Latent Dirichlet Allocation (LDA)

1. Exploring the books data:

Step 1: After loading the text files of all 6 books, lines of each book are read to create a list of texts and stored in the data frame.

Step 2: Each book is selected from the list of books, and a data frame is created for each book with the columns including, text, title, and line numbers.

Step 3: The lines that have the word ‘CHAPTER’, followed by any Roman numbers are identified to give a chapter ID to each chapter of the book, the chapter id is then stored in a new column.

Step 4: In the book ‘A Tale of Two Cities’, after chapters 1-6, the chapter number starts again from 1. So, the sequence of the chapter number in the column ‘chapter id’ is modified to show the continuous chapter id from 1 till the last chapter of the book.

Step 5: The first few rows as well as the last few rows of the text file of all the books are about the contents of the document and other additional details, which are not useful for us as we run the LDA on our data. So, these rows are removed.

Step 6: A new column ‘document’ is created, where the title and the chapter id are merged into a single column, and separated with the ’_’.

Step 7: The text column Is tokenized to split the text into the words, and the stop words are then removed from the data.

The dataset has 115423 rows and 3 columns,

- document – This column contains name of the book along with the chapter number separated by an underscore.
- word – this column contains the words repeated in each chapter of the book
- n – this column contains the number of times the words are repeated.

These are the top 12 words repeated in the entire dataset (Fig.1):

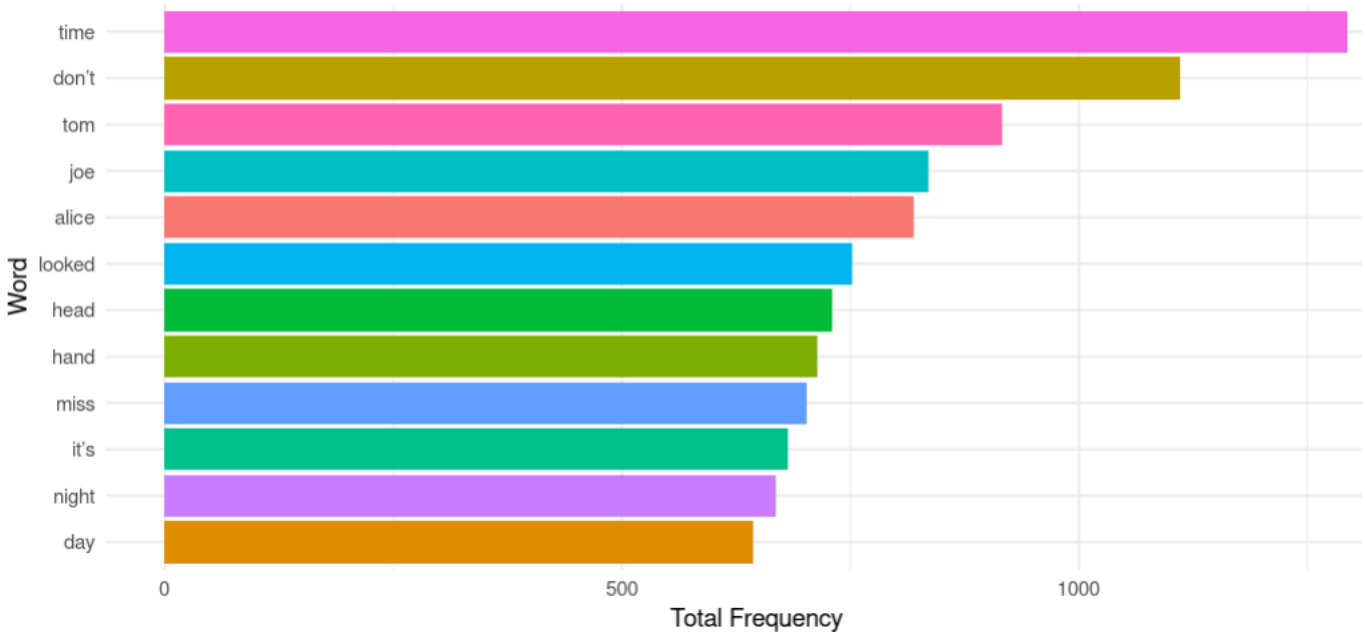

***Fig.1**. Top 12 most frequent words used in all 6 books combined.*

2. Applying the Document term matrix:

A document term matrix (DTM) is a matrix that describes the frequency of terms or words that occur in a book. In a DTM, rows represent books, and columns represent terms. Each entry in the matrix represents the frequency of the word in a document.

Applying the document term matrix to my dataset:

	de	queen	joe	en	alice	biddy	estella	tom	knight
Adventures of Huckleberry Finn_50	96	0	0	72	0	0	0	2	0
Through the Looking Glass_9	3	89	0	0	72	0	0	0	0
Great Expectations_57	0	0	88	0	0	15	1	0	0
Great Expectations_7	0	0	70	0	0	3	0	0	0
Great Expectations_17	0	0	5	0	0	63	6	0	0

***Fig.3.** sample of the Document Term Matrix*

The document term matrix of my data contains a total of 205 rows and 19356 columns.

3. Latent Dirichlet Allocation:

LDA is a kind of statistical model that is usually used in natural language processing to assign text in a document to a specific topic or set of topics. It facilitates the identification of abstract subjects from a vast amount of textual data. A three-topic LDA model is applied to my data.

Beta Distribution:
The beta matrix represents the probability distributions of words over the 3 topics.

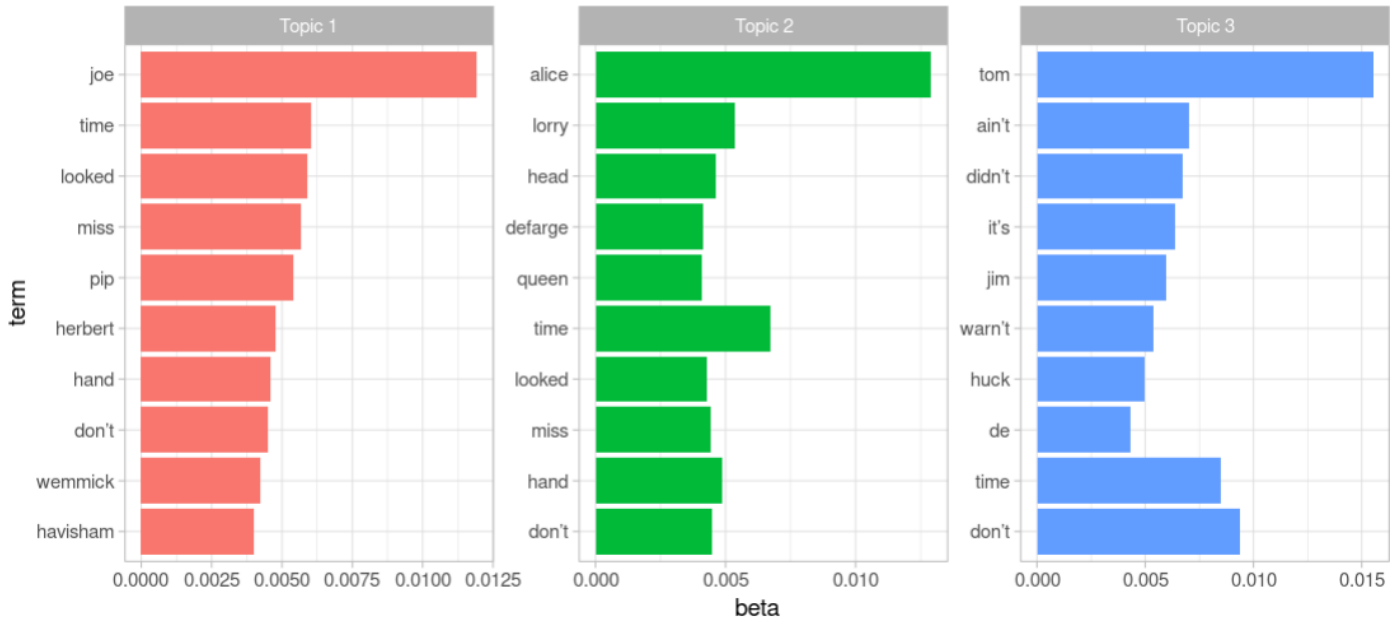

Fig.4. *Top 10 words distributed over the topics*

In the above plot (Fig.4), the visualization shows the word distribution in each topic of 3 topic LDA model. The y-axis represents the top 12 repeated terms and the x-axis represents the beta distribution. From the plot, we can see that ‘joe’, ‘alice’, and ‘tom’ are the most used words from the topic 1, 2, and 3.

- By analyzing the plot, we can notice that topic 1 seems to capture the terms associated with the different characters. The terms such as "time", "looked", "miss", "hand", "head", and "night" suggests interactions among the characters.

- Even in topic 2, words such as "Alice", "Lorry", "Defarge", and "Queen" come across as the characters, and words like "day", "eyes", "doctor", "dear", "time", "looked", "miss", "hand", "head", "night" indicates as a possible description of the characters or the narration.

- This topic is loaded with the contractions and informal terms like, "ain't", "didn't", "it's", "warn't", "that's", "couldn't", "wouldn't", possibly indicating the dialogues of "Tom", "Jim", "Huck".

Gamma distribution across all the books:

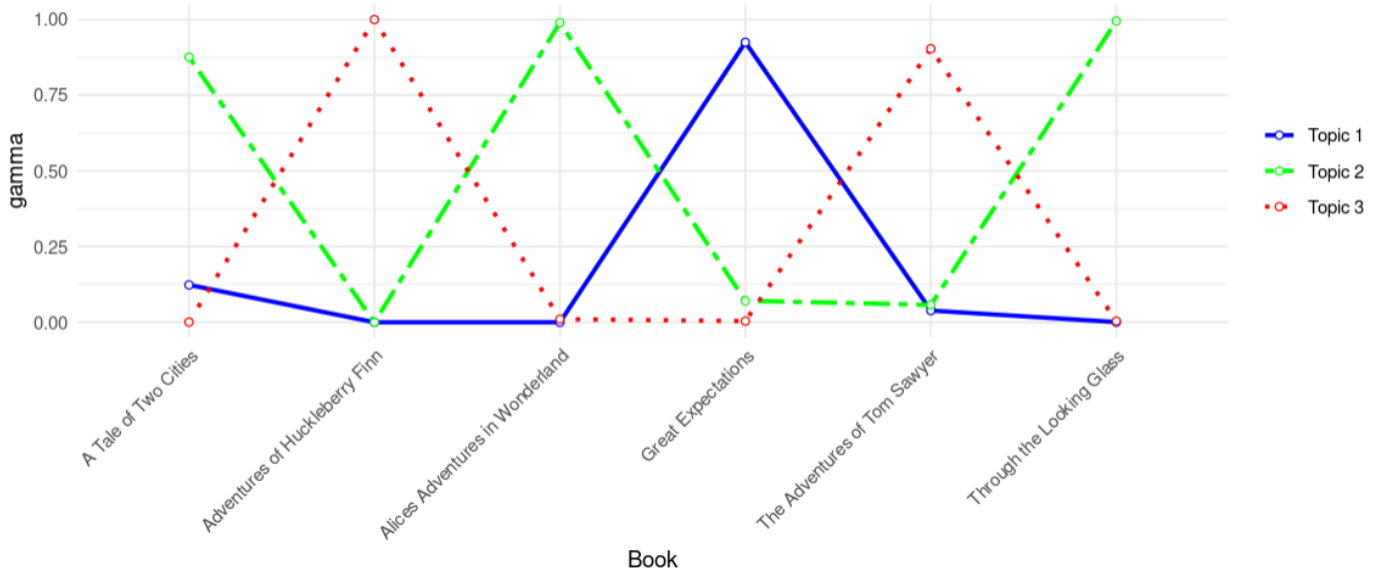

Fig.5. Gamma distribution

The plot (Fig.5) uses gamma values from an LDA model to show the distribution of three topics over a number of books. The proportion of each topic in each book is shown by the gamma values.

1. Topic 1 shows low to moderate presence across most books but has its most presence in the book "Great Expectations."

2. Topic 2 is most dominant in the books "Alice's Adventures in Wonderland," and “Through the Looking Glass”, a little lesser proportion in "A Tale of Two Cities" and moderate levels in other books.

3. Topic 3 has its highest presence in "The Adventures of Huckleberry Finn" and "The Adventures of Tom Sawyer", indicating a significant presence in these books while being minimal or moderate in others.

Gamma distribution in a single book – “A Tale of Two Cities”

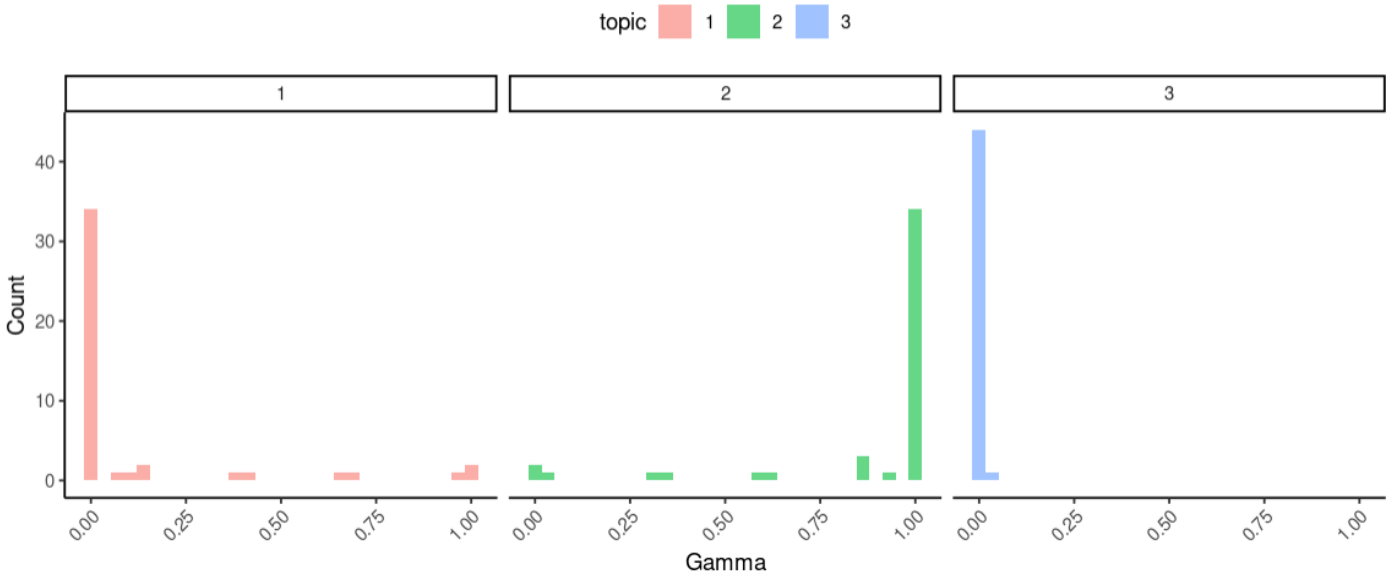

Fig.6. Gamma distribution in a single book

Topic 1: Majority of chapters have a very low gamma value, near 0, indicating that this topic is only slightly present in most chapters. The fact that the gamma value rarely approaches 1 suggests that the chapter could be fully represented by the topic.

Topic 2: The distribution for Topic 2 reveals few chapters with gamma values close to 0. On the other hand, a significant increase in the number of documents with a gamma value of 1.0 indicates that this topic could be the only one covered in most of the chapters.

Topic 3: Nearly all chapters have a gamma value of 0 or close to 0, indicating that while Topic 3 is either completely absent or present in some chapters in only a minor way.