**EDS 6397 – NLP**

**Assignment 1 – Named Entity Recognition)**

**Varun Vaddi - #2347481**

## INITIAL SETUP:

Initially, retrieved the **300 tweets** assigned to me from **11,101-11,400** as mentioned in the Roster and saved the file in CSV format as '**NER_Tweets_Data.csv**' with column name as '**tweets'**.

NER_Tweets_Data

| | tweets |
|---|---|
| 2 | EU Commission President Jose Manuel Barroso has said the 25-member bloc will make it clear to the Russian President that the EU is not " satisfied " with Ukraine 's disputed election . |
| 3 | Prime Minister Viktor Yanukovych -- who was backed by Mr. Putin -- has been declared the winner of Sunday 's voting , but the EU is rejecting the results because of allegations of widespread voting fraud . |
| 4 | On other issues , EU leaders say they will call on Russia to sign and ratify border agreements with Estonia and Latvia , and urge it to seek a political settlement in war-torn Chechnya . |
| 5 | Palestinian President Mahmoud Abbas has fired three top security officials after militants broke the ceasefire he reached with Israel and fired mortars at Jewish settlements in the Gaza Strip . |
| 6 | Mr. Abbas dismissed the men , including security chief Abdel Razek Majaidie , hours after Hamas militants fired the mortars . |
| 7 | No one was wounded in the shelling , which Hamas said was a response to Israeli gunfire that killed a Palestinian man Wednesday in Gaza . |
| 8 | Despite the violence , Israeli and Palestinian officials met late Thursday to discuss security issues . |
| 9 | Meanwhile , a deadly shootout was also reported at a Palestinian jail in Gaza City . |
| 10 | Palestinian sources say at least three people described as Israeli collaborators were killed when dozens of masked gunmen stormed a jailhouse and set free several prisoners . |

Copied all the 300 rows from the CSV file and pasted into a text file and saved it as '**NLP_Tweets.txt'.**

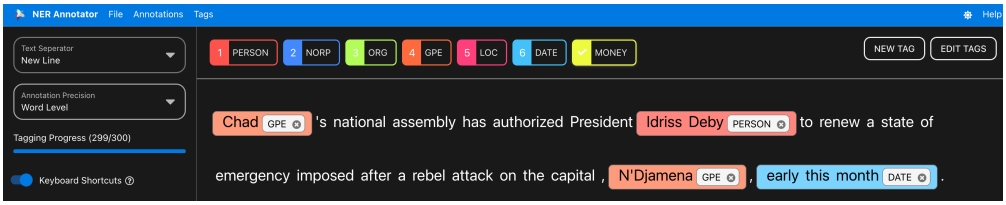
NLP_Tweets.txt

```
On other issues , EU leaders say they will call on Russia to sign and ratify border agreements with Estonia and Latvia , and urge it to seek a political settlement in war-torn Chechnya .
Palestinian President Mahmoud Abbas has fired three top security officials after militants broke the ceasefire he reached with Israel and fired mortars at Jewish settlements in the Gaza Strip .
Mr. Abbas dismissed the men , including security chief Abdel Razek Majaidie , hours after Hamas militants fired the mortars .
No one was wounded in the shelling , which Hamas said was a response to Israeli gunfire that killed a Palestinian man Wednesday in Gaza .
Despite the violence , Israeli and Palestinian officials met late Thursday to discuss security issues .
Meanwhile , a deadly shootout was also reported at a Palestinian jail in Gaza City .
Palestinian sources say at least three people described as Israeli collaborators were killed when dozens of masked gunmen stormed a jailhouse and set free several prisoners .
Germany , France and Lithuania are the latest European countries to congratulate Viktor Yushchenko on his victory in the Ukrainian presidential election .
In a letter to Mr. Yushchenko Tuesday , German Chancellor Gerhard Schroeder expressed optimism that Ukraine will continue what he called its transition toward the rule of law and a market economy .
Lithuanian President Viktor Adamkus congratulated Mr. Yushchenko in a telephone conversation .
In Paris , the foreign ministry said France and its European Union partners are determined to support democracy and modernization in Ukraine .
Meanwhile , Russia has questioned the objectivity of monitors for the Organization for Security and Cooperation in Europe .
That organization characterized last month 's presidential election re-run as flawed , while saying that Sunday 's poll vastly improved upon the last vote .
Russia backed Mr. Yushchenko 's opponent .
```

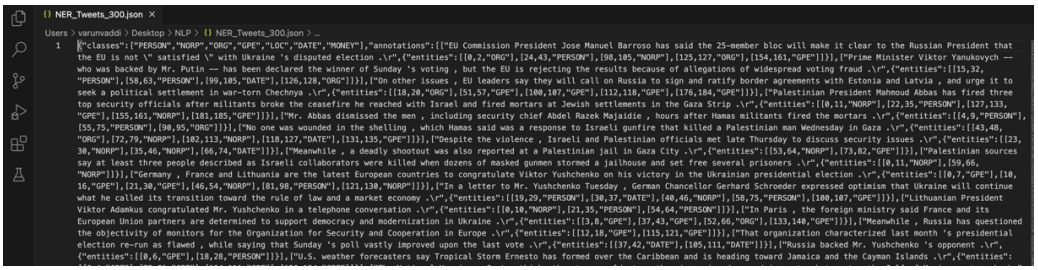
## PHASE-1:

Uploaded the '**NLP_Tweets.txt'** file into the **NER Annotator tool** and selected '**Text separator**' as '**New Line**' and '**Annotation Precision**' to '**Character Level**'.

Added all 7 entity tags by **PERSON, NORP, ORG, GPE, LOC, DATE, MONEY** and started annotating the tweets by assigning the tags to identified Named Entities and clicked on **Save** after each tweet.




**Exported** the Annotations and saved it as file naming as '**NER_Tweets_300.json**'.

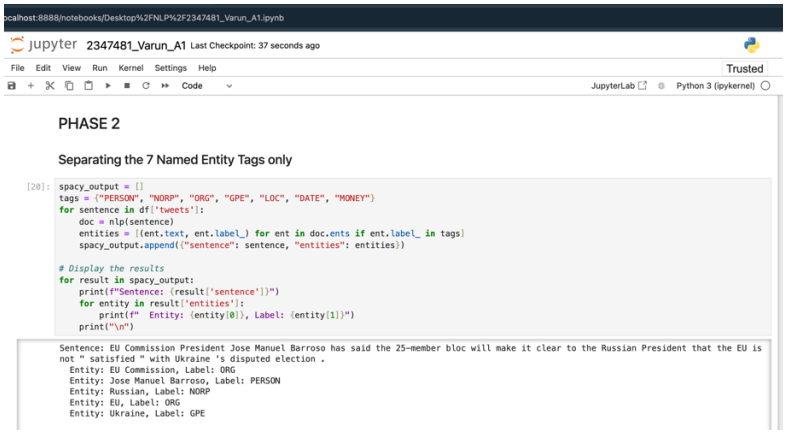



# PHASE-2:

**Opened** the JSON file in Jupyter notebook.

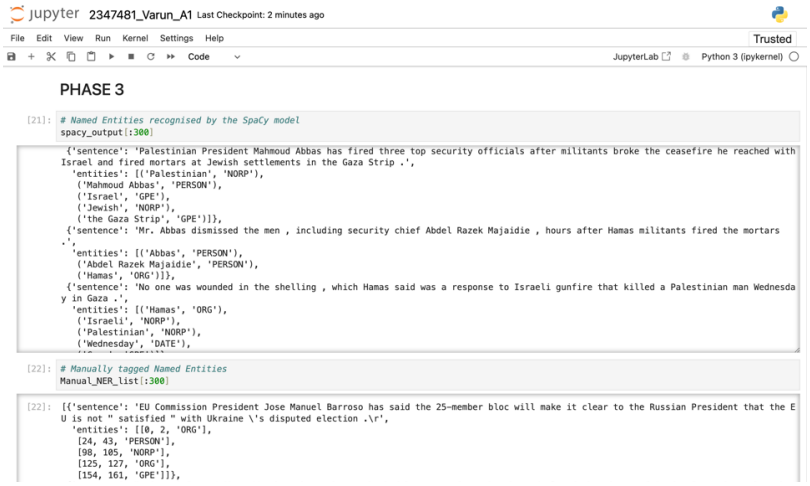
**Downloaded** the '**en_core_web_lg**' package and **loaded** it into the spacy model.

Then, I **filtered** only the 7 Tags that we used for the Manual NER Annotation.




# PHASE-3:

Spacy output format contains a tuple of word & tag **(word, tag)** whereas the manual format contains start index, end index & tag **(start, end, tag)**.



Since, both the formats are different, I have developed a new function - **format_conversion_from_spacy_to_manual()** to convert spaCy format from word to start & end indexes.

```
[23]: #Since, JSON output is having start & end indices instead of named entity, we need to convert the spaCy output from named entity to index posi
      def retrieve_entity_positions(sentence, entity_text):
          """Find the start and end positions of the entity in the sentence."""
          start = sentence.find(entity_text)
          if start == -1  # If the entity is not found, then return None
              return None
          end = start + len(entity_text)
          return start, end

[24]: def format_conversion_from_spacy_to_manual(spacy_NER):
          """Convert spaCy format (word, tag) to manual format (start, end, tag)."""
          annotations_after_conversion = []

          for entry in spacy_NER:
              sentence = entry['sentence']
              entities = []

              for entity_text, entity_tag in entry['entities']:
                  positions = retrieve_entity_positions(sentence, entity_text) #get each entity start & end index positions just like manual JSON ou
                  if positions:
                      start, end = positions
                      entities.append([start, end, entity_tag])

              annotations_after_conversion.append({'sentence': sentence, 'entities': entities})

          return annotations_after_conversion
```

Developed another function – **calculate_metrics()** to calculate **True Positives(TP), False Positives(FP),** and **False Negatives(FN)** for each tag.

Then used **TP, FN, FP** to calculate the **Precision, Recall** and **F1-score** are calculated for each tag.

```
[25]: def calculate_metrics(manual_NER, spacy_NER):
          tp, fp, fn = 0, 0, 0
          tp_tweet, fn_tweet, fp_tweet = 0, 0, 0
          temp_counter = 0

          for manual, spacy in zip(manual_NER, spacy_NER):
              # Convert manual annotations into a set of tuples (start, end, label)
              set_manual_entities = set([tuple(entity) for entity in manual['entities']])

              # Convert spaCy annotations into a set of tuples (start, end, label)
              set_spacy_entities = set([tuple(entity) for entity in spacy['entities']])


              # True Positives
              tp_tweet = len(set_manual_entities & set_spacy_entities)
              tp = tp + tp_tweet
              # False Positives
              fp_tweet = len(set_spacy_entities - set_manual_entities)
              fp = fp + fp_tweet
              # False Negatives
              fn_tweet = len(set_manual_entities - set_spacy_entities)
              fn = fn + fn_tweet
              temp_counter += 1

              precision_tweet = tp_tweet / (tp_tweet + fp_tweet) if (tp_tweet + fp_tweet) > 0 else 0
              recall_tweet = tp_tweet / (tp_tweet + fn_tweet) if (tp_tweet + fn_tweet) > 0 else 0
              f1_tweet = 2 * (precision_tweet * recall_tweet)/ (precision_tweet + recall_tweet) if (precision_tweet + recall_tweet)> 0 else 0

              print("------------------------------------------------------------------------------------------------------------------------")
              print(f"tweet: {temp_counter}")
              print(set_manual_entities)
              print(set_spacy_entities)
              print("-------------------------------")
              print(f"TP: {tp_tweet} | FP: {fp_tweet} | FN: {fn_tweet}")
              print("-------------------------------")
              print(f"Precision of tweet: {precision_tweet}")
              print(f"Recall of tweet: {recall_tweet}")
              print(f"f1-score of tweet: {f1_tweet}")


          return tp, fp, fn
```

## Assessment of SpaCy:

From my short stint working on this assignment, I have observed that spaCy identifies Named Entities different from how I annotated the tokens manually, reason for our Precision & recall being low.

For example, I have annotated **Cayman Islands** as **GPE**, but SpaCy annotated **the Cayman Islands** as a **GPE**. In other case, I have annotated **Carribean** as **GPE**, but SpaCy identified it as **LOC**.

The Ambiguity of a location/ place being put under GPE/ LOC is very high especially in case of less popular places. We sometimes need to have prior knowledge or info about the news in the tweet inorder to identify correctly the names of cities and persons.

```
-------------------------------------------------------
Manual NER vs SpaCy NER:
-------------------------------------------------------

Overall Precision: 0.36363636363636365
Overall Recall: 0.36681222707423583
Overall f1-score: 0.3652173913043478
-------------------------------------------------------
Overall True Positives (TP): 252
Overall False Positives (FP): 441
Overall False Negatives (FN): 435
```