26/8/2017 ABHISHEK KULKARNI

NLP-PROJECT REPORT

It was a good few months where we learnt some concepts and terminologies in Natural Language Processing along with implementing Machine Learning on text data. I have tried to implement some of these concepts in the classification project which has been done.

The data resource from which I acquired the data was Kaggle, the data consists of around 35,000 reviews of hotels which are based in the United States. The data consists of around 15-16 attributes, out of these attributes we will be processing or considering only 2 columns that of the hotel reviews and review ratings. The ratings that we have considered for this project are 1 to 5. Each review has been rated from 1 star to 5 star which in turn has a reflection on the hotel. Our main aim is to predict the rating from the review that we have, or we can say that we are classifying the review from 1 star to 5 star.

First we imported the data using pandas as a dataframe. We have done some exploratory data analysis, we have renamed the columns we are considering for further processing namely ‘reviewstext’ and ‘reviewsrating’. Then we checked the unique ratings that we have, they are ranging from 0 to 10. We have considered only 1 to 5 stars and also eliminated the ratings such as 1.1, 1.2, 2.2 etc. Also we have eliminated the rows which have ‘n/a’ values, that reduced the data by only 22 rows. Still we have significant amount of data for further processing which is around 35,000 rows. Also, we see that the 5 star and 4 star reviews are much more than the 1 star, 2 star and 3 star reviews combined. We can see that the stars are negatively skewed.

As we cannot process text data the way it is, so we converted the data to numeric data using Countvectorizer as well as Tfidfvectorizer where we converted the words to features. We tried various features in countvectorizer like not considering stopwords and considering the n-grams.

Then we applied various algorithms such as Multinomial NB, Linear SVC and SVM on the data to classify the reviews. But we have achieved an accuracy close to 50% in spite of changing various parameters.

After trying these combinations, I wanted to check whether I am getting the same low accuracy if I consider only two reviews i.e. 1 star and 5 star. Then I applied the same techniques that I had applied when we had considered all the 5 stars i.e Countvectorizer and Tfidfvectorizer to convert words to features along with ngrams. What I found was in this I was getting an accuracy of around 94% when I use MultinomialNB for classification, and it reduced if we considered other parameters.

What I have concluded from getting high accurcay for two classes of data and low accuracy for more than two classes of data is maybe the data that we have for predicting 5 classes is not that high, we may need more data to classify 5 classes of data with high accuracy.