Business Analysis of Bars on Yelp

STAT 628 Module 3 Group 2

1. Introduction and Background

In this project, our goal is to use the data provided by the Yelp, which includes some reviews, tips, business and user information, to do some statistical analysis. Then we need to use a simple and understandable way to provide these business owners with some suggestions to help them improve their performance.

2. Data Preprocessing

2.1 Data Structure

Due to the plenty of dataset on Yelp, we decide to use all the bars business under the restaurant category as our whole dataset. Then we roughly clean the data by category and table. As we can see from table 1, after cleaning the data, there is still enough data for our analysis.

File	Original	Cleaned up Bar Business
Business	192,609	8,155
Review	6,685,900	946,522
User	1,637,138	453,004
Tip	1,223,094	170,127

Table 1 Cleaned Up Dataset

2.2 Text Cleaning

For some text data, we need to clean them in order to simplify the following process of constructing model. We do the following steps to clean the reviews and tips.

Table 2 Workflow for Text Cleaning

And we can see how one sentance changes from the original to the final as table 3 shows.

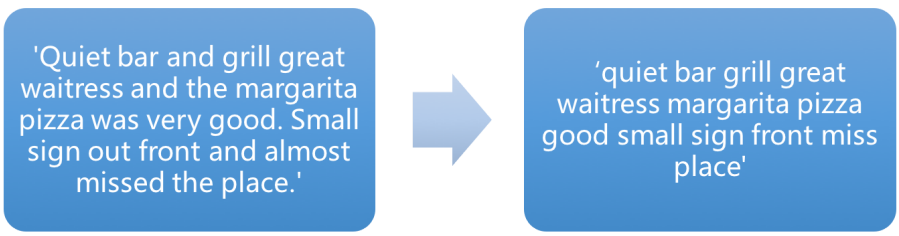


Table 3 Sample for Text Cleaning

3. Experiments and Algorithms

3.1 Feature Extraction——TF-IDF

We use tf-idf to do feature extraction because compared with naïve word frequency, tf-idf emphasizes those words with high frequency but not too high and this will help us exclude some words frequently appear but are not very informative such as 'she', 'he', 'the' etc..

Table 4 shows the result of tf-idf of all bars reviews, and we rank these words by weighting their importance in the whole document. We divide them into 4 groups: service, place, food, feeling.

Service		Place		Food			Feeling	
Time	Staff	Atmosphere	Spot	Menu	Price	Drinks	Positive	Negative
Wait	Waitress	Music	Outside	Chicken	Cheap	Beer	Great	Terrible
Table	Friendly	Quiet	Inside	Burger	Not Cheap	Craft Beer	Delicious	Bad
Hour	Attentive	Live	Town	Cheese	Priced	Blood Mary	Nice	Awful
Slow	Rude	Ambience	Street	Salad	Expensive	Old Fashion	Good	Disappointed
Busy	Helpful	Not Clean	Parking	Fries	Affordable	Cocktails	Love	Crowded

Table 4 Frequency Word for TF-IDF

Based on these words, we can give these business owners some rough suggestions. For instance, for menu part, bar owners could provide more craft beer since most customers are satisfied with that. Also, for place and service they should keep the bar clean and add more tables to shorten the waiting time, and train server to be more polite.

3.2 Generalized Linear Regression——Business Attributes

Based on the business data, we find that there is one column called attributes, and it has many types, so we do a generalized linear regression on it to decide which attributes will do more contribution to star rating. And in this way, we can provide some suggestions for the business owners to add, cancel or keep their attributes for higher rating.

First, we find that lots of business attributes have missing values (31 attributes). For attributes which have over 50% missing values rate, we directly delete those attributes. As for other attributes, we use 'missForest' package in R to implement nonparametric missing value Imputation based on Random Forest algorithm. Then we fit the model.

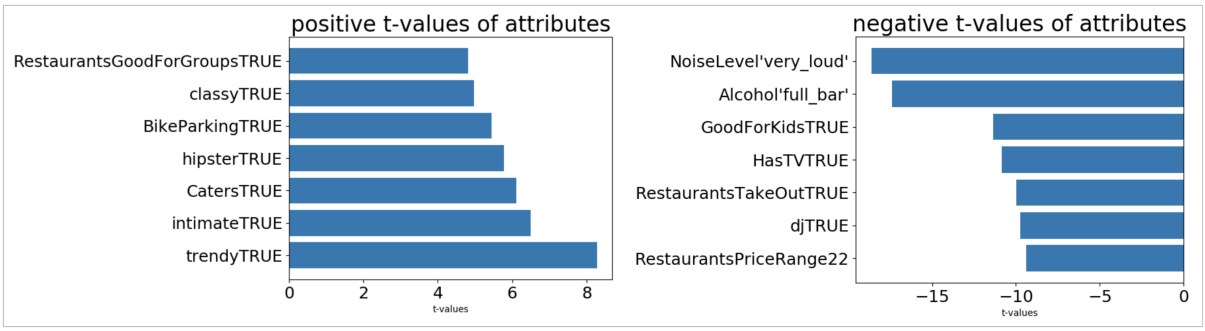


Figure 1 Business Attributes Rank

Larger the absolute value of t-value, more importance the attribute is. In this way, we can conclude from the figure 1, "NoiseLevel_very loud", Alcohol'full_bar', will lead to negative ratings most. We suppose that this is because too noisy will make customers annoyed. And Alcohol'full_bar' means this bar does not have its own

specialty, then it can not be attractive enough.

So we suggest that the business owners can take some actions to make the bar less noisy, and make his bar more special, such as adding his own specialty.

3.3 Topic Modelling——Non-negative Matrix Factorization

There are 11 states in total and in the report we take Ontario in Canada as an example to illustrate the model.

3.3.1 Model Construction

Topic modelling is an unsupervised text mining approach. And table 5 shows how topic modelling works. Our task is to provide suggestions for business owners using user reviews. To achieve this, the first step is to extract 'topic' from review corpus that represents different aspects of a restaurant such as service, flavor of food and atmosphere. NMF is a great option for finding latent topics.

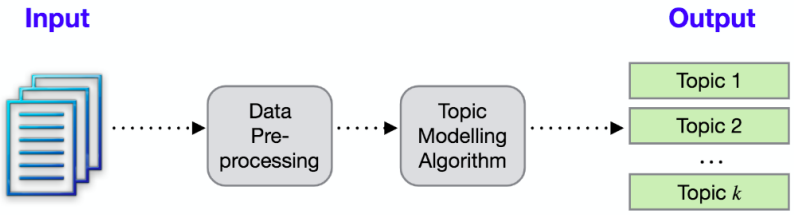


Table 5 Workflow of Topic Modelling

3.3.2 Model Evaluation

We divide these reviews into two parts, positive and negative and then do topic modelling on each of them. First, we need to decide the best topic number. We can conclude from table 6 topic coherence that k=3 is the best parameter for number of topic in positive reviews and k=5 in negative reviews. Then we construct the full model based on k.

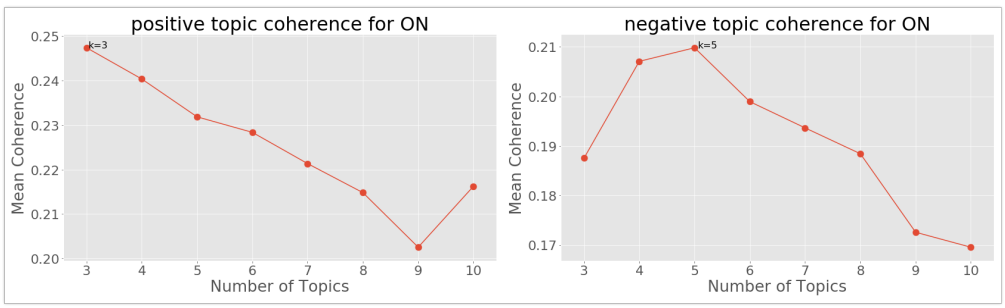


Table 6 Topic Coherence in ON

By using non-negative matrix factorization with best topic number, we construct two models for positive and negative reviews, and table 7 and table 8 show the result.

For positive one, we should suggest the bars to maintain it, such as maintaining their food to be tasty, and making their staff still friendly.

Positive Topic	Words
Environment	drink, bar, night, beer, table, look, menu
Service	time, service, amaze, experience, staff, friendly, recommend
Food	dish, fry, sauce, chicken, taste, menu, delicious

Table 7 Positive Topic in ON

For negative topics, business owners should make some improvements, such as adding tables to decrease waiting time, providing special sauce for wings and offering more featured drinks which will make it more special and attractive. Also, they should train the bartender, waitress and manager to provide better service.

Negative Topic	Words
Table Availability	table, wait, minutes, seat, sit, host, ask
Service Time	time, service, wait, experience, staff, slow, long
Food Flavor	fry, taste, chicken, dish, menu, wing, sauce
Bartender Proficiency	drink, bar, beer, night, look, friends, bartender
Service Quality	ask, server, tell, bill, manager, leave, waitress

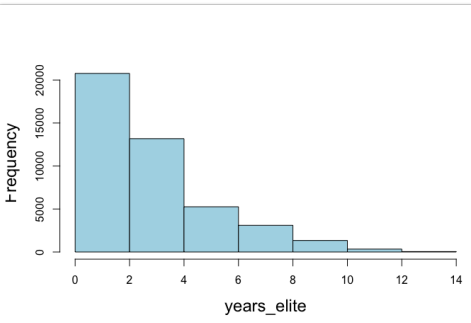
Table 8 Negative Topic in ON

3.4 Score Model

3.4.1 User Efficiency

On Yelp's official website, they define Elite Squad to be a way of recognizing people who are active in the Yelp community and role models on and off the Yelp site. Elite-worthiness is based on a number of things, including well-written reviews, high quality photos, a detailed personal profile, and a history of playing well with others. So we have enough reason to use the feature Elite to judge user's efficiency and then give each of their reviews a weight to be implemented in the following model construction.

First we draw a histogram to show the distribution of elite as figure 2 (left) shows below.



Elite Years	User Weight
0	1
1	1.1
2	1.2
3	1.3
…	…
11	2.1
12	2.2

Figure 2 Elite (left) and Weight (right) Distribution of User

Then, we give them weight based on years of Elite as figure 2 (right) shows. We calculate user weight according to the following formula.

$$Weight = 1 + 0.1 * Elite$$

3.4.2 Business Score Model

This score model is based on NMF (non-negative matrix factorization) and user weight. Table 9 shows how it works. NMF will extract topics from each state’s negative reviews and decompose each topic further into words and corresponding weight. Also, we assign different weight for users, increasing with their years of being elite. Then we can combine the weight of each topic generated by each review and the weight of user of that review to calculate the score for the business.

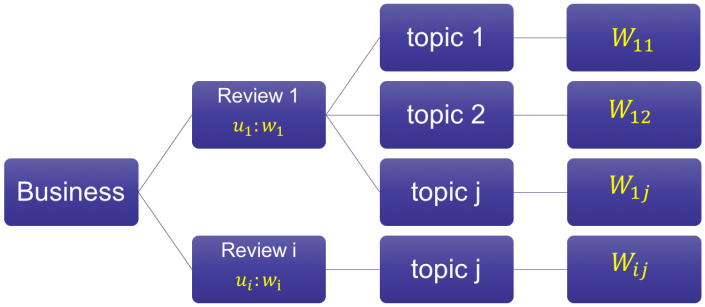


Table 9 Workflow of Score Modelling

The following formula shows how we calculate the score for each company.

$$Business\ topic_1\ score = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} user_i * weight_i^j}{\sum_{i=1}^{n} user_i}$$

In this report, we take one business called Sparrow in QC as an example to illustrate the score model in detail.

Since the topics are extracted from negative reviews, with the 'score' of each topic goes higher, the performance of business at that aspect tends to worse. The absolute value of score doesn't provide much information, so we convert the scores of each business into their rank in the state to see how good it performs in its competitors. The result is shown in the table 10 and table 11.

	Words	Category
Topic 1	table, wait, minutes, ask, time, service, minutes, waitress	Place
Topic 2	drink, bar, night, beer, bartender, pay, leave	Service
Topic 3	fry, taste, chicken, dish, menu, service, burger	Menu

Table 10 Topic Categories in QC

Business Name	State	Topic 1 Place Rank	Topic 2 Service Rank	Topic 3 Menu Rank
Sparrow	QC	66%	55%	**71%**

Table 11 Score Model of Sparrow

The suggestions are provided according to the rank of each business and the corresponding words in that topic. For example, for topic 3, we noticed that word 'fry' has the largest weight. It is very common for bars to get negative review if their fryer oil is not fresh enough. So, for all the business that has low rank in topic three, we'll provide the suggestion: change the fryer oil frequently.

4. Conclusions

Based on the result of all the models and analysis, we deivide our suggestions into two parts. One is general suggestions based on the analysis of linear regression of attributes, the other is specific suggestions for each business based on the topic model and score model. And specific advice are from three perspectives: food, service and environment. Business owners could choose the suitable advice for themselves based on their rank of each topic.

4.1 General Suggestions

Make the target customer more clear and pricing more pointed.
Add booths to provide more private room for customers.
Increase the diversity of music and make the atmosphere to be more attractive.

4.2 Specific Suggestions

4.2.1 Food

1.The fryer oil needed change regularly.
2.Add more special sauce for wings to improve its taste.
3.Provide more craft beer and add own specialty, such as special drinks.
4.Control the quality if cooked food such as steak and burger.

4.2.2 Service

1.Service Etiquette
Improving customer service: such as training the waitress and manager to be politer and serve more carefully. And pay more attention to customers' feedback.
2.Service and Order Time
Hiring more staff to speed up service and offering some gifts or discounts for customers who wait too long.
3.Bartender Proficiency
Hiring more professional bartenders.

4.2.3 Environment

1.Table Availability
Adding tables to decrease waiting time and increasing table turnover rate and providing some entertainment and seats for waiting customers.
2.Shows: adding more shows in club at night.
3.Music: controlling the music volume.

5. Shiny

Shiny App Link (https://ylingbfcalculator.shinyapps.io/Yelp_data_analysis/).

6. Strength and Weakness

6.1 Strength

1.We combine different models to provide more comprehensive advice from several perspectives, and this will help these business owners to get customized advice for their own business.

2.We use topic modelling that devides positive and negative reviews into different topics by importance weighting and relevance between them, and it can help us to give more detailed and actionable suggestions for business owners.

3.Score model combined with the user information will help us to rank these business by states more accurately, and business owners can have a general idea of their performance compared with other bars in the same state which will help them to be more clear about their goal and motivate them to make changes for their problems.

6.2 Weakness

1.Due to the lack of negative reviews in VA and WA, we can not construct topic model for these two states.

2.For bars have little amount of reviews (less than 10), our score model is not that accurate because the information contained in the reviews is not enough.

7. Contribution

Name	Contribution
Naiqing Cai	Responsible for data preprocessing, presentaion slides and summary report.
Jitian Zhao	Responsible for feature extraction, model construction and statistical analysis.
Zihao Li	Responsible for data preprocessing, feature extraction and statistical analysis.
Yaobin Lin	Responsible for feature extraction, statistical analysis and shiny app.

8. Reference

[1]. Translation in Text Cleaning: (https://textblob.readthedocs.io/en/dev/)
[2]. Text Cleaning: (https://www.nltk.org/_modules/nltk/stem/snowball.html)
[3]. Topic Model Tutorial (https://github.com/derekgreene/topic-model-tutorial)