AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Project Report

Assignment Title:	Final Project Report			
Assignment No:	01		Date of Submission:	30 May 2025
Course Title:	Introduction to Data Science			
Course Code:	CSC4180		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 concerned teacher.

Group Name/No.:	02

No	Name	ID	Program	Signature
1	MD. TAMJID HOSSAIN	22-46460-1	BSc [CSE]	
2	WASIF ASAD ALVI	22-46451-1	BSc [CSE]	
3	RIFAT TALUKDAR	22-46428-1	BSc [CSE]	
4	MD. TANZIUL HAQUE	22-46435-1	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	

Table of Contents

Project Report

1. Objective

The objective of this project was to perform a complete text analysis on Bangla news articles scraped from an online news portal. It will help to develop practical skills in web scraping, text preprocessing, visual analytics, and topic modeling using real-world news articles. In the whole project we have done the below steps:

- We have scrapped Bangla news article texts from multiple pages linked in the home page of NTV BD news portal (https://www.ntvbd.com/).
- Then we performed preprocessing steps on the scrapped article texts. For example, we've performed punctuation removal, number or digit removal, unnecessary space removal, and some basic Bangla stopwords removal.
- Then we've done word frequency analysis and visualization using wordcloud
- After that, we've performed topic modeling using Latent Dirichlet Allocation (LDA) and visualized each topic's top words using bar plots.
- Lastly, we've visualized and interpreted discovered topics using Document-topic distribution and stacked bar plots.

2. Project Tasks

A. Web Scrapping

We've selected Bengali news portal NTV BD to scrape all the article texts linked in the homepage of the site. The link to the news portal is: (https://www.ntvbd.com/). Using the rvest library in R, we collected URLs and extracted article content from the <p> tags in individual article pages. The steps we followed are given below:

- First, we loaded the homepage HTML using the rvest package (read_html()).
- We extracted the links from headline tags (h2 a, h3 a, h4 a) using html_nodes().

- The links and titles were stored in a data frame, cleaning incomplete URLs by prepending the base URL when necessary.
- We filtered out invalid or empty links, removed duplicates, and saved these URLs along with their titles to a CSV file (ntv_links_with_labels.csv) encoded in UTF-8 to preserve Bangla characters.
- Next, we read the first 60 article links back from the CSV to limit the scraping load.
- For each article URL, a custom function extract_article_details() was used:
    - It loaded the article page HTML.
    - Removed unwanted nodes (e.g., inactive tab panels).
    - Extracted all paragraph (<p>) tag texts, concatenated them, and returned as the article content.
    - Errors were handled gracefully to avoid script interruption.
- All scraped article texts along with their URLs were saved in a CSV file (ntv_article_texts.csv) again with UTF-8 encoding.

This procedure ensures structured data collection from multiple article pages linked on the homepage, ready for text analysis.

Here is a code snippet example:

```
paragraphs <- page %>%
      html_nodes("p") %>%
      html_text() %>%
      paste(collapse = " ")
```

Files generated:

- ntv_links_with_labels.csv – 60 URLs and titles
- ntv_article_texts.csv – Full scraped article texts

Here is some of the examples of the article texts from our generated csv files:

	full_url	article_text
3	https://www.ntvbd.com//bangladesh/news-1564189	বিশ্বজুড়ে ক্রমবর্ধিষ্ণু অস্থিরতার আবহে এশিয়ার দেশগুলোর মধ্যে পারস্পরিক
4	https://www.ntvbd.com//world/news-1564197	অল ইন্ডিয়া মজলিস-ই-ইত্তেহাদুল মুসলিমিন (এআইএমআইএম)-এর প্রধান ও
5	https://www.ntvbd.com//bangladesh/news-1564181	১৯৮৮ সাল থেকে বিশ্বজুড়ে সংঘাতপূর্ণ অঞ্চলে বাংলাদেশি শান্তিরক্ষীরা অত্যন্ত
6	https://www.ntvbd.com//world/news-1564177	ট্রাম্প প্রশাসন 'চীনা শিক্ষার্থীদের জন্য আগ্রাসীভাবে ভিসা বাতিল করবে' বলে
7	https://www.ntvbd.com//world/news-1564149	মার্কিন প্রেসিডেন্ট ডোনাল্ড ট্রাম্পের আরোপিত ব্যাপক শুল্কের আদেশ বাতিল
8	https://www.ntvbd.com//world/news-1564193	অ্যামনেস্টি ইন্টারন্যাশনাল জানিয়েছে, গত দুই বছরে নাইজেরিয়ার মধ্য ও উত্ত
9	https://www.ntvbd.com//economy/news-1564129	বাংলাদেশে আনুষ্ঠানিকভাবে চালু হতে যাচ্ছে গুগলের ডিজিটাল পেমেন্ট সেবা
10	https://www.ntvbd.com//economy/news-1564173	বাংলাদেশ জুয়েলার্স অ্যাসোসিয়েশনের (বাজুস) সহ-সভাপতি মো. রিপনুল হাস
11	https://www.ntvbd.com//bangladesh/news-1564165	বাংলাদেশের অন্তর্বর্তী সরকারের প্রধান উপদেষ্টা অধ্যাপক মুহাম্মদ ইউনূস বলে
12	https://www.ntvbd.com//world/news-1564137	মার্কিন সরকারের একটি বিশেষ পদ থেকে পদত্যাগ করছেন ধনকুবের ইলন মা
13	https://www.ntvbd.com//bangladesh/news-1564145	দ্য নিঞ্জন ফাউন্ডেশনের প্রধান ইয়োহেই সাসাকাওয়া গতকাল বুধবার (২৮ মে)
14	https://www.ntvbd.com//bangladesh/news-1564161	অন্তর্বর্তী সরকারের প্রধান উপদেষ্টা অধ্যাপক ড. মুহাম্মদ ইউনূস বলেছেন, 'মান
15	https://www.ntvbd.com//bangladesh/news-1564133	দেশের ৬ অঞ্চলে বৃষ্টি অথবা বজ্রবৃষ্টিসহ অস্থায়ীভাবে ঝড়ো হাওয়া বয়ে যেতে
16	https://www.ntvbd.com//bangladesh/news-1564157	সকাল থেকে টানা বৃষ্টিতে ঢাকার বায়ুমানের উন্নতি হয়েছে। আজ বৃহস্পতিবার (
17	https://www.ntvbd.com//bangladesh/news-1564205	বঙ্গোপসাগরে সৃষ্ট লঘুচাপের কারণে চাঁদপুরে ভারী বৃষ্টিপাত ও ঝোড়ো হাওয়া ব
18	https://www.ntvbd.com//world/news-1564109	ইসরায়েলের প্রধানমন্ত্রী বেনিয়ামিন নেতানিয়াহু বুধবার (২৮ মে) এক ঘোষণা জ
19	https://www.ntvbd.com//bangladesh/news-1564009	চলতি বছরের ডিসেম্বরের মধ্যে জাতীয় সংসদ নির্বাচন আয়োজনের জন্য অন্ত
20	https://www.ntvbd.com//world/news-1564097	ইসরায়েলের প্রধানমন্ত্রী বেনিয়ামিন নেতানিয়াহুকে ইরানে সামরিক হামলা চালা
21	https://www.ntvbd.com//sports/news-1564121	ব্যর্থতার বৃত্ত থেকে বের হতে পারল না বাংলাদেশ। বদলেনি ভাগ্য। পাকিস্তানের
22	https://www.ntvbd.com//sports/news-1563941	হামজা চৌধুরীর পর সবচেয়ে বেশি উন্মাদনা শমিত সোমকে ঘিরে। কানাডা নয়
23	https://www.ntvbd.com//bangladesh/news-1563969	রাজনীতির মাঠে নিষ্ক্রিয় হলেও ভালোবাসায় অটল পটুয়াখালীর কৃষক সোহাগ
24	https://www.ntvbd.com//entertainment/news-1563785	বলিউডের পরিচিত মুখ অভিনেত্রী দীপিকা কক্কর। তিনি মরণব্যাধি ক্যানসার
25	https://www.ntvbd.com//health/news-1563613	ক্যানসার বিশ্বব্যাপী এক মরণব্যাধি। সাম্প্রতিক বছরগুলোতে তরুণের ক্যানসারে

Figure 1: Article text example from CSV

B. Text Processing

After scrapping the article texts from each of the news links in the homepage, we've performed basic text processing for each of the articles. Since there's a complexity of processing Bangla text processing we've had to use the Stringr library to process the texts. We've also used tm, and tokenizers libraries. For example,

```
cleaned_articles <- article_text %>%
  str_replace_all("[[:punct:]]", " ") %>%
  str_replace_all("[[:digit:]]", " ") %>%
  str_replace_all("\\b[a-zA-Z]+\\b", " ") %>%
  str_replace_all("\\s+", " ") %>%
  str_remove_all(paste0("\\b(", paste(bangla_stopwords, collapse =
"|"), ")\\b"))
```

From the code snippet, it can be seen that we've performed steps like:

- Punctuation, digits, newlines, and tabs removal
- English words removal (using regex: \\b[a-zA-Z]+\\b)
- Custom Bangla stopwords removal
- Normalized whitespaces

After performing the preprocessing steps the article texts looks like this:

```
> print(substr(cleaned_articles, 1, 500))
 [1] "বিশ্বজুড়ে ক্রমবর্ধিষ্ণু অস্থিরতার আবহে এশিয়ার দেশগুলোর পারস্পরিক নির্ভরতাকে সংঘাতের বদলে সহযোগিতার নতুন দিগন্তে রূপান্তরের আহ্বান জানিয়েছেন বাংলাদেশ সরকারের প্রধান উপ
দেষ্টা অধ্যাপক মুহাম্মদ ইউনূস আজ বৃহস্পতিবার মে টোকিওর ইম্পেরিয়াল হোটেলের ফুজি রুমে অনুষ্ঠিত নিক্কেই ফোরামের তম ফিউচার অব এশিয়া সম্মেলনে অশান্ত বিশ্বে এশিয়ার চ্যালেঞ্জ শীর্ষক
মূল বক্তব্য উপস্থাপন তার বক্তব্যে তিনটি শূন্য দারিদ্র্য বেকারত্ব নেট কার্বন নির্গমন সাহসী দৃষ্টিভঙ্গি তুলে ধরে এশিয়ার নতুন ভবিষ্যৎ নির্মাণের আ"
 [2] "অল ইন্ডিয়া মজলিস ই ইত্তেহাদুল মুসলিমিন এআইএমআইএম প্রধান হায়দরাবাদ নির্বাচিত সংসদ সদস্য আসাদউদ্দিন ওয়াইসি পাকিস্তানের ইসলামনির্ভর প্রচারণার বিরুদ্ধে কড়া অবস্থান
নিয়েছেন বলেছেন ভারতে কোটি গর্বিত মুসলমান বাস ইসলামিক স্কলাররা বিশ্বের যেকোনো দেশের চেয়ে বেশি জ্ঞানী তারা আরবি ভাষায় দক্ষতায় অনন্য আজ বৃহস্পতিবার মে সৌদি আরবে অ
নুষ্ঠানে ভাষণ দিতে গিয়ে এসব কথা খবর এনডিটিভির পাকিস্তানের বিরুদ্ধে কথা বলার সময় আসাদউদ্দিন ওয়াইসি পাকিস্তান বিশ্বজুড়ে বিশেষ আরব মুসলিম বিশ্বের কাছে ভুল বার্তা ছড়াচ"
 [3] " সাল বিশ্বজুড়ে সংঘাতপূর্ণ অঞ্চলে বাংলাদেশি শান্তিরক্ষীরা অত্যন্ত গুরুত্বপূর্ণ ভূমিকা পালন আসছেন উল্লেখ ভূমিকার প্রশংসা ইউরোপীয় ইউনিয়ন ইইউ জাতিসংঘ শান্তিরক্ষী দিবস উপল
```

Tokenization:

- We also used tokenize_words() from the tokenizers package to split text into individual words/tokens.

```
[[18]]
  [1] "ইসরায়েলের"    "প্রধানমন্ত্রী"   "বেনিয়ামিন"     "নেতানিয়াহুকে"   "ইরানে"          "সামরিক"         "হামলা"          "চালানো"
  [9] "বিরত"          "থাকার"           "সতর্ক"          "করেছেন"          "মার্কিন"        "প্রেসিডেন্ট"    "ডোনাল্ড"        "ট্রাম্প"
 [17] "পরমাণু"        "চুক্তি"          "নিয়ে"          "ওয়াশিংটন"       "তেহরানের"      "চলমান"          "আলোচনার"       "ধরনের"
 [25] "পদক্ষেপ"       "অনুপযুক্ত"       "বলে"            "উল্লেখ"          "করেছেন"        "ট্রাম্প"        "খবর"            "এএফপির"
 [33] "স্থানীয়"      "সময়"           "আজ"             "বুধবার"          "মে"             "সাংবাদিকদের"   "আলাপকালে"      "ট্রাম্প"
 [41] "নিশ্চিত"       "গত"              "সপ্তাহে"        "নেতানিয়াহুর"    "ফোনালাপে"      "ইসরায়েলি"     "প্রধানমন্ত্রীকে"  "এমন"
 [49] "কোনো"          "পদক্ষেপ"         "নিতে"           "বলেছিলেন"        "যা"             "আলোচনায়"      "ব্যাঘাত"        "ঘটাতে"
 [57] "পারে"          "ট্রাম্প"         "আমি"            "শুধু"            "বলেছি"          "এটি"            "এখন"           "উপযুক্ত"
 [65] "নয়"           "আমরা"           "খুব"            "ভালো"            "আলোচনা"        "করছি"          "আমরা"          "সমাধানের"
 [73] "খুব"           "কাছাকাছি"       "আমি"            "মনে"             "করি"           "তারা"           "ইরান"          "চুক্তি"
 [81] "করতে"          "চায়"           "যদি"            "আমরা"           "চুক্তি"         "করতে"          "পারি"          "তাহলে"
 [89] "জীবন"          "বাঁচানো"        "যাবে"           "সালে"            "ট্রাম্পের"      "প্রথম"          "মেয়াদে"       "পরমাণু"
 [97] "চুক্তি"        "সরে"            "আসে"            "যুক্তরাষ্ট্র"     "সম্প্রতি"       "তেহরান"         "ওয়াশিংটন"      "পরমাণু"
[105] "চুক্তি"        "নিয়ে"          "দফায়"          "আলোচনা"          "যা"             "চুক্তি"         "বন্ধের"         "সর্বোচ্চ"
[113] "পর্যায়ের"     "যোগাযোগ"       "আগে"            "ইরান"            "ইঙ্গিত"         "দিয়েছিল"       "যুক্তরাষ্ট্রের"  "চুক্তি"
[121] "হলে"           "তারা"           "জাতিসংঘের"      "পরমাণু"          "পর্যবেক্ষকদের"  "স্থাপনা"        "পরিদর্শনের"     "অনুমতি"
[129] "দিতে"          "পারে"           "ইসরায়েল"       "তার"             "চিরশত্রু"       "ইরানের"        "বিরুদ্ধে"       "বারবার"
[137] "সামরিক"        "পদক্ষেপের"      "হুমকি"          "আসছে"            "গত"             "সপ্তাহে"        "মার্কিন"        "গণমাধ্যমের"
[145] "খবরে"          "বলা"            "হয়েছিল"        "যুক্তরাষ্ট্র"     "ইরান"           "আলোচনার"       "মধ্যেও"         "ইসরায়েল"
[153] "ইরানের"        "পরমাণু"         "স্থাপনাগুলোতে"  "হামলা"           "চালানোর"        "প্রস্তুতি"      "নিচ্ছে"         "ট্রাম্প"
[161] "অবশ্য"         "সামরিক"         "পদক্ষেপের"      "সম্ভাবনা"        "পুরোপুরি"       "উড়িয়ে"        "দেননি"         "বলেছেন"
[169] "প্রথমে"        "চুক্তি"         "যথেষ্ট"         "সময়"            "দিতে"           "চান"            "উল্লেখ"         "কোনো"
[177] "হামলা"         "হলে"            "সেটির"          "নেতৃত্ব"         "দেবে"           "ইসরায়েল"       "যুক্তরাষ্ট্র"    "নয়"
[185] "পশ্চিমা"       "শক্তিগুলো"      "দীর্ঘদিন"       "ধরে"             "ইরানকে"         "পারমাণবিক"      "অস্ত্র"         "তৈরির"
[193] "চেষ্টার"       "অভিযোগ"         "আসছে"           "ধারাবাহিকভাবে"  "এসব"            "অভিযোগ"         "অস্বীকার"       "পারমাণবিক"
[201] "কর্মসূচিকে"    "সম্পূর্ণরূপে"   "শান্তিপূর্ণ"     "বেসামরিক"       "উদ্দেশ্যে"      "পরিচালিত"       "বলে"            "দাবি"
[209] "তেহরান"
```

Figure 2: Sample tokeniztion

C. Exploratory Text Analysis

Corpus and Document-Term Matrix:

- Created a corpus object using the tm package from cleaned articles.
- Generated a Document-Term Matrix (DTM) to quantify word occurrences across documents.
- Filtered documents without any terms.

Frequency Analysis:

- Calculated total frequency of each word and sorted them.

Example Code Snippet:

```
# DTM matrix and word frequency
m <- as.matrix(dtm)
row_totals <- rowSums(m)
dtm_filtered <- dtm[row_totals > 0, ]
m_filtered <- as.matrix(dtm_filtered)
word_freqs <- sort(colSums(m_filtered), decreasing = TRUE)
df <- data.frame(word = names(word_freqs), freq = word_freqs)
head(df)
```

Top 5 words with their frequency:

```
> head(df)
                word freq
বাংলাদেশ বাংলাদেশ   54
মে               মে   51
হয়েছে       হয়েছে   51
তার             তার   46
করতে           করতে   41
প্রধান         প্রধান   41
```

Figure 3: Top 5 most frequent words with their frequence

Visualization:

- We created word clouds and bar plots of the top frequent words using wordcloud and barplot.

Figure 4: Word cloud visualization

From the wordcloud, we can see that the word "বাংলাদেশ" has the highest frequency so it is bigger on the wordcloud as well. Similarly words with higher frequency has bigger sizes and lower frequency has smaller sizes.

Barplot of top 20 most frequent words from the articles:

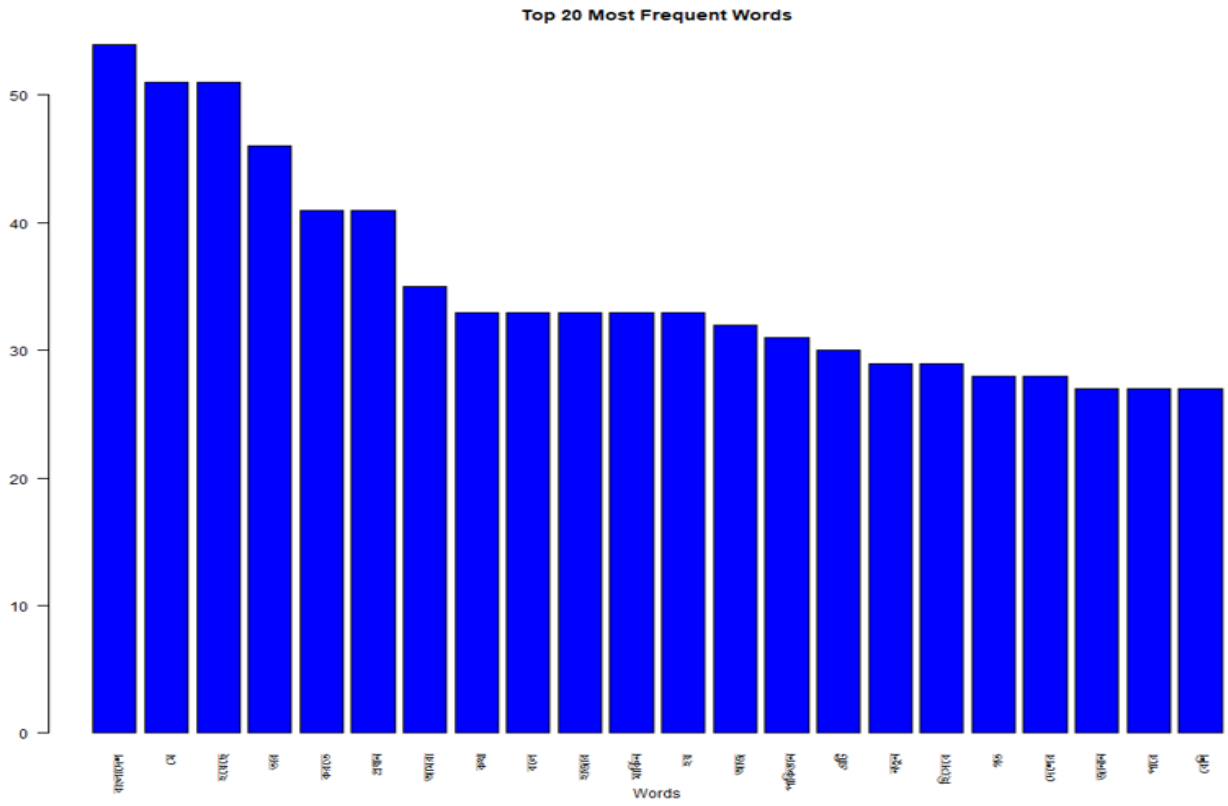

Figure 5: Top 20 most frequent words barplot

From the barplot, it can also be seen that the words "বাংলাদেশ", "মে", "হয়েছে", "তার" etc have the highest frequencies.

D. Topic Modeling

For topic modeling we used tm, topicmodels, ggplot2 libraries. Then we followed the below steps for topic modeling:

- Applied Latent Dirichlet Allocation (LDA) with 5 topics using the topicmodels package.
- Extracted top terms per topic and plotted the top words for each topic.
- Visualized document-topic distributions to understand how topics are spread across articles.

Top terms per topic:

Below is a code snippet example for the top terms per topics:

```
# Topic Modeling with LDA
num_topics <- 5
lda_model <- LDA(dtm_filtered, k = num_topics, control = list(seed
= 58))

# Top terms per topic
top_terms <- terms(lda_model, 10)
print(top_terms)
```

The result is:

```
> print(top_terms)
      Topic 1    Topic 2      Topic 3      Topic 4   Topic 5
 [1,] "বাংলাদেশ" "প্রধান"     "পয়সা"      "মোহাম্মদ" "মার্কিন"
 [2,] "তার"      "বাংলাদেশ"  "সিরিজ"      "হামলার"  "ট্রাম্প"
 [3,] "ত্বকের"   "বলে"       "পাকিস্তান"  "টিকিট"   "প্রশাসন"
 [4,] "নতুন"     "উপদেষ্টা"  "কোম্পানিটির" "বেশি"    "ক্ষমতা"
 [5,] "করতে"     "করতে"      "সমাপ্ত"     "হয়"      "পদক্ষেপ"
 [6,] "এশিয়ার"  "আমরা"      "এটি"        "সিনওয়ার" "শুল্ক"
 [7,] "বছর"      "সরকারের"   "হয়েছে"     "মে"       "দেশের"
 [8,] "হিসেবে"   "বাংলাদেশের" "বছরে"      "দেওয়া"   "পারে"
 [9,] "জাতীয়"   "বিএনপির"   "শেয়ারপ্রতি" "কথা"     "বড়"
[10,] "জানান"    "রহমান"     "বাংলাদেশ"   "তার"      "হাজার"
>
```

Figure 6: Top 10 terms per topic

We can easily see the top terms of each of topics here. By seeing the terms we can assume what is the topic that was meant for those words. For example,

- Topic 1 can be assumed for National & Domestic Affairs
- Topic 2 can be assumed for Politics and Governance
- Topic 3 can be assumed for Finance & Economy
- Topic 4 can be assumed for Conflict, or Social Issues
- Topic 5 can be assumed for International Affairs.

Barplots for top words per topic:

Topic 1:

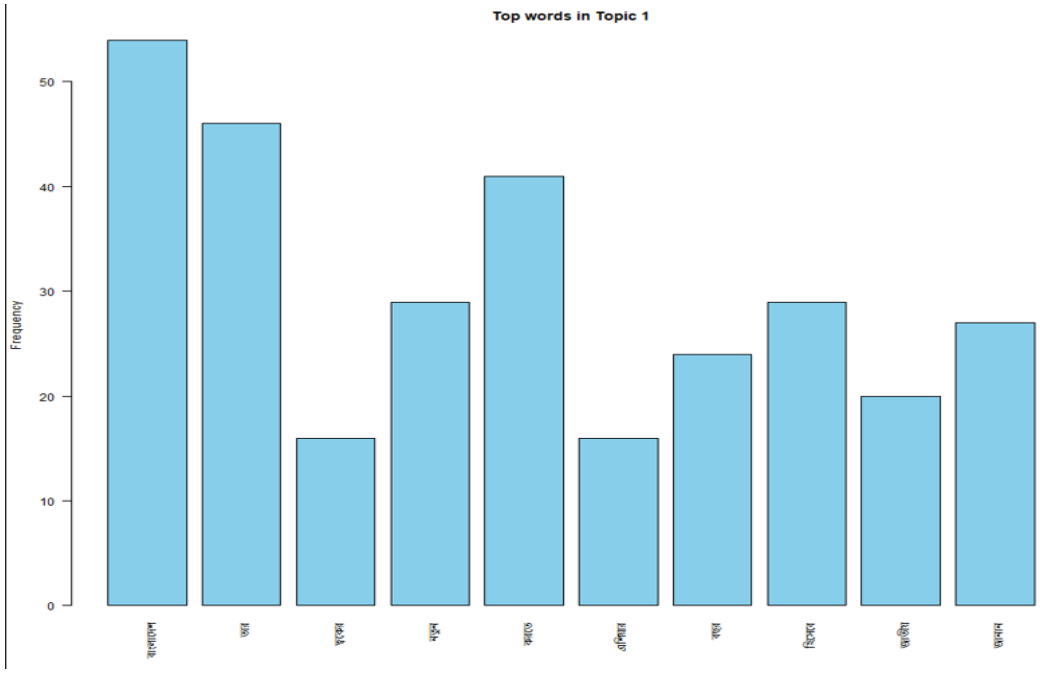

Figure 7: Top words for Topic 1

Topic 2:

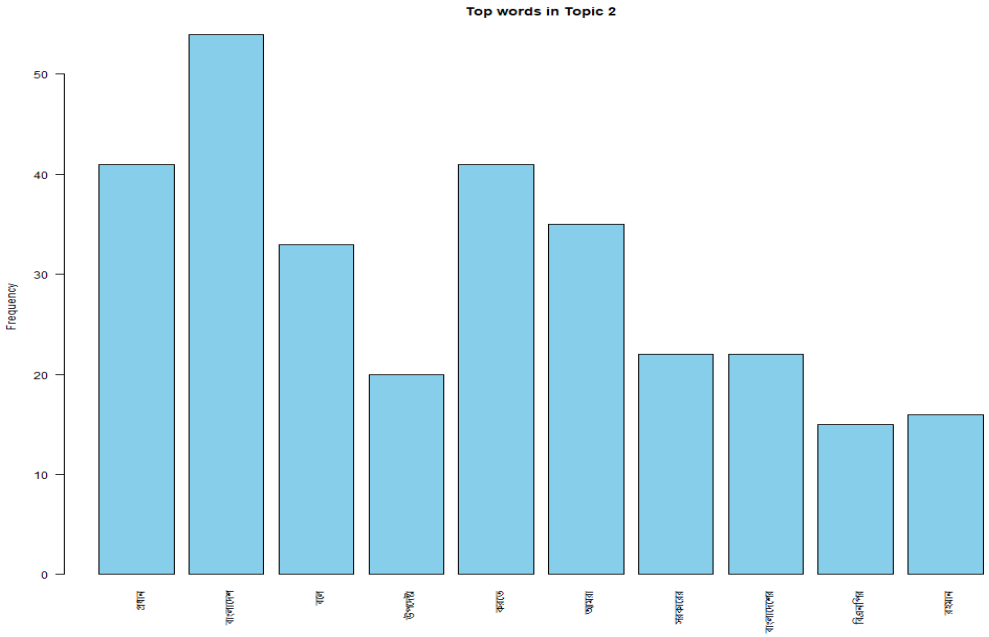

Figure 8: Top words for Topic 2

Topic 3:

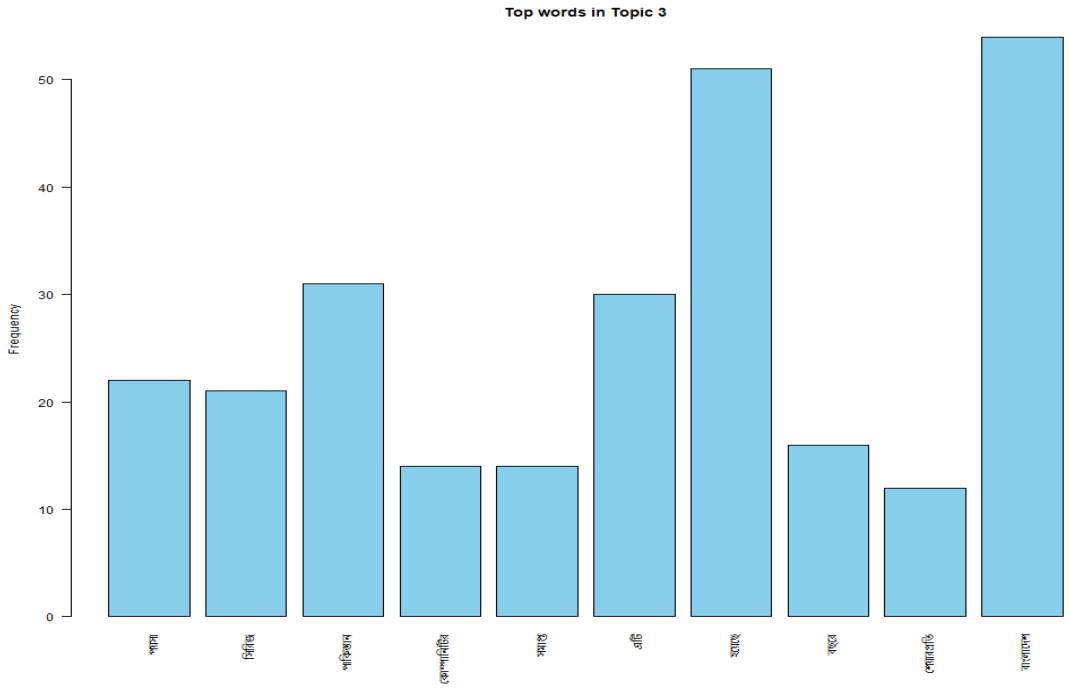

Figure 9: Top words for Topic 3

Topic 4:

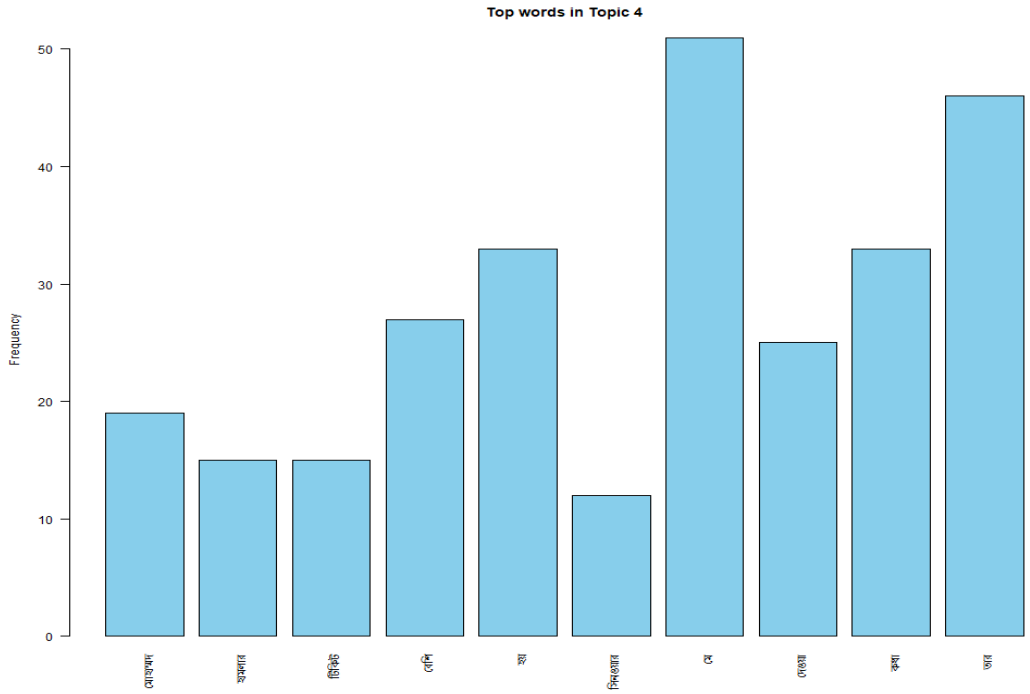

Figure 10: Top words for Topic 4

Topic 5:

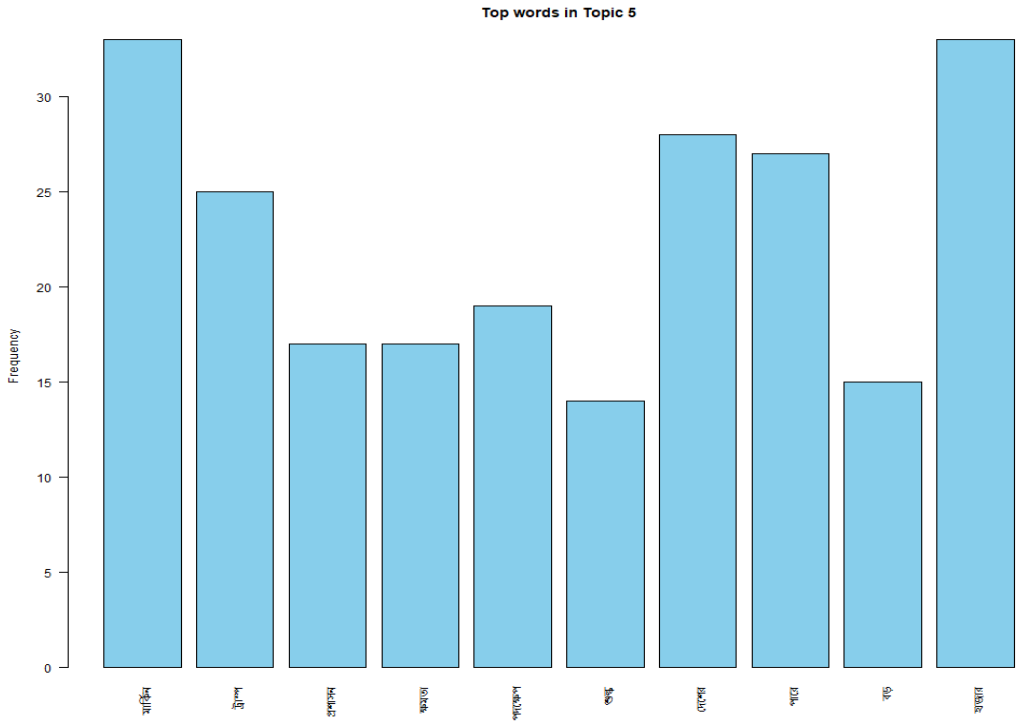

Figure 11: Top words for Topic 5

From the barplots, we can easily see and understand the top words of respective topics and their frequencies.

Document-topic distribution using Stacked Bar:

We visualized a document-topic distribution using a stacked bar. A document–topic distribution shows how much each topic contributes to each document in a corpus.

Example code snippet:

```
# Document-topic probabilities using tidytext
doc_topics <- tidy(lda_model, matrix = "gamma")

# View a sample
head(doc_topics)

# Sort Documents and Label Topic
doc_topics <- doc_topics %>%
  mutate(document = factor(as.integer(document), levels =
sort(unique(as.integer(document)))),
         topic = paste("Topic", topic))

# Plot Stacked bar chart
png("Document-Topic Distribution.png", width = 1000, height = 800)
```

```
ggplot(doc_topics, aes(x = factor(document), y = gamma, fill =
topic)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Document-Topic Distribution",
    x = "Document",
    y = "Topic Probability",
    fill = "Topic"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90))
dev.off()
```

Result of the document-topic distribution:

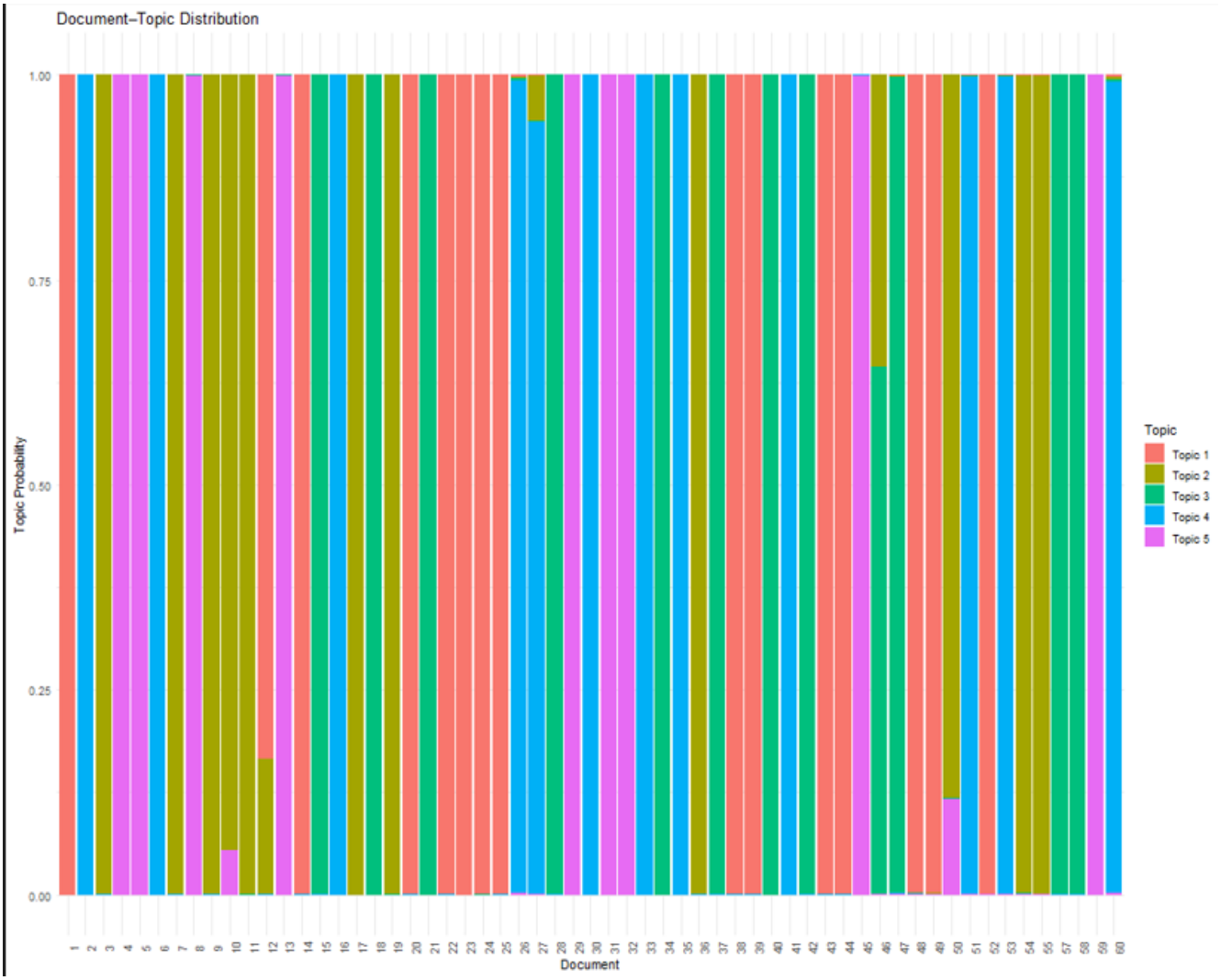

Figure 12: Document-topic distribution

From the image it can be seen that:

- Most documents are dominated by a single topic. This is shown by bars that are nearly entirely one color.
- A few documents have mixed topic distributions, where multiple colors appear in a single bar.
- For example, Document 10 and 47 show contributions from two or more topics, indicating a more diverse content.

Conclusion

This project was done for collecting and analyzing Bengali news articles from the NTV online news portal. Through web scraping, we gathered real-time data and applied text preprocessing techniques for Bengali language content. The use of topic modeling via Latent Dirichlet Allocation (LDA) allowed us to analyze news articles and their topics.

Overall, this project showcases the practical application of data science and NLP techniques in a real-world context involving the Bangla language. It highlights the importance of preprocessing and data visualization.