STAT 605 Project

Xiangyu Wang xwang2439

Zijin Wang zwang2548

Jingshan Huang jhuang456

Yicen Liu liu943

Yuan Cao cao234

December 11, 2020

1 Introduction

As we all know, Yelp is a world-class platform with more than 100 million store reviews worldwide. 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 and business attributes among different types of them which can reflect user special 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

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 and we also 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. Finally, we filter 1-5 star reviews given in 2015, 2016, 2017 and 2018, then we split the review.tsv into 5(star rating levels) $\times$ 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.

Due to the huge difference of the most frequent word in different categories(verified in section 2.6), we analyse the word frequency result with showing some review examples and give reasonable suggestion for different categories’ businesses. See result sample in attached file, Analysis for Active Life Business.

2.4 Trend Detection

In this part, we compared the number of reviews of different star levels in four years for each business category.

Specifically, we 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.

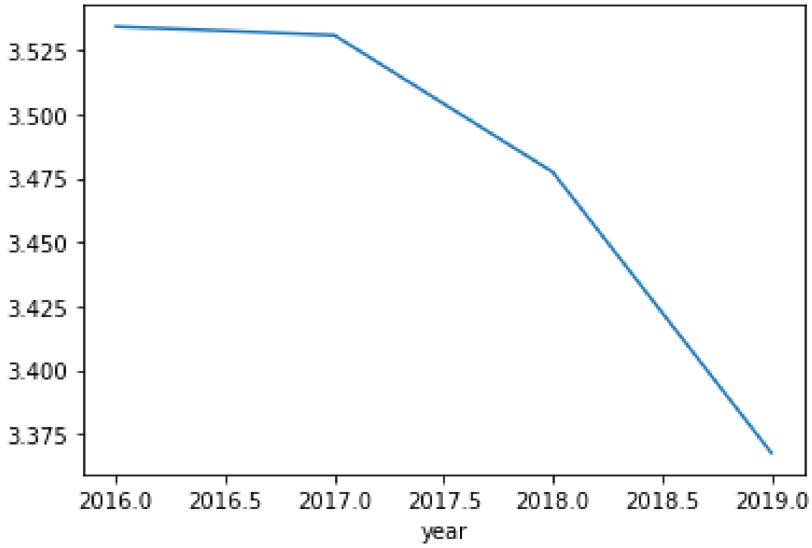
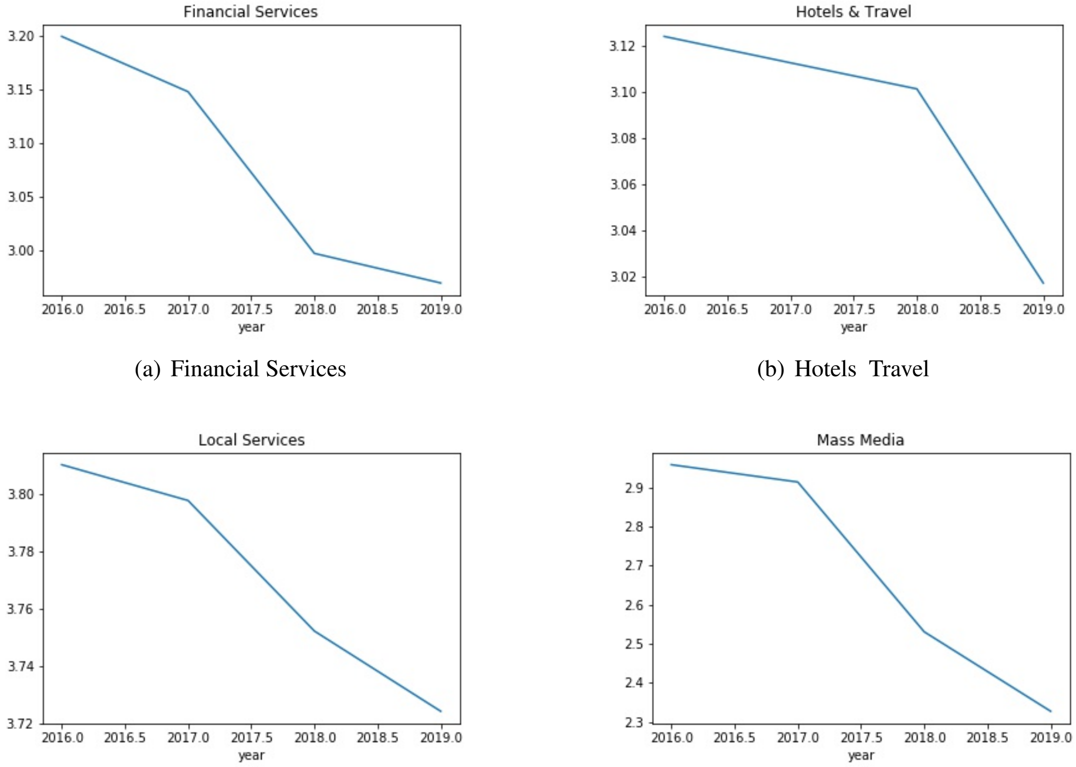


Figure 1: average star in 4 year

Here we also want to show you some trend plot with statistical significance. In fact, we find out many businesses show down trend significantly in these four years, such as ”Financial Services”,”Hotels Travel”, ”Local Services”,”Mass Media” and ”Shopping”.Figure 2.But only one category ”Local Flavor” shows upward trend .Figure 3.



(a) Financial Services

(b) Hotels Travel

(c) Local Services

(d) Mass Media

Figure 2: Business categories with decreasing mean star over 2015-2018

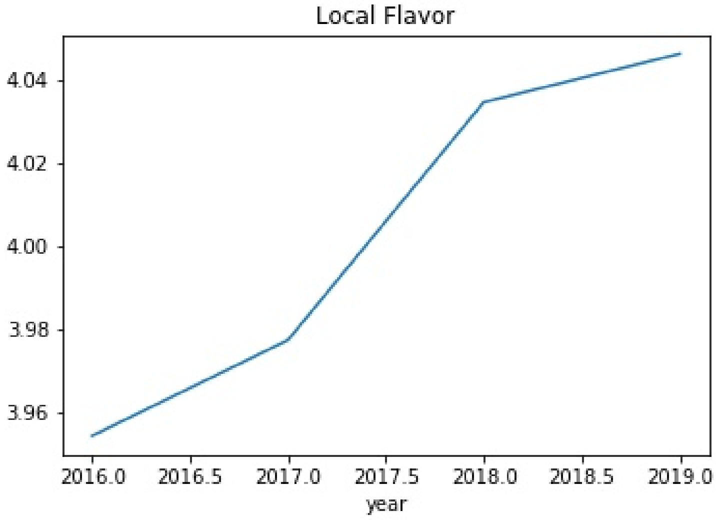


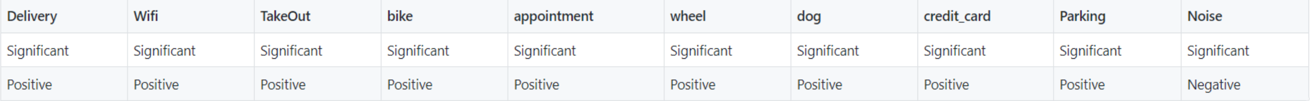
Figure 3: Business category with decreasing mean star over 2015-2018

2.5 Business Attributes Analysis

In this analysis we first split the data into 22 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 attributes 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.

At first, we use the whole data set instead of splitting them by category. As you can see from Figure 4, all attributes are significant statistically but only ”Noise” shows negative effect. This table can tell business owners that, for examples, having free WiFi is good for improving review star level while existing loud noise could lower your average star.

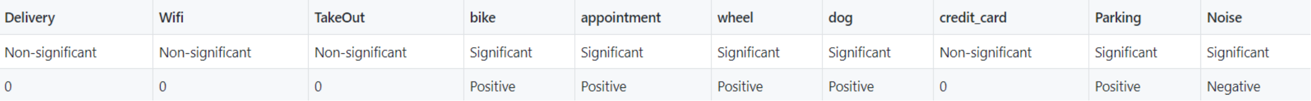


Delivery	Wifi	TakeOut	bike	appointment	wheel	dog	credit_card	Parking	Noise
Significant	Significant	Significant	Significant	Significant	Significant	Significant	Significant	Significant	Significant
Positive	Positive	Positive	Positive	Positive	Positive	Positive	Positive	Positive	Negative

Figure 4: How business attributes affect star

For further result, we found out these attributes shows different significance if you focus only on some specific category Figure 5. So we also have results 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.

We hope these special case could be consider by business owners so that they can pay more attentions on those attributes that matter more.



Delivery	Wifi	TakeOut	bike	appointment	wheel	dog	credit_card	Parking	Noise
Non-significant	Non-significant	Non-significant	Significant	Significant	Significant	Significant	Non-significant	Significant	Significant
0	0	0	Positive	Positive	Positive	Positive	0	Positive	Negative

Figure 5: How Active Life business attributes affect their star

2.6 NLP model

After we get high frequency bigrams for each category, we want to measure the effect of some key bigrams to ratings. Figure 6 and 7 show popular bigrams for two categories, Financial Service and Religious Organizations. In order to assess influence of them based on popular bigrams for each category, we use several logic variables that whether a review contains these bigrams as our predictors and build an ordinal logistic regression model for each of them. The model treats ratings as an ordinal variable, and evaluates bigram effect by using odds ratio of increasing one level rating when a bigram appears or disappears.

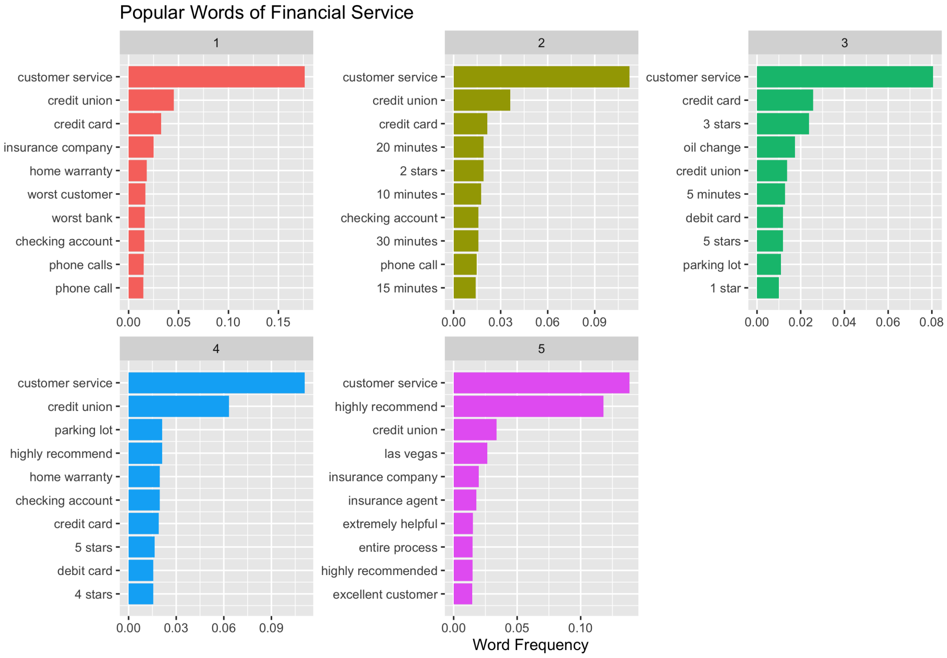


Figure 6: Financial Service

The interpretation of model results is intuitive and simple. For example, the model results after fitting financial service data is shown in Table 1, all coefficients are significant under 0.05 level and odds ratios are exponential of coefficients. For odds ratio less than 1, such as "customer service", it means that *"keeping all other variables same, the odds of increasing one level rating is* $1/0.622 = 1.6$ *times when 'customer service' disappears in a review of financial service"*. Vice versa, for odds ratio greater than 1 such as "insurance company", it means *"keeping all other variables same, the odds of increasing one level rating is* 1.46 *times when 'insurance company' appears in a review of financial service"*. Based on these results, we may suggest business oweners of financial service to realize the importance of customer service and try their best to decrease the complaints about that. Althogh this bigram shows frequntly in all rating levels, it appears more in low ratings. Also, optimizing service process to decrease waiting time, paying attention to basic service such as credit card and bank account are also useful suggestions.

For these NLP models, we do not consider involving "category" as a predictor like the attributes analysis above. The reason is that we believe customers care different aspects for different business category. This diversity can also be found intuitively in Figure 6 and 7. In fact, we find popular bigrams for all reviews, which is shown in Figure 8. Most of these bigrams are "overall", also quite different

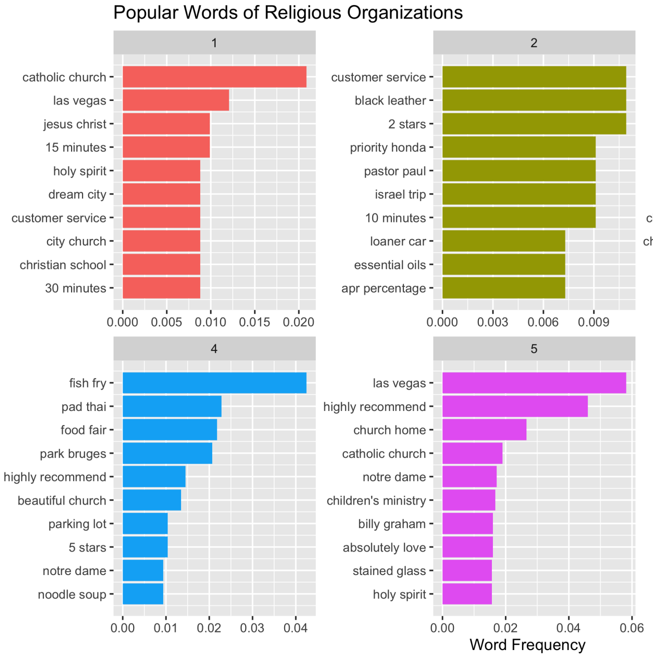


Figure 7: Religious Organization

Bigram	Coefficient	Odds Ratio
customer service	-0.474	0.622
minutes (all related)	-0.674	0.509
credit card	-1.108	0.33
checking account	-0.437	0.646
insurance company	0.37	1.46
credit union	-0.456	0.634

Table 1: Model Results of Financial Service

if compared with high frequncy bigrams in our example Figure 6 and 7. Thus, fitting a model with all review data will bring trouble for selecting popular bigrams and cannot generate useful information for every business category.

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:

- Introduction
- Trend Analysis
- Word Frequency Count Analysis

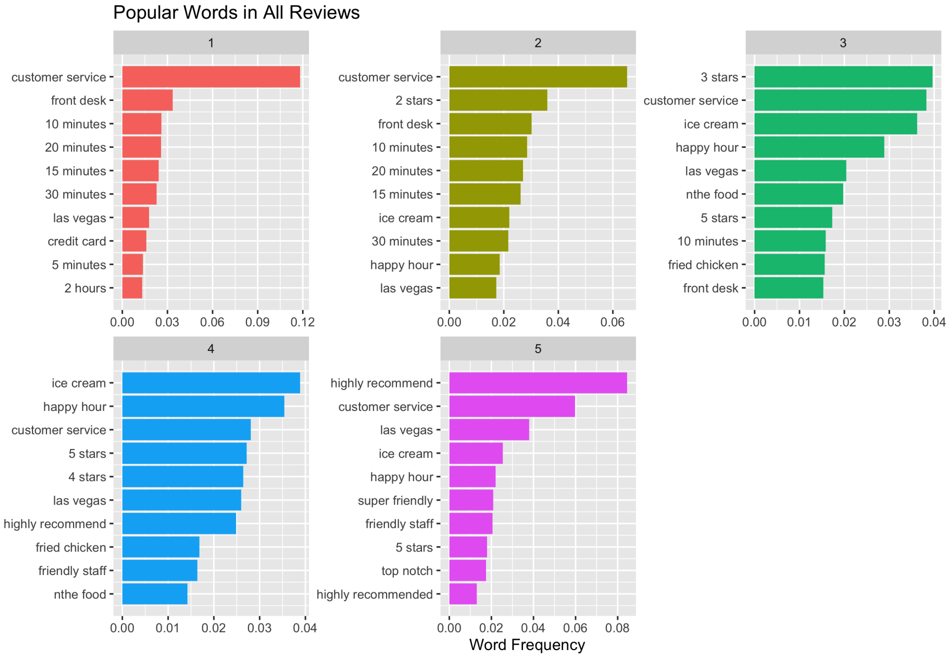


Figure 8: Popular Bigrams for All Categories

- 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.