Natural Language Processing (CS 563)

Assignment-2: NER

(Read all the instructions carefully & adhere to them.)

Date: Feb 09, 2022 **Deadline: Feb 20, 2022**

Total Marks: 20

Instructions:

1. The assignment should be completed and uploaded by **Feb 20, 2020, 11:59 PM IST.**
2. Markings will be based on the correctness and soundness of the outputs. Marks will be deducted in case of plagiarism.
3. Proper indentation and appropriate comments are mandatory.
4. You should zip all the required files and name the zip file as **<1st_member_roll_no>_<2nd_member_roll_no>_assignment_<#>.zip , eg. 1701cs11_1701cs31_assignment_01.zip**
5. Upload your assignment (**the zip file**) in the following link:
    - https://www.dropbox.com/request/Vfm58D5KmknihAbWQJZ0
6. For any queries regarding this assignment you can contact:
    - Deeksha Varshney (deeksha.varshney2695@gmail.com) or Gopendra Vikram Singh (gopendra.99@gmail.com)

Setups (Write codes for the following):

1. Identify all the named entities, i.e., whether a token is a named entity or not.
2. Identify the fine-grained named entity types in a sentence.
    - Ex- *"Junk food may not kill us directly …."- **Velasquez-manof** #diet*
    - Total 10 NER tags (e.g. *person, product, company, geolocation, movie, music artist, tvshow, facility, sports team and others.*)

Dataset (NER in Twitter):

- Train.txt (Train the ner model using this file)
    - Format: Each line contains <Word \t Tag>
    - Sentences are separated by a blank line.
- Valid.txt (Use this file for validating the model)
- Test.txt (This file should be used to generate the predictions and compute the accuracy)
- **Download link:** https://drive.google.com/file/d/1_wRTAZL7xwuRf_sd6Shy5DJhHS5UYtmw/view?usp=sharing

Using the above-mentioned Dataset, perform the tasks mentioned in Setups using the HMM-based Model:

1. **HMM Parameter Estimation**

   Input: Annotated tagged dataset

   Output: HMM parameters

   Procedure:

   Step 1: Find states.

   Step 2: Calculate Start probability (π).

   Step 3: Calculate transition probability (A)

   Step 4: Calculate emission probability (B)

2. **Features for HMM**
   a. Use bigram and trigram models
   b. Introduce no context and the additional context in the form of preceding tag while computing the emission probabilities

3. **Testing**

   After calculating all these parameters apply these parameters using the Viterbi algorithm, and determine the best sequence of the named entity.

Evaluation:

1. Report Accuracy, Precision, Recall and F-Score on the test set provided.
2. Submit Test Set Predictions.
3. Comparison between the results: Bigram vs trigram model (without context during emission probability); Bigram vs trigram model (context during emission probability);
4. 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).