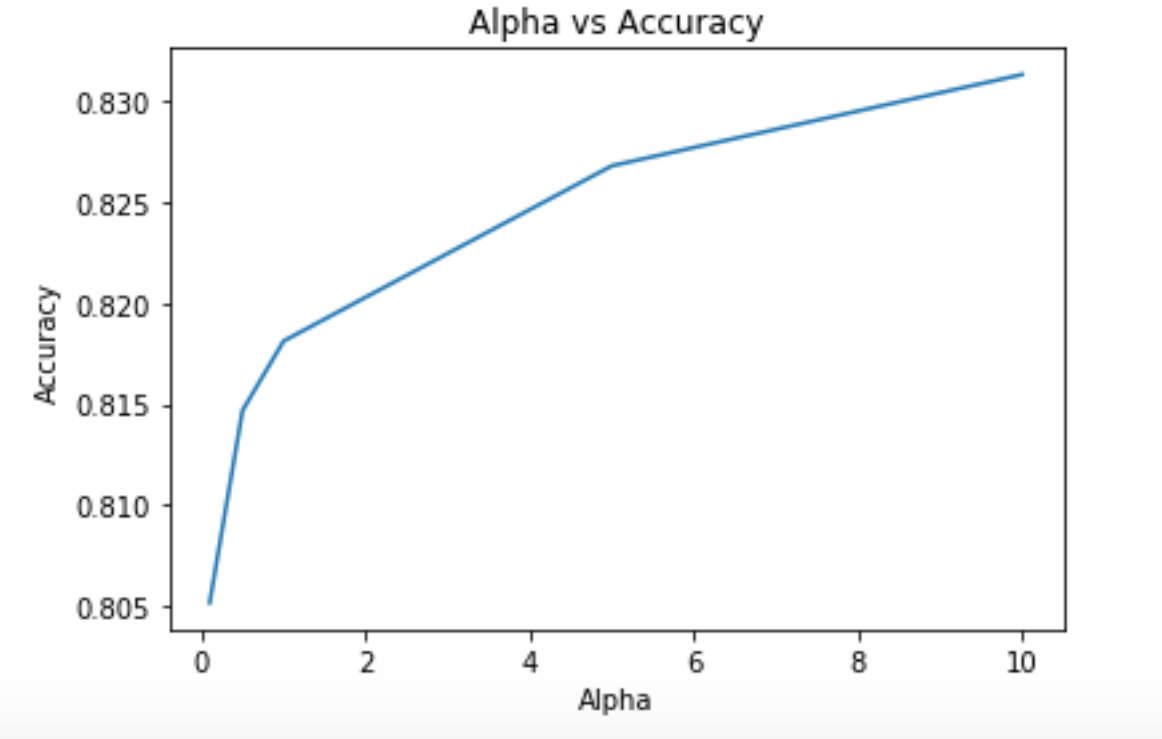
Sagar Mane **A20379756**

Naïve Bayes Text Classification

Classification and Evaluation:

Preprocessing: IMDB data preprocess by removing stop words, punctuation, number and all non-alpha character for better result.

Alpha	Accuracy	Precision	Recall
0.1	0.80516	0.847879616963	0.74376
0.5	0.81468	0.855426041384	0.75736
1.0	0.81812	0.857502472355	0.76304
5.0	0.8268	0.856768558952	0.7848
10.0	0.83132	0.850647701295	0.80376



Above graph represents accuracy for different alpha range from 0.1 to 10.0. In IMDB dataset we get lots of similar words in positive and negative review which sometime miss-predict class label.

Alpha value start increasing, Accuracy also start increasing but at some point, its start decreasing. That point nothing but threshold value for alpha. Module overfit after that alpha value.

Probability Prediction:

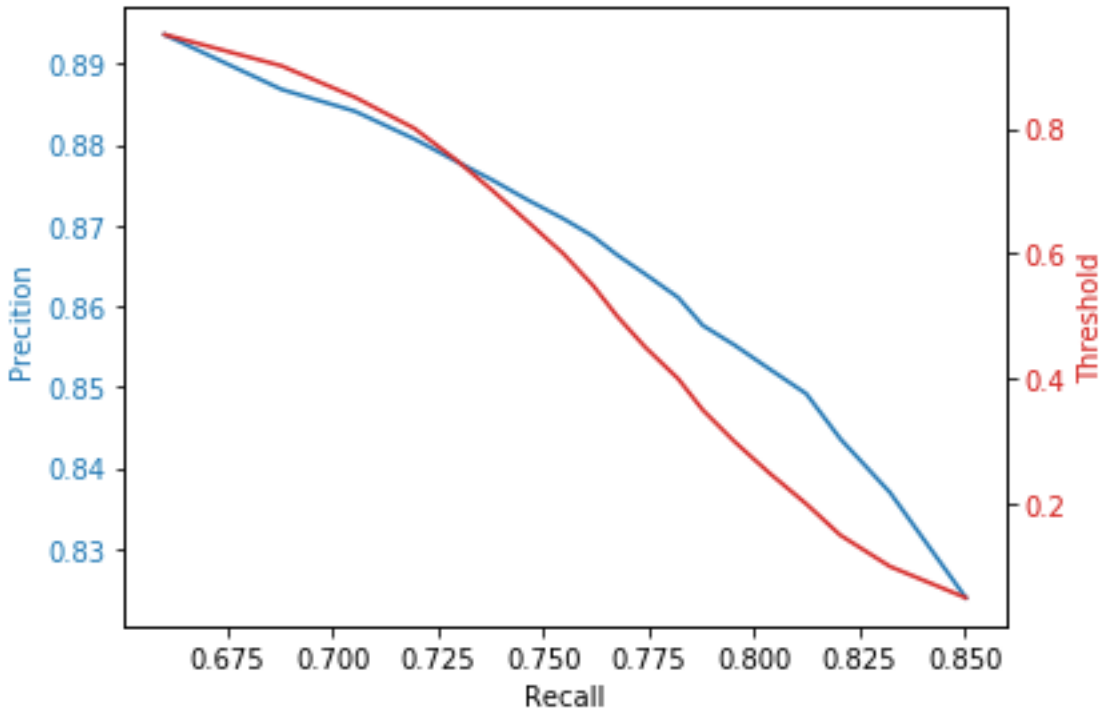
Review	Original Class	Predicted Class	Positive Class Probability	Negative Class Probability
1	- 1.0	- 1.0	0.044265041072575633302	0.95573495892742436669
2	- 1.0	- 1.0	3.01374165975993028e-09	0.99999999698625834029
3	- 1.0	1.0	0.85639397725846045989	0.14360602274153954018
4	- 1.0	- 1.0	0.000349876185763448498	0.99965012381423655149
5	- 1.0	- 1.0	0.11554119143580156655	0.88445880856419843339
6	- 1.0	- 1.0	2.13547316722068667e-13	0.99999999999978645272
7	- 1.0	- 1.0	2.06358679719775622e-18	0.99999999999999999794
8	1.0	1.0	0.99999999998618860749	1.3811392519914501e-11
9	- 1.0	- 1.0	7.80465444683892573e-07	0.99999921953455531609
10	- 1.0	- 1.0	0.049115479855332274097	0.95088452014466772589

Above table represents the log probability of first 10 reviews. We observed that most of the class label are predicted correct as expected but some of them like review no. 3 is miss-predicted because of the most similar words in both positive and negative reviews.

Sagar Mane **A20379756**

Precision VS Recall

Threshold	Accuracy	Precision	Recall
0.05	0.8342	0.8239	0.8502
0.1	0.835	0.8369	0.8322
0.15	0.8341	0.8435	0.8204
0.2	0.834	0.8491	0.8123
0.25	0.8321	0.8522	0.8034
0.3	0.8303	0.8552	0.7952
0.35	0.8285	0.8611	0.7819
0.4	0.8279	0.8611	0.7819
0.45	0.8261	0.8639	0.7742
0.5	0.8245	0.8663	0.7674
0.55	0.8232	0.8687	0.7616
0.6	0.8214	0.8708	0.7546
0.65	0.8189	0.8731	0.7464
0.7	0.8165	0.8756	0.738
0.75	0.8139	0.8779	0.7292
0.8	0.8109	0.8807	0.7194
0.85	0.8064	0.8842	0.7052
0.9	0.8004	0.8868	0.6878
0.95	0.7907	0.8936	0.6601



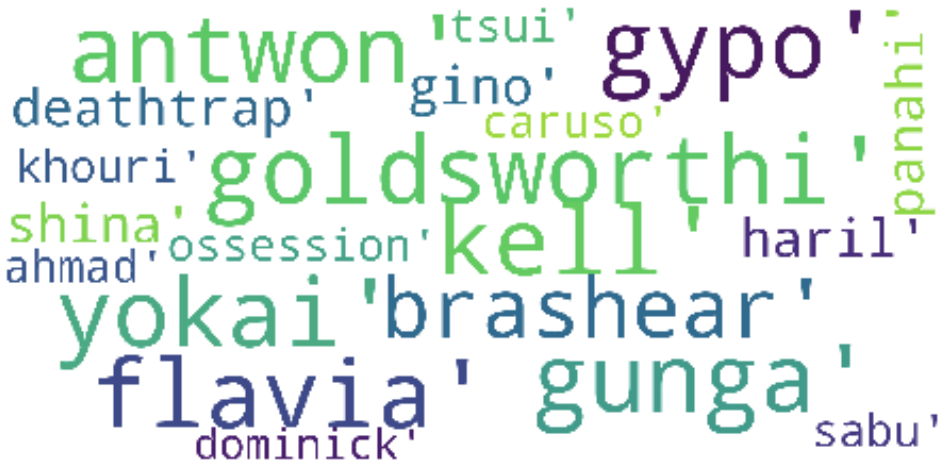
Above plot represents Precision-Recall curve for different threshold values from 0.1 to 0.95. We observed that for high threshold we get more precision i.e True Positive rate is high. If threshold value is less we get low precision i.e. False Positive rate is high.

Sagar Mane **A20379756**

Features:

Top 20 Positive Features/Words

```
antwon 4.45104989718
goldsworthi 4.14089496888
gunga 4.06211409102
gypo 4.06211409102
yokai 4.04558478907
flavia 3.90248394543
kell 3.84306052496
brashear 3.82244123776
deathtrap 3.80138782856
gino 3.77988162334
shina 3.75790271662
haril 3.68890984513
panahi 3.68890984513
ossession 3.64011968096
tsui 3.61480187298
caruso 3.61480187298
khouri 3.61480187298
sabu 3.58882638658
ahmad 3.58882638658
dominick 3.56215813949
```



Top 20 Positive Words

Top 20 Negative Features/Words

```
boll 4.33234633501
slater 3.96078277858
tashan 3.87740116964
kareena 3.76233183986
kornbluth 3.76233183986
thunderbird 3.69941801444
sarn 3.68634593288
delia 3.6596776858
saif 3.60410783464
darkman 3.57512029777
orca 3.54526733462
zenia 3.51449567595
seagal 3.50822606294
hackenstein 3.48274697764
beowulf 3.46648645676
savini 3.44995715481
shaq 3.44995715481
kibbutz 3.41605560314
mraovich 3.41605560314
wayan 3.38096428333
```

Top 20 Negative Words