AUCSC 460
Homework #2
**Due date: March 29 by 11:59 pm**
**(The late submission policy is the same as homework #1.)**

# Spam filtering using Bayes model

## 1 Description

In this assignment, you will classify an email message as 'spam' or 'ham' using Bayes model. The Attached pdf file (asn2_bayes_model.pdf) shows how to use Bayes model for this assignment.

## 2 Dataset

If you unzip the data file, there are two folders: 'Train' and 'Test'.

In 'Train', there are two sub-folders: 'train_Original' and 'train_Lemmatized'. 'train_Original' folder includes the original emails, and 'train_Lemmatized' folder includes the modification of the original emails by lemmatizing words (e.g. 'boys' → 'boy', 'stopped' → 'stop') and removing functional words (words that do not have meanings such as 'a', 'the', and 'from'). To count the number of occurrences of words, you have to use the folder 'train_Lemmatized'. If you would like to see the original email, you can access the folder 'train_Original'. But, you do not need the 'Original' folder in this assignment.

In the folder 'train_Lemmatized', there are 259 email messages (220 legitimate messages and 39 spam messages). The files of which names start with 'spm' are spam messages. You will calculate probability of each word's occurrence in each spam, and ham group, and store the probability in the file (See the file asn2_bayes_model.pdf to learn how to compute probability).
You have to create two files: 'probability_spam_words.txt', and 'probability_ham_words.txt'. Then you have to write each word's probability in each group in the corresponding file. The format of writing probability in each file is up to you. At the start of each file, please write comments that explain the format you used to write each word's probability.
Please note the probability of a word cannot be zero. Refer to asn2_bayes_model.pdf for more information.
You will construct hw2_prob.py. In this python file, you will write the code which computes each word's probability in each group (ham and spam) and store all the words' probabilities in the two files ('probability_spam_words.txt', and 'probability_ham_words.txt').

The files ‘probability_spam_words.txt’, and ‘probability_ham_words.txt’ will be used when you classify new(unseen) messages in the folder ‘Test’.

## 3 Classification

In ‘Test’, there are two sub-folders: ‘test_Original’ and ‘test_Lemmatized’. ‘test_Original’ is only for your reference in case that you want to see the original email. You will use ‘test_Lemmatized’ folder for classification.
In ‘test_Lemmatized’, there are 30 emails. Using the probabilities that you already calculated, classify each email message as ‘spam’ or ‘ham’.

## 4 Classification result

I will execute your program through the following command

**python hw2_classify.py 1.txt output1.txt**
**python hw2_classify.py 2.txt output2.txt**
….

Then your program will compute the Bayesian model probabilities, and the output should be written in output<number>.txt. (Here, number is the test file number). Following is the example of the output file.  (Round off 8 decimal places.)

(1) P(Spam, all words)
P(Spam) = 0. 23455445
P('you'|Spam) = 0.00000234
P('lose'|Spam) = 0.00000023
P('weight'|Spam) = 0.00000094
P('you'|Spam) = 0.00000234
……..
log P(Spam, all words) = -49.2300
(2) P(Ham, all words)
P(Ham) = 0.76544555
P('you'|Ham) = 0.00000837
P('lose'|Ham) = 0.00000007
P('weight'|Ham) = 0.0000018
P('you'|Ham) = 0.00000837
……..
log P(Ham, all words) = -68.2797

**Conclusion: This message is classified as Spam.**

*<Example of output1.txt>*

## 5 Deliverables

1. **Submit zip file including 35 files:**
   1) **hw2_prob.py,**
   2) **hw2_classify.py,**
   3) **probability_spam_words.txt,**
   4) **probability_ham_words.txt,**
   5) **output1.txt ,**
   6) **output2.txt…**
   **…and**
   **34) output30.txt.**
   35) evaluation.pdf
   Please measure the accuracy and explain your performance in evaluation.pdf. In the attached data.zip file, you can see evaluation.txt which shows correct answers.
   (e.g., My classification performance is 30/30.. (100%) or my classification performance is 25/30. I misclassified two 'ham' messages as 'spam', and three 'spam' messages as 'ham', etc.)