Lab 3: Problem Statement

Rudra Murthy V, Diptesh Kanojia
Course: Natural Language Processing (NLP)

February 4, 2022

This weeks lab consists of the Named Entity Recognition (NER) task. We have provided you with sample code which will help you understand a basic implementation with the help of the Twitter-NER data. This should help you take you a few steps towards your own implementation, on one of the datasets provided with this lab.

You are provided with a Jupyter Notebook which consists of the following code snippets:

- Reading the Train/Test data in the CoNLL 2003 format.
- Generating features from the Twitter data provided.
- A CRF-based NER Tagger Implementation which included hyperparameters to be tuned.
- Evaluation & Analysis, *i.e.,* Evaluating the output of the CRF-based NER tagging approach in terms of F1-scores.

Problem 1. You assignment is to create your own new notebook which performs the CRF-based NER tagging using ONE of the datasets provided. The NER datasets provided for this task are described in the Git README.md file. The discussion from our lectures should help you understand CRF and use the feature sets for your implementation. **You are encouraged to download the dataset in your native language for this task.** You are already provided with an example training/testing process which should help you with some of the code-snippets to be used as-is from our notebook. However, please be careful with which dataset you are using (Read the README file on Git).
You need to train/test the CRF-based NER tagger for the dataset chosen. Further, you are required to train/test this tagger with different values for both the hyperparameters (c1, c2). After obtaining different weighted F1 scores for each hyperparameter setting, you are required to plot a graph between c1 value, c2 value, and weighted F1 score.

Problem 2. You notebook should also output/show a confusion matrix for all the tags apart from the plotted graph. After you show the graph in the notebook, you should also display/output a confusion matrix for the best hyperparameter setting.

Submission You are requested to submit one file - your jupyter notebook.

These submissions can be made at the Google Classroom portal akin to the first lab.