Exploratory Research on Yelp’s Local Recommendation System Using

Song Jiang Mathematics and Statistics

Department Georgia State University

25 Park Pl NE, Atlanta, GA 30303 [sjiang11@student.gsu.edu](mailto:sjiang11@student.gsu.edu)

Tianqi Luan Mathematics and Statistics

Department Georgia State University

25 Park Pl NE, Atlanta, GA 30303 [tluan1@student.gsu.edu](mailto:%20tluan1@student.gsu.edu)

**Abstract—The Yelp Dataset concludes data collected from 8,021,122 reviews and 209,393 businesses located in 10 major metropolitan areas. As a comprehensive dataset includes multiple aspects with respect to the businesses. We’re interested to find the reliability of Yelp’s recommendation algorithm by building our own food recommendation algorithm based on Machine Learning.**

**Keywords—Yelp, EDA**

1. INTRODUCTION

As a leading company in local business recommendation in north America. As of the second quarter of 2019, Yelp reported having a monthly average of 61.8 million unique visitors via desktop computer and 76.7 million unique visitors via its mobile website according to their 2019 Q2 Shareholder Letter. As of June 30, 2019, Yelp stated on their Investor Relations page that it had 192 million reviews on its site.[1] From massive amount of user base, business information and reviews, Yelp collected large amount of data and developed their own business recommendation system. Since recommendations are highly subjective, there has been a non-stop exploration on how to make a more comprehensive recommendation system.[2, 3]

The dataset used for this project is readily available from Yelp Challenge at this link.[4] The dataset contains information about businesses as well as users and the reviews they left in 10 metropolitan areas across four countries. There are a total of 6 datasets, among which we will mainly work with checkin.json, business.json and user.json for our model construction. Review.json is the largest one, containing over 4.7 millions reviews on Yelp from 2004 to 2017, and has a size of 3.82 GB which could be a perfect dataset for practicing Natural Language Processing (NLP). Each review is accompanied by its business id, user id, date, stars (on a scale of 1 to 5), review id and its original text. All ids in this dataset are represented by randomly assigned combinations of numbers and letters for protecting the privacy of users while serving as unique identifiers.

1. Data Description

All .json files are converted to .csv files for the ease of further analysis. All the files are then trimmed down in order to compensate for the lack of computing power of group members PCs and laptops while maintaining enough data points for training and testing models. The rows left are mainly data from Arizona state.

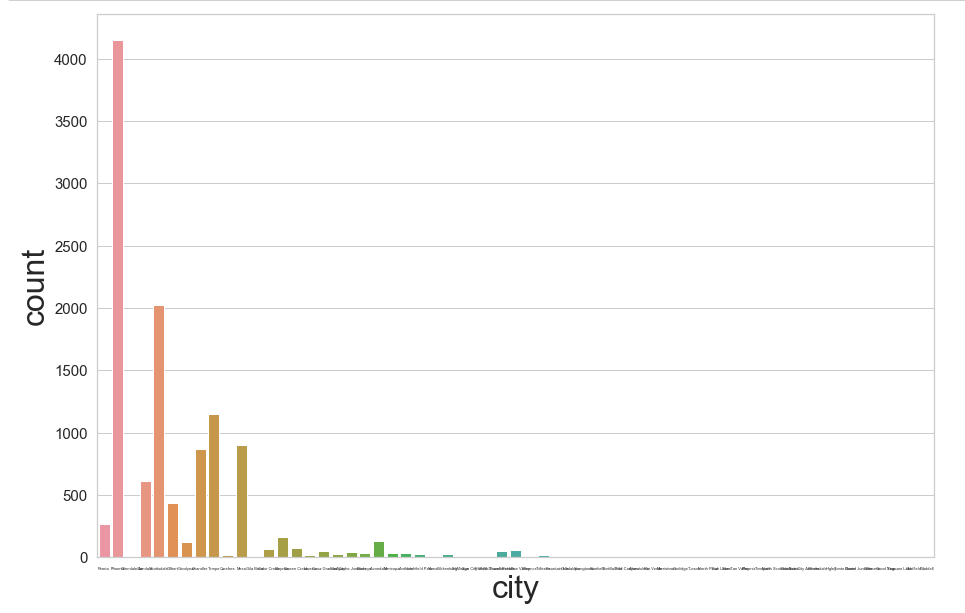


Figure 1. Business number correspond to cities

As shown in (Figure 1), most business registered are concentrated in the larger cities. Phoenix, as the largest city in Arizona has the most business registered, takes almost half of all AZ’s business counts. As part of the business.csv dataset. The attribute “neighborhood” was dropped because the column is empty originally(Figure 2) and also overlapping with the purpose of the attribute “city”.

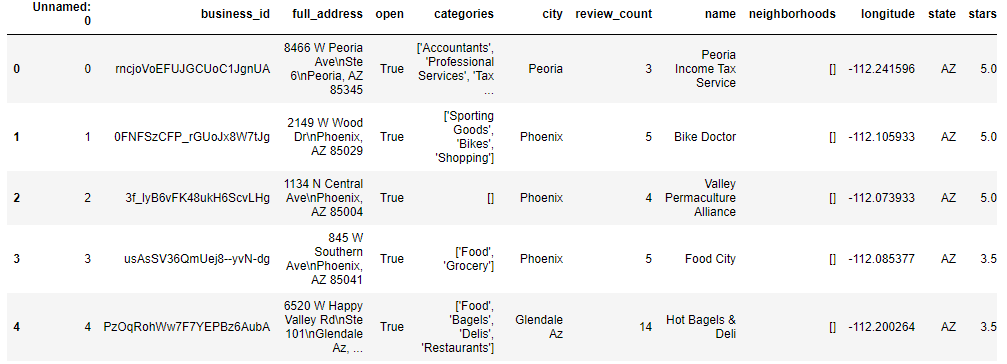


Figure 2. Original business.cs

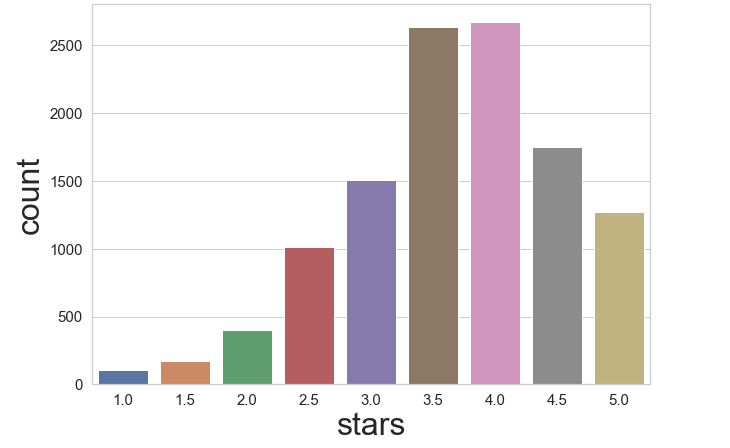


Figure 3. Business ratings

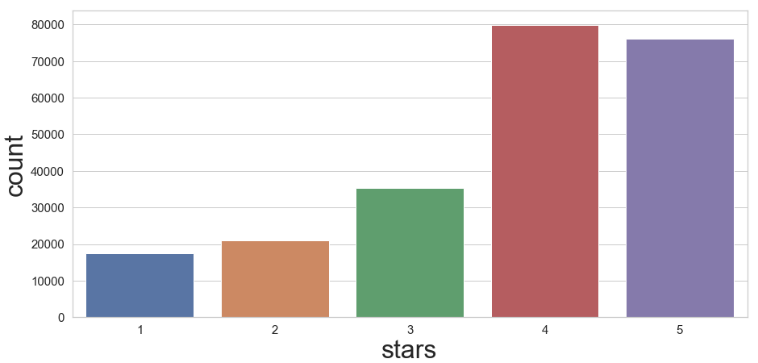
Another crucial attribute when it comes to business is the stars. Yelp uses numerical rating system for all businesses, users can rate a business from 1 to 5 stars. In Figure 3 is the histogram contains all ratings among all business. The overall shape of this histogram is a Gaussian distribution with a bias to higher stars. This phenomenon is also reflected in review.csv which is data from the stars user given in each individual reviews(Figure 4).

Figure 4. Stars given by individual reviews

We can see most of the reviews are either 4 stars or 5 stars in which 4 stars being the most populated and 5 stars a close second. But this doesn’t necessarily mean each single user is going to preferably give a business positive review. From further investigation into the user.csv dataset we learned that ratings are highly user dependent and different users have drastic diverse tendency of giving higher ratings.

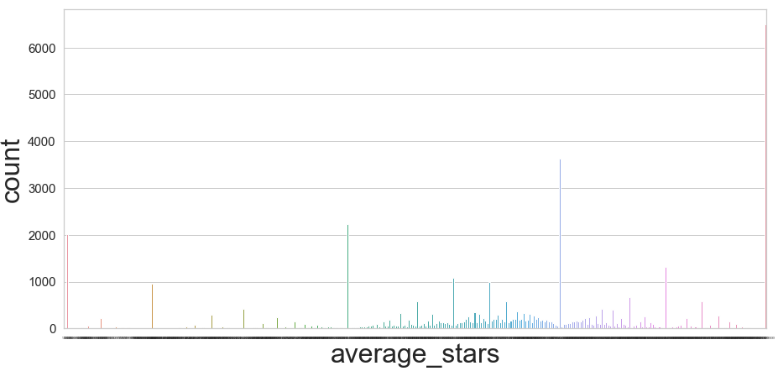


Figure 5. All reviews stars given by different users

In Figure 5 we explored the tendency of individual behaviors impact on business reviews. From the five sharp peaks we can conclude that there are users specifically give certain stars in all of their reviews. A good example is the tallest peak of five stars, which means there are a large amount of users would leave 5 stars in all of their reviews. This phenomenon partially explained the positively biased overall rating numbers.

This type of behavior carries on in some of the other fields. The most typical one being the “vote” attribute in review.csv and user.csv. Each user can vote a review as “funny”, “useful” or “cool”, and user will receive the vote as part of their user credit. Yet, there are a big percentage of user reviews do not get any votes either because their quality is not so high or overall not so impressive.

 From review.csv it is clear that 36% of the reviews have no votes at all and “useful” being the most voted. This kind of lack of participation behavior is similar to the star rating system we discussed earlier. This being said, we need to look into the “vote” attribute in user.csv to evaluate our verdict.

As we can see in “vote” attribute in user.csv, 11.8% users never given any votes, and “useful” is the vote most users tend to give to any reviews. As a conclusion, users participation is highly diverse in Yelp reviews and ratings, most “casual” users do not have the tendency to leave reviews or votes so recommendation for them could be harder than the “involved” users who leave more data on either website or mobile APP for further evaluation.

1. CONCLUSION
2. REFERENCES

[1] Jeffrey M. O'Brien, "Yelp's Ambitious Plan to Take Over the Local Ad Market". CNNMoney.com. p. 2.

[2] Sawant, S., Pai, G., “Yelp Food Recommendation System”. pdfs.semanticscholar.org

[3] A. Sihombing and A. C. M. Fong, "Fake Review Detection on Yelp Dataset Using Classification Techniques in Machine Learning," *2019 International Conference on contemporary Computing and Informatics (IC3I)*, Singapore, Singapore, 2019, pp. 64-68.

[4] <https://www.kaggle.com/yelp-dataset/yelp-dataset>