# Project 2  : Naive Bayes Classifier

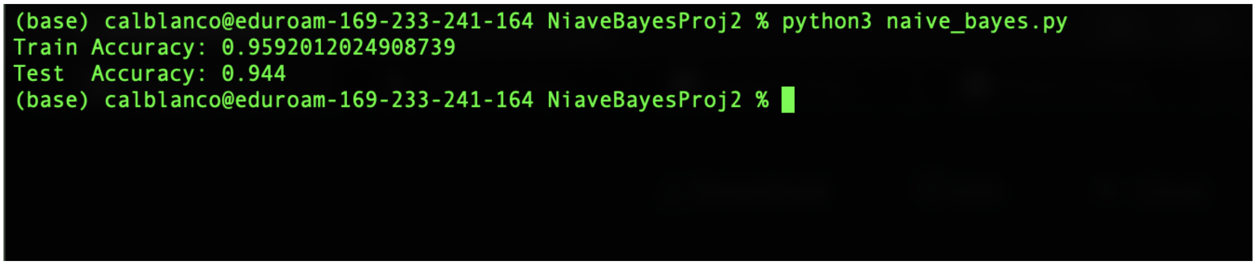
Cal Blanco
wcblanco@ucsc.edu

## How to run/compile program:

The only file I made changes to is naive_bayes.py. All you need to do to run it is open your terminal and type : *python3 naive_bayes.py*

Should be able to work as long as the prerequisites in the given functions are met (have a folder called data with subfolders for training and testing data)

## Expected output:



```
(base) calblanco@eduroam-169-233-241-164 NiaveBayesProj2 % python3 naive_bayes.py
Train Accuracy: 0.9592012024908739
Test  Accuracy: 0.944
(base) calblanco@eduroam-169-233-241-164 NiaveBayesProj2 % █
```

On the given data set, this has been the output for every run I have done.

My program works / I think my program works because I know that my code logically makes sense.

1. The Training data is loaded and scraped for data **[fit()]**
    a. Each file in the training spam/ham set is used to generate a word_set for the file **[load_data(), word_set()]**
    b. Those word sets are then iterated over to generate overall word_count for the spam and ham **[get_counts()]**
2. To predict if a file is spam or ham : **[predict()]** *Logs are used here to prevent underflow with small numbers this switches the probability math from products to sums because log(xy) = log(x) + log(y)*
    a. Get the word_set for the input file/email

b. Use bayes theorem on input words and check word_count dictionaries to determine if the words present in the file are more common in spam or in ham

The whole thing runs without errors which also leads me to believe that the output is fine or  correct.

I feel like this is a very light writeup but based on the prompts I really did not know what else to put in here. I did run this program on some of my own spam emails and it worked!! Which was super cool. This has been an awesome project and it has made me wonder if I could use it to classify other things. Maybe like coming up with how an article might be opinionated. Instead of Ham/Spam have Positive/Neutral/Negative and then basically use the same strategy.