Stat628-Module 3

Group 1

1 Introduction/Motivation

It is difficult for business owner to get the idea of how to make improvement based on Yelp reviews. Because Yelp only proves the overall rating instead of the rating for some detailed parts, and it is hard for business owner to obtain valid information from a number of reviews, especially when they also want to refer to reviews of some similar businesses. We aim to build a model to extract useful information from reviews, and provide suggestions to business owners to improve their Yelp rating. This project focuses on Chinese restaurants.

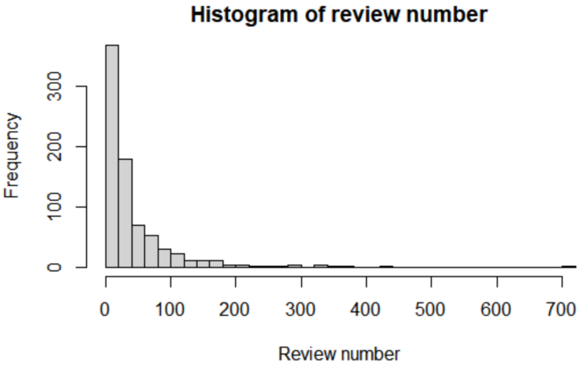


Figure 1: review number

2 Data Cleaning

Original Data:

We use the Yelp review data from open source. This dataset contains 942,027 reviews for 36,327 restaurants in Madison, Cleveland, Pittsburgh and Urbana-Champaign. 'business.json' contains business data including location, attributes and categories(58 variables). 'review.json' contains review information along with user_id and the corresponding business_id.[1]

Pre-processing

1. We are interested in Chinese restaurants so we grep those businesses with category "Chinese" and retains corresponding reviews. We get 778 businesses and 34,028 reviews in total.

2. Among all 778 businesses, we choose not to analyze those restaurants which have only a few reviews(see fig 1) as we won't be able to give a convincing suggestion. We have deleted restaurants having reviews less than 22(50% quantile), resulting 378 restaurants to review in total.

Cleaning

There is one medical shop appears in the dataset which doesn't fit the category of "Restaurant". We guess incorrectly categorized and included in the original dataset, so we just delete it. Business ID is of this shop is, *t4rBeSFDfwvaIbLftLlhig*.

Text Preprocessing

1. Before doing natural language processing for review text, we replace some words in the reviews. For example, some customers will write "hot pot" and we will replace it into "hot-pot" so that it will not be separate into two tokens and lose its meaning when we do the analysis.

2. Then we follow standard practice in text preprocessing[2] to process the text data. Remove stop-words and sentiment-words, we get 30,020 tokens and their corresponding number of occurrences.

3. We singularize all the tokens and aggregate them again, and 24,584 tokens left. We do the singularization here instead of doing it before tokenization to save lots of calculations.

3 Exploratory Data Analysis

Word:

Among all 24,584 words, some of the words has higher frequency than others and vice versa. Many words are meaningless and some of them are spelled incorrectly. Only 3,390 words appear more than 20 times. Naturally, we should pay attention to those words which appear more frequently in our analysis.

word	food	chicken	...	coco	...	gazed
n	35947	15865	...	12	...	1

Attributes:

File business.json contains 39 business attributes, such as Noise Level, WiFi, Smoking. Association test can be performed on those words and ratings to see whether those factors will influence customer's rating for a specific business. But not all of those words appears very frequently, moreover, only 6 words appears more than 500 times. Thus, we will consider more words into our analysis instead of just focusing on attributes.

word	service	price	...	music	...	bike
n	10234	5914	...	337	...	10

Distribution of words:

Compute the mean occurrence of a word across reviews over rating, we can find that the distribution of frequency over star rating is not uniform. That means, there may be an association between words and rating. Thus, we can do further analysis to test whether there is relationship between the specific words and rating. Or, do customers pay particular attention to some words which affects their ratings for the restaurant.

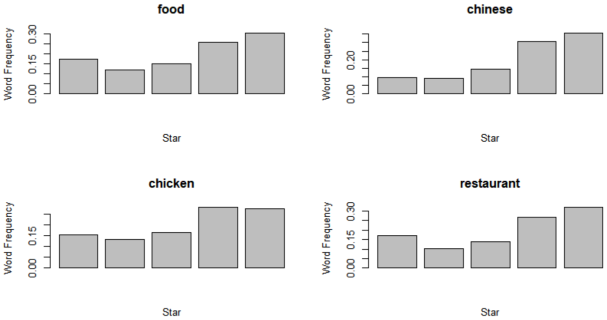


Figure 2: Distribution across all reviews

4 Suggestion Model

4.1 Main Model

The main idea of this model is to see for a specific business, whether its performance on a specific aspect is above or below average. To do this, we first decide the word list, and for each word in it, it is the aspect we can give suggestion about and people frequently mention in their reviews. Then find the keywords for some specific business (frequently used words in its reviews), and they also appear in our word list. Then we do the hypothesis testing to see whether the rating mean for a keyword of a business is different with that of the whole dataset.

Step 1. Decide Word List

We do the following steps to decide the word list we want to work on: (1) Choose those words appear more than 500 times, 383 tokens left; (2) Delete some useless words which are meaningless for our analysis, such as 'eating', 'restaurant', 'lot', 191 tokens left.

Step 2. Decide Keywords

For a specific business, we choose the keywords in the word list that most frequently used in its reviews. First, we tokenize all the reviews and for each words count the numbers that these words appear in the reviews. Then, we collect 18 most frequently used words appeared in the review which is also present in our word list. Then, we ware using these keywords as a basis to give suggestion.

Step 3. Get Samples

To compare the distribution means of the ratings for the reviews containing the keywords for this specific business and all other businesses in our dataset, we need to get the samples for these two distributions. What we need to do is that for each keyword, collect the ratings with reviews containing this keyword for this specific business and all other businesses in our dataset. Then, we can perform hypothesis testing on the two samples for the two distributions.

Step 4. Hypothesis Testing

After getting the samples, we can do the hypothesis testing. To figure out the difference of the two distribution means, we perform one-side Wilcoxon rank sum test with hypothesis:

- $H_0 : R_0 = R_1$
- $H_1 : R_0 < R_1$

where R_0 is the distribution mean of the ratings of this keyword for this business, and R_1 is that for the whole dataset.

Now we can get the p-value for the test. If the p-values are low enough, we can say the performance for this aspect for this keyword of this restaurant is below average.

Step 5. Give Suggestion

By using word count, we are showing a word cloud. It will help restaurants to focus on certain meaningful words, people use most frequently in their reviews.

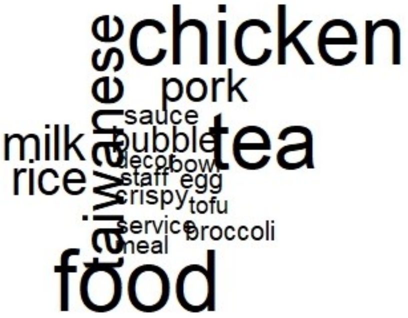
With keywords and their p-values, we can give suggestions to the business based on those significant keywords. We can also give score for each keywords based on the p-value. For example, we can use 100 times the p-value of the first test to compute the score. If the score is 95 or higher, we know the business performs significantly above average for this aspect. If the score is 5 or lower, it performs significantly below average. If the performance of the business for a keyword (for example, Food) is significantly below average, we will give suggestion that they should pay more attention to improving their food quality to improve their rating. If their performance is above average, it means that this is their advantage and we will suggest them keep on it.

4.2 Example

In this subsection, we give an example of a restaurant, implement the analysis on it and show the results and our suggestion to this business.

We choose the restaurant Taiwan Little Eats with business_id ”HVpwpXneaCWMeEBF7H8jpQ” as an example. We first find the 18 most frequently used words used in its reviews within our wordlist. They are *food, chicken, tea, Taiwanese, milk, rice, pork, bubble, crispy, sauce, broccoli, egg, meal, staff, service, bowl, decor, tofu.* We can see the word-cloud below:

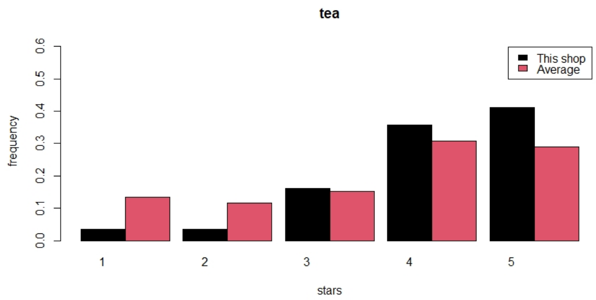
keyword	count	p-value	score
food	53	0.995	99.5
chicken	51	0.983	98.3
tea	49	0.998	99.8
taiwanese	33	0.639	63.9
milk	27	0.995	99.5
rice	27	0.484	48.4
pork	26	0.389	38.9
bubble	18	0.970	97.0
crispy	12	0.667	66.7
sauce	12	0.654	65.4
broccoli	11	0.294	29.4
egg	11	0.324	32.4
meal	10	0.540	54.0
staff	10	0.487	48.7
service	10	0.968	96.8
bowl	10	0.214	21.4
decor	9	0.761	76.1
tofu	9	0.109	10.9

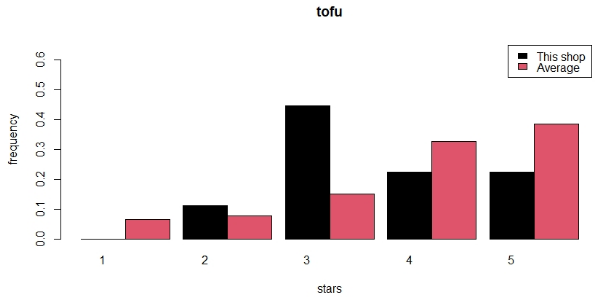


The table shows the keywords with their counts, p-values and significance.

From this table, we can see that for food, chicken, tea, milk, rice, bubble, and service, Taiwan Little Eats performs significantly above average. For tofu, it performs relatively poor(although not that significant).

We have the histogram of the ratings for tea and tofu:



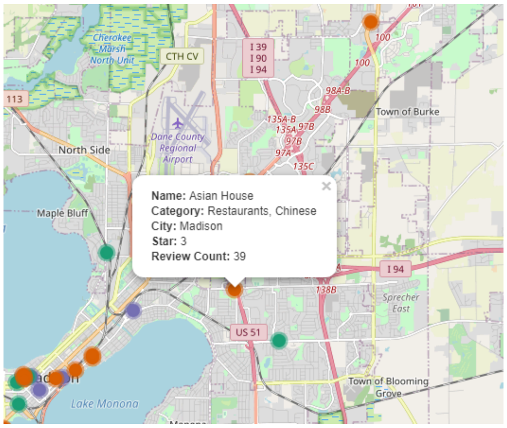


We can clearly see that the distribution mean of the rating for keyword Tea is above average while that of keyword tofu is below average. Now, we can give suggestions to the Taiwan Little Eats restaurant:

- Most people focus on chicken, tea(milk tea), rice and pork in your restaurant.
- People are satisfied with your food quality overall, and they especially like your chicken, tea(milk tea) and bubble.
- Your service is satisfactory, and please keep it on.
- Your tofu may have some problems. Please pay more attention to it.

5 Shiny App & Visualization

There are four panels in our shiny app: ”Yelp Map”, ”Customer Reviews”, ”Business Suggestions” and ”GUIDE”.
GUIDE is the guide of our shiny app, it introduce the function and the usage.
Yelp Map is an interactive map, user can click the markers on the map to get basic information of the restaurant.
Customer Reviews will show the top 12 key words that shown most in the customer reviews and make a word cloud to make it easy to know. Then we make 12 plots to show the star given from the customer when he/she mention this key word in review. It will easily show what the customers focus on this restaurant and the level.
Business Suggestions will show the corresponding score and suggestions based on Wilcoxon Rank test on each key word, and we use the p-value to generate score of the aspect of each key aspect.


Name: Asian House
Category: Restaurants, Chinese
City: Madison
Star: 3
Review Count: 39

The above plot is the reflection we get if click on a marker on Yelp Map panel. This Shiny app is based on the Yelp data. We have selected only restaurants which fits only the category of ”Chinese food”, containing about 375 restaurants. This Shiny map is interactive and uses real geographical location of the US map. We can click on a marker to get an information about a restaurant and Shiny app shows reviews and suggestions in other panels.

6 Conclusion

In our project, We first extract the information about which aspects customers care about more for all Chinese restaurants or a specific and we detect keywords as the important aspects. Then for a chosen aspect, we test this business’s performance by comparing all other similar businesses that has the same aspect based on the reviews, and give out a score to quantitative the performance based on hypothesis test. We believe those scores are good reference for business owners to make improvement, and we also give some specific suggestions based on the scores and our knowledge.

One drawback of our analysis is that, we didn’t include those businesses with few reviews since we can not give them convincing suggestions.

Appendix

Contribution

Qingchuan Ji: write the part 5 and 6, revise part 2 in the summary. Give presentation on the part 4: shiny application in the presentation. Write the interactive map, shiny application code, shiny guide, the code generating suggestions and plots parts. Test code on the Shiny.io server.

Luyang Fang: write part 2 and 3, revise part 4 in the summary; give presentation on data processing; write code to process data and decide word list; revise the code for generating suggestions; assist on Shiny part.

Xiaotian Wang is responsible for the idea, code, implement, report and presentation of the suggestion model. He also provide some idea about text preprocessing.

Ashvini Fulpagar contributed in ideas, contributed in a code for categoryselection, some part of sentimentanalysis, examined shiny app, project management, Github, Presentation on demo, main readme

References

[1] Yelp Dataset JSON Document: https://www.yelp.com/dataset/documentation/main

[2] Tidytext Guidance: https://www.tidytextmining.com/tidytext.html