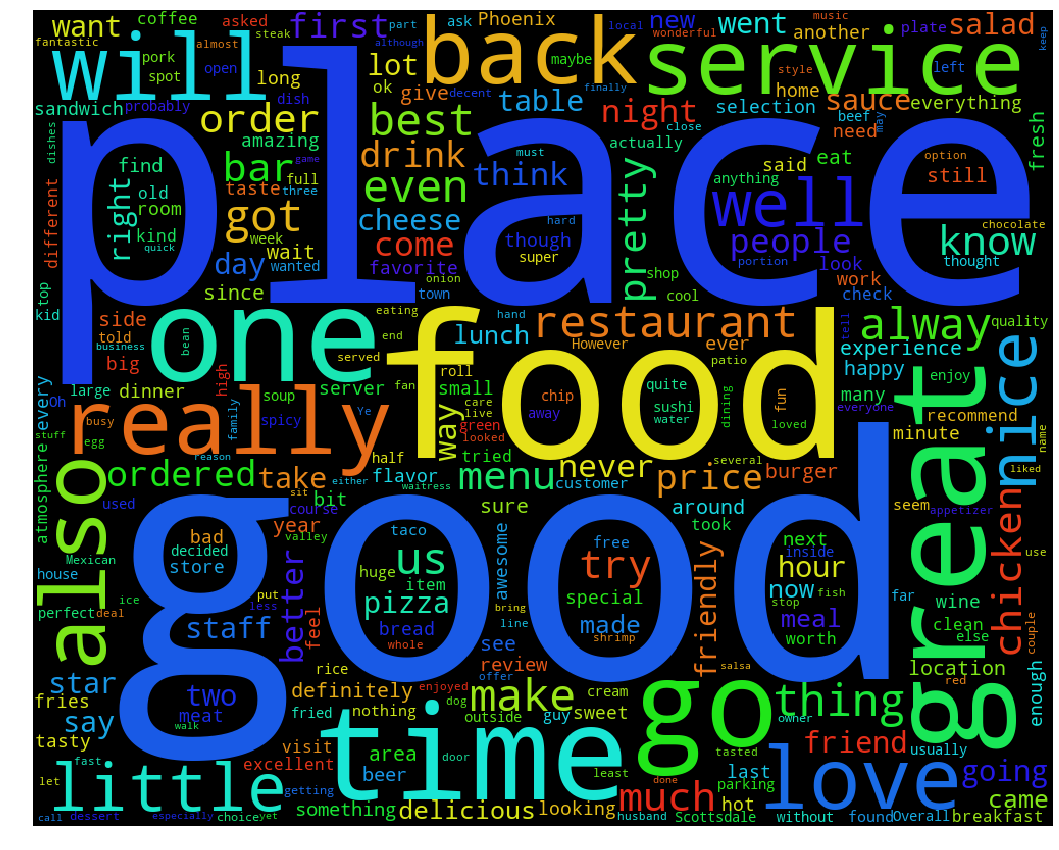
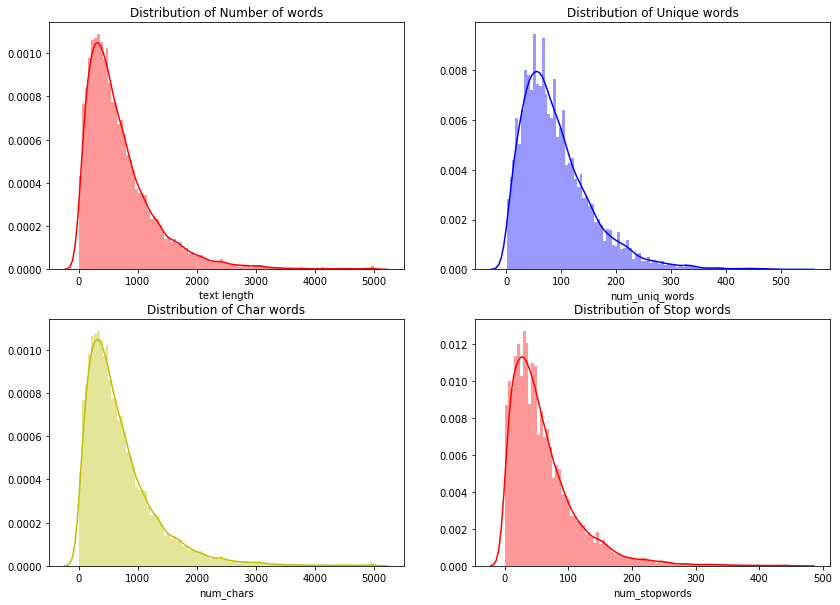
**Additional Analysis:**

Text analytics serves as an important aspect in addition to the numerical-centric analysis, it provides additional information relating to the customer satisfactory, subjective opinion and attitude towards the restaurant. In this section, review data exploration analysis, sentiment analysis and topic modeling are performed to extract further insights.

Review Data Exploratory Analysis:

Unlike numeric data, text data is unstructured. These exploratory data analytics focuses on the word frequency, length of the words, number of characters and number of stop words. Following graphs show the word cloud and density plot of the basic attributes of review data.



****

As shown above, the average number of characters in the reviews data is around 1000 characters, which is equivalent to 100 words. And based on the word cloud, there are several positive adjectives like ‘good’ and ‘great’.

Sentiment Analysis:

One of the assumptions based on the previous data analytics and heuristic is that the data collected is biased towards the positive side. A possible approach to verify this assumption is to use sentiment analysis. Sentiment analysis or opinion mining is a technique with the objective to systematically identify and extract the subjective information in text data. More specifically, the purpose of sentiment analysis is to predict whether the business review data extracted from Yelp is positive or negative.

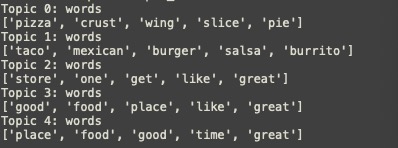
In order to perform sentiment analysis, first step is to remove the stop words like ‘the’, or ‘is’, and then each reviews are mapped into word vectors for the ease of analysis. Multinomial Naïve Bayer classifier is used to predict the subjectivities of the reviews. In order to simplify the model, the reviews data are just trained with two classes labels: positive and negative. Following graphs shows the confusion matrix of the predictive model.



The predictive model achieves an average predicting accuracy of 92%, however, based on above confusion matrix, the model is biased towards positive reviews, which corresponds to the prior assumption.

Topic Modeling:

Another way to extract useful information from text data is topic modeling. Topic modeling is a way to discover the abstract ‘topics’ based on the clustering of similar words. Latent Dirichlet Allocation is used to find the top topics in the business reviews data. Following figures illustrates the output top words in the top topics.





Based on the above plots, though the top topics are hard to interpret, there are some positive adjectives in the top words lists, this suggests that the top topics are inclined towards positive side.