden
Hospitals Home Services
Hotels & Travel Medical Centers ☐ Nightlife Transportation Review To 2012-04-20 Shopping getSubCategory 2008-01-01 ▼ 1 Number of Friends > ▼ 2 Average Stars > 1 ▼ 3 ▼ 5 ▼ 2

Figure 2- Yelp Application Main UI (User Search)

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

- 28. Transportation

Appendix-D (Sample query output)

Test case 1:

Main Categories: Automotive, Car Rental

Search For Attribute: AND

Output:

	⊕ B_NAME	♦ CITY	♦ STATE	♦ STARS	
1	A-1 Van Rental & Leasing	Phoenix	AZ	5	
2	Big Two Toyota-Scion Of Chandler	Chandler	AZ	4	
3	Dollar Rent A Car	Las Vegas	NV	1.5	
4	Escape Eagles	Las Vegas	NV	4.5	
5	JAM Scooters	Las Vegas	NV	5	
6	Smart Motors Toyota Scion	Madison	WI	4.5	
7	Thrifty Car Rental	Glendale	AZ	2.5	
8	Thrifty Car Rental	Las Vegas	NV	1.5	

Test case 2:

Main categories: Hospitals, Health & Medical

Subcategories: Diagnostic Imaging, Diagnostic Services Search For Attribute: OR

Output:

Full Address	City	Review Count	Business Name	Longitude	Latitude	State N	Stars	Whet
9090 W Post R	Las Vegas	5	High Risk Pregnancy Center	-115.29	36.074	NV	5.0	false
16641 N 40th	Phoenix	4	Valley Radiologists Ltd	-111.99	33.638	AZ	5.0	false
7751 W Flami	Las Vegas	5	Labcorp	-115.25	36.114	NV	1.5	false
708 N. Rainbo	Las Vegas	3	Any Lab Test Now Las Vegas	-115.24	36.180	NV	3.5	false
3501 N. Scotts	Scottsdale	12	Scottsdale Medical Imaging L	-111.92	33.488	AZ	3.0	false
6591 W Thund	Glendale	3	Insight Imaging - Thunderbird	-112.20	33.610	AZ	2.5	false
4501 E Thoma	Phoenix	3	Any Lab Test Now Phoenix	-111.98	33.479	AZ	4.5	false
2767 N Tenay	Las Vegas	18	Steinberg Diagnostic Medical	-115.24	36.210	NV	2.5	false
5125 N 16th St	Phoenix	12	Baby's First Ultrasound	-112.04	33.511	AZ	3.0	false
8551 W Lake	Las Vegas	9	Pueblo Medical Imaging	-115.28	36.198	NV	3.5	false
9070 W Post R	Las Vegas	8	Steinberg Diagnostic Medical	-115.29	36.074	NV	3.5	false
6502 Grand Te	Madison	3	Storkvision	-89.492	43.059	WI	2.5	false
2850 Siena He	Las Vegas	9	Steinberg Diagnostic Medical	-115.11	36.004	NV	3.0	false
2080 W South	Apache Jun	3	Sonora Quest Laboratories	-111.56	33.393	AZ	1.5	false
6037 S Fort Ap	Las Vegas	28	Baby's First Image 3D/4D Ultr	-115.29	36.079	NV	4.5	false
2811 W Horizo	Henderson	5	Desert Radiologists	-115.11	35.998	NV	4.0	false
3483 S Mercy	Gilbert	6	Sonora Quest Laboratories	-111.74	33.286	AZ	3.0	true
565 Marks StSt	Henderson	15	Miracle In Progress 3D/4D Ult	-115.03	36.062	NV	4.0	false
2611 Horizon	Henderson	6	West Valley Imaging - Hender	-115.10	36.001	NV	3.5	false
2051 N Rainb	Las Vegas	5	The Vaccine Center and Trav	-115.24	36.198	NV	4.0	false
880 Seven Hill	Henderson	4	Wellhealth Womens Specialty	-115.12	35.995	NV	4.0	false
5040 N 15th Av	Phoenix	7	Insight Imaging - Camelback	-112.08	33.492	AZ	2.5	false
16515 S 40th	Phoenix	3	Sonora Quest Laboratories	-111.99	33.297	AZ	4.5	false
2000 E Southe	Tempe	10	Sonora Quest Laboratories	-111.90	33.393	AZ	2.0	false
500 E Windmil	Las Vegas	8	The Vaccine Center and Trav	-115.15	36.042	NV	3.5	false
7901 W Tropic	Las Vegas	3	Quest Diagnostics	-115.26	36.270	NV	2.5	false
11209 N Tatu	Phoenix	4	Sonora Quest Laboratories	-111.97	33.587	AZ	4.0	false
1409 E Lake M	North Las V	4	North Vista Hospital	-115.12	36.195	NV	3.0	false
2950 S Maryla	Las Vegas	6	Steinberg Diagnostic Medical	-115.13	36.136	NV	3.5	false
9465 E Ironwo	Scottsdale	3	Labcorp	-111.87	33.577	AZ	4.5	false
6707 N 19th Av	Phoenix	4	Sonora Quest Laboratories	-112.09	33.535	AZ	3.0	false
4220 S Grand	Las Vegas	21	Miracle In Progress 3D4D Ultr	-115.30	36.112	NV	3.5	false

2680 S Val Vist Gil	lbert	3	Simonmed	-111.75	33.302	AZ	1.5	false
2020 Palomin La	is Vegas	6	Desert Radiologists	-115.17	36.162	NV	2.5	false
2501 E Southe Te	mpe	4	LabCorp	-111.88	33.392	AZ	3.0	false
2625 Box Cany La	is Vegas	11	Katz Joel, MD	-115.24	36.207	NV	3.5	false
16222 N 59th GI	endale	4	Sun Radiology	-112.18	33.632	AZ	2.0	false
9201 E Mounta Sc	ottsdale	9	Simonmed Imaging	-111.88	33.574	AZ	3.0	false
6320 W Union Gl	endale	3	Sonora Quest Laboratories	-112.19	33.655	AZ	1.5	false
3145 E Warm La	is Vegas	8	Sneak Peek Video-Ultrasound	-115.10	36.056	NV	4.5	false
3440 W Cheye No	orth Las V	4	Labcorp	-115.18	36.217	NV	3.5	false
5410 W Thund Gl	endale	3	Simonmed	-112.17	33.611	AZ	1.5	false
4 Sunset Way He	enderson	7	Steinberg Diagnostic Medical	-115.06	36.071	NV	2.0	false
5850 Polaris A La	is Vegas	18	Concentra Medical Centers	-115.18	36.083	NV	2.0	false
5495 S Rainbo La	is Vegas	6	Nevada Imaging Centers	-115.24	36.089	NV	2.5	false
515 W Buckey Ph	noenix	3	Labcorp	-112.08	33.437	AZ	3.0	false
5447 S Rainbo La	is Vegas	3	My Baby Bump Imaging	-115.24	36.089	NV	5.0	true
9700 N 91st St Sc	ottsdale	5	Sonora Quest Laboratories	-111.88	33.573	AZ	3.0	true
1760 E Pecos Gil	lbert	4	Marquis Diagnostic Imaging	-111.75	33.296	AZ	4.0	false
1300 N 12th St Ph	noenix	6	Sonora Quest Laboratories	-112.05	33.463	AZ	3.0	false
8551 W Lake La	is Vegas	10	Labcorp	-115.28	36.198	NV	3.0	false
3025 S Rainbo La	is Vegas	12	West Valley Imaging	-115.24	36.134	NV	4.0	false
6925 N Duran La	is Vegas	4	Steinberg Diagnostic Medical	-115.28	36.288	NV	2.0	false
3921 E Baseli Gil	lbert	4	Labcorp	-111.74	33.378	AZ	2.5	false
6400 S Easter La	is Vegas	5	Psychemedics Corporation	-115.11	36.073	NV	4.5	false
2865 Siena He He	enderson	3	LabCorp	-115.11	36.003	NV	3.5	false
740 Dorrell Ln No	orth Las V	8	3D Baby Kisses	-115.13	36.287	NV	5.0	false
3811 E Bell Rd Ph	noenix	3	Sonora Quest Laboratories	-111.99	33.639	AZ	5.0	false

Test case 3:

Main Categories: Food, Restaurants Sub Categories: Bakeries and Donuts

Attributes: Take-out_true , Takes Reservations_false

Search For Attribute: AND Review From: 2014-01-15 Review To: 2014-06-11 Review Stars > 0 Review Votes >= 0

Output:



Test case 4:

Member since: 2014-01-01 Review Count > 5 Number of Friends > 30 Average number of stars > 4.8

vote funny > 0vote useful > 0vote cool > 0

Search for attribute: AND

Output:

User Name	Date created	Review Count	Number of friends	average stars	total votes
Gregory	2014-02-01	7	72	5.0	14