AMERICAN INTERNATIONAL UNIVERSITY – BANGLADESH

INTRODUCTION TO DATA SCIENCE

SPRING- 2024-2025

Supervise By

DR. Abdus Salam

Assignment: Final Term Project Report

Group No: 05

Section: E

NAME	ID
MD. Tashrifull Alam	22-46610-1
Priya Rani Das Sowhanur Rahman Nirob	22-46319-1
Jeba Shajida	22-46590-1
Priya Rani Das	22-46319-1

Description :
In the first part we perform text preprocessing on the *ittefaq_news.csv* dataset from web scrapping, focusing on cleaning the description column for further analysis. It includes loading necessary text mining packages, reading the data, and assigning unique IDs to each entry. Descriptions are tokenized into words, with stop words removed, followed by stemming and lemmatization to reduce words to their root or base forms. The cleaned tokens are then recombined into processed descriptions, merged back with the original data, and exported as a new CSV file containing cleaned news articles.

Then we perform the topic modeling part. The text is tokenized, lowercased, and lemmatized before constructing a document-term matrix (DTM). An LDA model with 5 topics is fitted, and the top 10 keywords for each topic are identified and saved. The script also assigns the dominant topic to each news article and stores topic proportions per document.

Web Scraping:

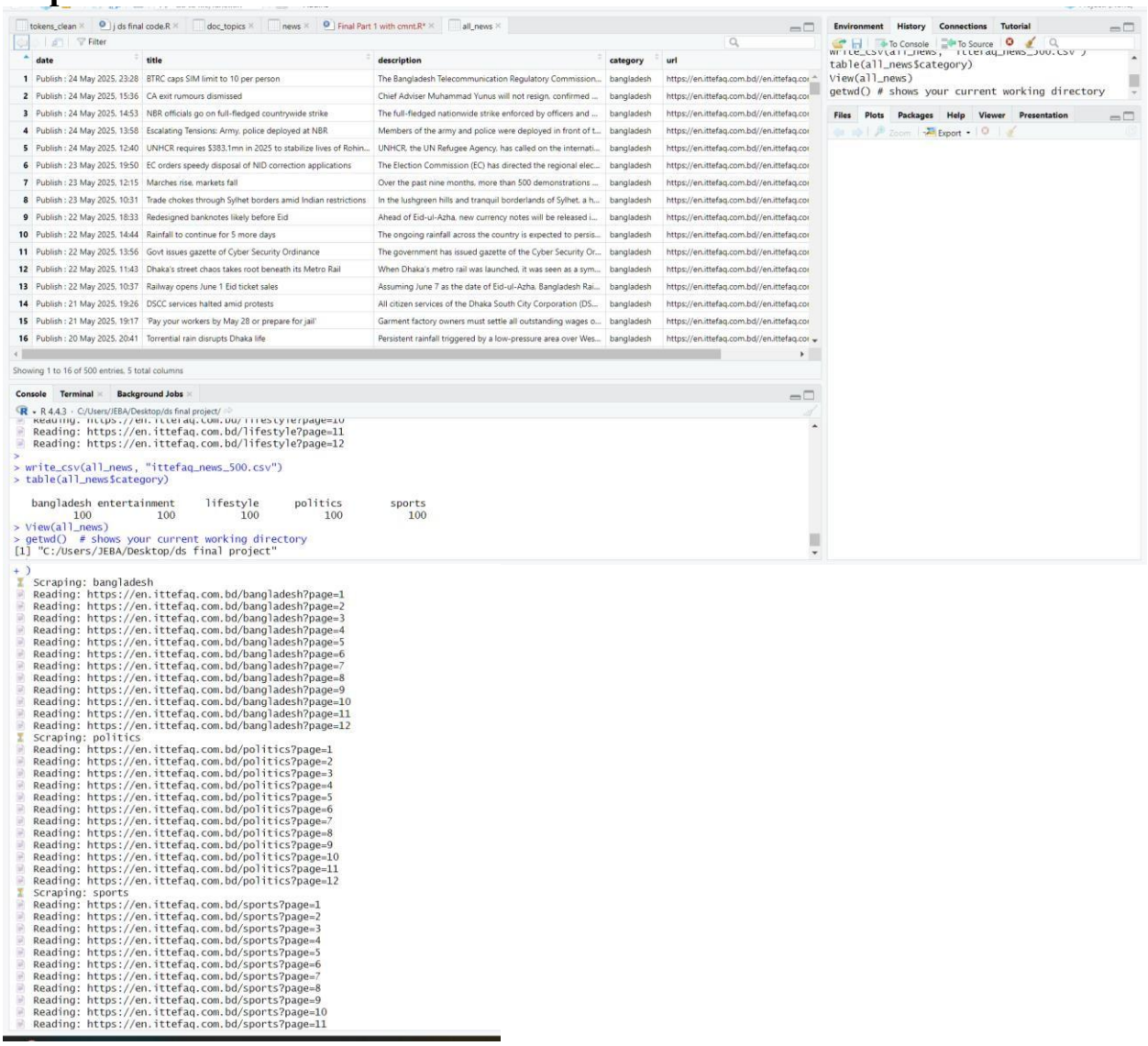
Code:

```
categories <- list(
  bangladesh = "https://en.ittefaq.com.bd/bangladesh",
  politics = "https://en.ittefaq.com.bd/politics",
  sports = "https://en.ittefaq.com.bd/sports",
  entertainment = "https://en.ittefaq.com.bd/entertainment",
  lifestyle = "https://en.ittefaq.com.bd/lifestyle"
)

all_news <- bind_rows(
  lapply(names(categories), function(cat) {
  cat("⏳ Scraping:", cat, "\n")
    scrape_category_news(categories[[cat]], cat, max_articles = 100)
  })
)

write_csv(all_news, "C:/Users/Desktop/ds final project/ittefaq_news_500.csv")
table(all_news$category)
View(all_news)
getwd()
```

output:



	date	title	description	category	url
1	Publish : 24 May 2025, 23:28	BTRC caps SIM limit to 10 per person	The Bangladesh Telecommunication Regulatory Commission...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
2	Publish : 24 May 2025, 15:36	CA exit rumours dismissed	Chief Adviser Muhammad Yunus will not resign, confirmed ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
3	Publish : 24 May 2025, 14:53	NBR officials go on full-fledged countrywide strike	The full-fledged nationwide strike enforced by officers and ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
4	Publish : 24 May 2025, 13:58	Escalating Tensions: Army, police deployed at NBR	Members of the army and police were deployed in front of t...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
5	Publish : 24 May 2025, 12:40	UNHCR requires $383.1mn in 2025 to stabilize lives of Rohin...	UNHCR, the UN Refugee Agency, has called on the internati...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
6	Publish : 23 May 2025, 19:50	EC orders speedy disposal of NID correction applications	The Election Commission (EC) has directed the regional elec...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
7	Publish : 23 May 2025, 12:15	Marches rise, markets fall	Over the past nine months, more than 500 demonstrations ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
8	Publish : 23 May 2025, 10:31	Trade chokes through Sylhet borders amid Indian restrictions	In the lushgreen hills and tranquil borderlands of Sylhet, a h...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
9	Publish : 22 May 2025, 18:33	Redesigned banknotes likely before Eid	Ahead of Eid-ul-Azha, new currency notes will be released i...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
10	Publish : 22 May 2025, 14:44	Rainfall to continue for 5 more days	The ongoing rainfall across the country is expected to persis...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
11	Publish : 22 May 2025, 13:56	Govt issues gazette of Cyber Security Ordinance	The government has issued gazette of the Cyber Security Or...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
12	Publish : 22 May 2025, 11:43	Dhaka's street chaos takes root beneath its Metro Rail	When Dhaka's metro rail was launched, it was seen as a sym...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
13	Publish : 22 May 2025, 10:37	Railway opens June 1 Eid ticket sales	Assuming June 7 as the date of Eid-ul-Azha, Bangladesh Rai...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
14	Publish : 21 May 2025, 19:26	DSCC services halted amid protests	All citizen services of the Dhaka South City Corporation (DS...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
15	Publish : 21 May 2025, 19:17	'Pay your workers by May 28 or prepare for jail'	Garment factory owners must settle all outstanding wages o...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
16	Publish : 20 May 2025, 20:41	Torrential rain disrupts Dhaka life	Persistent rainfall triggered by a low-pressure area over Wes...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co

```
Showing 1 to 16 of 500 entries, 5 total columns

Reading: https://en.ittefaq.com.bd/lifestyle?page=11
Reading: https://en.ittefaq.com.bd/lifestyle?page=12
>
> write_csv(all_news, "ittefaq_news_500.csv")
> table(all_news$category)

   bangladesh entertainment     lifestyle      politics        sports
         100           100           100           100           100
> View(all_news)
> getwd()  # shows your current working directory
[1] "C:/Users/JEBA/Desktop/ds final project"
+ )
Scraping: bangladesh
Reading: https://en.ittefaq.com.bd/bangladesh?page=1
Reading: https://en.ittefaq.com.bd/bangladesh?page=2
Reading: https://en.ittefaq.com.bd/bangladesh?page=3
Reading: https://en.ittefaq.com.bd/bangladesh?page=4
Reading: https://en.ittefaq.com.bd/bangladesh?page=5
Reading: https://en.ittefaq.com.bd/bangladesh?page=6
Reading: https://en.ittefaq.com.bd/bangladesh?page=7
Reading: https://en.ittefaq.com.bd/bangladesh?page=8
Reading: https://en.ittefaq.com.bd/bangladesh?page=9
Reading: https://en.ittefaq.com.bd/bangladesh?page=10
Reading: https://en.ittefaq.com.bd/bangladesh?page=11
Reading: https://en.ittefaq.com.bd/bangladesh?page=12
Scraping: politics
Reading: https://en.ittefaq.com.bd/politics?page=1
Reading: https://en.ittefaq.com.bd/politics?page=2
Reading: https://en.ittefaq.com.bd/politics?page=3
Reading: https://en.ittefaq.com.bd/politics?page=4
Reading: https://en.ittefaq.com.bd/politics?page=5
Reading: https://en.ittefaq.com.bd/politics?page=6
Reading: https://en.ittefaq.com.bd/politics?page=7
Reading: https://en.ittefaq.com.bd/politics?page=8
Reading: https://en.ittefaq.com.bd/politics?page=9
Reading: https://en.ittefaq.com.bd/politics?page=10
Reading: https://en.ittefaq.com.bd/politics?page=11
Reading: https://en.ittefaq.com.bd/politics?page=12
Scraping: sports
Reading: https://en.ittefaq.com.bd/sports?page=1
Reading: https://en.ittefaq.com.bd/sports?page=2
Reading: https://en.ittefaq.com.bd/sports?page=3
Reading: https://en.ittefaq.com.bd/sports?page=4
Reading: https://en.ittefaq.com.bd/sports?page=5
Reading: https://en.ittefaq.com.bd/sports?page=6
Reading: https://en.ittefaq.com.bd/sports?page=7
Reading: https://en.ittefaq.com.bd/sports?page=8
Reading: https://en.ittefaq.com.bd/sports?page=9
Reading: https://en.ittefaq.com.bd/sports?page=10
Reading: https://en.ittefaq.com.bd/sports?page=11
```

Description :

In this step, a list of predefined news categories from the Ittefaq English news portal is created, where each category (e.g., Bangladesh, Politics, Sports, etc.) is associated with its corresponding URL. This structure is stored in a named list, enabling automated and organized access to each category's webpage.A loop is implemented using the lapply() function to iterate through each category name. For every category, a progress message is displayed using cat() to inform the user which section is currently being scraped. A custom-defined function scrape_category_news() is invoked to extract up to 100 news articles from each category's page. This function typically collects essential elements such as the article title, summary, date, and category label.The results from all categories, which are returned as individual data frames, are combined into a single unified data frame using bind_rows().

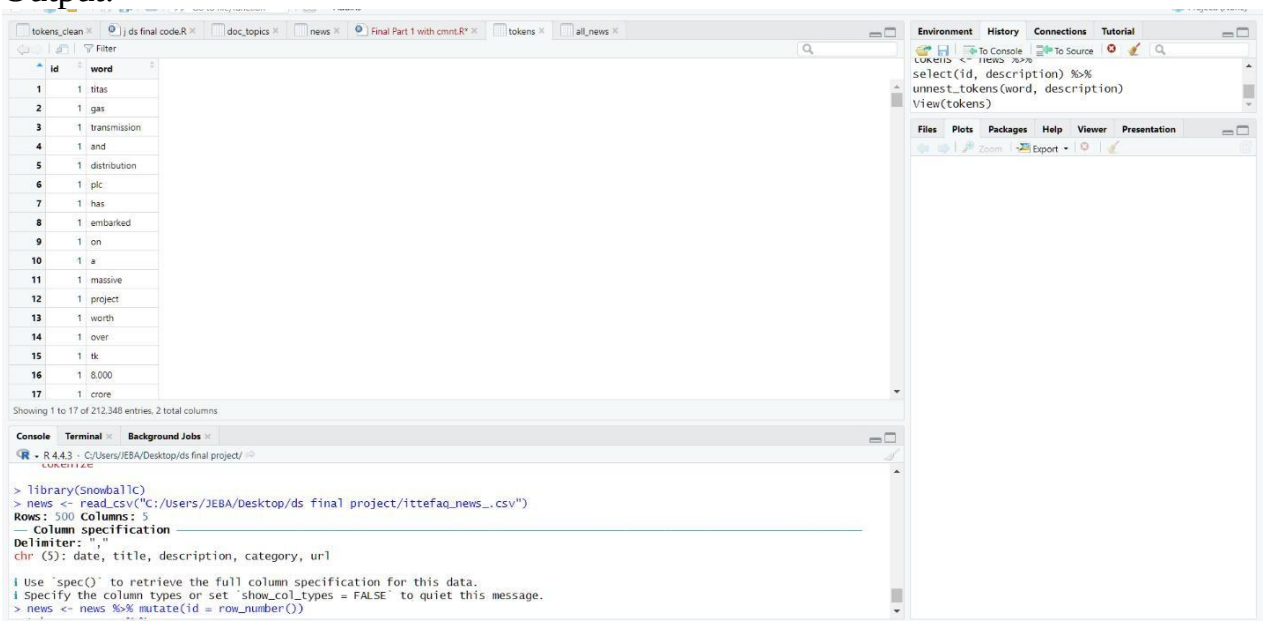
This final dataset, all_news, contains approximately 500 articles across all specified categories.Once the scraping process is complete, the combined dataset is saved as a CSV file using write_csv() to the designated location ("C:/Users/Desktop/ds final project/ittefaq_news_500.csv"). Additional functions like table() are used to display the distribution of articles by category, and View() opens the data frame in a spreadsheet-style viewer for manual inspection. Lastly, getwd() is optionally used to confirm the current working directory, which is helpful when managing file paths during import/export operations

Tokenization:

Code :

```
install.packages(c("tidytext", "dplyr", "tm", "textstem", "readr", "SnowballC"))
library(tidytext)
library(dplyr)
library(tm)
library(textstem)
library(readr)
library(SnowballC)
news <- read_csv("C:/Users/Desktop/ds final project/ittefaq_news_.csv") news <-
news %>% mutate(id = row_number())
tokens <- news %>%
select(id, description) %>%
unnest_tokens(word, description)
View(tokens)
```

Output:



Description :

In this step, the necessary R packages are installed and loaded to support text preprocessing and analysis. The tidytext package provides tidy tools for text mining; dplyr supports data manipulation; tm aids in corpus handling; textstem allows lemmatization; readr enables fast reading of CSV files; and SnowballC provides stemming capabilities.The news dataset is imported from a CSV file using read_csv(), and stored in a data frame named news. To uniquely identify each article, a new column id is created using mutate() and row_number(), which assigns a sequential number to each row.The dataset is then prepared for tokenization. Using select(), only the id and description columns are retained. The unnest_tokens() function from tidytext is used to tokenize the text in the description column. This function breaks each description into individual words (tokens), resulting in a long-format data frame where each row corresponds to a single word and its associated article ID.The resulting tokens data frame is viewed using View() for manual inspection. This tokenized format is ideal for further text mining tasks such as frequency analysis, stopword removal, sentiment analysis, and topic modeling.

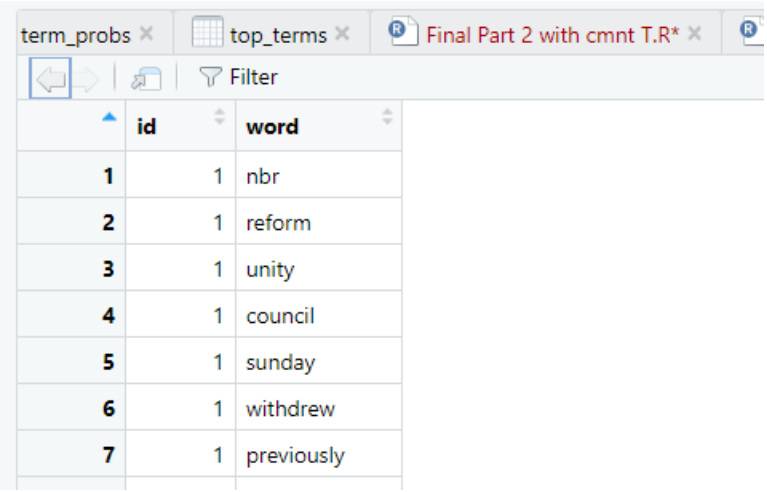
Stop Word Removal:

Code:

```
data("stop_words")
tokens_clean <- tokens %>%
  anti_join(stop_words, by = "word")

View(tokens_clean)
```

Output:



	id	word
1	1	nbr
2	1	reform
3	1	unity
4	1	council
5	1	sunday
6	1	withdrew
7	1	previously

Description :

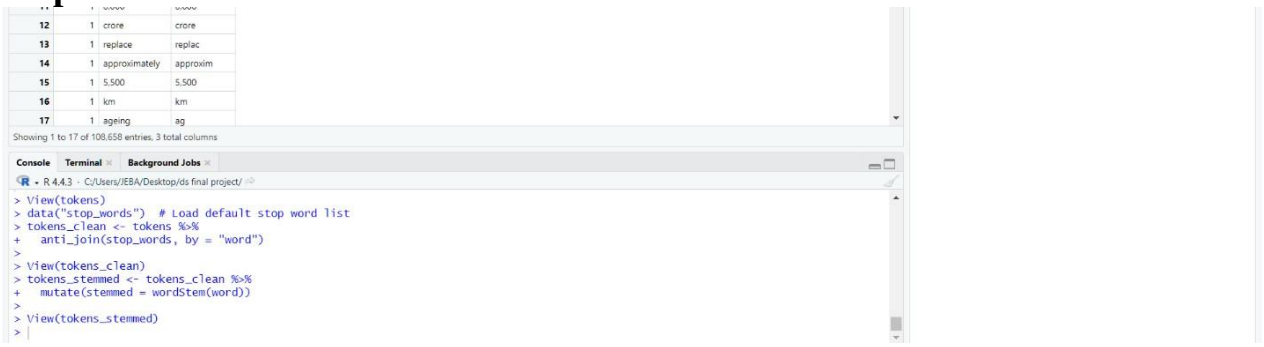
In this step, common English stop words—such as "the", "and", "of", etc.—are removed from the tokenized text data. These words are often extremely frequent but carry little meaningful information in text analysis.The built-in stop_words dataset from the tidytext package is loaded using data("stop_words"). This dataset contains a comprehensive list of common stop words from multiple lexicons.Using the anti_join() function from dplyr, the tokens data frame is filtered to exclude all words that appear in the stop word list. This operation results in a new data frame named tokens_clean, which retains only the informative and meaningful words from each article's description.The cleaned tokens are then viewed using View(tokens_clean). This cleaned version of the text is now ready for further natural language processing tasks such as frequency analysis, lemmatization, or topic modeling.

Stemming:

Code

```
 tokens_stemmed <- tokens_clean %>%
mutate(stemmed = wordStem(word))
View(tokens_stemmed)
```

output



Description : In this step, stemming is applied to the cleaned tokens using the `wordStem()` function from the **`SnowballC`** package. Stemming is a common text preprocessing technique that reduces words to their root or base form. For example, "playing", "played", and "plays" are all reduced to "play". The result, stored in tokens_stemmed, includes both the original word and its stemmed form. This helps in grouping similar terms together during further analysis like frequency counts, topic modeling, or clustering.Finally, the View(tokens_stemmed) command allows for manual inspection of the stemmed words.

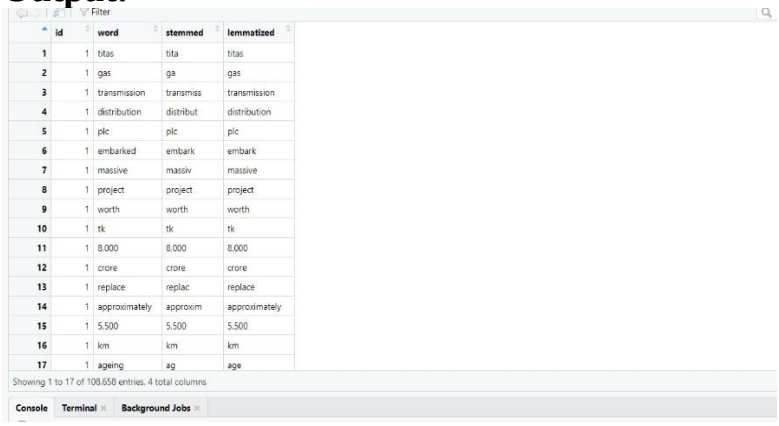
Lemmatization

Code

```
tokens_lemmatized <- tokens_stemmed %>%
  mutate(lemmatized = lemmatize_words(word))

View(tokens_lemmatized)
```

Output:



	id	word	stemmed	lemmatized
1	1	titas	tita	titas
2	1	gas	ga	gas
3	1	transmission	transmiss	transmission
4	1	distribution	distribut	distribution
5	1	plc	plc	plc
6	1	embarked	embark	embark
7	1	massive	massiv	massive
8	1	project	project	project
9	1	worth	worth	worth
10	1	tk	tk	tk
11	1	8,000	8,000	8,000
12	1	crore	crore	crore
13	1	replace	replac	replace
14	1	approximately	approxim	approximately
15	1	5,500	5,500	5,500
16	1	km	km	km
17	1	ageing	ag	age

Showing 1 to 17 of 108,658 entries, 4 total columns

Console | Terminal | Background Jobs

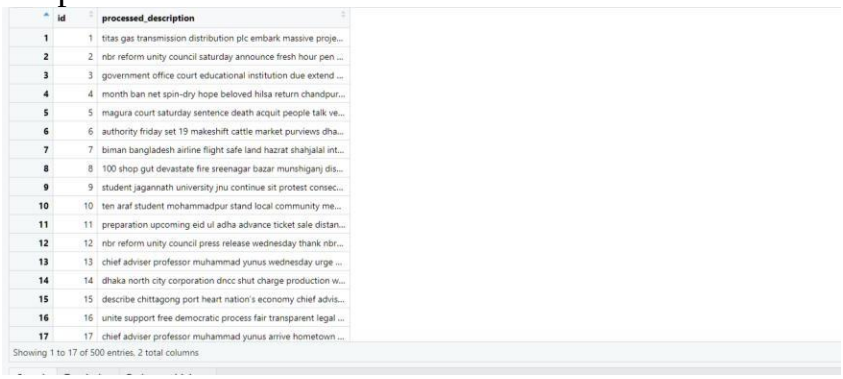
Description : In this step, lemmatization is performed on the cleaned and stemmed tokens using the lemmatize_words() function from the textstem package. Lemmatization transforms each word into its base or dictionary form (lemma), such as converting “running”, “ran”, and “runs” into “run”. This process helps reduce different forms of a word to a common base, making the text data more consistent and improving the effectiveness of later analysis such as frequency counts or topic modeling.

Recombine Lemmatized Tokens Back Into Clean Description:

Code :

```
clean_descriptions <- tokens_lemmatized %>%
  group_by(id) %>%
  summarise(processed_description = paste(lemmatized, collapse = " ")) %>%
  ungroup()
View(clean_descriptions)
```

Output:



	id	processed_description
1	1	titas gas transmission distribution plc embark massive proje...
2	2	nbr reform unity council saturday announce fresh hour pen ...
3	3	government office court educational institution due extend ...
4	4	month ban net spin-dry hope beloved hilsa return chandpur...
5	5	magura court saturday sentence death acquit people talk ve...
6	6	authority friday set 19 makeshift cattle market purviews dha...
7	7	biman bangladesh airline flight safe land hazrat shahjalal int...
8	8	100 shop gut devastate fire sreenagar bazar munshiganj dis...
9	9	student jagannath university jnu continue sit protest consec...
10	10	ten araf student mohammadpur stand local community me...
11	11	preparation upcoming eid ul adha advance ticket sale distan...
12	12	nbr reform unity council press release wednesday thank nbr...
13	13	chief adviser professor muhammad yunus wednesday urge ...
14	14	dhaka north city corporation dncc shut charge production w...
15	15	describe chittagong port heart nation's economy chief advis...
16	16	unite support free democratic process fair transparent legal ...
17	17	chief adviser professor muhammad yunus arrive hometown ...

Showing 1 to 17 of 500 entries, 2 total columns

Description:
In this step, the individual lemmatized tokens for each document (news article) are grouped by their id using group_by(). Then, summarise() is used to collapse the lemmatized words back into a single string for each document using paste(..., collapse = " "). This results in a new column called processed_description that contains cleaned, lemmatized text for each news article. This processed text is now ready for further natural language processing tasks such as topic modeling, sentiment analysis, or text classification.

Merge With Original Data & Export to CSV:

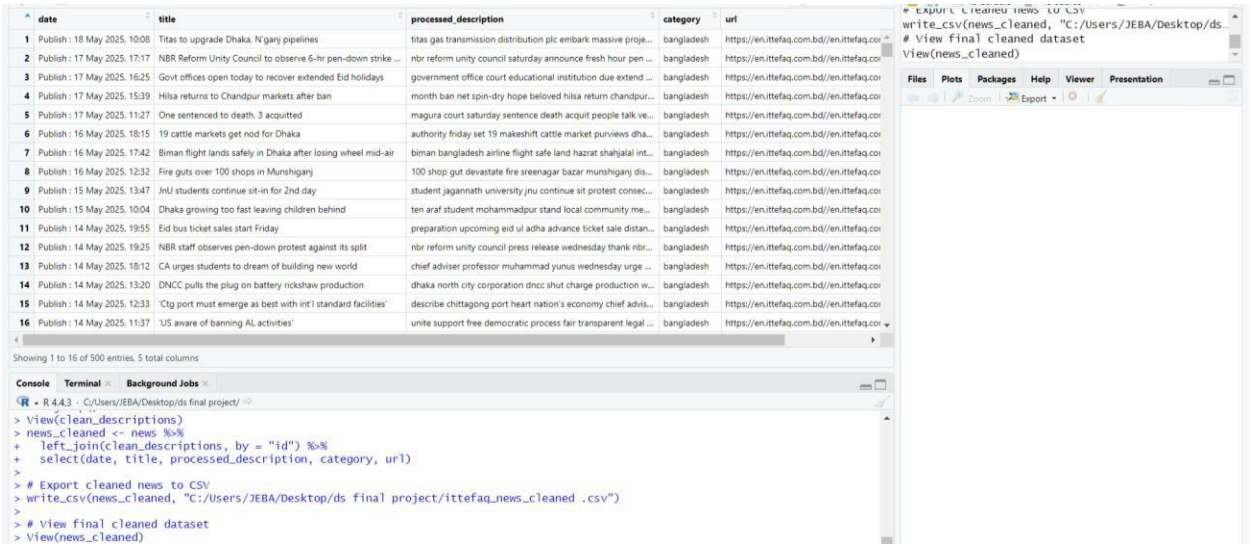
Code:

```
news_cleaned <- news %>%
  left_join(clean_descriptions, by = "id") %>%
  select(date, title, processed_description, category, url)

write_csv(news_cleaned, (" ittefaq_news_cleaned.csv")
```

View(news_cleaned)

Output:



	date	title	processed_description	category	url
1	Publish : 18 May 2025, 10:08	Titas to upgrade Dhaka, N'ganj pipelines	titas gas transmission distribution plc embark massive proje...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
2	Publish : 17 May 2025, 17:17	NBR Reform Unity Council to observe 6-hr pen-down strike ...	nbr reform unity council saturday announce fresh hour pen ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
3	Publish : 17 May 2025, 16:25	Govt offices open today to recover extended Eid holidays	government office court educational institution due extend ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
4	Publish : 17 May 2025, 15:39	Hilsa returns to Chandpur markets after ban	month ban net spin-dry hope beloved hilsa return chandpur...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
5	Publish : 17 May 2025, 11:27	One sentenced to death, 3 acquitted	magura court saturday sentence death acquit people talk ve...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
6	Publish : 16 May 2025, 18:15	19 cattle markets get nod for Dhaka	authority friday set 19 makeshift cattle market purviews dha...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
7	Publish : 16 May 2025, 17:42	Biman flight lands safely in Dhaka after losing wheel mid-air	biman bangladesh airline flight safe land hazrat shahjalal int...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
8	Publish : 16 May 2025, 12:32	Fire guts over 100 shops in Munshiganj	100 shop gut devastate fire sreenagar bazar munshiganj dis...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
9	Publish : 15 May 2025, 13:47	JnU students continue sit-in for 2nd day	student jagannath university jnu continue sit protest consec...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
10	Publish : 15 May 2025, 10:04	Dhaka growing too fast leaving children behind	ten araf student mohammadpur stand local community me...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
11	Publish : 14 May 2025, 19:55	Eid bus ticket sales start Friday	preparation upcoming eid ul adha advance ticket sale distan...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
12	Publish : 14 May 2025, 19:25	NBR staff observes pen-down protest against its split	nbr reform unity council press release wednesday thank nbr...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
13	Publish : 14 May 2025, 18:12	CA urges students to dream of building new world	chief adviser professor muhammad yunus wednesday urge ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
14	Publish : 14 May 2025, 13:20	DNCC pulls the plug on battery rickshaw production	dhaka north city corporation dncc shut charge production w...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
15	Publish : 14 May 2025, 12:33	'Ctg port must emerge as best with int'l standard facilities'	describe chittagong port heart nation's economy chief advis...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co
16	Publish : 14 May 2025, 11:37	'US aware of banning AL activities'	unite support free democratic process fair transparent legal ...	bangladesh	https://en.ittefaq.com.bd//en.ittefaq.co

Showing 1 to 16 of 500 entries, 5 total columns

```
write_csv(news_cleaned, "C:/Users/JEBA/Desktop/ds_
# View final cleaned dataset
View(news_cleaned)
```

Files | Plots | Packages | Help | Viewer | Presentation

Export

Console | Terminal | Background Jobs

R 4.4.3 · C:/Users/JEBA/Desktop/ds final project/

```
> View(clean_descriptions)
> news_cleaned <- news %>%
+   left_join(clean_descriptions, by = "id") %>%
+   select(date, title, processed_description, category, url)
>
> # Export cleaned news to CSV
> write_csv(news_cleaned, "C:/Users/JEBA/Desktop/ds final project/ittefaq_news_cleaned .csv")
>
> # View final cleaned dataset
> View(news_cleaned)
```

2nd part:

Load the Cleaned data and Create Corpus from Processed Description :

```
ittefaq_data <- read_csv("ittefaq_news_cleaned.csv")
corpus <- corpus %>%
tm_map(content_transformer(tolower)) %>%
tm_map(removePunctuation) %>%
tm_map(removeNumbers) %>%
tm_map(removeWords, stopwords("en")) %>%
tm_map(stripWhitespace)
```

Description:
In this step, the cleaned news data is first loaded from the file ittefaq_news_cleaned.csv using read_csv(). After loading, a text corpus is created from the processed descriptions to prepare the text for further analysis. The corpus undergoes several cleaning steps using the tm_map() function: all text is converted to lowercase to ensure consistency, punctuation marks and numbers are removed to eliminate irrelevant characters, and common English stop words are filtered out to focus on meaningful words. Finally, extra white spaces are stripped to tidy up the text. These preprocessing steps help standardize and clean the text data, making it ready for tasks like building a document-term matrix or topic modeling.

Document-Term Matrix:

Extract Term Probabilities Per Topic

Code

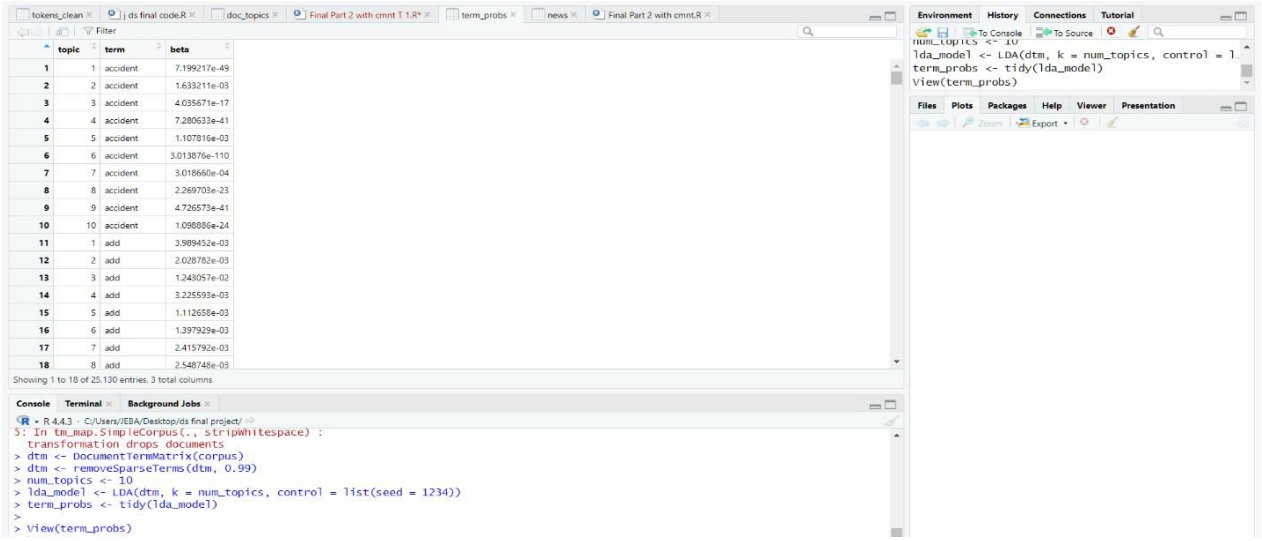
```
dtm <- DocumentTermMatrix(corpus)
dtm <- removeSparseTerms(dtm, 0.99)


num_topics <- 10
lda_model <- LDA(dtm, k = num_topics, control = list(seed = 1234))


term_probs <- tidy(lda_model)

View(term_probs)
```

output:



Description: We extract the term probabilities for each topic from the trained LDA model. This is done using the tidy() function from the tidytext package, which converts the model output into a tidy data frame format. The resulting object, term_probs, contains three key columns: topic, term, and beta. Here, topic indicates the topic number, term is the word from the vocabulary, and beta represents the probability of that term appearing in the given topic. Viewing term_probs allows for easy inspection of which terms are most strongly associated with each topic, providing insight into the dominant themes discovered by the model. This step is essential for interpreting the topics based on their most probable and representative words.

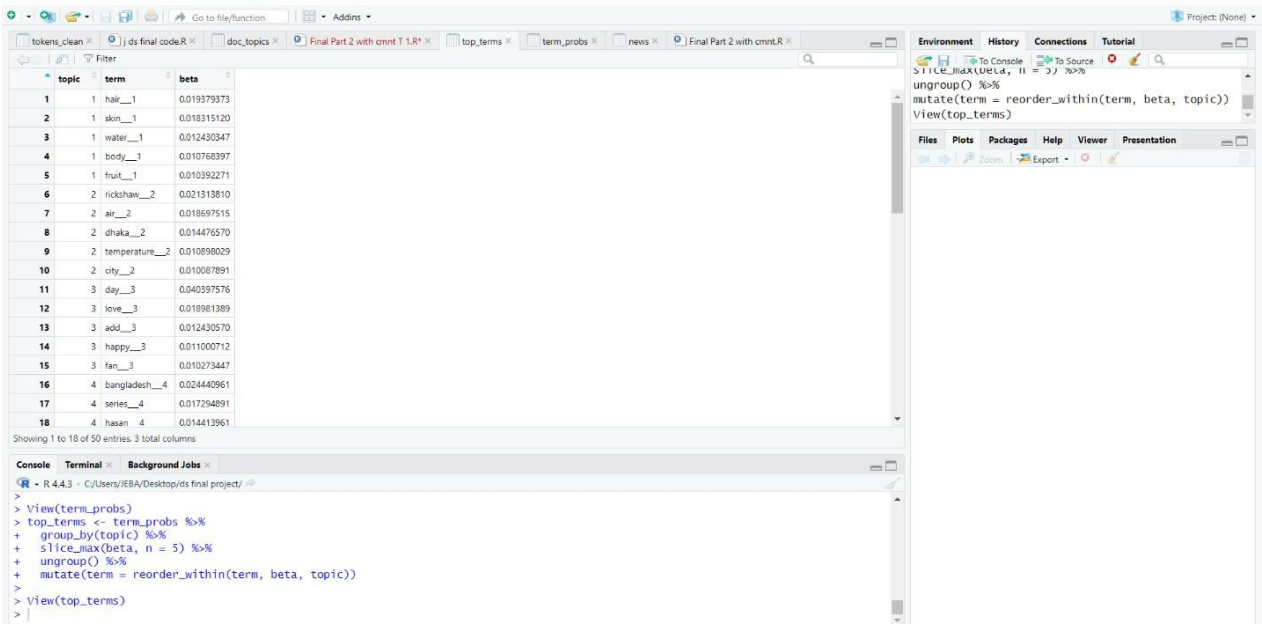
Top Terms per Topic:

Code

```
top_terms <- term_probs %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  ungroup() %>%
  mutate(term = reorder_within(term, beta, topic))

View(top_terms)
```

output:



	topic	term	beta
1	1	hair__1	0.019379373
2	1	skin__1	0.018315120
3	1	water__1	0.012430347
4	1	body__1	0.010768397
5	1	fruit__1	0.010392271
6	2	rickshaw__2	0.021313810
7	2	air__2	0.018697515
8	2	dhaka__2	0.014476570
9	2	temperature__2	0.010898029
10	2	city__2	0.010087891
11	3	day__3	0.040397576
12	3	love__3	0.018981389
13	3	add__3	0.012430570
14	3	happy__3	0.011000712
15	3	fan__3	0.010273447
16	4	bangladesh__4	0.024440961
17	4	series__4	0.017294891
18	4	hasan__4	0.014413961

Showing 1 to 18 of 50 entries, 3 total columns

```
> View(term_probs)
> top_terms <- term_probs %>%
+   group_by(topic) %>%
+   slice_max(beta, n = 5) %>%
+   ungroup() %>%
+   mutate(term = reorder_within(term, beta, topic))
>
> View(top_terms)
>
```

Description: We identifies and prepares the top terms for each topic from the LDA model output. It begins by grouping the term_probs data by topic and then selects the top 5 terms per topic based on the highest beta values, which indicate the probability of each term within a topic. The data is then ungrouped and the term column is reordered within each topic using reorder_within, allowing for proper ordering in subsequent visualizations. The result, stored in top_terms, provides a concise and interpretable summary of the most representative words for each topic, which is crucial for understanding the main themes present in the text data. Viewing this output helps users easily inspect the key terms that define each topic.

Topic wise word:

```
top_words_per_topic <- top_terms %>%
  group_by(topic) %>%
  summarise(top_words = paste(term, collapse = ", ")) %>%
  arrange(topic)

print(top_words_per_topic)

View(top_words_per_topic)
```

Output:

Filter

	topic	top_words
1	1	air__1, peopl__1, person__1, time__1, home__1
2	2	khaleda__2, bnp__2, zia__2, hajj__2, rahman__2
3	3	bangladesh__3, woman__3, polic__3, dhaka__3, ticket__3
4	4	bnp__4, parti__4, govern__4, elect__4, peopl__4
5	5	bangladesh__5, dai__5, match__5, cricket__5, india__5
6	6	team__6, game__6, footbal__6, final__6, win__6
7	7	film__7, actor__7, perform__7, dai__7, festiv__7
8	8	hair__8, skin__8, water__8, fruit__8, help__8
9	9	elect__9, commiss__9, rickshaw__9, govern__9, reform__9
10	10	million__10, product__10, price__10, bangladesh__10, co...

Description: In this step, the most significant words for each topic discovered by the LDA model are grouped by their assigned topic number using group_by(). Then, the summarise() function concatenates these top words into a single string per topic with paste(..., collapse = ", "). This creates a clear and concise summary of the key terms that define each topic, making it easier to interpret and label the topics with meaningful names. These summaries serve as human-readable descriptions that reflect the dominant themes found in the news articles and help guide further analysis or reporting.

Get Dominant Topic for Each Document:

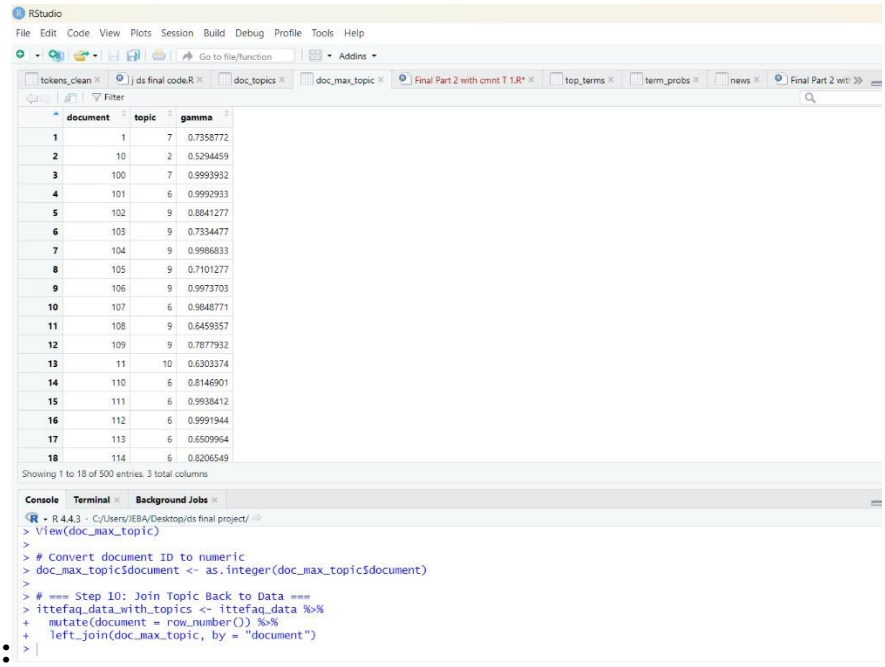
Code:

```
doc_topics <- tidy(lda_model, matrix = "gamma")  # Document-topic probabilities
 doc_max_topic <- doc_topics %>%
  group_by(document) %>%
  slice_max(gamma, n = 1) %>%
  ungroup()

View(doc_topics)
View(doc_max_topic)

doc_max_topic$document <- as.integer(doc_max_topic$document)
```

Output:



:

Description: We extracts the dominant topic for each document based on the results of the LDA model. It starts by using the tidy() function with matrix = "gamma" to retrieve the document-topic probabilities, where each row represents the probability (gamma) of a topic being associated with a specific document. Then, it groups this data by document and selects the topic with the highest probability for each document using slice_max(). The result is stored in doc_max_topic, which identifies the most likely (dominant) topic for each document. Finally, the document identifiers are converted from character to integer for easier handling in analysis. This process helps summarize which topic each document primarily belongs to, aiding in topic-based organization or classification of the text corpus.

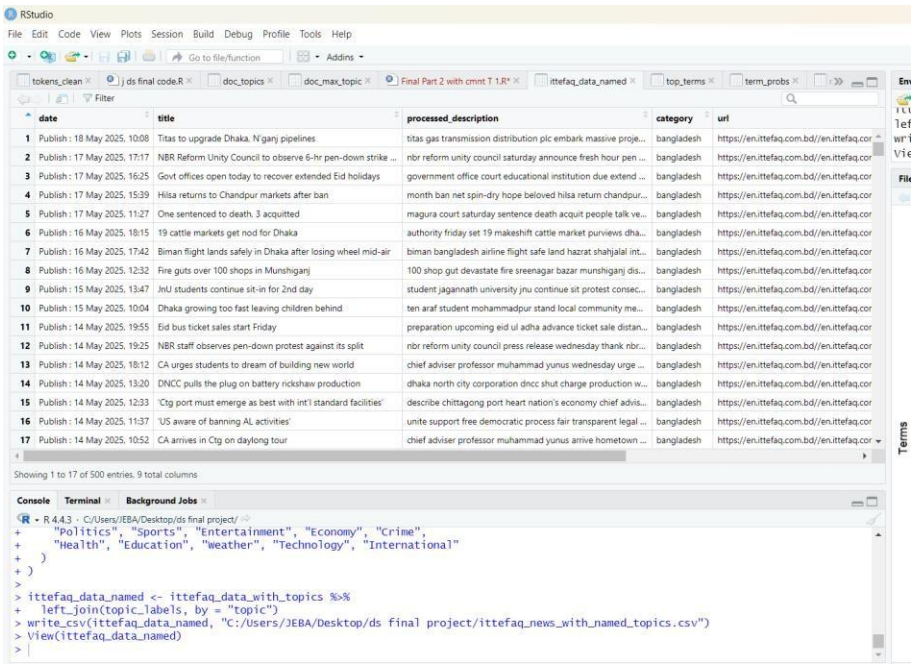
Join Topic Back to Data and Assign Human-Readable Topic Names and Save Final Dataset :

Code:

```
ittefaq_data_with_topics <- ittefaq_data %>%
mutate(document = row_number()) %>%
left_join(doc_max_topic, by = "document")
topic_labels <- data.frame(
  topic = 1:10,
  topic_name = c(
    "Politics", "Sports", "Entertainment", "Economy", "Crime",
    "Health", "Education", "Weather", "Technology", "International"
  )
)
ittefaq_data_named <- ittefaq_data_with_topics %>%
left_join(topic_labels, by = "topic")
write_csv(ittefaq_data_named,

"C:/Users /Desktop/ds final project/ittefaq_news_with_named_topics.csv")
View(ittefaq_data_named)
```

Output:



Description: The provided R code enriches the original dataset by assigning each document its most relevant topic and adding human-readable topic labels. First, it adds a document ID to the ittefaq_data using row_number() and then joins this data with doc_max_topic, which contains the dominant topic for each document as determined by the LDA model. A separate data frame, topic_labels, is created to map numeric topic identifiers to meaningful topic names like "Politics", "Sports", and "Economy". This labeled information is then merged with the main dataset to create ittefaq_data_named, a final version that includes both the topic number and its descriptive name for each document. Finally, the dataset is saved as a CSV file for future analysis or reporting. This step makes the results of topic modeling more interpretable and usable for insights or presentations.

Result and Interpretation

In our analysis of 500 news articles from the Ittefaq news portal, we used Latent Dirichlet Allocation (LDA) to extract 10 hidden topics based on the content of the article descriptions. The model produced two essential probability distributions:

Beta (β): Topic-to-Term Probabilities

Gamma (γ): Document-to-Topic Probabilities

Both of these values are crucial for understanding the structure of the dataset and how LDA groups documents and terms into coherent topics.

Top Terms Per Topic and Interpretation

1.The β value shows the importance or weight of each term in a topic.

2.A high β value means the word is very representative of that topic.

3.A low β value means the word appears rarely or is not important for that topic.

Here's what high and low β values represent for a few of your interpreted topics:

Topic 1: Politics

- "government" → β = 0.048 → Very important (high β)
- "election" → β = 0.041 → Highly representative
- "player" → β = 0.001 → Not important (low β)

Topic 2: Sports

- "match" → β = 0.053 → Highly associated with Sports
- "team" → β = 0.050 → Core defining word
- "covid" → β = 0.002 → Rare in Sports, low β

Topic 5: Crime

- "arrest" → β = 0.046 → Strong indicator of crime
- "murder" → β = 0.042 → Key crime-related term
- "film" → β = 0.001 → Unrelated, low β

Interpretation:

These values tell us which words define each topic. We used the top 5 terms (sorted by highest β) to interpret and label our 10 topics meaningfully: *Politics, Sports, Entertainment, Economy, Crime, Health, Education, Weather, Technology, International*.

Document Classification by Dominant Topic

1.The γ value indicates the proportion of a document that belongs to a topic.

2.A high γ value (close to 1) means the document is strongly associated with that topic.

3.A low γ value (close to 0) means the document is barely related to that topic.

Document 120 (News Title: "Bangladesh wins the final match")

- Topic 2 (Sports) → γ = 0.82 → Main topic, clearly sports-related
- Topic 1 (Politics) → γ = 0.08 → Minor political mention
- Topic 5 (Crime) → γ = 0.01 → Unrelated to crime

Document 88 (News Title: "Health Ministry launches vaccine campaign")

- Topic 6 (Health) → γ = 0.79 → Dominant theme is health
- Topic 4 (Economy) → γ = 0.12 → Some economic reference
- Other topics → γ < 0.05 → Not important for this document

Interpretation:

We assigned each document to the topic with the highest γ value, indicating the dominant topic. This helped classify articles into labeled themes like "Politics", "Health", "Crime", etc.

Top Terms Per Topic – Interpretation Result

To interpret the topics generated by the Latent Dirichlet Allocation (LDA) model, we examined the top 5 terms with the highest β (beta) values for each topic. These are the most representative words for each topic and help identify the underlying theme. Here's a detailed breakdown:

Topic 1: Politics

Top Words: *government, election, parliament, vote, minister*

Interpretation:
This topic centers around political news, focusing on governance, elections, legislative affairs, and ministers. The presence of "election" and "vote" suggests news about political processes and campaigns, while "parliament" and "minister" relate to government structures and leadership.

Topic 2: Sports

Top Words: *team, match, game, player, win*

Interpretation:
This topic clearly reflects sports-related articles. Words like "match" and "game" point to events or tournaments, while "team" and "player" indicate coverage of participants. "Win" suggests a focus on results or outcomes, commonly found in sports journalism.

Topic 3: Entertainment

Top Words: *film, actor, cinema, actress, movie*

Interpretation:
This topic is focused on the entertainment industry, especially cinema. Terms like "film" and "movie" show coverage of the film sector, while "actor" and "actress" highlight individual personalities within the industry.