Assignment 8

Ryan Dill
CS 432
April 17, 2018

1. Create two datasets; the first called Testing, the second called Training.

The Training dataset should:

a. consist of 10 text documents for email messages you consider spam (from your spam folder)

b. consist of 10 text documents for email messages you consider not spam (from your inbox)

The Testing dataset should:

a. consist of 10 text documents for email messages you consider spam (from your spam folder)

b. consist of 10 text documents for email messages you consider not spam (from your inbox)

Upload your datasets on github

For this part I copy-pasted the contents of my emails into .txt files. This seemed like the easiest way to accomplish this task. I used my personal email that I give out when filling out on-line forms, since I know its going to be full of spam. I divided the emails into two folders, Test and Train. Train contains NotSpam(01-10) and Spam(01-10) and Test contains NotSpam(11-20) and Spam(11-20)

2. Using the PCI book modified docclass.py code and test.py (see Slack assignment-8 channel)
Use your Training dataset to train the Naive Bayes classifier (e.g., docclass.spamTrain())
Use your Testing dataset to test (test.py) the Naive Bayes classifier and report the classification results.

For this part I used the code from https://github.com/arthur-e/Programming-Collective-Intelligence/blob/master/chapter6/docclass.py and https://cs532s18.slack.com/files/U8K4TSGJ1/F9Z33U1B6/test.py Below is a snippet of the code used.

```
def getwords(doc):
  splitter=re.compile('\\W*')
  words=[s.lower() for s in splitter.split(doc)
          if len(s)>2 and len(s)<20]

  # Return the unique set of words only
  toreturn = dict([(w,1) for w in words])
  return toreturn
```

Unfortunately, I have not been able to get this code working as intended. I will return to and update it if I have time before the end of the semester.