STAT628 Module 3 Executive Summary

Xiao Li, Mingyu Wang, Chufan Zhou

Dec 2022

1 Introduction

From the recommended data from Yelp, we want to get some useful and analytic insights. What we focus on is the cinema business and giving data-driven action plan to the owners to help them improve their ratings on Yelp in the future.

2 Data preprocessing

2.1 Data filtering

To obtain the exact data, we first find those business related to cinema, what we do is getting those business by filtering out all rows whose categories contains "cinema" in "business.json". We extract 359 cinemas and get their business ID's, use them to find the corresponding reviews in "review.json". Each review, a piece of text data, belongs to one customer. Each business has many reviews. Finally, we get 360 cinemas and 15716 reviews for subsequent analysis.

2.2 Text cleaning

For each review, we use tokenizers::tokenize_words[1] to divide sentences according to spaces, and make the word segmentation results are lowercase without numbers and punctuations. The tokenize_words function uses the stringi package and C++ under the hood, which makes it very fast to save time.
Then we exclude some frequent words that don't have statistical significance, which named stopwords. We choose stopwords from two sources, one is using the default English stopwords by stopwords::stopwords("en"), the other one is find some frequent English stopwords from the website[2].
Then we consider stemming and lemmatisation. The former is the process of removing the prefix and suffix of the word to get the root word, here we use tm::stemDocument. The latter is based on dictionaries, transforming the complex form of words into the most basic form, here we use textstem::lemmatize_words. These two steps can facilitate subsequent processing and analysis.

3 Exploratory Data Analysis

In the EDA part, we want to see the general distribution of the rating stars. Sentimental analysis also gives us rating of each sentence segmented from the reviews. The trend of sentimental scores is also what we are interested in.
It can be seen from Figure 1 that most people are willing to give high scores (4 or 5 stars) while there are not many people giving moderate evaluations. From the figure 2, most customers would like to give scores near 0, and the distribution is somewhat right-skewed. Moderately preferred reviews account for the majority.

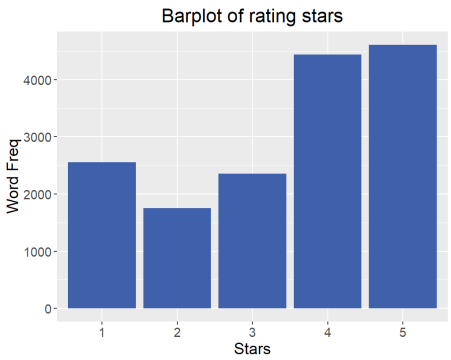

Figure 1: Barplot of stars

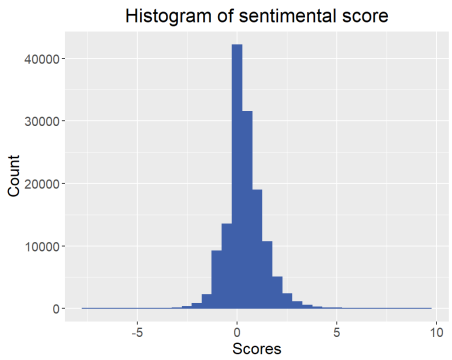

Figure 2: Histogram of sentimental scores

Both of the plots tells us that customers are more inclined to give positive reviews. However, the former is bipolar distribution and the later is central distribution. Since editing long review texts is time consuming, which let customers spend more time to think about their evaluation of the experience. The sentimental scores can give us more information to some extent, so we use sentimental analysis of text reviews to analysis.

4 Feature Extraction

After text cleaning, we pick the words that appear in at least 5 percent of comments, which means they are common enough to be analysed, then we deleting words that are irrelevant or not helpful to our analysis and choose only 40 words. The words we choose and their frequency are shown below.

music	water	3d	busy	dirty	mall	expensive	lobby	imax	family
213	540	544	560	699	701	755	785	867	875
money	far	crowd	room	location	bathroom	beer	snack	kid	chair
883	913	1052	1238	1287	1328	1370	1395	1473	1550
friendly	film	sound	old	park	new	comfortable	screen	service	drink
1651	1751	1753	1858	1991	2366	2675	2778	2814	3225
price	show	popcorn	clean	food	ticket	great	seat	theater	movie
3396	3424	3580	3586	4736	4806	4841	7650	9530	15683

Figure 3: Frequent Words

To decide which words are "important" enough for cinemas, We use the "distance score" and "stars score" to measure the importance and choose those words that are most important.

To calculate the distance score, we count the star ratings of all reviews where a certain word appears, calculate the difference in the proportion of each star rating and add their absolute values.

And to calculate the stars score, we calculate the average score of all reviews when a word appears, and then subtract the average score of all reviews from it.

$$\text{distance} = \sum_{i=1}^{5} |O_i - S_i| \qquad \text{stars} = \frac{1}{5}\sum_{i=1}^{5} S_i - A$$

O_i refers to overall distribution of star rating; S_i refers to distribution of star rating of specific word.

It is worth noting that "distance score" will only have positive values while "stars score"

will have both positive and negative values, so we sum each word's rank in "distance score" with its rank in "absoluted stars score", then get top several words as important factors affecting cinema ratings that be classified into 3 aspects including movies, food and ambiance. For instance, the "movie" aspect includes "price", "expensive", "3d", etc. The "ambiance" aspect includes "dirty", "service", "crowd", etc. The "food" aspect includes "popcorn", "snack", "drink", etc.

5 Recommendations Based on Statistical Analysis

Based on the dataset, we will divide our recommendation/plan for thoes business owners into two parts, one is derived from these attributes indicators in business.json, the other is derived from sentimental analysis of the review text in review.json.

5.1 ANOVA mothed for attributes indicators

In the business.json dataset, every cinema has a line with many defferent attributes like whether accept credit card, whether allow dogs etc. These attributes are True of False valued mixed with many NA. We select 5 attributes as our candidate variables to check whether they are related to user's rating.
For each variable chosed, we ignore the NAs and consider the stars as response to conduct ANOVA between False and True value. Here we choose $\alpha = 0.05$ and finally we obtain 3 effective variables(Table 1).

Table 1: Analysis of variance

attributes	mean of True	mean of False	Pvalue
BikeParking	3.93	3.45	0.044
OutdoorSeating	4.25	3.67	0.00408
GoodForKids	3.57	4.02	0.0035

From the first line of the table, If a cinema has bike parking facilities, then its user rates tend to increase by 0.48. This does make sense because people for those people who don't have cars, riding a bike is a convenient way to reach cinemas, cheap and quick.
The second line indicates that people love outdoor seats especially for dining hall inside a cinema.
The third line is a interesting story because if a cinema is good for kids, then the rate will decrease by 0.45. If we look at relevant reivews, one can find many moviegoers are complaining about the screaming kids and their uncontrollable actions like walking around the seats or touching the screen

5.2 Sentimental Analysis for Review Text Segmentation

Our next part considers making use of the review texts by doing sentimental analysis sentence-wisely with syuzhet package[3]. Firstly, we split every review into single sentences. Then we calculate sentiment score for each sentence.
When we are going to evaluate the performance of a cinema by users' review, there will be three parts needed to consider: movie, ambiance and food. For every part, we choose key words as below

Table 2: key words

Parts	key words
movie	"price","expensive","cheap","brightness","imax","3d","sound"
ambiance	"dirty","clean","mall","crowd","park","service","bathroom"
food	"food","popcorn","snack","drink","water","beer","soda"

In order to figure out these key words, we build a importance evaluate model which take their distribution distance, P-value and frequency into account. You can notice there are some adjectives with opposite meanings like cheap and expensive. They are selected simultaneously to make our advice more reliable. For every given cinema, we can calculate the mean sentiment scores of each key words and then get the score of three main parts by weighted summation, the weight is based on their frequency.

6 Strengths and Weaknesses

The strength of our statistical analysis is that we not only analyze those common indicator variables in business operations, but also try to find the potential factors from reviews. Besides, we separate the text segmentations into three parts including movies, food and ambiance, so we can give advice from certain three aspects.
The weakness is that the three aspects are selected based on experience. Maybe some words should not be classified into them and it is possible that we miss some words, which can lead to bias in subsequent recommendations. And in the sentimental analysis part, the division can be more detailed like dividing by phrase, which can make the analysis results more accurate.

7 Conclusion

In conclusion, we give suggestions based on two aspects, one is from ANOVA of the business indicator variable, the other is from sentimental analysis of reviews. Each business owner can not only see the overall evaluation of three aspects, but also get exact advice on some specific operations. By using the Shiny, the business owners can get clear barplots of their performance and some corresponding advice easily. Following these suggestions can help them improve the rating stars in Yelp in the future.

Contributions

1) Xiao Li edited the feature extraction part of summary, created the code for importance analysis, and is responsible for construction and maintenance of Shiny and the following representation.
2) Chufan Zhou wrote the introduction, preprocessing, strength and conclusion part of summary, created the code related to data preprocessing and text cleaning like word segmentation, wrote the slides for presentation.
3) Mingyu Wang edited the recommendation part of summary, created the code for ANOVA method and sentimental analysis and assist in the construction of Shiny, edited the sentimental analysis part in the slide.

References

[1] https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html

[2] https://blog.csdn.net/shijiebei2009/article/details/39696523

[3] https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-with-r/