APAN 5205 Final Project Report
Instructor: Vishal Lala
Name: Shuhan Zhang, Junzhe Gong, Sheng Luo, Zhenghui Xue

Yelp Open Data Analysis Project

Background

Yelp.com is a well-known crowd-sourced business platform founded in 2004 by Jeremy Stoppelman and Russel Simmons. It provides the services with reviews and ratings, and the users may evaluate establishments, post reviews, and discuss their adventures of being served. This platform includes companies in the restaurant, retail, hotel, and tourism industries, among other sectors. As of 2021, Yelp hosted over 200 million reviews and had approximately 35 million unique monthly visitors on its mobile app. Yelp also offers company and user review data sets that are open to the public. For those local businesses, as long as they review and analyze the dataset wisely, they could find out valuable business insights from customers' reviews. Then, it would be effective to make or keep the corresponding business features in this market.

Relevant Techniques

To better serve Yelp's business, there are several relevant techniques helpful in the project implementation. These techniques include clustering, text mining, association rules, and a recommender system. There are rationales for considering them in further analysis. The clustering method allows us to group similar data points based on their features. In the context of Yelp reviews, clustering could help to identify patterns and trends in the data, grouping customers' preferences and consumption behaviors. The text mining method mainly refers to sentiment analysis in this scenario. The review contents need to be broken down and tokenized by words to extract the sentiments of every sentence. This technique could help Yelp reviews, gauge overall customer satisfaction, and provide some insights into customer attitudes towards the target business. Association rule mining is discovering relationships between variables or items in a dataset. In the context of Yelp reviews, this could reveal relationships between different review attributes(e.g. Ratings, categories) or customer behaviors. Lastly, the recommender system is a technology that could provide suggestions to individual user preferences through personalized information filtering.

Project Purpose

In today's business landscape, Yelp generates a vast amount of user review data daily, making it crucial for retailers to fully understand user feedback and make appropriate adjustments to meet market needs. This project aims to explore and utilize the wealth of customer review data on Yelp to improve companies' service quality. To achieve this goal, it is vital to analyze ratings, extract valuable insights from customer review text content, and investigate the relationships between these reviews and quantitative metrics, such as rating stars and comments on usefulness. Three techniques are employed in this project: clustering, sentiment analysis,

association rules, and a recommender system. The choice of these techniques is based on the dataset's content and the research questions of interest. By leveraging the synergy between these techniques, the project aims to provide businesses with a comprehensive understanding of customer sentiment and preferences, ultimately leading to enhanced service quality and increased customer satisfaction on Yelp.

Research Questions

Before starting this project, several objectives are set as follows:

1. What features or patterns do the segmentation customers have?
2. What keywords lead to the positive attitude and tone of customer reviews on Yelp? And what factors lead to the negative ones? Would the influential factors vary across various types of businesses?
3. Does the richness of categories stimulate the popularity of the business? Does the rule exist for specific combos of categories?
4. Is it possible to return personalized suggestions on business selection? How could companies better target the customer base with this recommendation?

Dataset Exploration

There are 2 datasets we are going to use in our project after browsing through the raw datasets downloaded from open data sources, including the Yelp academic dataset in business and review. The dataset ‘business’ contains essential information about different establishments, including address, state, business categories, and some attributes. Datasets ‘review’ is about customers' comments on product and service quality including, business IDs, review IDs, User IDs, stars, joy, cool, funny, and review text accordingly. All of them are stored in JSON files, so the first step is converting these data into data frames in R. Then we use the ‘sapply’ function to check if there are any missing values in each column in every dataset and find that only ‘business’ set has NA values in three columns including attributes, categories and hours. Thus, we delete the rows that contain missing values in the data frame and randomly select 10,000 user review data and use 47,350 business info data for further analysis.

Link of the dataset after cleaning:
https://drive.google.com/drive/folders/1hF1h_RfLAl82ojxXLuYnKdlj9G1rrjem

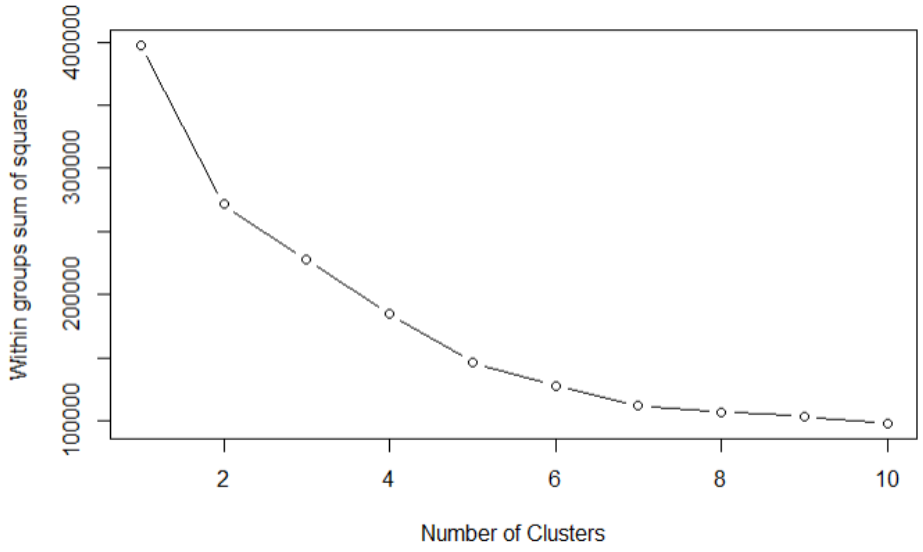
Report Results

1. Clustering

Our research aims to perform a cluster analysis on the Yelp.com dataset to understand the similarities and differences between user reviews and ratings of different merchants to derive business insights about different merchants. We explore whether there is clustering in the restaurant rating data, i.e., whether the ratings of different restaurants are concentrated in some

specific score ranges. Clusters of restaurants with similar characteristics, such as price, type of cuisine, and service quality, are also identified. This can help restaurant owners better understand their business's market positioning and provide more targeted services. and examine the differences between restaurant clusters and identify the key associated factors. For example, customers in certain clusters focus more on service quality, while customers in other clusters focus more on food quality.

The ggplot2 and dplyr libraries in R language were used to process and visualize the data. During data pre-processing, we perform operations such as cleaning, de-duplication, and missing value processing to ensure the accuracy and completeness of the data. The filter() function removes outliers in the dataset with scores and funny, useful, and cool scores greater than 10 and selects columns for clustering. Data normalization transforms the data into a uniform format for further analysis and modeling. The selected columns are normalized to ensure that scaling and offsets do not affect the clustering results. We also determined the number of clusters using the "elbow rule" to determine the optimal number of clusters to distinguish differences between groups better. Based on the number of clusters shown, we chose the optimal number of clusters to be 2 for the study.

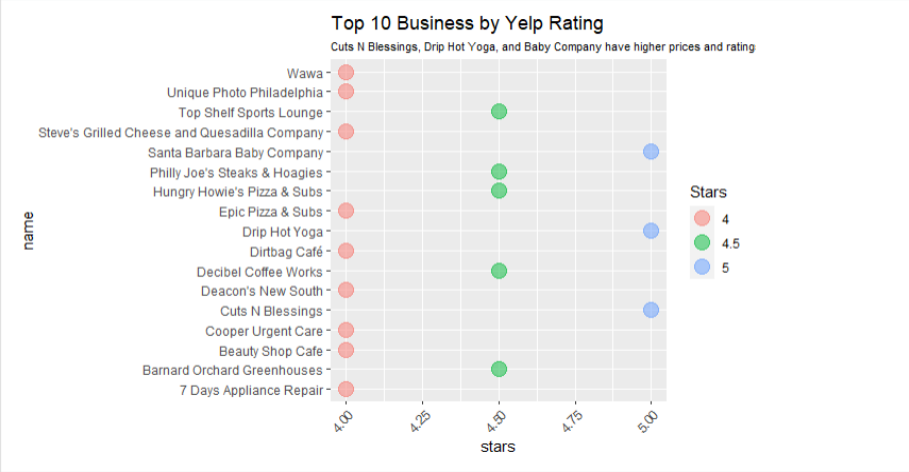


Running the clustering algorithm involves clustering the data using the k-means algorithm and adding the clustering results to the original dataset. Randomly select k initial centroids; k points are usually randomly selected from the sample as initial centroids. Assign the sample points to the class where the nearest centroid is located, i.e., calculate the distance from each sample point to the k centroids and assign them to the class where the closest centroid is located. Based on the assigned sample points, recalculate the centroids of each category, i.e., calculate the mean of all sample points within each category as the new centroids of that category. Steps 2 and 3 are repeated until the centroids no longer change or a preset number of iterations is reached. And to choose the number of clusters, we use l the elbow method, which is to calculate the sum of squared errors (SSE) for different numbers of clusters (k values), then draw

a line graph of k values and SSE, and find the inflection point where the SSE changes significantly, and the k value corresponding to the inflection point is the optimal number of clusters. In our code, k=2 is chosen as the optimal number of clusters at the inflection point, and the kmeans() function is used for clustering.

Visualizing the clustering results, we use the geom_point() function from the ggplot2 library to plot a scatter plot of the data points for the different clusters and annotate and embellish the graph as necessary. In our code, the ggplot() function and the geom_point() function are used to plot the scatterplot with added captions, axis labels and axis range restrictions, and different color markings for the data points of different clusters.

The characteristics of cluster 1 are more useful, funny, and cool; cluster 2 is not that useful, funny, and cool. The four characteristics of funny, stars, useful, and cool are used as clustering.



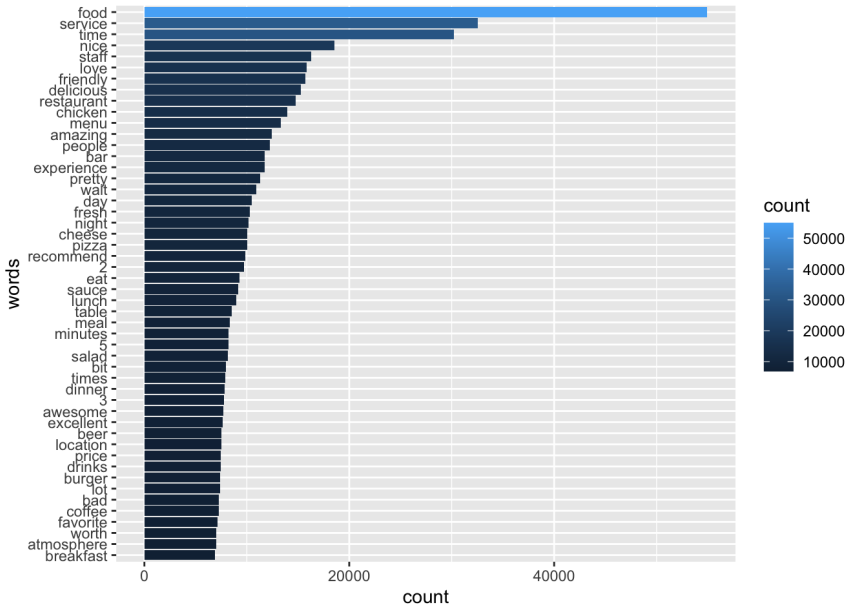
Also, based on the information from business.csv, we extracted the information from the name and city columns to make the top 10 rated companies. We used the functions in the ggplot2 library to plot a scatter plot and a bar chart. In the scatter plot, the x-axis represents the business's name, the y-axis represents the city of the business, and the color represents the business's star rating.

The scatter plot gives a general idea of the distribution of merchants, and we can see that the distribution of merchants is relatively scattered, with no obvious trend of concentration. The bar chart provides a clearer comparison of the star ratings of different merchants, which shows that the star ratings of merchants in different cities vary greatly. Also, it shows the number of high-rated and low-rated merchants in each city.

2. Sentiment Analysis

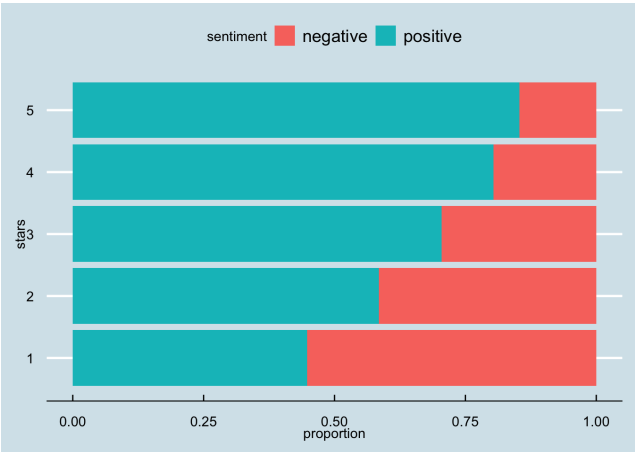
We focus on the review dataset for text mining and sentiment analysis in this CSV file. We calculate the average number of characters, words, and sentences per review and only the 'text' column from the dataset. Then we identify the shortest and longest reviews in the dataset, based on the number of words in each review to clarify the outliers and particularly noteworthy reviews. By calculating the correlations between the length and complexity of the reviews and the star rating that each review received, we identify whether there is a relationship between the way reviews are written and the rating they receive.

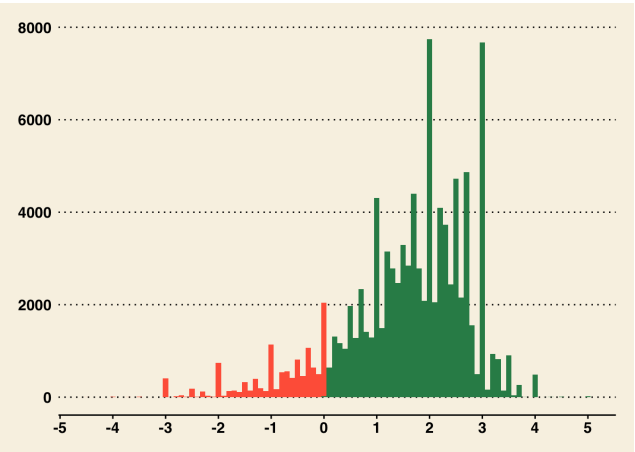
We also analyze the most common words in the reviews, excluding commonly occurring "stop words" like "the" and "and" to give a sense of the topics and themes that are most commonly discussed in the reviews. Thus, we visualize the most common words creating a horizontal bar chart that displays the 50 most commonly occurring words in the reviews to extract insights and patterns and help to understand better the factors that may influence a review's star rating. By using the unnest_tokens function to split the reviews into individual words, we select words that are grouped by frequency and arranged in descending order. This aims to help identify recurring themes or issues mentioned by customers.



To deeper understand the level of engagement and satisfaction of customers, we calculate the proportion of positive and negative sentiment words for each star rating. By using the Bing lexicon for sentiment analysis, it aims to determine whether customers with different star ratings have different sentiments towards the business and further identify specific areas for improvement or strengths of the business on Yelp.

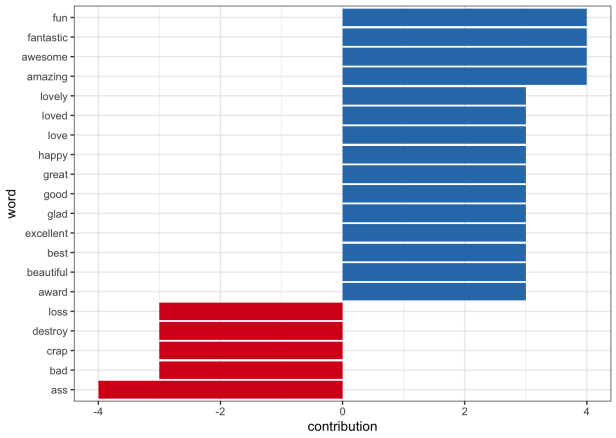
Additionally, the proportion of positive and negative sentiment words in each review calculates the overall positivity and negativity of each review which could particularly inform the business's response to customer feedback. Using the Afinn lexicon for sentiment analysis that assigns a numerical score to each word based on its sentiment, we provide a summary of the overall sentiment of the reviews.



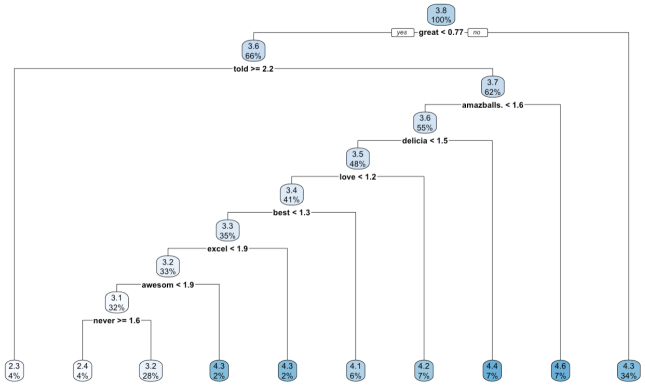


The histogram of the review sentiment scores aims to visualize the distribution of sentiment scores and help evaluate the overall sentiment of the customer base as well. Through generating a word cloud for a specific business, the same approach could also be applied to all

other businesses for target interpretation, and we could recognize specific areas of focus for a particular business. The bar graph of the most positive words used in reviews for a specific business can also help distinguish specific strengths or positive aspects of the business, which can inform marketing strategy or business decisions.



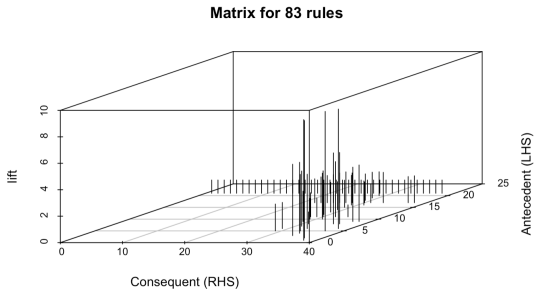
Furthermore, we conduct the predictive models using text features. After converting the text to lowercase, removing URLs, punctuation, stop words, and stemming the words to their root form, the preprocessed text is converted into a document term matrix. Next, a dictionary of frequently occurring words is created to ensure that words with the same stem are treated as the same feature, improving the performance of the model. To remove the sparse terms, words that occur in very few documents, we reduce the dimensionality of the matrix. The tf-idf and simple frequency DTMs are combined with the star ratings for each review. The combined data is split into train and test sets and then a decision tree model and a linear regression model are used to evaluate the test data using the root mean square error (RMSE) metric, which determines the difference between the predicted and actual star ratings. The output from the model is a list of frequently occurring words and their associated frequencies in the review dataset. We do suggest if the reviews or comments for establishments contain words like "food," "place," "service," "order," and "great" are often associated with positive sentiment, while words like "wait," "price," and "didn't" may suggest negative sentiment.



3. Association Rules

Association rules help to discover hidden patterns and relationships. By analyzing the frequency of items that co-occur together, it can identify patterns that are not immediately apparent. In this project, we leverage association rules to see which categories combined closely and provided a reference for new businesses. We break the categories of all businesses and utilize the unique categories to create dummy variables. Then we build the association rules and extract the top associations with high confidence. For example, we would recommend bars have more nightlife elements and shopping malls with more fashion elements. As the plot below we have generated a matrix for 83 rules, which are the most frequent combinations of categories of all Yelp businesses. The new businesses could use these recommendations as a reference when they have some main categories, the rules we provided could help them detect some relevant categories, giving them more opportunities to strive in the competitive marketplace.

	lhs <chr>		rhs <chr>
[71]	{ Bars}	=>	{ Nightlife}
[82]	{ Bars, Restaurants}	=>	{ Nightlife}
[41]	{ Specialty Food}	=>	{ Food}
[83]	{ Nightlife, Restaurants}	=>	{ Bars}
[72]	{ Nightlife}	=>	{ Bars}
[45]	{ Home & Garden}	=>	{ Shopping}
[55]	{ American (New)}	=>	{ Restaurants}
[57]	{ Fashion}	=>	{ Shopping}
[69]	{ American (Traditional)}	=>	{ Restaurants}
[63]	{ Sandwiches}	=>	{ Restaurants}



4. Recommendation

For cluster recommendation in each cluster, the characteristics and attributes of each cluster can be analyzed. For example, there are many high-end restaurants with higher prices, luxury decorations, and high-quality service in Cluster 1. There may be many fast food restaurants with affordable prices and simple service in Cluster 2.

Analyzing the differences and similarities between different clusters can help understand the differences between different clusters. There may be significant differences in ratings and reviews between Cluster 1 and Cluster 2, while certain similarities exist between Cluster 1 and Cluster 2. This can help merchants understand their competitors and market positioning and develop business strategies accordingly. The merchant's desired business strategy can be obtained based on clustering to make the restaurant more approachable or luxurious.

```
 [1] "Al-Sham Restaurant"
 [2] "Hardee's"
 [3] "Scout's Pub"
 [4] "Joe Dunn Arts"
 [5] "Stone Creek Zionsville"
 [6] "McDonald's"
 [7] "Taziki's Mediterranean Cafe - West End"
 [8] "Bro's Pizzeria & Bar"
 [9] "Sheraton Westport Chalet Hotel St. Louis"
[10] "Glazer Children's Museum"
[11] "Dolan's Irish Pub"
[12] "University Family Fun Center"
[13] "The Local"
[14] "Sligo"
[15] "Fez Moroccon Restaurant"
[16] "Rebecca Gay Doula CBE CLC"
[17] "Manhattan Bagel"
[18] "Ramp Riders"
[19] "Francesco's Pizzeria & Ristorante"
[20] "Taco Bell"
[21] "Grotto Pizza"
[22] "The Roost"
[23] "Disc Replay"
[24] "Toyama Japanese Steak House"
[25] "Hyatt Place Philadelphia/King of Prussia"
[26] "IHOP"
[27] "Western Village Inn and Casino"
[28] "Temaki Sushi"
[29] "Commonwealth Kitchen"
[30] "Material Culture"
[31] "La Chancla"
[32] "Sears Appliance and Hardware Store"
```

These merchants are our recommended merchants in the clustering analysis, which are all in clustering 1.

Based on the clustering results, business insights about different merchants are presented, such as which merchants have similar characteristics for customer reviews and ratings and which merchants, Al-sham Restaurant and others, have significant competitive advantages in their respective markets. For example, for high-end restaurants in cluster 1, it can be analyzed what their competitive advantages are and how they can further improve their service quality and reputation, etc. For Chicago Pizza fast food restaurants in Cluster 2, it can be analyzed how to gain competitive advantages in terms of price and service.

Based on cluster analysis results, more market analysis and business suggestions can be provided to merchants to help them better understand the market, develop marketing strategies, improve product and service quality, and enhance competitiveness.

Interpretation

Clustering could make it possible to group the type of reviews, generating the characteristics of each group of reviews. In other words, after making the reviews segmentation, it’s easier to generate the overall market needs. For instance, preferences on prices and locations grouped could separate the customer’s preferences on them to see whether they would prefer or choose the opposite things. According to the result of our analysis, there is one chart named Top10

Business by Yelp rating that could be referred to for market promotion. Moreover, the clustering results could be interpreted interestingly. Higher star rating groups reflected the higher price level, and vice versa. It could be rephrased as the social feature of the restaurants. Higher consumption levels could guarantee better-serving quality, which is essential on some specific occasions. The suggested strategy feedback to the business owners would be a differentiated strategy. Offering better serving quality even in the lower price-level restaurants.

Sentiment analysis is the process of analyzing the user's attitude to their ratings. The trained sentiment model is then applied to the assessment sentence containing each core evaluation category to determine if the sentence is positive or negative and calculate the appropriate sentiment score. Through the analysis, it's more accessible to reflect the impact on the star ratings from their attitude. One valuable finding is that several tokenized keywords are associated with positive sentiment, while some suggest negative sentiment, which could help businesses to inspect the reviewers' comments straightforwardly. In one aspect, these words could reflect the enhancement that the restaurant needs to pay attention to and the potential advantages that the restaurant has obtained. For instance, examining the average waiting time in a restaurant, and finding an optimal method to shorten the waiting time would be a highly effective strategy for companies.

The association rule is a tool to return insights into the relationship between two variables. The overall dataset has a column named category reflecting the market position of the business. This could be used to analyze whether different categories would have inner effects on the final ratings. Some rules that would be interesting to notice, like ”Home & Garden → shopping”, may be ignored. And there are some high lift values in the list of rules which should be examined in the further business operation.

Last, the recommendation system is the extended part of the clustering technique. After the reviews are grouped, the system could return a suggested result to users with the training from the previous part. This is the more general module to return some key prediction results on the clustering results. In the future, it could give more reference value to business owners. Restaurants, on the one hand, could modify their own business to match the taste of the target groups, on the other hand, it is more accessible to reach out to the potential new customer bases with an image of those groups.

Conclusion

In all, this project conducted three techniques to take a deeper analysis of the customers' reviews. There are some valuable insights generated in this project. They could provide some modified suggestions to the companies. Coming back to the research questions. Segmentation customers are grouped to generate consumption patterns. However, it could generate more precise segmentation with more detailed information in the dataset. The tone analyzed in this project is relatively more obvious but could be used for self-checking. The interesting thing is that the association of categories does exist to offer different insights into traditional things. Business companies may even use a more precise prediction system to analyze the features and patterns of customers and give them more proper recommendations.