Assignment 1, NLP

Dhruv Ahlawat, CS1210556

May 7, 2024

1 Model details

Model - Naive bayes estimator (scikit-learn)
Input - combination of 5-grams and 6-grams (characters).

2 Notable technique

The most notable method I used to improve my accuracies can essentially be explained as a form of **label smoothing**. Since it was quite clear that the training data had several incorrect labels as real datasets usually have, I did not just train on the original labels.

after training my model once, since the accuracy was high (91-93%), I used this same model to relabel all the training data, then I increased the stored Naive Bayes weights by a factor of α which was around 1.1, before continuing training again on this modified dataset, ($\alpha > 1$ so there is higher weights to the original training data). I did this process for a number of iterations and fine-tuned the value of α to get the highest final accuracy over the validation set. This increased my accuracy to about 95% on the val set, almost a 2-3% increase.
Note: I did not train the Naive Bayes model again, sklearn also has a feature that lets you "continue" the training and that is what I did. Hence I manipulated its stored counts first before training further.