Assignment 1
Cogsci 188 Summer Session
Creating Feature Vectors
Due August. 21st Monday at 2:00 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)
```

This function takes an input of directory name as a string. It writes two files in the current directory. First file is Vocab.txt and second is fvectors_<dirName>.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. 3rd.

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 …

   for mark in ["0","1" ……]:
```

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.
   allFiles = os.listdir(dirName)
   for f in allFiles:
     #extract label from the file name, f.
     #open file in read format
     #read file as a string
     #clean up the string by calling cleanup method
      …
      …
```

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

```
def directory2Vocab(dirName):
   vocabSet = set()   #start with an empty set
   #open a new file, vocab.txt, in the "w" mode.
   allFiles = os.listdir(dirName)
   for f in allFiles:
     #open file in read format
     #read file as a string
     #clean up the string by calling cleanup method
     #split the cleaned strings into words in a list
     #add the set made from word list into the vocabSet set
     #write the contents of the vocab set in the vocabfile
      …
      …
```

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

```
def createFeatureVectors(dirName):
   directory2Vocab(dirName)
   directory2Features(dirName, "vocab.txt")
```

Assignment submission details

The assignments can be done in groups of up to two students. Following is the turning process. Submit the assignment on Ted.ucsd.edu by uploading both the files and pasting the comment part of your assignment in the textbox.