Stat 628 Report

Chenhao Fang, Yuechuan Chen, Mengqi Li

November 2020

1 Introduction

Yelp is a widely used app in the United States that provide a platform for people to write reviews of businesses. The data contains the information of different businesses and customers reviews of these businesses. Our analysis focus on restaurants served desserts and our goal is to provide some useful insights and suggestions for businesses owner to help them improve their service.

2 Data Cleaning

2.1 Data preprocessing

There are 36,327 businesses and 942,027 reviews in the original dataset. We focus on restaurants serve desserts which include 619 businesses and 36,118 reviews.
We first create a new dataset by merging the business and review file and select the variables that we want: 'business_id', 'business_name', 'business_stars', 'business_attributes', 'business_categories', 'review_text', 'review_stars'.

2.2 Data cleaning on 'review_text'

- Change all the text data to the lower case.
- Remove all punctuation characters in the text data.
- Tokenize using regular expression.
- Stopwords use built-in dictionary in nltk.corpus.stopwords.
- Use nltk.stem.wordnet lemmalizer to lemmatize the words.

3 Word Cloud

Word Cloud is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency. It works in a simple way: the more a specific word appears in a text, the bigger and bolder it appears in the word cloud. By analyzing the Word Clouds which can turn large text into a pile of information, we can understand how customers feel about the restaurants.
The Figure 1 shows the Word Cloud of all the restaurants that has a rating greater than

4 and figure 2 shows that for restaurants with rating smaller than 4. By checking these two figures, we can see that the restaurants with higher ratings comes with more positive words like "great", "good", "amazing" and so on, while the one with lower ratings comes with words like "even", "bad", "never", "dont" and "didnt". They do make a difference and therefore, in our shiny app, we will draw Word Cloud for every restaurants so that they can have a rough idea about how customers feel about their shops.

Figure 1: Bakeries with high ratings.

Figure 2: Bakeries with low ratings.

4 Naive Bayes Classification

We use Naive Bayes classifier to build the model and get the most informative words.

4.1 Introduction to Naive Bayes Classification

Naive Bayes classification is nothing but applying Bayes rules for forming classification probabilities.

- Divide reviews into two categories:
  - We use the variable $star\_x$ to classify the value of 4 or 5 as positive evaluation, and the value equal to 1, 2, 3 as negative evaluation.
- Bayes rule:

  For a document d and class c:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

  In this case, the class comprises two parts. Positive and negative.
  - The LHS term $P(c|d)$ is the probability of class c given a document d. This term is also known as Posterior.
  - $P(d|c)$ should be similar.
  - The term $P(c)$ which is referred as Prior is the original belief i.e., original label of the document being positive or negative.

- $P(d)$ is referred as the normalization constant, This term is divided with the result produced by the multiplication to ensure the outcome can be presented in a probability distribution.

- The use of Bayes' rule in this problem:

  Take every words appeared in a specific review as $x_1, x_2, x_3, \cdots$ and under some specific assumptions, we have:

  $$P(d|c) = P(x_1, x_2, x_3, \cdots, x_n|c) = P(x_1|c)P(x_2|c)\cdots P(x_n|c)$$

  We counted the number of times each word appears in each review, and then use the Bayes rule on the whole reviews of a store. Finally, we can get the probability of each word's appearance given positive or negative reviews.

- The score of each word:

  After we have the probabilities of each word's appearance given positive or negative reviews, we can divide the two probabilities and get a ratio. we used a formula to turn the ratio to a score ranging from 0 to 10. And if the score of a word is greater than 5, we assume it as a positive word, vice versa. (After this, we can get tables like Table 1,2,3 shown in page 4)

4.2 Advantages

- The model is easy to understand and doesn't have many parameters to tune.
- We can analyze each word and give a specific score.

4.3 Disadvantages

- Naive Bayes classifier has conditional independence assumption which is hard to satisfy. However, since we aim to give out suggestions, the minor difference in the final scores may not matter too much.
- If a restaurant has a small number of reviews, Naive Bayes' judgment on the words that appear less frequently will become inaccurate.

5 Recommendation

We can get the results based on Naive Bayes model using all dessert business data. Then we pick up some representative words and give advice to dessert businesses.

Table 1: Service

Words	Characteristic
slowest	negative
rudeness	negative
unprofessional	negative
handmade	positive
diversity	positive
customizable	positive

As for service, we can see from table 1 that dessert restaurants which treat customers rudely, make food unprofessional and serve food slowly are more likely to have lower Yelp rating. So you should ask your waiters to be kind and patient to the customers and serve food fast. Meanwhile, you should train your chefs well so that they can make desserts professionally. In addition, the dessert restaurants which provide handmade, customize and diversity of desserts tend to get more positive reviews.

Table 2: Environment

Words	Characteristic
loudly	negative
cigarette	negative
antique	positive

As for environment, people generally don't like noisy places so you should keep your restaurant quiet. And people also don't like places filled with smoke, so you should not allow your customers to smoke in the restaurant. Plus, placing antiques appropriately could be a good idea to attract customers.

Table 3: Food

Words	Characteristic
doughnut	positive
mochi	positive
croissant	positive
focaccia	positive
gingerbread	positive
raisin	positive
fusilli	negative
cookiewich	negative
coriander	negative

As for food, your restaurants had better include desserts like doughnut, mochi as basic. To enrich your menu, you could also consider croissant, focaccia, gingerbread and raisin flavor desserts. And you should be caution if you serve fusili, cookiewich and coriander in your restaurants.

6 Conclusion

To help dessert business owner improve their Yelp rating, we analyze customers reviews on Yelp and give suggestions to business owner. We first clean the data using nltk library, then use Naive Bayes classifier to build the model and get the most informative words. Finally, we give advice to dessert business owner to improve their service and get higher Yelp rating.

7 Contributions

- Mengqi Li: data cleaning, word clouds code, write introduction, data cleaning, recommendation and conclusion for report.
- Yuechuan Chen: build Naive Bayes model, write Naive Bayes classification for report.
- Chenhao Fang: word clouds code, data visualization, build, debug and deploy the shiny app, write word cloud part of report, final editing for report, write all Github README.md files, editing the presentation video.

8 Reference

- https://www.nltk.org/_modules/ nltk/ classify/ naivebayes.html
- https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-pytho