CS 589 Project 2 Report

1.

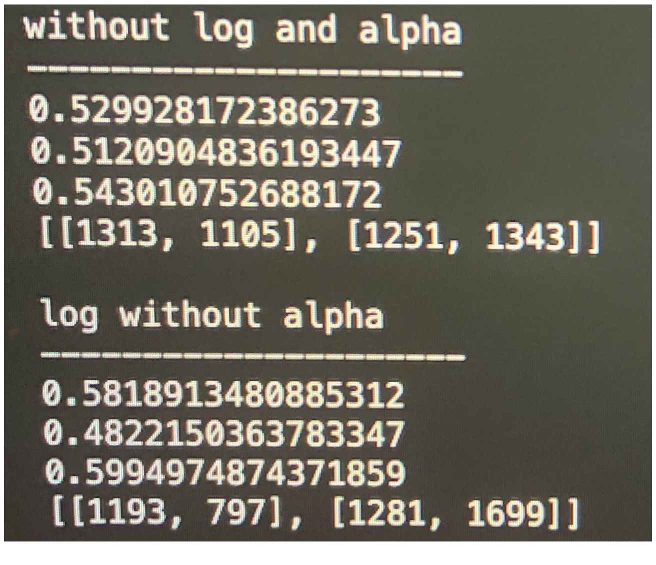
Results for NB with Equation 1

Accuracy: 52.99%
Precision: 51.20%
Recall: 54.30%
Confusion Matrix: [[1313,1105],[1251, 1343]]
Confusion Matrix Format: [[TP, FP][FN, TN]]

Results for NB with Logarithmic Probability

Accuracy: 58.18%
Precision: 48.22%
Recall: 59.94%
Confusion Matrix: [[1193, 797],[1281, 1699]]
Confusion Matrix Format: [[TP, FP][FN, TN]]

Here, accuracy increased, precision decreased and recall increased. This means it does have an effect on the performance of the Naive Bayes model. This improvement results can be attributed to how we are handling the case where the probability of the word occurring goes on to be an extremely small number. With logs we are summing up, this helps in better results.



```
without log and alpha
---------------------
0.529928172386273
0.5120904836193447
0.543010752688172
[[1313, 1105], [1251, 1343]]

log without alpha
---------------------
0.5818913480885312
0.4822150363783347
0.5994974874371859
[[1193, 797], [1281, 1699]]
```

2.

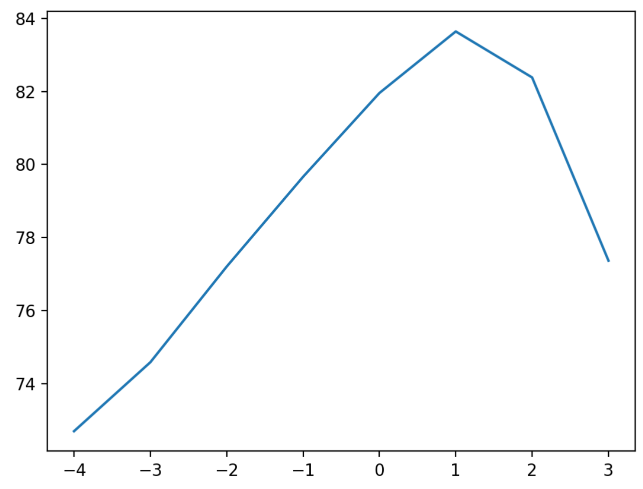
Results for NB with alpha = 1 and Logarithmic Probability

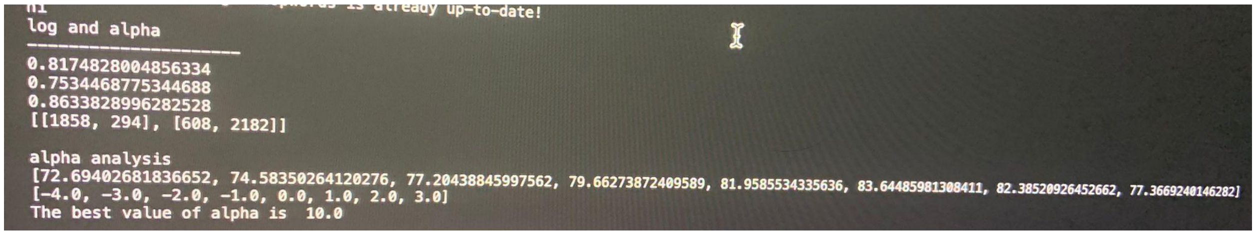
Accuracy: 81.74%
Precision: 75.34%
Recall: 86.33%
Confusion Matrix: [[1858, 294],[608, 2182]]
Confusion Matrix Format: [[TP, FP][FN, TN]]

Results for NB with alpha = 0.0001 to alpha = 1000 and Logarithmic Probability
alpha values for each value of alpha:
[72.69, 74.58, 77.20, 79.66, 81.95, 83.64, 82.38, 77.36]

Alpha Value With Best Accuracy: 10

when alpha is too low it doesn't make much difference to not adding smoothing. So it's the lowest at that point. When alpha is too high, it gives way bias where we are artificially giving that point a very high probability of occurrence. This also leads to making words with high frequencies insignificant.

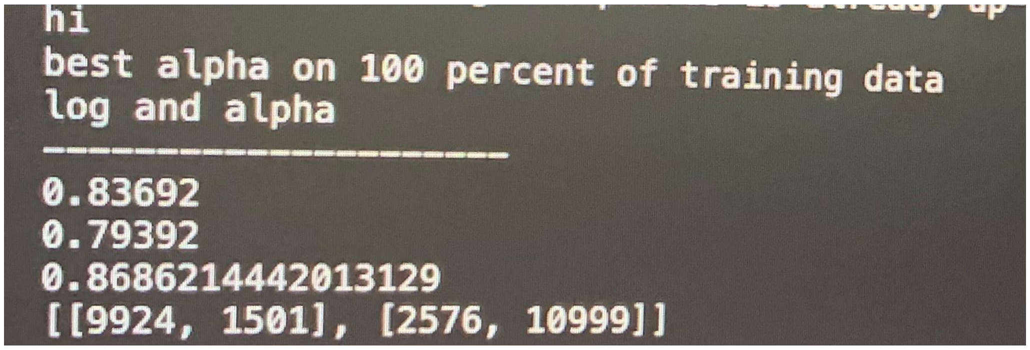




```
hi                                      is already up-to-date!
log and alpha
----------------------
0.8174828004856334
0.7534468775344688
0.8633828996282528
[[1858, 294], [608, 2182]]

alpha analysis
[72.69402681836652, 74.58350264120276, 77.20438845997562, 79.66273872409589, 81.9585534335636, 83.64485981308411, 82.38520926452662, 77.3669240146282]
[-4.0, -3.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0]
The best value of alpha is  10.0
```

3.
<u>**Results for NB with alpha = 10 and Logarithmic Probability**</u>
<u>**for 100% Training Data and 100% Test Data**</u>
Accuracy: 83.69%
Precision: 79.39%
Recall: 86.86%
Confusion Matrix: [[9924, 1501],[2576, 10999]]
Confusion Matrix Format: [[TP, FP][FN, TN]]

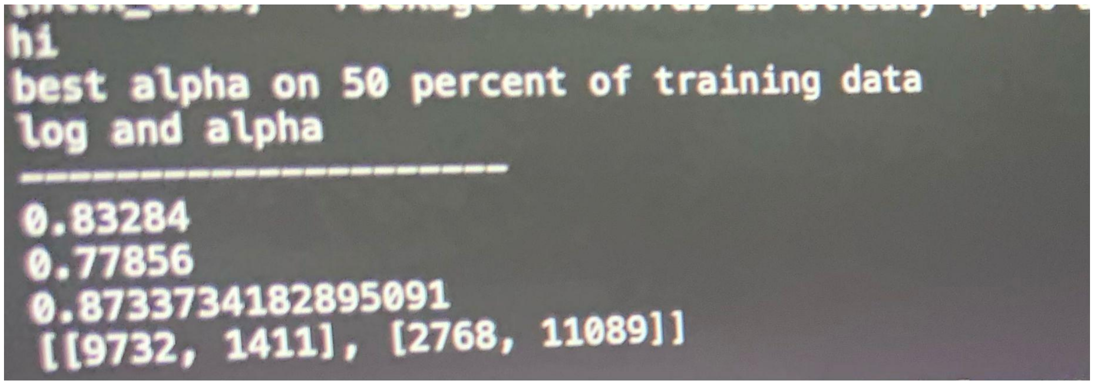


```
hi
best alpha on 100 percent of training data
log and alpha
----------------------
0.83692
0.79392
0.8686214442013129
[[9924, 1501], [2576, 10999]]
```

4.
<u>**Results for NB with alpha = 10 and Logarithmic Probability**</u>
<u>**for 50% Training Data and 100% Test Data**</u>

Accuracy: 83.28%
Precision: 77.85%
Recall: 87.33%
Confusion Matrix: [[9732, 1411],[2768, 11089]]
Confusion Matrix Format: [[TP, FP][FN, TN]]

Just a slight increase in recall and slight decrease in accuracy and precision. But over all there is not much difference in the results.



```
hi
best alpha on 50 percent of training data
log and alpha
--------------------
0.83284
0.77856
0.8733734182895091
[[9732, 1411], [2768, 11089]]
```

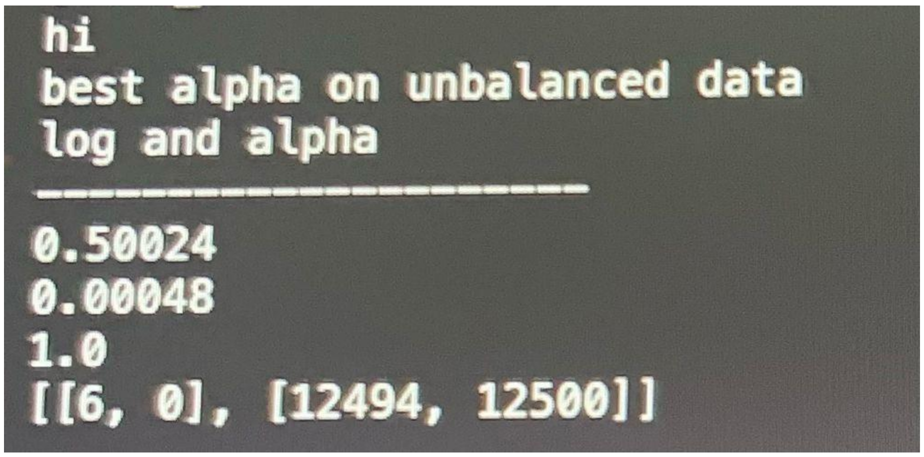
5.

In a movie review dataset, we want to avoid all the cases where the false positives and false negatives. Accuracy is the only metric giving both an equal weightage. Improving the accuracy hence will result in better results.

6.
<u>Results for NB with alpha = 10 and Logarithmic Probability</u>
<u>for 100% Training Data and 100% Test Data</u>

Here, a drastic decrease in the accuracy to 50% and precision to 0% while the recall is at 100 percent. This shows that there is a bias in the dataset. Only 6 were classified as positive. All the others were classified as negative.

Accuracy: 50.24%
Precision: 0.048%
Recall: 100%
Confusion Matrix: [[6, 0],[1249, 12500]]Confusion Matrix Format: [[TP, FP][FN, TN]]



```
hi
best alpha on unbalanced data
log and alpha
--------------------
0.50024
0.00048
1.0
[[6, 0], [12494, 12500]]
```