

Assignment 1
Cogsci 188 Summer Session 2019
Creating Feature Vectors
Due August. 20st Monday at 6:00 PM

Motivation:

Creating feature vectors from text documents using bag of words format and word counts.

Task:

You will write one main function for this assignment as follows:

```
def createFeatureVectors(dirName, vocabFile)
```

This function takes an input of directory name and vocabulary filename `vocabFile` as strings . It writes two files in the current directory. First file is `vocabFile` and second is `<dirName>_fVectors.txt` (where `< dirName >` is replaced by the actual input string passed in for the directory Name).

Filenames in the `dirName` have the following format:

`<filename>_<Label>.txt`

Where `< Label >` is the actual stars given to that particular review on imdb. In our feature vector file, the first number in each line will be the label of the feature vector (review stars of the file). All the files, positive and negative reviews are in one same directory. Your function only looks at files in one directory.

Please note that this is a little different than what was discussed in the lecture on Aug. 6rd.

On top of your python files, please write your name and your partner's name (if you are working in groups of 2). Your comments will look as follows:

```
#Assignment 1. File createFeatureVectors.py
#Student 1 Name: <First and Last Name of 1st group member>
#Student 2 Name: <First and Last Name of 2nd group member>
#<Student 1> and <Student 2> attest that this assignment was done
by them two and reflects their original work and based on their
understanding of the concepts. Both students have equally
contributed to the solution of this assignment.
```

Then you will have your own methods as you need to declare the methods before you can use them, so your smaller functions will go first in the file. First you will need to import the “os” library. So your code following the comments will look as follows:

```

import os
import os

#cleanup method cleans up the intext string of punctuation, numbers
and stop words etc. and returns a lowercase string
def cleanup(intext):
    intext = intext.replace("!", " ExclamationMark ")
    intext = intext.replace "?", " QuestionMark ")
    for mark in string.punctuation:
        #remove punctuations
    ...

```

Your next method may be a method that only writes the vocab file from a given directory. So it may look something like this:

```

def directory2Vocab(dirName, outputFileName):
    vocabSet = set()      #start with an empty set
    #open a new file, vocab.txt, in the "w" mode.
    fileName = outputFileName
    vocabFile = open(fileName, 'w')
    allFiles = os.listdir(dirName)
    for review in allFiles:
        ...
        ...

```

Your next method may be the method makes the feature vectors given a vocab file. So it may look something like this:

```

def directory2features(dirName, vocabfilename):
    # open and read the vocabfilename
    # put each word in vocabfile as an element in a list.
    vocFile = open(vocabfilename, 'r') #Open the vocab file
    vocabWords = vocFile.read().splitlines()      ...
    ...

```

Now your final method will only need to call your earlier two methods, so it will look something as follows:

```

def createFeatureVectors(dirName, vocabFile):
    directory2Vocab(dirName, vocabFile)
    directory2Features(dirName, vocabFile)

```

Starting Assignment 1

The assignment 1 submission is open now. Here is a step by step instruction on how to submit:

1. login to your student account at datahub.ucsd.edu
2. spawn your server under cogs188
3. click the Assignments tab
4. Fetch Assignment1
5. You can now access the createFeatureVectors notebook under Assignment 1 under the downloaded assignments section.
6. Put your name and your partner name at the top by editing the second cell.
7. Fill in the function and run the simple tests provided.
8. You can't edit the test cell, they are read-only.
9. Once you are ready, you can click the validate button at the top.
10. After you successfully validate your functions, you can go back to the Assignments tab and click submit.

Example test cases

We have provided you 2 example files:

1. Vocab_Example.txt
2. MovieReviews_Example_fVectors.txt

Vocab_Example.txt file is generated by running
directory2Vocab("MovieReviews_Example", "Vocab_Example.txt")

MovieReviews_Example_fVectors.txt is generated by running
directory2features("MovieReviews_Example", "Vocab_Example.txt")