Our objective was to build a language detector using python and machine learning.

Linguistics study said distribution of each Roman alphabet letter appears differently to each language.

This might be because one language would use certain word or certain expression more frequently over another and end up contains certain letter more frequently on their written content.

So we build a python script that counts number of each alphabet letters appears on each txt files that are written in 3 different languages. And learn distinctive pattern of each languages and be able to distinguish language of the test txt file.

After we collect our train data, we named each txt file with prefix of initial of each language and use them as label. Then we collect our test data and did same process to our test dataset. So later when we test our test dataset, the scrip can decide if it distinguish the language correctly by checking if label match to the result.

The first algorithm that we used is SVC algorithm. Initially when we first tried our script with 5 sets of train data of each language, we already have accuracy score of 0.875 which I thought it was pretty high. It identified on Indonesian txt file as English.

Then we decided to add more train data to see if accuracy would increase after with additional training. And we got 100% accuracy with more train data. Our script successfully distinguished language of 15 test.txt file.

We also tried to use Random Forest Classifier to see how it’s different from using SVC algorithm. The accuracy result also 100%. However it took slightly longer to finish the learn data compare to SVC so we have concluded that SVC algorithm is the best fit algorithm to be used for our project purpose.