**

COMP814 – Text Mining

Named Entity Recognition Lab

# Objective

1. To be able to process NER output from Spacy to extract higher level information.

# Task

1. You will need the sample code snippets from lectures for the following tasks.
2. Firstly ensure that you can run the lecture code as is, with the same data as in the demo.
3. Download sampleText2.txt and run the NER tool to the content of this file and examine the output.
4. Firstly manually evaluate the outputs from NLTK and Spacy and compute the **Precision** for each. Which one is better?
5. **The Information Extraction objective is to format the date entities into actual dates so that they can be processed as date series.**

1. Filter the entities of type “DATE” from the Spacy output.
2. Study the material [here](https://www.dataquest.io/blog/python-datetime-tutorial/) to learn how to process date and time using python.
3. Use the following specifications to format the dates.
   1. Today 🡺 YYYY-MM-DD (today’s date)
      1. Similarly yesterday would be one day before today’s date. Etc.
   2. Last week 🡺 YYYY-MM-DD TO YYYY-MM-DD ( where the 2 dates should indicate today’s date to 7 days prior date.
   3. Seven-day 🡺 7 DAY
      1. Last seven days should be an interval from today’s date to previous seven days.
   4. About six weeks 🡺 (approximate this to two dates with 6 week interval.
   5. March 🡺 (assume the current year March so output interval from 1st of march this year to 31 march.)
   6. Coming days 🡺 (approximate 4 day interval from current date.
4. If there are other ones not covered, you can use the heuristics used above to give it either a definite date, an interval or a fixed time in days.
5. Insert these converted dates just after the “DATE” object tokens in a file containing the rest of the text using upmarked formatting such as the following.
   1. <YYYY-MM-DD>
   2. <YYYY-MM\_DD TO YYYY-DD-MM>
6. Upload the link to the code as an attempt for this lab.
7. **BONUS Task**

* Use the date information from above and combine it with the Date **nsubj**, **Root**, and **dobj** corresponding to each of the DATE entities in a csv file. Use the code snippet from lectures.
* Output one entity per line.
* Download several more new articles focussing on COVID and output the **nsubj**, **Root**, and **dobj** triples and the corresponding connected dates. What useful information does this give you?