**Project 2: Bayesian learning for classifying netnews text articles**

**Rathna Sindura Chikkam**

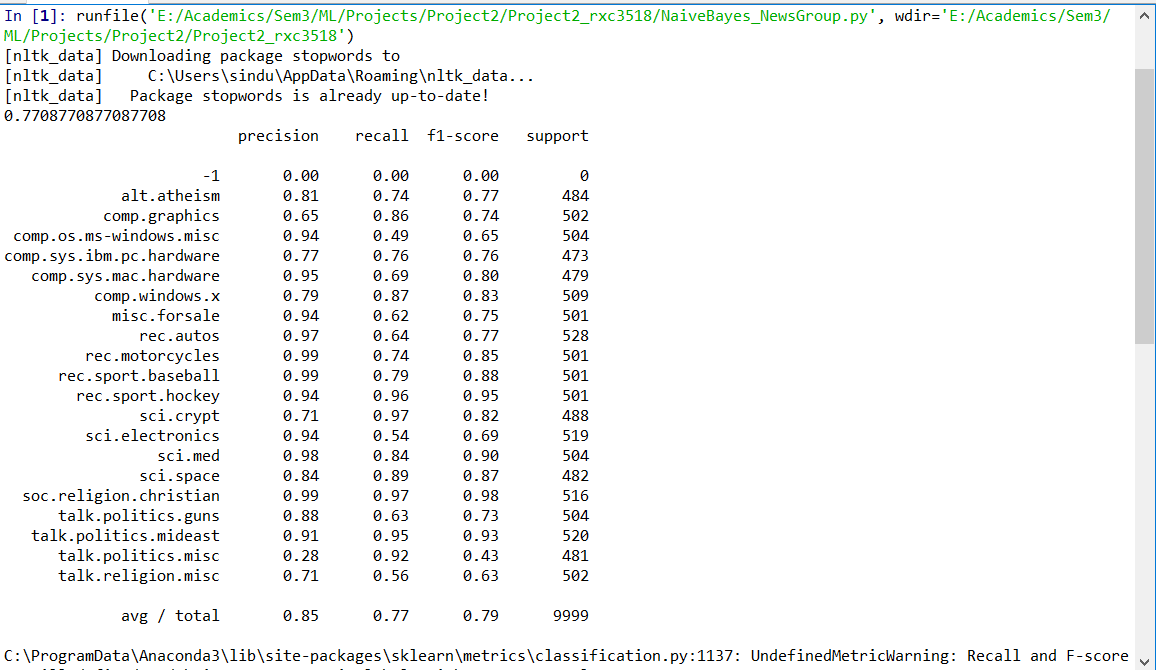
**ID: 1001553518 | rxc3518**

**Procedure:**

For this project to perform Naïve Bayes classification on text articles, I have followed Bag of Words approach to compartmentalize the data. Below are the steps:

1. First, I have read the file names of all the files present in 20\_newsgroup folder and have added them into one big master list of filenames and their categories.
2. From the master list of files and their respective categories, I am using train\_test\_split cross validation package from Scikit-learn to split the entire data into train and test document sets along with their train and test categories. I have used the split size to be 0.5 as mentioned in the project description. I have also written few helper function likes flatten, doc\_tokenize, hasNumbers that helps in list of words formation, tokenzing each document and performing regex validation to find numbers in words.
3. In Step 3, I am performing preprocessing on the bag of words formed from train set of documents that we had split earlier. I have used nltk – natural language toolkit to remove stop words that have no significance in category classification, have removed words of length 1 and 2, have performed alphanumeric words removal and have made sure that no words has any digit in it.
   1. I am also performing a nltk frequency count of words and their number of appearances in order to remove least common words and pull the first 10,000 words as features – I have performed a feature selection of choosing first 10,000 most frequently occurred words from the entire dataset.
4. From the 4th step, I have pulled a dictionary set from every document present in train and test split of data.
5. In the 5th step, I have written helper functions to fit the train data set, find class probability while performing classification on test data set, and have used accuracy\_score and classification\_report packages from Scikit-learn to know the accuracy of our Naïve bayes claasifier.

Accuracy obtained is: 77% approximately as shown below:



References:

<https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/>

<https://towardsdatascience.com/unfolding-na%C3%AFve-bayes-from-scratch-2e86dcae4b01>

<https://github.com/gokriznastic/20-newsgroups-text_classification/blob/master/Multinomial%20Naive%20Bayes-%20BOW%20with%20TF.ipynb>