**A Bite of Data,**

**Dianping.com Restaurants Comments Analysis**

**GAO, Yuan**

**20293988**

**May 20, 2016**

1. **Abstract**

Current study is an exploratory research on dianping.com restaurant comments. By adopting text analytic techniques such as word frequency analysis, sentiment analysis, the study quantified the emphases and sentiments in the comments and compared their differences over various types of restaurants. The results show some hidden information and patterns in web comments.

1. **Introduction**

Dianping.com is the most popular website providing reviews and ratings on places to eat, shop, beautify, exercise, rest, and play. It has become an increasingly important reference for people to make various consuming decisions. At the same time, the large amount of comments from its users contains important information concerning customers’ patterns of preference and behavior. Current study selected 64 Tianjin restaurants on dianping.com, taking their ratings, categorical tags and comments as research objects. By adopting word frequency analysis, sentiment analysis, the study finds out some differences between various types of restaurant including customers’ purposes for going to restaurants, their different emphases in the comments and some correlation between sentiments, per capita price and evaluations.

**3. Research Questions**

A. Does the emphases of user’s comments on dianping.com vary by the type of restaurant they evaluate?

B. Does the sentiments in the comments correlate with per capita prices or ratings of the restaurants on various aspects?

**4. Data collection and management**

The data of the current study are collected from dianping.com. 64 restaurants in Tianjin were selected for current study. The restaurants are in 8 different types, including Tianjin cuisine restaurant, western restaurant, fast food restaurant, cafeteria, hotpot restaurant, Japanese restaurant, Korean restaurant and coffee shop. In each type, eight most popular restaurants were chosen in different price ranges. Then the comments from the users of each restaurant were extracted respectively from their webpages using the Octoparse web data extractor.

Finally, in total of 106,095 comments were extracted accompany with the IDs of the users, the dates and the ratings given by the users.

Then, I used the jiebaR package in R to process the comment texts for further analysis. This part includes Chinese text segregation, stop words removing and word frequency calculation.

**5. Data Analysis.**

1. **Tag Analysis**

Commenters will give each restaurant on dianping.com some category tags. These tags contain information mainly about the purposes when they go to the restaurants. For current analysis, first, I collected the numbers of all tags of each selected restaurant. Then, based on the numbers, I calculated the proportion of each tag for every restaurant. The result is shown in the following table:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Friend | Family | Business dinning | Dating | Casual | Leisure & rest | afternoon tea | Breakfast | Midnight snack |
| Tianjin cuisine | 0.32 | 0.22 | 0.20 | 0.10 | 0.11 | 0.04 | 0.00 | 0.01 | 0.00 |
| Western restaurant | 0.22 | 0.10 | 0.04 | 0.28 | 0.10 | 0.21 | 0.03 | 0.01 | 0.01 |
| Cafeteria | 0.36 | 0.20 | 0.09 | 0.17 | 0.09 | 0.06 | 0.00 | 0.02 | 0.00 |
| Fast food | 0.25 | 0.05 | 0.02 | 0.28 | 0.14 | 0.18 | 0.03 | 0.01 | 0.05 |
| Coffee shop | 0.21 | 0.04 | 0.04 | 0.24 | 0.07 | 0.22 | 0.18 | 0.01 | 0.00 |
| Japanese restaurant | 0.27 | 0.07 | 0.09 | 0.39 | 0.06 | 0.11 | 0.00 | 0.00 | 0.00 |
| Korean restaurant | 0.44 | 0.12 | 0.04 | 0.22 | 0.12 | 0.06 | 0.00 | 0.00 | 0.00 |
| Hotpot restaurant | 0.33 | 0.21 | 0.15 | 0.15 | 0.11 | 0.06 | 0.00 | 0.00 | 0.00 |

Base on the data, I created a pie chart for each restaurant, which will illustrate the proportions more intuitively.

Most people (32%) go to a Tianjin cuisine restaurant for a friends gathering. The following two popular purposes are family gathering (22%) and business dining (20%).

As we can expect, western restaurant is a popular choice for dating, which account for 28% of the purposes. There are also large proportions of people going to a western restaurant for friends gathering (22%) or for a rest (11%).

The largest proportion of people (36%) go to a cafeteria for a friend gathering.

The most popular purposes for going to a fast food are dating, friends gathering and leisure & rest.

The largest proportion of people in Tianjin choose a Japanese restaurant for dating, which is even larger than the one of western restaurant.

An overwhelmingly large proportion (44%) of people going to Korean restaurants for friends gathering, which is followed by dating (22%).

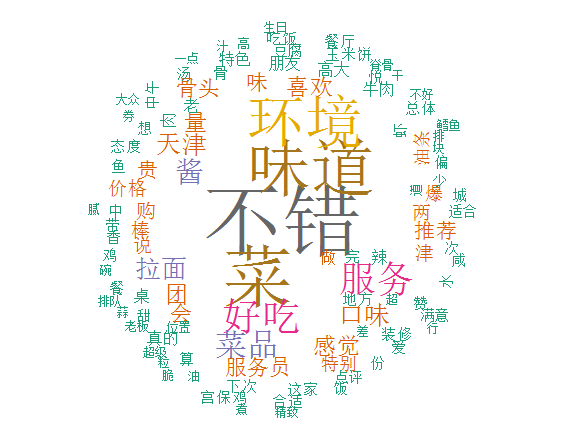
Hotpot restaurants provide a good environment for friends and families to gather together. Sometimes it is also suitable for business dining.

The following bar chart compares the proportions by purpose. Generally, according to the chart, friends gathering and dating are the two most popular purposes for Tianjin people going to a restaurant. To be specific, for friend gathering, most people choose a Korean restaurant or a cafeteria. Tianjin cuisine restaurant and hotpot restaurant are two most popular choices for family gathering and business dining. When it comes to dating, most people would like to go to a Japanese restaurant. Western restaurant and fast food restaurant are also common places for dating. Leading by fast food restaurant, almost all types of restaurants are good for a casual dining. Coffee shop, fast food restaurant and western restaurant are chosen by most people for leisure and a rest. Coffee shop seems to be the only good choice for afternoon tea and only a few restaurants provide breakfast and midnight snack.

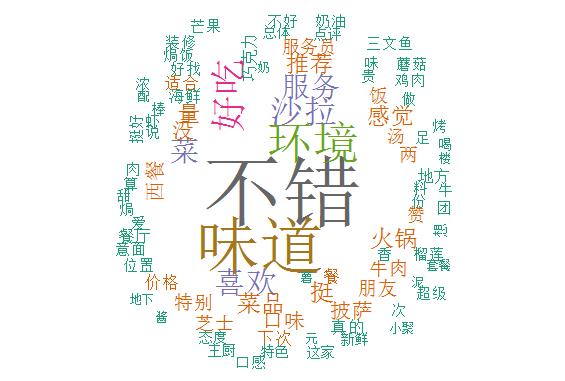
1. **Word Cloud**

In this part, word clouds for each type of restaurant were created using the jiebaR package in R. The results are shown as follows:

Word cloud of Tianjin cuisine restaurant



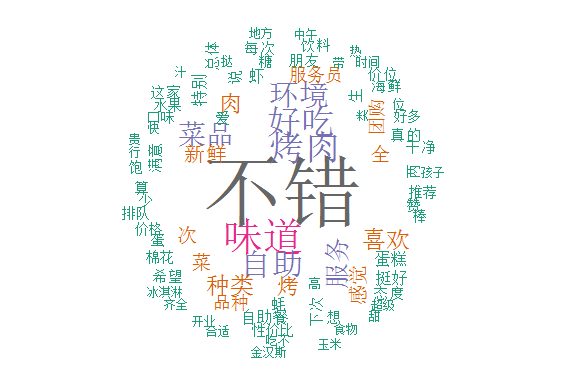
Word cloud of western restaurant



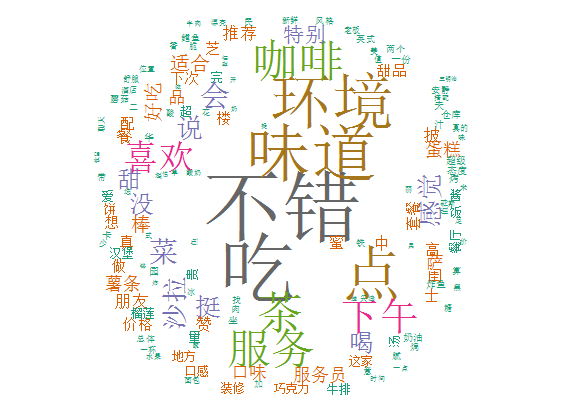
Word cloud of fast food restaurant



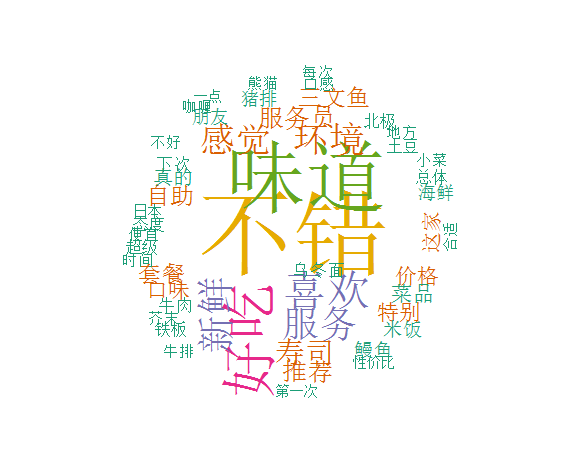
Word cloud of cafeteria



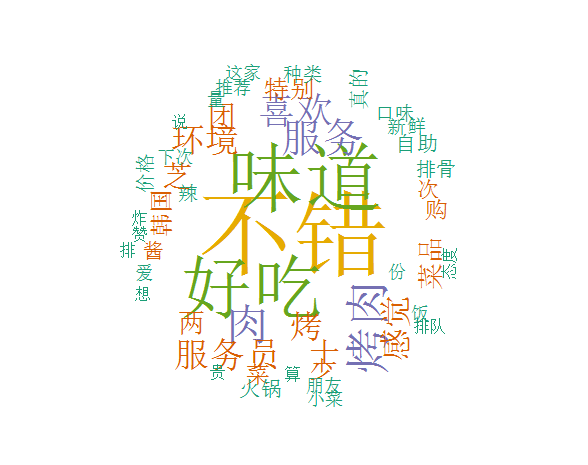
Word cloud of coffee shop



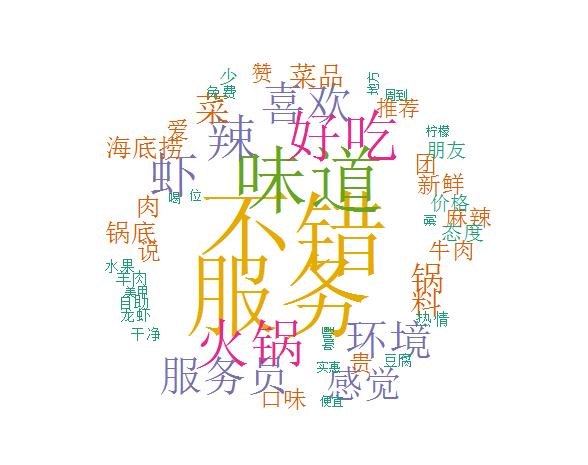
Word cloud of Japanese restaurant



Word cloud of Korean restaurant



Word cloud of Hotpot restaurant



The word clouds show that most commenters feel the selected restaurants are “not bad”. They frequently mention “taste”, “service”, “environment” and the recommended dishes for each restaurants.

1. **Emphasis Analysis**

The word clouds only provide a general idea about the most frequent word in the comments. This section will further the study on word frequency in the comments to explore the question: Does the emphases of user’s comments on dianping.com vary by the types of restaurant they evaluate?

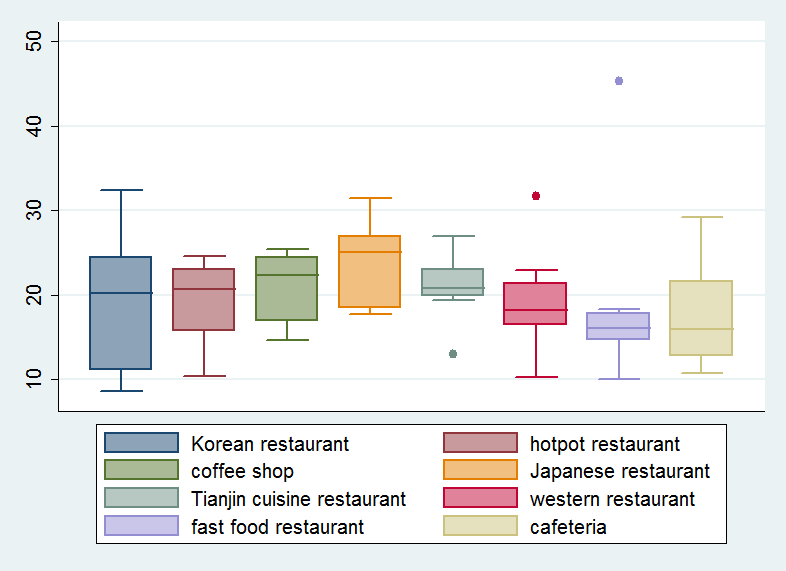
In order to find out the emphases, firstly, I selected the top 100 most frequent words of each restaurant type. Then, I classified these words into four categories: taste, environment, service and price. I summed up the frequency of words in each category respectively and obtain the total word frequency of each of the four comment aspects. The following bar chart manifests the result:

The bar charts show that throughout all types of restaurants, the proportion of comments about taste is overwhelmingly higher than the other three aspects. Specifically, comments of a fast food restaurant, coffee shop and Korean restaurant emphasize slightly more on taste than that of other types. For the other three aspects, customers of Tianjin cuisine restaurant and coffee shop pay more attention on environment than others. Customers of hotpot restaurant mentioned most frequently about service. Coffee shop commenters care much less about price than others.

1. **Sentiment Analysis**

In this part, the ROST content mining system was employed to do the sentiment analysis of the comments. Based on the build-in sentiment dictionary, the system gave every comment a score reflecting the strength of its positive or negative sentiment. Then the average sentiment score was calculated for each restaurant.

The following box plot shows the distribution of sentiment scores by restaurant type.



From the plot, it can be seen that user’s sentiments about all types are positive in general. The sentiments about Korean restaurant vary the most, while the sentiment of Tianjin cuisine restaurant varies the least. In general, Japanese restaurants enjoy the most positive comments. Comments on fast food restaurants are the least positive.

1. **Factors behind sentiment**

In this part, I tried to find some factors in different types of the restaurants that may impact the sentiment in the comments. The result of a preliminary test shows no strong correlation between sentiment and the ratings of any single aspect of the restaurant or per capita price.

It can be concluded that sentiments in the comments are not determined by factors from a single aspect. Following analysis shows that sentiments of different types of restaurants are impacted by different aspects.

Based on the result in section C, emphasis analysis, the correlation between sentiment and different aspects of the restaurant were tested. Here are some key findings:

A quite strong positive correlation between sentiment and price has been found among Korean restaurants. The correlation coefficient is 0.98241.

Another strong positive correlation between sentiment and price has been found among cafeterias, with a correlation coefficient of 0.868901.

The correlation coefficient of other types are shown in the follow table. There is no strong correlation in these types.

|  |  |  |
| --- | --- | --- |
| Hot pot restaurant | | 0.480636 |
| Coffee shop | | 0.480636 |
| Japanese restaurant | | 0.2226 |
| Tianjin restaurant | 0.290629 | |
| Western restaurant | | 0.570201 |
| Fast food restaurant | | 0.354759 |

Then a strong positive correlation (0.943228) between user’s evaluation about environment and sentiment has been found within coffee shop. No strong correlation was found in other types of restaurants.

The result of the correlation analysis is consistent with the emphasis analysis. Korean restaurant and cafeteria, whose comments show greater emphasis on price, also show stronger correlations between their price and sentiment than other types of restaurant. Comments on coffee shops emphasize the most strongly on environment. The sentiments in these comments also strongly correlate with ratings on environment.

**6. Summary**

The results of current study provide a series of interesting findings. First, the proportions of customers’ purposes for going to a restaurant vary by the types of restaurants. The top two purposes for dining in restaurants are friend gathering and dating. Second, user’s emphases in comments also vary by the types of restaurants, but taste is the dominant concern of customers throughout all types of restaurants. Last, no overall correlation across all types of restaurants has been found between sentiment and price or between sentiments and ratings. Instead, the results show some correlations within certain types of restaurants. Sentiments in the comments of Korean restaurants and cafeterias show positive correlations with their respective per capita prices. Sentiments in coffee shop comments are strongly correlated with their environmental ratings. This finding is consistent with the result of the emphasis analysis.

**7. Discussion**

Users’ comments contain a large amount of information that remains to be discovered. Current study is only a preliminary attempt, but there are still some useful findings. The result of purpose analysis can serve as a reference for people to choose the most appropriate restaurant for different purposes. In addition, it provides the restaurant operators information about their target customers. Furthermore, emphasis analysis shows more about the customer’s demand for different types of restaurants. It tells restaurant operators the focuses of operation and directions for future improvement.

However, there are still some limitations in the study. First of all, the number of restaurants included in current study is limited, it may not reflect the whole pattern of all restaurants. Second, due to technique limitation, I didn’t find a good way to combine the emphases of different aspects in the comments with the result of sentiment analysis. So I failed to calculate the weights of commenters’ sentiments about each single aspect.

**8. Appendix**

Here is a piece of R code for text segmentation, word frequency calculation and creating the word cloud of Tianjin cuisine restaurant. The code for other types of restaurant are similar.

*install.packages("devtools")*

*install.packages("jiebaR") #! for Chinese text segmentation*

*library(tm)*

*library(devtools)*

*library(jiebaRD)*

*library(jiebaR) #! Check out "http://qinwenfeng.com/jiebaR/" for more information*

*library(RColorBrewer)*

*library(wordcloud)*

*setwd("C:\\Users\\Gaoyuan\\Documents\\Quantitative Data Analysis\\comments")*

*#> Keywords*

*keys = worker("keywords", topn = 5)*

*keys<="tianjinfood\_comments.txt"*

*#> Text segmentation*

*seg = worker()*

*seg<= "tianjinfood\_comments.txt" # NOT "<-".*

*#equivalent to: segment("ST\_excerpt.txt" , cutter)*

*segdir<-file.path("C:\\Users\\Gaoyuan\\Documents\\Quantitative Data Analysis\\comments", "segmented")*

*segtext<-Corpus(DirSource(segdir, encoding="UTF-8"))*

*summary(segtext)*

*inspect(segtext[2])*

*st.dtm<-DocumentTermMatrix(segtext)*

*st.freq<-colSums(as.matrix(st.dtm))*

*inspect(st.freq)*

*st.dtms<-removeSparseTerms(st.dtm, 0.1) # This makes a matrix that is 10% empty space, maximum.*

*st.freqs<-colSums(as.matrix(st.dtms))*

*inspect(st.freqs)*

*set.seed(166)*

*wordcloud(names(st.freqs), st.freqs, max.words = 50, colors = brewer.pal(6, "Dark2"), random.order=F)*