STAT 628 Module 3 Report

November 18, 2019

1 Introduction

In our Yelp Data Analysis project, we focus on bars in the categories of business data and use natural language processing methods to extract information in customer reviews to provide useful suggestions to bars' owners to improve their ratings on Yelp and suggestions to customers for their decision.

2 Data Processing

We started with the business data. Since we like to study bars, we removed the missing values and picked bars and nightlife in the categories. In the meanwhile, we want to focus on cities in the U.S. so removed cities in Canada and Europe.

After cleaning the business data, we combined it with the review data and reduced the 6.6 million reviews to around 1 million.

Later on, we built some rating distribution plots for words to decide what topics and words we want to further research.

3 Information Extracting

We first define what kind of features we need to use for our analysis. We refine information based on 4 dimensions: foods, drinks, services and atmosphere. For each dimension, we choose a list of words(both common and proper nouns) that frequently exists in the bar industry. Here we extract 57 words in total.

Then we separated review into sentences, and separated sentences based on transitional conjunctions for the convenience of analysis below. We get 5.8 million separated sentences in total.

We select the separated sentences that mention 57 words above. After that, we make sentiment analysis on these filtered reviews. We define whether each separated sentence has a positive sentiment or a negative sentiment.

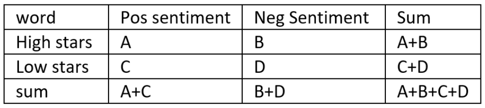
Finally, for all analyzed merchants, we count the number of reviews mention certain words, the number of reviews mention certain words with either positive or negative sentiment separately.

4 Significance Test

Since we have extracted information in terms of "adj + noun", then we checked the significance of the their influence on review stars by using T-test, chi-square test and fisher's exact test grouped by cities. By doing these tests, we can get what features that customers care about most, and what pros and cons can improve or decrease stars for a bar (business) most. For each city and each word in city list, we did four tests, details are as follow:

T-test: We use Welch's T-test, there are two groups. The first group is a list of review stars, and in these reviews the certain word is positively mentioned. Similarly, the second group is also a list contains review stars where the certain word is negatively mentioned. The null hypothesis is there is no significant differences in reviews stars no matter the certain word is positively or negatively mentioned in reviews. Then we do the test and get p-values.

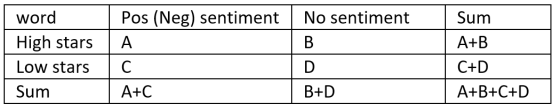
chi-square test: We use the chi-square test to test for the independence of the sentiment of a certain word and the review stars. The contingency table is shown below.



word	Pos sentiment	Neg Sentiment	Sum
High stars	A	B	A+B
Low stars	C	D	C+D
sum	A+C	B+D	A+B+C+D

"A" is the number of the certain word that is positively mentioned and its review stars is 5. "B" is the number of the certain word that is negatively mentioned and its review stars is 5. "C" is the number of the certain word that is positively mentioned and its review stars is below 3."D" is the number of the certain word that is negatively mentioned and its review stars is below 3. The null hypothesis is that the sentiment of certain word is independent with review stars. Then we do the test and get p-values.

fisher's exact test The two tests above compare positive sentiment with negative sentiment, the two tests follow compares positive (negative) sentiment with none sentiment. That is the none sentiment is the certain is not mentioned. The purpose of these two tests is to test how significant it is that a word is positively (negatively) mentioned in review can improve (decrease) reviews stars compared to reviews that the certain word is not mentioned. Two tests we use are fisher's exact test.



word	Pos (Neg) sentiment	No sentiment	Sum
High stars	A	B	A+B
Low stars	C	D	C+D
Sum	A+C	B+D	A+B+C+D

"A" is the number of the certain word that is positively (negatively) mentioned and its review stars is above 3. "B" is the number of the certain word that is not mentioned and its review stars is above 3. "C" is the number of the certain word that is positively (negatively) mentioned and its review stars is below or equal to 3."D" is the number of the certain word that is not mentioned and its review stars is below or equal to 3. The null hypothesis is that the sentiment of certain word is independent with review stars. Then we do the test and get p-values.

5 Recommendation System

Based on our test results of each city, we made a list of significant key words. The list contains four categories: foods, drinks, service and atmosphere. For some business, there are very limited number of reviews, so we use 30 as a threshold. For business with more than 30 reviews, if the positive mentions make up to more than 75% of mentions with attitude for a certain key word, then we note it as business with a positive key words, vice versa. For business with reviews fewer than 30, if the positive mentions make up to more than 50% mentions with attitude for a certain key word, then we note it as business with a positive key words, vice versa.

The recommendation system is city-specific, for the test results are not consistent among different cities.

For customers

1. Our recommendation system will tell customers what food or drink is mentioned with positive attitude.
2. Our recommendation system will warn customers from ordering food or drink with negative reviews.
3. For service part, our recommendation systems will let customers know if there are good or bad waiter/waitress, Wi-Fi, whether there is enough parking lots or not, whether the restroom is nice, if these key words are mentioned in reviews.
4. For atmosphere part, our recommendation systems will let customers know if there are good or bad music, band, environment, atmosphere, if these key words are mentioned in reviews.

For business

1. Our recommendation systems will let business know whether there is some positive or negative impression for food from reviews, and from what aspects the business can improve it.
2. Our recommendation systems will let business know whether there is some positive or negative impression for drinks from reviews, and from what aspects the business can improve it.
3. Our recommendation systems will let business know whether customers are satisfied with their service, and from what aspect can make it better, like Wi-Fi, parking lot, restroom.
4. Our recommendation systems will let business know whether customers like the atmosphere here, and from what aspect can make it better, like music, light.

6 Conclusion

We found there are some difference among cities in whether they care about the positive mention or negative mention for key words.

Positive keywords

1. For the most of the key words, if they are positively mentioned, then the probability of giving a review with 5 stars will significantly increase, for all the cities.
2. For Madison, a positive mention of sauce, price, light, environment, and atmosphere can't significantly increase the probability of giving a 5 stars review.
3. For Pittsburgh, a positive mention of sandwich, spark, negroni, brandy and booking can't significantly increase the probability of giving a 5 stars review.
4. For Charlotte, a positive mention of negroni, sandwich, cheese, chicken daiquiri, manhattan and wine can't significantly increase the probability of giving a 5 stars review.
5. For Phoenix, a positive mention of sandwich, brandy and wine can't significantly increase the probability of giving a 5 stars review.

Negative keywords

1. For all the cities, if there are some key words of servie negatively mentioned, then the probability of giving of review with lower than 2 stars will significantly increase.

2. For Madison, a negative mention of vodka will significantly increase the probability of giving a review with lower than 2 stars, which is not significant in other cities. A negative mention of burger, mozzarella stick, pizza, fried, dip, restroom and parking can't significantly increase the probability of giving a review with lower than 2 stars, which is significant in other cities.
3. For Pittsburgh, a negative mention of nugget, peanut and cocktail will significantly increase the probability of giving a review with lower than 2 stars, which is not significant in other cities. A negative mention of manner can't significantly increase the probability of give a review with lower than 2 stars, which is significant in other cities.
4. For Charlotte, a negative mention of popcorn, rum and booking will significantly increase the probability of giving a review with lower than 2 stars, which is not significant in other cities. A negative mention of wine can't significantly increase the probability of give a review with lower than 2 stars, which is significant in other cities.
5. For Phoenix, a negative mention of quesadillas, bacon, slider, oyster, spark, music, light and onion rings will significantly increase the probability of giving a review with lower than 2 stars, which is not significant in other cities.

7 Pros and Cons

Pros:

1. Our model is concise and robust since we have four tests on each word's significance.
2. Our analysis not only compares positive sentiment with negative sentiment, but we also add no sentiment into comparison.
3. Our project not only analysis data business-wise but also city-wise.
4. The shiny app is very user-friendly and beautiful.

Cons:

1. Our word list could be enlarged, then we can extract more information.
2. The adjective list of positive words and negative words can also be enlarged, since there might be some positive sentiments and negative sentiments that we have missed.
3. Our model can't predict reviews stars based on reviews.
4. Due to the limitation of time, we only analysis four cities, and we could add more cities in our project.

8 Contribution

- Yijie Liu: Complete the Tf-idf part of initial analysis, and contribute to merging reviews into businesses and significance test. For report, he contributes the part of significant test. And he also contributes to the slides.
- Jiantong Wang: Complete the recommendation system(suggestion giving), and contribute to set up the ShinyApp. For report, he contributes the part of recommendation system and conclusion.
- Xiaoxiang Hua: Complete the time-related analysis in exploratory data analysis, make sentence tokenization on all analyzed reviews and make sentiment analysis on all features that we care about.

- Yihsuan Tsai: Complete data cleaning and visualize the distribution of the word. Built the Shiny App. For report, he contributes the part of introduction and data processing. And he also contributes to the slides.