Rushikesh Dixit

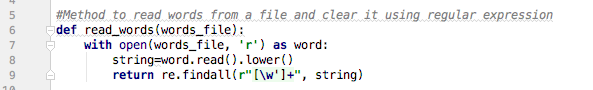
1001434559

Machine Learning

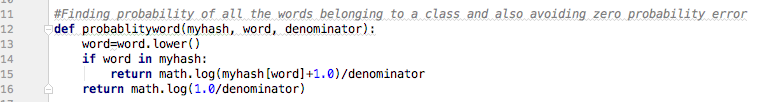
Assignment 2

**Initial Setup**

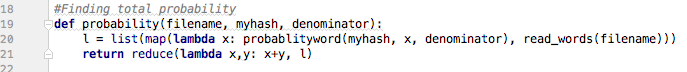
The Naïve Bayes Classifier is implemented using python on the 20 newsgroups dataset using half as the training data as well as half as test data.



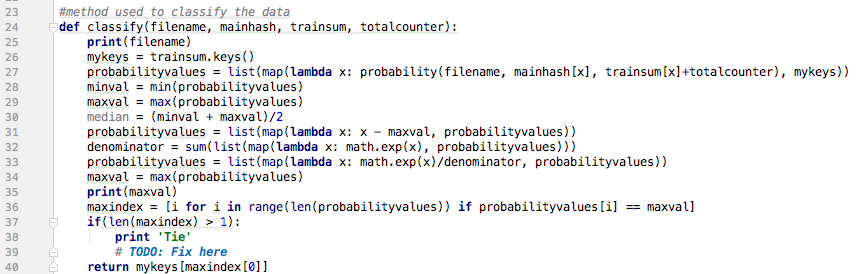
The above is a read words method which opens the file passed as parameter and reads it in lowercase. Also the file is read using the specified regular expression which helps to get rid of unneeded special characters.



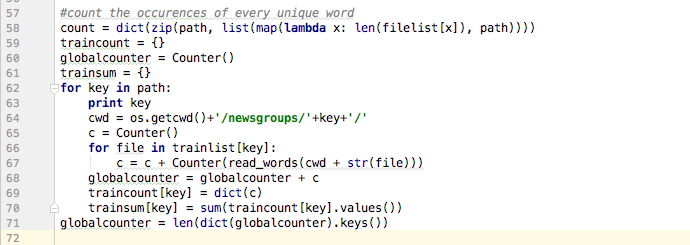
Here the above method calculates the probability of each word and also avoids the zero probability error by adding 1 to numerator and adding the total count of words into the denominator.



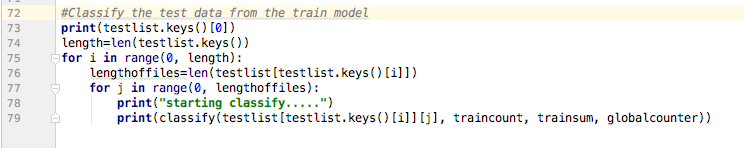
Here the total probability is calculated by using the previous two methods: read words and probabilityword.



The above line of code is used to define the classify method which classifies the test data by calculating max probability values.

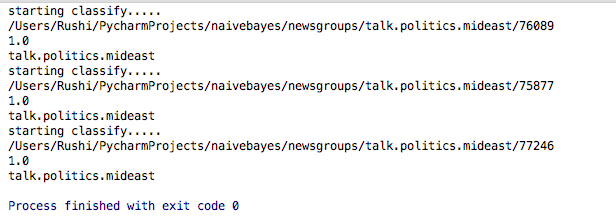


This above snippet is used to count the occurrences of unique word from each document for each class using the counter() method from the Collections package.



The last part shown above classifies the test data from the testlist which includes half of the documents from each class.

**Output:**



The above screenshot shows the output of the classifier where it shows the file and its path that is being tested and shows the output of the classifier as talk.politics.mideast.