Yelp Restaurant Review Word Association and Recommender Systems

Ehsan Hosseinzadeh Khaligh, Shah Newaz

Department of Computer Science, California State University, Sacramento
6000 J Street Sacramento, CA 95819

ehosseinzadeh@csus.edu

shahnewaz@csus.edu

***Abstract*—Nowadays many businesses and restaurants struggle to increase their sales. It is really hard to come up with new strategies to increase the sales without knowing exactly what customers are looking for. Machine learning algorithms give us the capability to create models, and do predictions based on the created models. Predictions may help to come up with new strategies. By looking at the results and asking the below questions, businesses and restaurants will know which improvements they need to make. Do you want to open a business of some sort? For example, let's say that you are a coffee connoisseur and you want to open a coffee shop . Ever wondered why there is a coffee shop right next to the library? or a park right next to a home depot? Is this business good for kids? Does this business provide WiFi for their customers? Does it have adequate Parking? How much of the business's success is dependent on just the location? What are the differences in similar business between cities and states? What are the items that customers order and buy together? We will investigate these kinds of characteristics by applying different data science algorithms to the Yelp dataset. In this paper, we look at two algorithms which are association and recommendation systems. By looking at Yelp restaurant reviews and applying association algorithm, we can find word relations. Word relations help to identify string rules between the words which may be the answer of mentioned questions. At the same time, applying simple recommendation algorithm to the Yelp review dataset helps to find highly recommended restaurants. Other businesses can also see these highly recommended restaurants. They can use the recommended restaurants as a base to learn from their strategies and techniques to improve the quality of their own business. The outcome would be to gain knowledge regarding what approaches other businesses need to take in order to increase their sales.**

I. Introduction

Increasing sales is one of the critical roles for a business to continue functioning efficiently. Companies use other advertising companies like Yelp to promote their businesses and products so that they can increase their sales. In turn, Yelp makes a profit by selling advertisements. Customers also use Yelp to seek out businesses that fit their need. The customers get what they want, the businesses get the customers and Yelp not only make money through advertisement, but they also store data for other data science researchers. Having a large set of data helps to create models and apply machine learning algorithms. By representing each restaurant as a parent node and customers as child nodes, we can have the basic structure which can be used in many machine learning models and algorithms. Yelp provides "Restaurant" datasets which have a user, photos, review, tip, check-in information. The datasets contain more than five million records of data for data mining.

II. Related Work

Many big private tech companies use big data as a key ingredient for part of their recommendation functionalities. For example, YouTube has a recommendation column that will list many videos based on the previous data they have amassed regarding your search habit.

Google uses GSA(Google Search Appliance) to recommend words or sentences that you would search for next. Google Search can also use exact keypresses to recommend full sentences based on the data collected from your previous interactions. Google search focuses more on displaying recommendation based on behavioral patterns and what you search most often for the keywords themselves.

There is also the elasticsearch, which started out as a large open-source project. Large companies like Amazon use elasticsearch for their recommendation functionality. Not all companies have a budget like Amazon but since elasticsearch is open-source, smaller companies with smaller budgets can still use the features provided by elasticsearch.

There is also the public data sites like data.gov that has open public data that you can use. The site itself has

recommendation/relational features that will help you find the exact data that you would be looking for.
For this project, we didn't have to dig around the public data sites to get free data. Fortunately, we were provided a large amount of data by Yelp, but we won't build an application like a Google or Amazon search recommendation engines to go through the dataset. We don't have the resources to operate at that scale. We will just use the Yelp data to show what businesses and keywords relate together. We will also use recommendation to show the top restaurants.

III. Data Description

Our dataset is provided by Yelp as a part of their round 11 challenge. The challenge kick off started on January 18. 2018 and ran through June 30, 2018. The dataset includes information about local businesses in 11 metropolitan areas across four countries.

Concurrently, the dataset contains five unique files which are check-in, photos, review, tip, and user. Each file is in JSON format. Each file contains JSON object per line.

For this project, we use to review and business JSON files. The business file is represented as a 'business.json' with the following attributes: business id, name, neighborhood, address, city, state, postal code, latitude, longitude, stars (integer values between and including 1 and 5), review count, categories, hours, categories, attributes, and etc. The review is represented as a 'review.json' with the following attributes: review id, business id, stars s (integer values between and including 1 and 5), date, text, useful s (integer values between and including 1 and 5), funny, and cool.

Different categories of businesses are represented in this yelp data such as restaurants, shopping, hotels, travel, and etc. For this project, we focus on only restaurants, food-related businesses, and reviews associated with each restaurant.

TABLE I
Business json file example

Attributes	Values
business_id	o9eMRCWt5PkpLDE0gOPtcQ
name	Messina
neighborhood	N/A
address	Richterstr. 11
city	Stuttgart
state	BW
postal_code	70567
latitude	48.7272
longitude	9.14795
stars	4.0
review_count	5
is_open	1
attributes	GoodForMeal ...
categories	Italian, Restaurants
hours	Monday: 18:00-0:00

TABLE II
Review json file example

Attributes	Values
review_id	v0i_UHJMo_hPBq9bxWvW4w
user_id	bv2nCi5Qv5vroFiqKGopiw
business_id	0W4lkclzZThpx3V65bVgig
stars	5
date	2016-05-28
text	Love the staff, love the meat, love the place. Prepare for a long line around lunch or dinner hours ...
useful	0
funny	0
cool	0

IV. Experimental Setup and Approach

In this project, we used Jupyter-notebook to write all the python code. The prerequisite for this project required us to use Anaconda distribution to install the python and Jupyter-notebook dependencies. After the installation of software, we used Apriori and association_rules provided by

the mlxtend library for the association. For our second part, we have used the recommendation system (Simple Recommenders) to show the highly recommended restaurants.

1. Association

Association rules matter when we are trying to find things that relate to each other. For example, when a business announces that they have something for sale, they might raise the price of something or promote something that is closely related to the item on sale. Not just for businesses, association rules can also be applicable to consumers. In our project, we use a data science technique called 'Apriori' to figure out the closest related businesses.

1.1 Preprocessing

First, we wrote a simple python code to convert the 'review.json' file to a text file format. Because the 'review.json' is not a standard JSON format, Python Pandas library cannot load the data and it will be expensive to standardize the JSON. By using the read_table function provided by Pandas and passing 'sep=",\""' argument we simply loaded the data into a dataframe and we used 'apply()' function to remove the quotations, attribute names, and the punctuations. This approach is cheaper in comparison to standardizing the JSON and takes less time to load the data.

In the next step, we 'grouped by' the same user_ids to have had distinct user_ids in each row.

Furthermore, we wrote a function to remove all the special characters in the 'text' column/attribute and another function to remove all the stop words. We used 'nltk' python library to import a list of stop words in English to use in the function. Special characters and stop words are noises which may affect the model and the final result.

In the final step of preprocessing, we wrote a function to store the words in a list type.

1.2 Statistical analysis

Before creating a model and applying machine learning techniques, it is a good idea to find a few statistics of data/data-frame such number of users, number of unique words used by each user, a total of words used by each person including repetitive words.

There are statistics of our cleaned data:

1.2.1 Number of unique words by each user

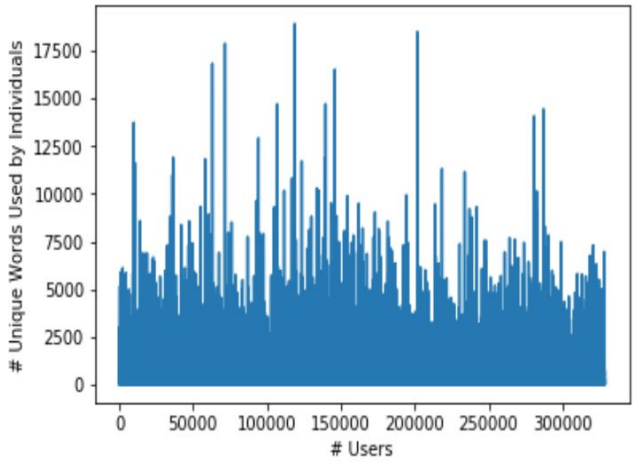

Fig 1, Number of unique words by each use

1.2.3 Total number of words used by each user

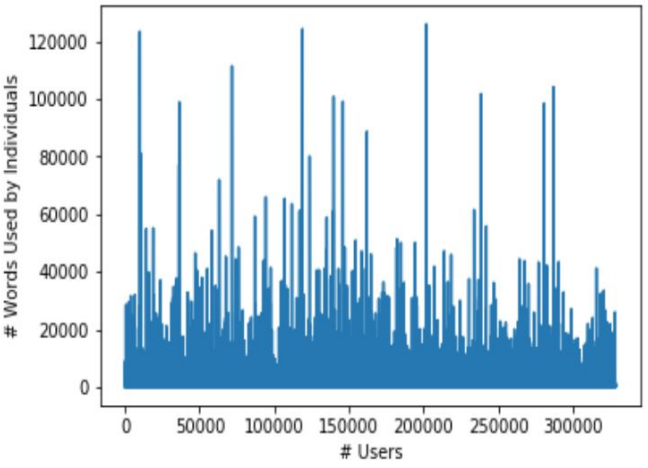

Fig 2, number of words used by each user including repetitive words

1.2.3 total number of users 328727

1.3 apriori algorithm Approach

In order to apply Apriori algorithm, we vectorized our data using CountVectorizer provided by Sklearn feature extraction and then used MLXTEND apriori implementation to run the Apriori algorithm. Our minimum support for this algorithm is 0.05.

We filtered the items sets whose length is greater/equal to two and the support is greater/equal to 0.08.

In the final step, we used 'association_rules' provided by MLXTEND with 'lift' metric and minimum threshold equal to 0.04 to create the rules. Finally, we selected the items whose

lift is greater/equal to 1 and the confidence is greater/equal 0.8.

2.Simple Recommendation

Simple Recommendation technique is one of the most widely used data science technique is many well-known companies. Any website with any sort of a search mechanism would be applying some sort of the simple recommendation technique. There are even music apps like Pandora or Spotify that uses some sort of a recommendation algorithmic technique to tell you what music you might like listening to. Yelp also can use it to tell you what food you might enjoy as you type in the yelp search field. In our project, we used a weighted rating equation to find the top five restaurants from the Yelp dataset.

2.1 Preprocessing

Similar to the Association (last part) we use the converted 'review.json' to a text file format to load the data into a dataframe. For simple recommender, we only need 'user_id', 'business_id', 'stars', and 'useful' columns/attributes. The cleaning part of the column such as removing quotations and attribute names is similar to the last part.

2.2 Applying Simple Recommendation

After cleaning the data, we used 'mean()' function provided by Pandas to find the average rate of stars for all the users. Then we used 'quantile' function provided by Pandas followed by the 90th percentile as our cutoff. The items must have values more than at least 90% of all reviews. We applied this function to the 'useful' rating column.

Furthermore, we used the following weighted rating (WR) equation to rank the restaurants:

$$WR = (V / (V + M) * R) + (M / (M + V) * C)$$

Where:

V is the 'stars' value for each record
M is the 'useful' rate, received by the user in the 90th percentile
R is the 'useful' value for each record
C is the mean/average rate of 'stars'

Also, IMDB uses this equation for their most recommended movies.

By applying WR for each restaurant we get a score. Finally, by storing values we can find highly recommended restaurants.
Depending on the client's review, a restaurant with a higher score will be of a higher quality than the average restaurant.

Finally, by looking at 'business.json' or 'business.txt' files we can find more information about that restaurant by search the 'bussiness_id'.

V. Results and Analysis

1. Association

The following table display the rules associated with the words:

ANTECEDENTS	CONSEQUENTS
great, customer service	service
customer	service
good, restaurant	food
great, restaurant	food
place, restaurant	food
service, restaurant	food
highly, great	recommend
the, service, restaurant	food

2. Recommendation

The top 5 recommended restaurants are:

1. name: Amy's Baking Company
   neighborhood: N/A
   address: 7366 E Shea Blvd, Ste 112
   city: Scottsdale
   state: AZ
   postal code: 85260

2. name: In-N-Out Burger
   neighborhood: Westside
   address: 2900 W Sahara Ave
   city: Las Vegas
   state: NV
   postal code: 89102

3. name: 99 Cents Only Stores
   neighborhood : University
   address: 1325 E Flamingo Rd
   city: Las Vegas
   state: NV
   postal code: 89119

4. name: Fremont Street Experience
   neighborhood: Downtown
   address: 425 Fremont St
   city: Las Vegas
   state: NV
   postal code: 89101

5. name: F Pigalle
   neighborhood: Downtown
   address: 508 E Fremont St
   city: Las Vegas
   state: NV
   postal code: 89101

VI. Conclusion and Future work

1. Association

As we can see in the association result table in the previous section, all of the rules are meaningful and we can use the rules to apply it in future machine learning models.

For example, when users type 'great customer' and we know it follows by 'the word service', Yelp recommendation system can suggest this sentence 'This restaurant/business has a great customer service'.

In future, we plan to apply the Apriori algorithm for a different type of businesses and other data sets provided by Yelp. Also, try applying Apriori with different implementation provided by Dr. Christian Borgelt in University of Konstanz, Germany.

2. Recommendation

These are the top recommended restaurants

2.1. 'Amy's Baking Company': This business had the highest score and fairly high useful score. Not sure what happened to this business but it showed up 6 times in results. We know this because the business id stayed the same but the user_id changed multiple times. Upon further research, we found out that the business had changed owners multiple times and has been closed since August 2015. Currently, Yelp has been monitoring the page. People still reviews the yelp page for some reason. On top of the yelp page for Amy's Baking Company, it says "This business is being monitored by Yelp's Support team for content related to media reports."

2.2. 'In-N-Out': This business had the second highest score and useful score. In-n-out is one of the more popular fast food burger chains. Even though it's a fast food chain business and there is a lot of In-n-out places around the country, the spot in las vegas scored fairly high scores in comparison to other fast food chains. People from all over the world come to Vegas and are impressed when they find a fast food place with food better than regular fast food restaurant. This convinces a lot of people to leave a lot of reviews.

2.3. '99 Cents Only Stores': This business seems like it got a lot of reviews because of the amount of competition with the 'Dollar Tree' store nearby. Also, this store is located in Las Vegas, a lot of tourists go here to buy many essential items for cheap prices.

2.4. Fremont Street Experience: This is one of the unique places in Las Vegas. Lots of lights, lots of food, music and a very busy place in general. It's connected to a lot of bars, casinos, and restaurants.

2.5. F Pigalle: This was an 'adult' fondue place in Las Vegas. The restaurant abruptly closed down for what was assumed to be because of renovation. Rumor has it that the place is cursed and lots of other businesses opened and quickly closed here in the past.

3. Final Conclusion

According to our results from part 1 and 2, we see that the top 5 restaurants and businesses value customer service a lot. Yelp is driven by the opinions of lots of people. Our results show that the touristy area generally gets a high amount of yelp feedbacks regardless of how good your actual business is. The good touristy location will get you more reviews, more reviews show that you have a lot of customers with a high amount of sales.

4. future work

For future work, instead of first applying Apriori and Simple Recommendation, we shall try going through the data and look at all the restaurants that have the highest amount of reviews. Then clean the data accordingly, apply our algorithms, followed by more research with the results to get a more concrete conclusion. We shall plan to continue working on this project and revise the paper further before considering for publication.

VII. Restriction and complains

Because of computer system limitation and having a very large dataset. We split the original 'review.txt' file into 5 equal parts and ran the experiments and algorithms in one of the files.

The Jupyter notebook has a running limitation and when we ran Apriori algorithm for very large data frame size such (30000,500000), it stops kernel. This is a restriction by notebook which may happen for similar size datasets.

VIII. Lessons Learned and task Division

1.Lessons Learned:

1.1 Working on a research project first requires deeply understanding the problem and read as much as possible similar published papers.

1.2 We learned about Jupiter notebook limitation on running times and how to avoid 'kernel died' in the notebook.

1.3 we explored different types of recommendation systems such as simple recommender and content-based recommender.

1.4 We learned more about different types of implementation of apriori algorithm.

1.5 We learned how to work with a large set of Yelp data. The Yelp data size was about 8 gigabytes. We had to reduce the size of data and clean the data by removing unneeded parts before we could continue and apply our algorithms.

2.Task Division

Name	ID	Parts	Hours
Ehsan Hosseinzadeh	218865310	All	40
Shah Newaz	205053957	All	40

IX. References

@spantreellc, Spantree -. "Elasticsearch vs Google Search Appliance." *Spantree Technology Group, LLC*, www.spantree.net/blog/2013/09/04/elasticsearch-vs-google-search-appliance.html.

Saam, Conrad, et al. "Just How Much Does Yelp Cost?" *Mockingbird Marketing*, 3 May 2014, mockingbird.marketing/just-much-yelp-cost/.

Yelp.com, www.yelp.com/biz/amys-baking-company-scottsdale.

Borgelt, Christian. *Christian Borgelt's Web Pages*, www.borgelt.net/apriori.html.

Yelp.com, www.yelp.com/biz/f-pigalle-las-vegas.

Lilly, Caitlin. "Downtown Las Vegas Eatery F. Pigalle Abruptly Closes Its Doors." *Las Vegas Review-Journal*, 19 Feb. 2017, www.reviewjournal.com/entertainment/food/downtown-las-vegas-eatery-f-pigalle-abruptly-closes-its-doors/.