Proposed strategy 1

The model used till now was the n -gram approach.

One approach which can be implemented is the Markov model.

If a document is given, each word can be bifurcated as a Markov chain of letters.

When the entire document is taken into account as one complete set of Markov chains, the set of starting and transitional probabilities can be calculated and referred to as a Markov Model for that particular language.

This proposed model in our research project which will not only identify the languages with a lower error rate, but will also result in faster identification speed as compared to N-gram model.

The occurrences of letters in a word can be regarded as a stochastic process and hence the word can be represented as a Markov chain where letters are states. The occurrence of the first letter in the word is characterized by the initial probability of the Markov chain and the occurrence of the other letter given the occurrence of its previous letter is characterized by the transition probability.

References:

[1]<https://pdfs.semanticscholar.org/2bf0/8addb83f51befa8b4bc7ed16b54ed34018d0.pdf>

[2]<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.630&rep=rep1&type=pdf>

[3] http://www.cs.princeton.edu/courses/archive/spr05/cos126/assignments/markov.html