**COMP 237 - NLP Project Assignment**

NIYANTA (301224927) Date :07/04/2022

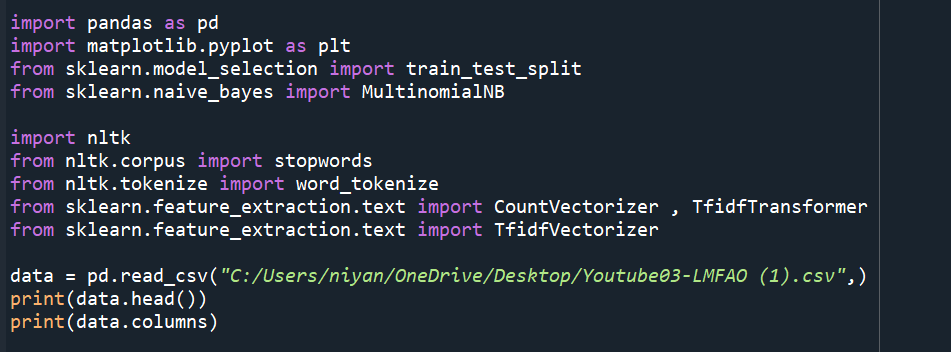
**Overview:** In this project I built a simple text classifier using the “Bag of words’ language model & the Naive Bayes classifier. As you know movies on Youtube are very common and once a movie is made available many comments are posted over the internet by viewers in response to the movie. But in some instances, these comments are auto generated and considered spam.

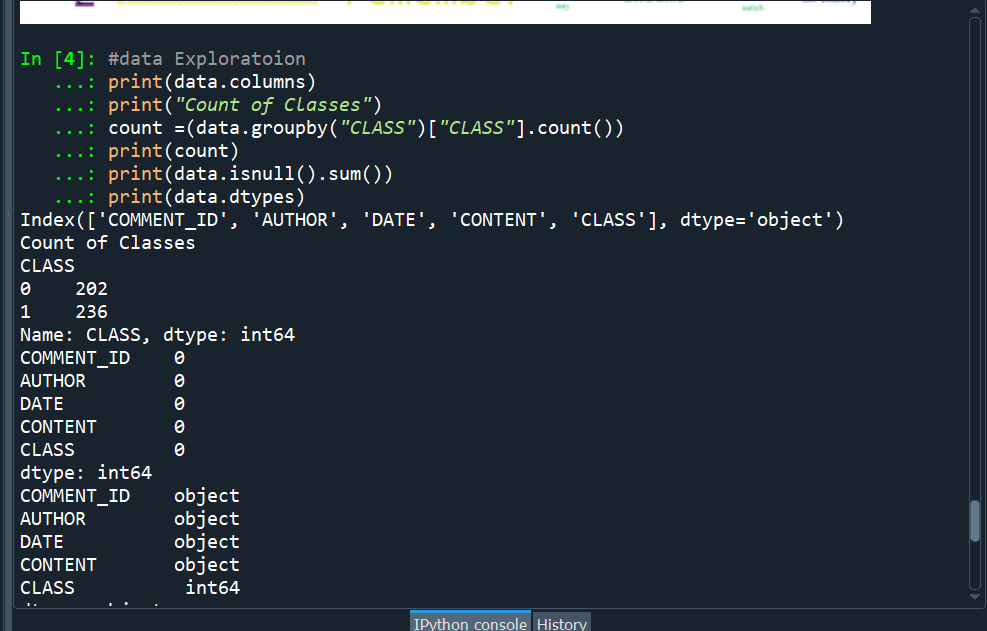
The purpose of your model is to filter the spam comments by training a Naive Bayes classifier. The data is available for movies at the UCI machine learning repository.

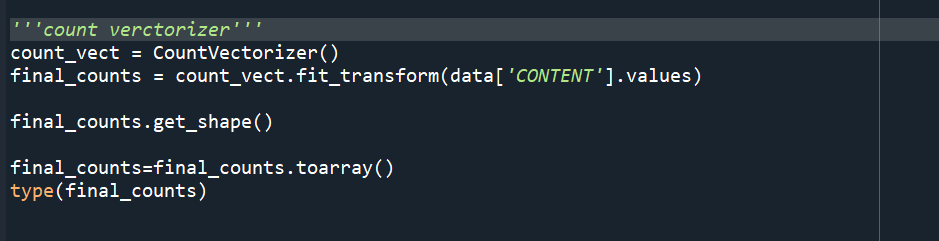
Assigned Dataset: [Youtube03-LMFAO.csv](https://e.centennialcollege.ca/d2l/common/viewFile.d2lfile/Database/MjkyMzAwNTA/Youtube03-LMFAO.csv?ou=746933)

Requirements:

1. Load the data into a panda’s data frame.



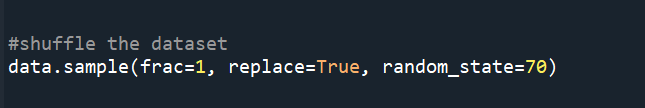
1. Carry out some basic data exploration and present your results. 
2. Using nltk toolkit classes and methods to prepare the data for model building, refer to the third lab tutorial in module 11 (Building a Category text predictor ). Use count\_vectorizer.fit\_transform().

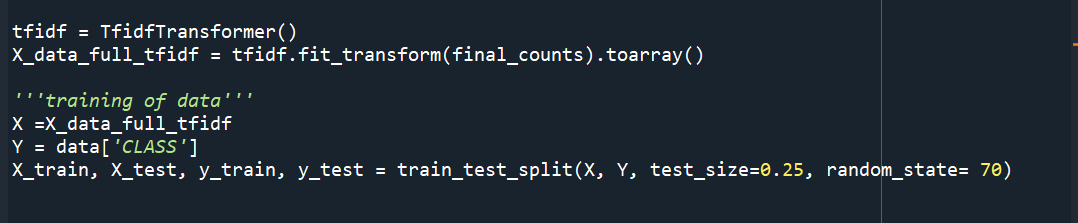


1. Present highlights of the output (initial features) such as the new shape of the data and any other useful information before proceeding.
2. Downscale the transformed data using tf-idf and again present highlights of the output (final features) such as the new shape of the data and any other useful information before proceeding.

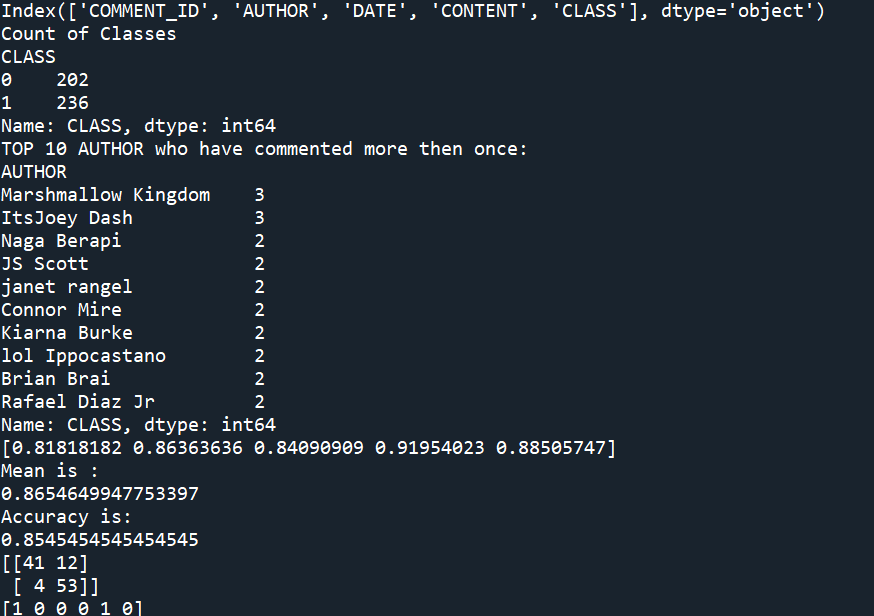


1. Use pandas.sample to shuffle the dataset, set frac =1 .Using pandas split your dataset into 75% for training and 25% for testing, make sure to separate the class from the feature(s). (Do not use test\_train\_ split)

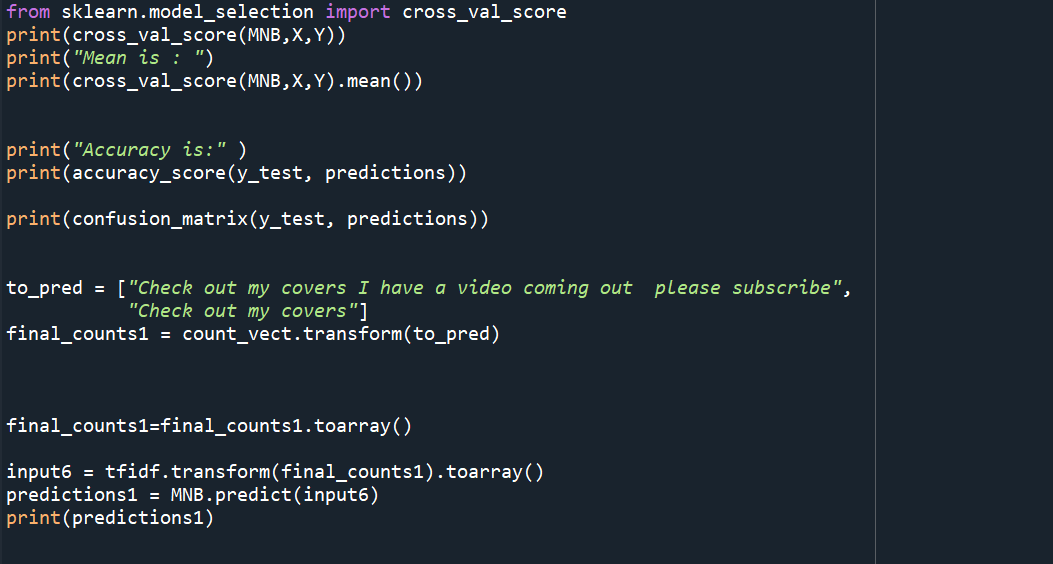




1. Fit the training data into a Naive Bayes classifier.
2. Cross validate the model on the training data using 5-fold and print the mean results of model accuracy.



1. Test the model on the test data, print the confusion matrix and the accuracy of the model.



1. As a group come up with 6 new comments (4 comments should be non spam and 2 comment spam) and pass them to the classifier and check the results. You can be very creative and even do more .



1. Present all the results and conclusions.

We trained the model by removing the stop words and model accuracy is 85%. However, when we trained the data without removing the stop words the data accuracy is more. The given dataset is slightly unbalanced i.e there are more class “1” values then “0”. Only the two columns i.e CLASS and CONTENT are relevant. We ignore the column AUTHOR because most of the people commented only once. There is a lot of possibility because of overfitting of words like please, http,subscribe. Apart from this, humans use more emotional words like awesome,wow,best etc., but bots use words like check,subscribe,and youtube.

