Yelp Data Analysis Project Report

Junhua Liang*, Virginia Wu*, Xin Lu*, Wenhao Li*

* Courant Institute of Mathematical Sciences, New York University

Abstract- Feast is important for everyone in the world, especially in this era as the development of the society's productivity. People who travel to a new city usually want to have a taste of the best restaurant in this area, and those who want to repair their house's roof would also check the rating of the store. Merchants also want to open their restaurants in the best location in the city. Thus, it's important to have related analytics to help them. One direction for analytics is using people's reviews and their check-in information to get the pattern of consumption in the area and build related models to help recommend restaurants for people and find locations for merchants. In this project, we will use Yelp's open datasets to visualize specific information and build models for recommendations and rating predictions.

***Keywords*- Spark, machine learning, data visualization, big data.**

I. INTRODUCTION

This report details the results of our team's final project for CSCI-GA 2437 Big Data Application Development at New York University. Our team chose Yelp Data Analysis as our topic, we would perform an analysis on Yelp's open-source datasets. Yelp is an American company that develops the Yelp.com website and the Yelp mobile app, which publish crowd-sourced reviews about businesses. Many people would use the app for recommendations when they need to find a restaurant or other service. Some of them will also give feedback to the app after they experience the service of that business. Hence, it collects a large amount of data, and there would be huge amounts of information behind them. Our team would transfer these raw data into graphs (geography graph, bar graph, line graph, etc.) that people could understand using Python and Excel, and retrieve some patterns from them. These patterns could help improve the quality of the app's recommendation systems and help merchants make better decisions when they are going to open a new restaurant in some city. In addition, many people who use Yelp just want to find the best businesses (restaurants) when they want to eat with their friends, and thus the recommendation system is really important for this app. Our team will build a simple recommendation system using one matrix factorization method called Alternating Least Square (ALS) [1], and will also test the model compared to the past records of the people. Plus, we will also explore some interesting patterns from the data, such as the friend relationship behind the data. We will try to visualize the friend relationship using a Force Atlas 2 Graph, and mining useful information from it. Most of our tasks will make use of Spark[2] for big data processing and analytics

learning from this course (DataFrame, SparkML[3], Optimization, etc.), we will experience the power of big data while mining the big dataset of Yelp. Most of our applications will run on NYU Peel Cluster.

The report consists of mainly five major sections. The number of pages may vary depending on the content of the report. These are:

1) Abstract
2) Introduction
3) Data Sources
4) Project Design
5) Results and Analysis
6) Summary

II. DATA SOURCES

This section introduces the data sources our team used in this project.

Our team project used Yelp[1] open dataset for analysis. This dataset is a subset of Yelp's businesses, reviews, and users dataset for education and academic purposes, and it contains 6990280 reviews, 150346 businesses, 200,100 pictures, and 11 metropolitan areas. In this project, we mainly used the reviews and businesses part of the dataset for data visualization, building recommendation systems, and analyzing the friend relationship behind the information. The dataset contains five parts: user, business, review, tip, and check-in. Table 1 shows the size of each of them.

Table 1. Size and description of Yelp's dataset.

Dataset	Description	Size
yelp_academic_dataset_user.json	Users' information, including their id and metadata.	4.34 GB
yelp_academic_dataset_business.json	Businesses' information, including their id and metadata.	118.9 MB
yelp_academic_dataset_review.json	Review and stars for businesses by users.	5.34 GB
yelp_academic_dataset_checkin.json	Businesses' id and their check-in date.	287 MB
yelp_academic_dataset_tip.json	Tip's information on specific users and businesses.	180.6 MB

The data format of these files is all JSON, and we will introduce how we would process them in the next section.

III. PROJECT DESIGN

This section introduces the design architecture and challenges in our project.

A. Design Diagram

Figure 1 shows the design diagram of our project.

[1] Source: https://www.yelp.com/dataset

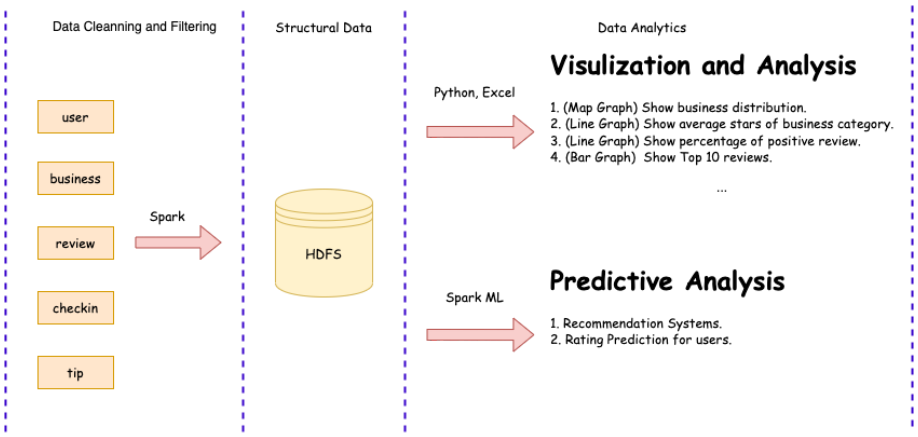

Figure 1: Design diagram of Yelp Data Analysis.

In the first step of our project, we assigned the five datasets to each of our team members to do data cleaning and filtering. In this project, we used Spark, the core skills taught in this course, to process the dataset. We filtered useless columns and transformed some columns from the raw format to one convenient for us to analyze. For example, we turned the review column in the review dataset from text to an array of numbers using word embedding (word2vec). Since Spark has a rich API for data processing and machine learning, we could turn our dataset into a good format easily. Figure 2 shows the data ingestion procedure for the review dataset.

```
+--------------------+----+-------------------+-----+--------------------+-----+--------------------+------+--------------------+
|         business_id|cool|               date|funny|           review_id|stars|                text|useful|             user_id|
+--------------------+----+-------------------+-----+--------------------+-----+--------------------+------+--------------------+
|nphMADgWEWqQ_CM_-...|   0|2020-03-13 21:14:56|    0|BFP8USpSMVbewMBwB...|  1.0|Needs to lear how...|     1|r80c-gIg2e4cCJyIH...|
|nC46s-L4KIcYQ_HmQ...|   0|2020-08-15 00:46:30|    0|74zHjNSFdRXCp3LgF...|  4.0|We visited last w...|     1|33qdjvUesOUv4Q7DR...|
|HPqTJ_yF2ZJw87yHW...|   0|2017-06-06 23:40:10|    0|DX-3tr_qusLpw2_kL...|  2.0|We went to this p...|     0|Q0MehSQNivqQRc3ZT...|
|D1WqtPVxtjMQxH9Fs...|   5|2020-03-07 17:32:37|    5|HVXeXdDwdLCcxquzM...|  5.0|We love this plac...|     5|Tv_HJNg8Weyk_8V9_...|
|8SOgWpYKJgiEfTuyX...|   0|2020-12-22 00:20:43|    0|GrQpD7COMh6yKKxH5...|  4.0|First of all I wa...|     0|FnUs-mADLCFRTWyGo...|
+--------------------+----+-------------------+-----+--------------------+-----+--------------------+------+--------------------+
```

Data Ingestion

```
+--------------------+--------------------+--------------------+--------------------+-----+-------------------+
|             user_id|         business_id|           review_id|            word2vec|stars|               date|
+--------------------+--------------------+--------------------+--------------------+-----+-------------------+
|clWLI5OZP2ad25ugM...|x4XdNhp0Xn8lOivzc...|G_5UczbCBJriUAbxz...|[-0.0047346777282...|  5.0|2013-08-15 15:27:51|
|-sryo4gDYxbZ1T5Bz...|ltBBYdNzkeKdCNPDA...|elqRpX9T3YwL07uLN...|[-0.0010586014164...|  2.0|2015-02-02 04:29:13|
|3n_QYiWlZdt5C9ogb...|wP9Cx-jxLflhv6c1m...|0kMlS1dVfIw1bH1jK...|[0.00261225699830...|  5.0|2011-08-25 15:42:07|
|KepicN2NnZ5aPXVlB...|2myVMP_l-9f9JIZew...|jHQiGVYdbT-9lQJmS...|[0.00389007436111...|  3.0|2011-10-28 02:09:13|
|i8kKtg9HlYiHps1Or...|Vnob_w_Aohf7ZDqxc...|bUR41rh0qluvTTgdg...|[0.00293917025306...|  5.0|2016-01-30 01:14:25|
+--------------------+--------------------+--------------------+--------------------+-----+-------------------+
```

Figure 2. Data ingestion for review dataset.
Filtered the funny column and transformed the text column to word2vec.

After data ingestion for each dataset, we saved these data as schema into the HDFS[4] (provided by NYU Peel Cluster) for later processing. After that, we retrieved data patterns from these data by data visualization using Python and Excel. We analyzed them using various kinds of figures. For example, we used location information in the

business dataset to construct a geography graph, and thus got the location distribution of these business stores. In addition, we also built a recommendation system based on the review dataset. We used the Alternating Least Square (ALS) algorithm, a matrix factorization recommendation system algorithm, to build the model, and recommended new items for users. Finally, we also did one interesting task, exploring the friend relationship using these data and found potential valuable information for use.

B. Code Challenges

The project relied heavily on the NYU Peel Cluster as we used Spark and HDFS to process and store the dataset. Hence, we met lots of code challenges while finishing this project. First, in order to use the code Highline in IDE and connect to the remote service, we did not code as fluently as the local computer. Second, since the NYU Peel Cluster did not support Tableau for data visualization, we should export analytics results in CSV format and then use Python and Excel to visualize the results. Third, the NYU Peel Cluster had limited memory volume for students, and thus we could not train a good model with a high accuracy rate in limited iterations, because training too long would cause a stack overflow.

C. Obstacles

Except for the code challenges mentioned in the previous section, our team still met many obstacles. The datasets are really large compared to other teams, especially the review and user datasets. Thus, when running our program on the clusters, it may take a long time to finish each step. Besides, it's not convenient to use the Zeppelin, because it's too slow to show the results. Second, the training set for the recommendation system is too large to train for more iterations in Spark. It's easy to cause stack overflow if the number of iterations is greater than 25, this situation greatly affects the accuracy of the model. We think the model should have better performance if it had enough training iterations. Third, for the friend relationship analysis, each user has a list of friends' IDs. However, some of those ids are not included in our primary key. In other words, we don't have data for that user. However, while rendering the Force Atlas Graph, these data points are necessary. To solve this problem, we could just create a dummy point for that user and assign it with 0 weight.

IV. RESULTS AND ANALYSIS

A. Experimental Setup

Most of our applications are run on NYU Peel Cluster. The Spark SQL version for the service is 2.4.0. The Scala version is 2.11.12. Most codes are run through Zeppelin. The Hadoop version is 3.0.0. Data visualization is run on local machines using Microsoft Excel for Mac Version 16.60 and Python 3.9 using folium, Wordcloud, matplotlib, and pandas.

B. Data Visualization

1. Data Overview

We visualize the latitude and longitude data for each business and the number of locations in each area on a US map using Python Folium. Users can zoom or move the cursor to find the restaurant they want. Looking at the map, most of the businesses are located around the East Coast and West Coast.

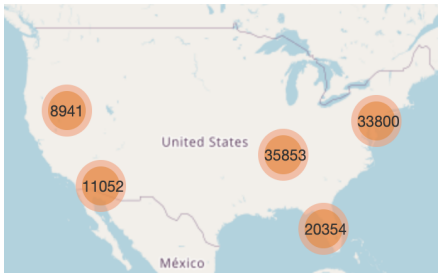

Figure 3. Data on the US map

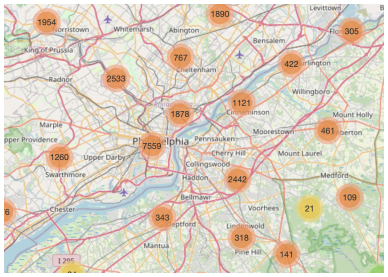

Figure 4. Data in a specific area

Figure 5.

We rank the Top 10 cities with the most businesses in the Yelp dataset, and Philadelphia ranks first.

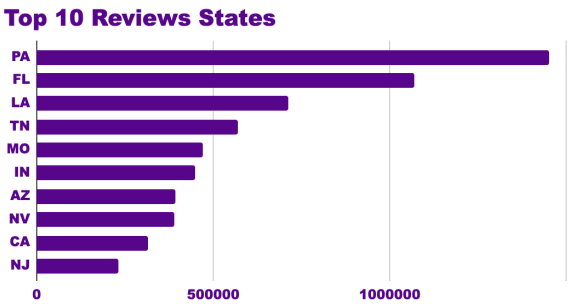

Figure 6.

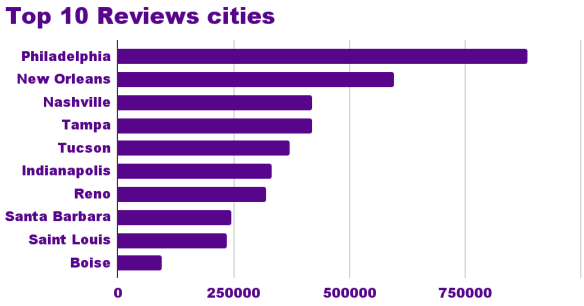

Figure 7.

Figure 6 and 7 shows the top 10 states and cities with the most reviews. We can see the dataset mainly includes business data for 10 US states and 10 metropolitan areas (It can also be seen in figure 8). Among them, Pennsylvania is the state, and Philadelphia is the city with the most reviews data.

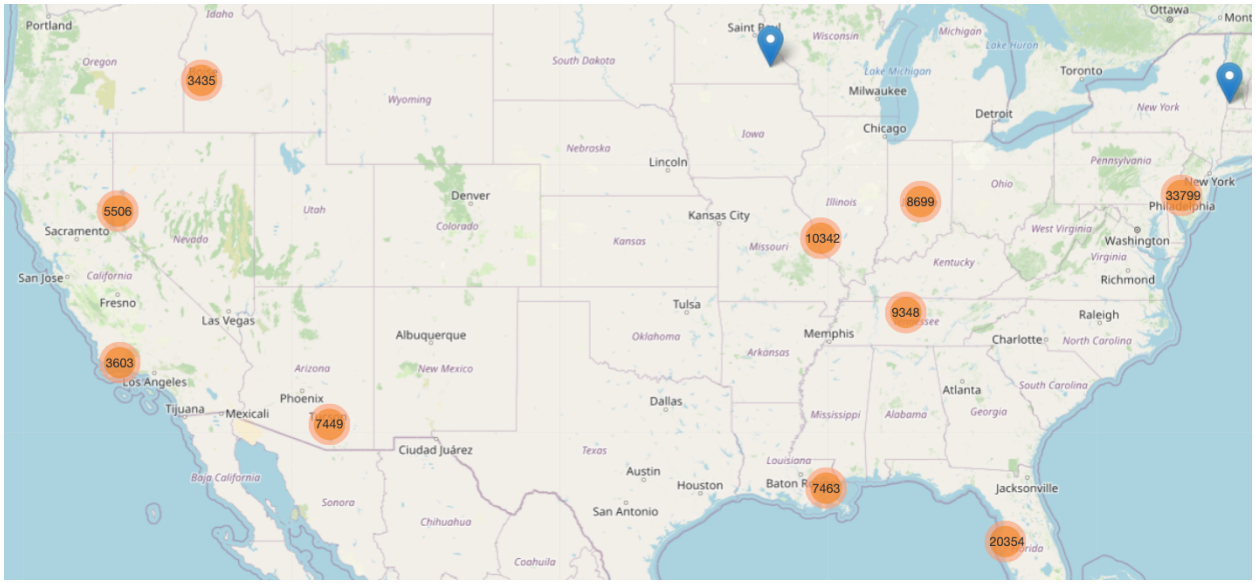

Figure 8. Data points across the US

We are also interested in how the stars in reviews are distributed in terms of months to see if seasonal factors have an impact on ratings.

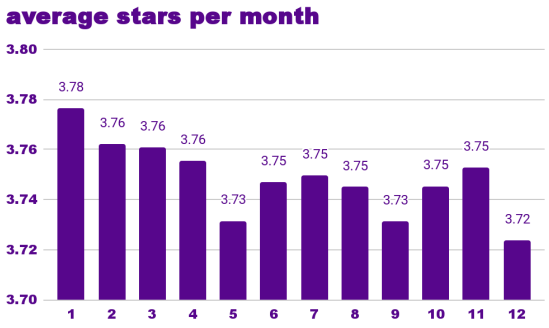

Figure 9.

Figure 9 shows the trend that the average stars for all businesses reach the highest in January and lowest in December.

Figure 10 shows the Average Stars in the Business Category. People score higher on religious organizations. The worst rating category is Real Estate.

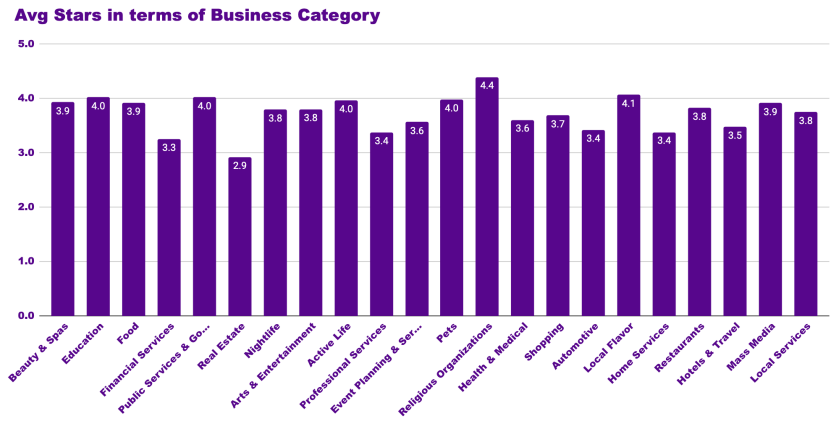

Figure 10.

The following figure 11 shows the names of businesses with the most Five Star Reviews. Reading Terminal Market and Oceana Grill are the two most popular businesses from the Yelp reviews with five-star ratings.

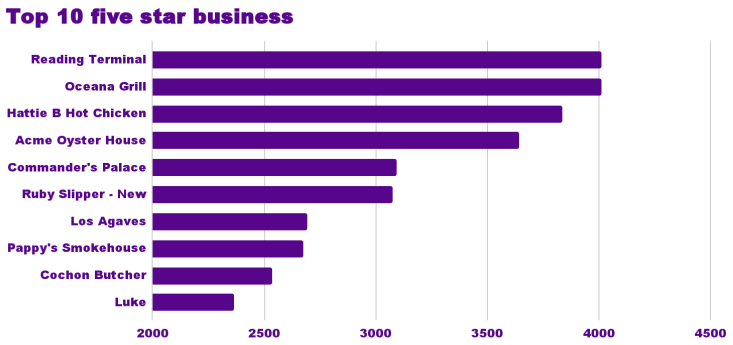

Figure 11.

Yelp is a collection of different businesses in different areas. As figure 12 shows, in the Yelp dataset, the most popular business category is “Restaurants”. The dataset has a huge collection of restaurants. So we take a closer look into this category and get more interesting results.

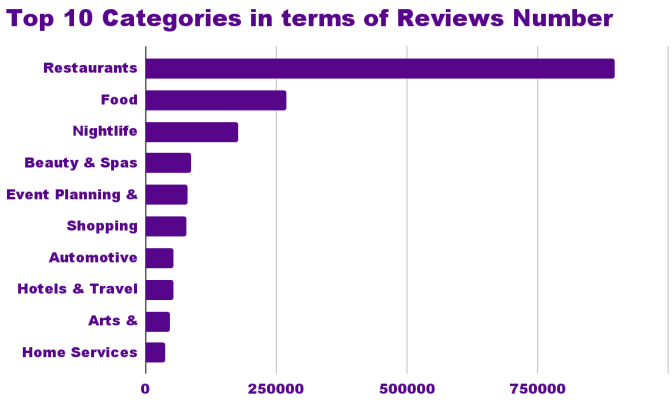

Figure 12.

The number of reviews can reflect the popularity of businesses to some extent. Figure 13 shows the names of restaurants with the most reviews and figure 14 shows the names of restaurants with the most 5-star reviews. “Oceana Grill” is the most reviewed restaurant. As we can see, the restaurants with the most reviews are almost the same as the restaurants with the most 5 stars. We may safely conclude that there is a positive correlation between the number of restaurant reviews and the number of positive restaurant reviews.

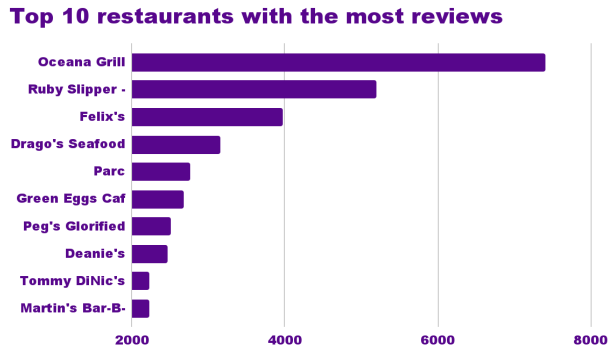

Figure 13.

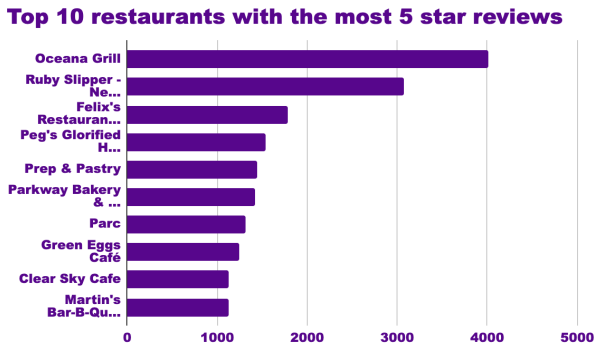

Figure 14.

According to the rating stars of the review, we classify the reviews into positive(stars>=4), negative(stars<3), or neutral(3=<stars<4) and we evaluate the positive review percentage in figure 15. As we can see, the Religious Organizations category has the most positive reviews and the Real Estate category has the least positive reviews, this result is the same as the average stars of the business category, which means that our analysis in the review and classify method is quite reasonable.

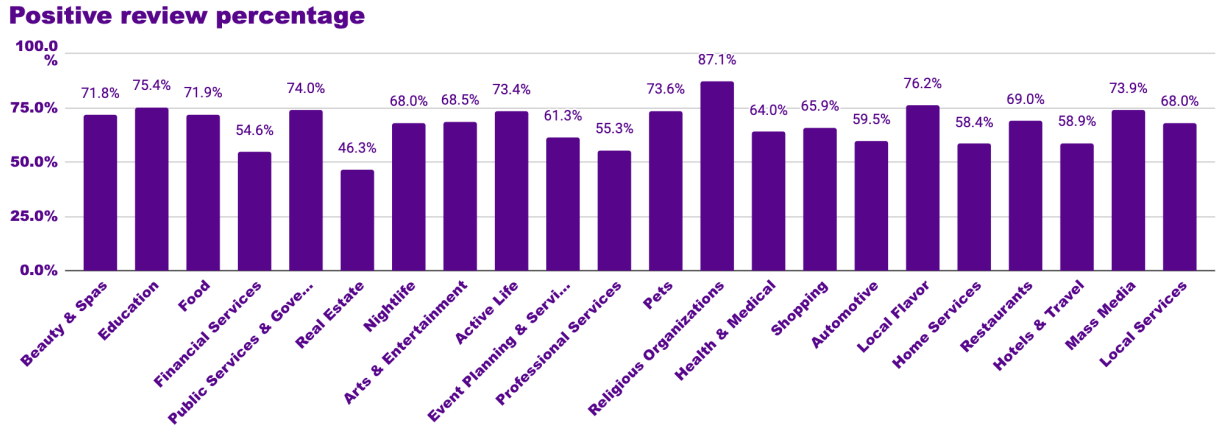

Figure 15.

2. Restaurant Analyze

2.1. Check-in Time

The following figure 16 shows the check-in time distribution throughout the day. The time of check-in is mainly distributed from 5 to 7 p.m. and from 11 p.m. to 1 a.m. at midnight. But abnormally, the check-in numbers during lunchtime are not as much as we expected. On the contrary, it is relatively small.

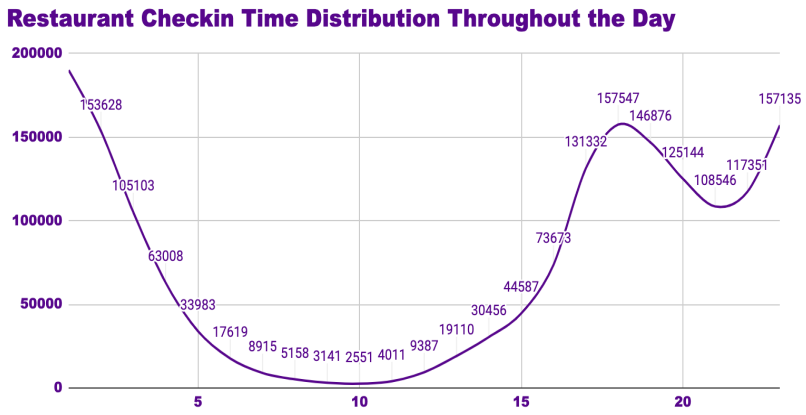

Figure 16.

So, we further investigated this curious result. We looked at check-in data for each day to see if there was a correlation between the time people went to the restaurant and the day of the week.

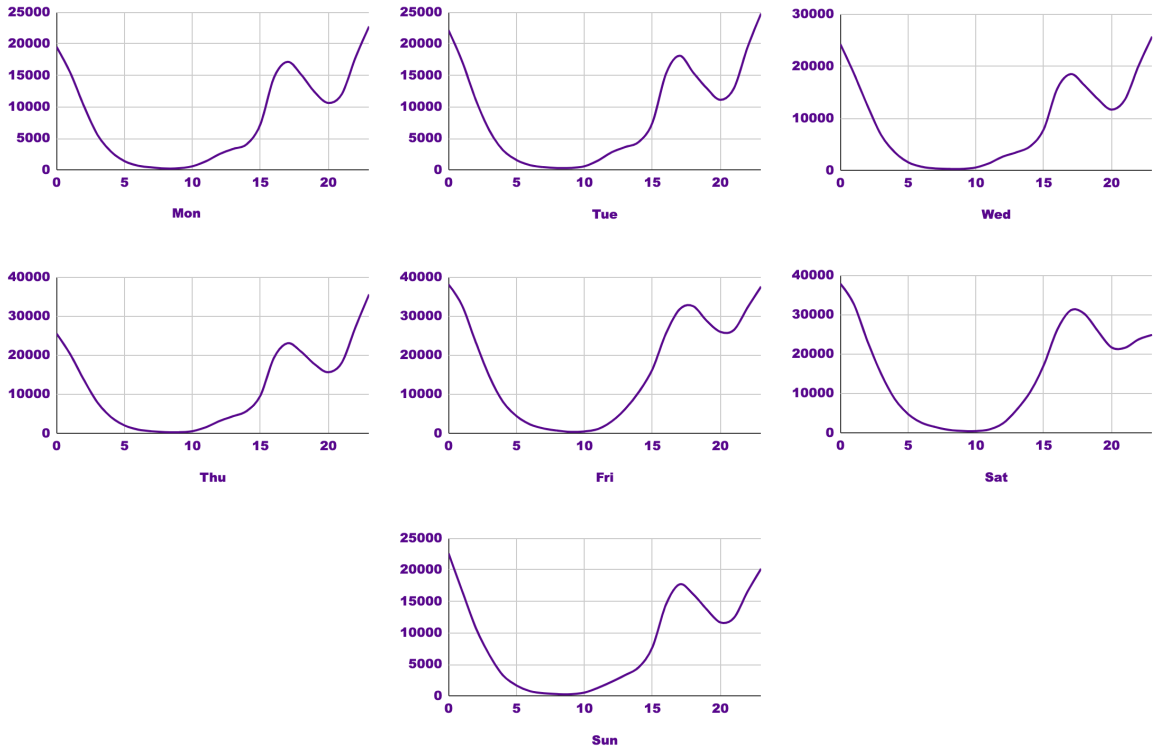

Figure 17. The distribution of check-in time in each day

From the data of each day, there is no significant difference in the check-in time. So this may be a refreshing conclusion that we can draw from this dataset.

2.2. Cuisines

From the business dataset, we analyze the top 10 cuisines, which shows in figure 18. It shows that whether traditional or new, people who use the Yelp app like American food.

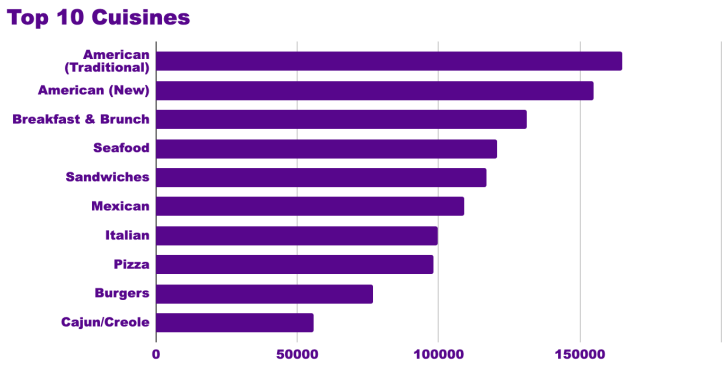

Figure 18.

2.3. Review Length

Figure 19 compares the length of positive and negative reviews. It shows that people were more likely to use more words to express dissatisfaction when they scored low, while they were more stingy with their evaluations when they scored high.

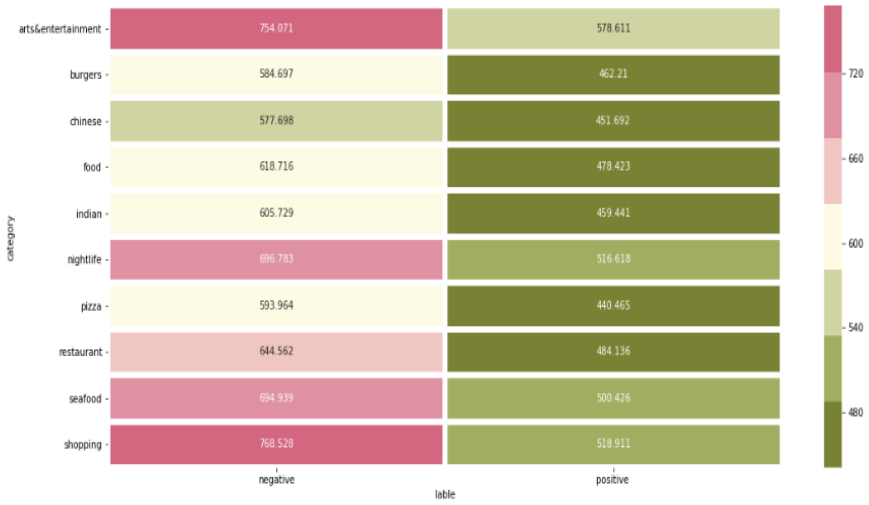

Figure 19.

2.4. Word Cloud

A tag cloud[5] (also known as a word cloud) is a visual representation of text data, which is often used to depict keyword metadata on websites or to visualize free-form text. Tags are usually single words, and the importance of each tag is shown with font size or color. In our analysis, we use the tag cloud to display the words most used by users in positive and negative reviews, which is more visually. To more intuitively show positive and negative evaluation, We filtered out adverbs, prepositions, and some neutral words.

Figures 20 and 21 show the words people use the most when they review restaurants. “Delicious” has undoubtedly become the most commonly used and most concerning word. Simultaneously, words like “Friendly”, “Fresh”, and “Worth” also occupy a certain concern. While “Bland”, “Rude”, and “Cold” became the most negatively evaluated word.

This method can also better help us analyze people’s evaluation of the restaurant and make improvements. Here we use a Chinese restaurant as an example.

Figure 20. Keywords for Restaurants Positive Reviews

Figure 21. Keywords for Restaurants Positive Reviews

Figure 22. Keywords for Chinese Food Positive Reviews

Figure 23. Keywords for Chinese Food Positive Reviews

Chinese restaurants have positive reviews mainly for their friendly service and fresh food. People who like hot food may also enjoy eating in a Chinese restaurant. From the negative word cloud, we could observe that fried is one of the main problems for Chinese cuisine types, which means customers expect food should be less oily.

Our review analysis extracts certain features from the review dataset. Restaurant owners can use a certain number of Yelp reviews to get basic information about why customers like or dislike restaurants, whether they are getting good reviews for fresh food or poor reviews since the prices are too high. At the same time, they can compare their restaurant to a similar restaurant.

C. Recommendation System

In this project, our team applied the Alternating Least Square (ALS) model in Spark to build a simple recommendation system. ALS is a simple matrix factorization method used for recommendation systems just using the information of user_id and business_id as well as the rating to predict the hidden rating for a recommendation. After finishing training, our team recommended 5 items for 4 randomly selected people. Table 3 shows the result.

Table 3. The top 2 items are recommended by the system of 4 randomly selected people.

name	p1_name	p1_categories	p2_name	p2_categories
Benny	It's Sugar	Candy Stores, **Specialty Food**	El Pastorcito	**Mexican**
Varusha	Not Just Crab	Seafood, **Nightlife**	Zitner's Candy	Candy Stores
Crosby	June's Restaurant	Ice cream & Frozen Yogurt	Pizza Hut	Italian
Daniel	City Grocers	Food, Grocery	Amore Pasta	Italian

To verify the recommendation system, we compared recommendation results with the users' records and verified whether they are relative. The 4 people's records are in Table 4.

Table 4. Records of the 4 randomly selected people.

name	p1_name	p1_categories	p2_name	p2_categories
Benny	The Seed	Juice Bars, **Specialty Food**	The Cheesecake Factory - Reno	**Mexican**
Varusha	Roller Kingdom	**Active Life**, Skating Rinks		
Crosby	Polish Cottage	Polish, Restaurants	Pasco Kitchen & Lounge	American
Daniel	Pokey's	Pubs, Nightlife, Bars		

From the result, we observed that the recommendation system did recommend some appropriate items for Benny and Varusha, but not Crosby and Daniel. This is because the model could not train for enough iterations due to machine limitations in NYU Peel Cluster. However, we could see the importance of the recommendation systems of Yelp. People could go to places that are suitable for them with better recommendations.

D. Friends Relationship

We analyzed yelp_academic_dataset_user.json and found the most popular users who have the most followers. 57624 users' data are used to render a Force Atlas 2 Graph[6]. Each vertice represents a user and an edge represents a friend relationship. We listed the top 10 users as influencers who have the most followers.

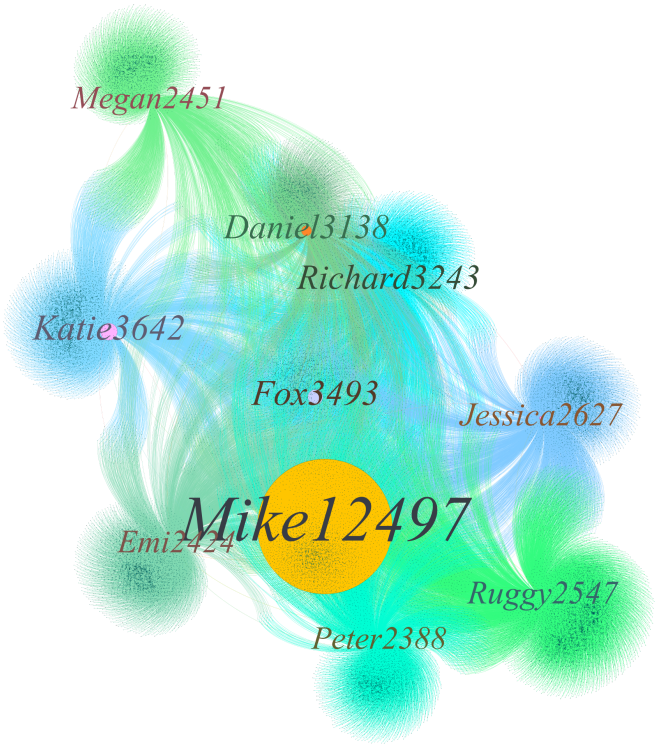

Figure 24. Force Atlas 2 Graph

For instance, Mike has 12497 followers. As we can see the vertice for Mike locates at the center, which means that most of his followers have also followed other influencers

V. SUMMARY

Yelp's dataset could explore many interesting patterns. First, the dataset is large enough for us to apply Spark to processing and analyzing, as well as exploring many interesting patterns from it. Second, the dataset contains much information in its raw data. We explore lots of valuable information through visualization. Information such as geography distribution in different areas and the top 10 cities with the most business could help merchants who want to open new stores in these areas make better decisions. Positive reviews and most used words could help users recognize whether they should go to the recommendation stores. Third, building a good recommendation system is important for Yelp's development. In our project, we could not obtain a good model with high accuracy due to machine limitations and data incompletion, and as you can see, it would cause disastrous results for the users. Hence, it's important to improve the recommendation systems with better datasets and better training machines. Finally, many interesting patterns could be explored from the dataset, and we explored the friend relationship from it. This information could be used to improve the system's quality and correctness. In general, this data analytics could help both the users and merchants a lot.

ACKNOWLEDGMENT

Our team's experimental setup was supported by NYU Peel Cluster, thanks to them for providing the machines for us to run the application. This course was organized and instructed by Professor Tang Yang, thanks for his contributions to this course and the instructions to the students. Yelp provides open-source datasets for us to utilize and analyze, thanks to the company. Thanks for everyone's contribution to the team project.

REFERENCES

[1] Zhou, Y., Wilkinson, D., Schreiber, R., Pan, R. (2008). Large-Scale Parallel Collaborative Filtering for the Netflix Prize. In: Fleischer, R., Xu, J. (eds) Algorithmic Aspects in Information and Management. AAIM 2008. Lecture Notes in Computer Science, vol 5034. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-68880-8_32

[2] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, Ion Stoica. Apache Spark: a unified engine for big data processing. Commun. ACM 59(11): 56-65 (2016)

[3] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Talwalkar, A. (2016). Mllib: Machine learning in apache spark. The Journal of Machine Learning Research, 17(1), 1235-1241.

[4] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1-10). IEEE.

[5] M. Halvey and M. T. Keane, An Assessment of Tag Presentation Techniques Archived 2017-05-14 at the Wayback Machine, poster presentation at WWW 2007, 2007.

[6] Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PloS one, 9(6), e98679.