# Introduction to Data Science

## Final Term Project

## Project Title: Text Analysis and Topic Modeling

## Objective:

The goal of this project is to develop practical skills in web scraping, text preprocessing, visual analytics, and topic modeling using real-world news articles. Students will extract news content from a news portal, clean and analyze the text, generate insightful visualizations, and uncover underlying topics in the corpus.

## Project Tasks

### 1. Web Scraping

Select a news portal (preferably Bengali news portal).

Scrape news article texts from multiple pages linked in the home page.

Save the extracted content into a .csv file with at least the following columns: url, article_text

### 2. Text Preprocessing

Perform the following preprocessing steps:

Remove punctuation, numbers, symbols, extra spaces

Convert to lowercase (if English)

Tokenize the text

Remove stop words (custom stop word list)

Apply stemming/lemmatization (if applicable)

***Document the steps with examples and code.***

### 3. Exploratory Text Analysis

Generate a word cloud to visualize the most frequent words.

Create a bar chart of the top 20 most common words.

Analyze which words appear frequently and comment on any patterns.

Save the visualizations as PNG/JPG images.

**4. Topic Modeling**

Construct a Document-Term Matrix (DTM) or Term-Document Matrix (TDM).

Apply Latent Dirichlet Allocation (LDA) to identify 3–5 main topics from the dataset.

Visualize:

Top words per topic using bar plots.

Document-topic distribution using appropriate charts (e.g., stacked bar or pie).

Use packages like topicmodels, tidytext, or textmineR.

**5. Reporting and Visualization**

Combine all graphs and visual insights into a structured PDF report.

Each section (scraping, preprocessing, word analysis, topic modeling) should have:

Plots/graphs/tables

Short interpretation of results

**Submission Requirements**

R Script files

CSV file of scraped text

A pdf report highlighting key findings in each step

**Deadline:** 30th May 11:59 PM

**Late submission will be penalized.**