University of Management and Technology, Lahore Campus

Assignment – 01: Fall 2025

Course Name:	NLP	**Course Code:**	CC438	Credit Hours:	3
Course Instructor/s:	Mr. Waqar Ashiq	**Program Name**	BS Software Engineering		
Semester/ Section:	BSSE W1	**Maximum Marks:**	10		
Submission Deadline:	November 04, 2025	**Moderator Signature**			

Instructions:

- Understanding the problem is a part of the task.
- Your assignment should be submitted in PDF format.
- There is a ZERO tolerance policy for plagiarism. Tasks found with copy-pasted material or copied from fellows will be marked as ZERO.
- Submit your assignments on LMS within the given time frame. Emailed Tasks will not be accepted.
- I can take either quiz or in-class viva for this assignment by calling anyone from the class.
- Answer must be to the point.

Assignment: Text Preprocessing and Regular Expressions in NLP

Course: NLP

Topic Covered: Text Preprocessing – Regular Expressions (Regex)

Submission: A Jupyter Notebook file (.ipynb) uploaded on LMS/email

NOTE: This Topic is Part of Quiz, Exam & Viva

Task Description

You are provided with a **Jupyter Notebook** containing examples of common **regular expressions** used in NLP (removing numbers, special characters, URLs, mentions, hashtags, etc.).

Your task is to:

1. **Understand the Example Notebook:**
   - Carefully read and run the given notebook on your systems.
   - Explore each regex function and understand how it modifies text.
2. **Create Your Own Notebook:**

    - Make a **new Jupyter Notebook** with the title: `YourRollNumber_Assignment_01.ipynb`
3. **Apply all preprocessing steps and Regex on a New Dataset:**
    - Collect or create a **small text dataset** (e.g., 10–15 short text sentences such as tweets, reviews, or news headlines).
    - Apply all text preprocessing steps discussed in the class (i.e. Tokenization, Stop words removal, stemming, lemmatization etc...)
    - Apply regex operations to clean and analyze the text. Examples:
        - Remove numbers, punctuation, and special symbols.
        - Extract hashtags, mentions (@user), or emails.
        - Find all words starting with a capital letter.
        - Tokenize sentences into words using regex.
        - Replace multiple spaces with a single space.
4. **Document Your Work:**
    - Clearly explain each regex pattern in **Markdown cells**.
    - Show **before and after results** of applying regex on text.

Submission Instructions

- File Name: `YourRollNumber_Regex_Assignment.ipynb`
- Upload the notebook to LMS/email before the deadline.
- Late submissions will result in **mark deduction**.

✅ **Reminder:** This assignment will also help you in **Quizzes, Exams, and Viva**, so ensure you understand the logic instead of just copying code.