# Parts of Speech Tagger

- By Baani(2016234)

## 1. Training the data

The BERP dataset was used to train the data.
The dataset was trained using bigram probabilities of word given tag, and tag given tag, i.e the emission and transition probabilities respectively.

## 2. Viterbi Algorithm

The viterbi algorithm which using dynamic probability to calculate the hidden POS tags for a given test corpus, using **Hidden Markov's model** was used.
Unknown probabilities were handled by using **Add-1 smoothing (Laplace smoothing)**.

## 3. Assumptions

i. The tags that appear in the training data constitute all the tags that exist (no new tags will appear in testing). However, new words may appear in the test set.
ii. All the sentences are treated as sentences in the lower case.
iii. All the terminal punctuations are mapped to '.', since its the only terminal punctuation tag in the training corpus.
iv. All the trained models have been dumped as json files in the Train.py file, and have been used in the test POSTagging.py file

## 4. Guidelines

a. **The Dataset:** The BERP dataset has been used to train the data. The same is present in the file Training set_HMM.txt
b. **The Training Script:** The script Train.py is used to train the model, and saves all the models as json dump files in the data folder.
c. **The Test Script:** The script POSTagging.py is used for a given input in the same directory by the name test_set.txt, to predict the POS Tags.