POS TAGGER

Hidden Markov Model (HMM):-

This model works by calculating two kinds of probabilities: the emissions probability and the transition probability.
Emission probability: The emission probability is the probability of observing a specific **output symbol (word)** given a particular **hidden state (tag)**.
Transition probability: The transition probability is the probability of transitioning from one **hidden state (tag)** to another. We add two extra tags <S> (start tag) and <E> (end tag) at the start and end of every sentence respectively. This is generally done to make the model more accurate.
The above probabilities are calculated for every tag in the training data.
After calculating both the probabilities, we construct the emission matrix and transition matrix by summing up the probabilities.
Now, we apply the Viterbi Algorithm to both the matrix to predict the tags of the given text. Each tag is represented as a node in the graph and transition probabilities are represented as the weight of the edges while emission probabilities are assigned to each node corresponding to a word.
Now, we remove all the edges and nodes with a probability of 0 and calculate the maximum of each path (obtained by multiplying all the probabilities of that path).
The predicted tags are returned as output.

Conditional Random Field (CRF):-

Similar to HMM, this is also a graphical model but it uses an undirectional graphical approach and conditional probability.
Undirectional approach means now we can consider all the possible combinations of the tags instead of just focusing on which tag comes after another.
This model is more complex to implement as it considers a number of factors like the length of the words, the order of the words, and the context of the words.
Here we calculate the conditional probability of every tag given the previous and the next tag in the text. We can extend this model to consider multiple words in a sentence as per our choice. In my implementation, I have considered two words before and two words after a given tag to make it more efficient as compared to the other model.

Observations:-

After calculating the precision, recall and F1 score of both models, it was observed that the accuracy of the CRF model is slightly better than the HMM model. If we compare the F1 score of both models, a significant difference is observed between the two in the case of the English language but in the case of the Hindi language, the difference is not that significant. But in the case of large data sets, the difference might be huge.