

## CS 563: Natural Language Processing

### Assignment-2: NER

**Deadline: 2 March 2023**

- 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/9AqrVqNzPi2KiZh99cLF>

#### **Problem Statement:**

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

#### **Implementation:**

#### **HMM based Model:**

- HMM Parameter Estimation
  - Input: Annotated tagged dataset
  - Output: HMM parameters
  - Procedure:
    - Step1: Find states.
    - Step2: Calculate Start probability ( $\pi$ ).
    - Step3: Calculate transition probability (A)
    - Step4: Calculate emission probability (B)
- Features for HMM:
  - Train two HMM models based on:
    - First order markov assumption (Bigram) where current word NER tag is based on the previous and current words

- Conditionally independent (Unigram) where current word NER tag is just based on the current word and not on previous word

### **RNN based Model:**

- Explain and draw the architecture of RNN that you are proposing with justification
- Describe the features of RNN

### **Testing:**

- After calculating all these parameters apply these parameters to the Viterbi algorithm and test sentences as an observation to find named entities

### **Dataset:**

- Dataset consists of tweets and each word is tagged with its corresponding NER tag
- NER-Dataset-Train.txt → Contains train set
- Tweet NER dataset: [Link to dataset](#)
- Format of dataset:
  - Each line contains <Word \t Tag> (word followed by tab-space and tag)
  - Sentences are separated by a new line

### **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).
- Briefly discuss about Unigram vs Bigram assumption while training HMMs
- 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:**

Aizan Zafar ([aizanzafar@gmail.com](mailto:aizanzafar@gmail.com)),  
Ramakrishna Appicharla ([ramakrishnaappicharla@gmail.com](mailto:ramakrishnaappicharla@gmail.com)),  
Mamta ([mamta20118@gmail.com](mailto:mamta20118@gmail.com)) and,  
Arpan Phukan ([arpanphukan@gmail.com](mailto:arpanphukan@gmail.com))