CS 564: Foundations of Machine Learning
Assignment 3

Deadline: 17 November 2022

- Markings will be based on the correctness and soundness of the outputs.
- Marks will be deducted in case of plagiarism.
- Proper indentation and appropriate comments (if necessary) are mandatory.
- Use of frameworks like scikit-learn etc is allowed.
- *All benchmarks(accuracy etc), answers to questions and supporting examples should be added in a separate file with the name 'report'.*
- *All code needs to be submitted in '.py' format.* Even if you code it in '.IPYNB' format, download it in '.py' format and then submit
- You should zip all the required files and name the zip file as:
    - <roll_no>_assignment_<#>.zip, eg. 1501cs11_assignment_01.zip.
- Upload your assignment (the zip file) in the following link:
    - https://www.dropbox.com/request/HAHdJCOh1PEtqOBdLJ8y

Problem Statement:
- The assignment targets to implement Hidden Markov Model (HMM) to perform Named Entity Recognition (NER) task

Implementation:
- HMM Parameter Estimation
    - Input: Annotated tagged dataset
    - Output: HMM parameters
    - Procedure:
        - Step1: Find states.
        - Step2: Calculate Start probability (π).
        - Step3: Calculate transition probability (A)
        - Step4: Calculate emission probability (B)
- Features for HMM
    - Please build features according to your understanding and choice
- Testing
    - After calculating all these parameters apply these parameters to the Viterbi algorithm and test sentences as an observation to find named entities.

Dataset:

- NER-Dataset-Train.txt —> Contains train set
- NER-Dataset-TestSet.txt —> Contains test set
- Link to dataset:
  https://www.dropbox.com/sh/sc14co1b30nru66/AAAhhoCTz6NxKsCw1eBQC_B
  ma?dl=0

Documents to submit:

- Model code
- Perform 5 fold cross-validation on the Training datasets and report both average & individual fold results (Accuracy, Precision, Recall and F-Score).
- Submit Test Set Predictions
- Write a report (doc or pdf format) on how you are solving the problems as well as all the results including model architecture (if any).

For any queries regarding this assignment, contact:

Abhisek Tiwari (abhisektiwari2014@gmail.com),
Ratnesh Kumar Joshi (ratneshkr.joshi@gmail.com), and
Ramakrishna Appicharla (ramakrishnaappicharla@gmail.com)