**Mat345 Project 2**

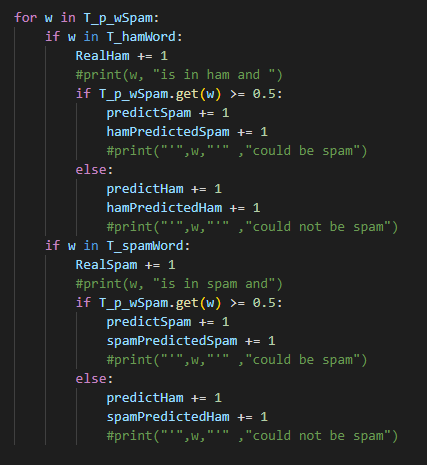
1. Uijin Lee
2. **The data set I used** : Sets A
3. **The tools I used in each part of the project**: Visual code with python
4. **Any decisions specific to your spam filter, such as: words that you might have decided to exclude, threshold for the probability of spam (especially if different than 0.5), etc.**

I made a list of words to ignore that are frequently used words, whether spam or ham.

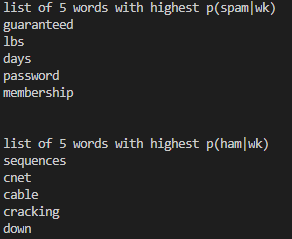
Ex)



And after setting the spam filter, testing words can be recognized by spam if the probability is over 0.5

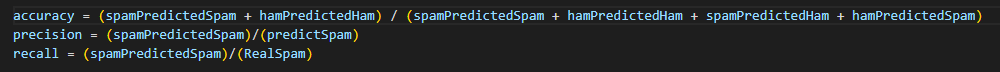


1. **The 5 most ”spammiest” and the 5 most ”hammiest” words from the training stage: answers (c) and (d) from Step 3.**

****

1. **The accuracy, precision and recall rates from the testing stage: answers (a), (b), and (c) from Step 4.**

****

****

1. **Conclusions on the performance of your spam filter and possible steps you would take to improve it**

I have worked hard to build a spam filter through training and testing for this task, but honestly, I am not sure if I made it correctly. Accuracy and precision have a fairly high probability, but to improve this, more data and more accurate threshold values seem to be set.