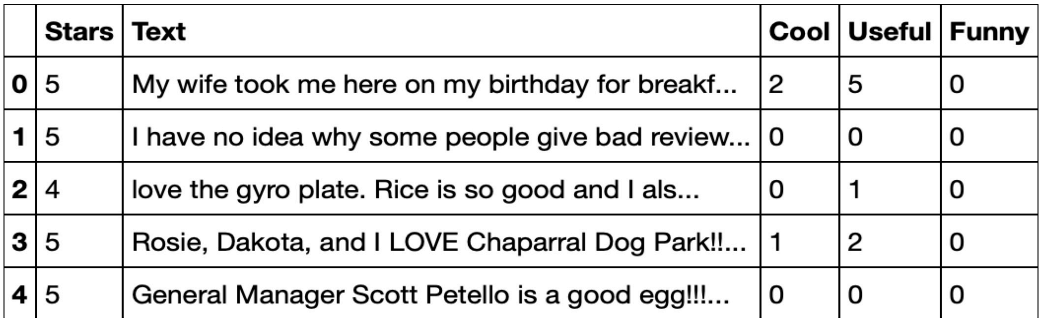
Customers Feedback Analysis

For our Yelp, the centre is customers. We regard customers as the real ‘god’ and try our best to improve Yelp’s restaurants to make sure that customers can get the best service. One of the useful ways is finding the restaurants with higher scores, setting them as examples, and making other lower ratings’ ones modifying themselves. And restaurants’ ratings are highly related to customers’ reviews. Hence, how to analyze these reviews and predict ratings based on reviews becomes an important issue. And our analytics group focuses on using NLP to conduct text analysis and applying MultinomialNB to predict ratings (stars from one to five).

Preprocessing the dataset

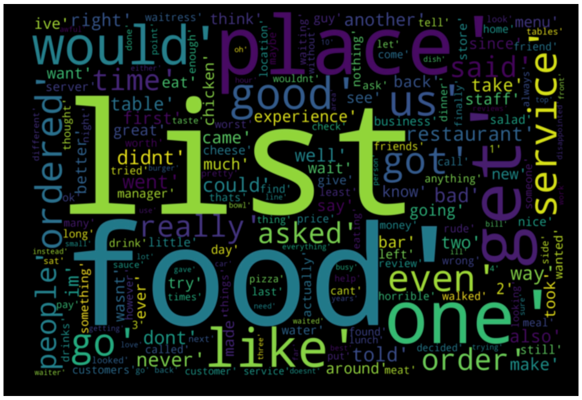
The dataset we will use is from Yelp’s website and contains customers’ information, restaurant stars, texts (reviews), etc. Other unuseful variables are removed and only ‘Stars’, ‘Text’, ‘Cool’, ‘Useful’, and ‘Funny’ are chosen for further analysis. The analyzing dataset is like:

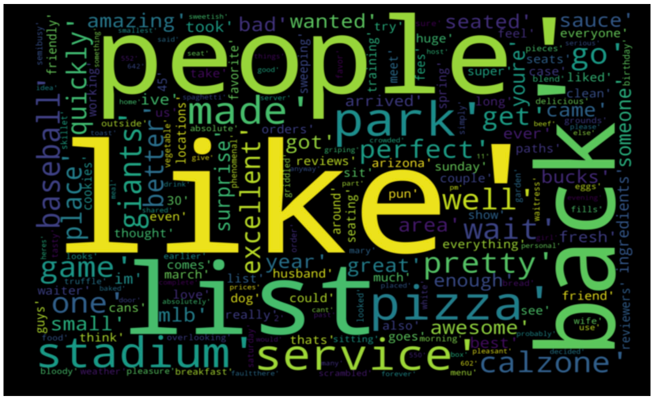


	Stars	Text	Cool	Useful	Funny
0	5	My wife took me here on my birthday for breakf...	2	5	0
1	5	I have no idea why some people give bad review...	0	0	0
2	4	love the gyro plate. Rice is so good and I als...	0	1	0
3	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	1	2	0
4	5	General Manager Scott Petello is a good egg!!!...	0	0	0

Natural language processing for reviews’ (text) analysis

In this part, NLP is used to analyze feedback data and tend to get some insights from these reviews. The final goal is to get word clouds to see the most frequent words of reviews for restaurants with the highest and lowest ratings. By doing this, we believe we will find reasons why some restaurants can get higher stars while others can only get lower stars. In order to get more accurate word clouds, the most basic step is to remove stopwords (e.g. I, am, are, they) and punctuations. If stopwords are kept, they will definitely be the most frequent words and affect the final result. The word clouds from customers’ reviews that belong to those restaurants with stars 1 and stars 5 appear as follows:

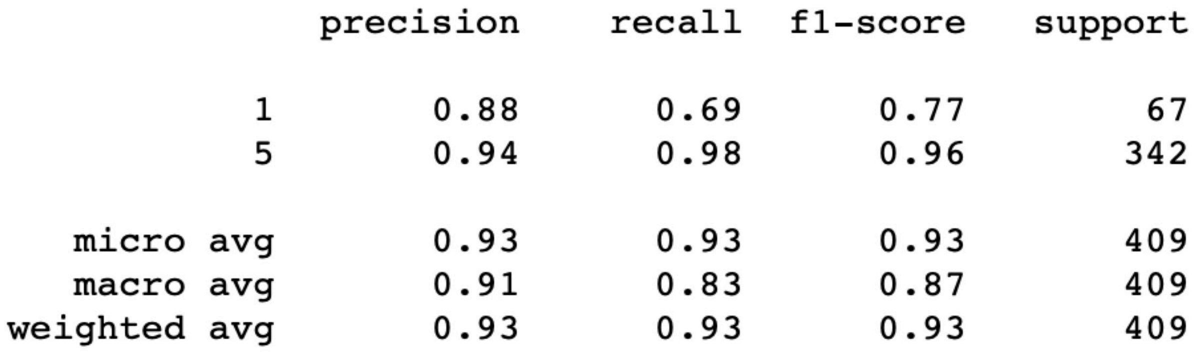




With these word clouds, some points about why some restaurants can perform better than the others can be determined. And these will be shown in ‘Conclusions’ part.

Building up machine learning models to predict ratings

The reason why restaurants’ ratings are needed to be predicted is that we want to find some restaurants that will get lower ratings in advance, provide an alarm to them, and help them improve. By doing exploratory data analysis, all numerical variables – ‘Cool’, ‘Useful’, and ‘Funny’ are not related with ‘Stars’, which means machine learning model must be built based on ‘text’. However, texts cannot be recognized by machine learning models because they are not numbers. Therefore, texts should be transferred to vectors, making sure that they can be put into models. What’s more, we find that the model can predict stars 1 and stars 5 with higher accuracy, which is enough to help us target at those worse restaurants and better restaurants. The final accuracy is **93%** and the result appears as follows:



```
              precision    recall  f1-score   support

           1       0.88      0.69      0.77        67
           5       0.94      0.98      0.96       342

   micro avg       0.93      0.93      0.93       409
   macro avg       0.91      0.83      0.87       409
weighted avg       0.93      0.93      0.93       409
```

Conclusions

From two word clouds, we can find that ‘food’, ‘list’, and some words related to ‘service’ are the most frequency words. As is known to all, food and service mean everything for every restaurant. The restaurants that are rated as better ones perform well because of wonderful food and service, while those getting lower ratings are not so good because of bad food and service. Hence, for any restaurant, they should concentrate on improving their food and service and just ignore some pricing strategies, coupons, etc. Also, lists or menus play an important part and restaurants should make their menus more attractive and readable.

From the technical aspect, we also summarized some points. First, because the model can predict the highest and lowest scores with the best accuracy, we can just apply our model to analyze these two kinds of restaurants. Second, when predicting ratings, other numerical variables can be safely removed and we should only define customers’ reviews as the independent variable. Last but not least, TF-IDF (term frequency–inverse document frequency) is an advanced way to improve text-related machine learning models’ accuracy. However, in our Yelp reviews’ analysis, it makes no contribution and we should not count on it to help us increase accuracy.

(Github: https://github.com/Zhenyu0521/Text-Analysis/blob/master/NLP for Yelp Reviews/NLP_for_Yelp_Reviews.ipynb)