Spring 2025

EDS 6397 – Natural Language Processing

Assignment #1 – Named Entity Recognition

You may use AI-assistants in code development, but you are discouraged from doing so. You may NOT use AI-assistants to write the report. It needs to be entirely written by you.

Assignment Overview

The objective of this assignment is to familiarize you with an online tool for labeling named entities as well as performing Named Entity Recognition using Spacy's built-in pipeline.

Input Data Description

Data given to you is cleaned-up tweets downloaded from Kaggle in an xlsx files named Assignment 1 Data.xlsx. You need to be able to directly read this file and access its sheets (specifically the one named Tweets.)

Each student should use only 300 tweets. The xlsx file has two sheets; "Tweets" which has the tweets, and "Roster" which has the names of students along with assigned tweet ID to each student. Make sure you use only the tweets that are assigned to you. **Failure to do so might cost you up to 50 points out of 100 for this assignment**.

Assignment Description

Phase 1: NER Annotation (labeling)

Here we will use an online tool that helps us label some data. We will not use the labeled data for training here. We will use it is phase 3 for accuracy assessment of Spacy's NER pipeline.

Step 1: Access the NER Annotator Tool

- In your web browser go to NER Annotator tool https://tecoholic.github.io/ner-annotator/
- Note: This is a free online tool that doesn't need you to create an account. The down-side is that your annotations will not be saved. So, make sure you follow step 7 and save your work every few tweets in order to prevent having to redo annotation in the event you lose internet connectivity or experience power outage, computer malfunction, etc.

Step 2: Load Your Data

- The homepage will show an upload logo with text written "Text File". Copy your assigned tweets from the csv file into a .txt file and save it. Then upload the .txt file to the NER Annotator tool. Once uploaded, click the "Begin" button on the top right of your screen to begin and you will be moved to a new page.

Step 3: Configuration

- First option you will see on the top left of your screen is "Text Separator". Make sure it has "New Line" on it, since your new tweets will start after a new line.
- Second option you will see below "Text Separator" is the "Annotation Precision". For that option select "Character Level".

Step 4: Select the Entities You Want to Annotate

- On the right side of the tool, you will find options to add new entities (NEW TAG) or edit (EDIT TAGS).
- To add a new entity type (such as LOC, PERSON, ORG), click on the "+" button and enter the name of the entity in the dialog box that appears. You need to include these 7 tags: outputs, only separate the following NER tags: PERSON, NORP, ORG, GPE, LOC, DATE, MONEY. You can check Lecture 3 slides for more explanation on these tags.
- Once you have added your entities, they will appear in a list on the top of the screen.

Step 5: Start Annotating

- Select the appropriate entity type from the list on the top.
- Highlight the text you want to annotate. For example, if you want to annotate "New York" as a location, simply select "New York" with your mouse.
- The highlighted text will now be marked with the selected entity type. The annotation will be color-coded based on the entity type for easy identification.

Step 6: Review and Edit Annotations

- Continue highlighting and tagging all relevant entities in your text.
- To remove an annotation, click "x" on the highlighted text.
- Once done with annotating the current tweet, press the "SAVE" button to proceed onto the next one.

Step 7: Export Your Annotations

- Once you have finished annotating your text, you can export the annotations.
- Click on "Annotations" on the top of the screen and then select "Export" button.
- The annotations will be saved as Spacy's JSON format on your local PC.

Finally, you need to write a code to read the JSON file(s) of your NER tags in Python to be able to use them in step 3.

Phase 2: Named entity recognition using NER pipeline

Develop a code similar to the code sample presented in Lecture 3 slides that uses Spacy's built-in pipeline to perform named entity recognition. You will only pass the 300 tweets that are assigned to you to the pipeline. Then from the outputs, only separate the following NER tags:

PERSON, NORP, ORG, GPE, LOC, DATE, MONEY

We will need these tagged tokens in the last phase (accuracy assessment.)

Phase 3: Accuracy assessment

We need to use the manually labeled (annotated) tweets here to see what percentage of tags Spacy missed (a measure of recall) and what percentage of tags were mis-identified (a measure of precision), for each tag category.

Note that we need to look at the occurrence of each tag, not the number of tweets. For example, you have 300 tweets, but might have 340 GPE tags, or only 30 MONEY tags. We perform accuracy assessment on the number of times the tags happen in the whole corpus you are given, not the number of sentences or tweets in the corpus.

You need to write a code that goes through the tweets. For each tweet, check which tags are present, and if they are identified by Spacy or not (if not, then that shows omission error which is related to recall.) Also keep track of wrongfully tagged entities (commission error which is related to precision.)

Report the Precision, Recall, and F1 Score for each tag (separately) and discuss the performance of Spacy's NER pipeline in one paragraph.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

What to submit:

1- Your Python Jupyter Notebook (including code to read the JSON annotations file, the code reading your assigned tweets from the csv file, your code to perform NER in Spacy, and your accuracy assessment code.) make sure you use some cells for adding brief explanations.
2- One JSON file showing your manually annotated (=labeled = tagged) entities from the 7 entity groups specified in phase 2.
3- A report (1-2 pages) that explains the steps you took to complete this task as well as your assessment of Spacy's accuracy in NER. Make sure you write the report without using AI assistance like ChatGPT. It is prohibited.

Due Date: February 7, 2025 by 11:59 PM (submit through Teams)

Late submissions will be penalized at a rate of 10 points per day.