Part 1

Wrote a script by using and modifying code provided in slack.

#traintest.py

**import** sys

**import** docclass

**import** os

cl **=** docclass**.**naivebayes**(**docclass**.**getwords**)**

#remove previous db file

#check\_output(['rm', 'testobj.db'])

cl**.**setdb**(**'testobj.db'**)**

docclass**.**spamTrain**(**cl**)**

#train

#for file in folder train

**for** file **in** os**.**listdir**(**'trainspam'**):**

sys**.**stdout**.**write**(**file**)**

#read file, train

infile **=** open**(**'./trainspam/' **+**file**,**'r'**)**

cl**.**train**(**infile**.**read**(),** 'spam'**)**

infile**.**close**()**

**for** file **in** os**.**listdir**(**'trainnotspam'**):**

sys**.**stdout**.**write**(**file**)**

#read file, train

infile **=** open**(**'./trainnotspam/' **+**file**,**'r'**)**

cl**.**train**(**infile**.**read**(),** 'not spam'**)**

infile**.**close**()**

#test

**for** file **in** os**.**listdir**(**'testspam'**):**

sys**.**stdout**.**write**(**file **+** " "**)**

#read file, train

infile **=** open**(**'./testspam/' **+**file**,**'r'**)**

**print(** cl**.**classify**(**infile**.**read**())** **)**

infile**.**close**()**

**for** file **in** os**.**listdir**(**'testnotspam'**):**

sys**.**stdout**.**write**(**file **+** " "**)**

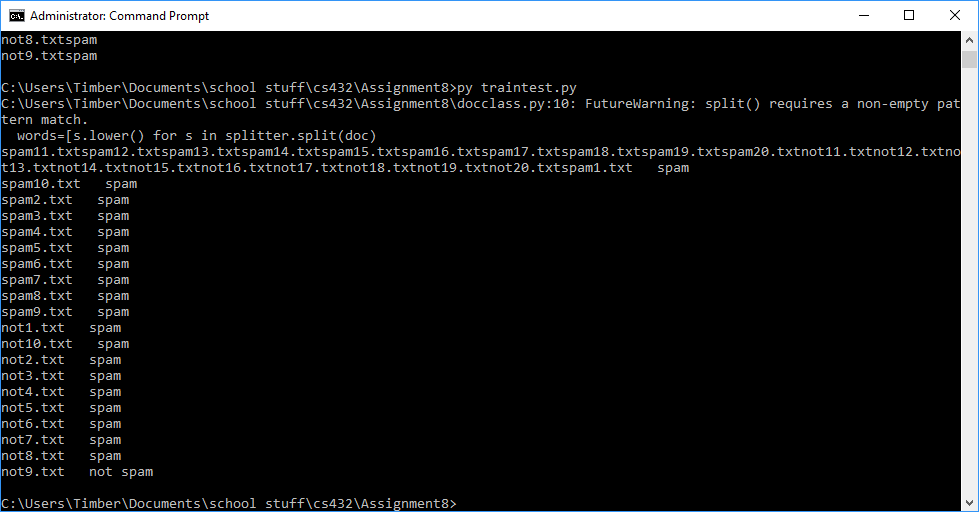
#read file, train

infile **=** open**(**'./testnotspam/' **+**file**,**'r'**)**

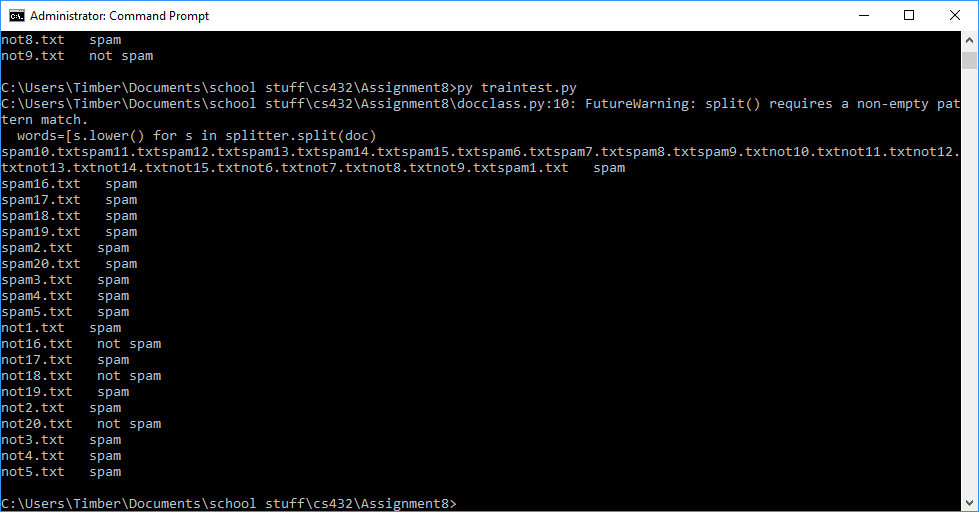
**print(** cl**.**classify**(**infile**.**read**()))**

infile**.**close**()**

The datasets used for spam e-mails consist mostly of newsletter type messages. The content of not-spam dataset is mostly responses to support tickets and purchase conformations and receipts. When running the script, the algorithm tends to classify most of the e-mails as spam. It correctly identifies all spam as such, but will false-positive most of the not spam e-mails as spam.



As a check, I swapped 5 items from each testing and training set and re-ran the script. Swapping items between testing and training sets did not improve performance much.



My guess for why performance is poor is that most of my e-mails are very similar. I Have never used e-mail for personal communication, so the entirety of by inbox consists of messages from organizations and businesses, so the structure of the e-mails will be more similar than if they were compared with personal e-mails.

Confusion Matrix for first iteration:

|  |  |  |
| --- | --- | --- |
|  | Spam | Not Spam |
| Spam | TP: 10 | FN: 0 |
| Not Spam | FP: 9 | TN: 1 |

Precision: 10/19 Recall: 10/10

Second iteration:

|  |  |  |
| --- | --- | --- |
|  | Spam | Not Spam |
| Spam | TP: 10 | FN: 0 |
| Not Spam | FP: 7 | TN: 3 |

Precision: 10/17 Recall: 10/10