

NLP Assignment: Working with IMDB Movie Reviews Dataset

Objective:

In this assignment, you will work with the IMDB Movie Reviews dataset to practice core NLP and data preprocessing concepts discussed in class. The dataset contains two columns:

- text: Movie review text
- label: Sentiment (0 = Negative, 1 = Positive)

You will focus on data loading, exploration, preprocessing, and basic text analysis.

Dataset:

IMDB Movie Dataset (CSV)

Columns: text, label

Tasks:

Task 1: Load and Understand the Dataset

- Load the dataset using pandas
- Display the first 5 and last 5 rows
- Check dataset shape, column names, and data types
- Explain what each column represents

Task 2: Mini Exploratory Data Analysis (EDA)

- Count the number of positive and negative reviews
- Create two new columns:
 - text_length: number of characters
 - word_count: number of words

- Find the review with the maximum word count

Task 3: Text Cleaning Using Regex

Apply the following preprocessing steps using regex:

- Remove URLs
- Remove numbers
- Remove extra spaces

Final Submission:

- Python code(.ipynb file)
- Name the file: <YourName>.ipynb

Evaluation Criteria (15 Marks):

- Correct implementation (8)**
- Code readability & comments (3)**
- Explanations & insights (4)**