Module 3 Report

Keyu Hu, Kelly Zhao, Zhifeng Chen

1. Introduction

In this project, we aim to provide useful suggestions for BBQ restaurant owners to help them improve their business. In order to achieve this, we analyze the **business.json** file and the **review.json** file from the Yelp dataset.

2. Data Cleaning

2.1. Data Selection

The original Yelp dataset has 8.62 million reviews and 16 thousand businesses. So, we filter business and associated reviews with keywords “barbeque”. This leaves us with 247,896 reviews from 1483 businesses.

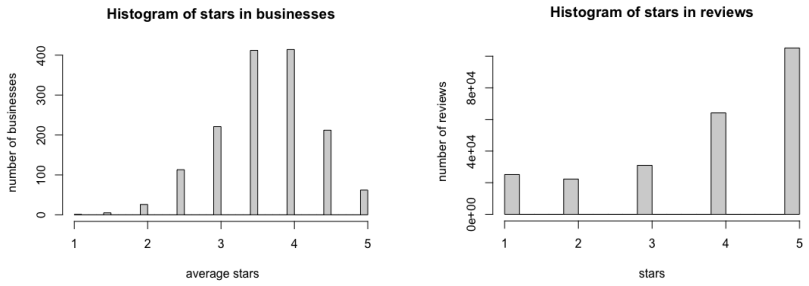
2.2. Text Cleaning

We did the following steps to clean the texts in the review dataset.

- Remove the punctuations, digits and hashtags
- Convert to lower case
- Remove the stopwords
- Tokenization (split the texts into individual words)
- Stemming (extract the base form of words by removing affixes)

2.3. Data distribution

The rating distribution in our cleaned data is shown as follows.



3. Model

3.1. Logistic Regression

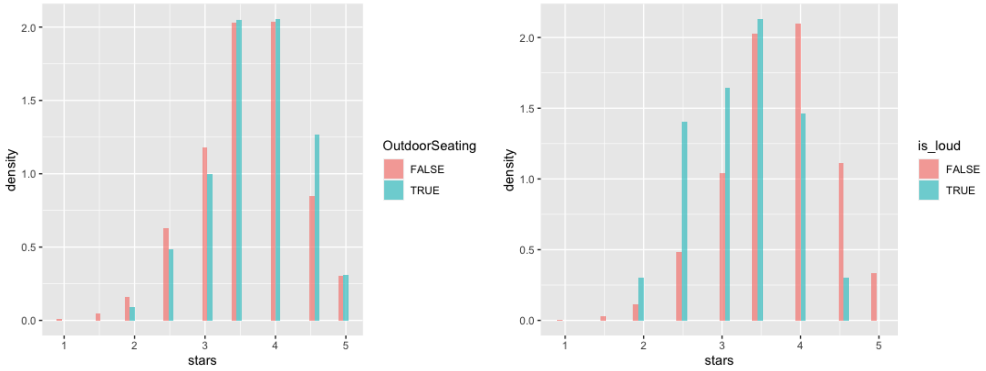
We classify businesses into two categories according to their average stars. If a business has an average rating equal or higher than 4.0, it is classified as a good restaurant(1), otherwise it is classified as a bad restaurant(0). We fit a logistic regression model by the following variables of interest and then test if these variables have significant effect on the classification.

- Outdoor seating: available or unavailable
- Free wifi: available or unavailable
- Noise level: loud or quiet
- Free parking: available or unavailable
- Ambiance: casual or not casual
- Price range: range from 1 to 4

3.2. Regression Model Result

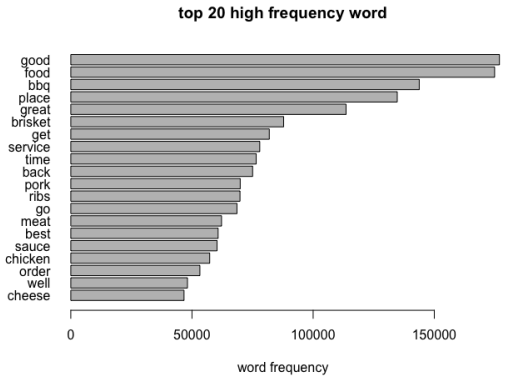
Variable	Coefficient	P-value
outdoor seating - available	0.06	0.02 *
Noise level - loud	-0.20	2.5e-5 ***
free wifi - available	-0.04	0.12
parking - available	0.01	0.74
ambience - casual	0.02	0.53
price range	-0.004	0.90

The model shows that outdoor seating and noise level has significant effects on the attraction of barbeque restaurants while others not. We make a plot to see if the two factors really make a difference in the rating distribution.



3.3. Text Mining

3.3.1. Word Frequency

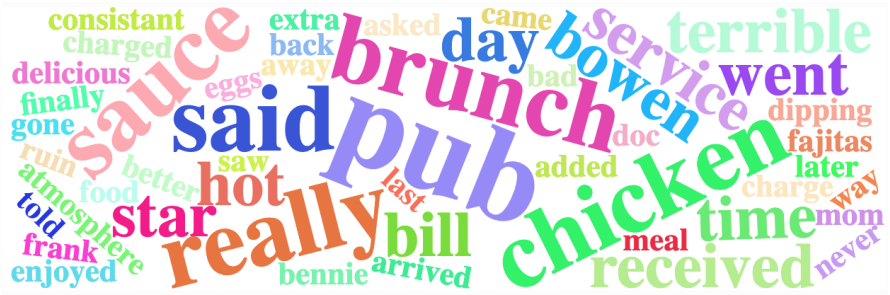


3.3.2. Word Cloud

We divide reviews into 2 parts. If a customer gives more than 3 stars, we regard the review as a positive review. Otherwise, we regard it as a negative review. We will show 2 word clouds for each business owner: Positive and Negative.

This is an example from Bowen Island Pub:

Positive review



Negative review

The word clouds above show that the beer, burger, wings are frequently mentioned in positive reviews, while the chicken, sauce and eggs are mainly mentioned in negative reviews. This implies that beer, burger, wings are attractive to customers while the taste of chicken, sauce and eggs may need to be improved. The pub and service are controversial.

3.3.3. Word2Vec

We use the Word2Vec model and see what the words are most likely around our target words. We select these target words according to the word frequency and all of the target words are nouns.

Here are 10 target words we use in our model:

word	freq
food	174918
place	134729
service	77981
pork	69908
ribs	69824
meat	62200
chicken	57336
beef	31684
staff	30635
atmosphere	16680

Word2Vec uses a neural network model to learn word associations from the corpus. There are two model architectures behind Word2Vec: continuous Bag-of-Words (CBOW) and skip-gram. We use CBOW here, because CBOW runs fast on a large scale dataset. We can easily do the job with Gensim in

Python. First, we get positive reviews with stars of 4 or 5 and do some basic preprocessing. We can build the model and see what the customers love most about the "price" of the BBQ restaurant.
There are top 10 words related to "price": **reasonable, fair, tag, reasonably, steep, higher, moderate, lower, high, decent and low.** Obviously, people really appreciate their price when they mention reasonable and fair in their reviews. We also do the same for other target words.

3.3.4. Generating specific Suggestions

Now we know why people love the above aspects about barbeque restaurants, we move forward to generate individual suggestions for each barbeque restaurant. As we did for the word cloud, we divide the reviews into two categories: positive and negative. Then, we build a function that takes in a Business ID and returns one suggestion for each target word. This function first collects all the reviews with the specified Business ID. Then, for each target word, the function looks into the review texts and selects the ones that contain the target word. Next, the function calculates the positive vs negative proportion among the selected reviews. If this proportion is lower than the overall positive vs negative proportion of all the reviews of the specified Business, we would suggest that the target aspect needs to be improved.

4. Shiny APP

Please check our shiny App: https://zhaozihan.shinyapps.io/yelp_analysis/

5. Conclusion

We have analyzed some business attributes' effect on average rating in yelp and concluded that outdoor seating and noise level are two important factors for barbeque restaurants. Barbeque business owners should offer outdoor seating to customers and reduce noise in the restaurants to boost their attraction to customers.
Also, we build an algorithm to automatically generate specific suggestions for each business according to their customer reviews.
However, a drawback of our algorithm is that, when a customer mentions a specific aspect of a business, we judge his sentiment towards this aspect just based on the stars he gives. This could be inaccurate in some cases. Maybe it would be more accurate to analyze the sentiment towards an aspect by some NLP method, but our approach is simple, easy to understand and computationally economical.

6. Contribution

Keyu Hu: text cleaning, word frequency, report and slides
Kelly Zhao: data cleaning, build model, shiny app, report and slides
Zhifeng Chen: regression model, report and slides