**AMERICAN INTERNATIONALA close up of a sign

Description automatically generated**

**UNIVERSITY-BANGLADESH**

**Faculty of Science and Technology**

**Project**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assignment Title: | Final Term Project Report | | | |
| Assignment No: | 01 | | Date of Submission: | 24 June 2024 |
| Course Title: | Introduction to Data Science | | | |
| Course Code: | CSC4180 | | Section: | F |
| 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.: | 03 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Name** | **ID** | **Program** | **Signature** |
| 1 | NAFIZ AL NAHIAN ALVY | 22-46036-1 | BSc [CSE] |  |
| 2 | A.H.M TANVIR | 22-47034-1 | BSc [CSE] |  |
| 3 | SADMAN SHAKIB | 21-44673-1 | BSc [CSE] |  |
| 4 | ANNAN KUMAR SAHA | 22-46753-1 | BSc [CSE] |  |

|  |  |  |
| --- | --- | --- |
| ***Faculty use only*** | | |
| FACULTY COMMENTS | **Marks Obtained** |  |
|  |
|  |
|  | **Total Marks** |  |
|  |
|  |

# Document-Term Matrix (DTM) and Latent Dirichlet Allocation (LDA) for Topic Modeling

## Introduction

The **Document-Term Matrix (DTM)** and **Latent Dirichlet Allocation (LDA)** are critical methodologies in natural language processing for analyzing textual data and uncovering latent topics within document collections. This report provides a formal and structured explanation of the DTM, its construction process, and its application in topic modeling using LDA, while retaining the original visualizations and media references.

## Document-Term Matrix (DTM)

### Definition

The **Document-Term Matrix (DTM)** is a mathematical matrix that quantifies the frequency of terms (words) across a collection of documents. It is generated using the DocumentTermMatrix function from the tm package in R, based on a preprocessed text corpus.

* **Rows**: Each row corresponds to an individual document within the corpus.
* **Columns**: Each column represents a unique term identified in the corpus after preprocessing.
* **Entries**: Each cell contains the frequency of a term in a specific document, typically using term frequency (TF) weighing by default.

### Operational Mechanism

The construction of a DTM involves several systematic steps:

**Step 1: Text Preprocessing**

To ensure data quality and minimize noise, the text undergoes preprocessing:

* Conversion to lowercase for uniformity.
* Removal of punctuation to eliminate non-semantic characters.
* Exclusion of numbers unless contextually relevant.
* Elimination of stop words (e.g., “the,” “is,” “and”) to focus on significant terms.
* Application of stemming to reduce words to their root form (e.g., “running” to “run”).
* Removal of excess whitespace to standardize formatting.

**Step 2: Vocabulary Creation**

A vocabulary is established, comprising all unique terms across the preprocessed documents.

**Example**: For the documents:

* Document 1: “cat eats mouse”
* Document 2: “dog eats bone”

The resulting vocabulary would be: [“bone”, “cat”, “dog”, “eats”, “mouse”].

**Step 3: Frequency Counting**

The frequency of each term in each document is calculated to populate the DTM.

**Example DTM**:

| **Document** | **bone** | **cat** | **dog** | **eats** | **mouse** |
| --- | --- | --- | --- | --- | --- |
| Doc1 | 0 | 1 | 0 | 1 | 1 |
| Doc2 | 1 | 0 | 1 | 1 | 0 |

**Step 4: Weighting (Optional)**

The matrix values can be weighed using various methods:

* **Term Frequency (TF)**: Raw counts of term occurrences (default).
* **Term Frequency-Inverse Document Frequency (TF-IDF)**: Adjusts for term commonality across documents.
* **Binary Weighting**: Assigns 1 if a term is present, 0 otherwise.

### Application in Topic Modeling

The DTM serves as the primary input for **Latent Dirichlet Allocation (LDA)**, a statistical method for identifying abstract topics in documents. By providing a numerical representation of text, the DTM enables LDA to:

* Detect patterns of co-occurring words within documents.
* Assign probabilistic topic distributions to each document based on word frequencies.

**Code**:  
**A black and white text

AI-generated content may be incorrect.**

**Output**:  
**A black text on a white background

AI-generated content may be incorrect.**

### 

### Output Description

The DTM output provides the following metrics:

| **Term** | **Description** |
| --- | --- |
| **documents: 5** | The corpus contains 5 documents after filtering. |
| **terms: 1546** | There are 1,546 unique terms across the documents after preprocessing. |
| **Non-/sparse entries: 2214/5516** | Of 5 × 1546 = 7,730 possible entries, 2,214 are non-zero, indicating term presence. |
| **Sparsity: 71%** | 71% of the matrix entries are zero, a common characteristic of text data. |
| **Maximal term length: 14** | The longest term in the vocabulary contains 14 characters. |
| **Weighting: term frequency (tf)** | Matrix entries represent the raw frequency of terms in each document. |

## Latent Dirichlet Allocation (LDA) for Topic Modeling

### Definition

**Latent Dirichlet Allocation (LDA)** is a generative probabilistic model designed to identify latent topics within a document collection. It assumes that documents are composed of multiple topics, and each topic is characterized by a distribution over words.

### Key Concepts

1. **Documents, Topics, and Words**:
   * A document is a collection of words (e.g., a news article).
   * A topic is a probability distribution over words, with certain words more likely to appear (e.g., “dog,” “pet,” “bark” for a pet-related topic).
   * Each document is a mixture of topics with varying proportions.
2. **Generative Model**:
   * LDA posits that documents are generated through a probabilistic process and infers topics by reversing this process.
3. **Parameters**:
   * **KKK**: Number of topics, specified by the user.
   * **α**: Controls the distribution of topics in documents (lower α results in fewer dominant topics per document).
   * **β**: Controls the distribution of words in topics (lower β results in fewer dominant words per topic).

### Operational Mechanism

**1. Generative Process**

LDA assumes the following process for generating a corpus of ( D ) documents:

* Select the number of topics ( K ).
* For each topic (k=1,2,...,Kk = 1, 2, ..., Kk=1,2,...,K):
  + Draw a distribution over words, ϕk\phi\_kϕk​, from a Dirichlet distribution with parameter β\betaβ. This defines the probability of each word in the vocabulary for topic kkk. For example, topic 1 might assign high probabilities to "dog," "cat," and "pet."
* For each document (d=1,2,...,Dd = 1, 2, ..., Dd=1,2,...,D):
  + Draw a distribution over topics, θd\theta\_dθd​, from a Dirichlet distribution with parameter α\alphaα. This defines the proportion of each topic in document ddd. For example, document 1 might be 70% topic 1 and 30% topic 2.
  + For each word wiw\_iwi​ (where i=1,2,...,Ndi = 1, 2, ..., N\_di=1,2,...,Nd​, and NdN\_dNd​ is the number of words in document ddd):
    - Choose a topic ziz\_izi​ from the topic distribution θd\theta\_dθd​. For example, pick topic 1 with 70% probability.
    - Choose a word wiw\_iwi​ from the word distribution ϕzi\phi\_{z\_i}ϕzi​​ for the selected topic. For example, if topic 1 is selected, choose "dog" with its corresponding probability.

**2. Inference**

LDA aims to infer:

* **Topic-Word Distributions (ϕk\phi\_kϕk):** Probability of words for each topic.
* **Document-Topic Distributions (θd\theta\_dθd​):** Proportion of topics in each document.
* **Word-Topic Assignments (ziz\_izi​):** Topic assigned to each word.

Due to computational complexity, approximate inference methods are employed:

* **Gibbs Sampling**: Iteratively samples topic assignments for words.
* **Variational Inference**: Optimizes a simpler distribution to approximate the true posterior.

**3. Output**

LDA produces:

* **Topic-Word Distributions (ϕk\phi\_kϕk​):** Probability distribution over the vocabulary for each topic.
* **Document-Topic Distributions (θd\theta\_dθd​):** Proportion of topics in each document.
* **Word-Topic Assignments (ziz\_izi​):** Topic assignments for each word.

### Rationale for Using LDA

LDA is employed for topic modeling due to:

1. **Unsupervised Learning**: It requires no labeled data, ideal for discovering hidden structures.
2. **Probabilistic Framework**: It provides interpretable probabilities for topics and words.
3. **Flexibility**: It accommodates varying topic counts and diverse text data.
4. **Scalability**: Approximate inference methods enable processing of large datasets.

**Code**:  
**A screenshot of a computer program

AI-generated content may be incorrect.**

**Output:**  
**A screenshot of a computer code

AI-generated content may be incorrect.**

## 

## LDA Output Analysis

### Parameters and Interpretation

1. **Alpha (****α):**
   * **Value**: 0.01352813, significantly lower than the default (50/K = 10) for (K = 5).
   * **Implication**: Low α promotes sparsity, encouraging documents to focus on a single topic, which may enhance topic distinctiveness but increase sensitivity to noise.
2. **Beta (β):**
   * **Description**: Represents the probability of words given a topic.
   * **Example for Topic 1**:
     + Top words: “school” (β=0.0367), “state” (β=0.03), “ban” (β=0.025), “student” (β=0.0183), “cellphon” (β=0.0117).
     + **Theme**: School policies, likely related to cellphone regulations.
   * Other topics exhibit similar patterns, with β values indicating word importance.
3. **Gamma (γ):**
   * **Description**: Indicates the proportion of each topic in each document.
   * **Example**:
     + Document 4: (γ=1.0) for Topic 1, indicating exclusive focus on school policies.
     + Document 5: (γ=1.0) for Topic 2, suggesting a distinct theme.
     + Documents 1, 2, 3: Low γ values (0.0000183–0.0000273), indicating weaker topic associations, possibly due to limited dataset size.
   * **Implication**: The dominant topic for each document is determined by the highest γ

### 

### Result Analysis:

**A graph of different colored bars

AI-generated content may be incorrect.**

**General Graph Analysis**

* **Graph Title: “Top Terms per Topic”**
* **X-axis (Beta):** Represents the probability (beta) of a word given a topic, i.e., how representative a term is for that topic.
* **Y-axis (Terms):** Top 10 terms for each topic.
* **Facets:** Each subplot represents a different topic, named using the top 5 words from that topic.
* **Color:** Each topic has its own unique color, making them visually distinct.

### Topic Descriptions

The LDA model identified five topics, characterized as follows:

1. **Topic 1: “gaza aid hama accord farra”**
   * **Themes**: Middle Eastern conflict, Gaza, aid, political agreements.
   * **Top Terms**: gaza, aid, said, news, najjar, israel, farra, accord.
   * **Interpretation**: News articles focusing on Gaza-related aid and diplomacy.
   * **Visualization**: Word cloud emphasizing “gaza” as the most representative term.  
     A close up of words. AI-generated content may be incorrect.
2. **Topic 2: “said trump russia putin monda”**
   * **Themes**: U.S. politics, Trump, Russia, diplomacy.
   * **Top Terms**: said, trump, russia, putin, monday, ukraine.
   * **Interpretation**: Political discussions involving Trump, Russia, and potentially Ukraine.
3. **Topic 3: “school state ban student bill”**
   * **Themes**: Education policy, legislation.
   * **Top Terms**: school, state, ban, student, bill, law, use.
   * **Interpretation**: Legislative measures impacting schools, such as cellphone bans.
4. **Topic 4: “servic recruit bonus year armi”**
   * **Themes**: Military recruitment, incentives.
   * **Top Terms**: servic, recruit, bonus, year, armi, marin, forc.
   * **Interpretation**: Articles addressing military service and recruitment incentives.
5. **Topic 5: “ukrain russian talk presid said”**
   * **Themes**: Russia-Ukraine conflict, diplomacy.
   * **Top Terms**: ukrain, russian, talk, presid, said, putin, trump.
   * **Interpretation**: News coverage of the Ukraine-Russia conflict and diplomatic efforts.

**Word Cloud Graph**:

**A close up of words

AI-generated content may be incorrect.**

**Word Cloud**:

* **Main Focus**: “gaza” (largest, highest β).
* **Other Terms**: “aid,” “hama,” “said,” “accord,” “israel,” “farra,” “najjar,” “news.”
* **Interpretation**: Highlights the prominence of terms in Topic 1, with “gaza” as the most significant.

## 

## Conclusion

The **Document-Term Matrix (DTM)** provides a structured numerical representation of text, facilitating topic modeling with **Latent Dirichlet Allocation (LDA)**. The DTM captures term frequencies, while LDA infers latent topics by modeling documents as mixtures of topics and topics as distributions over words. The resulting outputs, including topic-word distributions, document-topic distributions, and word-topic assignments, enable the identification of meaningful themes. Visualizations such as word clouds and top-term graphs enhance the interpretability of these findings, highlighting key terms and their relevance to identified topics.