# Analysis of sentiment data on reviews of restaurant businesses provided by Yelp

## Introduction

Yelp, as an online company, accepts and publishes user ratings and reviews for any local business. To date, the total number of comments on yelp has exceeded hundreds of millions. This report will use review data extracted from reviews data about restaurants, cafes, etc.,

By using the data about the positive words of the review, the negative words, the length of the review and the number of upvotes for the review to mine useful results to guide subsequent activities and based on the given criteria to get the best user and business activity.

## Statistical Summaries for Useful Variables

At the beginning of the report, the stars of the review, the length of review, the number of positive words in the review, the number of negative words in the data, and the net number of sentiment words (calculated by subtracting the number of negative words from the number of positive words).

The Following shows the result of descriptive statistics for these data:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Vars | N | Mean | Sd | Median | Trimmed | Mad | Min | Max | Range | Skew | Kurtosis |
| Stars | 1 | 1569264 | 3.74 | 1.31 | 4 | 3.93 | 1.48 | 1 | 5 | 4 | -0.84 | -0.45 |
| Review\_length | 2 | 1569264 | 125.60 | 115.50 | 92 | 105.94 | 77.10 | 0 | 1047 | 1047 | 2.30 | 7.99 |
| Pos\_words | 3 | 1569264 | 7.07 | 5.93 | 6 | 6.19 | 4.45 | 0 | 94 | 94 | 2.14 | 8.12 |
| Neg\_words | 4 | 1569264 | 2.55 | 3.25 | 2 | 1.93 | 2.97 | 0 | 65 | 65 | 2.77 | 13.08 |
| Net\_sentiment | 5 | 1569264 | 4.52 | 5.24 | 4 | 4.09 | 4.45 | -59 | 139 | 139 | 1.34 | 5.85 |

It is not difficult to see from the table that for the length of the review, the average intersection is nearly 20% higher than the median, while the average and median of other variables do not converge much.

The standard deviation of the review length is the largest among these variables, with a specific value of 115.50, while the standard deviation of other ratio variables is only between 1 and 6, which shows that the length of the review has a very high degree of dispersion. This situation may be caused by some users who are accustomed to writing long comments, which will be statistically reflected as outliers, and the analysis of outliers will be given later.

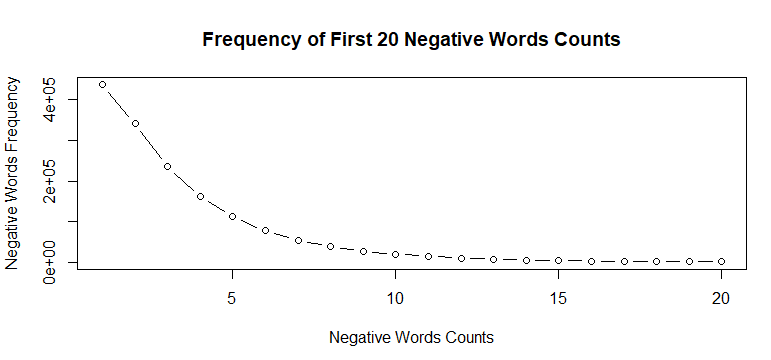
From the skewness values given in the table, only the distribution of stars shows a left-skewed distribution, while the distributions of the other four show a right-skewed distribution, and among these four variables, only the skewness of net sentiment is less than 2. From the analysis of the value of kurtosis, only the distribution of stars shows a leptokurtic distribution, and the other four show a platykurtic distribution.

## Further Analysis of Positive and Negative Words

By analyzing the frequency of occurrence of positive and negative words and selecting the first 20 data, tow figures can be obtained as follows:

图表, 折线图

描述已自动生成



It is not difficult to see from the picture that the distribution of positive words in comments shows a trend from low to high and then to low as the number of positive words increases. The negative has always shown a downward trend.

## Further Analysis of Net Sentiment

A review may have both positive and negative words, so further analysis of net sentiment will be very valuable:

形状

中度可信度描述已自动生成

It is not difficult to see from the picture that net sentiment is mainly concentrated between -6 and 13. Within this range, it shows a trend of first rising and then falling, and the distribution of this score is more positive.

## Average Review Length Per Star Category

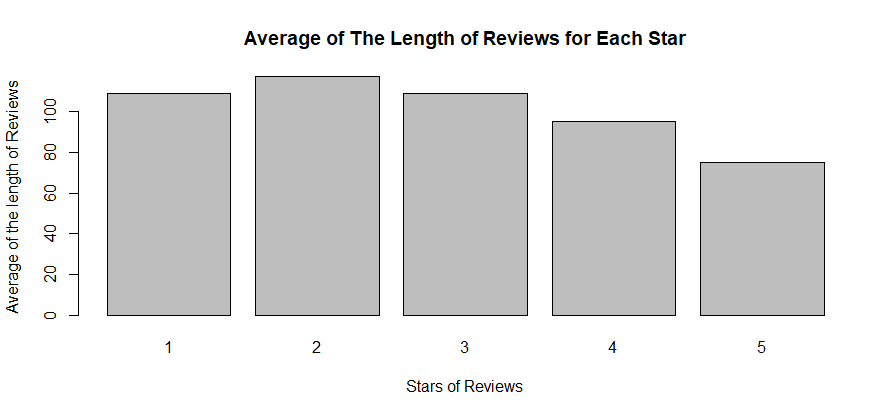
Based on the results of the descriptive statistical analysis on the length of reviews provided in the "Statistical Summaries for Useful Variables" section of this report, the skewness value of the length of reviews is greater than 2, so when calculating the average of the length of reviews, the median value is used instead of the mean value.

The following table shows the analysis results:

|  |  |  |
| --- | --- | --- |
|  | Stars | Avg\_length |
| 1 | 1 | 109 |
| 2 | 2 | 117 |
| 3 | 3 | 109 |
| 4 | 4 | 95 |
| 5 | 5 | 75 |

By observing the Avg\_length variable provided in the table, it is not difficult to see that the reviews increase with the number of stars, it shows a trend of first rising (from 109 when stars is 1 to 117 when stars is 2) and then falling (from 117 when stars is 2 to 75 when stars is 5).

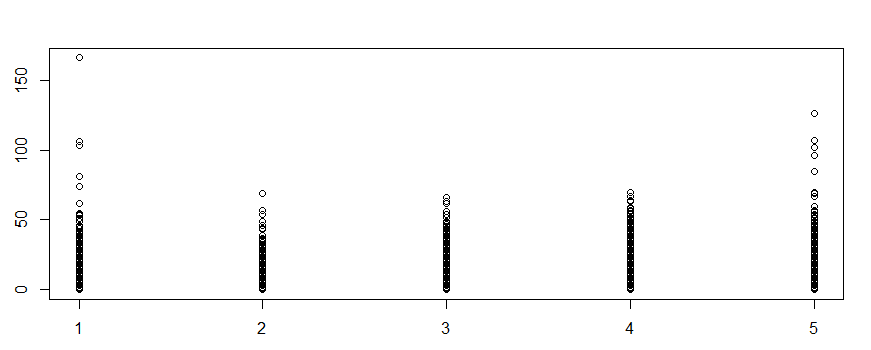
The same conclusion can be obtained by visualizing the resulting histogram:



## Relationship Between Useful of Reviews and The Star-Rating and Length of Reviews

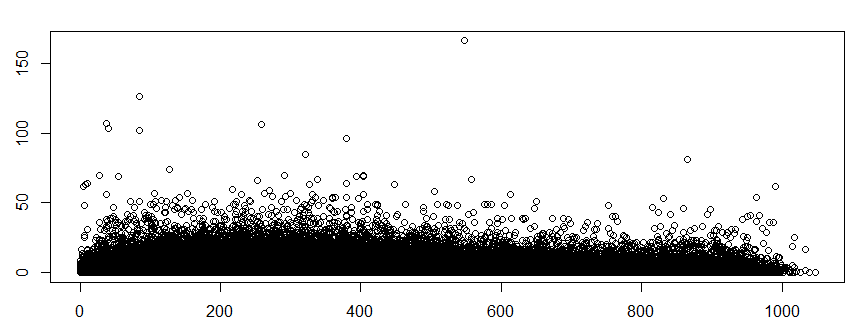
Viewers who read reviews will judge whether the reviews are useful or not. The relationship between the useful of reviews and stars and the relationship between the length of reviews needs to be studied.

The following figure shows the distribution of the useful frequency of reviews in different stars:



It seems that the distribution trend seems to show a trend of first falling and then rising, but if the influence of outliers is excluded, it can be considered that there is no change in the trend. It can be considered that there is no relationship or a very weak relationship between the useful of reviews and stars. The data analysis shows that the relationship between the useful of reviews and stars is -0.04897358, which supports a similar conclusion.

Next is an analysis of the relationship between the useful of reviews and the length of reviews:



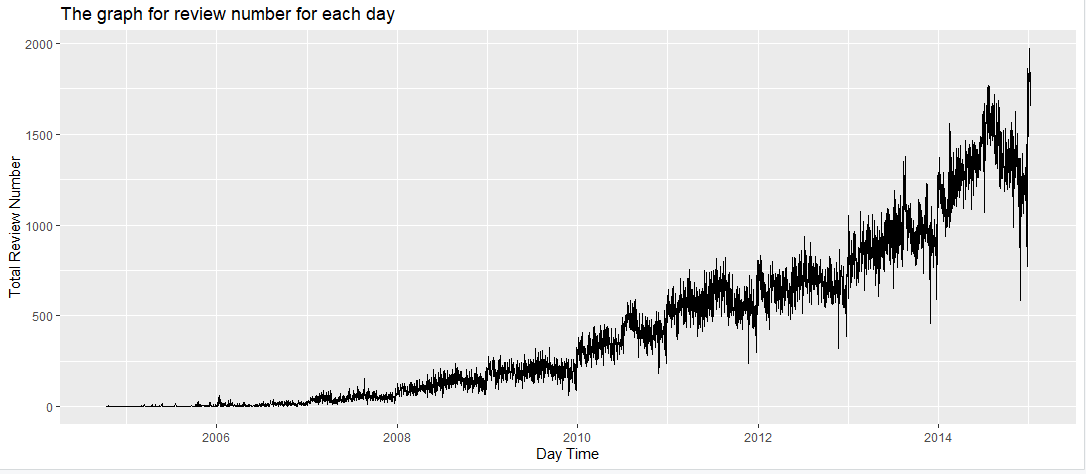
If the influence of outliers is excluded, we can think that there is a weak positive correlation between the useful of reviews and the length of reviews. The result of the mathematical analysis is that the relationship index between the two is 0.3258155, which also supports this conclusion.

Based on the useful of reviews and the relationship coefficient between the two, we will not conduct further analysis on stars, but on the length of revies. Based on the result, it can be known only 10.62% of the variance of useful can be explained because the value of adj R square is 0.1062. A model that simply uses the length of reviews as a variable is not a very good model, because the explanatory power of this model is not enough.

You can try to use both stars and the length of reviews, but the result does not change much, the adj R square is still 10.60%. This may be because the linear relationship between stars and the useful of reviews is too weak, so using this variable will not bring any improvement.

## The Trend of Reviews Per Day

The graph below shows the trend in the number of daily reviews over time:



From the trend chart, the number of comments showed a slight upward trend between 2006 and 2008, there was a clear upward trend between 2008 and 2010. From 2010 to 2014 and after, there was a substantial upward trend. There is good reason to believe that the number of daily reviews on YELP for food and beverages has generally been on the rise.

## Best Business and Best User

To select the best users and the most desired businesses, criteria for "what is best" need to be specified. First, the best users and businesses can't appear only once or only a few times in the data, because too little data is statistically inaccurate. Then we need to find the best one among those users and businesses that have enough occurrences.

Based on the above logic, the criteria for "best" are very clear. For users, find the top 10 users in the data, and find the one with the highest average of the useful of reviews. For businesses, find the 10 most frequently occurring businesses in the data set, and then find the business with the highest average number of stars among them.

The table below shows the 10 users with the most reviews and the average helpful rating of their reviews.

|  |  |  |  |
| --- | --- | --- | --- |
|  | User\_id | Number | Avg\_vote\_useful |
| 1 | uZbTb-u-GVjTa2gtQfry5g | 819 | 1.69 |
| 2 | 3gIfcQq5KxAegwCPXc83cQ | 874 | 1.66 |
| 3 | DrWLhrK8WMZf7Jb-Oqc7ww | 877 | 0.918 |
| 4 | **fczQCSmaWF78toLEmb0Zsw** | **908** | **8.68** |
| 5 | PV5voYSD43Cn\_3gHmxG7DA | 922 | 2.29 |
| 6 | ia1nTRAQEaFWv0cwADeK7g | 941 | 4.85 |
| 7 | glRXVWWD6x1EZKfjJawTOg | 1080 | 3.99 |
| 8 | Iu3Jo9ROp2IWC9FwtWOaUQ | 1186 | 3.89 |
| 9 | ikm0UCahtK34LbLCEw4YTw | 1225 | 4.25 |
| 10 | kGgAARL2UmvCcTRfiscjug | 1427 | 6.16 |

Based on the evaluation logic of the best user, the user whose user\_id is fczQCSmaWF78toLEmb0Zsw will be considered as the best user. Although the number of occurrences of this user is not the highest, the number of occurrences of this user is among the top 10, and statistically speaking, 908 is a sufficiently large sample size. This user's comments have the highest average of helpful approvals among the 10 users.

The following table shows the top 10 businesses with the most reviews:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Business\_id | Number | Avg\_stars |
| 1 | AtjsjFzalWqJ7S9DUFQ4bw | 2241 | 3.95 |
| 2 | CZjcFdvJhksq9dy58NVEzw | 2252 | 3.59 |
| 3 | aGbjLWzcrnEx2ZmMCFm3EA | 2269 | 3.87 |
| 4 | tFU2Js\_nbIZOrnKfYJYBBg | 2337 | 4.00 |
| 5 | YNQgak-ZLtYJQxlDwN-qIg | 2619 | 3.72 |
| 6 | sIyHTizqAiGu12XMLX3N3g | 2657 | 3.88 |
| 7 | Xhg93cMdemu5pAMkDoEdtQ | 2675 | 3.08 |
| 8 | zt1TpTuJ6y9n551sw9TaEg | 3352 | 3.75 |
| 9 | **2e2e7WgqU1BnpxmQL5jbfw** | **3517** | **4.31** |
| 10 | 4bEjOyTaDG24SY5TxsaUNQ | 7137 | 4.14 |

Based on the best business criteria, the listing with business\_id of 2e2e7WgqU1BnpxmQL5jbfw will be selected as the best business.

## Conclusion

The number of daily reviews of yelp has shown an upward trend over time, and it is reasonable to think that the trend of the number of registrations and users of the website has also shown an upward trend in general. Net Sentiment counts will be between -6 and 13 and appear to be mostly positive.