RESTAURANT TREND ANALYSIS AND RECOMMENDATION ENGINE BASED ON YELP DATA HISTORY

Ashish Paralkar (amp3453@rit.edu) Arjun Dhuliya(amd5300@rit.edu)
Vignesh Nair (vn3246@rit.edu) Jaideep Murkute (jvm6526@rit.edu)

1. OVERVIEW

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. In the dataset you'll find information about businesses across 11 metropolitan areas in four countries. However, we are only concerned with restaurants in this case. The restaurant business is a booming industry. With the advancement of technology, people have started relying on social media and the internet for information and make decisions based on the information. Yelp is a ratings website which allows user's to give reviews and ratings over various business establishment.

Our intention in this project is to find out various trends of restaurants. Analyze the yelp data set pertaining to restaurants, and provide useful insight into the same. Our project caters to the needs of a customer as well as a potential restaurant business owner. For the customers, we have performed and given access to various functionalities which can give us the best possible choice according to a particular users preference. For the business owners, we have performed analysis which may give them an insight on the where,why and what type of restaurant to open up and to also maximize profits in their establishment. In this paper we discuss the following:

1. Design : Design of the database, the problems with the original database with respect to our objective.
2. Normalization: How the database was modified to suit to our needs.
3. Observations: We used R and Weka to perform various analysis on Restaurant trends using a sample dataset.
4. Map Plotting: We created a system which plots the nearest restaurants based on the location provided and the ratings along with them, so the user can decide accordingly.
5. Sentiment Analysis: We performed a basic level sentiment analysis on the reviews posted by the user.
6. Future work: Future improvements on the project is discussed here
7. Conclusion : Our concluding statements with regard to this project.

2. DESIGN

The project entails analysis of datasets, acquired from yelp under the domain of restaurants. While the dataset contains information of businesses of all kinds, but this project's main target is the data related to restaurants. This projects main objective is to analyze this data and see what the current trend is and provide valuable recommendations based on user history, geo-location, restaurant type and budget. The dataset chosen for this project has got information of all types of businesses. Currently since our focus is on restaurants, the other types of data was filtered out. Unnecessary or blank attributes or rows which are either redundant or not relevant was sifted through. The initial dataset had 11 tables.The following diagram represents the relations between the initially tables:

The dataset is available in the following
`www.kaggle.com/yelp-dataset/yelp-dataset/data`

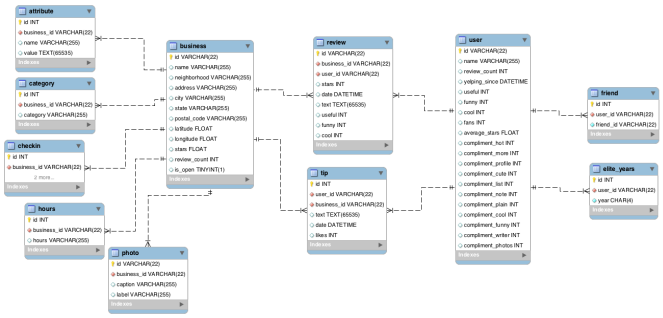

Figure 1: Flow diagram of the dataset

3. NORMALIZATION

The data we imported was in excel format. Initially it was not normalized.

1. Yelp_businesses: Not available 1NF. This is because same business ID has multiple business categories associated with it in a single attribute value, which violates 1NF property. We need to create a relation that will match each business ID with each category.Following functional dependencies were also found: (city) -> state, state, postal_code, latitude, longitude (For simplicity, just enlisting non-primary key dependencies).
2. yelp_business_attributes: Normalized in BCNF. Functional Dependencies: NA
3. yelp_business_hours: Normalized in BCNF Functional Dependencies: NA
4. yelp_checkedin: Normalized in BCNF Functional Dependencies: NA
5. yelp_user: Not available in 1NF. User ids have been maintained with list of all friends' ids as a single attribute value. We need to create a new relation that will map each user ID to another user's ID whom user is friends with. Functional Dependencies: NA
6. yelp_tip: Normalized in BCNF Functional Dependencies: NA
7. yelp_review: Normalized in BCNF Functional Dependencies: NA

To achieve this normalization, we initially ran through the setup of MySql software on our Windows 10 OS. We ran through python scripts for the execution of this normalization. Details are found in the file: scripts.pdf. On inspection we realized that certain tables need not have been separated or were largely unnecessary. Tables Attribute, check-in could simply be combined and merged into the table business. Photos table was not necessary in our objective of data analysis so it was scraped off. So was elite table. Since the tables were not even in 1NF (because the category attribute was multivalued)We had to separate out the multiple attributes. Finally we ended up with the following tables: List of Tables after normalization:

1. yelp_business:Primary Key: business_id
2. yelp_business_category:composite primary key: (business_id, category)
3. yelp_review:Primary key: review_id
4. yelp_tip:Composite primary key: (business_id, user_id)
5. yelp_user:Primary key: user_id
6. yelp_checkedin:Composite key: (user_id, business_id)
7. yelp_hours:Composite key: (_id, business_id)

Diagrammatically is represented as:

There is also another data available in sql format in which there was at least 1NF normalized content. On the laptops of other team members, we loaded the data via sql and normalized it in a similar fashion.

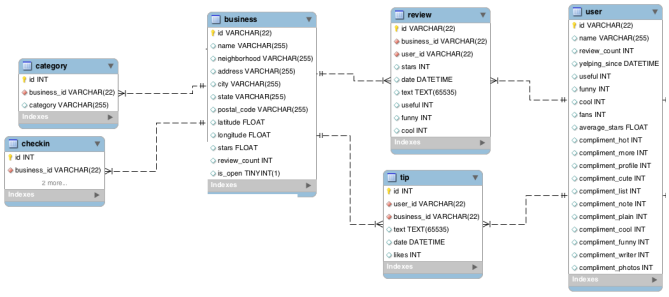

Figure 2: Flow diagram of the dataset after normalization

4. OBSERVATIONS

4.1 Trending of new users and reviews

We initially wanted to check out how many users are making use of Yelp dataset sample and what is the general trend behind it. The data available from January 2005 to May 2017 , spanning for over 11 years. We performed time series function using R and got the following results: From

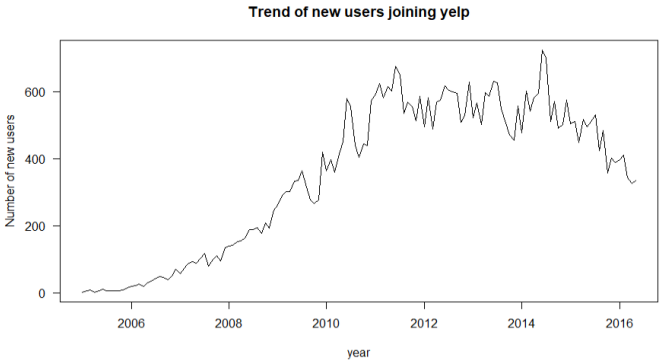

Figure 3: Trend of new users joining yelp

Figure 3 and 4 the data we can observe that in the end of 2014, there was the highest jump in the number of users joining yelp. Thus we can conclude yelp especially started gaining popularity around this time. People started relying on users comments and ratings to make a decision of whether to visit a business location or not. Similarly, we also notice that the number of reviews posted had a huge jump in the beginning of 2015. This makes sense since there was a jump in the number of users that signed was just before this time-line.

4.2 Top 25 cities with highest number of restaurants

From the sampled data, we wanted to figure out how populated a city is with restaurant. This data simply gives a user, say , to figure out how many options do they have in a particular city. Suppose a restaurant chain wants to

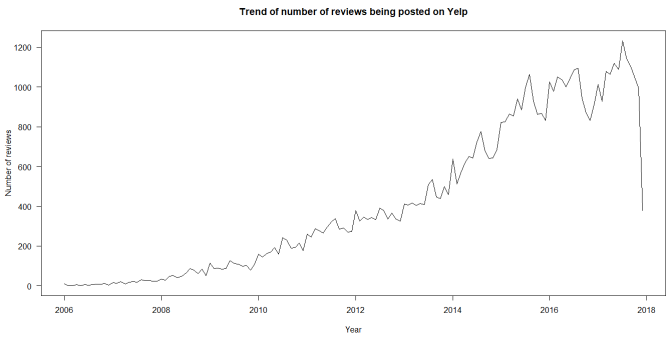

Figure 4: Trend of posted reviews on yelp

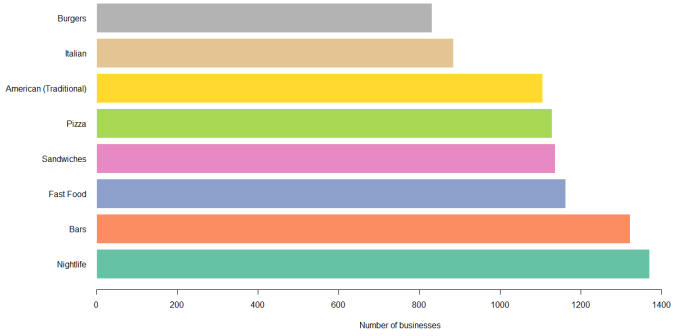

Figure 6: Top 8 type of restaurant businesses

open up another franchise in a new city, data like this can help them figure out what are their chances and help them make decisions based on their business strategy. From the

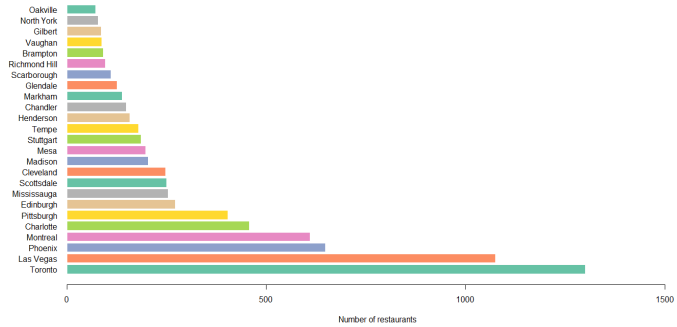

Figure 5: Top 25 cities

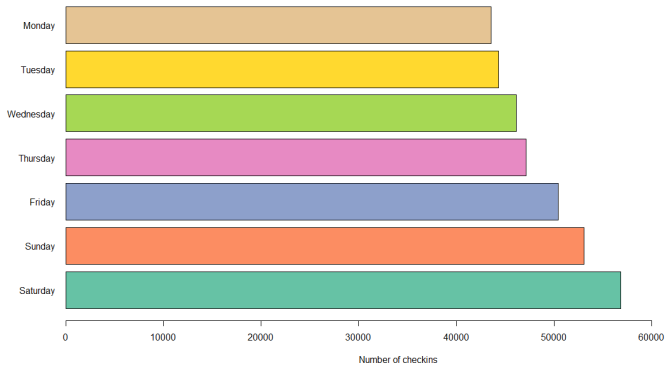

Figure 7: Busy days of restaurants

sampled data set, it appears that Toronto has the highest number of restaurants. This, like any other sampled data is prone to errors because not all of the data has been processed.

4.3 Top 8 restaurant categories

The next observation involves finding out the types of restaurant businesses. A real world application of this could be, to figure out whether or not you want your establishment to serve a particular type of cuisine or provide a particular type of service. For example, burgers seem to be the least popular category in this case and nightlife the most. Perhaps a safer bet for more profits would be if a nightlife establishment can be setup,

4.4 Busiest hours and days of the week

The next observation is finding out the general time when the restaurants are packed or busy. For this we used the checkin data of users and classified the data according to hour of the day and also day of the week.

From Figure 7 , this data we observe that after 10pm, restaurants usually get very busy.

From Figure 8, this data we observe that Saturdays are the busiest days. In the real world, this information can be mostly useful to a customer. For example, a user may avoid visiting a restaurant on saturdays after 10 pm because it will be too busy. From the point of view of business, if a restaurant realizes that a certain time of the day is usually empty, they could come up with business strategies like happy hours during that time so that profit occurs in that time slot. They could also jack up the price during busy hours of the restaurant.

5. MAP PLOTTING

This is the visual aspect of our project. The objective here is to give Latitude and Longitude, and based on these coordinates we find out the nearest restaurants, plot them using google maps api and also display the ratings of these restaurants so that the user can make an informed decision on the same. To accomplish this,we performed the following:

1. We have a basic html layout which uses the Google maps javacript API in which a user can enter latitude and longitude by clicking on the map.

2. The latitude and longitude is processed and fed as an SQL query to the server (In this case our yelp database).

3. This server takes these values and queries yelp database in the backend and checks for every restaurant if it lies in the 2.5 KM vicinity.

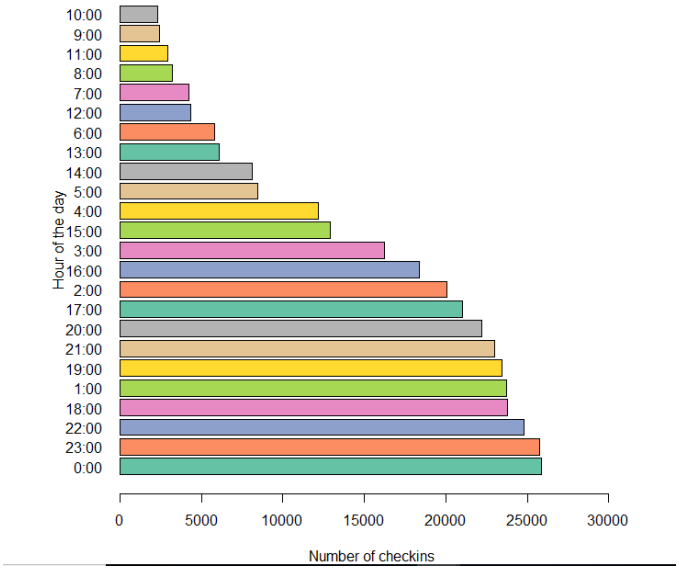

Figure 8: Busy hours of restaurants

4. To calculate the radial distance, we used the haversine formula.[1]

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

where φ *is latitude,* λ *is longitude,* R *is earth's radius (mean radius = 6,371km); note that angles need to be in radians to pass to trig functions!*

Figure 9: Formula [1]

The example we used here is for the city of Toronto in figure 10

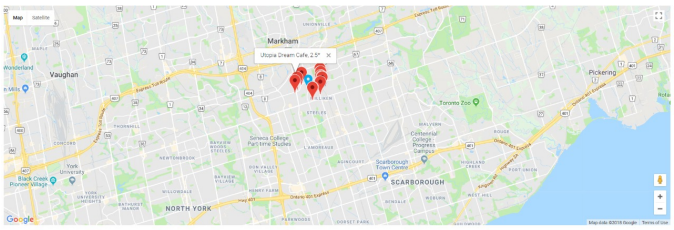

Figure 10: Plotting nearby restaurants in Toronto

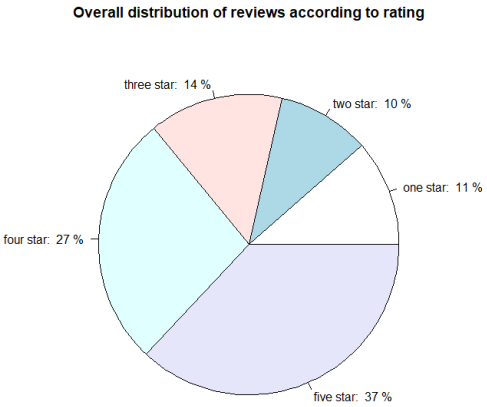

Figure 11: Ratings analysis

6. SENTIMENT ANALYSIS

First,let us discuss about the general trend in the ratings provided by the users. Generally people give 5 stars generously to restaurants. This can indicate that if satisfied by the performance of the restaurant, it is very likely that a user will give their positive review to the restaurant. We performed a basic level sentiment analysis on the users comments, tips and reviews of the restaurant. For this we took a sample of 1000 comments, and labeled it manually by reading each one of them. We then ran a program which learns this dataset and tested it on part of the same data for accuracy. This data was also co-related to ratings provided by the same user of the restaurant. The algorithm we used to classify the comments was Naive Bayes.

General steps performed were as follows:

1. Scan all data and preprocess it. This involved converting to lowercase, remove spaces,special characters,remove stop-words and remove stems.
2. Create list of all the words.
3. Find out frequency distribution and find thousand or desired value of words that will be most frequent and we will consider those relevant for our purpose.
4. Scan through all of the reviews again one by one and create "one-hot" representation of each review.
5. Train Naive Bayes classifier with given "one-hot" vector and label positive or negative that we have put manually.
6. Once trained on training data, we pass test data to naive bayes function. For test data, classifier predicts output and then checks if it was right or wrong.

We came up with the following results:

1. Most common 10 words in reviews with 5-star rating: [('food', 14631), ('great', 13068), ('place', 12840), ('good', 10378), ('service', 7735), ('best', 6250), ('time', 5879), ('one', 5799), ('delicious', 5750), ('back', 5686)]
2. Most common 10 words in reviews with 3-star rating: [('good', 7951), ('food', 7676), ('place', 5266), ('like', 4605), ('would', 3980), ('service', 3679), ('really', 3320), ('one', 3188), ('get', 3084), ('time', 3064)]
3. Most common 10 words in reviews with 1-star rating: [('food', 6316), ('place', 3749), ('service', 3372), ('us', 3210), ('would', 3029), ('like', 2987), ('back', 2848), ('one', 2812), ('order', 2735), ('get', 2727)]

Since we did not have a large sample of data and the fact that the sentiment analysis was very basic level, we can come up with certain conclusions based on certain assumptions. We have not performed word association, ideally in this case word association, confidence and support frequent dataset analysis would be required to get more accurate data, but for now we are drawing conclusions from the basic level analysis we have performed.
In case of 5 star ratings, we can conclude that word choices like "best" "back" and "service" are generally in the same realm of the user comment, it is likely to be an extremely positive review.
In case of 3 star rating, we can conclude that word choices like "good" "would" "place" are in user comment, it can be an average review. Neither positive nor negative basically.
In case of 1 star rating, we can conclude that word choice like "back" "would" "us" are in the comments of the user, it is likely to be a poor review.
The accuracy of the program was 66% when tested on sample dataset.

7. FUTURE WORK

Since this project was started off on a base level, there is a lot that can be done to improve the efficiency, accuracy, usage and scope of the project.

Efficiency:.

One of the biggest problems that we faced for handling this project is the size of the project and the efficiency behind it. While formulating our queries we did try as much as possible to save time and efforts, we were largely limited by the hardware capabilities and the software capabilities to process the amount of data and its complexity. To avoid the slow-ness we did use a sample dataset which was randomly extracted but we believe that also caused certain accuracy problems because we cannot ascertain how much a particular sample affected the results of our analysis. So in order to improve the efficiency, we can possibly use better servers and hardware as well as come up with better queries for the same.

Accuracy:.

There was a fluctuation in accuracy, especially in the sentiment analysis. We believe the analysis can largely be improved by adding more factors and making a more comprehensive algorithm run through it. We can make use of Boolean Multinomial Naive Bayes algorithm instead of the original Naive Bayes.

Usage:.

Since our projects main focus was finding out trend analysis of a restaurant, having a UI was not really a major objective of our project.While map plotting provided visualization of data, alot more can be done to make it user friendly. We could simply use the queries generated by us to get a well furbished android/apple/Web UI so users can portably scan and use the yelp dataset.

8. CONCLUSION

Due to the advancement of technology, services like Yelp prove to be a huge factor in not only identifying, but also adding to, the influence of Trends of Restaurants. We realize that by figuring out various attributes of the restaurant, we can help future potential business owners to set up or improve their businesses. We also provided an insight on how users can use the yelp dataset, to get them the best possible restaurant they want as per their location.Sentiment analysis is a very useful tool for the machine to give even more information about the vibe and service of a restaurant. We can conclude successfully that even Novice level Data Science tools can give us in depth insight about a database and has a lot of undiscovered potential to prove useful to the society.

9. REFERENCES

[1] Calculate distance, bearing and more between latitude/longitude points. https://www.movable-type.co.uk/scripts/latlong.html.