Natural Language Processing (CS 563)

Assignment-1: NER

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

Date: Jan 26, 2020

Deadline: Feb 04, 2020
Total Marks: 50

Instructions:

1. The assignment should be completed and uploaded by **Feb 04, 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 ***roll_no*.zip**, eg. **1601cs11.zip.**
5. Upload your assignment (**the zip file**) in the following link: https://www.dropbox.com/request/qXcL4W7cD0jXQycAzKi9

Setups:

1. Identify all the named entity, i.e., whether a token is a named entity or not.
2. Identify the named entity types in a sentence.

Dataset:

- NER-Dataset-Train.txt (Identify the presence of named entity in a tweet.)
- NER-Dataset-10Types-Train.txt (Identify the presence of named entity

and classify them into predefined 10 subtypes. 10 Types are person, product, company, geolocation, movie, music artist, tvshow, facility, sports team and other .)

- NER-Dataset-TestSet.txt (Separate prediction files will be generated using both training set on this Dataset)

- Format:
  - Each line contains <Word \t Tag>
  - Sentences are separated by a blank line.

Using the above mentioned Dataset, perform the tasks mentioned in Setups for the following two models:

1. **HMM based Model**

   a. **HMM Parameter Estimation** **(9 Marks)**

      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)

   b. **Features for HMM** **(6 Marks)**

      Please build features according to your understanding and choice

   c. **Testing** **(10 Marks)**

      After calculating all these parameters apply these parameters to the Viterbi algorithm and testing sentence as an observation to find named entities.

2. **RNN based Model**

   a. **Model Architecture** **(2 Marks)**

      Draw a model architecture of the model you are proposing

   b. **Features for RNN** **(3 Marks)**

      Please build features according to your understanding and choice

Evaluation (For both models): (10+10=20 Marks)

1. Perform 5 fold cross-validation on the Training datasets and report both average & individual fold results (Accuracy, Precision, Recall and F-Score).
2. Submit Test Set Predictions (a total of 4 files, 2 for HMM based models and 2 for RNN based model). [***have to upload***]
3. 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). [***have to upload***].