**Assignment 5**

Student Name: Aghor, Mrunalini Vijay

UTA ID: 1001362150

Project No.: Home Work Assignment 5

Due date: April 15, 2017

**Steps for Execution:**

1. **Put the folders arxiv, jdm and plos along with the python file. These folders are inside articles folder. Put ‘stoplist.txt’ in the same folder.**

2. Open command prompt on the same location and type this command,

**python final1.py**

3. Wait for the processing of the training step to finish.

**Approach followed:**

1**) Pre- Processing step:**

**Building vocabulary and removing stop words** :

The first step of my project is to form a vocabulary. I have taken 50 % data from all the three folders (JVM, ARXIV, PLOS) for forming the vocabulary.

I have used bag if words approach for forming the vocabulary. I am storing all the words from training set into a list along with its frequencies.

So my dataset has (key, value) pairs of words in the training text files of all the classes along with their probabilities. Probabilities are calculated in the following way:

If a word appears 3 times in a document and the total number of words in that document is 10. Probability of that word will be 3/10.

I have removed all the **stop words** from the vocabulary which are given in the stoplist.txt file.

Example of how vocabulary looks like:

{'specialized,': 6.303183107469272e-05, 'barrages': 6.303183107469272e-05, 'localizes': 6.303183107469272e-05, 'reconstruction,': 6.303183107469272e-05…………………………….}

**Convert training data into features:**

In this step I have converted the training data into features. I am reading contents of one file from one folder. Comparing each word in that file to the vocabulary. In the vocabulary, if that file word is present I save value as ‘1’ on that position and ‘0’ if the word is not there.

I am also appending ‘A’ , ‘J’ or ‘P’ at the end of each vector, and these letters denote if it belongs to class Arxiv, Jdm or Plos.

Example of feature vector will look like this:

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 'A']

This we get a vector like this for all the files in the training data set.

**2) Classification step:**

In classification step we classify if the text document belongs to class ARXIV, JDM or PLOS. It is done using our training data.

I have made the same feature vector for testing as we had made in training. This time I have included the word probability from the vocabulary as well. I compared all words of a single test file with vocabularies of all the three classes. If I found a match I stored it as the probability of that word, if I dint find a match I stored it as 0.

Likewise I had three vectors (one for each class) for a single file. Then I added all the values in each vector. Now I had total probability of that text file with each class. I found the largest among the three probabilities and classified that document in that class.

For example if I get three values of addition of all values in a vector of a document.

For ARXIV = 0.7

For JDM = 0.4

For PLOS = 0.6

I classified the file in folder ‘ARXIV’. This same procedure was followed for all the files of all the classes.

**Accuracy:**

After classifying all the documents from testing data, I calculated the accuracy of my classification in percentage.

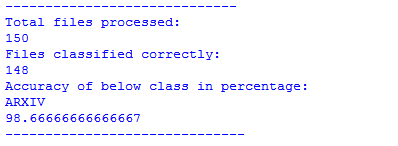
It was calculated by this formula:

**Accuracy = (Number of files correctly classified \* 100)/ (total number of files classified)**

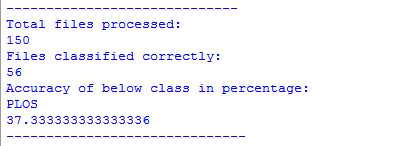
**RESULTS:**

These are the results I got when I implemented the classifier, I have attached images of the actual output.

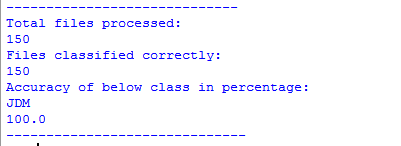
**For class ARXIV: (Accuracy : 98.66 %)**

****

**For class PLOS: (Accuracy = 37.33 %)**

****

**For class JDM: (Accuracy = 100%)**

****