Summary

Abstract

In this project, we used data from Yelp, including business information and customers' reviews to manipulate. We hope to find meaningful and useful features which could affect the rating and provide reliable suggestions to business owners. We chose gyms as our topic and extracted data about gyms from raw data. The two main goals of our projects are:

1. Determine attributes that have influences on the rating of the gyms.
2. Provide suggestions for gym owners based on our research.

For the first goal, we derived information from data 'business.json' and 'review.json' and found key features that may affect the rating. Then we used statistical methods to help us determine whether these features matter.

For the second goal, we did sentiment analysis on each certain gym and give them suggestions based on our result in the first goal. Then we published the results on a shiny app that allows people to get information via visualization.

Goal 1

We conduct this part in 4 steps:

1. data processing
2. text-processing & tokenization
3. variable selection
4. statistical analysis

Step 1:

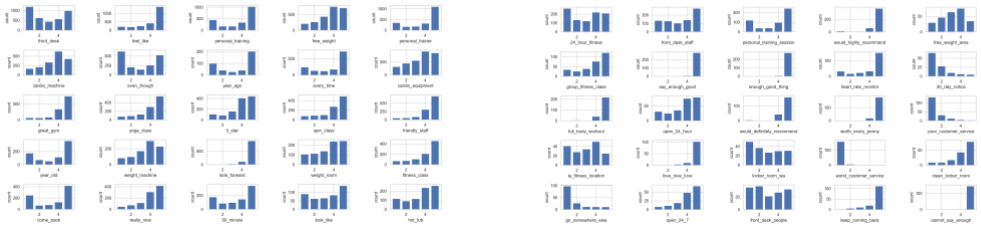
We extracted gym business information from 'business.json'. Then based on the key variable 'business_id' we did the data extraction from other .json files, especially the review data.

Dataset	Size
business_Gym.json	2052
review_Gym.json	58459

tip_Gym.json	9724
user_Gym.json	47124

Step 2:

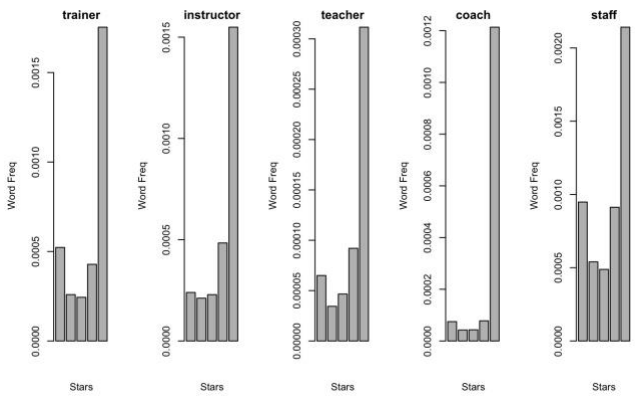
We did text processing on reviews of gyms. We removed reviews that are not in English. We removed all punctuations except prime, changed them to lowercase, removed stopwords, and did lemmatization. Based on these we get word frequency and n_grams. So we get a word box about the gym reviews for further exploration.

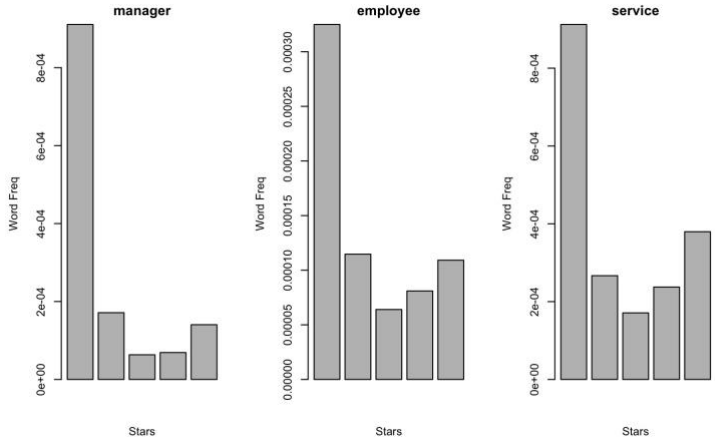


Step 3:

We explored the relationship between the high-frequency words and the rating stars (from 1 to 5). Then we merged some words into groups since they have nearly the same distributions in rating. All these may have influences on the rating of the gyms and need next step exploration.

Keywords Groups	# relative reviews	% mentioning in reviews
class	21972	37.59%
trainer, instructor, teacher, coach, staff	30622	52.38%
manager, employee, service	10197	17.44%
equipment, machine, space, treadmill	21319	36.47%
shower, towel, water, bathroom	9320	15.94%
locker, facility, parking	13160	22.51%
money, fee	17932	30.67%



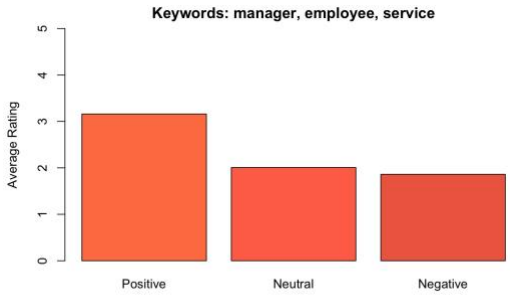


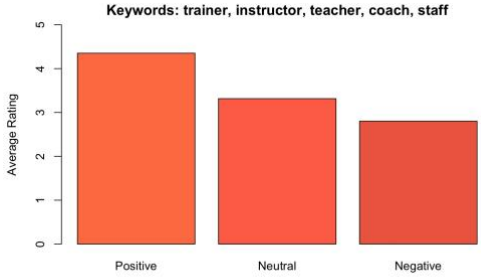
Also, there are some variables in business information data that could have effects on rating. We did regressions on these variables to find whether they matter.

Attribute	Effect	P-value
Wi-Fi (*TRUE*)	Positive Effect	0.0123
Dog-friendly(*TRUE*)	Positive Effect	0.000341
Parking Space in big cities(*TRUE*)	Positive Effect	0.0447

Step 4:

For the 4th step, we first extracted the reviews related to certain keyword groups and extracted the key sentences in the reviews. Then we did sentiment analysis on these sentences and tagged them with ‘positive’, ‘negative’, and ‘neutral’. At last, we did statistical tests on them to find whether the average ratings of these tags are the same.





Also, we utilized the attributes in the business dataset to build linear models. Some attributes are important. Dog-friendly policy and generous Wi-Fi plan help improve the rating. In big cities, providing parking space can help improve the rating while it does not work in smaller towns.

Goal 2

We conduct this part in 2 steps:

1. review-based suggestion data building
2. connection to R shiny app

In step 1, for each certain gym, we extracted the reviews and did sentiment analysis on each keyword group. We compared its sentiment score with the average level. If it is significantly higher than average, it means that this certain gym did a good job on this theme of keyword group, and vice versa.

In step 2, we mainly contain 3 sections: Home: data table; Top words graphs: top mentioned bar plots in single, bigram, trigrams; Gym location maps: gyms displayed in maps; Evaluations on certain gyms: performances and suggestions. As for the first part, we display the ratings, popularity, attributes and key words of the gym, where popularity is a log transformed function of review numbers. The ratings and popularity are to a scale of 0-5, and the key words are transformed to a scale 0-1. As for the second part, we automatically generate suggestions for certain gyms.

Conclusion

As we presented above, wifi service, dog-friendly and parking are significant in the linear model. Also, as most of the reviewers mentioned, the keyword groups also matter in how customers evaluate gyms.

Contribution

Suhui Liu: Data processing, text processing, part of the report, Shiny app.

Shuguang Chen: Data processing, part of variable selection, part of statistical analysis, part of the report, presentation slides.

Yifan Du: Data extraction, data processing, text processing, part of variable selection, part of statistical analysis, part of the report.