Shalini Mukeshkumar Jain
CWID:- A20405095

Homework - 1
Naive Bayes Classification
Natural Language Processing
CS-585

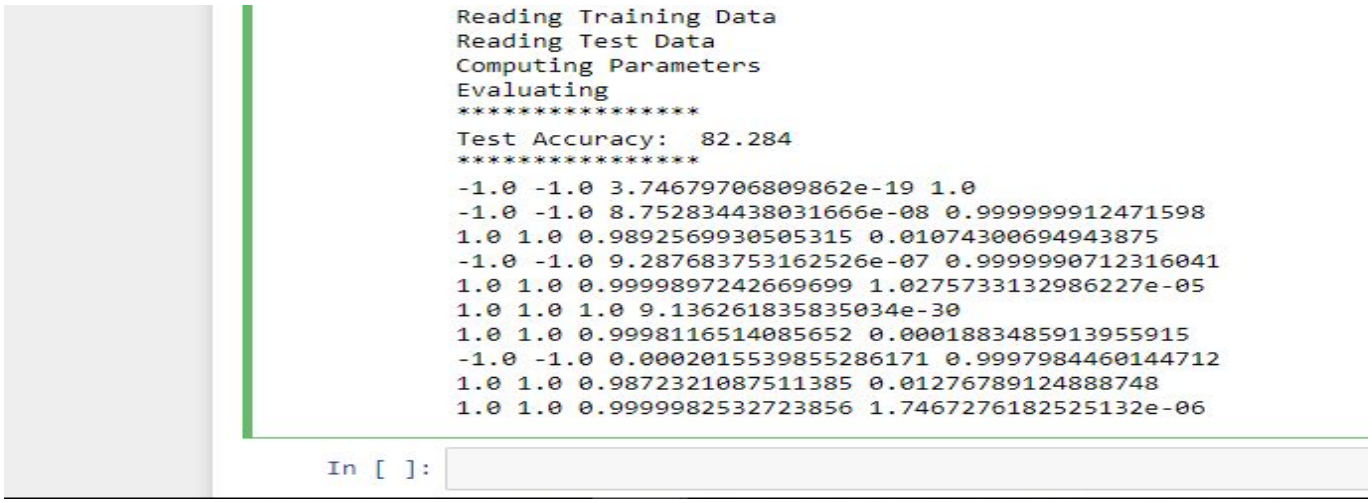
1) Classification and Evaluation

-In this section we are implementing two functions NaiveBayes.Train and NaiveBayes.PredictLabel with various ALPHA values which impacts the Accuracy as shown below:-

ALPHA	ACCURACY
0.1	81.132%
0.5	82.148%
1.0	82.284%
5.0	82.77%
10.0	82.812%

2) Probability Prediction

-In this function we are implemented two methods NaiveBayes.PredictProb and LogSum we calculated the Probability of the word in the given class using logsum exp trick. The screenshot of the first 10 review in the test data is as follows: (where ALPHA=1.0)



3) Precision and Recall:

-In this section we are using Probability Threshold and calculating precision and recall and precision/Recall curve. Here i took Probability Threshold arrays having multiple values=[0.2,0.4,0.6,0.8] and then bases of this we are plotting the curve.

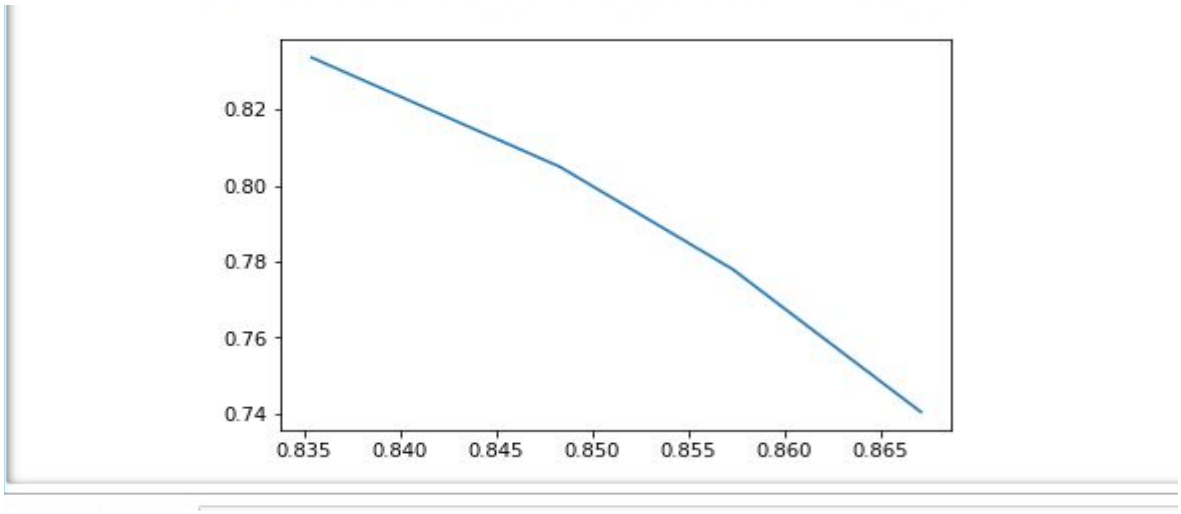
ProbThresh: [0.2 , 0.4 , 0.6 , 0.8]
Precision : [0.8354044736631123 , 0.8483136593591906, 0.8572813822284908, 0.867072599531616]
Recall : [0.8336, 0.80488, 0.778, 0.74048]
Test Accuracy : 81.34 %

The snippet from my code:

```
[0.8354044736631123, 0.8483136593591906, 0.8572813822284908, 0.867072599531616]
[0.8336, 0.80488, 0.778, 0.74048]
Test Accuracy:  0.81348
```

The below snipper is the curve graph:which shows the exponentially decrement relationship

x-axis:-Precision
y-axis:-Recall



4)Features:

-In this we are printing the most positive and most negative 20 words with their weight.
Using the Vocab for wordId and wordweight and then numpy.argsort() to sort as per the most highest wordweight.

20 most Positive words and their weights:

Words-->:edie 4.39585132134 ,Words-->: gundam 4.30255200, Words-->: antwone 4.10414509859 ,Words-->:: yokai 3.84821172445, Words-->/>8/10 3.84821172445, Words--> gunga 3.82715831525, Words--> />7/10 3.82715831525 Words--> />10/10 3.80565211003,Words--> gypo 3.78367320332,Words--> din 3.78367320332,Words--> othello 3.73821082924,Words-->7/10. 3.61459687327,Words--> tsui 3.560529652,Words--> paulie 3.54654341003,Words-->

blandings 3.53235877503,Words--> goldsworthy 3.47351827501,Words--> gino 3.44274661634,Words--> kells 3.44274661634,Words--> />9/10 3.44274661634,Words-->: harilal 3.41099791803

20 most Negative words and their weights:

Words--> />4/10 -4.06604055429,Words--> seagal -4.05722992461,Words--> 2/10 -3.91480958457,Words--> boll -3.9045530844,Words--> uwe -3.89419029736,Words-->*1/2 -3.85163068294,Words--> unwatchable. -3.82965177623,Words-->:thunderbirds -3.76065890474,Words--> />3/10 -3.73656135316,Words--> gamera -3.73656135316,Words--> 4/10 -3.67364752775,Words--> wayans -3.6339071991,Words-->awful! -3.57833734794,Words--> slater -3.48872518925,Words-->/>avoid -3.48872518925,Words--> tashan -3.45697649094,Words--> segal -3.45697649094,Words--> drivel. -3.45697649094,Words--> aztec -3.42418666812,Words--> kareena -3.42418666812