Course: **Natural Language Processing [A]** 03-July-2023 **(Spring 2023)**

Resource Person: **Muhammad Shakeel** **ASSIGNMENT–1 (Text Preprocessing)**

Total Points: 20
Submission Due: Saturday, July 15, 2023

Instructions: Please Read Carefully!

- **This is an individual assignment. Everyone is expected to complete the given assignment on their own, without seeking any help from any website or any other individual.** There will be strict penalties for any work found copied from any source and the university policy on plagiarism will be strictly enforced.
- You are expected to submit this assignment as:
    a. Create a **single Python Jupyter Notebook file** for the assignment solution.
    b. Create a **PDF** file of your Notebook file.
    c. Create a **zip file** having name as your ID. Add both files (Jupyter Notebook file and the PDF) in the zip file and TURN IN against this assignment. **Do not create a .rar file.**
- Assignment is to be submitted on the **Google Classroom only**.

ASSIGNMENT DESCRIPTION [20]

You have been given a Twitter corpus as a CSV file containing large number of tweets. You are only going to use the **TweetText** field of this file in this assignment. Please open the CSV file and understand its structure.
For this assignment, you will need to create a **Jupyter Notebook file** that uses the **spaCy library** and performs the following operations on this corpus:

1. Counts and prints the number of **@, #, commas,** and **URLs** in the corpus using appropriate messages.
2. Remove **@, #, commas,** and **URLs** from the first **100 tweets,** and then print each resulting tweet line-by-line.
3. Performs Tokenization of the first **100** tweets.
4. Performs Lemmatization of the first **100** tweets.

You will need to use **spaCy** for performing the Tokenization and the Lemmatization operations. You may use **Google Colab** to setup spaCy as it will be easier to work with in this fashion.

For each operation, use the markdown cells to describe how have you achieved each operation. You should first describe your logic and then write code below your descriptions. Give proper headings to your descriptions. When writing about the spaCy tokenization and lemmatization functions, please give details on how those functions work and how they are called.

Finally, display all code cell outputs of your Notebook and convert to a PDF file so that one can easily read your complete code, their outputs, and descriptions.

Submit your Notebook and the resulting PDF files as one zip file.

END OF ASSIGNMENT