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.

Updated Strategy:

Formally, a Markov chain is a probabilistic automaton. The probability distribution of state transitions is typically represented as the Markov chain’s ***transition matrix****.*If the Markov chain has **N** possible states, the matrix will be an **N x N**matrix, such that entry **(I, J)** is the probability of transitioning from state **I**to state **J**. Additionally, the transition matrix must be a **stochastic matrix**, a matrix whose entries in each row must add up to exactly 1. This makes complete sense, since each row represents its own probability distribution.

Additionally, a Markov chain also has an ***initial state vector***, represented as an **N x 1** matrix (a vector), that describes the probability distribution of starting at each of the **N** possible states. Entry **I**of the vector describes the probability of the chain beginning at state**I**.

The task is to compute in a best way, given the parameters of the model, the probability of a particular output sequence.

{\displaystyle Y=y(0),y(1),\dots ,y(L-1)\,}

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>

[4]

http://jedlik.phy.bme.hu/~gerjanos/HMM/node4.html