****

**FACULTY OF SCIENCE & TECHNOLOGY**

**INTRODUCTION TO DATA SCIENCE**

**Spring 2024-25**

**Section: B**

**Group 10**

**Supervised by  
TOHEDUL ISLAM**

**Assistant Professor, Computer Science**

**Submitted by:**

|  |  |
| --- | --- |
| *Name* | *ID* |
| Md Tareq Jamil Sarkar | **22-46619-1** |
| Md Mehedi Hasan | **22-46322-1** |
| Hasin Almas Sifat | **22-48679-3** |

Submitted On: 23 june 2025

**Introduction**

First, we scraped online news articles related to Middle East issues from the Prothom Alo English website. The objective is to gather unstructured text data from multiple webpages, clean and preprocess the textual content, and apply topic modeling using Latent Dirichlet Allocation (LDA). By identifying latent themes from the articles, we aim to explore topic distributions and term relevance across documents.

**Dataset Description**

The data consists of 10 different English news articles collected from:

<https://en.prothomalo.com/international>

Each article includes the title, content paragraphs, and meta structure. After scraping, the texts were aggregated and structured into a CSV file containing the title and corresponding paragraph texts.

**Required Libraries**

install.packages("rvest")

install.packages("tm")

install.packages("SnowballC")

install.packages("tidytext")

install.packages("dplyr")

install.packages("textstem")

install.packages("topicmodels")

install.packages("textmineR")

install.packages("LDAvis")

install.packages("ggplot2")

install.packages("slam")

install.packages("wordcloud")

These libraries were used for:

* **Scraping** (rvest)
* **Text mining & cleaning** (tm, tidytext, textstem)
* **Topic modeling** (topicmodels)
* **Visualization** (ggplot2, wordcloud, LDAvis)

**Web Scraping**

Ten separate news articles were scraped using rvest. HTML tags such as <title> and <p> were extracted and converted into text using:

html\_node(link, 'p') %>% html\_text()

The title and paragraphs were combined into data frames and merged using rbind() to form a complete dataset. The final result was exported into:

write.csv(scraped\_data, "scraped\_data.csv")

A screenshot of a computer

AI-generated content may be incorrect.

Figure: 1

Displays the scraped\_data.csv file, which is generated

**Text Cleaning and Tokenization**

After scraping:

* Articles were grouped by title and paragraphs combined.
* Text was converted to lowercase.
* Tokenized using unnest\_tokens().
* Common stopwords were removed.
* Numbers, punctuation, and short words were filtered.
* Lemmatization was applied using lemmatize\_words() from textstem.

Clean tokens were saved in:

write.csv(clean\_tokens, "lemmatized\_data.csv")

A screenshot of a computer

AI-generated content may be incorrect.

Figure: 2

Displays the lemmatized\_data.csv file, which is generated

**Using Document-Term Matrix (DTM) for Topic Modelling**

The DTM is a structured format where each row represents a document (news article), and each column represents a unique word (term). The values indicate how often each word appears in a document.

We use the DTM because topic modeling (like LDA) works on numerical representations of text. DTM gives a quantitative view of text data that LDA can use to identify patterns and topics.

Code :

dtm\_input <- clean\_tokens %>%

count(Title, lemma) %>%

cast\_dtm(document = Title, term = lemma, value = n)

**Topic Modeling using LDA**

LDA is a statistical model used to discover **hidden topics** in a collection of documents. It treats each document as a mix of topics and each topic as a mix of words.

We use LDA to **automatically find topics** in text data without labeling. It groups words that frequently appear together and assigns them to a topic.

**Code**:

**lda\_model <- LDA(dtm\_input, k = 3, control = list(seed = 1234))**

**Output :**

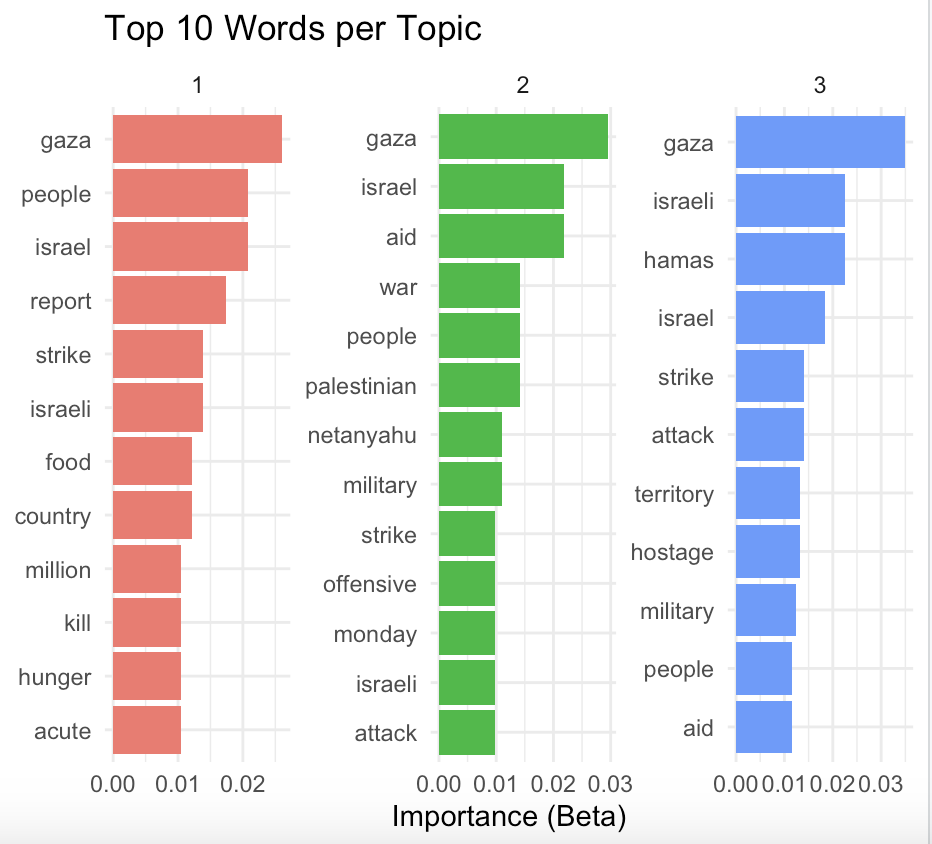
****

Figure: 3

Each topic is a theme the model found based on word usage patterns.

**Top Terms per Topic**

Each topic is defined by the top 10 most probable words (β values).

**Model Evaluation**

To decide how many topics (k) give the best result. Two common metrics:

* Perplexity: Lower is better. Shows how well the model predicts unseen data.
* Log Likelihood: Higher is better. Shows how probable the model's structure is.

To evaluate model performance:

Perplexity and Log-Likelihood were plotted against topic counts from k = 2 to k = 10.

ggplot(eval\_df, aes(x = Topics, y = Perplexity)) + geom\_line()

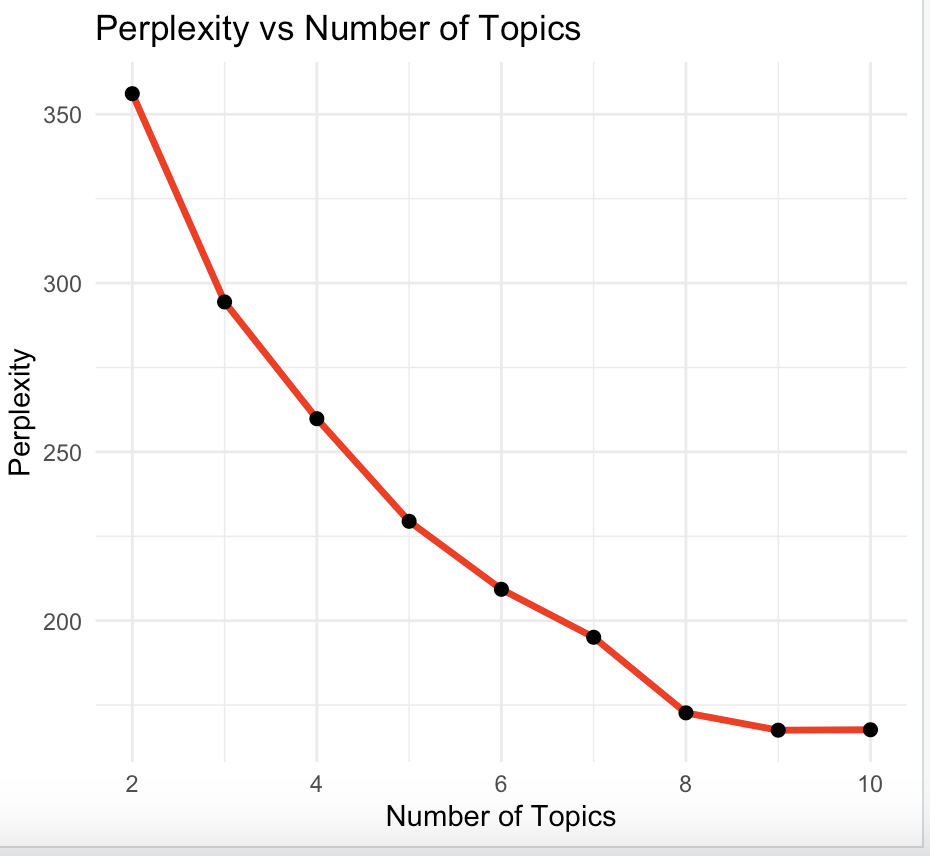


Figure: 4

A graph with a line

AI-generated content may be incorrect.

Figure: 5

These graphs help to identify the optimal number of topics (usually where perplexity plateaus).

**Visualization**

Using LDAvis, we visualized topic distances and word relevance interactively.

serVis(json\_lda, open.browser = TRUE)

A screenshot of a graph

AI-generated content may be incorrect.

Figure: 6

**Word Clouds**

Generated one word cloud for each topic using wordcloud() to visually display high-probability terms:

wordcloud(words = topic\_words$term, freq = topic\_words$beta, ...)



Figure: 7

**Conclusion**

We successfully scraped, cleaned, and analyzed online text data using R. By applying LDA, we identified key themes discussed in Middle East news articles. The project demonstrates how web scraping and topic modeling can be combined to derive insights from unstructured text. The process involved handling HTML data, preprocessing for text mining, and implementing statistical modeling for thematic analysis.The final models and plots provide a powerful summary of public discourse on current international issues.