**Executive Summary**

Hangyu Kang, Xiangyu Wang, Ruyan Zhou

**Introduction**

There are more than 40 hotels in Madison area, we had wondered which hotels are evaluated as good hotels and what made them obtain good reputation. Hence, **f**or the project, we had planned to analyze yelp reviews about the hotels in Madison area.

Our questions for the analysis are the same as the following:

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

**Background Information/Data Cleaning/Data Pre-Processing**

We have a given 4 datasets from the Yelp. There are the files that shows details of the businesses, the reviews about them, detail information about the users who leave the reviews and useful tips for the business owners. In our first step of data cleaning, we merged review\_city.json file to business\_city.json file according to the corresponding business id. Secondly, we filtered out all the rows according to the categories that contains ‘Hotels’ and then selected out the rows according to the business name that contains ‘hotel’, ‘inn’, ‘Hilton’, ‘Courtyard’, ‘Edge’, ‘Place’, ‘Graduate’, ‘Wyndham’, ‘Suites’, ‘Motel’ and ‘Lodge’. Selecting out the row by their name is a necessary process because there are some businesses that are not related to accommodation business but are related to trip and restaurants come with the selected rows. After that, we selected out the rows that city column is equal to ‘Madison’ and dropped all columns except rated star, review text, address, postal code, average rates, attributions of the businesses and categories of businesses. For the next step, we made word embedding dataset using the text review column. We break the review text into word, getting rid of stop words and special characters, lemmatizing the words, and making it into matrix form. The refined dataset we used for analysis is the same as the Table 1.

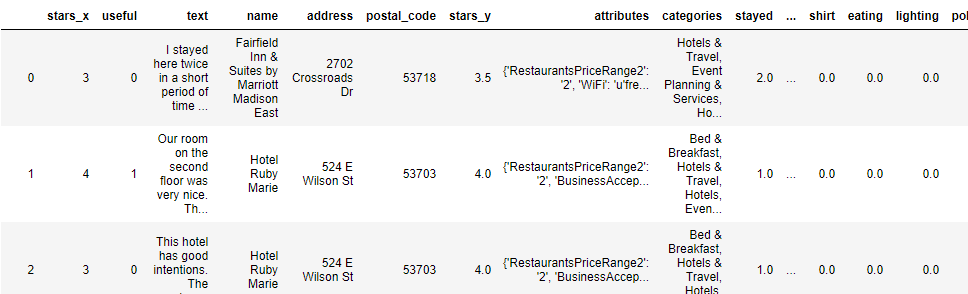


Table 1

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’ are counted as 2 in a row, then we made it to 1. Secondly, we excluded unmeaningful words, by selecting the words which are more used than 3rd quantile (53.00).

**Exploratory Data Analysis (EDA)**

We sorted the words according to their frequency (Figure 1), dividing those words into four different categories (Service and furniture, Facility, Location, Atmosphere), dumping words if it is unnecessary words. We used our rational to do this process.



Figure 1

And we produced frequency bar plots for each word like Figure 2, 3.

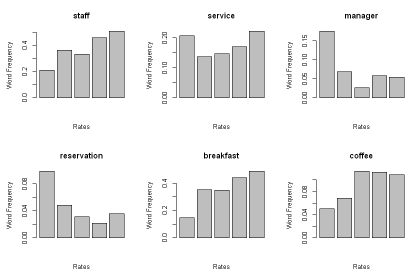
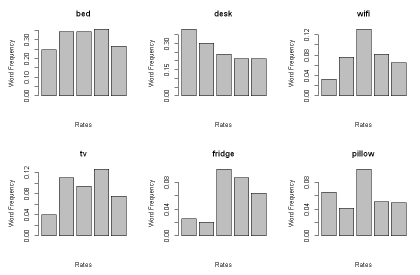
 

Figure 2 Figure 3

Based upon Figure 2 and 3, we concluded the roles of employees are important for determining the evaluation. Some plots are omitted here, but they showed that other services like reservation, provided food, breakfast, buffet, drinks like beer are played important role to determine the evaluation. However, the rates are not significantly differed by the provided furniture like bed, fridge. Price related words price and money are also somewhat related to the rate.

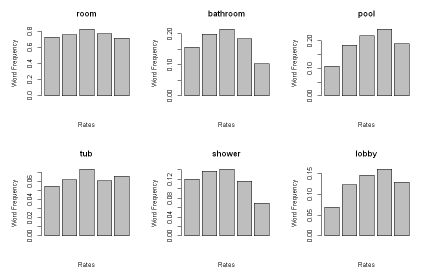
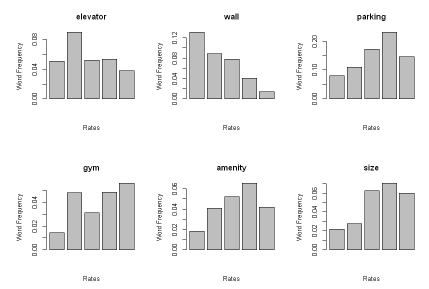
 

Figure 4 Figure 5

From Figure 4 and 5, we noticed that room itself does not play important role to the rating. However, amenities like gym, pool and parking places may affect to the rating. Interestingly, wall shows some significant difference. Some of bar plots also omitted here.

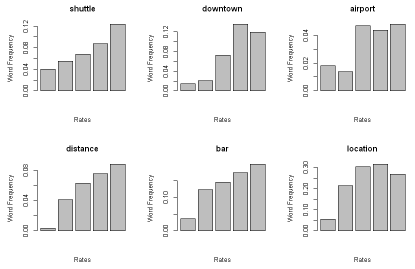
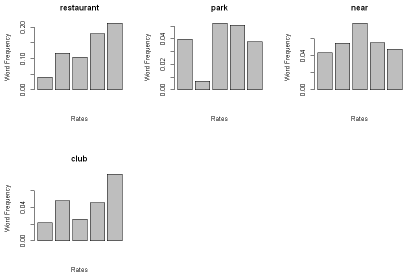
 

Figure 6 Figure 7

From Figure 6 and 7, we noticed that the words related to the location of the hotels and neighborhood restaurants, bars and clubs shows some relationship to the rates. The words associated with transportations also are relatable with the rate.

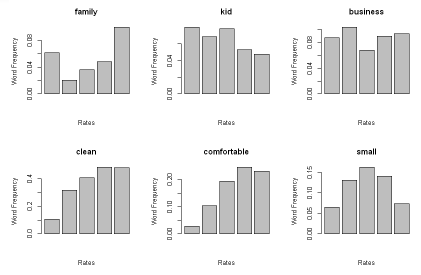
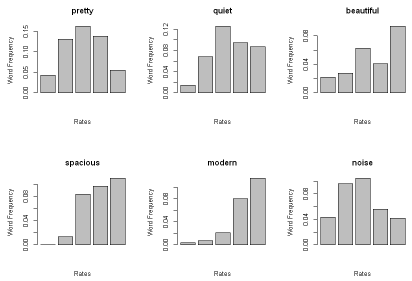
 

Figure 8 Figure 9

From Figure 8 and 9, we noticed customers expect more family-friendly and modern environments to the hotels in Madison. Also, we observed cleanness, roominess and smell of the place is also played important 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 effect on customers’ rating.

Multinomial-logistic regression is used for analyzing categorical variables like stars in our dataset. Basically, it uses proportion of the frequency of the word in each star rate, and see if the ratio are really different from each other. And we calculated p-value of the coefficients. If the p-value of the word is less than 0.05, then we can conclude that the word frequency of the word in each star rate are statistically significantly different. In other words, it effects on the customer’s evaluation.

Backward AIC is the process to see which variables effect on the outcomes (in our case, star rate) by dropping one variable for each step from the full model (containing all words). If AIC value decreased drastically comparing to others, then we can conclude the variable is important.

From this process, we concluded 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 About Businesses Hotel (Accommodation) Market on Yelp**

There are 45 hotels in Madison area. The summary of their average rating is the same as the Table2.



Table 2

We separated 45 hotels into two large groups according to the Median value (3.5) 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 words above, we compared high rated hotels and low rated hotels in Madison area.

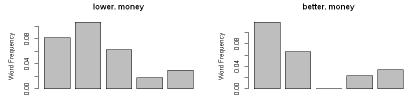
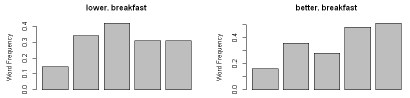
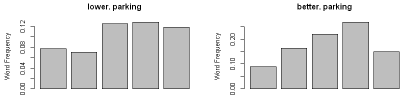
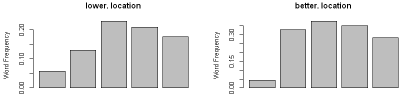
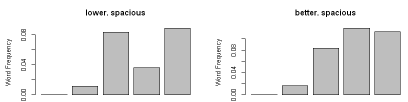
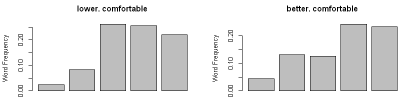
    

Figure 10

From the Figure 10, we noticed that the frequency proportions between high rated hotels and low rated hotels quite different in the words, ‘money’, ‘breakfast’, ‘parking’, ‘location’, ‘comfortable’, ‘spacious’.

are quite different

**Contribution**

HK