REPORT P2-A

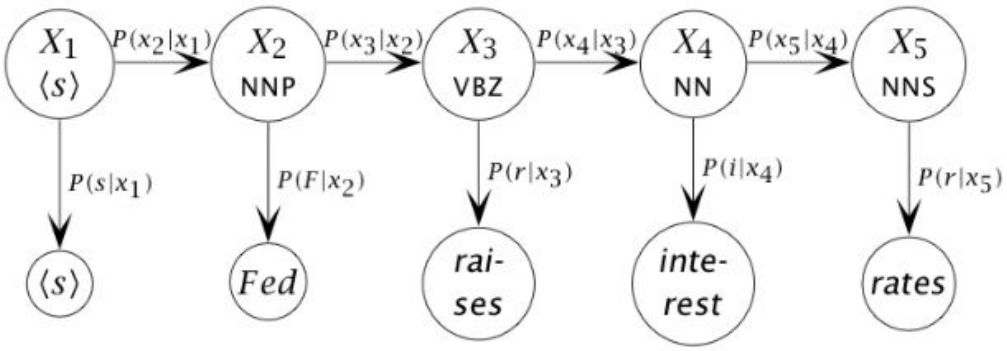
In the assignment we are supposed to find out the Part of Speech tags using Hidden Markov Model. The brown corpus is used for training and testing the Model.

The training data is in the form of [(x1,y1),(x2,y2),...] where xi is the word in the sentence and yi is the corresponding tag to the word.

Hidden Markov Models (HMMs) model : A system of discrete temporal unobserved (hidden) variables and discrete temporal observed variables. The observed and unobserved variables are related through emission probabilities.

Viterbi algorithm : Finding most likely sequence in HMM
The Viterbi algorithm is a dynamic programming algorithm that efficiently computes the the most likely states of the latent variables of a Hidden Markov Model (HMM), given an observed sequence of emissions.

In tagging problem, each xi would be a sequence of words x1 x2 x3 … xn and each yi would be a sequence of tags y1 y2 y3...yn



Implementation Details :

1. Pre-processing
2. Data Structure
    - A dictionary is maintained to store the trigrams, bigrams, state transition and states.

    (Here tags are the states and words are transition).
    - Smoothing is used so as to handle the unknown words.