Introduction

Today, people get used to looking up information online before they visit somewhere. Yelp provides a good platform where people can submit reviews of businesses using a one-to-five star rating system. Business owners expect higher ratings to increase revenue, therefore advice for improvement based on existing reviews is needed. In our analysis, we use data-driven methods to generate actionable solutions to help improve ratings in Yelp. Moreover, we provide a web application to demonstrate our findings and personalize advice for every single business owner.

Background and thesis statement

Yelp releases open datasets (including four aspects: business, user, review and tip) for our study. The business data includes location data, attributes, and categories. The user data includes the user's friend mapping and all the metadata associated with the user. The review data includes full review texts and the "user_id" that wrote the review and "the business_id" the review is written for. The tip data includes shorter reviews that convey quick suggestions. Considering that the dataset is huge, we focus on a subset of it. We are interested in cinemas because there aren't many analyses on cinemas and the features of cinema is characteristic, easy to filter. The selected dataset consists of 490 cinemas across 11 states in North America (3 in Canada, 8 in the US), 28747 reviews and 11356 tips from 24359 users. To provide suggestions for the business owners, it's vital to find people's attitudes towards their performance in key aspects. In our analysis, we filter out the valid reviews and then analyze the sentiment behind the reviews.

Data Processing

User Selection

We know that there exists some users who intentionally give extreme stars for personal purposes, it’s unfair to include their reviews in our analysis. We find the spam users based on the "user.json" data. Figure 1 shows the 4 variables we use: compliments, followers, average_stars and review_count. We combine all kinds of compliments to be one and combine fans and friends as followers. The spam users are those who receive very few compliments, have very few followers, give extreme stars (very high or very low) and have very few reviews.

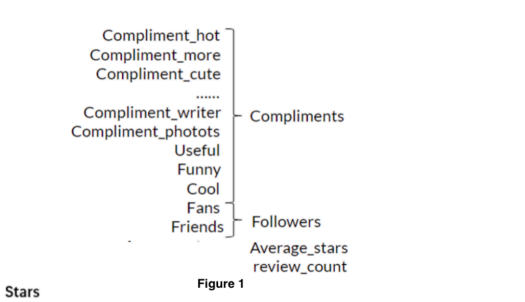


Figure 1

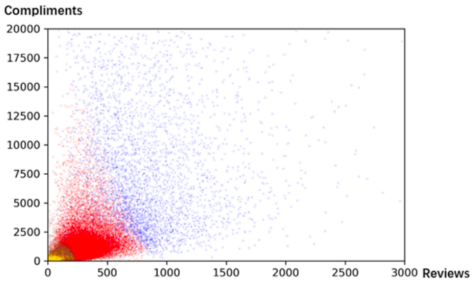


Figure 2

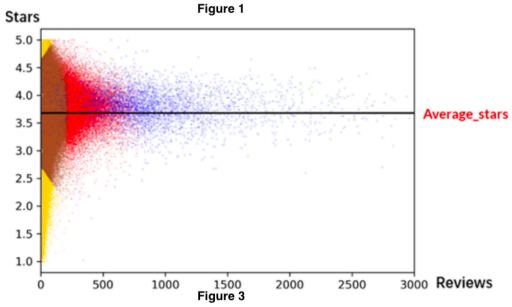


Figure 3

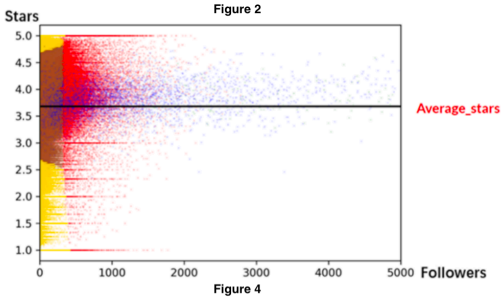


Figure 4

Since no user is labeled as bad user in the data set, we recognize it as a unsupervised learning task. The method we choose is k-means clustering. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. For the parameter of the number of clusters, we tried from 2 to 10. We check the result of clustering by drawing plots. When the number of clusters is 5, we receive the best result. From all 1,637,138 users, 635,755 of them are recognized as spam users. As shown in figure 2, 3 and 4 (different colors for different clusters), the group of golden points share the same characteristic of extreme stars and low amounts of compliments, followers and reviews. After deleting these users, 25816 reviews and 9643 tips are remained for analysis.

Text Processing

Next we process the reviews and tips in order to extract the key attributes later. Our goal in this part is to clean and split the words. Our work takes the following steps:

- Change all letters into lower case;
- Delete stop words (stop words list);
- Split the sentences into words;
- Delete punctuation;
- Turn the tense into the present tense;
- Correct the spelling error.

Word Counting

Then we count the words. Actually, the frequency of word occurance in the reviews and tips contains a lot of information. High-frequency noun can be utilized as indicators of the cinemas' performance, such as seat, popcorn. High-frequency adjectives may contains sentiment, which helps us to analyze the sentiment of the review or tip, especially those adjectives with high variance across different ratings. To count the words reasonably, we take the following steps:

- Count the times every word appearing in the reviews under different ratings, denote it as N_1.

- Put weights on count according to the number of compliments for the reviews or tips (funny, useful, cool) and the time of the review or tip: $count = N_1 \times N_2 \times w_T$ where N_1 denotes the times the word appears in all the reviews; N_2 denotes the total number of compliments of those reviews; w_T denotes the time weight ranging from 0 to 1, 0 being the earliest date the review when the review is given and 1 being the latest date.
- Treat “should/would/not+adj.” and “never+verb.” as one word instead of two.
- Scale the count for comparison.

Extracting the Key Attributes

So far, we've count the frequency of words appearing in the reviews and generated a list of the most frequent words. The list gives us an insight into what people look for and feel when they visit a cinema. We select important business attributes customers values the most from the list. We divide them into 6 aspects:

- Price: the ticket price and the the cost of other goods sold in the cinema
- Location: the location of the cinema and the important geographical elements nearby (hotels, restaurants and car parks), etc.
- Facility: the quality of movie, screen, sound and seats, etc.
- Environment: the wait time to buy tickets, the style of the cinema, etc.
- Food and drinks: the concession stand, the food and drinks offered.
- Services: the quality of service from staff, the online services (eg. online reservation), etc.
- Promotion: special offers for vip, free passes and discount, etc.

For each aspect, we use several key words to represent it for the following study. For example, "Price" is represented by "price", "cost", "ticket" etc.

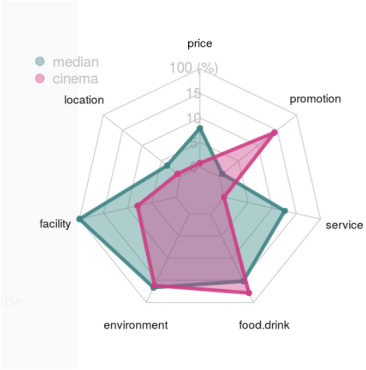
Sentiment Analysis

To study people's attitudes towards the selected key words, we first find the segments where it appears, specifically, we extract the parts between the punctuation right before and after the words. We use a value called "polarity" ranging in [-1, 1] to to denote the sentiment. -1 is the most negative and 1 is the most positive. Written text can be broadly categorized into two types: facts and opinions. Opinions carry people's sentiments, appraisals and feelings towards the world, which are what we focus on in this section. Python's Pattern library provides a lexicon of adjectives that occur frequently in product reviews, annotated with scores for sentiment polarity (en_sentiment). Important attributes in this lexicon are: "pos" - the part-of-speech tags; "sense" - the situation where the word is used; "polarity" - the same as our definition; "intensity" - the effects on sentiment of modifier words (eg. very, little, ...). For a single word, we take the average of polarity values for all the senses as its polarity. Then we take the sum of the polarity values of every word appearing in the segment extracted. If the word comes with a modifier word, we multiply the polarity by the intensity of the modifier word. If the word comes with a negation (eg. never, not, ...), we multiply the polarity by -0.5 and divide it by the modifier word's intensity.

We then sum the polarity up to find the score for each aspect and get a rating chart like following:

id	name	price	location	facility	environment	food&drink	service	promotion
OEQrPxeku4BfHMCSi8UASQ	Chandler Cinemas	0.000	0.000	0.686	1.005	0.225	-0.063	0.313
zRV7bzP_CfTg-_R9U-VsVg	Visulite Theatre	0.009	0.8	9.596	5.817	6.180	3.115	0.000
pDA8NJUwGl1IoLDeaVfo0Q	AMC Ridge Park Square Cinema 8	0.400	0.450	0.624	0.234	0.000	0.683	0.000

The performance of a business will be shown in a radar chart, take the cinema named "Chandler Cinemas" as an example:



Advice Generation

Based on the scores, we first generate comments on the present performance of the cinema. We compare the scores with median scores and give comments like: "*The rating of food&drinks, promotion of your cinema is higher than median. The rating of facility, environment, service of your cinema is lower than median.*" Then, we choose the worst 2 attributes to give them advice like: "*Your customers are complaining about the price*" and "*Your cinema may not be in the best area.*" For the best 2 attributes which have higher scores than median, we give compliments like: "*Your cinema has create a pleasant environment for the customers.*"

Strength and Weakness

Strength

1. Find spam users and exclude them in the following analysis;
2. Evaluate the quality of reviews and tips according to users, time and the number of compliments;
3. Make full use of both tips and reviews;
4. Give results both visually and verbally.

Weakness

1. It's hard to fully evaluate the effiency of the advice;
2. The Python lexicon library is not built on our data;
3. The advice may be not specific enough. For example, we can only give advice on food and drink

but not on specific kind of food such as popcorn, orange juice etc;
4. For the reviews with no key words appearing, we can't find the opinion for any attribute.

Conclusion

We filtered out useful reviews and tips by using K-means method on the user data. Then, we proccess the text and count the word frequencies to extract 7 key business attributes and their relevant key words. Next, we conducted sentiment analysis on the segments where key words appear. We used polarity to denote the sentiment, and scored every attribute. Finally, based on the scores, we demonstrated the present performance of the businesses and give corresponding advice for improvement.

Duties

Yuchen Zeng: Text processing, word counting and shiny app.

Chong Wei: User selection, sentiment analysis.

Jingwen Yan: Extracting the business attribute, sentiment analysis.

References

[1] https://www.clips.uantwerpen.be/pages/pattern-en

[2] Feldman, Ronen. "Techniques and applications for sentiment analysis." Commun. ACM 56.4 (2013): 82-89.