Extracting Most Popular Dishes From Restaurants Based on User Reviews

Prachi Bhopatkar
Dept. of Computer Science,
NYU Polytechnic School of Engineering
New York City, United States
Email: psb300@nyu.edu

Mansi Patel
Dept. of Computer Science,
NYU Polytechnic School of Engineering
New York City, United States
Email: mp3950@nyu.edu

Karan Kapoor
Dept. of Computer Science,
NYU Polytechnic School of Engineering
New York City, United States
Email: karan.kapoor@nyu.edu

***Abstract*—This paper summarizes the analytics done on user reviews to understand "must have" dishes in a particular restaurant. An anlytics was performed on User data, Restaurant data, User reviews data and Restaurant menus data that we obtained from Yelp and Foursuare. We are using some Sentiment Analysis tools to process user reviews and calculate average scores for each of the dishes. We are suggesting "must have" dishes from a particular restaurant depending on these scores.**

***Keywords*- Analytics, Sentiment Analysis, User Reviews, Restaurant Menu, User Data, Restaurant Data, Hive, Map Reduce, Must-have Dishes**

I. Introduction

We arbitrate to choose a restaurant by its overall ratings. But sometimes we are indifferent about the choice of food in the corpus of dishes inside the menu.

The idea of "Munchies" came to our mind when we were travelling. We were searching for a good restaurant for dinner. As always, we searched on Yelp to check nearby restaurant ratings and selected one with really good rating. After going to the restaurant we weren't sure about what dish to order. There wasn't enough time to go through all Yelp reviews to figure out "must-have" dish from the restaurant. That time we thought what if Yelp provides us a number of dishes which matches with our taste and at the same time are speciality of that restaurant? That time we decided to develop "Munchies" app.

In this analytics we are working on different datasets:
1. User data which provides information like user name, user's reputation, user rating, etc. [Data set size: 200MB]
2. Restaurant specific data like Restaurant name, hours of operation, address, overall rating, etc. [Data set size: 60MB]
3. User reviews data which includes users comments about different dishes, service at the restaurant and their overall experience at the restaurant.[Data set size: 2GB]
4. We have collected menus from different restaurants using APIs provided by Google and Foursquare[Data set size: 4MB].

Furthermore, we have analysed the user data in conjunction with reviews data to visualize some interesting facts. Such as the most active users and the distribution of number of friends amongst the users

II. Motivation

In today's tech savvy world social media, applications like yelp, trip advisor, etc. are playing major role in custome's decision process.People tend to check reviews about movies, restaurants, shopping malls, etc. before trying these things out.

Restaurant ratings available on public platforms are generalized and based on various things like dcor, hygiene,quality of service, food, etc. These ratings does not reflect what is a speciality of a particular restaurant.For example, If I am looking for a good restaurant in downtown New York City, application like yelp will suggest restaurants with good customer ratings, but it won't suggest particular dishes which I might like.

We have Munchies app in the Google play store which can be downloaded by any user free of cost and can be easily used. In the Munchies application we are processing Yelp restaurant ratings as well as user reviews about the restaurant's overall service, quality of food and "must-have" dishes. We are suggesting highly recommended dishes from each restaurant.

We believe our application will certainly help users in selecting best restaurants and dishes which can please their taste buds.

III. Related Work

We read multiple papers on Hive and sentiment analysis which helped us understand concepts of analytics, sentiment analysis and hive. Below are summaries of some of the papers that we found useful during our analytics:

Sentiment analysis is a powerful tool while doing competitive analysis, detection of unfavourable rumours, market analysis, etc. While doing sentiment analysis, special attention is required towards textual context,relationship to the subject, domain knowledge, common sense as well as linguistic knowledge. Other than adjectives, other content words such as nouns, adverbs, and verbs are also used to express sentiments. [1]

There are various predictions methods like Regression method, Bayes classifier, K- nearest neighbour classifier, Artificial Neural network, Decision tree, Model based prediction. Predictions using social media is comparatively new method

an has relatively low accuracy, but it allows us to use wisdom/opinions of crowd in constructive manner. As many users share their opinion via social media, social media provide us aggregated source of different view points which gives us opportunity to analyze actions and predict future human related events.[2]

In this alaytics we used different data sources having different formats. We had to write different map reduce jobs, hive queries to extract restaurant specific data and process the available data to get expected results. languages like HiveQL make such analysis easy.HiveQL enables users to plug in custom map-reduce scripts into queries. The language includes a type system with support for tables containing primitive types, collections like arrays and maps, and nested compositions of the same.[3]

There are different applications like Yelp, Foursquare, Seamless, etc available which provide restaurant ratings and help user select perfect restaurant. But none of these applications suggest speciality of a restaurant. This is what makes Munchies different. Along with restaurant's overall ratings, Munchies display "must-have" dishes from that restaurant. In this analytics we have used Sentiment Analysis tools like Alchemy to analyse tone of the user review.We are calculating average scores for each of the dishes and displaying top rated dishes to the user as "must-have" dishes. We believe that our suggestions would help users select best of the best from a restaurant.

IV. Design

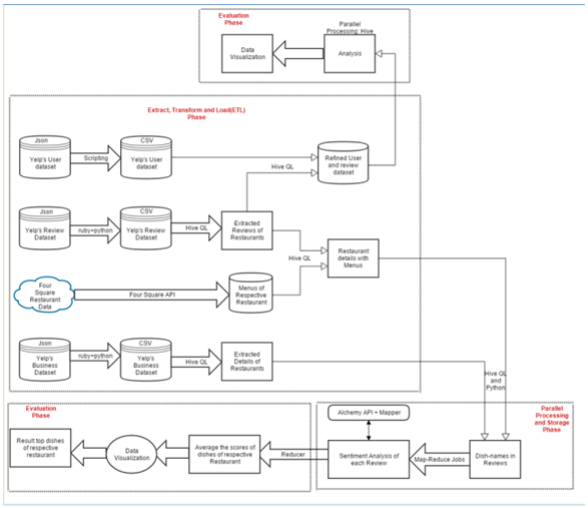

fig.1 Flow Design

The design of the analytics mainly focuses on the Yelps Data sets and the scraped menu data set using Foursquare API. As shown in Fig. 1, the data sets are cleansed and then further processed to visualize the data and combine with other data sets to have meaningful information. Although all the data sets were readily available but the main challenge was to scrape the menus of the respective restaurants. This obstacle was overcomed by using the Foursquare API. As mentioned in Table 1, amongst the total 61184 business entries, only 3503 were restaurants. We were successfully able to scrape the complete menu of 1100 restaurants.

Entity	No. of rows
user data	366715
review data	1569264
business data	61184
filtered restaurant data	3503
menus scraped for	1100

Table 1: Information about data sets

The eateries were characterized from the business data set based as "Restaurants", "Bars", "Nightlife", "Food", "Coffee and Tea", "Pubs", "Breweries", "Bakeries" and "Desserts" categories. As mentioned in Table 2, Yelp has provided the data sets for 7 states. The successful attempts to scrape the menus of restaurants are listed in the table against respective city which summed up to around 1100.

states	number of restaurants(with successful menu scraping)
AZ	637
IL	34
NC	117
NV	22
PA	118
SC	4
WI	50

Table 2: Successful menu scraping for restaurants

A cumulative attempt of data analysis, by using Hive, python UDFs [5] and Map Reduce [4] jobs we were able to derive the required analytics. The analytics are designed to run on Cloudera QuickStart VM and the data visualization is performed using the Tableau. We used Amazon EMR to do a part of our analysis.

V. Results

Our first analytics was targeted to find the most positive reviewed dishes in a restaurant., within provided Yelp reviews. In the first course the dish names are searched in the reviews, splitted over full stops for the respective restaurants. The output of this job is sent to a mapper reducer job. The mapper takes the restaurant ID, dish name and review sentence as input and outputs the alchemy score for respective dish of a restaurant while the reducer will further combine and average those results.

Table 3 has the details of restaurant Waldhorn Restaurant, located in Pineville, NC. Table 3, shows the output of mapper reducer function. This restaurant has the highest alchemy average for the largest number of unique dishes.

Restaurant_id	Dish Name	Alchemy Output
2_0P2AmSSFtPgddio0MgYA	German Food	0.946145
2_0P2AmSSFtPgddio0MgYA	Sweet/ Dessert	0.943347
2_0P2AmSSFtPgddio0MgYA	Apfelstrudel	0.934015
2_0P2AmSSFtPgddio0MgYA	Cheese Grits	0.932138
2_0P2AmSSFtPgddio0MgYA	International Hoagie	0.92997
2_0P2AmSSFtPgddio0MgYA	Portobello Mushroom Sandwich	0.924842
2_0P2AmSSFtPgddio0MgYA	Red Cabbage	0.924842
2_0P2AmSSFtPgddio0MgYA	Rindsrouladen	0.923473
2_0P2AmSSFtPgddio0MgYA	Sauerbraten	0.922865
2_0P2AmSSFtPgddio0MgYA	Schokoladenmus	0.920764
2_0P2AmSSFtPgddio0MgYA	Spaten Oktoberfest	0.91635
2_0P2AmSSFtPgddio0MgYA	Ungarische Gulaschsuppe	0.900421
2_0P2AmSSFtPgddio0MgYA	Warm Sour Kraut	0.894123
2_0P2AmSSFtPgddio0MgYA	Wiener Schnitzel	0.87921

Table 3: Highest rated review dished of a restaurant

For the usability of this analytics, we have made an android application which show the restaurants in the 5 miles vicinity of the user. When a user selects a restaurant, the best reviewed dishes from their menu is shown to the user.

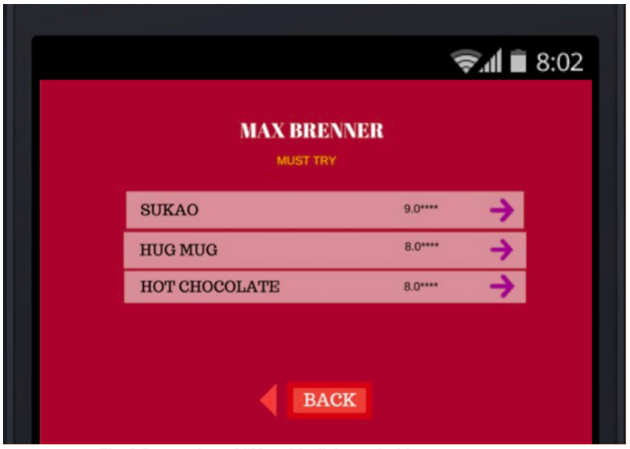

Fig. 2 Screenshot of "Munchies" the android app

In Fig. 2 the user has selected Max Brenner restaurant, which is displaying 3 "Must Try" dishes, on the basis of our analysis. The analysis output for Max Brenner restaurant is:

Restaurant_id	Dish Name	Alchemy Output
YTqSkgOiil9HG3hNH7ovfg	sukao	0.872533
YTqSkgOiil9HG3hNH7ovfg	hug mug	0.774194
YTqSkgOiil9HG3hNH7ovfg	hot chocolate	0.7639899999
YTqSkgOiil9HG3hNH7ovfg	Chocolate Chunks Pizza	0.4608065

According to the output pattern of Alchemy API, the average below 0.4 is neutral and towards negative. Thus, our analytics cater the dishes more than 0.4 average as positive and show the users the best reviews dishes in the app.

In our second analytics, we have used user dataset of Yelp. The user's data set the total number of users are 174094. After the analytics it can be deduced that average number of friends of a user is 14.73(Table 4).

No. of users	Average number of friends of a user
174094	14.79

Table 4: User analytics for friends

Furthermore the information derived from the dataset was divided in 9 bins, as shown in Table5. It was analyzed that 25.65 percent of users do not have any friends. And, 49.34 percent have less than 10. The analytics is plotted in Fig. 3.

Range	Number of friends
More than 500	363
More than 400 and less than 500	150
More than 200 and less than 400	1022
More than 100 and less than 200	2420
More than 50 and less than 100	5496
More than 20 and less than 50	15155
More than 10 and less than 20	18912
More than 1 and less than 10	85907
ZERO	44669

Table 5: User ranges for number of friends

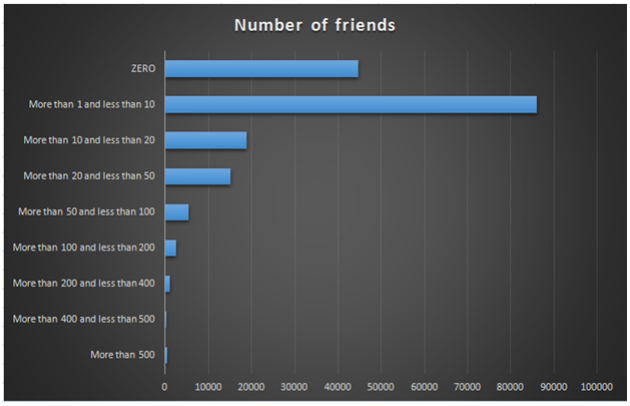

Fig. 3 User range for number of friends

It was analysed that out of the top(wrote most most number of reviews) 8 reviewers only 2 users had one friend each, while the other 6 had no friends. Thus, it was deduced

Yelping-since	No. of Reviews	Name	No. of friends
2013-08	8843	Scott	ZERO
2014-07	4573	Robin	ZERO
2013-04	4534	Bobby	ZERO
2014-01	4323	Chase	ZERO
2014-07	4270	Robert	ZERO
2012-02	4270	Julie	ONE
2012-07	3865	Randy	ONE
2010-06	3392	Schnaubbi	ZERO
2007-09	3380	Jon	ZERO

Table 6: Most active users' correlation with number of friends

In the data set there are only 7 users who havent written any reviews and for the rest of the users the number of reviews range from 0 to 8843. During our theoretical analysis and reading through research work [6], we were working on the correlation of active users(who have written most number of reviews) with the number of friends they have and how they have social networking. But it turned out that the most active users have one or zero friends, as shown in Table 6.

VI. FUTURE WORK

Our analytics over the business and reviews data turned out to have a good correlation. Our correlation was hugely dependable on the menu data, which was scraped from Foursquare API. This API had a restriction of downloading limited set of information, due to which the number of scraped menus were below our expectations. A cleaner and faster way can be determined to scrape the menus. Secondly, finding the patterns(dish names) in the reviews can be improved.

VII. CONCLUSION

We were successfully able to find the dish names in the user reviews. The respective dishes and reviews were send to Alchemy API to calculate the average scores for the dish. We are able to display the top dishes for a restaurant in the android application. In the second analytics, over the user data where we wanted to show the correlation between the activity of users and number of friends turned out to be opposite. The most active users were found to have to least or no friends at all.

ACKNOWLEDGMENT

Yelp for providing us the proprietary data for meaningful analysis. Prof. Suzanne McIntosh for her constant support every week during break hours/after class hours. Amazon for providing voucher to use the AWS services.

REFERENCES

[1] Sentiment Analysis: Capturing Favorability Using Natural Language Processing

[2] A survey of Prediction using social media - Sheng Yu, Subhash Kak; Oklahoma StateUniversity

[3] A Performance Evaluation of Hive for Scientific Data Management

[4] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implementation, 2004

[5] Hive Operators and UDFs documentation https://cwiki.apache.org/confluence/display/Hive/LanguageManual

[6] Nasrullah Memom, Jenniffer jie Xu, David L. Hicks. DataMining of Social Network Data