AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Choose an item.

Choose an item.

Assignment Title:	Topic Modeling Project Report		
Assignment No:	3	Date of Submission:	25 June 2025
Course Title:	Introduction to Data Science		
Course Code:	Click here to enter text.	Section:	D
Semester:	Spring 2024-25	Course Teacher:	Tohedul Islam

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.:

No	Name	ID	Program	Signature
1	Kaniz Faria Ahamed	22-46429-1	BSc [CSE]	
2	Jinia Sultana Sama	22-46301-1	BSc [CSE]	
3	Airin Akther	22-46744-1	BSc [CSE]	
4	Tahmeed Ali Patwary	21-44428-1	BSc [CSE]	

Faculty use only		
FACULTY COMMENTS		
	Marks Obtained	
	Total Marks	

Topic Modeling

- **Topic Modeling:**
  In unsupervised learning, subjects that are fundamental and analyzed across multiple documents are found by Topic Modeling. It is done without using any labeled data. Instead of using categories that are predefined, it inspects how words are likely to be appear together across multiple documents and groups these unique words into topics based on those patterns. By analyzing these word patterns, topic modeling can instinctively give away what subjects are being discussed in the text. (Steyvers, 2007).

- **DTM (Document Term Matrix):**

  Document-Term Matrix (DTM) is a mathematical representation of a collection of documents in which rows represent a document, and columns represent rear terms. The cells of the matrix generally reflect how frequently the terms appear in the respective documents. This arrangement is usually used to make algorithms evaluate and interpret unstructured textual data in natural language processing (NLP) and text mining. (Feinerer, 2008).

- **How DTM works:**

  Document-Term Matrix (DTM) processes a set of text data and extracts meaningful words from them which are often known as terms. In this process texts are split into small tokens, then common stop-words are removed, and sometimes words are converted to their root forms.

  In a DTM, documents are represented as rows and terms as columns, where each cell shows how often a particular term occurs in a document. This numeric representation analyzes text data using the standard procedure.

  Since quite a few terms can be absent from various documents, it can cause the resulting matrix to become sparse. Weighting methods such as Term Frequency-Inverse Document Frequency are used to emphasize terms that are more important in specific documents in the entire collection to enhance the analysis. (Cambridge, 2009).

- **Code Screenshot:**

```r
libs <- c("readr", "tm", "topicmodels", "tidyverse", "tidytext", "reshape2", "wordcloud")
for (lib in libs) {
  if (!require(lib, character.only = TRUE)) {
    install.packages(lib)
    library(lib, character.only = TRUE)
  }
}

processed_news <- read_csv("aljazeera_processed_final_file.csv")
corpus_text <- processed_news$processed

top_word_count <- 7
topic_number <- 7

corpus <- Corpus(VectorSource(corpus_text))
doc_term_matrix <- DocumentTermMatrix(corpus)


doc_term_matrix <- removeSparseTerms(doc_term_matrix, 0.95)
doc_term_matrix_dtm<- as.matrix(doc_term_matrix)
```

Output screenshot:

doc_term_matrix_dtm | IDS final project(group-5).R*

Filter | Cols: 1 - 50

	accord	across	add	address	affair	afternoon	agency	aid	air	allow	alone	along	announce	area
1	2	1	2	1	1	1	1	9	1	4	1	1	1	
2	1	0	2	0	0	0	0	1	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	2	0	1	0	0	0	
5	1	1	0	0	0	0	0	9	0	2	0	0	0	
6	1	1	1	0	0	0	0	0	0	1	1	0	0	
7	1	0	1	0	0	0	0	0	1	0	0	0	1	
8	2	1	2	1	1	1	1	9	1	4	1	1	1	
9	1	0	2	0	0	0	0	1	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	2	0	1	0	0	0	
12	2	0	0	0	0	0	0	0	0	1	0	0	0	
13	1	0	1	0	0	0	0	0	1	0	0	0	1	
14	0	0	1	0	0	0	0	0	0	0	1	0	0	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	1	0	0	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	0	0	0	0	0	0	0	0	0	

Description:

The Document-Term Matrix (DTM) contains 51 rows, each representing a single news article, and columns representing distinct words. A sparsity threshold of 0.95 was applied here, indicating that only words that appear at least 5 times or more were retained. This helps eliminate rare terms and reduce dimensionality. The matrix shows the frequency of each retained word in every article, providing a structured format for further text analysis.

- **LDA (Latent Dirichlet allocation) definition:**
  Latent Dirichlet allocation is a generative probabilistic model built on two important assumptions. The first assumption is that every document within a corpus is composed of several underlying topics. On the other hand, the second assumption conveys the idea that each topic is defined by a blend of words from the overall vocabulary. Hence, the goal of LDA is to uncover these topics and determine the distribution of topics for each document, as well as words within each topic (Bystrov, 2024).

- **How LDA Algorithm works:**
  Among the various methods available, Latent Dirichlet Allocation (LDA) is regarded as one of the most commonly used and effective algorithms for topic modeling. It follows a probabilistic framework based on the following ideas:

  - *Document-to-Topic Distribution:* Each document is represented as a distribution across various topics. This indicates that a document can belong to several topics, each with a certain probability.

  - *Topic-Word Distribution:* Each topic is regarded as a distribution over words. This suggests that a topic is defined by a collection of words, each linked with a probability of occurring in that topic.

  - *Generative Process:* LDA uses a generative process to produce documents. This involves choosing a distribution of topics for each document, selecting a topic based on the distribution for each word in the document, and generating the word from the chosen topic's word distribution.

    (GeeksforGeeks, 2024)

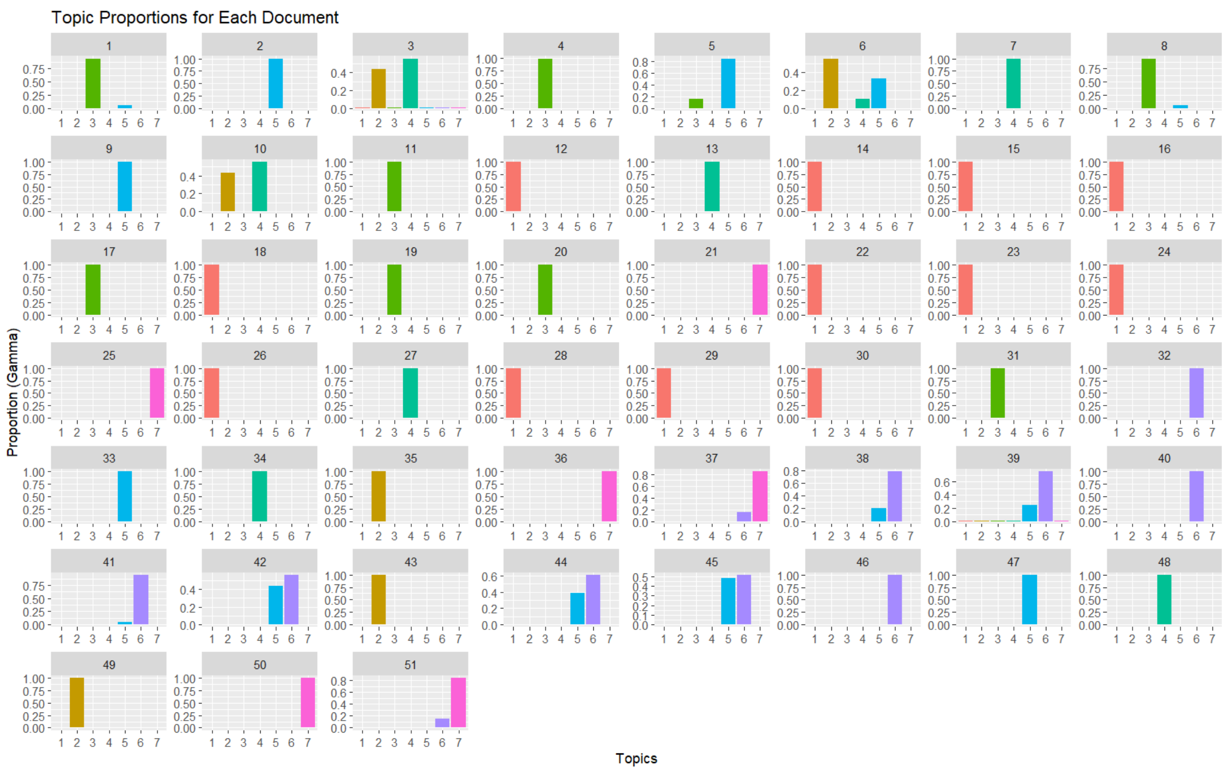
- **Code screenshot:**

```r
lda_model <- LDA(doc_term_matrix, k = topic_number, control = list(seed = 123))
unwanted_terms <- c("say", "jazeera","bbc","get")


topics <- tidy(lda_model, matrix = "beta") %>%
  filter(!term %in% unwanted_terms) %>%
  group_by(topic) %>%
  slice_max(beta, n = top_word_count, with_ties = FALSE) %>%
  ungroup()

print(topics)

topic_proportions <- tidy(lda_model, matrix = "gamma") %>%
  mutate(document = as.numeric(document))

topic_labels <- c(
  "1" = "league Sports",
  "2" = "India Politics",
  "3" = "Middle East",
  "4" = "kashmir issue",
  "5" = "Military Conflict ",
  "6" = "India Pakistan conflict",
  "7" = "athletic sport"
)

topics <- topics %>%
  mutate(topic_name = topic_labels[as.character(topic)])


p_beta <- topics %>%
  ggplot(aes(x = reorder(term, beta), y = beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic_name, scales = "free") +
  coord_flip() +
  labs(
    title = "Top Terms for Each Topic",
    x = "Terms",
    y = "Beta"
  )

p_gamma <- topic_proportions %>%
  ggplot(aes(x = factor(topic), y = gamma, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ document, scales = "free") +
  labs(
    title = "Topic Proportions for Each Document",
    x = "Topics",
    y = "Proportion (Gamma)"
  )

print(p_beta)
print(p_gamma)

ggsave("top_terms.png", plot = p_beta, width = 12, height = 8, dpi = 300)
ggsave("topic_proportions.png", plot = p_gamma, width = 20, height = 10, dpi = 300)
```

- **Output screenshot:**

```
> lda_model <- LDA(doc_term_matrix, k = topic_number, control = list(seed = 123))
> unwanted_terms <- c("say", "jazeera","bbc","get")
>
>
> topics <- tidy(lda_model, matrix = "beta") %>%
+   filter(!term %in% unwanted_terms) %>%
+   group_by(topic) %>%
+   slice_max(beta, n = top_word_count, with_ties = FALSE) %>%
+   ungroup()
>
> print(topics)
# A tibble: 49 x 3
   topic term        beta
   <int> <chr>      <dbl>
 1     1 military 0.0251
 2     1 pakistan 0.0229
 3     1 army     0.0145
 4     1 munir    0.0107
 5     1 india    0.0105
 6     1 khan     0.00994
 7     1 indian   0.00951
 8     2 india    0.0185
 9     2 pakistan 0.0169
10     2 indian   0.0150
# ℹ 39 more rows
# ℹ Use `print(n = ...)` to see more rows
>
> topic_proportions <- tidy(lda_model, matrix = "gamma") %>%
+   mutate(document = as.numeric(document))
>
```

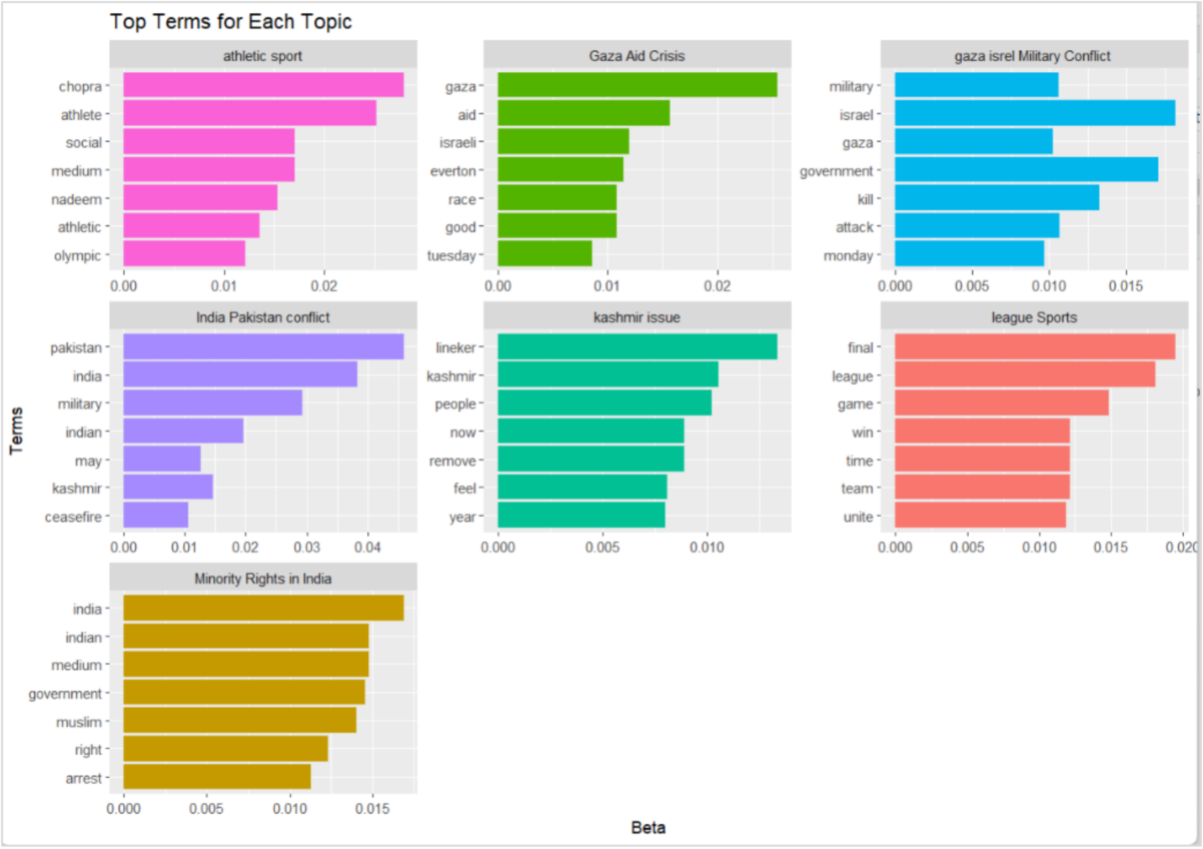
Topic Proportion for each document:



Description:

The visualization above shows the topic proportions for each document. Each small plot corresponds to one document, where the x-axis (independent variable) shows the topic number from 1 to 7 and the y-axis (dependent variable) represents the gamma (γ) value, which reflects the proportion of each topic within the document. As can be seen, in most of the documents, they are mainly dominated by a single topic, shown by a single tall bar. For example, Document 1 is mostly composed of Topic 3, while Document 50 is dominated by Topic 7.

Top Terms for each Topic:



Description:

This visualization on top shows the most significant keywords or terms associated with each of the seven topics derived from the LDA model. In this, each subplot represents a unique topic, where the dependent variable lists the top terms and the independent variable shows their corresponding beta (β) values, suggesting how strongly each term is associated with that topic. Greater beta values represent higher relevance to the topic. For instance, Topic 5 is labeled as “Gaza Israel Military Conflict,” which includes terms like military, Israel, and attack, suggesting a geopolitical focus. Overall, this visualization presents the main theme of each topic by clearly showing its most used terms.

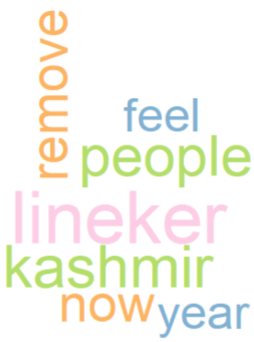
- **Evaluation matrix (Word Cloud):**

Topic 1 - league Sports

Topic 2 - Minority Rights in India

Topic 3 - Gaza Aid Crisis

Topic 4 - kashmir issue



D

Topic 5 - gaza isrel Military Conflict

Topic 6 - India Pakistan conflict

Topic 7 - athletic sport

Description:

The word clouds present seven clearly defined topics generated through LDA, each revealing a unique theme. Topics 1 and 7 are focused on sports, emphasizing terms like *game, final, athlete, and Olympic,* indicating coverage of major competitive sports events and figures. Topics 2, 4, and 6 deal with political and social issues in South Asia, where Topic 2 discusses Minority rights in India, Topic 4 highlights the Kashmir conflict, and Topic 6 focuses on India-Pakistan tensions, using keywords such as *ceasefire and military*. Meanwhile, Topics 3 and 5 address the Gaza crisis, with Topic 3 focusing on *humanitarian efforts* and Topic 5 emphasizing military conflict. Overall, these word cloud visualizations reveal the main concerns and discourse patterns within the dataset.

References:

Bystrov, V. N.-K.-B. (2024). Choosing the number of topics in LDA Models–a Monte Carlo comparison of selection criteria. *Journal of Machine Learning Research*.

Cambridge, U. P. (2009). *Introduction to information retrieval*.

Feinerer, I. H. (2008). *Text mining infrastructure in R. Journal of statistical software*, 25, 1-54.

GeeksforGeeks. (2024, June 11). *Topic Modeling Using Latent Dirichlet Allocation (LDA)*. Retrieved from https://www.geeksforgeeks.org/nlp/topic-modeling-using-latent-dirichlet-allocation-lda/

Steyvers, M. &. (2007). *Handbook of latent semantic analysis.* Psychology Press.