**AMERICAN INTERNATIONALA close up of a sign

Description automatically generated**

**UNIVERSITY-BANGLADESH**

**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]** | **Text Preprocessing** |
| 2 | **KAMRUZZAMAN SONY** | **22-46797-1** | **BSc [CSE]** | **Topic Modeling** |
| 3 | **MRIDUL KANTI KUNDU** | **22-47182-1** | **BSc [CSE]** | **Exploratory Text Analysis** |
| 4 | **MAYSHA MARIUM** | **22-47197-1** | **BSc [CSE]** | **Web Scraping** |
| 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** |  |
|  |
|  |

# 1.Introduction

The purpose of this project is to analyze real-world news articles through web scraping, text preprocessing, visual analytics, and topic modeling. We have extracted news content from the Bengali news portal [**https://www.banglanews24.com/**](https://www.banglanews24.com/), which provides a rich source of current news articles in Bengal. It involves extracting news content from a Bengali news portal, cleaning and analyzing the text, creating meaningful visualizations, and uncovering hidden topics within the news dataset. This approach aims to gain valuable insights from the information and apply them to natural language data sourced from online media, thereby enhancing understanding of the news content.

# 2.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.

Figure 01: CSV Preview of Scraped News Articles

### **3.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.

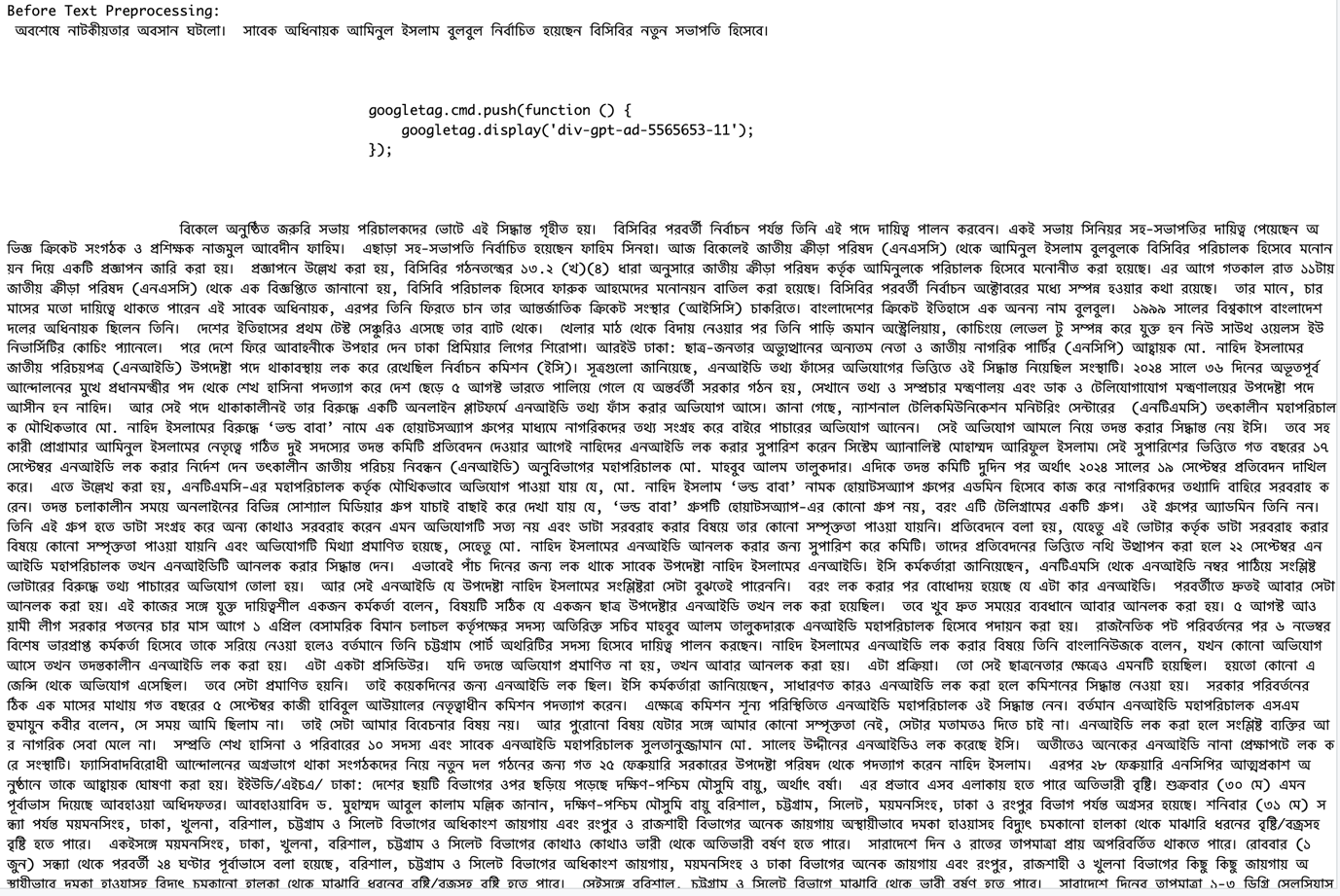


Figure 02: Raw Bengali Text Before Preprocessing

#### **3.1 Clean the Text**

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

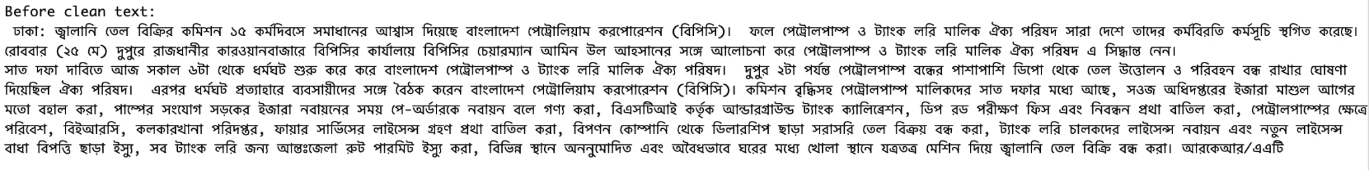
* **Before Clean Text:**

Figure 03: Bengali Text Before Cleaning Text

* **Remove non-Bengali characters:** Any character that is not a Bengali letter or whitespace is removed.

A close up of text

AI-generated content may be incorrect.

Figure 04: Bengali Text After Remove Non-Bengali Characters

* **Remove punctuation marks:** All punctuation symbols are eliminated.

A close up of text

AI-generated content may be incorrect.

Figure 05: Bengali Text After Remove **punctuation marks**

* **Remove Bengali digits:** Bengali numerals (০ to ৯) are removed.

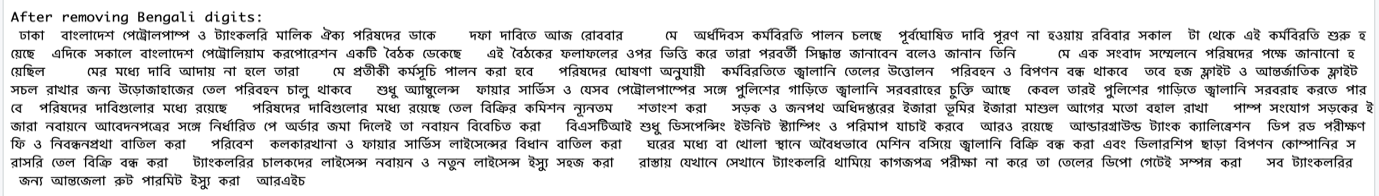


Figure 06: Bengali Text After **Remove Bengali digits**

* **Remove short words:** Words with lengths between 1 to 3 characters are removed to reduce noise.

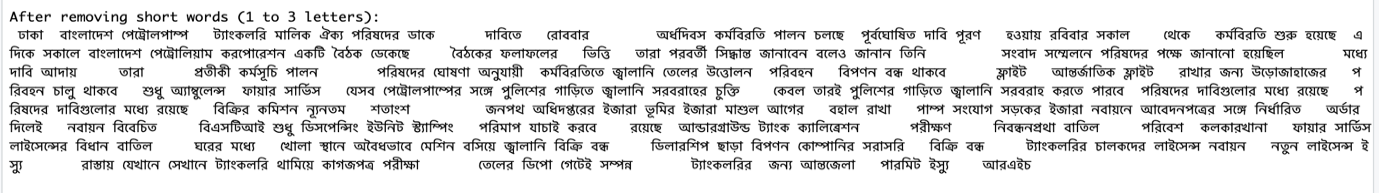


Figure 07: Bengali Text After Remove short words

* **Squishing whitespace:** Multiple spaces are replaced by a single space, and leading/trailing spaces are trimmed.

A close up of text

AI-generated content may be incorrect.

Figure 08: Bengali Text After **Squishing whitespace**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 09: CSV Preview of Cleaned Bengali News Text

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 close-up of a pattern of green and white text

AI-generated content may be incorrect.

Figure 10: Custom Bengali Stop Words List

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.



Figure 11: Bengali Text After Removing Bengali Stop Words

#### **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.

Figure 12: Tokenized Bengali Words

# 4.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.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 13: CSV Preview of Bengali Words by Frequency

4.1 **Word Cloud with Most Frequent Word**

A circle of multicolored text

AI-generated content may be incorrect.

Figure 14: Word Cloud of Most Frequent Bengali Words

This word cloud provides a visual summary of word frequencies found in a collection of Bengali text documents, most likely related to news or political topics.

1. **Most Frequent Words:**
   * **জিয়াউর (Ziaur):** This is the largest word, showing it is the most commonly mentioned term in the text. It probably refers to “Ziaur Rahman” or the Zia political faction, indicating a political focus.
   * **সরকার (Government):** Indicates discussions involving government policies, activities, or official statements.
   * **ঢাকা (Dhaka):** The capital of Bangladesh, frequently mentioned in political news and national events.
   * **ঘোষণা (Announcement):** Points to frequent mentions of official declarations or statements.
2. **Political Context:**
   * Words such as **নির্বাচন (Election), বিএনপি (BNP - Bangladesh Nationalist Party), সদস্য (Member), দলের (Party’s)** suggest that the text mainly covers political events, party actions, and elections.
   * Terms like **স্বাধীনতা (Independence), শ্রদ্ধা (Respect/Honor), শহীদ (Martyr)** reflect themes of nationalism and historical significance within political narratives.
3. **Geographical and Social References:**
   * The prominence of **ঢাকা (Dhaka)** highlights the central role of the capital in the news or discussion topics.
   * Other terms like **পানি (Water)** and **বৃষ্টি (Rain)** hint at social or environmental issues alongside political reporting.
4. **Other Frequent Terms:**
   * Names and titles like **প্রধান (Chief/Head), মেয়র (Mayor), আহমেদ (Ahmed)** suggest mentions of specific leaders or officials.
   * Words such as **কারণ (Reason), বিরুদ্ধে (Against), প্রথম (First)** indicate the presence of narrative or argumentative elements typical in news stories.

This word cloud reflects a text dataset rich in political and governmental themes, focusing on prominent figures likely connected to the Zia family or faction, the government, and national matters centered mainly on Dhaka. It also includes related topics such as elections, official announcements, social/environmental concerns, and expressions of national pride and historical remembrance.

4.2 **Bar Chart of Top 20 Words**

A graph of words with numbers

AI-generated content may be incorrect.

Figure 15: Bar Chart of Top 20 Most Frequent Bengali Words

This bar chart presents the top 20 most frequent Bengali words from a text corpus, likely centered on political and social topics. The word “জিয়াউর” (referring to Ziaur Rahman or his political faction) appears most frequently, followed by terms like “তারা” (they), “যেখান” (where), and “ঢাকা” (the capital city), indicating a strong focus on political figures, places, and government-related discussions. Words such as “সরকার” (government), “নির্বাচন” (election), and “বিএনপি” (BNP) further highlight themes of political events, party activities, and electoral processes. Additionally, nationalistic terms like “স্বাধীনতা” (independence) and “শ্রদ্ধা” (respect) appear alongside social and environmental references such as “পানি” (water) and “বৃষ্টি” (rain), reflecting a broader context of news coverage. The presence of official titles and narrative words like “প্রধান” (chief), “মেয়র” (mayor), “কারণ” (reason), and “বিরুদ্ধে” (against) suggests detailed reporting and argumentation within the texts. Overall, the chart offers a clear quantitative visualization that complements the word cloud, emphasizing the political and social nature of the dataset with a strong geographic focus on Dhaka.

# 5.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.

Figure 16: Bar Charts of Top Words per Topic

### **Topic 1 (Red Bar) Word Analysis:**

* **জিয়াউর, জিয়া**: These words refer to prominent political figures in Bangladesh, specifically related to Ziaur Rahman and his political legacy, indicating a political theme.
* **দিয়ে**: A verb indicating an action, likely used in narratives describing events or processes.
* **পানি**: Refers to water, suggesting environmental or natural resource topics.
* **তাদের, সালে, শুরু**: Words related to time and groups of people, indicating the start or timeline of certain events.
* **সালেন, জাতীয়, শহীদ**: Words that relate to national events or commemorations, such as martyrdom or national days.

### **Topic 2 (Olive Bar) Word Analysis:**

* **ঘোষণা**: Means "announcement," indicating official statements or declarations.
* **জিয়াউর, জিয়া, রহমানের**: Names related to political leaders, linking this topic to political history or current politics.
* **স্বাধীনতার**: Refers to "independence," indicating discussions about national freedom or historical struggles.
* **মেজর, মার্চ, সালের**: Military rank and time indicators, likely referencing specific dates or periods.
* **হিসেবে**: Means "as" or "in the capacity of," showing roles or positions.

### **Topic 3 (Green Bar) Word Analysis:**

* **শনি, বৃষ্টি**: Related to weather, specifically rain and days of the week (Saturday).
* **ঢাকা**: The capital city of Bangladesh, a central location for political, social, and environmental news.
* **বৃহস্পতিবার, গেছেকারণে**: Time references and causal words, indicating when and why something happened.
* **বিভিন্ন, করেছে, প্রধান**: Describing various events or actions with importance.

### **Topic 4 (Blue Bar) Word Analysis:**

* **বিএনপি**: Refers to the Bangladesh Nationalist Party, a major political party.
* **হিসেবে**: Denotes roles or perspectives.
* **জিয়াউর, রহমানের**: Political leader names, indicating political discussions.
* **শহীদ**: Refers to martyrs, highlighting historical or commemorative themes.
* **করেছেন, হচ্ছে**: Action verbs indicating ongoing or completed activities.
* **সারেক, ঢাকা**: Names of places or organizations, indicating local context.

### **Topic 5 (Purple Bar) Word Analysis:**

* **সরকার, সরকারের**: Refers to government and related administrative matters.
* **ঢাকা**: Political and administrative hub.
* **ক্রিকেট**: Refers to the sport, indicating social or entertainment-related topics.
* **নির্বাচন**: Refers to elections, pointing to democratic processes and political events.
* **কিছু, ফারুক, সংস্কার, করার, জুলাই**: Various terms relating to actions, individuals, reforms, and time (July).

Each topic’s key words reflect important aspects of Bangladesh’s political landscape, social environment, and daily life. Political figures and parties such as **জিয়াউর** and **বিএনপি**, as well as terms like **সরকার** and **নির্বাচন**, highlight the political discourse. Environmental and social themes emerge from words like **পানি**, **বৃষ্টি**, and **ঢাকা**, while entertainment and cultural life appear through words like **ক্রিকেট**. Time-related words such as **মার্চ**, **শনি**, and **জুলাই** emphasize temporal contexts within the data. Overall, the dataset provides a rich representation of the multifaceted realities of Bangladesh’s society, politics, and environment.

#### **5.4 Document-Topic Distribution**

A graph of different colored vertical lines

AI-generated content may be incorrect.

Figure 17: 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.