COMPONENTS:

My project is separated into three parts:
1) hw5.py: is used to build the TFIDF vectors for categories and for 20000 article objects, and pickle and store them under the sub dir of my folder.
2) tester.py: is used to classify some particular articles and print out the correct rate.
3) hCluster.py is used to combine similar categories and print out the hierarchical clustering.

IMPLEMENTATION DETAILS:

1) I remove stop words and punctuations. My stop words list comes from the first link of Wikipedia for stop words:
http://en.wikipedia.org/wiki/Stop_words
http://norm.al/2009/04/14/list-of-english-stop-words/
2) I also remove email address with regular expression.
3) I treat WF as TF.
4) I correctly rate is very good, in the range of 85%-95%

HOW TO RUN:

Since I have upload all the pickled files, (including TFIDF vectors for categories and for 20000 article objects list), you don't need to run the hw5.py to build the files. But if you really want, you can run:
"python hw5.py #pathFiles# #pathPickle#" in which, #pathFiles# stands for the location for "20_newsgroups" directory is("/Users/xo/Downloads/20_newsgroups/", end with "/"; #pathPickle#" is for where you want to pickle the vectors)

You can directly run the "python tester.py 50", here 50 is the count of article, to get the correct rate.

And you just need to run "python hCluster.py #path#" to get the hierarchical clustering. You should give the path of the pickled TFIDF vectors (i.e. where you store the "pickleTFIDF/" directory. For example, you store it in "/Users/xo/Study/2013Fall/AI/homework5/pickleTFIDF/", then you should write "/Users/xo/Study/2013Fall/AI/homework5/", with the "/" at the end)