CSE556 : Natural Language Processing
Assignment 02

Deadline: 4th Oct, 11:59:59 PM Max Marks: 100

Instructions:

- The assignment is to be attempted individually.
- Language allowed: Python
- You are allowed to use libraries such as NLTK for data preprocessing.
- For Plagiarism, institute policy will be followed.
- You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Output.pdf.
- Mention methodology, preprocessing steps and assumptions you may have in README.pdf.
- Mention your sample outputs in the output.pdf.
- You are advised to prepare a well documented code file.
- Submit code, readme and output files in ZIP format with the following name: **A2_<roll_no>.zip**
- Use classroom discussion for any doubt.

Dataset: Brown PoS tag corpus. (Attached)

Dataset Format:

- Each line represents one sentence.
- Sentences are already tokenized.
- Words in a line have the format word_tag.

1. Design and implement Hidden Markov Model (HMM) based Part-of-Speech (POS) tagger implementing Viterbi algorithm with the following assumptions.
   - **Markov assumption length 1 -** Probability of any state s_k depends on its previous state only, i.e., $P(s_k \mid s_{k-1})$
   - **Markov assumption length 2 -** Probability of any state s_k depends on its previous two states only, i.e., $P(s_k \mid s_{k-2} s_{k-1})$

   Perform 3-fold cross validation on the dataset (Brown_Train.txt) and report the following for each fold and an average score

   1.1. Precision, recall and F1-score.
   1.2. Tag-wise precision, recall and F1-score
   1.3. Confusion matrix (Each element Aij of matrix A denotes the number of times tag i classified as tag j)
   1.4. Statistics of tag set.

2. Work out the mathematics of HMM for Markov assumption length 2, and include it in your assignment report.
3. Which word types are most frequently tagged incorrectly by the HMM, and why?