Lab 2: NLP
Spring 2023 – Jan 30

In this lab, we will be working with the Python regular expression library "re".

Step 1: In this lab, we will be using the TLS-Covid19 dataset containing Covid-19 news published in 2020. The data is provided on the course website. Spend some time understanding the data.
Hint: There are news on different topics (such as asia, australia, and beijing) from two news agencies: CNN and Guardian. Each topic has its unique directory. Ignore the timelines and just focus on input_docs directory.

Step 2 (60%): The goal of this lab is to develop a tool that can extract temporal expressions with the "re" library from the text. After you have identified the temporal expression(s), you should normalize them in mm/dd/yyyy format.

Temporal Pattern:

1. **Month Number – January 13**
2. **Number Month – 5 November**
3. **Month Year – January 2020**
4. **Extra Pattern of your choice to extract temporal expressions**

Step 3: **(40%)**
In the next step, our goal is to analyze the types of words occurring before and after numbers. You will provide a word cloud of the top-20 common terms appearing before and after the numbers, and write a paragraph analyzing this data.

Submit your code and results (PDF). The PDF file should have the link to your GitHub. (One submission per team)