Analysis on Pennsylvania Restaurant

Chaoran Wang*, Lu Li*, Qiaoyu Wang*, and Yuhan Meng*

*Department of Statistics

Abstract

In this project, we have data close to 8GB which include review.json, tip.json, user.json and business.json. In order to focus more on the details and information of data, we restrict our research on the food and restuarants in Pennsylvania in term of the information about review and business. Initially, we set explicit goal to lead our project. Three main goals are:
1.Establish the top ten rankings in terms of 4 aspects including atmosphere evaluation, food quality, service quality and price level.
2.Provide important fetures to customers including basic information and the evaluation from other users for their consideration.
3.Make insightful suggestions to business owners about their business hours, time period and attributes.
Firstly, we filter out food data in Pennsylvania and use statistical test to select significant information. Most of time in preparation process is spent on processing text data such as deleting stopwords and restoring abbreviations. Secondly, we take the weight of "useful" into consideration to better balance the proportion of each review in whole text. Then we apply LDA topic model using the word frequency in all reviews of single one business owner to get the topics of each business owner. Finally, we combine statistical methods such as random forest, linear regression and LOESS to analyse the significant information in first step and give advice for business owners.

1 Motivation

As living standards rise, more and more people choose to eat outside. Then problems will come to us that which restaurant should we choose and how can a restaurant improve its star level. Based on the yelp dataset,our group decided to do some analysis. There are four files in yelp dataset. Due to the time restriction, we mainly focuse on the business file and review file. Business file cotains basic information about business such as id, address, category,star level,opening hour,closing hour. For review file, it includes customers opinion based on their experience like how they feel about the service, how they like the food, how they think about the price and also publish the star level. Since there are so many businesses in the yelp dataset, we focused on the restaurants in Pennsylvania. What's more, the population in PA is the sixth in the US. According to our analysis, we hope that we can establish the top ten rankings in terms of 4 aspects including atmosphere evaluation, food quality, service quality and price level. What's more ,providing important fetures to customers including basic information and the evaluation from other users for their consideration is also one of our goals. Finaly, we make efforts on making insightful suggestions to business owners about their business hours, time period and attributes.

2 Data Process

As we mention before, we mainly focuse on the business file and users file and make some analysis based on them. Since these two files contain too much information and both of them are also really messy, we need to do some data processing to extrate the data we need. In this part, we briefly talk

about the way we use to clean the data and we divide the data processing into two parts as we show in the following.

2.1 Business data

There are 192608 business in the business.json and each business is classified to servel categories. For instance,'Emerald Chinese Restaurant' has categories like Specialty Food, Restaurants, Dim Sum, Imported Food, Food, Chinese, Ethnic Food, Seafood. The reason why we choose business in food and restaurant categories is that the number is in these two are the 1st and 3rd among all the categories. And we choose the food and restaurant in Pennsylvania because the population in Pennsylvania is the 6th in the US and the number of restaurant is suitable enought for us to analysis.The process we filter the data is show below.
∗ Choose businesses whose category include food and restaurant.
∗ There are some businesses are not restaurant but include food such as grocery and gas station. We delete the businesses are not related to the restaurant.
∗ Filter the restaurant in Pennsylvania.
∗ We also try to delete the business with only one or two business. However the minimum number of reviews for business is 3 and about 20% of the businesses have only reviews. Hence we decide not to delete any business.

2.2 Review data

Most of data in review.json are stored in the review text which are actually difficult to deal with and have many relevent papers concerning the methods to process natural language. We have to transform the text in some steps to make computer language more easy to analyse them.
∗Transform word into lower form and expand contractions.
∗Break paragraphs to sentence and then remove stopwords.
∗Add NOT mark after negative words and finally remove punctuations.

3 Exploratory Analysis

Based on our final goals and existing data, we plan to research on the relationship between data and business stars and select significant ones for our further study. We focus on information in business.json and review.json such as various attributes location and categories. This step is aimed at get an initial understanding of data.

Feature	Source	Methods	Result	Significant
Working Time	Business.json	Ordered Logistic Regression	P-value <= 1.38×10^{-13}	✓
Location	Business.json	Mann-Kendall Trend Test	P-value = 0.122	×
$\frac{Number\ of\ positive\ words}{Number\ of\ negative\ words}$	Review.json	Scatter plot	Positively connected	✓
Type of food	Review.json	Histogram	Distributed diversely	✓

4 Model estabishment

After the basic exploration about the data, we find different type of food like Chinese,Italian and different adjectives have different distribution over star level. Besides, $\frac{number\ of\ positive\ words}{number\ of\ negative\ words}$ for each star level also leads to quite different result. Working hour is significant for star levels while location shows no influence on star level. What's more, we also curious about attributes term in business file since it contains a lot of specific information about business. We decide to provide advice and establish our models based on these aspects.

4.1 Topic classification

There are lots of interesting information in review document including 'useful', 'funny' and 'cool'. 'useful' especially takes a quite large proportion of these words which comment on single review. Hence, we try to take the influence of 'useful' into consideration to balance the weight of each review. We calculate $\frac{Number\ of\ review}{Number\ of\ useful}$=0.863 as $\frac{Willingness\ of\ making\ review}{Willingness\ of\ making\ 'useful'}$ and we can use this conversion rate to adjust the frequency of each word.

$$New\ frequency = (1 + number\ of\ useful \times Conversion\ Rate)\ \times\ Previous\ frequency$$

Review	Previous fre	Useful	Rate	New frequency
The pizza is so good. I love this salad whose flavor is good.	pizza:1; salad:1; good:2	0	0.863	pizza:1; salad:1; good:2
The great food here is my favourite and the wings is so good.	food:1; wing:1; good:2; great:1; favourite:1	5	0.863	food:5.315; wing:5.315; good:10.63; great:5.315; favourite:5.315

The categories shown in business.json is complex and hard to distinguish easily, leading us to assign a new topic to each business owner based on their all reviews. Based on the words and their frequency in each business owner which we extracted from reviews, we apply Latent Dirichlet allocation algorithm on the frequency matrix with dimensions (4524×30000) which is the number of business owners and the reduced number of words. According the result from LDA algorithm, we name 5 topics as Brunch, Bar, Dessert, Fast food and Foreign flavor.

Topic order	LDA result	Topic we name it
1	sanwich get ti go fri coffe friend no breakfast tri	Brunch
2	bar ti drink beer servic menu us restaur tabl realli	Bar
3	cream ice chocol flavor tri cake coffee get also love	Dessert
4	pizza sauc ti salad restaur get chees go tri servic	Fast food
5	chicken restaur taco ti fri dish get also go sauc	Foreign flavor

4.2 Ratings for different topics

Based on our experience we choose four aspects to rate restaurants including atmosphere evaluation, food quality, service quality and price level. For different topic we would like to find the top ten restaurant in four aspect separatly. The way to rate each restaurant is shown below.

∗ Order the frequency of words appears in reviews for each topic.
∗ Find nouns and adjectives from the first 500 most frequency words.
∗ For nouns, we classify them into 4 aspects. For atmosphere evaluation,service quality and price level, the keyswords are the same for different topics, while for food quality, the keyswords are different. For instance, even though pizza,hamburger,cake,ice-cream all belong to food we choose pizza,hamburger as keywords for fast food topic and we select cake and ice-cream as keywords for desert topic.
∗ We cut the sentences in reviews according to conjunctions and then count the frequency of each adjective appears before and after the nouns in two words.
∗ Sum up the the frequency of different adjective for each aspect for each restaurant .
∗ Construct linear model to compute the weight for adjectives.
Y:the star level for each business, X: frequncy matrix only about adjective words.

Y	NOTbad	NOTgood	...	awesome
3.5	19	11	...	55
4.5	0	0	...	3

∗ Keep the adjectives with absolute coefficent larger than 0.01. ∗ Compute the score for each business in four aspects separately and use the min-max normalization. ∗ Find the top ten restaurants in each aspect in each topic. Here are the top five food quality in Bar topic:

names	stars	food_scores
NU Jewish Bistro	4.00	1.00
Jamison's On West Liberty	4.5	0.99
Aurochs Brewing	5.0	0.98
Arsenal Cider House & Wine Cellar	5.0	0.94
Full Pint Brewing Company	4	0.94

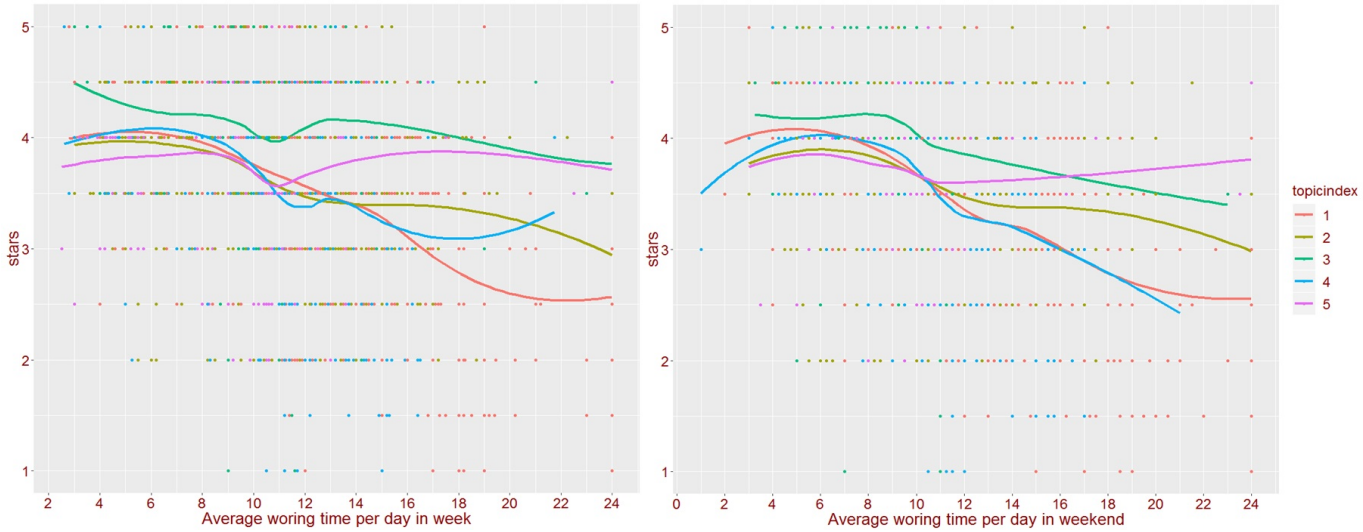
4.3 Analysis of working hour for different topics

Since we have proved the connection between working hours and their stars, we are aimed at materializing this connection and providing suggestions for business owner. Firstly, in view of experience that there is a difference in the working hour between weekdays and weekends. We do the two sample t-test on the working hour between weekdays and weekends, the P-value <= 2.2×10^{-16} which means there is a significant difference in them. Hence, we decide to reserch on the link between working hour and stars based on the topic and period(weekdays or weekends). Our idea is to use different methods to give advice in the aspect of working time and working period respectively. As for working time, we calculate the working time of each business owner and utilize LOESS(locally estimated scatterplot smoothing) to find the best working hours in each topics.

Topic order	Topic name	Working time in Weekdays	Working time in Weekends
1	Brunch	4-7 hours	4-7 hours
2	Bar	4-7 hours	5-8 hours
3	Dessert	3-6 hours	6-9 hours
4	Fast food	5-8 hours	5-8 hours
5	Foreign flavor	6-9 hours	5-8 hours

When it comes to the best working period, we seperate their working period in the unit of one hour and select working period through two methods which are random forest and distribution analysis. Given the stars and their working period of business owner, we use random forest by putting stars as feature which is the criterion of classification to find significant period and make a linear regression between stars and significant time period to get the beneficial period with positive coefficient. What'more, through the distribution of four-star and five-star working period, we conclude the golden time which

means the common period chosen by 90% of high-star business owners for their operation. We conbine the result of two analysis to give advice for working time.



Topic order	Topic name	Working period in Weekdays/o'clock	Working period in Weekends/o'clock
1	Brunch	6:00 - 7:00,12:00 - 16:00	12:00 - 15:00
2	Bar	16:00 - 17:00 ,18:00 - 21:00	16:00 - 17:00 ,18:00 - 21:00
3	Dessert	9:00 - 10:00,13:00 - 17:00,18:00 - 19:00	9:00 - 11:00,13:00 - 16:00,20:00 - 21:00
4	Fast food	9:00 - 11:00,18:00 - 20:00	18:00 - 20:00
5	Foreign flavor	10:00 - 15:00,18:00 - 21:00	10:00 - 11:00,17:00 - 21:00,24:00 - 25:00

4.4 Significant attributes for different topics

As attributes in business data set provide many specific descriptions about different businesses, we would like to explore whether some attributes significantly influence the star rating and are key to businesses whthin each topic.
We firstly find out all the existed attributes in business data set and then do some reorganization like defining GoodForMeal:dessert as a new attribute named as GoodForMeal_dessert. After this, we have 65 distinct attributes. For each business, if it does not have specific information of some attributes, we record these attibutes as Na. Then whthin each topic, we treat the attributes as predictors and treat the star rating as the response variable. Then get the important score for all the attributes using GUIDE. Finally whthin each topic, we select out different important attirbutes according to the important score. To futher explore the relationship between the star rating and important attirbutes, we use boxplot, linear regression to decect meaningful relationships. However considering there exites many missing value in some attributes, we use the following three criterions to do the data analysis and make suggestions.
∗If the missing value is the main part of all the data related to one attirbute, we think it's unreasonable to do the data imputation or make inference based on limited data without missing value. So we will not futher study this attribute.
∗If the missing value takes about one third or less part of all the data related to one attirbute, we will draw the boxplot, and make some inference by comparing the distribution of star rating under different levels of this attirbute. Some suggestions may like that It's better for bar type resturant to be no smoking, as the star rating for no smoking is significantly larger than the star rating under permiting smoking or smoking outdoor.
∗If the missing value is very rare for one attribute, we will exclude these missing value and use the remaining data to fit the linear regression model between the star rating and this attribute and also draw

the corresponding boxplot. Then for the significant attributes, we will interpret the coefficient as how much the star rating will change. Based on the coefficients and boxplot, we will make some suggestions for the businesses whthin each topic. For example, we may suggest that it's better to have a hipster ambience in your brunch resturant as it will increase your star rating about 0.68.
Finally, according to the values of attributesin one business , we will make some suggestions according

5 Conclusion

5.1 Summary and discussion

To sum up, we concentrate our research on provide basic information for customers and give advice to business owners for their improvement not only in star but also in other specific aspects.As for information for customers, given location and their star that we can easily get from business owner, we analyze reviews to group them by LDA algorithm and intentionally evaluate their atmosphere, food quality, service and price which are essential for a customer to choose the restaurant. As for advice to business owners, in addition to giving advice on the 4 aspects above, we also utilize various statistical methods such as random forest, LOESS and linear regression to provide suggestions on working hours and various attributes. Our model has good perfomance and will be of profound meaning for the improvement of business owners and choice of customers.

shiny app link :https://chaoranwang.shinyapps.io/shiny/

5.2 Weakness and strength

Strength: we combine statical method with reality and establish robust models.
Weakness: lack of comparisons of other models and may need other kind of data to support our opinion

6 Reference

- https://github.com/rasbt/python-machine-learning-book-2nd-edition/blob/master/code/ch08/ch08.
- https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72
- https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32
- http://snowball.tartarus.org/algorithms/english/diffs.txt
- www.stat.wisc.edu/ loh/guide.html

7 Contribution

Chaoran Wang(cwang647@wisc.edu): Clean and tokenize "review" data; analyse working time by LOESS and random forest and linear regression; establish shiny app; make slides.
Lu Li(lli468@wisc.edu): Make word clouds; LDA ; use guide to select out important attributes and make corresponding suggestions; write part of the jupyter notebook; make slides.
Qiaoyu Wang(qwang382@wisc.edu): Analyse location; compute the weight for each reviews; work with nouns and adjectives in all reviews; write part of the jupyter notebook; make slides.
Yuhan Meng(meng46@wisc.edu): Clean "business" data; analyse special works; rank the top 10 restaurants; write part of the jupyter notebook and markdown for the Github; make slides.