AMERICAN INTERNATIONAL 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	

1. 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.

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.

Data Preview:

url *(character)*	article_text *(character)*
https://www.banglanews24.com/banglanews-special...	রাজধানীর প্রাণকেন্দ্রে নির্মল বাতাসে প্রাণভরে শ্বাস নেওয়ার জা...
https://www.banglanews24.com/politics/news/bd/15...	ঢাকা: প্রধান উপদেষ্টা অধ্যাপক ড. মুহাম্মদ ইউনূসের একটি সা...
https://www.banglanews24.com/climate-nature/new...	ঢাকা: স্থল গভীর নিম্নচাপ ধীরে ধীরে দুর্বল হলেও এর প্রভাবে এ...
https://www.banglanews24.com/national/news/bd/1...	প্রধান উপদেষ্টা অধ্যাপক ড. মুহাম্মদ ইউনূস ও জাপানের প্রধা...
https://www.banglanews24.com/environment-biodiv...	ঢাকা: গভীর নিম্নচাপ স্থলভাগে উঠে আসার পর দেশে মৌসুমে...
https://www.banglanews24.com/saradesh/news/bd/...	নোয়াখালী: বঙ্গোপসাগরে সৃষ্ট গভীর নিম্নচাপের প্রভাবে নোয়াখা...
https://www.banglanews24.com/cricket/news/bd/15...	জাতীয় ক্রীড়া পরিষদ (এনএসসি) কর্তৃক ফারুক আহমেদের প...
https://www.banglanews24.com/politics/news/bd/15...	ঢাকা: সমগ্র জাতি একটি সুষ্ঠু, সুন্দর, নিরপেক্ষ নির্বাচনের জ...
https://www.banglanews24.com/opinion/news/bd/15...	শহীদ প্রেসিডেন্ট জিয়াউর রহমান আমাদের জাতীয় ইতিহাসের ...
https://www.banglanews24.com/saradesh/news/bd/...	ভোলা: নিম্নচাপের প্রভাবে ভোলায় নদ-নদীর পানি অস্বাভাবিক...
https://www.banglanews24.com/saradesh/news/bd/...	দিনাজপুর: দিনাজপুরের পার্বতীপুরে ট্রাক আটকে চাঁদাবাজির স...
https://www.banglanews24.com/daily-chittagong/ne...	চট্টগ্রাম: জোড়া লাগানো যমজ শিশু রিয়াশাদ ও রেনিশকে জ...
https://www.banglanews24.com/politics/news/bd/15...	ঢাকা: মহান স্বাধীনতার ঘোষক ও বিএনপির প্রতিষ্ঠিতা শহীদ প্রে...

Previewing first 50 entries.

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.

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.

Data Preview:

url (character)	cleaned_text (character)
https://www.banglanews24.com/banglanews-special...	রাজধানীর প্রাণকেন্দ্রে নির্মল বাতাসে প্রাণভরে শ্বাস নেওয়ার জা...
https://www.banglanews24.com/politics/news/bd/15...	ঢাকা প্রধান উপদেষ্টা অধ্যাপক মুহাম্মদ ইউনূসের সাক্ষাৎকারের...
https://www.banglanews24.com/climate-nature/new...	ঢাকা স্থল গভীর নিম্নচাপ ধীরে ধীরে দুর্বল হলেও প্রভাবে এখনো ...
https://www.banglanews24.com/national/news/bd/1...	প্রধান উপদেষ্টা অধ্যাপক মুহাম্মদ ইউনূস জাপানের প্রধানমন্ত্রী ...
https://www.banglanews24.com/environment-biodiv...	ঢাকা গভীর নিম্নচাপ স্থলভাগে আসার দেশে মৌসুমের রেকর্ড বৃ...
https://www.banglanews24.com/saradesh/news/bd/...	নোয়াখালী বঙ্গোপসাগরে সৃষ্ট গভীর নিম্নচাপের প্রভাবে নোয়াখা...
https://www.banglanews24.com/cricket/news/bd/15...	জাতীয় ক্রীড়া পরিষদ এনএসসি কর্তৃক ফারুক আহমেদের পরি...
https://www.banglanews24.com/politics/news/bd/15...	ঢাকা সমগ্র জাতি সুষ্ঠু সুন্দর নিরপেক্ষ নির্বাচনের অপেক্ষমাণ ...
https://www.banglanews24.com/opinion/news/bd/15...	শহীদ প্রেসিডেন্ট জিয়াউর জাতীয় ইতিহাসের অবিস্মরণীয় জাতি...
https://www.banglanews24.com/saradesh/news/bd/...	ভোলা নিম্নচাপের প্রভাবে ভোলায় নদীর পানি অস্বাভাবিকভাবে ...
https://www.banglanews24.com/saradesh/news/bd/...	দিনাজপুর দিনাজপুরের পার্বতীপুরে ট্রাক আটকে চাঁদাবাজির তা...
https://www.banglanews24.com/daily-chittagong/ne...	চট্টগ্রাম জোড়া লাগানো শিশু রিয়াশাদ রেনিশকে জন্মের ঘণ্টার ...
https://www.banglanews24.com/politics/news/bd/15...	ঢাকা মহান স্বাধীনতার ঘোষক বিএনপির প্রতিষ্ঠিতা শহীদ প্রেসি...

Previewing first 50 entries.

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

```
bangla_stopwords <- c(
  "ও","করে","এবং","না","থেকে","তিনি","করা","ন","একটি","বাংলাদেশ","হবে","এক","ই","সঙ্গে",
  "বাংলাদেশের","নিয়ে","পর","রহমান","করতে","মে","কোনো","র","দেশের","আমাদের","যে","পারে","মধ্যে","হিবে","টাকা",
  "এর","এবং","না","থেকে","তিনি","করা","এ","হয়","তার","বলেন","কিন্তু","তবে","হয়েছে","ছিল","যা","সে",
  "এই","হয়","জন্য","না","ওরে","তুমি","আমি","তারা","আমরা","আপনি","তাহলে","করেন","হয়েছে","ড"
)
```

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.

```
  [1] "রাজধানীর"      "প্রাণকেন্দ্রে"   "নির্মল"       "বাতাসে"       "প্রাণভরে"
  [6] "শ্বাস"          "নেওয়ার"         "জায়গার"      "অভাব"         "অভাবের"
 [11] "জায়গা"         "কিছুটা"          "হলেও"         "পূরণ"         "করছিল"
 [16] "পানি"           "সবুজে"           "মেশা"         "হাতিরঝিল"     "রাস্তার"
 [21] "দুধারে"         "সবুজে"           "গাছের"        "ছায়ায়"      "দুদণ্ড"
 [26] "শান্তি"         "খুঁজতে"          "অনেকেই"       "ছুটে"         "আসেন"
 [31] "হাতির"          "ঝিলে"            "ছুটির"        "দিনে"         "ঘুরতে"
 [36] "আসেন"           "পরিবারসহ"        "প্রতিদিন"      "সন্ধ্যার"      "হাতিরঝিলের"
 [41] "কৃত্রিম"        "আলোর"           "ঝলকানিও"      "নগরের"        "মানুষের"
 [46] "কাছে"           "অন্যতম"          "আকর্ষণের"      "কেন্দ্র"       "এখানে"
 [51] "দর্শনার্থী"      "পথচারী"         "কিংবা"        "আশপাশের"      "বাসিন্দা"
 [56] "সবারই"          "চলাচল"           "চেপে"         "আয়তনের"      "হাতিরঝিল"
 [61] "এফডিসি"         "গুলশান"          "পর্যন্ত"       "সংযুক্ত"      "জলাশয়ের"
 [66] "পানি"           "দুর্গন্ধমুক্ত"     "রাখতে"        "কর্তৃপক্ষের"    "উদ্যোগ"
 [71] "থাকলেও"         "দায়িত্ব"         "পাওয়া"       "ঠিকাদারি"     "প্রতিষ্ঠানের"
 [76] "কারণে"          "সেটি"            "দৃশ্যমান"      "ঝিলের"        "পানি"
 [81] "পরিষ্কার"        "রাখতে"           "বছরের"        "ঠিকাদারি"     "প্রতিষ্ঠান"
 [86] "মেহজাবিন"       "এন্টারপ্রাইজকে"    "কোটি"         "বরাদ্দ"       "দেওয়া"
 [91] "অগ্রগতির"       "তেমন"            "প্রমাণ"        "মধুবাগ"       "এলাকার"
 [96] "স্থায়ী"         "বাসিন্দা"         "শাওন"        "আহমেদ"       "চলতি"
[101] "হাতিরঝিলে"      "প্রাণভরে"        "শ্বাস"         "নেওয়ার"      "উপায়"
[106] "সকালে"          "বাচ্চাদের"       "দিয়ে"         "হাঁটতে"       "হতাম"
[111] "এমনই"          "অবস্থা"          "নিজের"         "ঠিকানা"       "এলাকা"
[116] "ছেড়ে"          "যেতাম"           "প্রকট"         "দুর্গন্ধে"      "হাতিরঝিল"
[121] "এলাকার"         "মানুষের"         "স্বাস্থ্যঝুঁকি"   "বাড়ছে"       "এখানে"
[126] "বিশুদ্ধ"         "বাতাস"           "নির্মল"        "পরিবেশ"       "স্থানীয়"
[131] "রফিকুল"         "নামে"            "আরেকজনের"       "হাতিরঝিলে"      "প্রকল্পের"
[136] "শুরুর"          "দিকে"            "ওয়াটার"       "ট্যাক্সিতে"      "গন্তব্যে"
[141] "ফিরতাম"         "অন্যরকম"         "একটা"         "সতেজতা"       "অনুভব"
[146] "চলাই"           "অস্বস্তিকর"       "উঠেছে"        "বাচ্চারা"      "সারা"
[151] "লেকের"          "পাশে"            "বইসা"         "খেলত"         "গন্ধ"
[156] "পানি"           "পোলাপান"         "আসতেই"        "হাতিরঝিল"      "এলাকার"
[161] "বাসিন্দা"        "গৃহিণী"          "হাসনা"        "ঝিলের"        "ধারেই"
[166] "ঝালমুড়ি"       "বিক্রি"          "হাসান"         "মিয়া"         "আগের"
[171] "চেয়ে"          "মানুষ"           "হাতিরঝিলে"      "দূষণ"         "দুর্গন্ধে"
[176] "এলাকাবাসীর"      "পাশাপাশি"        "আমরাও"        "অস্বস্তিতে"      "সালে"
[181] "জনসাধারণের"      "উন্মুক্ত"         "হাতিরঝিল"      "প্রকল্প"       "পানি"
[186] "পরিষ্কার"        "পরিশোধনের"       "পর্যাপ্ত"       "ব্যবস্থা"      "থাকায়"
[191] "পানির"          "গুণগত"           "খারাপ"         "প্রকট"         "আকার"
[196] "ধারণ"           "করেছে"           "পয়োবর্জ্য"      "ময়লা"         "আবর্জনা"
[201] "ড্রেনের"         "পানি"            "ঢুকে"         "বিষাক্ত"       "উঠেছে"
```

[206]	"ঝিলের"	"পানি"	"বাতাসে"	"ভাসছে"	"উৎকট"
[211]	"গন্ধ"	"আশপাশ"	"ঘুরে"	"এমনটি"	"দেখেছেন"
[216]	"প্রতিবেদক"	"হাতিরঝিল"	"ঘুরে"	"দেখা"	"গেছে"
[221]	"ঝিলের"	"প্রায়"	"থেকেই"	"পানির"	"দুর্গন্ধ"
[226]	"ভেসে"	"আসছে"	"পানিতে"	"নানা"	"বর্জ্য"
[231]	"পলিথিন"	"প্লাস্টিকের"	"পাইপও"	"ভাসতে"	"দেখা"
[236]	"গেছে"	"কারওয়ান"	"বাজারের"	"প্যান"	"প্যাসিফিক"
[241]	"সোনারগাঁও"	"হোটেলের"	"পেছনে"	"বর্জ্যের"	"মাত্রা"
[246]	"বেশি"	"কাজী"	"নজরুল"	"ইসলাম"	"অ্যাভিনিউ"
[251]	"থেকেও"	"দুর্গন্ধ"	"পাওয়া"	"ঝিলে"	"মাছও"
[256]	"ভাসতে"	"দেখা"	"পানির"	"ওপরে"	"বর্জ্যের"
[261]	"পুরু"	"স্তর"	"তৈরি"	"ওয়াটার"	"ট্যাক্সিতে"
[266]	"চলার"	"সময়"	"স্পষ্টভাবে"	"দেখা"	"যাচ্ছিল"
[271]	"হাতির"	"ঝিলের"	"পানি"	"জীববৈচিত্র্য"	"পরিবেশের"
[276]	"হুমকি"	"বলছেন"	"পরিবেশবিদরা"	"রাজউকের"	"কর্মকর্তারা"
[281]	"বলছেন"	"প্রায়"	"নর্দমার"	"পানি"	"হাতিরঝিলের"
[286]	"পানিতে"	"মিশে"	"নিষেধাজ্ঞা"	"থাকলেও"	"মানুষ"
[291]	"আবর্জনা"	"পলিথিন"	"এমনকি"	"শিল্প"	"বর্জ্য"
[296]	"ফেলছে"	"ঝিলের"	"পানিতে"	"পানি"	"দূষিত"
[301]	"হচ্ছে"	"ময়লা"	"আবর্জনা"	"মিথেন"	"গ্যাস"
[306]	"সৃষ্টি"	"হচ্ছে"	"কারণেই"	"দুর্গন্ধ"	"ছড়িয়ে"
[311]	"পড়ছে"	"রাজউকের"	"তত্ত্বাবধায়ক"	"প্রকৌশলী"	"যান্ত্রিক"
[316]	"সাব্বির"	"তাহের"	"বাংলানিউজকে"	"হাতিরঝিলের"	"চারপাশে"
[321]	"প্রায়"	"নর্দমা"	"গর্ত"	"রয়েছে"	"প্রতি"
[326]	"মাসে"	"একবার"	"সেগুলো"	"পরিষ্কার"	"কঠিন"
[331]	"বর্জ্য"	"গর্তের"	"পেছনে"	"তারপরে"	"লেকের"
[336]	"পানিতে"	"ভেসে"	"রামপুরা"	"কাঁঠালবাগানে"	"দুটি"
[341]	"স্লুইস"	"রয়েছে"	"বর্ষাকালে"	"এগুলো"	"খুলতে"
[346]	"কঠিন"	"বর্জ্য"	"প্রবেশ"	"প্রতি"	"রক্ষণাবেক্ষণের"
[351]	"বরাদ্দের"	"অংশই"	"বিদ্যুৎ"	"খাতে"	"যায়"
[356]	"পানি"	"পরিশোধনের"	"বরাদ্দ"	"রাসায়নিক"	"প্রায়"
[361]	"হয়ে"	"সালের"	"আগের"	"অবশিষ্ট"	"রাসায়নিক"
[366]	"অল্প"	"পরিমাণে"	"ব্যবহার"	"হচ্ছে"	"বছরে"
[371]	"মাত্র"	"একবার"	"সালে"	"হাতিরঝিল"	"সংগৃহীত"
[376]	"নমুনায়"	"ক্ষতিকর"	"রাসায়নিক"	"পিএফওএ"	"পারফ্লুরোঅকটানোয়িক"
[381]	"অ্যাসিড"	"পিএফওএস"	"পারফ্লুরোঅকটেনসালফোনিক"	"অ্যাসিড"	"উভয়ই"
[386]	"দীর্ঘমেয়াদি"	"বিষাক্ততার"	"দায়ী"	"পিএফওএসের"	"স্তর"
[391]	"পরামর্শমূলক"	"স্তরের"	"চেয়ে"	"বেশি"	"মাত্রায়"
[396]	"পাওয়া"	"গেছে"	"বিভিন্ন"	"গবেষণায়"	"হাতিরঝিলে"
[401]	"পানি"	"দূষণের"	"বিষয়টি"	"এসেছে"	"এরপরও"
[406]	"পানি"	"দূষণ"	"যাচ্ছে"	"বাংলানিউজের"	"বলেছে"

[411]	"রাজউক"	"কর্তৃপক্ষ"	"লেকের"	"নির্বাহী"	"তত্ত্বাবধায়ক"
[416]	"প্রকৌশলী"	"যান্ত্রিক"	"সাব্বির"	"তাহের"	"জানান"
[421]	"হাতিরঝিল"	"লেকে"	"ময়লা"	"এরইমধ্যে"	"তিনটি"
[426]	"বিলে"	"মেহজাবিন"	"এন্টারপ্রাইজকে"	"পরিশোধ"	"পরিশোধ"
[431]	"হলেও"	"ঠিকাদার"	"কোম্পানি"	"মেহজাবিনের"	"দেখা"
[436]	"যায়নি"	"সরেজমিনে"	"হাতিরঝিলের"	"পানি"	"পরিষ্কার"
[441]	"পরিচ্ছন্ন"	"রাখতে"	"বসানো"	"কয়েকটি"	"স্পেশাল"
[446]	"স্যুয়ারেজ"	"ডাইভারশন"	"স্ট্রাকচার"	"এসএসডিএস"	"স্ক্রিনিং"
[451]	"মেশিন"	"মেশিন"	"অপারেটর"	"হিসেবে"	"থাকার"
[456]	"করছেন"	"ছয়জন"	"তাদের"	"বেতন"	"দেওয়া"
[461]	"হাজার"	"এসএসডিএস"	"হোটেল"	"সোনারগাঁও"	"এলাকায়"
[466]	"জাহিদ"	"হাসান"	"গোলাপ"	"এসএসডিএস"	"মগবাজার"
[471]	"সংলগ্ন"	"এলাকায়"	"সেলিম"	"এসএসডিএস"	"মধুবাগ"
[476]	"এলাকায়"	"মামুন"	"এসএসডিএস"	"নিকেতন"	"এলাকায়"
[481]	"ওহিদ"	"শহিদ"	"ছাড়া"	"এসএসডিএস"	"ম্যানুয়াল"
[486]	"সাতটি"	"মেশিন"	"এগুলোর"	"কাজের"	"চারজন"
[491]	"শ্রমিক"	"থাকার"	"যারা"	"ম্যানুয়ালি"	"করবেন"
[496]	"সরেজমিনে"	"একজনকেও"	"পাওয়া"	"যায়নি"	"এমনকি"
[501]	"অন্য"	"অপারেটরদের"	"কাছে"	"জানতে"	"চাইলে"
[506]	"জানান"	"ম্যানুয়াল"	"মেশিনে"	"কর্মচারীই"	"এসএসডিএসের"
[511]	"বর্জ্য"	"সরানোর"	"শ্রমিক"	"গাড়ি"	"লেবার"
[516]	"সপ্তাহে"	"একদিন"	"কারওয়ান"	"বাজার"	"গাড়ি"
[521]	"ভাড়া"	"তিনজন"	"শ্রমিক"	"ভাড়া"	"কাজটি"
[526]	"জানিয়েছেন"	"প্রকাশে"	"অনিচ্ছুক"	"কর্মচারী"	"শুরুতে"
[531]	"জলাশয়"	"পরিষ্কার"	"করার"	"চারজন"	"ডুবুরি"
[536]	"ছিলেন"	"যারা"	"বিষয়"	"জানতে"	"চাইলে"
[541]	"রাজউক"	"প্রকৌশলী"	"সাব্বির"	"তাহের"	"তেমন"
[546]	"কিছু"	"জানাতে"	"পারেননি"	"এমনকি"	"সরেজমিনে"
[551]	"পাওয়া"	"কর্মচারীর"	"সংখ্যার"	"দেওয়া"	"সংখ্যার"
[556]	"তথ্যেরও"	"অমিল"	"পাওয়া"	"ঠিকাদারি"	"প্রতিষ্ঠান"
[561]	"মেহজাবিন"	"এন্টারপ্রাইজের"	"মালিক"	"জাকির"	"হোসেনের"
[566]	"নিজের"	"অধীনে"	"থাকা"	"কর্মীর"	"দেখে"
[571]	"নামই"	"বলতে"	"পারেননি"	"পর্যায়ে"	"সুপারভাইজারকে"
[576]	"দিয়ে"	"জানতে"	"কাজের"	"চিত্র"	"সম্পর্কে"

[581]	"কাছে"	"জানতে"	"চাইলেও"	"সদুত্তর"	"দিতে"
[586]	"পারেননি"	"থেকেই"	"স্পষ্ট"	"নিলেও"	"লেকের"
[591]	"ময়লা"	"পরিষ্কারে"	"অগ্রগতি"	"বরাদ্দের"	"কোথায়"
[596]	"জানতে"	"চাইলে"	"প্রকাশে"	"অনিচ্ছুক"	"ঠিকাদার"
[601]	"কোম্পানির"	"কর্মচারী"	"বিভিন্ন"	"নামে"	"তোলা"
[606]	"হলেও"	"নিজের"	"সহকর্মীদের"	"এখনো"	"দেখেননি"
[611]	"অর্থাৎ"	"বিভিন্ন"	"কর্মচারীর"	"দিয়ে"	"তৈরি"
[616]	"তোলা"	"ঠিকই"	"নিয়োগই"	"রাজউক"	"চেয়ার"
[621]	"ম্যান"	"রিয়াজুল"	"ইসলাম"	"বাংলানিউজকে"	"জানান"
[626]	"দুর্নীতি"	"অনিয়ম"	"তদন্তে"	"কমিটি"	"তদন্ত"
[631]	"প্রতিবেদনের"	"প্রতিবেদন"	"অনুযায়ী"	"দোষীদের"	"বিরুদ্ধে"
[636]	"ব্যবস্থা"	"নেওয়া"	"দূষণে"	"ক্ষতি"	"হচ্ছে"
[641]	"পরিবেশবিদ"	"আহমদ"	"কামরুজ্জামান"	"মজুমদার"	"এলিভেটেড"
[646]	"এক্সপ্রেসওয়ের"	"পিলারই"	"সোনারগাঁও"	"হোটেলের"	"পেছনে"
[651]	"হাতিরঝিল"	"সংলগ্ন"	"এলাকা"	"সুয়ারেজ"	"সিস্টেম"
[656]	"ব্যাহত"	"হচ্ছে"	"মাটি"	"দিয়ে"	"ভরাটের"
[661]	"কারণেও"	"পরিবেশের"	"মারাত্মক"	"ক্ষতি"	"হচ্ছে"
[666]	"পানিতে"	"অক্সিজেনের"	"পরিমাণ"	"গিয়ে"	"ভেসে"
[671]	"উঠছে"	"ছাড়া"	"ঝিলের"	"সামগ্রিক"	"জীববৈচিত্র্যে"
[676]	"মারাত্মক"	"প্রভাব"	"পড়ছে"	"শহরের"	"প্রতিবেশ"
[681]	"ব্যবস্থায়"	"বিরূপ"	"প্রভাব"	"ফেলছে"	"কামরুজ্জামান"
[686]	"ঝিলে"	"আগের"	"ব্যাঙ"	"বিলীন"	"যাচ্ছে"
[691]	"উপকারী"	"ক্ষুদ্র"	"পোকামাকড়ও"	"নগরের"	"মানুষের"
[696]	"স্বাস্থ্য"	"ঝুঁকি"	"বাড়াচ্ছে"	"পানি"	"পরিষ্কার"
[701]	"পরিচ্ছন্ন"	"রাখতে"	"যেটুকু"	"বরাদ্দ"	"রাখা"
[706]	"সেটুকুও"	"ঠিকঠাক"	"কার্যকর"	"হাতিরঝিলের"	"পানির"
[711]	"দুর্দশা"	"নগরের"	"মানুষও"	"স্বস্তি"	"দুর্নীতির"
[716]	"কারণে"	"পরিবেশে"	"বিরূপ"	"প্রভাব"	"পড়ছে"
[721]	"রাজধানীর"	"মানুষের"	"স্বাস্থ্যকে"	"ঝুঁকিতে"	"ফেলছে"
[726]	"কামরুজ্জামান"	"আরএইচ"			

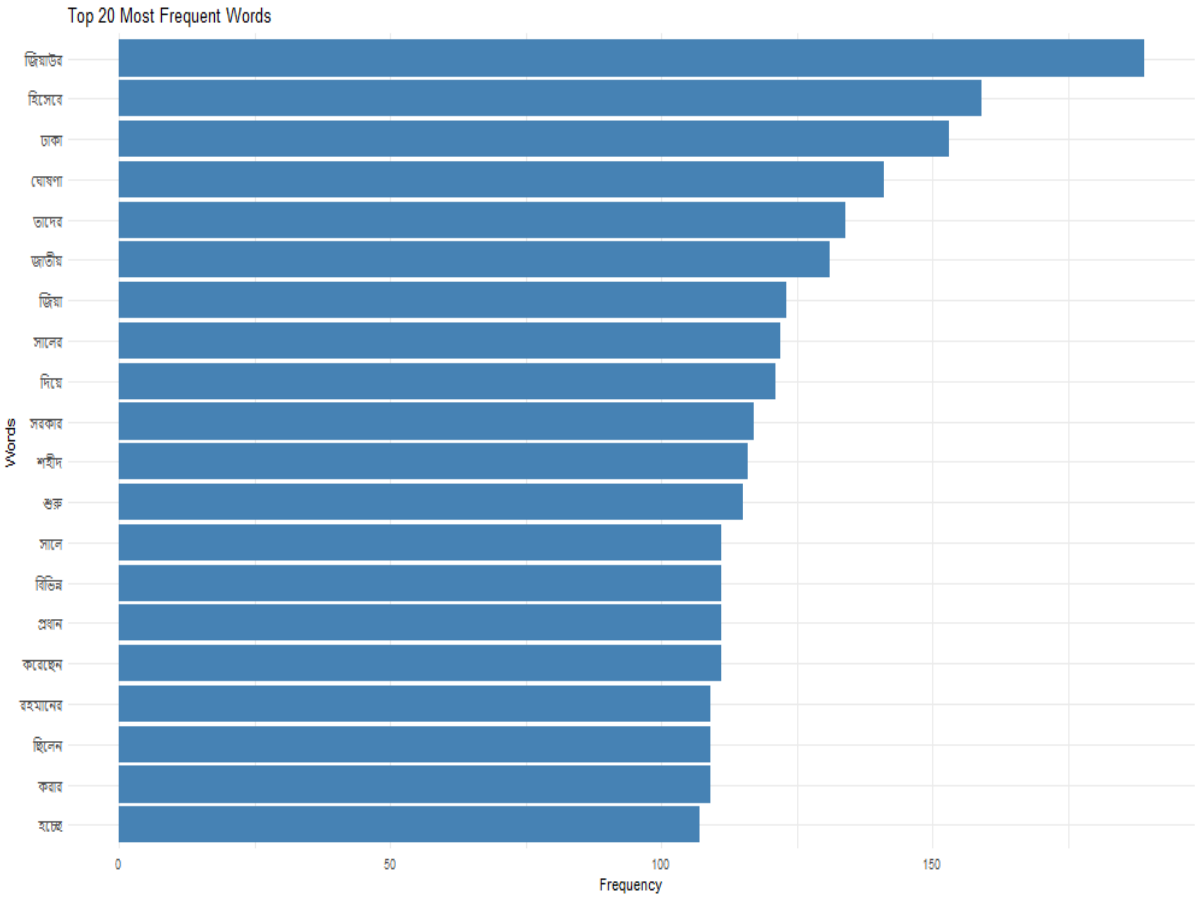
4. Exploratory Text Analysis

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.

4.1 Word Cloud with Most Frequent Word

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



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.

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)

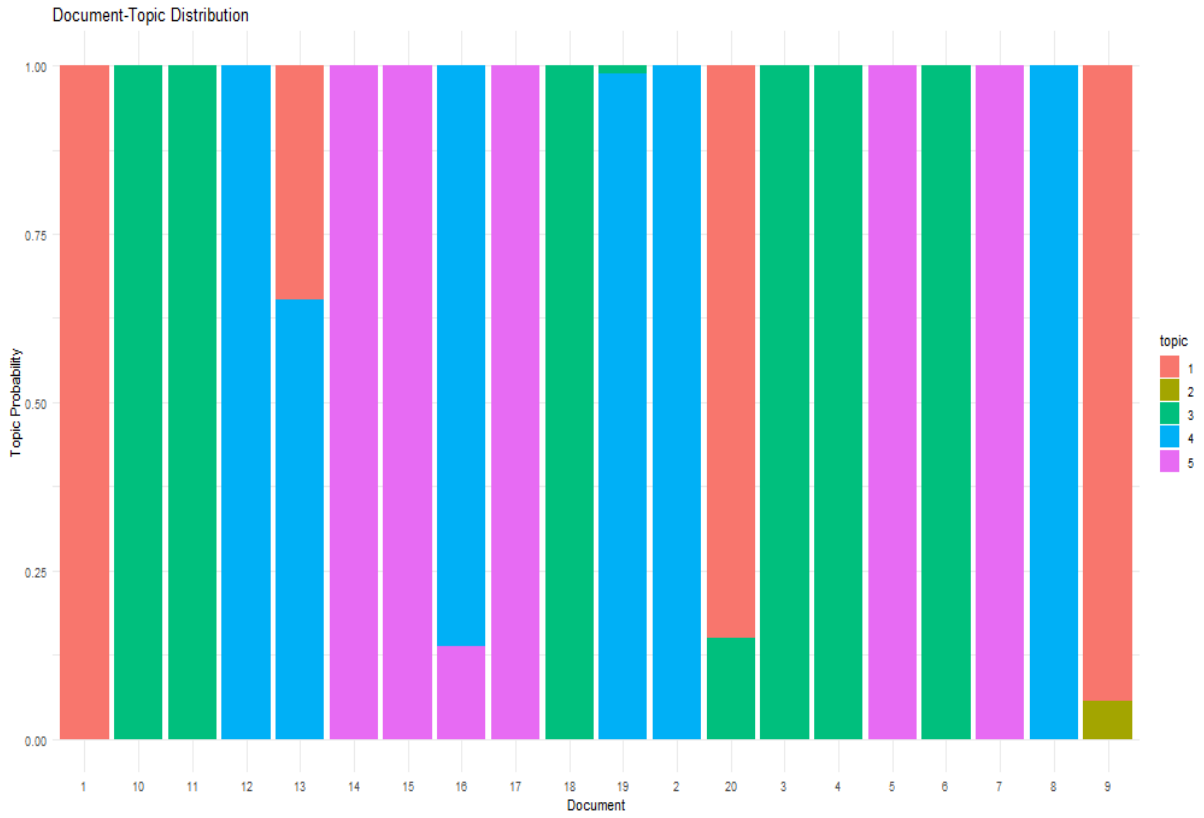
Latent Dirichlet Allocation (LDA) is a probabilistic model used to uncover hidden thematic structures in a large collection of text documents. By applying LDA to the Document-Term Matrix, the algorithm identifies 3 to 5 main topics. Each topic is represented as a distribution over words that frequently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

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.



6. 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.

Data Preview:

word *(character)*	freq *(double)*
জিয়াউর	189
হিসেবে	159
ঢাকা	153
ঘোষণা	141
তাদের	134
জাতীয়	131
জিয়া	123
সালের	122
দিয়ে	121
সরকার	117
শহীদ	116
শুরু	115
বিভিন্ন	111

সালে	111
করেছেন	111
প্রধান	111
করার	109
ছিলেন	109
রহমানের	109
হচ্ছে	107
পানি	101
করেছে	96
মাধ্যমে	93
রয়েছে	92
কিছু	91
সরকারের	91

নির্বাচন	90
চট্টগ্রাম	90
স্বাধীনতার	89
রাজনৈতিক	88
কারণে	87
পর্যন্ত	86
বিএনপির	85
বৃহস্পতিবার	81
প্রথম	80
দলের	79
বিএনপি	79
মেজর	79
বৃষ্টি	78

মেজর	79
বৃষ্টি	78
গেছে	77
জানান	76
দেখা	76
নতুন	76
মার্চ	73
সাবেক	71
তাকে	70
কাছে	69
দেওয়া	69
দাবি	69
অনেক	66