## Projects 4 & 5

These two projects have been combined to form one project. The project is about **Yelp**. Yelp is a business rating and reviews & online food delivery company. It provides crowd sourced reviews about local businesses as well as online service reservation requests.

From Yelp's "About Us" section:

Our purpose: To connect people with great local businesses

10 Things You Should Know About Yelp

1. Yelp was founded in 2004 to help people find great local businesses like dentists, hair stylists and mechanics.
2. Yelp had a monthly average of 30 million unique visitors who visited Yelp via the Yelp app and 74 million unique visitors who visited Yelp via mobile web in Q3 2017.
3. Yelpers have written more than 142 million reviews by the end of Q3 2017
4. In addition to reviews, you can use Yelp to find events, lists and to talk with other Yelpers.
5. Every business owner (or manager) can setup a free account to post photos and message their customers.
6. Yelp makes money by selling ads to local businesses - you’ll see these clearly labeled "Yelp Ads" around the site.
7. Paying advertisers can never change or re-order their reviews.
8. Yelp uses automated software to recommend the most helpful and reliable reviews for the Yelp community among the millions we get. The software looks at dozens of different signals, including various measures of quality, reliability, and activity on Yelp. The process has nothing to do with whether a business advertises on Yelp or not. Learn more here.
9. You can access Yelp via iPhone, Android, and more - see the full list of mobile apps here.
10. The Local Yelp brings locals updates on the latest and greatest business openings & other happenings.

*Reference:* https://www.yelp.com/about

Yelp database

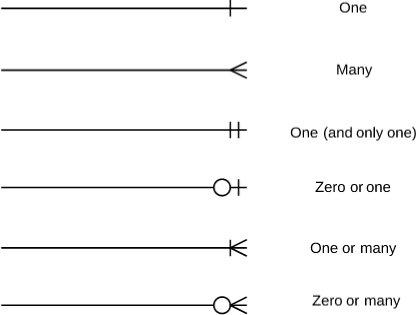
We will create a data model to keep track of information about Yelp users, local businesses on Yelp and reviews of those businesses written by Yelp users.

**Goals:**

* Draw the entity relationship diagram with the appropriate cardinality
* Add attributes to identify entities as needed (name all attributes and entities)
* Determine cardinality (1:N, N:N) and type of relationships

**Recall:**

General meanings:



Create entity relation diagrams for parts of Yelp's business, described below.

**PART A**

Businesses are reviewed on Yelp. Each business can have one or more reviews. Each review is associated with exactly one business.

**PART B**

Users login to Yelp to write reviews about local businesses. Each user in the database can write reviews sharing their experiences. Each review is associated with exactly one user.

**PART C**

One way to search businesses on Yelp is by category, which are keywords related to the business. For example, you could look for "burgers near me." This would look for businesses that have "burger" as one of their categories. Each business can have multiple categories associated with it, and each category can be associated with multiple businesses.

**PART D**

One way to filter businesses during a Yelp search is by attribute, which are features of the business. For example, you could look for restaurants that have "TRUE" for the "delivery" attribute. Each business can have multiple attributes associated with it, and each attribute can be associated with one or more businesses.

**PART E**

PARTS C and D have many-to-many relationships. How might we correct this? Hint: We might be interested in looking up all of the relevant attributes or categories for a given business. Conversely, if all of the businesses related to burgers close, we would want to make sure that "burger" is still a category that can be selected by a new business. Update the entity-relationship diagrams from PARTS C and D to remove the many-to-many relationships. Big Hint: This will involve adding an extra entity to each diagram.

**PART F**

Each user can receive different types of compliments from other users based on their business reviews, such as **cool**, **funny** and **hot**. Yelp does not care to keep track of who called a specific user **cool**, for example, however they want to store the number of times that each user received each type of compliment. Each user can receive compliments in each of the compliment categories. And each of the compliment categories can be related to each user.

**PART G**

Each user can provide feedback on reviews through voting. For example, a user can vote that a particular review is **useful** or **funny**. Yelp does not care to keep track of who called a specific review **useful**, for example, however they want to store the number of times that a user voted with each type of feedback. This is one way of understanding how engaged that user is with the platform. Each user can vote in each of the feedback categories, and each of the feedback categories can be utilized by each user.

**PART H**

PARTS F and G have many-to-many relationships. How might we correct this? Hint: We might be interested in looking up how many times a user has received each of the possible compliments Update the entity-relationship diagrams from PARTS F and G to remove the many-to-many relationships. Big Hint: This will involve adding an extra entity to each diagram.

**PART I**

Discuss the pros and cons of storing the user compliments and user feedback votes as attributes in the user table.

**PART J**

Put together an entity-relationship diagram for this system, using all of the entities discussed in PARTS A – H. Come up with candidates for primary and foreign keys for each entity. Be sure to include the cardinality of each relationship. Remember, which attribute or combination of attributes uniquely describe each entity?

**PART K**

Beyond the primary and foreign key attributes, what other attributes do you think would be stored in each of these tables? Modify the previous entity relationship diagram to include these attributes below.

**PART L**

Inspect the three .csv files provided, **yelp\_reviews.csv**, **yelp\_attributes.csv**, and **yelp\_categories.csv** and the data dictionary, **yelp\_dictionary.docx**. Using the column headers, fill in the any missing attributes in your entity-relationship diagram. Remove any attributes not found in the .csv files. Sketch a final entity relationship diagram with all attributes below. Be sure to include cardinality and keys.

*Dataset reference: https://www.yelp.com/dataset/challenge*

Yelp asks that you sign their terms of use to use these datasets. Please do so, here:

https://www.yelp.com/dataset/download

You do not need to download the dataset from the above link.

**PART M**

Use MYSQL queries to answer the following questions. For each question, write the SQL script along with the output obtained.

1. How many unique businesses are represented in the database? Find out how many of them are currently open.
2. How many unique users are represented in the user table?
3. Which user has received the highest average stars?
4. How many types of votes can a user send? List them.
5. How many types of compliments can a user receive? List them.
6. Return the reviews which have received the greatest value for stars, the greatest number of useful votes and the greatest number of cool votes. Make sure that you are returning all relevant reviews if there is a tie.
7. Which users have been "Yelping" the longest?
8. Which businesses have received the greatest value for stars? Sort these by the number of reviews, from highest to lowest. Return only open businesses.
9. Return the user who has written the greatest number reviews.
10. Return the number of open businesses in each Charlotte, NC neighborhood in the Yelp database. Sort from least number of businesses to greatest number of businesses.
11. Modify the previous query to return only neighborhoods with at least 100 businesses. Change the sorting to order the neighborhoods from greatest number of businesses to least number of businesses.
12. Return the neighborhood associated with the greatest number of reviews.
13. Return the average star rating for the businesses in each neighborhood. How could you modify this query to return the average star rating for the entire neighborhood? Hint: this will involve a weighted average.
14. Determine the average star rating given by the users who have written the greatest number of reviews. Return only the user's name, average stars and review count.
15. What are the top 10 features that most businesses are expected to offer? Sort them by the number of businesses, from highest to lowest.
16. Find out the categories associated with businesses that have been reviewed the most. Return only the business name, the associated category, the stars and the number of reviews received by the business.
17. The number of reviews and the stars rating received by a business are measures of its popularity. Identify the most popular businesses for the following categories? Sort these by the number of reviews and star ratings from highest to lowest.
18. Bars
19. Restaurants
20. Barbers
21. Apartments
22. Colleges & Universities
23. Car Rental
24. Any other business category that you are looking for 😉
25. Tagging businesses with categories can help narrow down search results. Taking hint from the above question find the most popular “Restaurants” that also have a “Bar”.
26. Using the given database investigate whether the following claims are true/false. Feel free to use multiple queries to answer this question.
27. Most people tend to share positive experiences only. Is there a bias for good reviews?
28. Most businesses on an average tend to receive more than an average star (>2.5) rating.
29. Businesses that are categorized as restaurants receive more reviews than businesses in other categories.
30. All major businesses except credit cards
31. All major businesses provide free Wi-Fi services.

**CHALLENGE (Bonus Question)**

1. A yelp user is free to rate businesses and write reviews. A user can also engage with other yelp users by voting their reviews as “funny”, “cool” or “useful”. An influential user is one who has received a lot of compliments from other yelp users.

You are tasked to investigate and recommend practices based using the “Charlotte\_Yelp” database by which users can become more influential in the network.

Hint: find metrics that could be correlated with compliments