**AMERICAN INTERNATIONALA close up of a sign

Description automatically generated**

**UNIVERSITY-BANGLADESH**

Choose an item.

**Final Project**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assignment Title: | Text Analysis and Topic Modeling | | | |
| Assignment No: |  | | Date of Submission: | 30 May 2025 |
| Course Title: | Introduction to Data Science | | | |
| Course Code: |  | | Section: | G |
| Semester: | Spring | 2024-25 | Course Teacher: | DR. ASHRAF UDDIN |

**Declaration and Statement of Authorship:**

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student’s work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaboration has been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the faculty for review and comparison, including review by external examiners.
7. I/we understand that Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of their arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

*\* Student(s) must complete all details except the faculty use part.*

\*\* Please submit all assignments to your course teacher or the office of the teacher concerned.

|  |  |
| --- | --- |
| Group Name/No.: | 07 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Name** | **ID** | **Program** | **Contribution** |
| 1 | **ARSHAD ABEDIN ABIR** | **22-47188-1** | **BSc [CSE]** |  |
| 2 | **KAMRUZZAMAN SONY** | **22-46797-1** | **BSc [CSE]** |  |
| 3 | **MRIDUL KANTI KUNDU** | **22-47182-1** | **BSc [CSE]** |  |
| 4 | **MAYSHA MARIUM** | **22-47197-1** | **BSc [CSE]** |  |
| 5 |  |  | **Choose an item.** |  |
| 6 |  |  | **Choose an item.** |  |
| 7 |  |  | **Choose an item.** |  |
| 8 |  |  | **Choose an item.** |  |
| 9 |  |  | **Choose an item.** |  |
| 10 |  |  | **Choose an item.** |  |

|  |  |  |
| --- | --- | --- |
| ***Faculty use only*** | | |
| FACULTY COMMENTS | **Marks Obtained** |  |
|  |
|  |
|  | **Total Marks** |  |
|  |
|  |

# Introduction

# The purpose of this project is to build practical skills in web scraping, text preprocessing, visual analytics, and topic modeling using real-world news articles. The project involves extracting news content from a Bengali news portal, cleaning and analyzing the text, generating insightful visualizations, and uncovering hidden topics within the news dataset. This approach will help develop a deeper understanding of text analysis techniques and their application to natural language data from online media.

# Web Scraping

# The first step of the project involves selecting a news portal, preferably a Bengali news website, to extract news articles. The web scraping process entails collecting news article texts from multiple pages linked on the homepage of the portal. After successfully scraping the data, the extracted content will be saved into a CSV file. This file will contain at least two essential columns: the URL of the news article and the corresponding article text. This structured data will then serve as the basis for further preprocessing and analysis.

A screenshot of a computer

AI-generated content may be incorrect.

### Text Preprocessing

Text preprocessing is a fundamental step to preparing raw text data for meaningful analysis. It involves cleaning and transforming the text to reduce noise, standardizing the format, and converting it into a suitable form for modeling. The main steps include cleaning the text, removing stop words, and tokenization.

#### 3.1 Clean the Text

The cleaning step removes unnecessary characters and standardizes the text. The following actions are performed:

* **Remove non-Bengali characters:** Any character that is not a Bengali letter or whitespace is removed.
* **Remove punctuation marks:** All punctuation symbols are eliminated.
* **Remove Bengali digits:** Bengali numerals (০ to ৯) are removed.
* **Remove short words:** Words with lengths between 1 to 3 characters are removed to reduce noise.
* **Trim extra spaces:** Multiple spaces are replaced by a single space, and leading/trailing spaces are trimmed.

A screenshot of a computer

AI-generated content may be incorrect.

This process helps reduce irrelevant content and prepares the text for deeper analysis.

#### 3.2 Remove Stop Words

A custom list of commonly occurring but non-informative Bengali stop words is defined. These include words such as

A group of colorful letters

AI-generated content may be incorrect.

among others. Removing these stop words from the cleaned text helps to focus the analysis on meaningful terms that contribute more significantly to the text's content.

#### 3.3 Tokenization

Tokenization is the process of splitting each cleaned document into individual words, called tokens. This allows the text to be analyzed at the word level. For example, using R's tokenize\_words() function, each article’s text is broken down into a list of words. These tokens are the basic units used for tasks such as frequency analysis, creating document-term matrices, and topic modeling.



A group of text on a white background

AI-generated content may be incorrect.

A group of text on a white background

AI-generated content may be incorrect.

A group of text on a white background

AI-generated content may be incorrect.

# Exploratory Text Analysis

A Term Document Matrix (TDM) was created to analyze term frequencies across documents. Word frequencies were calculated and the most common words identified. Visualizations such as word clouds and bar plots of the top 20 words were generated to aid intuitive exploration of the data.

***4.1 Word Cloud with Most Frequent Word***

A circle of multicolored text

AI-generated content may be incorrect.

Using the cleaned text corpus, a Term-Document Matrix (TDM) was created to count the frequency of each word across all documents. The word frequencies were then sorted in descending order. The wordcloud() function was applied to this data to generate a colorful word cloud, visually emphasizing the most frequent words by their size. This image was saved as wordcloud.png for easy reference.

***4.2 Bar Chart of Top 20 Words***

A graph of words with numbers

AI-generated content may be incorrect.

From the word frequency data, the top 20 most frequent words were extracted. A bar chart was plotted using ggplot2, displaying these words on the y-axis and their frequencies on the x-axis. The chart was designed with horizontal bars, sorted from highest to lowest frequency, providing a clear quantitative visualization. This plot was saved as top20\_words\_barplot.png.

 **Analyzing**

By reviewing the word cloud and bar chart, common and important terms within the news articles were identified, revealing key themes and patterns in the dataset. Both visualizations were saved as PNG images to include in reports and presentations.

# Topic Modeling

Topic modeling is a technique used to automatically discover the main themes or topics present in a large collection of text documents. It helps to organize and summarize the text by grouping words that frequently appear together into meaningful topics.

#### 5.1 Document-Term Matrix (DTM)

This step involves creating a structured matrix where rows represent documents and columns represent terms (words). Each cell in the matrix contains the frequency of a particular term in a specific document. This matrix serves as the foundational input for topic modeling algorithms, allowing them to analyze the distribution of words across documents systematically.

#### 5.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic model used to uncover hidden thematic structures in a large collection of text documents. By applying LDA to the Document-Term Matrix, the algorithm identifies 3 to 5 main topics. Each topic is represented as a distribution over words that frequently occur together, revealing underlying themes or subjects within the dataset.

#### 5.3 Top Words per Topic

After the topics are identified, bar plots are generated to display the top significant words for each topic. These visualizations make it easier to interpret the content of each topic by highlighting the most influential words, aiding in understanding the essence of the themes discovered.

A graph with different colored bars

AI-generated content may be incorrect.

#### 5.4 Document-Topic Distribution

To understand how topics are spread across individual documents, charts like stacked bar plots or pie charts are created. These visualizations show the proportion or probability of each topic within each document, illustrating the thematic composition of the documents and highlighting dominant topics.

A graph of different colored vertical lines

AI-generated content may be incorrect.

# Results

The most frequent terms corresponded to common news topics such as politics, national affairs, and social issues. The LDA model uncovered coherent topics representing major themes within the dataset. Document-topic analysis showed a mixture of articles dominated by single or multiple topics, indicating diverse news reporting styles and thematic overlaps.

A screenshot of a computer

AI-generated content may be incorrect.