STAT 605 Project

Xiangyu Wang `xwang2439` Zijin Wang `zwang2548` Jingshan Huang `jhuang456` Yicen Liu `liu943` Yuan Cao `cao234`

December 4, 2020

1 Introduction

As we all know, Yelp is a world-class platform with more than 100 million store reviews worldwide. So we are curious about whether the reviews do affect business market or not. Specifically, our main goal is to explore which aspects are frequently mentioned in the reviews in order to give business owners some feasible tips for them to improve average star. What's more, we also wonder whether there will be any difference shows in reviews among different years which can reflect user habits and preferences. To investigate those questions, we use CHTC to accomplish large scale parallel analysis on "paired words" which will be described in detail later.

2 Methodology

2.1 Description of data

Our Yelp dataset is obtained from Kaggle (`https://www.kaggle.com/yelp-dataset/yelp-dataset`). Yelp is founded in 2004 and it develops, hosts, and markets website and mobile app, which publish user-sourced reviews and ratings about businesses. This dataset is a subset of the businesses, reviews, and user data from Yelp. It contains 8,021,122 reviews, 209,393 businesses, 1,968,703 users and 1,320,761 tips from users and is stored in 4 separate files, respectively. Also, business attributes like hours, parking, availability, and ambience are included in business data.

2.2 Data processing

Firstly, in order to read and filter the data with **awk** more conveniently, we transform the business data and the review data format from .json to .tsv. Secondly, we perform some column transformations in the review data. We reassign the level of stars in the review data as the following rules: reviews with 1-2 stars become level 1 now, 3 stars become level 2 and 4-5 stars become level 3. By doing this, we modify the rating rank from 1-5 stars to simply "bad", "median" and "good". For the date data, we extract only the years by regular expression. Then we merge the column "categories" in the business data into the review data by matching the business_id in both .tsv file. Thirdly, we filter "bad", "median" and "good" reviews given in 2015, 2016, 2017 and 2018, then we split the review.tsv into 5(star rating levels) × 4 (years) = 20 subfiles.

2.3 Computations for Word Frequency

For each subfile mentioned above, we run a R script to count the frequency of “paired words”. The R script tokenize each review and delet all the stopwords, for example “the”, “is”, “at”, “which”, etc. Next, traverse all the review texts in the subfile and search for the consecutively two words, also known as “bigrams”, occurred most frequently. The reason we compute frequecy of the bigrams instead of just the single words is straightforward. Comparing with sigle words like “good”, the bigram “good service” obviously conveys more information to the analysists. For each of the 22 categories, the R script output the top 100 bigrams and write the result as a .csv file. Here we use CHTC to run the R script in 12 parallel jobs. Therefore, after the computing, we obtain 12 × 22 = 264 .csv files, for instance, “review_2015_1_Food_top100.csv”, “review_2016_2_Hotels & Travel_top100.csv”, etc.

2.4 Trend Detection

In this part, we first made some exploring data analysis. For each business category, we compared the number of reviews of different star levels in four years by barplot. Besides, we also explore the change of mean ratings during four years for this business category, i.e., the rating trend through four years, so that we could make business owners aware of changes in their business category and make better choice. For example, Figure 1 shows the changes of mean ratings for “Automotive”, it is obvious that mean rating is decreasing from 2015 to 2018, which suggests automotive businesses are becoming unsatisfying through these years. Under that circumstance, business owners may need to consider some innovation in service or communicate with customers more to understand what they need and care. In order to make this trend more convincing, we perform Mann-Kendall test to evaluate whether this trend is statistically significant. For “Automotive” businesses, our test result shows that the decreasing trend is significant under 0.1 level.

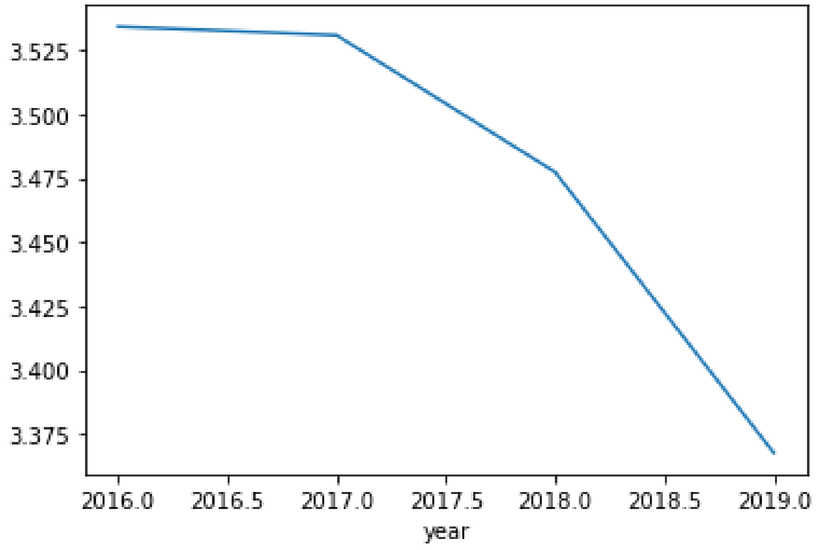

Figure 1: average star in 4 year

2.5 Business Attributes Analysis

In this analysis we first split the data into 23 subsets based on the business categories. Then we fitted an one-way anova model on each of the attribute in every subsets. The following shows the explanation of the attrbutes we have studied in our analysis.

- Delivery: whether the business offer delivery or not.
- Wifi: whether the business offer access to Wifi or not.
- TakeOut: whether the business offer the takeout option or not.
- Bike: whether the business offer the parking place of bike or not.
- Appointment: whether the business accept appointment or not.
- Wheel: whether the business offer wheel to disabled people or not.
- Dog: whether the business accept dog or not.
- Credit card: whether the business accept business card or not.
- Parking: whether the business offer parking place or not.
- Noise: The level of noise around the business.

We have a result of each type of the business which will show the significance and effects of each business attribute on each type of the business. For example, for the Active Life business, ”bike”, ”appointment”, ”wheel”, ”dog” and ”parking” are significant and have positive effects on the rating of business. Although ”Noise” is significant, it has negative effect on the business rating which does make sense. The full results are displayed in our github repository.

2.6 NLP model

After we get high frequency bigrams, we want to measure the effect of some key bigrams to ratings by building a ordinal logistic regression model. It treats ratings as a ordinal variable, and evaluate the bigram effect by using odds ratio of increasing one level rating. Thus, it is easy to interprete the results. We use the presence of a certain bigram, i.e., a logic variable that whether a review contains this bigram, as our predictor. The exponential of model coefficient is the odds ratio that we interested and we will give our advice for business owners based on odds ratio. For example, the coefficient of “customer service_TRUE” is 0.34, meaning if a review does not include “customer service”, then the odds to increase a level of rating (one star to two star) is $1/0.34 = 2.94$ times. Intuitively, this is because customer service shows up with higher frequency in low ratings as complaints than high ratings just as Figure 2 shows. Thus, we will suggest business oweners to realize the importance of customer service and try their best to decrease the complaints about that based on this coefficient result.

3 Conclusion

3.1 Findings

Since our ultimated goal is to give suggestions for business in each category, So our finding will be showed in a report.pdf form, it should follow by four sections:

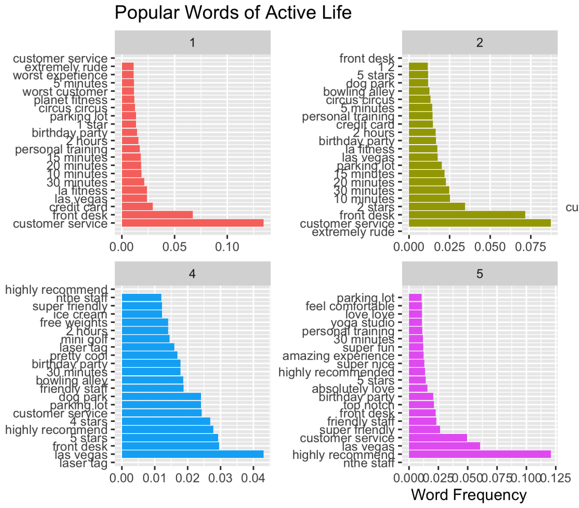

Figure 2: Word Frequency in Different Ratings

- Introduction
- Trend Analysis
- Word Frequency Count Analysis
- Business Attributes Analysis

A sample report of Active Life Business would be attached.

3.2 Difficulties

By now, we do encounter some difficulties, the most exhausting one is that computing the frequency of the bigrams requires extremely large memory. The reason is that the computer should memorize all the counts of possible word pairs in the same time and then sort the frenquecy. We solve this problem by requiring more memories(up to 4GB) in CHTC. However, the script still terminate due to lack of memory when it runs based on 2018 data. We split each 2018 .tsv file into 4 new subfiles to run the R script and merge the result respectively.

3.3 Future work

For future work, we first will adjust the ordinal logistic regression. Also, we will try some easily interpreted models as ordinal logistic regression, such as decision tree or random forest. In decision tree, we can use graph to see decision process directly and feature importance can be accessed in random forest model. We hope to compare the differences between models and try to give suggestion to other kinds of businesses.