Homework3

Allison Fenichel

UNI: alf2178

Files:

hw3.py

hw3p4i.py

hw3p4i.py

enron1/ham folder

enron1/spam folder

Word stemming: using porter stemming from python “stemming” package <https://pypi.python.org/pypi/stemming/1.0>

Part i (Bag of Words):

to run: go to command prompt. Change directory to folder where python files are saved. Type the following line of code to run algorithm:

**python hw3p4i.py m**

m =the number of emails in the training set from ham and from spam (i.e. if m=100 then 100 spam emails are scanned, and 100 ham emails are scanned). Must be an integer<1500.

Returns “word\_bag” which is matrix of bag of words for all training x.

example code to invoke part i

python hw3p4i.py 100

Part ii (Classifier):

to run: go to command prompt. Change directory to folder where python files are saved. Type the following line of code to run algorithm:

**python hw3p4ii.py i m**

m =the number of emails in the training set from ham and from spam (i.e. if m=100 then 100 spam emails are scanned, and 100 ham emails are scanned). Must be an integer<1500.

i=string input for filename containing email (i.e. i=4103.2001-03-29.farmer.ham.txt)

example code to invoke part ii

python hw3p4ii.py 4103.2001-03-29.farmer.ham.txt 10

Returns guess value of y given input value i on a training size defined by m