Executive Summary

TEAM MEMBERS: Hangyu Kang, Xiangyu Wang, Ruyan Zhou

Introduction

More than 40 hotels registered on Yelp in the Madison area, we wondered which hotels are evaluated as good hotels and what made them obtain a good reputation. Hence, we had planned to analyze yelp reviews about the hotels in the Madison area. The questions for the analysis are:

1. What affects customers' evaluation of hotels, especially in the Madison area?
2. What is the difference between high rated hotels and low rated hotels in the Madison area?
3. How can AC hotel (Our virtual client) be improved?

Background Information/Data Cleaning/Data Pre-Processing

We have a given four datasets from Yelp. In our first step of data cleaning, we merged the review_city.json file to the business_city.json file according to the corresponding business id and selected all the rows of which categories and names contain 'Hotels' or hotel-related words. Selecting the row by their name is necessary because some businesses that are not related to accommodation business but are related to trip come with the selected rows. Using the city column, we set the hotels associated with the Madison area and dropped all unnecessary columns. We made a word embedding dataset using the text review column. We broke the review text into words, getting rid of stop words and special characters, lemmatizing the terms, and making it into matrix form. Using the refined dataset, we made another word embedding dataset for the analysis. At first, we reduced the overlapping count to 1; for example, if the word 'stayed' counted as 2 in a row, we made it to 1. Secondly, we excluded unmeaningful words by selecting more used words than 3rd quantile (53.00).

Exploratory Data Analysis (EDA)

We sorted the words according to their frequency, dividing those words into four categories (Service, Facility, Location, Atmosphere), dumping words if they are judged as unnecessary words. We used our rationale to do this process. We produced relative frequency bar plots for each word like Figure 1, 2.

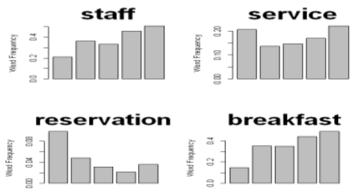


Figure 1. staff, service, manager, reservation, breakfast, coffee

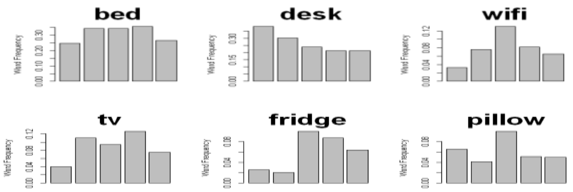
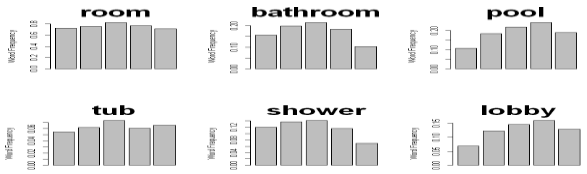


Figure 2. bed, desk, wi-fi, tv, fridge, pillow

Based on Figures 1 and 2, we concluded that employees' roles are essential for determining the evaluation. Some plots are omitted here, but they showed that other services like reservation, provided food, breakfast, buffet, and drinks like beer play an essential role in determining the evaluation. However, the rates are not significantly differed by the provided furniture like bed, fridge. Fare related words(price and money) are also somewhat associated with the rate.



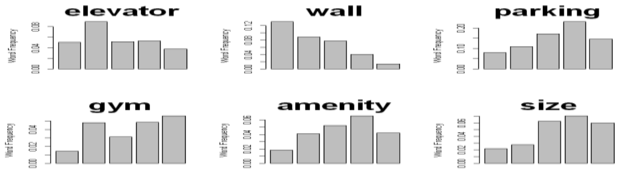
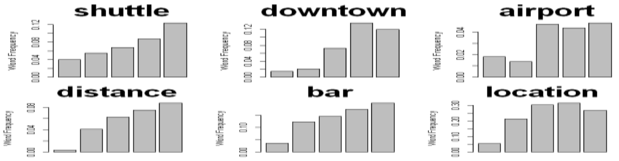


Figure 3. room, bathroom, pool, tub, shower, lobby

Figure 4. elevator, wall, parking, gym, amenity, size

From Figures 3 and 4, we noticed that room itself does not play an important role in the rating. However, amenities like gym, pool, and parking places may affect the rating. Interestingly, the wall shows some significant differences. Some of the unmeaningful bar plots are also omitted here.



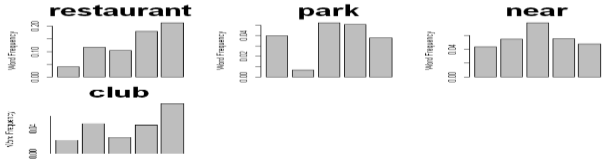
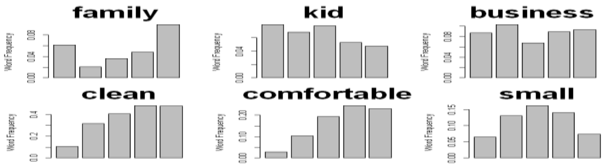


Figure 5. shuttle, downtown, airport, distance, bar, location

Figure 6. Restaurant, park, near, club

From Figures 5 and 6, we noticed that the words related to the hotels' location and neighborhood restaurants, bars, and clubs show some relationship with the rates. The terms associated with transportations also are relatable with the rate.



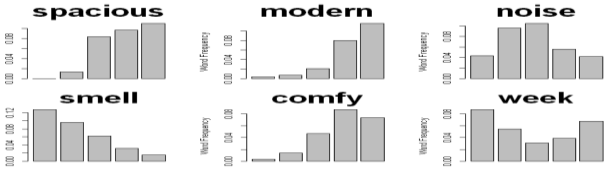


Figure 7. family, kid, business, clean, comfortable, small

Figure 8. spacious, modern, noise, smell, comfy, week

From Figures 7 and 8, we observed customers expect more family-friendly and modern environments to the Madison hotels. Also, we observed cleanness, roominess, and smell of the place is play essential roles in customers' evaluation.

For more accurate results, we performed multinomial-logistic regression and produced corresponding p-values. Also used backward AIC to select out the words that affect customers' rating. Multinomial-logistic regression is used for analyzing categorical variables like stars in our dataset. It uses the proportion of the word's frequency in each star rate and sees if the ratio is different from each other. And we calculated the p-value of the coefficients. If the word's p-value is less than 0.05, then we conclude that the word frequency of the word in each star rate is statistically significantly different. In other words, it affects the customer's evaluation. Backward AIC is the process of seeing which variables affect the outcomes (in our case, star rate) by dropping one variable for each step from the full model (containing all words). If the AIC value decreased drastically comparing to others, then we can conclude the variable is essential. From this process, we ended that the words which are staff, manager, breakfast, booked, money, desk, wall, parking, restaurant, downtown, bar, location, modern, clean, pretty, comfortable, spacious, quiet, comfy, smell and dirty are statistically significant words.

Key Findings Of Businesses Hotel (Accommodation) Market on Yelp

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1	2.77	3.51	3.28	4.00	4.77

Table 1. Summary of the average rate of hotels in Madison Area

We divided 45 hotels into two groups, and the threshold is the median value (3.51) of their average rate. If their average rate is lower than 3.5, we classified those hotels as low rated hotels; otherwise, we classified those hotels as high rated hotels. There are 22 low rated hotels and 23 high rated hotels in Madison Area. Each of them obtained 567 reviews and 965 reviews, respectively. Using the selected 21 words above, we compared high rated hotels and low rated hotels in the Madison area.

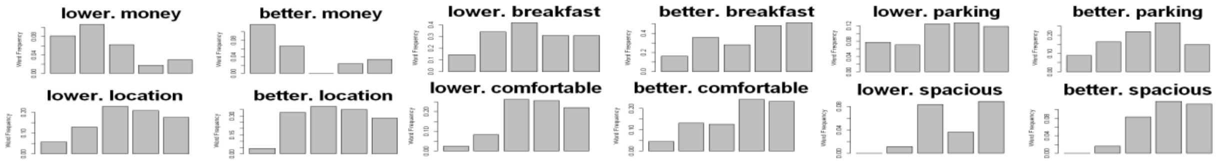


Figure 9. Comparison of relative frequency in words between the low rated hotel and high rated hotel

From Figure 9, we noticed that the relative frequency between high rated hotels and low rated hotels is quite different in words: ‘money’, ‘breakfast’, ‘parking’, ‘location’, ‘comfortable’, ‘spacious’. We performed a proportion test to see whether there are statistical differences in the frequency rate of words in high evaluation range (3,4,5) between the low rated hotels and high rated hotels.

Star	3	4	5
Money	0.03809	1	1
Breakfast	0.06923	0.002298	0.002886
Parking	0.1259	0.004478	0.6172
Location	0.04102	0.008122	0.09249
Comfortable	0.02812	0.8634	0.9627
Spacious	1	0.021	0.7003

Table 2. p-value for Proportion Test

Through the test, we observed that breakfast shows the most considerable difference. We also compared the distributions and the mean values of the star rate of corresponding words "money", "breakfast", "parking", "location", "comfortable" and "spacious" between low rated hotels and high rated hotels. For testing the differences, we produced box plots and performed t-tests.

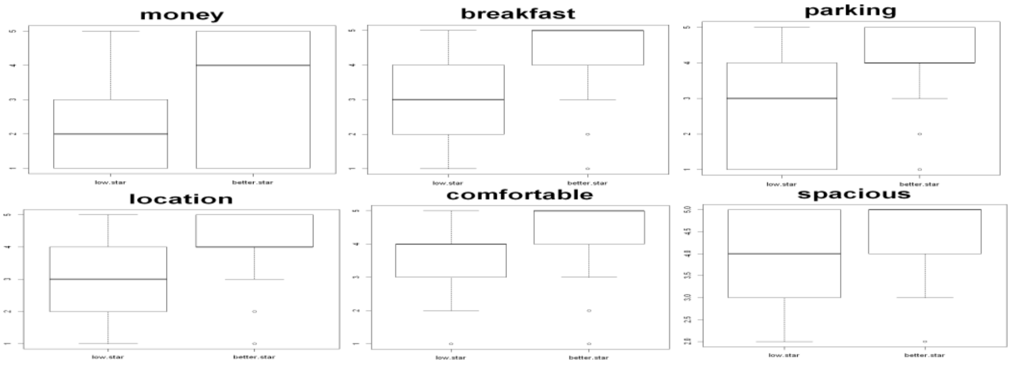


Figure 10. Box plots for six words (money, breakfast, parking, location, comfortable, spacious)

Words	Money	Breakfast	Parking	Location	Comfortable	Spacious
p-value	4.202e-09	2.91e-08	0.01437	4.202e-09	2.91e-08	0.01437

Table 3. p-values for t.test

From Figure 10 and Table 4, we observed all the mean values of the star rate between high rated hotels and low rated hotels are significantly different.

AC Hotel is one of the low rated hotels in the Madison area; however, it has the highest average star rate among the low rated hotels. They obtained an average of 3.46 stars and 44 reviews from the customers. We compared this AC Hotel with the high rated hotels. For the analysis, we used all 21 words, which are concluded as significant factors. We performed t-tests to see whether there are differences in average star rates corresponding words between AC Hotel and the high rated hotels. We also produced bar plots showing the frequency proportion of each word at each star rate to ascertain whether there are differences in their distribution. Some words like ‘booked’, ‘money’, ‘manager’, ‘spacious’, ‘pretty’ and ‘dirty’ are excluded since there are not enough words in AC hotel reviews.

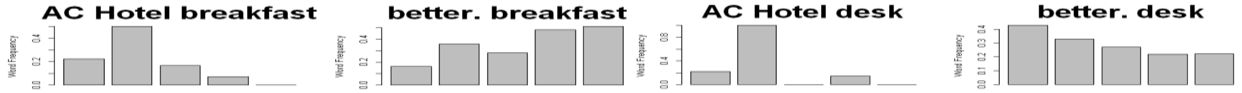


Figure 10. comparison of relative frequency in words between AC Hotel and high rated hotel

	staff	breakfast	desk	wall	parking	restaurant	location	bar	downtown
p-value	0.50	0.02	0.05	0.18	0.54	0.07	0.3	0.14	0.24
	comfortable	modern	clean	quiet	comfy	smell			
p-value	0.76	0.4	0.42	0.69	0.09	0.79			

Table 5. p-values for t.test

We observed a statistical difference in average star rates of “breakfast” and “desk” between the two groups.

Recommendation for Businesses Plan

Statistical analysis showed the averages of customers’ evaluation are different in the six words, ‘money’, ‘breakfast’, ‘parking’, ‘location’, ‘comfortable’, and ‘spacious’. We recommend the low-rated hotel owners adjust their fare properly, improve their quality of breakfast, secure the parking spots for their customers, provide transportation to overcome the locational problem, and change some interior parts of the room to make customers feel comfortable spacious.

The second part of the statistical analysis showed that the customers’ evaluation is different in the two words, ‘breakfast’, and ‘desk’ between AC Hotel and high rated hotels. We recommend the AC hotel owner improve their breakfast quality for customers and improve the desk's quality.

Conclusion

We figured out there are twenty-one words that affect customers’ evaluations. Among those twenty-one words, six words are influential in determining whether the hotel belongs to the high-rated hotel or not. The highest evaluated hotel among the low rated hotels is AC hotel, and as we compared to the high rated hotels, we found their significant weakness is breakfast and the desk.

Our approach's weakness is that we could not connect some adjectives and nouns, which are the subjects of our analysis. For this reason, we could exactly know whether the subjects earn a positive review or a negative review. Instead, we judge the customers’ contentment toward the subjects only by the rated stars. There is a possibility that customers had a good and bad part at the same time in the same review, but we could not handle that.

Contributions:

HK

- Had done the data cleaning process.

- Mainly worked on the analysis part

- Wrote key finding part, the recommendation for business plan part, and conclusion part in the executive summary

- Worked on Shiny App (Tips and a part of customers' evaluation)

- Wrote presentation slides 18-27 pages and record the voice for that parts.

XW

- I had done data cleaning process upon attribution of hotels in Madison, but sadly we could not find any useful information from it.

- Wrote introduction and data cleaning part of summary

- Main editor of Shiny App.

- Also wrote presentation slides 1-7 and recorded the voice for that parts.

RZ

- Tried to study the impact of attributions with XW, but hard to do analysis since majority of the attributes of Hotels are NA.

- Had done the sentiment analysis for clarifying reviews into positive or negative ones and embedded it into shiny app.

- Wrote EDA part in the summary, also the proofreading of the summary.

- Simplified the code of Shiny app and draw the plots except the word clouds in shiny app.

- Wrote presentation slides 8-17 and recorded the voice for that parts.