Spam Email Detection using Machine Learning Project in Python

This is a small project in python using machine learning to detect whether a given text is spam or ham.

Importing The Libraries

We used Pandas and NumPy for data manipulation, sklearn for preprocessing, model creation and model evaluation.

Analysing The Data

Using the df.head(), df being the Pandas DataFrame object where we have loaded the data from the CSV, function to view the data, here we see that there are only two columns. text, containing the text for detection and target, as the label to tell whether the text is spam or not.

Feature Engineering

We have split the data on a ratio of 80% — 20%, training and test respectively. Then we have initialized a count vectorizer from

sklearn.feature_extraction.text which we have used to convert collection of text into numereical vector form (0 ,1). Though we can mess with those to get even better results, keeping in mind to not overfit our model.

Model Creation

We have used multinomialNB imported from sklearn which calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.In naïve classifier a feature's existence or absence has no bearing on the inclusion or exclusion of another feature.

Then we have fitted in the model .

Result

We have got the result from model.score and then stored the result by multiplying with 100 to get in percengtage which we have got 98.2 % accuracy.

We have used pickle which is used to keep tracks of the objects it has already serialized and later we have referenced it to develop a web pages to predict the mail is spam or not by using streamlit .

TEAM MEMBERS AND CONTRIBUTIONS

SANJEET YADAV :20103129 (model development)

SATNAM SINGH:20103130 (dataset and report making)

SHIVAM JASWAL:20103133 (web page design)