Mid Project Report

Varun Gupta - 2018201003
Vatsal Soni - 2018201005
Darshan K - 2018201033
Dhawal Jain - 2018201065

Done task till now

First we collect yelp dataset from Kaggle in csv format.
https://www.kaggle.com/yelp-dataset/yelp-dataset

But Above dataset is of 4 GB and it would not fit into the ram so we took small dataset from below link.
https://www.kaggle.com/omkarsabnis/yelp-reviews-dataset/version/1

SInce dataset contains extra field, we removed fields which are not useful for this problem and then we analyzed data and convert review text in lowercase for further processing.

We found some of the words in review field which are not useful for rating prediction. So we removed it with the help of stop words technique.

We plotted the graph with respect to number of reviews vs rating of each review to know about distribution of the data. Then we came to know that 1 start had lowest number of reviews and 4 star rating had highest number of reviews.

In order to equally distribute review we balanced dataset so that it would not be biased when ML model applies on them.

We converted sentences to token and for that we used n_grams concept which breaks the statement into group of words. We used ngram_range(1,2) which breaks the statement into individual word and group of two words. Why? Because sometime individual word alone does not have its great importance but it become an essential element in combination with other neighbour word.

Also we determined importance of each word by TFIDF technique. This technique gives vector which contains word and number which denotes the importance of word.

Now we have input data which is nothing but the vector which we got it from previous step and we have output data.

We applied different-different ML model and each model gave different accuracy.

Logistic Regression - 52
SVM - 50
Naive Bayes - 32
Neural Network - 28

Also for zomato problem, We tried to get data of restaurant and reviews of each restaurant from zomato API but unfortunately we could not get data so we did it by web scraping.

Resources:

- https://triton.ml/blog/tf-idf-from-scratch
- https://www.datacamp.com/community/tutorials/stemming-lemmatization-python?fbclid=IwAR1KETcp_nU7a44T1-kjcK-Hniz_S2QuUEJys8-HoGZYWk9LFjlGvATEz_0
- http://cs229.stanford.edu/proj2017/final-reports/5244334.pdf

Future plan

- All above task we did for small dataset and now we are trying to do it for large dataset. By doing so we may get good accuracy. For this we will try to use tensorflow queuing to resolve this problem if time permits.

- Since it is observed that even by using some of the proven best algorithms for text processing has result in a significantly small accuracy so we will try to introduce a new feature that will take care of the semantic of the review and thus might help us to get good result.

- If everything goes well for the yelp dataset then we will switch to the zomato problem.