Project 4
Building a Spam Filter using a Naïve Bayes Classifier
Due before midnight on April 13, 2022

Problem Description and Data Set

Project 4 is to build a Naïve Bayes Spam filter. You will be able to download a labeled training set file and a labeled test set file from Canvas. Both files will have the same format. Each line will start with either a 1 (Spam) or a 0 (Ham), then a space, followed by an email subject line. A third file will contain a list of Stop Words—common words that you should remove from your vocabulary list. Format of the Stop Word list will be one word per line.

Assignment

Your program should prompt the user for the name of a training set file in the format described above and for the name of the file of Stop Words. Your program should create a vocabulary of words found in the subject lines of the training set associated with an estimated probability of each word appearing in a Spam and the estimated probability of each word appearing in a Ham email. Your program should then prompt the user for a labeled test set and predict the class (1 = Spam, 0 = Ham) of each subject line using a Naïve Bayes approach as discussed in the class videos. Note: We may or may not test your program on the same files that you used to create it!

Output to the screen of your program should include:

- How many Spam and Ham emails were in the Test set file that was read in.
- Number of False Positives, True Positives, False Negatives and True Negatives that your spam filter predicted.
- Accuracy, precision, recall and F1 values for your Spam filter on the Test Set file.

What to turn In Via Canvas

You are required to submit a project report, including:

- A brief introduction of your model.
- Number of False Positives, True Positives, False Negatives and True Negatives that your spam filter predicted.
- Accuracy, precision, recall and F1 values for your Spam filter on the Test Set file.
- Screenshot of your python console.
- A copy of your code

Your report should be named yourlastname_yourfirstname_P4.docx or .doc or .pdf. Your Python program should be named yourlastname_yourfirstname_P4.py, then zipped together with your project report and uploaded to Canvas

Notes and Suggestions

- If your program has problems with reading in the files, try opening the files like this: file = open(filename, "r", encoding = 'unicode-escape')
- Use the training file to figure out the percentage of emails expected to be Spam.
- You will probably have to use the natural log format of Bayes equation to avoid computer precision problems.
    - So instead of multiplying a lot of probabilities together, we can add their logs, then raise e to the power of the final sum.
        - Total probability = 0.8*0.0001*0.002*0.9
        - Total probability = $e^{\ln(0.8) + \ln(0.0001)+\ln(0.002)+\ln(0.9)}$
    - 2. Then use

$$\frac{1}{1 + e^{\ln(P(F|\neg E)P(E)) - \ln(P(F|E)P(E))}}$$