Shawn Davidson
Chris White
CSCI 404 – Fall 2018
Assignment 4

Write-Up:

First, we parsed the training data set and created a dictionary of the 2500 most common words that appeared in the corpus. Dictionaries of the same size were also created for spam and non-spam documents. Then, for the spam and non-spam dictionaries, the probability of each word occurring was calculated and stored in a set. After creating the dictionaries, the test documents were classified using the naïve bayes model, where the probability of the document being spam was compared to the probability of it not being spam.

Probability was calculated using:

C = {Spam, Not Spam}

$$\hat{c} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

- If a document was supposed to be spam and was calculated as spam, it was a true positive.
- If a document was supposed to be spam and was calculated as not spam, it was a false positive
- If a document was supposed to be not spam and was calculated as not spam, it was a true negative
- If a document was supposed to be not spam and was calculated as spam, it was a false negative

By running this program with a different amount of training documents we noticed that the overall F-score went down with fewer training documents. This is due to the fact that our programs computed a lot more documents as false negatives (documents calculated as spam that were not spam).

Test data set: 260 documents Training data set: 700 documents	Correct	Not Correct
Spam	tp = 128	fp = 2
Non-Spam	fn = 29	tn = 101

Precision= 0.984615384615

Recall= 0.815286624204

F Score= 0.891986062718

Shawn Davidson
Chris White
CSCI 404 – Fall 2018
Assignment 4

Test data set: 260 documents Training data set: 400 documents	Correct	Not Correct
Spam	tp = 128	fp = 2
Non-Spam	fn = 43	tn = 87

Precision= 0.984615384615

Recall= 0.748538011696

F Score= 0.85049833887

Test data set: 260 documents Training data set: 100 documents	Correct	Not Correct
Spam	tp = 130	fp = 0
Non-Spam	fn = 102	tn = 28

Precision= 1.0

Recall= 0.560344827586

F Score= 0.718232044199

Test data set: 260 documents Training data set: 50 documents	Correct	Not Correct
Spam	tp = 128	fp = 2
Non-Spam	fn = 105	tn = 25

Precision= 0.984615384615

Recall= 0.549356223176

F Score= 0.70523415978

Shawn Davidson
Chris White
CSCI 404 – Fall 2018
Assignment 4