# CLASS PROJECT 3

# TEXT ANALYTICS IN R

## CSCI 6444

## INTRODUCTION TO BIG DATA & ANALYTICS

## PROFESSOR: STEPHEN KAISLER


GROUP 12

NAME: RAJITH RAVIKUMAR

GWID: G45692898

NAME: SRAVANI BRAHAMMA ROUTHU

GWID: G28059824

## DATASET DESCRIPTION

For this project, our dataset consists of a text file containing segments extracted from the renowned literary work "Tarzan of the Apes" authored by Edgar Rice Burroughs. The selected corpus encompasses chapters I through XV of the novel. This decision was made to focus our text analytics efforts on a substantial yet manageable portion of the text, facilitating in-depth exploration and analysis within the scope of this project.

## OBJECTIVE

The objective of this project is to conduct text analytics on the extensive text file " Tarzan of the Apes." Through a series of analytical methods, we aim to delve into the text's content, discern underlying sentiments, and draw insightful observations. By employing various text mining techniques, we seek to gain a comprehensive understanding of the narrative, its thematic elements, and any implicit sentiments expressed within the text. Ultimately, this analysis endeavors to extract valuable insights and contribute to a deeper comprehension of the literary work "Tarzan of the Apes."

## LOADING THE DATASET

In the initial phase of this project, emphasis was placed on loading the dataset, "Tarzan of the Apes" into the analytical environment and structuring it into a coherent corpus suitable for text analytics. The commencement of our text analytics journey entails the installation and loading of essential libraries and packages within the R environment. These libraries and packages provide the necessary tools and functions to facilitate our analytical processes effectively.

```
#Installing the required Packages for the Project
install.packages("tm")
install.packages("plyr")
install.packages("dplyr")
install.packages("tidytext")
install.packages("ggplot2")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("SnowballC")
install.packages("stringr")

#Loading the packages as Libraries
library(tm)
library(plyr)
library(dplyr)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(SnowballC)
library(stringr)
```

Once the requisite libraries and packages are installed, we proceed by setting the working directory to ensure seamless access to files and resources. This step enhances organizational efficiency and streamlines subsequent data-loading procedures. Subsequently, we load the dataset containing the text of " Tarzan of the Apes " into our R session.

```
#Setting the Directory
setwd("C:\\Users\\Sravani\\OneDrive\\Documents\\2024_SPRING\\CS6444 - BIG DATA\\Projects\\Project_3\\eBook")
getwd()
eBook<-readLines("TarzanOfTheApes.txt")
eBook
```

Upon executing the last two lines of code in the above screenshot, the text file "TarzanOfTheApes.txt" is read into R and stored in the variable `eBook`. This variable contains the extracted content of the given book, allowing for subsequent processing and analysis within the R environment.

```
  [2] "Tarzan of the Apes"
  [3] ""
  [4] ""
  [5] "By"
  [6] ""
  [7] "Edgar Rice Burroughs"
  [8] ""
  [9] ""
 [10] ""
 [11] ""
 [12] "         CONTENTS"
 [13] ""
 [14] "      I  Out to Sea"
 [15] "     II  The Savage Home"
 [16] "    III  Life and Death"
 [17] "     IV  The Apes"
 [18] "      V  The White Ape"
 [19] "     VI  Jungle Battles"
 [20] "    VII  The Light of Knowledge"
 [21] "   VIII  The Tree-top Hunter"
 [22] "     IX  Man and Man"
 [23] "      X  The Fear-Phantom"
 [24] "     XI  \"King of the Apes\""
 [25] "    XII  Man's Reason"
 [26] "   XIII  His Own Kind"
 [27] "    XIV  At the Mercy of the Jungle"
 [28] "     XV  The Forest God"
 [29] "    XVI  \"Most Remarkable\""
 [30] "   XVII  Burials"
 [31] "  XVIII  The Jungle Toll"
 [32] "    XIX  The Call of the Primitive"
 [33] "     XX  Heredity"
 [34] "    XXI  The Village of Torture"
```

We initiate our analysis by segmenting the text into individual chapters, aiming to create a corpus where each of the fifteen chapters (Chapters I-XV) constitutes a distinct document. Employing the `which` method, we identify the indices corresponding to chapter names, such as "Chapter I" through "Chapter XV". Subsequently, we use these indices to partition the text based on chapter names, storing the content of each chapter in a separate variable. This approach enables us to systematically separate the chapters and verify the accuracy of the division process through examination of the output.

```r
#Separating the chapters into individual files
#Identifying the indices for chapters
indx_ch1 <- which(eBook == "Chapter I", arr.ind=TRUE)
indx_ch2 <- which(eBook == "Chapter II", arr.ind=TRUE)
indx_ch3 <- which(eBook == "Chapter III", arr.ind=TRUE)
indx_ch4 <- which(eBook == "Chapter IV", arr.ind=TRUE)
indx_ch5 <- which(eBook == "Chapter V", arr.ind=TRUE)
indx_ch6 <- which(eBook == "Chapter VI", arr.ind=TRUE)
indx_ch7 <- which(eBook == "Chapter VII", arr.ind=TRUE)
indx_ch8 <- which(eBook == "Chapter VIII", arr.ind=TRUE)
indx_ch9 <- which(eBook == "Chapter IX", arr.ind=TRUE)
indx_ch10 <- which(eBook == "Chapter X", arr.ind=TRUE)
indx_ch11 <- which(eBook == "Chapter XI", arr.ind=TRUE)
indx_ch12 <- which(eBook == "Chapter XII", arr.ind=TRUE)
indx_ch13 <- which(eBook == "Chapter XIII", arr.ind=TRUE)
indx_ch14 <- which(eBook == "Chapter XIV", arr.ind=TRUE)
indx_ch15 <- which(eBook == "Chapter XV", arr.ind=TRUE)
indx_ch16 <- which(eBook == "Chapter XVI", arr.ind=TRUE)
```

Following the execution of the provided lines of code, the content from each chapter of the book "Tarzan of the Apes" is individually extracted and stored into variables named 'chap1', 'chap2', and so forth up to 'chap15'. This segmentation allows for focused analysis and exploration of each chapter's content.

```r
#Extracting the text for chapters
chap1 <- eBook[(indx_ch1 + 1):(indx_ch2 - 1)]
chap2 <- eBook[(indx_ch2 + 1):(indx_ch3 - 1)]
chap3 <- eBook[(indx_ch3 + 1):(indx_ch4 - 1)]
chap4 <- eBook[(indx_ch4 + 1):(indx_ch5 - 1)]
chap5 <- eBook[(indx_ch5 + 1):(indx_ch6 - 1)]
chap6 <- eBook[(indx_ch6 + 1):(indx_ch7 - 1)]
chap7 <- eBook[(indx_ch7 + 1):(indx_ch8 - 1)]
chap8 <- eBook[(indx_ch8 + 1):(indx_ch9 - 1)]
chap9 <- eBook[(indx_ch9 + 1):(indx_ch10 - 1)]
chap10 <- eBook[(indx_ch10 + 1):(indx_ch11 - 1)]
chap11 <- eBook[(indx_ch11 + 1):(indx_ch12 - 1)]
chap12 <- eBook[(indx_ch12 + 1):(indx_ch13 - 1)]
chap13 <- eBook[(indx_ch13 + 1):(indx_ch14 - 1)]
chap14 <- eBook[(indx_ch14 + 1):(indx_ch15 - 1)]
chap15 <- eBook[(indx_ch15 + 1):(indx_ch16 - 1)]
```

Here, we tried to retrieve content from chapters 4 &15. Utilizing the stored variables 'chap4' and 'chap15', the text content from these respective chapters is accessed and made available for further analysis.

**NOTE:** Capturing the entirety of lengthy text passages in a screenshot is impractical, so displayed only a representative portion. This excerpt should effectively convey the context or essence of the complete passage being discussed.

```
#Retrieving the text from the chapters
chap4
chap15
```

Content retrieved from Chapter 4:

```
> #Retrieving the text from the chapters
> chap4
  [1] ""
  [2] "The Apes"
  [3] ""
  [4] ""
  [5] "In the forest of the table-land a mile back from the ocean old Kerchak"
  [6] "the Ape was on a rampage of rage among his people."
  [7] ""
  [8] "The younger and lighter members of his tribe scampered to the higher"
  [9] "branches of the great trees to escape his wrath; risking their lives"
 [10] "upon branches that scarce supported their weight rather than face old"
 [11] "Kerchak in one of his fits of uncontrolled anger."
 [12] ""
 [13] "The other males scattered in all directions, but not before the"
 [14] "infuriated brute had felt the vertebra of one snap between his great,"
 [15] "foaming jaws."
 [16] ""
 [17] "A luckless young female slipped from an insecure hold upon a high"
 [18] "branch and came crashing to the ground almost at Kerchak's feet."
 [19] ""
```

Content retrieved from Chapter 15:

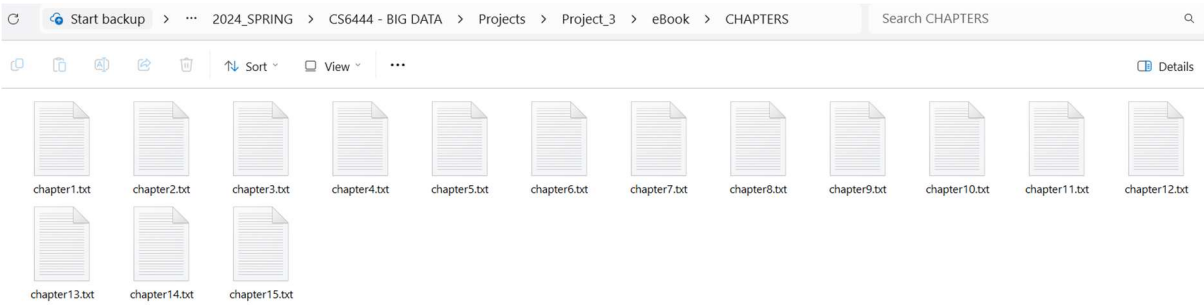
```
> chap15
  [1] ""
  [2] "The Forest God"
  [3] ""
  [4] ""
  [5] "When Clayton heard the report of the firearm he fell into an agony of"
  [6] "fear and apprehension.  He knew that one of the sailors might be the"
  [7] "author of it; but the fact that he had left the revolver with Jane,"
  [8] "together with the overwrought condition of his nerves, made him"
  [9] "morbidly positive that she was threatened with some great danger."
 [10] "Perhaps even now she was attempting to defend herself against some"
 [11] "savage man or beast."
 [12] ""
 [13] "What were the thoughts of his strange captor or guide Clayton could"
 [14] "only vaguely conjecture; but that he had heard the shot, and was in"
 [15] "some manner affected by it was quite evident, for he quickened his pace"
 [16] "so appreciably that Clayton, stumbling blindly in his wake, was down a"
 [17] "dozen times in as many minutes in a vain effort to keep pace with him,"
 [18] "and soon was left hopelessly behind."
 [19] ""
 [20] "Fearing that he would again be irretrievably lost, he called aloud to"
```

Following the segmentation of chapters, the next step involves creating a directory dedicated to storing individual chapters.

```r
#Creating Directory for chapters
dir.create("CHAPTERS")

#Writing chapters to the file
write.table(chap1, file = "chapters/chapter1.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap2, file = "chapters/chapter2.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap3, file = "chapters/chapter3.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap4, file = "chapters/chapter4.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap5, file = "chapters/chapter5.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap6, file = "chapters/chapter6.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap7, file = "chapters/chapter7.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap8, file = "chapters/chapter8.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap9, file = "chapters/chapter9.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap10, file = "chapters/chapter10.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap11, file = "chapters/chapter11.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap12, file = "chapters/chapter12.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap13, file = "chapters/chapter13.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap14, file = "chapters/chapter14.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
write.table(chap15, file = "chapters/chapter15.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
```

Subsequently, each chapter is written to a separate text file within this directory. This organized approach facilitates efficient management and accessibility of individual chapters for subsequent analysis or reference.




## CREATING A CORPUS

After successfully separating the chapters, the subsequent step involves creating a VCorpus, a crucial structure for text analysis. We utilize the VCorpus() method from the "tm" package to construct our corpus, which is then stored in the variable "TOTA" for further analysis. The structure of the corpus can be examined using the str() method, providing insights into its composition and organization.

```r
#Creating VCorpus
TOTA<-VCorpus(DirSource("./CHAPTERS",ignore.case=FALSE, mode="text"))
str(TOTA)
inspect(TOTA) #Calculating no. of characters
TOTA
```

The output provides insight into the composition of the VCorpus, which comprises 15 documents, each corresponding to a chapter of the literary work. These documents encapsulate the textual content of their respective chapters, with each containing approximately up to 586 characters. This segmentation enables focused analysis and manipulation of the textual data, facilitating in-depth exploration of themes, patterns, and linguistic characteristics within each chapter. Each chapter is further divided into two components: "content" and "meta". The "content" component contains the textual content of the chapter, while the "meta" component encompasses metadata such as author, date, description, heading, ID, language, and origin.

```
> #Creating VCorpus
> TOTA<-VCorpus(DirSource("./CHAPTERS",ignore.case=FALSE, mode="text"))
> str(TOTA)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 15
  ..$ :List of 2
  .. ..$ content: chr [1:420] "" "Out to Sea" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:215] "" "The Fear-Phantom" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:427] "" "\"King of the Apes\"" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter11.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:333] "" "Man's Reason" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter12.txt"
  .. .. ..$ language     : chr "en"
```

The output provides insight into the composition of the VCorpus, which comprises 15 documents, each corresponding to a chapter of the literary work. These documents encapsulate the textual content of their respective chapters, with each containing approximately up to 587 characters. This segmentation enables focused analysis and manipulation of the textual data, facilitating in-depth exploration of themes, patterns, and linguistic characteristics within each chapter. Each chapter is further divided into two components: "content" and "meta". The "content" component contains the textual content of the chapter, while the "meta" component encompasses metadata such as author, date, description, heading, ID, language, and origin.

```
> inspect(TOTA) #Calculating no. of characters
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 19449

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9856

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 19079

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 15186

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 25927

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18222

[[7]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8652

[[8]]
<<PlainTextDocument>>
Metadata:  7
```

```
> TOTA
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15
>
```

To retrieve text from the corpus, we can adhere to the specified guidelines outlined in the rubric.

```
#Extracting the text from the corpus
#Retrieves first document from the folder
etext<-TOTA[[1]]
etext
etext[1] #Content
#Retrieves ninth document from the folder
etext<-TOTA[[9]]
etext
etext[1] #Content
```

Here, it's evident that the first document, representing the initial chapter in the folder, contains 19,449 characters.

```
> #Extracting the text from the corpus
> #Retrieves first document from the folder
> etext<-TOTA[[1]]
> etext
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 19449
> etext[1] #Content
$content
  [1] ""
  [2] "Out to Sea"
  [3] ""
  [4] ""
  [5] "I had this story from one who had no business to tell it to me, or to"
  [6] "any other.  I may credit the seductive influence of an old vintage upon"
  [7] "the narrator for the beginning of it, and my own skeptical incredulity"
  [8] "during the days that followed for the balance of the strange tale."
  [9] ""
 [10] "When my convivial host discovered that he had told me so much, and that"
 [11] "I was prone to doubtfulness, his foolish pride assumed the task the old"
 [12] "vintage had commenced, and so he unearthed written evidence in the form"
 [13] "of musty manuscript, and dry official records of the British Colonial"
 [14] "Office to support many of the salient features of his remarkable"
 [15] "narrative."
 [16] ""
 [17] "I do not say the story is true, for I did not witness the happenings"
 [18] "which it portrays, but the fact that in the telling of it to you I have"
 [19] "taken fictitious names for the principal characters quite sufficiently"
 [20] "evidences the sincerity of my own belief that it MAY be true."
 [21] ""
 [22] "The yellow, mildewed pages of the diary of a man long dead, and the"
 [23] "records of the Colonial Office dovetail perfectly with the narrative of"
 [24] "my convivial host, and so I give you the story as I painstakingly"
 [25] "pieced it out from these several various agencies."
 [26] ""
 [27] "If you do not find it credible you will at least be as one with me in"
 [28] "acknowledging that it is unique, remarkable, and interesting."
 [29] ""
 [30] "From the records of the Colonial Office and from the dead man's diary"
 [31] "we learn that a certain young English nobleman, whom we shall call John"
 [32] "Clayton, Lord Greystoke, was commissioned to make a peculiarly delicate"
 [33] "investigation of conditions in a British West Coast African Colony from"
 [34] "whose simple native inhabitants another European power was known to be"
 [35] "recruiting soldiers for its native army, which it used solely for the"
```

Similarly, it's evident that the ninth document, corresponding to the 9th chapter, comprises 12,481 characters.

```
> #Retrieves ninth document from the folder
> etext<-TOTA[[9]]
> etext
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 12481
> etext[1] #Content
$content
  [1] ""
  [2] "Life and Death"
  [3] ""
  [4] ""
  [5] "Morning found them but little, if at all refreshed, though it was with"
  [6] "a feeling of intense relief that they saw the day dawn."
  [7] ""
  [8] "As soon as they had made their meager breakfast of salt pork, coffee"
  [9] "and biscuit, Clayton commenced work upon their house, for he realized"
 [10] "that they could hope for no safety and no peace of mind at night until"
 [11] "four strong walls effectually barred the jungle life from them."
 [12] ""
 [13] "The task was an arduous one and required the better part of a month,"
 [14] "though he built but one small room.  He constructed his cabin of small"
 [15] "logs about six inches in diameter, stopping the chinks with clay which"
 [16] "he found at the depth of a few feet beneath the surface soil."
 [17] ""
 [18] "At one end he built a fireplace of small stones from the beach.  These"
 [19] "also he set in clay and when the house had been entirely completed he"
 [20] "applied a coating of the clay to the entire outside surface to the"
 [21] "thickness of four inches."
 [22] ""
 [23] "In the window opening he set small branches about an inch in diameter"
 [24] "both vertically and horizontally, and so woven that they formed a"
 [25] "substantial grating that could withstand the strength of a powerful"
 [26] "animal.  Thus they obtained air and proper ventilation without fear of"
 [27] "lessening the safety of their cabin."
 [28] ""
 [29] "The A-shaped roof was thatched with small branches laid close together"
 [30] "and over these long jungle grass and palm fronds, with a final coating"
 [31] "of clay."
 [32] ""
 [33] "The door he built of pieces of the packing-boxes which had held their"
 [34] "belongings, nailing one piece upon another, the grain of contiguous"
 [35] "layers running transversely, until he had a solid body some three"
 [36] "inches thick and of such great strength that they were both moved to"
 [37] "laughter as they gazed upon it."
```

The subsequent phase involves the creation of initial Document-Term Matrices (DTM). In these matrices, each row signifies a document, while each column represents a term within the document. This framework facilitates data analysis through vector and matrix algebra, enabling comprehensive exploration of the textual data.

Additionally, we generate a Term-Document Matrix (TDM), where rows denote terms and columns represent documents. This alternate representation offers a complementary perspective on the textual data, enabling diverse analytical approaches. These matrices are constructed utilizing the DocumentTermMatrix() and TermDocumentMatrix() methods available in the "tm" package.

- Document Term Matrix:

```
#Creating the Document Term Matrix
TOTA_DTM<-DocumentTermMatrix(TOTA)
TOTA_DTM
inspect(TOTA_DTM)
str(TOTA_DTM)
```

```
> #Creating the Document Term Matrix
> TOTA_DTM<-DocumentTermMatrix(TOTA)
> TOTA_DTM
<<DocumentTermMatrix (documents: 15, terms: 7556)>>
Non-/sparse entries: 16261/97079
Sparsity           : 86%
Maximal term length: 22
Weighting          : term frequency (tf)
> inspect(TOTA_DTM)
<<DocumentTermMatrix (documents: 15, terms: 7556)>>
Non-/sparse entries: 16261/97079
Sparsity           : 86%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs            and but for from had his that the was with
  chapter1.txt   98  28  40   29  34  54   55 240  56   32
  chapter11.txt 125  21  36   25  42  90   23 328  43   45
  chapter12.txt  89  21  28   12  44  66   37 199  41   21
  chapter13.txt 159  20  20   27  55 105   33 367  55   39
  chapter14.txt 104  24  27   23  35  39   39 325  62   26
  chapter2.txt   96  29  32   17  27  26   34 211  32   29
  chapter4.txt   74  17  20   20  33  49   29 200  45   29
  chapter5.txt   95  23  29   12  30  53   31 161  49   24
  chapter7.txt  126  26  29   28  37  93   35 373  56   37
  chapter9.txt  139  28  26   19  54  96   32 263  62   33
> str(TOTA_DTM)
List of 6
 $ i       : int [1:16261] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:16261] 1 2 3 4 5 6 7 8 9 10 ...
 $ v       : num [1:16261] 1 1 1 1 1 1 1 1 1 1 ...
 $ nrow    : int 15
 $ ncol    : int 7556
 $ dimnames:List of 2
  ..$ Docs : chr [1:15] "chapter1.txt" "chapter10.txt" "chapter11.txt" "chapter12.txt" ...
  ..$ Terms: chr [1:7556] "'ancient" "'arf" "'e's" "'e;" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

Upon inspecting the Document-Term Matrix (DTM), we observe 15 rows, each signifying a chapter, with a total of 7556 terms. The maximum term length is 22 characters, and the Sample field offers a preview of the DTM. Notably, the DTM exhibits a high sparsity of 86%, indicating that a significant portion of the matrix comprises zero values. This phenomenon is anticipated due to the corpus's size and the likelihood that not every term appears in each document. The sparsity underscores the diverse vocabulary and distribution of terms across the corpus, highlighting the complexity of textual data analysis and the importance of handling sparsity efficiently.

- Term Document Matrix:

```
#Creating the Term Document Matrix
TOTA_TDM<-TermDocumentMatrix(TOTA)
TOTA_TDM
inspect(TOTA_TDM)
str(TOTA_TDM)
```

```
> #Creating the Term Document Matrix
> TOTA_TDM<-TermDocumentMatrix(TOTA)
> TOTA_TDM
<<TermDocumentMatrix (terms: 7556, documents: 15)>>
Non-/sparse entries: 16261/97079
Sparsity           : 86%
Maximal term length: 22
Weighting          : term frequency (tf)
> inspect(TOTA_TDM)
<<TermDocumentMatrix (terms: 7556, documents: 15)>>
Non-/sparse entries: 16261/97079
Sparsity           : 86%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
       Docs
Terms   chapter1.txt chapter11.txt chapter12.txt chapter13.txt chapter14.txt chapter2.txt chapter4.txt chapter5.txt chapter7.txt
  and            98           125            89           159           104           96           74           95          126
  but            28            21            21            20            24           29           17           23           26
  for            40            36            28            20            27           32           20           29           29
  from           29            25            12            27            23           17           20           12           28
  had            34            42            44            55            35           27           33           30           37
  his            54            90            66           105            39           26           49           53           93
  that           55            23            37            33            39           34           29           31           35
  the           240           328           199           367           325          211          200          161          373
  was            56            43            41            55            62           32           45           49           56
  with           32            45            21            39            26           29           29           24           37
       Docs
Terms   chapter9.txt
  and           139
  but            28
  for            26
  from           19
  had            54
  his            96
  that           32
  the           263
  was            62
  with           33
> str(TOTA_TDM)
List of 6
 $ i       : int [1:16261] 1 2 3 4 5 6 7 8 9 10 ...
 $ j       : int [1:16261] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:16261] 1 1 1 1 1 1 1 1 1 1 ...
 $ nrow    : int 7556
 $ ncol    : int 15
```

The creation of the Term-Document Matrix (TDM) involves the transformation of the VCorpus "TOTA" into a structured matrix format. Executing the TermDocumentMatrix() function on the corpus generates the TDM, denoted as "TOTA_TDM".

The TDM comprises 7,556 unique terms across 15 documents. This matrix structure consists of both non-zero and zero entries, with 16,261 non-zero entries and 97,079 sparse entries. The sparsity of the TDM is notably high at 86%, indicating that a significant portion of the matrix is populated by zero values. Furthermore, the weighting scheme employed in the TDM is based on term frequency (tf), emphasizing the frequency of occurrence of terms within each document.

## FINDING THE 10 LONGEST WORDS & SENTENCES IN EACH CHAPTER

Before proceeding with cleansing the corpus by removing stop words, punctuation, etc., we aim to identify the ten longest words and sentences in each chapter, as well as across all chapters. To accomplish this task, we employ the tibbles package, which facilitates the creation of data frames. However, as the corpus is currently in a different format, we first need to convert it into a tibbles data structure.

To initiate this process, we begin by installing the required packages "dplyr", "tidytext", and "ggplot2", followed by loading them as libraries in our project environment. These packages provide essential functionalities for data manipulation, text analysis, and visualization, essential for our analysis. Once the packages are installed and loaded, we can proceed with converting the corpus into a tibbles data structure and subsequently extracting the desired information regarding the longest words and sentences in each chapter and across all chapters.

After importing the necessary libraries, the next step involves converting our corpus, named "TOTA", into a tidy Corpus format. This transformation facilitates streamlined data manipulation and analysis.

```
#Converting our corpus "TOTA" into a tidy Corpus
TOTA_tidy<-tidy(TOTA)
TOTA_tidy
```

```
> #Converting our corpus "TOTA" into a tidy Corpus
> TOTA_tidy<-tidy(TOTA)
> TOTA_tidy
# A tibble: 15 x 8
   author datetimestamp       description heading id           language origin text
   <lgl>  <dttm>              <lgl>       <lgl>   <chr>        <chr>    <lgl>  <chr>
 1 NA     2024-05-04 17:03:30 NA          NA      chapter1.txt  en       NA     "\nOut to Sea\n\n\nI had this story from one who had no…
 2 NA     2024-05-04 17:03:30 NA          NA      chapter10.txt en       NA     "\nThe Fear-Phantom\n\n\nFrom a lofty perch Tarzan view…
 3 NA     2024-05-04 17:03:30 NA          NA      chapter11.txt en       NA     "\n\"King of the Apes\"\n\n\nIt was not yet dark when h…
 4 NA     2024-05-04 17:03:30 NA          NA      chapter12.txt en       NA     "\nMan's Reason\n\n\nThere was one of the tribe of Tarz…
 5 NA     2024-05-04 17:03:30 NA          NA      chapter13.txt en       NA     "\nHis Own Kind\n\n\nThe following morning, Tarzan, lam…
 6 NA     2024-05-04 17:03:30 NA          NA      chapter14.txt en       NA     "\nAt the Mercy of the Jungle\n\n\nAfter Clayton had pl…
 7 NA     2024-05-04 17:03:30 NA          NA      chapter15.txt en       NA     "\nThe Forest God\n\n\nWhen Clayton heard the report of…
 8 NA     2024-05-04 17:03:30 NA          NA      chapter2.txt  en       NA     "\nThe Savage Home\n\n\nNor did they have long to wait,…
 9 NA     2024-05-04 17:03:30 NA          NA      chapter3.txt  en       NA     "\nLife and Death\n\n\nMorning found them but little, i…
10 NA     2024-05-04 17:03:30 NA          NA      chapter4.txt  en       NA     "\nThe Apes\n\n\nIn the forest of the table-land a mile…
11 NA     2024-05-04 17:03:30 NA          NA      chapter5.txt  en       NA     "\nThe White Ape\n\n\nTenderly Kala nursed her little w…
12 NA     2024-05-04 17:03:30 NA          NA      chapter6.txt  en       NA     "\nJungle Battles\n\n\nThe wanderings of the tribe brou…
13 NA     2024-05-04 17:03:30 NA          NA      chapter7.txt  en       NA     "\nThe Light of Knowledge\n\n\nAfter what seemed an ete…
14 NA     2024-05-04 17:03:30 NA          NA      chapter8.txt  en       NA     "\nThe Tree-top Hunter\n\n\nThe morning after the Dum-D…
15 NA     2024-05-04 17:03:30 NA          NA      chapter9.txt  en       NA     "\nMan and Man\n\n\nTarzan of the Apes lived on in his …
>
```

In our tidy corpus, we observe that each document retains its metadata, as previously examined, alongside the corresponding text content. This combined information proves invaluable for conducting our analysis, particularly in identifying the ten longest words and sentences.

```
#Finding 10 longest words in all the chapters
TOTA_Words<-TOTA_tidy %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words
#Finding 10 longest sentences in all the chapters
TOTA_Sentences<-TOTA_tidy %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences
```

To initiate this analysis, we begin by computing the ten longest words and sentences across all chapters. This process involves tokenizing the words from the corpus and creating a tibble to store the results. The extracted words are then stored in a variable named TOTA_Words, facilitating subsequent analysis and visualization of the data.

```
> #Finding 10 longest words in all the chapters
> TOTA_Words<-TOTA_tidy %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> TOTA_Words
# A tibble: 43,619 x 3
   id           word            word_length
   <chr>        <chr>                 <int>
 1 chapter11.txt incomprehensible         16
 2 chapter3.txt  responsibilities         16
 3 chapter10.txt destructiveness          15
 4 chapter11.txt experimentation          15
 5 chapter13.txt personification          15
 6 chapter2.txt  notwithstanding          15
 7 chapter4.txt  notwithstanding          15
 8 chapter5.txt  resourcefulness          15
 9 chapter1.txt  responsibility           14
10 chapter1.txt  ridiculousness           14
# ℹ 43,609 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

The analysis reveals the ten longest words present across all chapters, with the longest word being "incomprehensible" containing 16 characters, and the tenth longest word, "ridiculousness" comprising 14 characters.

Similarly, by adjusting the parameters of the unnest_tokens() method, we can derive the ten longest sentences found across all chapters. This modified approach enables us to explore the composition and structure of the longest sentences present in the corpus, shedding light on the narrative style and linguistic intricacies embedded within the text.

```
> #Finding 10 longest sentences in all the chapters
> TOTA_Sentences<-TOTA_tidy %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(se
ntence)) %>% arrange(desc(sentence_length))
> TOTA_Sentences
# A tibble: 1,716 x 3
   id            sentence                                                                                                  sentence_length
   <chr>         <chr>                                                                                                               <int>
 1 chapter7.txt  "\n\nfrom this primitive function has arisen, unquestionably, all the forms\nand ceremonials of modern church and state, for through…          633
 2 chapter3.txt  "\n\nthe last entry in his diary was made the morning following her death,\nand there he recites the sad details in a matter-of-fact…          610
 3 chapter8.txt  "\n\nmany moons ago, when he had been much smaller, he had desired the skin\nof sabor, the lioness, or numa, the lion, or sheeta, th…          528
 4 chapter1.txt  "\n\nfrom the records of the colonial office and from the dead man's diary\nwe learn that a certain young english nobleman, whom we …          523
 5 chapter7.txt  "\n\nsquatting upon his haunches on the table top in the cabin his father\nhad built--his smooth, brown, naked little body bent over…          507
 6 chapter3.txt  "\n\nthe brilliant birds and the little monkeys had become accustomed to\ntheir new acquaintances, and as they had evidently never s…          492
 7 chapter12.txt "\nman's reason\n\n\nthere was one of the tribe of tarzan who questioned his authority, and\nthat was terkoz, the son of tublat, but…          468
 8 chapter9.txt  "\n\nand then lord greystoke wiped his greasy fingers upon his naked thighs\nand took up the trail of kulonga, the son of mbonga, th…          450
 9 chapter5.txt  "\n\nthe deep waters of the lake he had been taught by his wild mother to\navoid, and further, had he not seen little neeta sink ben…          447
10 chapter6.txt  "\n\nhe put the book back in the cupboard and closed the door, for he did\nnot wish anyone else to find and destroy his treasure, an…          441
# ℹ 1,706 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

The provided regular expression utilizes the period symbol (".") as a delimiter to identify sentences, ensuring correct segmentation of text into sentences. It accounts for exceptions where the period is preceded by titles like "Dr." or "Mr." to avoid erroneous splits. This robust expression facilitates accurate parsing of sentences from the text data.

Upon analysis, the longest sentence in Chapter 7 spans 633 characters, while the tenth longest sentence in Chapter 6 extends to 441 characters. These findings highlight the variability in sentence lengths across different chapters, reflecting the narrative intricacies and stylistic nuances present in the text.

Following are the code snippets to generate the 10 longest words & sentences in each chapter:

```
#Finding 10 longest words & sentences in 1st chapter
TOTA_tidy_Chap1<-tidy(TOTA[1])
TOTA_tidy_Chap1
TOTA_words_Chap1<-TOTA_tidy_Chap1 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap1
TOTA_Sentences_Chap1<-TOTA_tidy_Chap1 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap1

#Finding 10 longest words & sentences in 2nd chapter
TOTA_tidy_Chap2<-tidy(TOTA[8])
TOTA_tidy_Chap2
TOTA_words_Chap2<-TOTA_tidy_Chap2 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap2
TOTA_Sentences_Chap2<-TOTA_tidy_Chap2 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap2

#Finding 10 longest words & sentences in 3rd chapter
TOTA_tidy_Chap3<-tidy(TOTA[9])
TOTA_tidy_Chap3
TOTA_words_Chap3<-TOTA_tidy_Chap3 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap3
TOTA_Sentences_Chap3<-TOTA_tidy_Chap3 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap3

#Finding 10 longest words & sentences in 4th chapter
TOTA_tidy_Chap4<-tidy(TOTA[10])
TOTA_tidy_Chap4
TOTA_words_Chap4<-TOTA_tidy_Chap4 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap4
TOTA_Sentences_Chap4<-TOTA_tidy_Chap4 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap4

#Finding 10 longest words & sentences in 5th chapter
TOTA_tidy_Chap5<-tidy(TOTA[11])
TOTA_tidy_Chap5
TOTA_words_Chap5<-TOTA_tidy_Chap5 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap5
TOTA_Sentences_Chap5<-TOTA_tidy_Chap5 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap5

#Finding 10 longest words & sentences in 6th chapter
TOTA_tidy_Chap6<-tidy(TOTA[12])
TOTA_tidy_Chap6
TOTA_words_Chap6<-TOTA_tidy_Chap6 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap6
TOTA_Sentences_Chap6<-TOTA_tidy_Chap6 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap6

#Finding 10 longest words & sentences in 7th chapter
TOTA_tidy_Chap7<-tidy(TOTA[13])
TOTA_tidy_Chap7
TOTA_words_Chap7<-TOTA_tidy_Chap7 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap7
TOTA_Sentences_Chap7<-TOTA_tidy_Chap7 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap7

#Finding 10 longest words & sentences in 8th chapter
TOTA_tidy_Chap8<-tidy(TOTA[14])
TOTA_tidy_Chap8
TOTA_words_Chap8<-TOTA_tidy_Chap8 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap8
TOTA_Sentences_Chap8<-TOTA_tidy_Chap8 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap8

#Finding 10 longest words & sentences in 9th chapter
TOTA_tidy_Chap9<-tidy(TOTA[15])
TOTA_tidy_Chap9
TOTA_words_Chap9<-TOTA_tidy_Chap9 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_words_Chap9
TOTA_Sentences_Chap9<-TOTA_tidy_Chap9 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap9
```

```
#Finding 10 longest words & sentences in 10th chapter
TOTA_tidy_Chap10<-tidy(TOTA[2])
TOTA_tidy_Chap10
TOTA_Words_Chap10<-TOTA_tidy_Chap10 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words_Chap10
TOTA_Sentences_Chap10<-TOTA_tidy_Chap10 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap10

#Finding 10 longest words & sentences in 11th chapter
TOTA_tidy_Chap11<-tidy(TOTA[3])
TOTA_tidy_Chap11
TOTA_Words_Chap11<-TOTA_tidy_Chap11 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words_Chap11
TOTA_Sentences_Chap11<-TOTA_tidy_Chap11 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap11

#Finding 10 longest words & sentences in 12th chapter
TOTA_tidy_Chap12<-tidy(TOTA[4])
TOTA_tidy_Chap12
TOTA_Words_Chap12<-TOTA_tidy_Chap12 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words_Chap12
TOTA_Sentences_Chap12<-TOTA_tidy_Chap12 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap12

#Finding 10 longest words & sentences in 13th chapter
TOTA_tidy_Chap13<-tidy(TOTA[5])
TOTA_tidy_Chap13
TOTA_Words_Chap13<-TOTA_tidy_Chap13 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words_Chap13
TOTA_Sentences_Chap13<-TOTA_tidy_Chap13 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap13

#Finding 10 longest words & sentences in 14th chapter
TOTA_tidy_Chap14<-tidy(TOTA[6])
TOTA_tidy_Chap14
TOTA_Words_Chap14<-TOTA_tidy_Chap14 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words_Chap14
TOTA_Sentences_Chap14<-TOTA_tidy_Chap14 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap14

#Finding 10 longest words & sentences in 15th chapter
TOTA_tidy_Chap15<-tidy(TOTA[7])
TOTA_tidy_Chap15
TOTA_Words_Chap15<-TOTA_tidy_Chap15 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
TOTA_Words_Chap15
TOTA_Sentences_Chap15<-TOTA_tidy_Chap15 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
TOTA_Sentences_Chap15
```

The generated output below identifies the ten longest words and sentences present in the 1st chapter:

```
> #Finding 10 longest words & sentences in 1st chapter
> TOTA_tidy_Chap1<-tidy(TOTA[1])
> TOTA_tidy_Chap1
# A tibble: 1 × 8
  author datetimestamp       description heading id          language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>       <chr>    <lgl>  <chr>
1 NA     2024-05-04 17:03:30 NA          NA      chapter1.txt en       NA     "\nOut to Sea\n\n\nI had this story from one who had no business to tell it to me, or t…
> TOTA_Words_Chap1<-TOTA_tidy_Chap1 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> TOTA_Words_Chap1
# A tibble: 3,564 × 3
   id           word           word_length
   <chr>        <chr>                <int>
 1 chapter1.txt responsibility          14
 2 chapter1.txt ridiculousness          14
 3 chapter1.txt simultaneously          14
 4 chapter1.txt painstakingly           13
 5 chapter1.txt acknowledging           13
 6 chapter1.txt investigation           13
 7 chapter1.txt investigation           13
 8 chapter1.txt investigation           13
 9 chapter1.txt circumstances           13
10 chapter1.txt mortification           13
# ℹ 3,554 more rows
# ℹ Use `print(n = ...)` to see more rows
> TOTA_Sentences_Chap1<-TOTA_tidy_Chap1 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_len
gth=nchar(sentence)) %>% arrange(desc(sentence_length))
> TOTA_Sentences_Chap1
# A tibble: 139 × 3
   id           sentence                                                                                                                  sentence_length
   <chr>        <chr>                                                                                                                               <int>
 1 chapter1.txt "\n\nfrom the records of the colonial office and from the dead man's diary\nwe learn that a certain young english nobleman, whom we s…             523
 2 chapter1.txt "  without\nwaiting to rise he whipped a revolver from his pocket, firing point\nblank at the great mountain of muscle towering befor…             385
 3 chapter1.txt "\n\nwhen my convivial host discovered that he had told me so much, and that\ni was prone to doubtfulness, his foolish pride assumed …             362
 4 chapter1.txt "  what reason could he give the officer commanding\nher majesty's ship for desiring to go back in the direction from which\nhe had j…             357
 5 chapter1.txt "\n\nclayton asked no questions--he did not need to--and the following day,\nas the great lines of a british battleship grew out of t…             340
 6 chapter1.txt "  late in the afternoon he saw her upper works fade\nbelow the far horizon, but not before he learned that which confirmed\nhis grea…             337
 7 chapter1.txt "  the captain has brought this\ncondition upon himself, so why then should i risk subjecting my wife to\nunthinkable horrors in a pr…             302
 8 chapter1.txt "\n\non the second day after the wounding of black michael, clayton came on\ndeck just in time to see the limp body of one of the cre…             283
 9 chapter1.txt "\n\nso, whereas the captain might easily have been brought to regret his\nhasty speech had clayton attempted to conciliate him, his …             278
10 chapter1.txt "\n\ni do not say the story is true, for i did not witness the happenings\nwhich it portrays, but the fact that in the telling of it …             274
```

The below output shows the ten longest words and sentences present in the 6th chapter:

```
> #Finding 10 longest words & sentences in 6th chapter
> TOTA_tidy_Chap6<-tidy(TOTA[12])
> TOTA_tidy_Chap6
# A tibble: 1 × 8
  author datetimestamp       description heading id          language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>       <chr>    <lgl>  <chr>
1 NA     2024-05-04 17:03:30 NA          NA      chapter6.txt en       NA     "\nJungle Battles\n\n\nThe wanderings of the tribe brought them often near the closed a…
> TOTA_Words_Chap6<-TOTA_tidy_Chap6 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> TOTA_Words_Chap6
# A tibble: 2,341 × 3
   id           word           word_length
   <chr>        <chr>                <int>
 1 chapter6.txt impossibility           13
 2 chapter6.txt thousandfold            12
 3 chapter6.txt explorations            12
 4 chapter6.txt introduction            12
 5 chapter6.txt exhilaration            12
 6 chapter6.txt accidentally            12
 7 chapter6.txt excruciating            12
 8 chapter6.txt imagination             11
 9 chapter6.txt independent             11
10 chapter6.txt combination             11
# ℹ 2,331 more rows
# ℹ Use `print(n = ...)` to see more rows
> TOTA_Sentences_Chap6<-TOTA_tidy_Chap6 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_len
gth=nchar(sentence)) %>% arrange(desc(sentence_length))
> TOTA_Sentences_Chap6
# A tibble: 79 × 3
   id           sentence                                                                                                                  sentence_length
   <chr>        <chr>                                                                                                                               <int>
 1 chapter6.txt "\n\nhe put the book back in the cupboard and closed the door, for he did\nnot wish anyone else to find and destroy his treasure, and…             441
 2 chapter6.txt "\n\na mile back in the forest the tribe had heard the fierce challenge of\nthe gorilla, and, as was his custom when any danger threa…             348
 3 chapter6.txt "\n\nit was in the next visit to the vicinity, following the adventure with\nold sabor, that, as he approached the cabin, tarzan noti…             345
 4 chapter6.txt "\n\nthat her little tarzan could destroy a great bull gorilla she knew to\nbe improbable, and so, as she neared the spot from which …             344
 5 chapter6.txt "\n\nlike some huge phantom, kala swung noiselessly from tree to tree; now\nrunning nimbly along a great branch, now swinging through…             333
 6 chapter6.txt "\n\nthe boats, and trains, and cows and horses were quite meaningless to\nhim, but not quite so baffling as the odd little figures w…             316
 7 chapter6.txt "\n\nhe was alone, as was often the case when he visited the cabin, for the\napes had no love for it; the story of the thunder-stick …             297
 8 chapter6.txt "\nthe language of the apes had so few words that they could talk but\nlittle of what they had seen in the cabin, having no words to\…             297
 9 chapter6.txt "\n\nkala, alone, he was glad to have with him, but now that he was better\nshe was gone longer at a time, in search of food; for the…             292
10 chapter6.txt "  more and more weakly the torn and bleeding arm struck home\nwith the long sharp blade, then the little figure stiffened with a\nsp…             269
```

The below output shows the ten longest words and sentences obtained in the 13th chapter:

```
> #Finding 10 longest words & sentences in 13th chapter
> TOTA_tidy_Chap13<-tidy(TOTA[5])
> TOTA_tidy_Chap13
# A tibble: 1 x 8
  author datetimestamp       description heading id            language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>         <chr>    <lgl>  <chr>
1 NA     2024-05-04 17:03:30 NA          NA      chapter13.txt en       NA     "\nHis Own Kind\n\n\nThe following morning, Tarzan, lame and sore from the wounds of h…
> TOTA_Words_Chap13<-TOTA_tidy_Chap13 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> TOTA_Words_Chap13
# A tibble: 4,731 x 3
   id            word            word_length
   <chr>         <chr>                 <int>
 1 chapter13.txt personification          15
 2 chapter13.txt distinguishing           14
 3 chapter13.txt simultaneously           14
 4 chapter13.txt convalescence            13
 5 chapter13.txt reverberation            13
 6 chapter13.txt gesticulating            13
 7 chapter13.txt gesticulating            13
 8 chapter13.txt comparatively            13
 9 chapter13.txt eventualities            13
10 chapter13.txt intelligence             12
# ℹ 4,721 more rows
# ℹ Use `print(n = ...)` to see more rows
> TOTA_Sentences_Chap13<-TOTA_tidy_Chap13 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_l
ength=nchar(sentence)) %>% arrange(desc(sentence_length))
> TOTA_Sentences_Chap13
# A tibble: 209 x 3
   id            sentence                                                                                                                  sentence_length
   <chr>         <chr>                                                                                                                               <int>
 1 chapter13.txt "  adjusting his spectacles\nhe looked for a moment at the placard and then, turning away, strolled\noff muttering to himself:  \"mo…                363
 2 chapter13.txt "  we are absolutely in your power, but\nyou'll treat professor porter and miss porter with respect or i'll\nbreak that vile neck of…              344
 3 chapter13.txt "\n\ntarzan had never seen the effects of a firearm before, though his books\nhad taught him something of them, but when he saw the …              341
 4 chapter13.txt "\"\n\n\"tut, tut, child; tut, tut,\" responded professor porter, in a kindly\nand indulgent tone, \"do not trouble your pretty head…              337
 5 chapter13.txt "\n\nfor a moment longer tarzan watched while a number of boxes and bundles\nwere lowered into the waiting boats, then, as they shov…              331
 6 chapter13.txt "\n\nthen he determined to filch what few garments he could from one of the\nblack men of mbonga's village, for tarzan of the apes h…              301
 7 chapter13.txt "\n\nit took a dozen attempts on the part of the thoroughly frightened\nblacks to reenter their village, past the horrible, grinning…              299
 8 chapter13.txt "\n\nas long as they supplied him with arrows and food he would not harm\nthem unless they looked upon him, so it was ordered by mbo…              279
 9 chapter13.txt "  with this idea in mind she turned to search for some means of\nputting it into execution; but her first view of the interior of t…              268
10 chapter13.txt "\n\n\"ah, yes, indeed; yes indeed--most remarkable, most remarkable!\" said\nthe professor; \"but i can add nothing further to what…              264
```

The results are compiled into a table, showcasing the lengths of the longest words and sentences for reference. Screenshots depicting these findings have been provided on the following pages, offering a visual representation of the longest words and sentences identified across the chapters. These screenshots serve as a valuable reference for further analysis and interpretation of the textual data.

## 10 LONGEST WORDS IN EACH CHAPTER

| | Chapter 1 | Chapter 2 | Chapter 3 | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 | Chapter 9 | Chapter 10 | Chapter 11 | Chapter 12 | Chapter 13 | Chapter 14 | Chapter 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | responsibility | notwithstanding | responsibilities | notwithstanding | resourcefulness | impossibility | investigations | uncomfortable | consternation | destructiveness | incomprehensible | accomplishment | personification | comprehensible | irretrievably |
| 2 | ridiculousness | transportation | simultaneously | investigations | uncomfortable | thousandfold | establishments | possibilities | noiselessness | sentimentalist | experimentation | manifestation | distinguishing | simultaneously | apprehension |
| 3 | simultaneously | remorselessly | responsibility | understanding | experimenting | explorations | unquestionably | insignificant | transgressing | gesticulating | capriciousness | disobediences | simultaneously | insignificance | satisfaction |
| 4 | painstakingly | heartlessness | comparatively | precipitately | dissatisfied | introduction | convalescence | occasionally | bloodthirsty | superstitious | investigation | consternation | convalescence | bloodcurdling | occasionally |
| 5 | acknowledging | threateningly | uncomfortable | uncontrolled | recollection | exhilaration | investigation | considerable | convulsively | superstitious | investigation | superstitious | reverberation | complimenting | disappearing |
| 6 | investigation | investigation | acquaintances | successfully | aristocratic | accidentally | comparatively | overhanging | impenetrable | bloodthirsty | ornamentation | disappearance | gesticulating | disparagement | overwrought |
| 7 | investigation | contemplation | consciousness | intelligence | appraisement | excruciating | possibilities | superiority | appropriated | businesslike | determination | instinctively | gesticulating | instinctively | appreciably |
| 8 | investigation | circumstances | horizontally | intelligence | satisfaction | imagination | experimenting | conflicting | investigated | impenetrable | determination | dismemberment | comparatively | perspiration | neighboring |
| 9 | circumstances | forgetfulness | transversely | occasionally | unsuspecting | independent | inexhaustible | surrounding | appropriated | occasionally | loathsomeness | companionship | eventualities | disheartened | comparative |
| 10 | mortification | companionway | sufficiently | fearlessness | unbelievable | combination | remorselessly | destruction | laboriously | depredations | reconnoitered | circumstances | intelligence | unaccustomed | maintenance |

## 10 LONGEST SENTENCES IN EACH CHAPTER

| Chapter 1 | Chapter 2 | Chapter 3 | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 | Chapter 8 | Chapter 9 | Chapter 10 | Chapter 11 | Chapter 12 | Chapter 13 | Chapter 14 | Chapter 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| "\n\nfrom the records of the colonial office and from the... | "\n\nnotwithstanding that both clayton and his wife kept ... | "\n\nthe last entry in his diary was made the morning fol... | kerchak came to a halt a short distance from th... | waters of the lake he had been taught by hi... | the book back in the cupboard and closed the ... | primitive function has arisen, unquestiona... | moons ago, when he had been much smaller, he ha... | lord greystoke wiped his greasy fingers upo... | the still warm body of kulonga--on the... | "\n\nwhen he had followed kulonga through the forest he ... | "\nman's reason\n\n\nthere was one of the tribe of tarzan... | " adjusting his spectacles\nh e looked for a moment at t... | moment the lioness remained quiet with | "\n\nwhat were the thoughts of his strange captor or gui... |
| without\nwaiting to rise he whipped a revolver from | "\n\nto this end clayton selected four trees which formed... | "\n\nthe brilliant birds and the little monkeys had becom... | "\n\nhigh up among the branches of a mighty tree she hugg... | "\n\nso intent was he upon this personal appraisement of ... | "\n\na mile back in the forest the tribe had heard the fi... | "\n\nsquatting upon his haunches on the table top in the ... | "\n\nfor several months the tribe hovered near the beach ... | "\n\nthat they more greatly feared an attack from the rea... | "\n\nthe noted the extreme\ncare which the woman took that ... | " he was too shrewd to allow those long hairy arms to e... | "\n\nbut when, finally, he realized that his antagonist ... | " we are absolutely in your power, but\nyou'll treat pr... | "\n\nthe englishman, finally concluding that he was a | "\n\nhe was trying to tell the stupid white man to plung... |
| "\n\nwhen my convivial host discovered that he had told m... | "\n\n"what they did may we not do? and even better, for... | " grown carelessness\n\nfrom months of continued safety, durin... | "\n\nwhen the king ape released the limp form which had b... | ""\n\nbut when kerchak spoke to her about it kala threat... | "\n\nit was in the next visit to the vicinity, following ... | "\n\nde did not accomplish it in a day, or in a week, or ... | "\n\nthus far his plan had worked to perfection, but when... | "\n\nalso could he write, with printed letters, rapidly a... | " he killed for food most\nnoften, but, being a man, he ... | "\n\nthe ape was staggered, and what with the mortal wou... | " but when one day a\nmember of the tribe discovered th... | "\n\n\n

```r
#Data cleaning - Removing punctuations
TOTA<-tm_map(TOTA, content_transformer(gsub), pattern="'", replacement="")
removeNumPunct<-function(x) gsub("[^[:alpha:][:space:]]*","",x)
TOTA_cl<-tm::tm_map(TOTA, content_transformer(removeNumPunct))
TOTA_cl
str(TOTA_cl)
inspect(TOTA_cl)
```

```
> #Data cleaning - Removing punctuations
> TOTA<-tm_map(TOTA, content_transformer(gsub), pattern="'", replacement="")
> removeNumPunct<-function(x) gsub("[^[:alpha:][:space:]]*","",x)
> TOTA_cl<-tm::tm_map(TOTA, content_transformer(removeNumPunct))
> TOTA_cl
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15
> str(TOTA_cl)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 15
  ..$ :List of 2
  .. ..$ content: chr [1:420] "" "Out to Sea" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:215] "" "The FearPhantom" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:427] "" "King of the Apes" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
```

```
> inspect(TOTA_cl)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18862

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9650

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18652

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14802

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 25186

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 17761

[[7]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8427

[[8]]
```

To further ensure uniformity and consistency, we can confirm that our clean corpus is entirely in lowercase by passing it through the tm_map() function. This step guarantees that all text within the corpus is standardized to lowercase format, facilitating seamless text processing and analysis.

```
#Ensuring everything is in lowercase
TOTA_Low<-tm_map(TOTA_cl, tm::content_transformer(tolower))
TOTA_Low
str(TOTA_Low)
inspect(TOTA_Low)
```

```
> #Ensuring everything is in lowercase
> TOTA_Low<-tm_map(TOTA_cl, tm::content_transformer(tolower))
> TOTA_Low
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15
> str(TOTA_Low)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 15
  ..$ :List of 2
  .. ..$ content: chr [1:420] "" "out to sea" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:215] "" "the fearphantom" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chapter10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:427] "" "king of the apes" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 17:03:30"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
```

```
> inspect(TOTA_Low)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18862

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9650

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18652

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14802

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 25186

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 17761

[[7]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8427
```

With the removal of punctuation, numbers, and quotes, and the standardization of text to lowercase format, our corpus now contains uniform, cleaned data. We maintain the same number of characters across the documents, ensuring consistency in the text representation.

Adhering to the rubric, we now move forward to compute the Document Term Matrix (DTM) of the cleaned corpus. This process involves the transformation of the cleaned textual data into a structured matrix format, where each row represents a document and each column represents a unique term present in the corpus. This DTM serves as a foundational structure for subsequent text mining and analysis tasks.

```
#Computing Document Term Matrix for the clean corpus
TOTA_DTM<-DocumentTermMatrix(TOTA_Low)
TOTA_DTM
str(TOTA_DTM)
inspect(TOTA_DTM)
```

```
> #Computing Document Term Matrix for the clean corpus
> TOTA_DTM<-DocumentTermMatrix(TOTA_Low)
> TOTA_DTM
<<DocumentTermMatrix (documents: 15, terms: 5424)>>
Non-/sparse entries: 14067/67293
Sparsity           : 83%
Maximal term length: 20
Weighting          : term frequency (tf)
> str(TOTA_DTM)
List of 6
 $ i       : int [1:14067] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:14067] 7 8 10 11 25 26 36 37 44 45 ...
 $ v       : num [1:14067] 1 2 3 2 1 1 1 1 1 1 ...
 $ nrow    : int 15
 $ ncol    : int 5424
 $ dimnames:List of 2
  ..$ Docs : chr [1:15] "chapter1.txt" "chapter10.txt" "chapter11.txt" "chapter12.txt" ...
  ..$ Terms: chr [1:5424] "abandon" "abandoned" "abandoning" "abashed" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
> inspect(TOTA_DTM)
<<DocumentTermMatrix (documents: 15, terms: 5424)>>
Non-/sparse entries: 14067/67293
Sparsity           : 83%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs            and but for from had his that the was with
  chapter1.txt  104  33  40   29  34  54   56 240  57   32
  chapter11.txt 130  21  36   25  42  90   24 328  43   45
  chapter12.txt  91  21  29   12  45  67   39 199  42   21
  chapter13.txt 163  25  21   27  55 105   34 367  55   39
  chapter14.txt 107  26  28   23  35  39   40 325  62   26
  chapter2.txt  104  31  32   17  27  26   36 213  32   30
  chapter4.txt   75  17  20   20  33  49   29 200  45   29
  chapter5.txt   96  23  29   12  30  54   31 161  50   24
  chapter7.txt  128  27  30   28  37  93   36 373  56   37
  chapter9.txt  140  31  26   19  54  97   32 263  62   33
```

Indeed, although our sparsity has decreased by 3% to 83%, it remains relatively high. Sparsity refers to the proportion of zeros present in our Document-Term Matrix (DTM). This high sparsity is indicative of the substantial number of terms that do not appear in every document within our corpus.

To visualize the extent of sparsity in our corpus, we can examine our DTM using the as.matrix(TOTA_DTM) method. This visualization allows us to observe the abundance of zeros present in the matrix, confirming the 83% sparsity. The prevalence of 0's underscores the diverse vocabulary and distribution of terms across the corpus, highlighting the need for techniques to further reduce sparsity and enhance the efficiency of subsequent text analysis tasks.

To visualize the sparsity of the Document-Term Matrix (DTM), we can convert it into a regular matrix format using the as.matrix() function in R.

```
#Converting Document Term Matrix to a Regular Matrix to visualize the sparsity
TOTA_Matrix<-as.matrix(TOTA_DTM)
TOTA_Matrix
```

Running the above code will convert the Document-Term Matrix "TOTA_DTM" into a regular matrix format named "TOTA_Matrix". You can then visualize this matrix to observe the sparsity, which represents the abundance of zeros in the matrix.
Subsequently, we proceed to eliminate stop words from our Document-Term Matrix (DTM).

```
> #Removing stop words
> myStopWords<-c(tm::stopwords("english"))
> myStopWords
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"       "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"       "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"      "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"     "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"
 [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
[101] "who's"      "what's"     "here's"     "there's"    "when's"     "where's"    "why's"      "how's"      "a"          "an"
[111] "the"        "and"        "but"        "if"         "or"         "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"      "against"    "between"    "into"       "through"    "during"
[131] "before"     "after"      "above"      "below"      "to"         "from"       "up"         "down"       "in"         "out"
[141] "on"         "off"        "over"       "under"      "again"      "further"    "then"       "once"       "here"       "there"
[151] "when"       "where"      "why"        "how"        "all"        "any"        "both"       "each"       "few"        "more"
[161] "most"       "other"      "some"       "such"       "no"         "nor"        "not"        "only"       "own"        "same"
[171] "so"         "than"       "too"        "very"
> TOTA_Stop<-tm::tm_map(TOTA_Low,tm::removeWords,myStopWords)
> tm::inspect(TOTA_Stop[[1]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13612


  sea


   story  one    business  tell
    may credit  seductive influence   old vintage upon
 narrator   beginning      skeptical incredulity
  days  followed   balance   strange tale

  convivial host discovered    told   much
  prone  doubtfulness  foolish pride assumed  task  old
vintage  commenced    unearthed written evidence   form
 musty manuscript  dry official records   british colonial
office  support many   salient features   remarkable
narrative

   say  story  true     witness  happenings
  portrays   fact    telling
taken fictitious names   principal characters quite sufficiently
evidences  sincerity    belief   may  true

 yellow mildewed pages   diary   man long dead
records   colonial office dovetail perfectly   narrative
```

In this step, we have utilized the tm package to remove numerous stop words from our corpus. We can examine the first document to ascertain the impact of this removal process on our corpus. As observed, the elimination of stop words has enhanced the analyzability of our corpus.

With the stop words from the corpus "TOTA_Stop" successfully removed, we proceed to recreate the Term Document Matrix named "TOTA_Stop_TDM". This process ensures that the resulting TDM reflects the frequency of meaningful terms in the corpus, excluding common stop words that do not contribute significantly to the analysis.

```
> #Computing Term Document Matrix after removing stop words
> TOTA_Stop_TDM<-tm::TermDocumentMatrix(TOTA_Stop)
> TOTA_Stop_TDM
<<TermDocumentMatrix (terms: 5326, documents: 15)>>
Non-/sparse entries: 12944/66946
Sparsity           : 84%
Maximal term length: 20
Weighting          : term frequency (tf)
```

Next, we find frequent terms within the TDM. The findFreqTerms function is utilized to identify terms that occur with a specified frequency threshold, defined by the "lowfreq" parameter. The frequency thresholds are set at 5, 8, and 11 in different instances of the code. By extracting frequent terms based on varying frequency thresholds, analysts can gain insights into the distribution and prevalence of significant terms within the corpus.

- **<u>Terms with frequency of 5</u>**

```
> #Finding frequent terms
> #Frequency: 5
> freqTerms<-tm::findFreqTerms(TOTA_Stop_TDM, lowfreq = 5)
> freqTerms
  [1] "able"         "accompanied"  "accord"       "accustomed"   "across"       "act"          "added"        "advantage"    "adventure"
 [10] "africa"       "african"      "afternoon"    "agility"      "air"          "alarm"        "alice"        "alive"        "almost"
 [19] "alone"        "along"        "aloud"        "already"      "also"         "always"       "among"        "amphitheater" "ancient"
 [28] "anger"        "animal"       "animals"      "another"      "answer"       "answered"     "antagonist"   "anthropoids"  "ape"
 [37] "apeman"       "apes"         "apparent"     "appeared"     "approached"   "approaching"  "archer"       "arm"          "armed"
 [46] "arms"         "around"       "arrow"        "arrows"       "asked"        "attack"       "attempt"      "attempted"    "attempting"
 [55] "attention"    "aught"        "authority"    "away"         "awful"        "babe"         "baby"         "back"         "bad"
 [64] "balance"      "band"         "bared"        "battle"       "beach"        "beast"        "beasts"       "beat"         "beating"
 [73] "beautiful"    "became"       "become"       "bed"          "began"        "behind"       "belly"        "belongings"   "belt"
 [82] "beneath"      "bent"         "beside"       "best"         "better"       "beyond"       "birds"        "black"        "blackness"
 [91] "blacks"       "blade"        "bloody"       "blow"         "blows"        "boar"         "boat"         "boats"        "bodies"
[100] "body"         "bones"        "book"         "books"        "bore"         "born"         "bosom"        "bow"          "box"
[109] "boxes"        "boy"          "brain"        "branch"       "branches"     "breast"       "bright"       "brilliant"    "bring"
[118] "british"      "broke"        "broken"       "brothers"     "brought"      "brown"        "brute"        "brutes"       "bugs"
[127] "building"     "built"        "bull"         "bullet"       "cabin"        "call"         "called"       "came"         "can"
[136] "captain"      "carcass"      "care"         "carried"      "carry"        "cast"         "caught"       "cauldron"     "caused"
[145] "cautiously"   "ceased"       "center"       "chain"        "challenge"    "chance"       "change"       "charge"       "charged"
[154] "chest"        "child"        "children"     "circle"       "clawing"      "clay"         "clayton"      "claytons"     "clear"
[163] "clearing"     "close"        "closed"       "closely"      "closer"       "clung"        "coast"        "coat"         "cold"
[172] "colonial"     "come"         "comes"        "commenced"    "common"       "companions"   "conception"   "confronted"   "constant"
[181] "contents"     "continued"    "convinced"    "cooking"      "covered"      "cradle"       "crashing"     "creature"     "creatures"
[190] "crept"        "crew"         "cried"        "cries"        "crouched"     "crouching"    "cruel"        "cry"          "cupboard"
[199] "cut"          "cutthroats"   "daily"        "danger"       "dangers"      "dare"         "dark"         "darkness"     "day"
[208] "days"         "dead"         "deadly"       "dear"         "death"        "decided"      "deck"         "deep"         "denizens"
[217] "dense"        "depths"       "desire"       "diary"        "died"         "different"    "direction"    "directly"     "disappeared"
[226] "discovered"   "distance"     "distant"      "done"         "dont"         "door"         "doorway"      "dozen"        "dragged"
[235] "drawn"        "drew"         "dried"        "dropped"      "drum"         "drums"        "dry"          "dumdum"       "duties"
[244] "duty"         "ear"          "early"        "ears"         "earth"        "easily"       "easy"         "eat"          "edge"
[253] "effort"       "efforts"      "either"       "elephant"     "else"         "emerged"      "end"          "enemies"      "enemy"
[262] "english"      "englishman"   "enough"       "entered"      "entire"       "entirely"     "entrance"     "ere"          "escape"
[271] "esmeralda"    "even"         "ever"         "every"        "evidently"    "evil"         "examined"     "except"       "excited"
[280] "extreme"      "eye"          "eyes"         "face"         "faced"        "facing"       "fact"         "faint"        "fair"
[289] "fairly"       "fallen"       "falling"      "fangs"        "far"          "farther"      "fastened"     "father"       "fathers"
[298] "fear"         "feared"       "fearful"      "fears"        "fearsome"     "feast"        "features"     "feel"         "feeling"
[307] "feet"         "fell"         "fellow"       "fellows"      "felt"         "female"       "females"      "ferocious"    "ferocity"
[316] "fields"       "fierce"       "fifty"        "fighting"     "figure"       "figures"      "filled"       "final"        "finally"
[325] "find"         "finding"      "fine"         "fingers"      "finished"     "fire"         "firearms"     "first"        "fists"
[334] "fitted"       "fleeing"      "flesh"        "floor"        "foliage"      "follow"       "followed"     "following"    "food"
[343] "foot"         "force"        "forced"       "forehead"     "forest"       "forever"      "form"         "formed"       "former"
[352] "forth"        "forward"      "found"        "four"         "fresh"        "friends"      "fright"       "frightened"   "frightful"
[361] "full"         "fully"        "fuwalda"      "gain"         "game"         "gathered"     "gathering"    "gave"         "get"
[370] "giant"        "girl"         "give"         "given"        "gleaming"     "god"          "gone"         "good"         "gorilla"
[379] "grasp"        "grasped"      "grass"        "grasses"      "great"        "greater"      "greatest"     "greatly"      "grew"
[388] "greystoke"    "ground"       "growls"       "grown"        "guard"        "hair"         "hairless"     "hairy"        "half"
[397] "hand"         "hands"        "handsome"     "hanging"      "harbor"       "hard"         "harm"         "hastened"     "hated"
[406] "hatred"       "head"         "heads"        "hear"         "heard"        "heart"        "heavy"        "held"         "hide"
```

- **Terms with frequency of 8**

```
> #Frequency: 8
> freqTerms<-tm::findFreqTerms(TOTA_Stop_TDM, lowfreq = 8)
> freqTerms
  [1] "able"        "across"      "added"       "agility"     "air"         "alice"       "almost"      "alone"       "along"       "also"
 [11] "always"      "among"       "anger"       "animal"      "animals"     "another"     "anthropoids" "ape"         "apeman"      "apes"
 [21] "approached"  "approaching" "arm"         "armed"       "arms"        "around"      "arrow"       "arrows"      "attack"      "attempt"
 [31] "attention"   "away"        "awful"       "back"        "battle"      "beach"       "beast"       "beasts"      "beat"        "beautiful"
 [41] "became"      "behind"      "beneath"     "beside"      "best"        "better"      "beyond"      "black"       "blacks"      "blow"
 [51] "boar"        "boats"       "body"        "book"        "books"       "bow"         "box"         "boy"         "branch"      "branches"
 [61] "breast"      "bring"       "british"     "broken"      "brought"     "brown"       "brute"       "brutes"      "bugs"        "built"
 [71] "bull"        "bullet"      "cabin"       "called"      "came"        "can"         "captain"     "carried"     "carry"       "caught"
 [81] "cauldron"    "caused"      "cautiously"  "ceased"      "center"      "challenge"   "chance"      "chest"       "child"       "circle"
 [91] "clayton"     "claytons"    "clearing"    "close"       "closed"      "closely"     "closer"      "come"        "commenced"   "continued"
[101] "cooking"     "covered"     "creature"    "crew"        "cried"       "crouched"    "cruel"       "cry"         "danger"      "darkness"
[111] "day"         "days"        "dead"        "deadly"      "dear"        "death"       "decided"     "deck"        "deep"        "dense"
[121] "depths"      "desire"      "different"   "direction"   "directly"    "disappeared" "discovered"  "distance"    "done"        "door"
[131] "dozen"       "drew"        "dropped"     "ear"         "early"       "earth"       "easily"      "eat"         "edge"        "effort"
[141] "efforts"     "either"      "elephant"    "else"        "end"         "enemies"     "enemy"       "english"     "englishman"  "enough"
[151] "entered"     "entirely"    "ere"         "escape"      "esmeralda"   "even"        "ever"        "every"       "evidently"   "except"
[161] "eyes"        "face"        "faced"       "fact"        "faint"       "fair"        "fairly"      "fallen"      "fangs"       "far"
[171] "farther"     "father"      "fear"        "feared"      "fearful"     "feet"        "fell"        "fellow"      "fellows"     "felt"
[181] "females"     "fields"      "fierce"      "fifty"       "figure"      "filled"      "final"       "finally"     "find"        "first"
[191] "flesh"       "floor"       "foliage"     "follow"      "followed"    "following"   "food"        "foot"        "force"       "forehead"
[201] "forest"      "form"        "forward"     "found"       "four"        "friends"     "frightened"  "full"        "fully"       "fuwalda"
[211] "gathered"    "gathering"   "gave"        "giant"       "girl"        "give"        "god"         "gone"        "good"        "gorilla"
[221] "grasp"       "grasped"     "great"       "greater"     "greatest"    "greystoke"   "ground"      "hair"        "hairy"       "half"
[231] "hand"        "hands"       "hated"       "hatred"      "head"        "heard"       "heart"       "heavy"       "held"        "hide"
[241] "hideous"     "high"        "higher"      "hold"        "home"        "hope"        "horrible"    "horta"       "hour"        "hours"
[251] "house"       "however"     "huge"        "human"       "hundred"     "hung"        "hunting"     "hurled"      "hut"         "indeed"
[261] "instant"     "instead"     "interior"    "jane"        "john"        "jungle"      "just"        "kala"        "keen"        "keep"
[271] "kerchak"     "kerchaks"    "kill"        "killed"      "kind"        "king"        "knew"        "knife"       "know"        "known"
[281] "kulonga"     "kulongas"    "lady"        "laid"        "land"        "large"       "larger"      "last"        "later"       "launched"
[291] "lay"         "leaped"      "learned"     "least"       "leave"       "leaves"      "led"         "left"        "legs"        "less"
[301] "life"        "light"       "like"        "limb"        "lion"        "lioness"     "lips"        "little"      "lived"       "long"
[311] "longer"      "look"        "looked"      "looking"     "lord"        "lost"        "loud"        "low"         "lower"       "lying"
[321] "made"        "make"        "males"       "man"         "manner"      "many"        "mark"        "mass"        "matter"      "may"
[331] "mbonga"      "means"       "meant"       "meat"        "member"      "members"     "men"         "met"         "michael"     "might"
[341] "mighty"      "mind"        "minutes"     "moment"      "moments"     "monkeys"     "month"       "months"      "moon"        "morning"
[351] "mother"      "moved"       "much"        "muscles"     "must"        "mutiny"      "naked"       "near"        "nearly"      "neck"
[361] "never"       "new"         "next"        "night"       "noise"       "none"        "noose"       "nose"        "nothing"     "now"
[371] "number"      "occasionally" "officers"   "often"       "old"         "one"         "open"        "opened"      "opening"     "opportunity"
[381] "others"      "outside"     "pain"        "part"        "party"       "passed"      "path"        "people"      "philander"   "picture"
[391] "piece"       "place"       "placed"      "plunged"     "point"       "poor"        "porter"      "possibly"    "power"       "powerful"
[401] "presently"   "prey"        "primeval"    "professor"   "put"         "quick"       "quickly"     "quiet"       "quietly"     "quite"
[411] "race"        "rage"        "raised"      "ran"         "rapidly"     "ratfaced"    "rather"      "reached"     "read"        "ready"
[421] "realized"    "reason"      "repeated"    "replied"     "return"      "returned"    "revolver"    "revolvers"   "rifle"       "right"
[431] "roar"        "roaring"     "rolled"      "rolling"     "room"        "rope"        "rose"        "running"     "rushed"      "sabor"
[441] "sabors"      "safe"        "safely"      "safety"      "said"        "sailor"      "sailors"     "sat"         "savage"      "saw"
[451] "scarce"      "scarcely"    "scene"       "scream"      "search"      "second"      "securely"    "see"         "seemed"      "seen"
[461] "set"         "several"     "shadows"     "sharp"       "ship"        "shore"       "short"       "shot"        "shoulder"    "shoulders"
```

- **Terms with frequency of 11**

```
> #Frequency: 11
> freqTerms<-tm::findFreqTerms(TOTA_Stop_TDM, lowfreq = 11)
> freqTerms
  [1] "across"      "air"         "alice"       "almost"      "alone"       "along"       "also"        "always"      "among"       "animal"
 [11] "animals"     "another"     "ape"         "apeman"      "apes"        "approached"  "arm"         "armed"       "arms"        "arrow"
 [21] "arrows"      "attempt"     "attention"   "away"        "awful"       "back"        "beach"       "beast"       "beasts"      "behind"
 [31] "beneath"     "beside"      "beyond"      "black"       "blacks"      "boats"       "body"        "book"        "books"       "bow"
 [41] "boy"         "branch"      "branches"    "breast"      "british"     "broken"      "brought"     "brute"       "bugs"        "bull"
 [51] "cabin"       "called"      "came"        "captain"     "carried"     "caught"      "caused"      "chance"      "child"       "clayton"
 [61] "claytons"    "clearing"    "close"       "closed"      "closely"     "closer"      "come"        "commenced"   "continued"   "covered"
 [71] "creature"    "crew"        "cried"       "crouched"    "cruel"       "cry"         "danger"      "darkness"    "day"         "days"
 [81] "dead"        "death"       "deep"        "different"   "direction"   "directly"    "discovered"  "distance"    "door"        "dozen"
 [91] "drew"        "dropped"     "earth"       "eat"         "effort"      "end"         "enemy"       "english"     "enough"      "entered"
[101] "ere"         "escape"      "esmeralda"   "even"        "ever"        "every"       "evidently"   "except"      "eyes"        "face"
[111] "fact"        "fallen"      "fangs"       "far"         "farther"     "father"      "fear"        "feet"        "fell"        "fellow"
[121] "fellows"     "felt"        "females"     "fierce"      "figure"      "filled"      "finally"     "find"        "first"       "flesh"
[131] "floor"       "followed"    "following"   "food"        "forest"      "form"        "found"       "frightened"  "full"        "fuwalda"
[141] "gathered"    "gave"        "giant"       "girl"        "give"        "god"         "gone"        "good"        "gorilla"     "grasp"
[151] "grasped"     "great"       "greystoke"   "ground"      "hairy"       "half"        "hand"        "hands"       "head"        "heard"
[161] "heart"       "held"        "hide"        "hideous"     "high"        "hold"        "home"        "hope"        "hours"       "however"
[171] "huge"        "human"       "hundred"     "hung"        "hunting"     "hut"         "instant"     "instead"     "interior"    "jane"
[181] "john"        "jungle"      "just"        "kala"        "keep"        "kerchak"     "kill"        "killed"      "kind"        "king"
[191] "knew"        "knife"       "know"        "known"       "kulonga"     "lady"        "larger"      "last"        "lay"         "learned"
[201] "leave"       "led"         "left"        "legs"        "less"        "life"        "like"        "limb"        "lioness"     "lips"
[211] "little"      "long"        "longer"      "look"        "looked"      "lord"        "lost"        "low"         "made"        "make"
[221] "males"       "man"         "manner"      "many"        "mass"        "may"         "mbonga"      "means"       "meat"        "men"
[231] "michael"     "might"       "mighty"      "mind"        "moment"      "morning"     "mother"      "moved"       "much"        "must"
[241] "near"        "neck"        "never"       "new"         "next"        "night"       "noise"       "none"        "noose"       "nothing"
[251] "now"         "occasionally" "officers"   "often"       "old"         "one"         "open"        "opening"     "others"      "outside"
[261] "part"        "party"       "passed"      "people"      "picture"     "piece"       "place"       "point"       "poor"        "porter"
[271] "possibly"    "presently"   "prey"        "primeval"    "professor"   "put"         "quick"       "quickly"     "quiet"       "quite"
[281] "rage"        "raised"      "rapidly"     "rather"      "reached"     "read"        "realized"    "return"      "returned"    "revolver"
[291] "right"       "roar"        "room"        "rope"        "rose"        "rushed"      "sabor"       "safety"      "said"        "sailor"
[301] "sailors"     "sat"         "savage"      "saw"         "scarce"      "scarcely"    "scene"       "scream"      "search"      "second"
[311] "see"         "seemed"      "seen"        "set"         "several"     "shadows"     "sharp"       "ship"        "short"       "shot"
[321] "shoulders"   "side"        "sight"       "since"       "slowly"      "small"       "something"   "son"         "soon"        "sought"
[331] "sound"       "spear"       "sprang"      "standing"    "still"       "stood"       "stopped"     "strange"     "street"      "strength"
[341] "strong"      "suddenly"    "swung"       "take"        "taken"       "tarzan"      "tarzans"     "tearing"     "teeth"       "ten"
[351] "terkoz"      "terrible"    "terror"      "thing"       "things"      "though"      "thought"     "three"       "throat"      "thus"
[361] "time"        "times"       "tiny"        "together"    "told"        "took"        "top"         "torn"        "toward"      "trail"
[371] "tree"        "trees"       "tribe"       "tublat"      "turned"      "turning"     "twenty"      "two"         "understand"  "unknown"
[381] "upon"        "village"     "voice"       "warriors"    "watching"    "water"       "way"         "well"        "went"        "whatever"
[391] "white"       "wicked"      "wife"        "wild"        "will"        "window"      "within"      "without"     "woman"       "women"
[401] "wonderful"   "wondrous"    "word"        "words"       "work"        "years"       "yet"         "young"
>
```

- **Finding length of a string in Frequent terms**

```
#Finding length of a string in frequent terms
nchar(freqTerms[361])
freqTerms[361]
```

The provided code snippet aims to find the length of a specific string within the frequent terms list.

```
> #Finding length of a string in frequent terms
> nchar(freqTerms[361])
[1] 4
> freqTerms[361]
[1] "time"
```

We can inspect specific frequent terms, such as the 361st term mentioned above, which has a length of 4 characters. Moving forward, we can proceed to compute the term frequencies for each document. This involves generating a list of term frequencies for each document, which will later be unlisted and combined to obtain the overall term frequencies across the corpus.

```
> #Computing term frequencies for each document
> tfList<-list()
> for(i in seq_along(TOTA_Stop)){
+   tfList[[i]]<-termFreq(TOTA_Stop[[i]])
+ }
> print(tfList[[1]])
```

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| able | aboard | accentuated | accept | accord | accorded | achievement | acknowledging | act | advised |
| 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| affairs | afford | africa | african | aft | afternoon | agencies | aggregation | agin | ago |
| 1 | 1 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| alice | almost | alone | along | also | amazed | ambition | ammunition | among | amount |
| 11 | 4 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| anchor | ancient | angry | another | answered | antagonize | anxious | anything | apartment | appalled |
| 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 2 |
| apparent | appointed | appointment | appreciate | arf | arguments | arm | army | around | arrived |
| 2 | 1 | 2 | 1 | 1 | 1 | 2 | 3 | 1 | 1 |
| aruwimi | asayin | ask | asked | ass | assisting | associate | assumed | atlantic | atmosphere |
| 1 | 1 | 2 | 4 | 1 | 1 | 1 | 1 | 2 | 1 |
| attempt | attempted | attribute | aught | aunts | authority | average | averted | avoid | away |
| 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 |
| awful | back | backwards | bad | bags | balance | barkentine | battlefieldsa | battleship | bear |
| 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| beastly | became | beds | beginning | begun | behind | belaying | belief | believed | belongings |
| 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 |
| beneath | best | betrayed | better | big | billings | bit | black | blacks | blank |
| 1 | 2 | 1 | 1 | 1 | 2 | 2 | 7 | 1 | 1 |
| blasted | bloody | bloomin | blow | board | body | bore | borrowing | boxes | brasses |
| 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| breath | bright | british | brothers | brought | brutal | brutality | brute | bull | bullet |
| 1 | 1 | 9 | 1 | 3 | 2 | 2 | 2 | 1 | 1 |
| bullied | bullies | bully | business | busted | cabin | calculated | call | came | can |
| 1 | 1 | 1 | 2 | 1 | 4 | 1 | 1 | 3 | 3 |
| candor | captain | captains | capting | care | career | careful | carelessly | carriage | carried |
| 1 | 22 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 2 |
| carry | caused | centered | certain | chain | chance | change | characters | chartered | cheerful |
| 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| circumstances | class | clayton | claytons | cleared | close | closed | closer | clothing | coarse |
| 1 | 1 | 29 | 6 | 1 | 1 | 1 | 2 | 1 | 1 |
| coast | coastwise | collection | colonial | colony | come | comin | commanding | commenced | commission |
| 1 | 1 | 1 | 5 | 2 | 4 | 2 | 1 | 1 | 1 |
| commissioned | common | companionway | competent | complained | composed | comrade | conception | concern | conciliate |
| 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| condition | conditions | confidential | confirmed | congo | consequences | contemplate | contemplates | contents | control |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| conversation | convinced | convivial | corner | countenance | cousins | covers | craft | credible | credit |
| 2 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| crew | crews | crouched | crude | crushing | culminate | curse | cutthroats | daily | damned |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 |
| danger | dangers | day | days | dead | dealings | dear | death | decided | deck |
| 2 | 1 | 5 | 3 | 2 | 1 | 2 | 1 | 1 | 4 |

```
> print(tfList[[14]])
```

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| abundance | across | add | added | adventure | agility | ago | air | alert | alighting | also |
| 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 |
| always | among | ancestry | ancient | angry | animal | animals | another | answer | ape | apechild |
| 1 | 3 | 1 | 2 | 2 | 1 | 3 | 1 | 1 | 1 | 1 |
| apes | apparel | approached | ascendency | attention | attract | author | away | awesome | back | backlaid |
| 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 6 | 1 |
| badge | bared | bargain | bases | beach | beast | beat | became | beginning | behind | belched |
| 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 |
| bellowing | bending | beneath | bent | betokened | beyond | birds | biting | bitterest | blackness | blinding |
| 1 | 1 | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| boar | body | bolt | books | bottom | bound | braced | bracing | braid | branch | branches |
| 2 | 6 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 4 | 8 |
| bristling | broke | bronze | budged | busy | cabbage | cabin | came | cameall | cannonade | captive |
| 1 | 2 | 1 | 1 | 1 | 1 | 1 | 5 | 1 | 1 | 1 |
| carrying | cat | catlike | cause | ceased | challenge | charged | chest | children | clawing | close |
| 2 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 1 |
| closed | clothes | clouds | clung | clutched | coast | coat | coil | coils | cold | come |
| 1 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 |
| completely | conflicting | conform | confusion | considerable | constant | continued | cord | countless | cover | covered |
| 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 |
| cracked | crafty | crash | creature | crossing | crotch | crouched | cruel | curling | customs | danced |
| 1 | 1 | 2 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 |
| dancing | danger | darkness | darting | dawning | day | days | dead | deafening | death | deep |
| 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 |
| deluge | denizens | dense | descent | desire | desired | desires | destruction | details | detect | different |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| directly | disaster | distance | done | doubtless | dragged | dragging | dripping | dropped | dumdum | ears |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| earth | earthward | easy | eat | efficacy | eggs | elephant | else | end | enemies | enemy |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 3 |
| english | engulfs | equal | esteem | even | ever | every | expectancy | experience | explain | eyed |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 |
| eyes | face | faintly | fairly | fall | fallen | falling | fangs | far | farther | fastened |
| 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 |
| fat | fear | fearsome | feet | fell | felt | ferocity | fierce | fighter | figure | filled |
| 1 | 1 | 2 | 4 | 4 | 1 | 2 | 2 | 1 | 1 | 1 |
| finally | first | flashed | flashing | flowers | foliage | food | forest | forged | forgot | forth |
| 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 1 | 2 | 3 |
| forthnature | found | four | fours | freedom | fright | fruit | fury | futilely | gained | game |
| 1 | 4 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| giants | glistened | glossy | gorgeous | graceful | grasped | grasping | gray | great | greater | greatness |
| 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 14 | 1 | 1 |
| grimaces | ground | hairless | half | hand | hands | hanging | happily | hard | hatred | head |

We revisit the TOTA_Stop_TDM for further inspection.

```
> #Inspecting TOTA_Stop_TDM again
> tm::inspect(TOTA_Stop_TDM)
<<TermDocumentMatrix (terms: 5326, documents: 15)>>
Non-/sparse entries: 12944/66946
Sparsity           : 84%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
         Docs
Terms     chapter1.txt chapter11.txt chapter12.txt chapter13.txt chapter14.txt chapter2.txt chapter4.txt chapter5.txt chapter7.txt chapter9.txt
  apes               0           18           11            6            2           0            9           12           12           11
  clayton           29            1            0           10           24          18            5            0            0            0
  great              4           18           14            9           17           8           11           12           15           10
  jungle             0            6            5           20           16           3            2            6           14           17
  little            12           12           10           17            3          11           16           24           28           12
  man                8            5            7           31           19          10            3            1            2            8
  now                7            8            5            6           13           5            5            5           11            2
  one               11           15            9           19           10           2           10            8           22           13
  tarzan             0           39           43           33           14           0            0           20           22           34
  upon              11           20           11           31           20          15           14           13           22           23
>
```

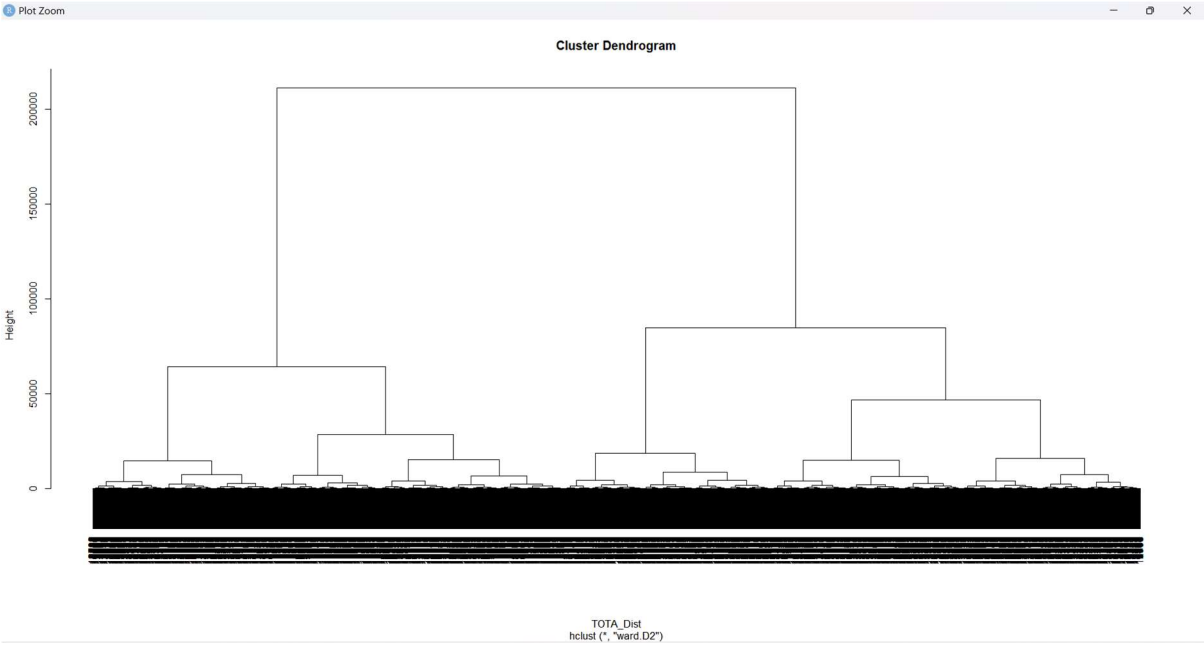
## GENERATING DENDRGRAM

The provided code segment aims to generate a dendrogram for visualizing the hierarchical clustering of documents based on their term frequencies post the removal of stop words. This process facilitates a deeper exploration of document similarities and relationships within the corpus. Initially, the term document matrix (TDM), named "TOTA_Stop_TDM", is transformed into a data frame labeled "TOTA_df". This conversion enables subsequent distance calculation and dendrogram generation.

Following the conversion, a distance matrix is computed using the dist() function, leveraging the data frame "TOTA_df". This matrix captures the pairwise distances between documents, essential for hierarchical clustering. The hclust() function is then employed to construct the dendrogram, utilizing the distance matrix. The method "ward.D2" is specified within hclust(), indicating the utilization of Ward's minimum variance method for hierarchical clustering, which tends to produce well-separated clusters.

```
> #Generating Dendrogram
> TOTA_df<-as.data.frame(TOTA_Stop_TDM[[1]])
> TOTA_Dist<-dist(TOTA_df)
> TOTA_DG<-hclust(TOTA_Dist, method="ward.D2")
> str(TOTA_DG)
List of 7
 $ merge      : int [1:12943, 1:2] -1 -3458 -6189 -7143 -10140 -2 -6190 -4 -7145 -5 ...
 $ height     : num [1:12943] 0 0 0 0 0 0 0 0 0 0 ...
 $ order      : int [1:12944] 3 9422 2633 4714 10142 7145 4 2634 8610 2635 ...
 $ labels     : NULL
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = TOTA_Dist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

Finally, the resulting dendrogram is plotted using the plot() function.




We noticed that our dendrogram contains numerous labels, making it challenging to analyze effectively. To address this issue, we opted to eliminate some of the sparse terms from the term document matrix (TDM). Utilizing the removeSparseTerms() function from the tm package, we experimented with different thresholds for maximal allowed sparsity.

- Through this experimentation, we determined that setting a maximal sparsity threshold of 0.85 resulted in a dendrogram that is both comprehensive and analyzable. It yielded a sparsity of 63%, striking a balance between reducing clutter in the dendrogram while retaining a substantial number of relevant terms for our analysis.
- Setting a maximum sparsity threshold of 0.5 produced a dendrogram that is both comprehensive and analyzable. This approach resulted in a 30% sparsity rate, balancing the reduction of clutter in the dendrogram with the retention of significant terms for analysis.

```
> #Eliminating some more sparse words
> TOTA_Stop_TDM <- removeSparseTerms(TOTA_Stop_TDM, sparse = 0.85)
> TOTA_Stop_TDM
<<TermDocumentMatrix (terms: 1470, documents: 15)>>
Non-/sparse entries: 8143/13907
Sparsity           : 63%
Maximal term length: 14
Weighting          : term frequency (tf)
> TOTA_Stop_TDM <- removeSparseTerms(TOTA_Stop_TDM, sparse = 0.5)
> TOTA_Stop_TDM
<<TermDocumentMatrix (terms: 310, documents: 15)>>
Non-/sparse entries: 3264/1386
Sparsity           : 30%
Maximal term length: 12
Weighting          : term frequency (tf)
>
```

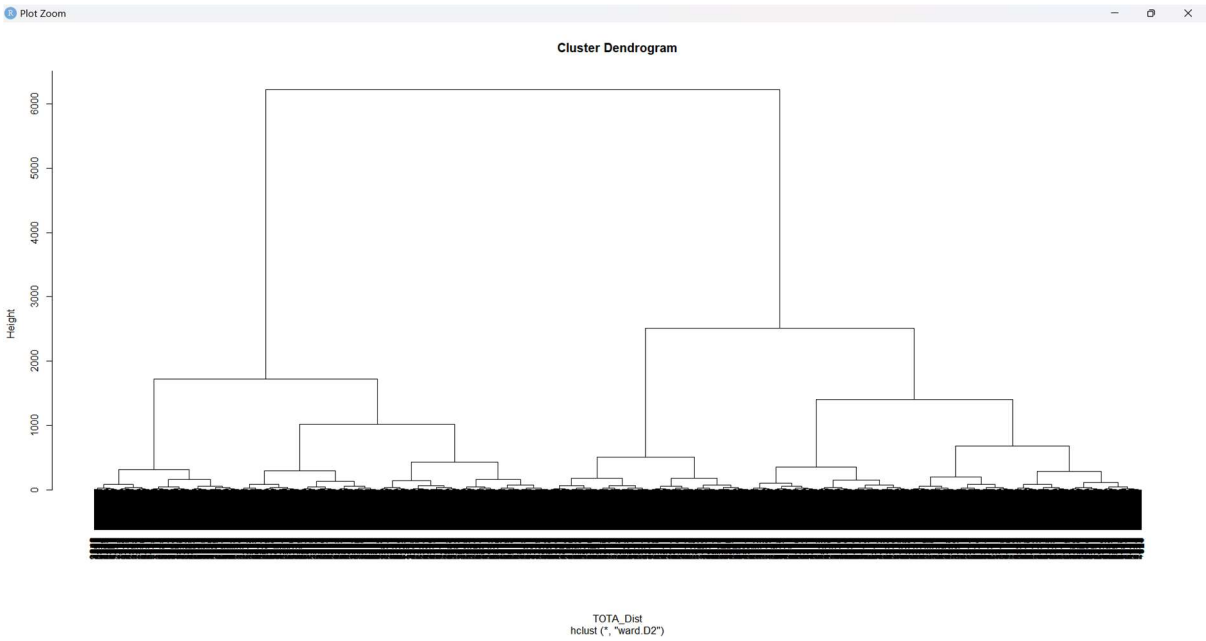
Inspecting TOTA_Stop_TDM with sparse=0.5

```
> TOTA_Stop_TDM <- removeSparseTerms(TOTA_Stop_TDM, sparse = 0.5)
> TOTA_Stop_TDM
<<TermDocumentMatrix (terms: 310, documents: 15)>>
Non-/sparse entries: 3264/1386
Sparsity           : 30%
Maximal term length: 12
Weighting          : term frequency (tf)
> tm::inspect(TOTA_Stop_TDM)
<<TermDocumentMatrix (terms: 310, documents: 15)>>
Non-/sparse entries: 3264/1386
Sparsity           : 30%
Maximal term length: 12
Weighting          : term frequency (tf)
Sample             :
         Docs
Terms     chapter1.txt chapter11.txt chapter12.txt chapter13.txt chapter14.txt chapter2.txt chapter4.txt chapter5.txt
  apes               0            18            11             6             2           0           9          12
  clayton           29             1             0            10            24          18           5           0
  great              4            18            14             9            17           8          11          12
  jungle             0             6             5            20            16           3           2           6
  little            12            12            10            17             3          11          16          24
  man                8             5             7            31            19          10           3           1
  now                7             8             5             6            13           5           5           5
  one               11            15             9            19            10           2          10           8
  tarzan             0            39            43            33            14           0           0          20
  upon              11            20            11            31            20          15          14          13
         Docs
Terms     chapter7.txt chapter9.txt
  apes              12           11
  clayton            0            0
  great             15           10
  jungle            14           17
  little            28           12
  man                2            8
  now               11            2
  one               22           13
  tarzan            22           34
  upon              22           23
>
```

Generating the new dendrogram:

```
> #Generating Dendrogram
> TOTA_df<-as.data.frame(TOTA_Stop_TDM[[1]])
> TOTA_Dist<-dist(TOTA_df)
> TOTA_DG<-hclust(TOTA_Dist, method="ward.D2")
> str(TOTA_DG)
List of 7
 $ merge      : int [1:3263, 1:2] -1 -383 -639 -847 -1108 -1509 -1714 -1925 -2141 -2581 ...
 $ height     : num [1:3263] 0 0 0 0 0 0 0 0 0 0 ...
 $ order      : int [1:3264] 3042 2849 2605 2372 2162 1949 1736 1348 1130 869 ...
 $ labels     : NULL
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = TOTA_Dist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

Plotting the new dendrogram:



## GENERATING THE WORD CLOUD

Now, we can move forward with generating a word cloud. To accomplish this task, we utilize the "wordcloud" package. Firstly, we need to install the package, and then we import it as a library into our project environment. This package facilitates the creation of visually appealing word clouds, which can effectively showcase the frequency and importance of terms within our corpus.

To prepare for generating a word cloud, we first gather a set of unique words from the entire corpus stored in "TOTA". By unlisting the "tfList" variable and summing up the frequencies for each unique term, we obtain a combined term frequency. This process ensures that we capture all unique terms across the 15 documents. Finally, we extract the names of the summed term frequencies, storing them in the "words" variable, which constitutes the set of words for our word cloud visualization.

```
> #WORD CLOUD
> #Getting sset of words from TOTA
> tf_combined<-unlist(tfList)
> tf_summed<-tapply(tf_combined, names(tf_combined), sum)
> words<-names(tf_summed)
> words
  [1] "abandon"          "abandoned"        "abandoning"       "abashed"          "abated"
  [6] "ability"          "able"             "aboard"           "abode"            "abruptly"
 [11] "absence"          "absent"           "absentmindedness" "absolute"         "absolutely"
 [16] "absorbed"         "abstruse"         "abundance"        "abundant"         "abuses"
 [21] "abysmal"          "accentuate"       "accentuated"      "accept"           "accession"
 [26] "accident"         "accidental"       "accidentally"     "accompanied"      "accompany"
 [31] "accomplish"       "accomplished"     "accomplishment"   "accord"           "accorded"
 [36] "account"          "accounted"        "accuracy"         "accurately"       "accustom"
 [41] "accustomed"       "achievement"      "acknowledging"    "acquaintances"    "acquirement"
 [46] "across"           "act"              "action"           "active"           "add"
 [51] "added"            "adding"           "addition"         "additional"       "addressed"
 [56] "adds"             "adequate"         "adjusted"         "adjusting"        "admiration"
 [61] "admired"          "admit"            "admitted"         "adopted"          "adult"
 [66] "adults"           "advance"          "advanced"         "advancing"        "advantage"
 [71] "adventure"        "adverbs"          "adversary"        "advised"          "aerial"
 [76] "aerie"            "affairs"          "affected"         "affection"        "afford"
 [81] "affrighted"       "affront"          "afraid"           "africa"           "african"
 [86] "aft"              "afternoon"        "age"              "agencies"         "ages"
 [91] "aggregation"      "agile"            "agility"          "agin"             "ago"
 [96] "agonized"         "agony"            "agree"            "agreeable"        "agreedthat"
[101] "ahead"            "aim"              "aimed"            "aimlessly"        "aint"
[106] "air"              "airraised"        "alarm"            "alas"             "alert"
[111] "alice"            "alight"           "alighting"        "alike"            "alive"
[116] "alla"             "allay"            "allegorical"      "allied"           "allow"
[121] "allowed"          "almost"           "aloft"            "alone"            "along"
[126] "aloud"            "alphabet"         "alphabetical"     "already"          "also"
[131] "altardrum"        "alternately"      "alternative"      "always"           "amazed"
[136] "amazing"          "ambition"         "ambitions"        "amen"             "american"
[141] "amidst"           "amiss"            "amity"            "ammunition"       "among"
[146] "amongst"          "amount"           "amphitheater"     "amuck"            "amused"
[151] "amusement"        "ancestor"         "ancestors"        "ancestry"         "anchor"
[156] "anchored"         "ancient"          "andoh"            "anger"            "angered"
[161] "angry"            "animal"           "animals"          "animosity"        "ankle"
[166] "anklet"           "anklets"          "announced"        "annoy"            "another"
[171] "anothera"         "answer"           "answered"         "answering"        "answers"
[176] "antagonist"       "antagonize"       "antelope"         "anthropoid"       "anthropoids"
[181] "anticipation"     "anxiety"          "anxious"          "anyone"           "anything"
[186] "anyway"           "apart"            "apartment"        "apathy"           "ape"
[191] "apechild"         "apehood"          "apeman"           "apemans"          "aperture"
[196] "apes"             "apesno"           "aplenty"          "appalled"         "apparel"
[201] "apparent"         "apparently"       "appearance"       "appeared"         "appearing"
[206] "appease"          "appetite"         "applied"          "appointed"        "appointment"
```

Following the extraction of unique words, we proceed to create the word cloud using the "wordcloud()" function. This function takes the summed term frequencies and the

corresponding words as inputs to generate the visualization. However, during this process, certain warnings were encountered indicating that some words were too lengthy to be accommodated within the word cloud and were consequently excluded.

```
> #Word cloud with different pallet
> pal2<-brewer.pal(9,"Spectral")
> #Generating word cloud
> pal<-brewer.pal(9,"BuGn")
> str(pal)
 chr [1:9] "#F7FCFD" "#E5F5F9" "#CCECE6" "#99D8C9" "#66C2A4" "#41AE76" "#238B45" "#006D2C" "#00441B"
> TOTA_WC<-wordcloud(words, tf_summed, colors=pal[-(1:4)])
There were 50 or more warnings (use warnings() to see the first 50)
>
```

beneath
tree
kerchak
tarzans
village
eyes
esmeralda
among
books
girl
knew
far
feet
water
passed
distance
strength
right
also
deep
thus
near
poor
wonderful
beyond
scream
bugs
said
felt
take
darkness
fellows
jane
occasionally
crouched
palisade
opened
windows
constant
mingled
remained
done
beings
samuel
plunge
glossy
quietly
minutes
practiced
gesticulating
around
common
rise
taste
iron
lived
strips
directions
scriptview
minds
blackness
jerk
leap
glimpses
hated
sink
greater
horrified
catch
wifes
brothen
depended
nearer
deer
dried
discover
larger
attempted
accustomed
mysteries
defenseless
incredible
ending
month
sake
nerves
fled
bar
desires
inert
duties
dancing
position
required
teeth
features
piece
gather
handsome
fury
tail
pencil
think
witnessed
wounded
free
pieces
tighter
working
heavy
cat
inside
belt
youngers
spectacles
cauldron
forth
headlong
lightly
less
descend
shot
joyguns
rest
fancy
locket
grew
reigned
similar
misery
cradle
third
manchild
difference
call
unless
fresh
tell
least
crying
answering
paces
shrewd
easy
suffering
mere
fronds
folk
brutes
say
tut
rubber
cautiously
entrance
band
south
anxious
spot
stop
mile
surely
speak
earlier
grasses
squareencircled
bared
revel
knowing
sitting
hugged
crept
din
peered
fly
flashing
vivid
author
huts
underbrush
fitted
edge
snipes
knowledge
world
army
simple
spoke
agony
bloody
single
dare
laid
lithe
lies
born
babe
armed
extreme
fate
wandering
stuff
jar
unprotected
pile
distant
infuriated
bit
eye
ear
hair
faces
kill
recover
crushed
plunged
threw
late
surprise
crouching
keeping
vanquished
talk
fifty
eat
flat
plantation
pots
sea
harbor
progress
plain
countenance
victim
reast
gry
yer
like
legs
set
enough
possibly
fear
except
years
look
see
twenty
night
neck
light
hours
bugs
kind
rushed
moment
tree
ape
trail
several
party
never
life
wondrous
another
yet
turned
little
enemy
kulonga
gorilla
claytons
almost
place
opening
turning
returned
willlips
jane
thing
tearing
quite
hated
commenced
fallen
scarce
entirely
instead
hold
times
open
two
looking
strong
gone
side
much
terrible
know
kala
trees
beach
earth
mother
evidently
noise
fellow
many
bull
find
well
sharp
sound
book
terkoz
fact
far
cleaning
ground
arrows
left
hour
end
dozen
limb
teeth
added
reached
longer
three
distance
later
realized
tarzan
cabin
head
saw
black
long
knew
body
savage
man
strange
jungle
fangs
white
god
street
scene
rifle
give
son
understand
floor
work
sprang
shadowsten
hope
grasp
ere
days
next
watching
rapidly
shot
make
top
ran
meat
captain
lost
right
near
girl
can
follow
fell
moved
task
animal
just
women
lion
hide
arrow
kill
looked
hour
different
better
spear
tiny
brought
felt
also
spicture
standing
number
heard
face
huntir
seemed
soon
went
fallen
eat
hide
hold
cleaning
hated
scarce
instead
entirely
tearing
place
almost
claytons
gorilla
kulonga
enemy
yet
another
wondrous
life
never
three
several
tree
ape
jungle
strange
ground
arrows
left
two
open
times
book
terkoz
sound
sharp
find
well
strong
side
many
man
black
saw
head
cabin
tarzan
kala
trees
beach
much
know
terrible
gone
side
fellow
noise
girl
near
right
lost
captain
make
savage
evidently
mother
knew
body
long
face
heard
number
standing
also
spicture
different
better
spear
tiny
brought
felt

The second-word cloud offers enhanced clarity and comprehension of the textual data represented within the corpus. By filtering out words with lower frequencies (set at a minimum of 10 occurrences), the visualization provides a more focused depiction of prominent terms. This refined presentation aids in discerning key themes, patterns, and trends within the corpus, facilitating a deeper understanding of the underlying textual content.

Word cloud generated with different pallet:

```
> #Word cloud with different pallet
> pal2<-brewer.pal(9,"Spectral")
> TOTA_WC<-wordcloud(words, tf_summed, colors = pal2)
```

prancnes
greystoke
learned
face
something
thus
years
larger
moment
even
death
warriors
discovered
giant
hideous
way
made
young
things
knew
arms
hundred
white
great
set
stopped
right
teeth
world
near
throat
arm
just
noose
immediately
confusion
covered
chance
effort
flesh
sea
dawning
gray
overwrought
smaller
amongst
rolling
foot
remain
fight
offering
kalas
remains
twelve
show
agony
muscled
faces
sides
spied
lines
sink
bound
cuthidden
dryhappenings
companion
pots
paw
name
throw
noon
matted
increasing
rubber
harm
goal
pitiful
inert
broad
fingers
lions
soft
scattered
similar
poisoned
antagonist
continue
squatting
progress
warrior
hard
belt
leap
tublats
big
guide
ning
fresh
ships
desires
forearm
notice
gathered
kerchaks
undergrowth
frightful
next
follow
swiftly
cries
rose
making
massive
greatly
crash
fruit
ago
narrative
house
victims
accept
platform
perfect
bullet
worn
panther
occurred
wait
bulls
constant
afraid
nose
sounds
circle
faced
photograph
din
smaller
bit
drop
rule
snarling
sea
flesh
stopped
leap
teeth
world
guide
tublats
big
ning
silence
walked
witnessed
limp
keeping
save
numa
huh
rose
legs
tell
worn
occurred
afraid
faced
circle
din
overwrought
dawning
gray
hurried
right
tell
bullet
worn
occurred
wait
bulls
constant
afraid
nose
sounds
circle
faced
photograph

Following keyword extraction, we proceed with tokenization by employing the lapply() function to tokenize each of the 11 documents within the corpus.

```
#kwic
TOTATokens <-lapply(TOTA_cl,function(x) quanteda::tokens(x$content))
TOTATokens
str(TOTATokens)
```

Inspecting the structure of TOTATokens using the str() function reveals extensive output for all fifteen documents within the corpus.

```
 ..$ text89 : chr [1:11] "savage" "blacks" "meant" "consternation" ...
 ..$ text90 : chr [1:5] "denizens" "of" "their" "new" ...
 ..$ text91 : chr(0)
 ..$ text92 : chr [1:12] "For" "three" "days" "the" ...
 ..$ text93 : chr [1:11] "this" "unknown" "and" "untracked" ...
 ..$ text94 : chr [1:14] "day" "they" "came" "upon" ...
 ..$ text95 : chr [1:11] "which" "seemed" "less" "thickly" ...
 ..$ text96 : chr "encountered"
 ..$ text97 : chr(0)
 ..$ text98 : chr [1:16] "Here" "they" "set" "to" ...
 ..$ text99 : chr [1:11] "clearing" "had" "been" "made" ...
 .. [list output truncated]
 ..- attr(*, "class")= chr "tokens"
 ..- attr(*, "types")= chr [1:1183] "Man" "and" "Tarzan" "of" ...
 ..- attr(*, "padding")= logi TRUE
 ..- attr(*, "docvars")='data.frame':  439 obs. of  3 variables:
 .. ..$ docname_: chr [1:439] "text1" "text2" "text3" "text4" ...
 .. ..$ docid_  : Factor w/ 439 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
 .. ..$ segid_  : int [1:439] 1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "meta")=List of 3
 .. ..$ system:List of 5
 .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
 .. .. .. ..$ : int [1:3] 4 0 2
 .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
 .. .. .. ..$ : int [1:3] 4 3 2
 .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "Sravani"
 .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
 .. .. ..$ directory      : chr "C:/Users/Sravani/OneDrive/Documents/2024_SPRING/CS6444 - BIG DATA/Projects/Project_3/eB
"
 .. .. ..$ created        : Date[1:1], format: "2024-05-04"
 .. ..$ object:List of 7
 .. .. ..$ unit        : chr "documents"
 .. .. ..$ what        : chr "word"
 .. .. ..$ tokenizer   : chr "tokenize_word4"
 .. .. ..$ ngram       : int 1
 .. .. ..$ skip        : int 0
 .. .. ..$ concatenator: chr "_"
 .. .. ..$ summary     :List of 2
 .. .. .. ..$ hash: chr(0)
 .. .. .. ..$ data: NULL
 .. ..$ user  : list()
```

The ‘kwic()’ function from the ‘quanteda’ package enables users to locate keywords in context within a given text corpus. By specifying a particular keyword as the search term, the function generates a keyword-in-context (KWIC) concordance. This concordance displays instances of the keyword found within the text, along with the surrounding context. We can utilize this function to perform detailed textual analysis and gain insights into how specific terms are used within the corpus.

Next, we utilize the dfm() function to generate sparse document-feature matrices for all the chapters. Additionally, we employ dfm_remove() to eliminate stop words from the matrices.

```
> TOTA_DFM_Chap1 <- quanteda::dfm(TOTATokens[[1]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap2 <- quanteda::dfm(TOTATokens[[2]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap3 <- quanteda::dfm(TOTATokens[[3]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap4 <- quanteda::dfm(TOTATokens[[4]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap5 <- quanteda::dfm(TOTATokens[[5]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap6 <- quanteda::dfm(TOTATokens[[6]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap7 <- quanteda::dfm(TOTATokens[[7]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap8 <- quanteda::dfm(TOTATokens[[8]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap9 <- quanteda::dfm(TOTATokens[[9]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap10 <- quanteda::dfm(TOTATokens[[10]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap11 <- quanteda::dfm(TOTATokens[[11]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap12 <- quanteda::dfm(TOTATokens[[12]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap13 <- quanteda::dfm(TOTATokens[[13]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap14 <- quanteda::dfm(TOTATokens[[14]]) %>% dfm_remove(myStopWords)
> TOTA_DFM_Chap15 <- quanteda::dfm(TOTATokens[[15]]) %>% dfm_remove(myStopWords)
> str(TOTA_DFM_Chap1)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    420 obs. of  3 variables:
  .. ..$ docname_: chr [1:420] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 420 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:420] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "Sravani"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/Sravani/OneDrive/Documents/2024_SPRING/CS6444 - BIG DATA/Projects/Project
k"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 5
  .. .. .. ..$ scheme   : chr "unary"
```

Subsequently, we calculate the frequency of terms in the Document-Feature Matrix (DFM) for the chapters.

```
> #Getting the frequency of terms in DFM
> TOTADocFreq_Chap1 <- quanteda::docfreq(TOTA_DFM_Chap1)
> TOTADocFreq_Chap2 <- quanteda::docfreq(TOTA_DFM_Chap2)
> TOTADocFreq_Chap3 <- quanteda::docfreq(TOTA_DFM_Chap3)
> TOTADocFreq_Chap4 <- quanteda::docfreq(TOTA_DFM_Chap4)
> TOTADocFreq_Chap5 <- quanteda::docfreq(TOTA_DFM_Chap5)
> TOTADocFreq_Chap6 <- quanteda::docfreq(TOTA_DFM_Chap6)
> TOTADocFreq_Chap7 <- quanteda::docfreq(TOTA_DFM_Chap7)
> TOTADocFreq_Chap8 <- quanteda::docfreq(TOTA_DFM_Chap8)
> TOTADocFreq_Chap9 <- quanteda::docfreq(TOTA_DFM_Chap9)
> TOTADocFreq_Chap10 <- quanteda::docfreq(TOTA_DFM_Chap10)
> TOTADocFreq_Chap11 <- quanteda::docfreq(TOTA_DFM_Chap11)
> TOTADocFreq_Chap12 <- quanteda::docfreq(TOTA_DFM_Chap12)
> TOTADocFreq_Chap13 <- quanteda::docfreq(TOTA_DFM_Chap13)
> TOTADocFreq_Chap14 <- quanteda::docfreq(TOTA_DFM_Chap14)
> TOTADocFreq_Chap15 <- quanteda::docfreq(TOTA_DFM_Chap15)
> str(TOTADocFreq_Chap5)
 Named int [1:1268] 1 2 3 33 1 1 1 2 3 2 ...
 - attr(*, "names")= chr [1:1268] "kind" "following" "morning" "tarzan" ...
> TOTADocFreq_Chap5
          kind      following        morning         tarzan           lame           sore         wounds
             1              2              3             33              1              1              1
        battle         terkoz            set         toward           west       seacoast       traveled
             2              3              2             15              2              1              1
        slowly       sleeping         jungle          night       reaching          cabin           late
             6              1             20              1              1             13              1
       several           days          moved         little         enough         gather         fruits
             3              2              2             17              1              1              1
          nuts       required        satisfy        demands         hunger            ten          quite
             1              1              1              1              1              2              6
         sound         except       terrible     halfhealed           scar       starting           left
             6              1              3              1              2              1              8
           eye            ran         across            top           head         ending          right
             1              4              2              2              5              1              3
           ear           mark           torn          scalp           away  convalescence          tried
             3              2              1              1              7              1              1
       fashion         mantle           skin          sabor           lain           time          found
             2              1              1              2              1              5              6
          hide          dried          stiff          board           knew         naught        tanning
             2              1              1              1              2              1              1
        forced        abandon       cherished           plan      determined          filch       garments
             2              1              1              1              1              1              1
           one          black            men       mbongas        village           apes        decided
            19             10             11              2             10              6              2
     evolution          lower         orders          every       possible         manner        nothing
             1              2              1              3              1              2              3
        seemed  distinguishing          badge       manhood       ornaments       clothing            end
```

Following term frequency calculation, we proceed to assign weights to these words.

```
> #Assigning weights to these words
> TOTAWeights_Chap1 <- quanteda::dfm_weight(TOTA_DFM_Chap1)
> TOTAWeights_Chap2 <- quanteda::dfm_weight(TOTA_DFM_Chap2)
> TOTAWeights_Chap3 <- quanteda::dfm_weight(TOTA_DFM_Chap3)
> TOTAWeights_Chap4 <- quanteda::dfm_weight(TOTA_DFM_Chap4)
> TOTAWeights_Chap5 <- quanteda::dfm_weight(TOTA_DFM_Chap5)
> TOTAWeights_Chap6 <- quanteda::dfm_weight(TOTA_DFM_Chap6)
> TOTAWeights_Chap7 <- quanteda::dfm_weight(TOTA_DFM_Chap7)
> TOTAWeights_Chap8 <- quanteda::dfm_weight(TOTA_DFM_Chap8)
> TOTAWeights_Chap9 <- quanteda::dfm_weight(TOTA_DFM_Chap9)
> TOTAWeights_Chap10 <- quanteda::dfm_weight(TOTA_DFM_Chap10)
> TOTAWeights_Chap11 <- quanteda::dfm_weight(TOTA_DFM_Chap11)
> TOTAWeights_Chap12 <- quanteda::dfm_weight(TOTA_DFM_Chap12)
> TOTAWeights_Chap13 <- quanteda::dfm_weight(TOTA_DFM_Chap13)
> TOTAWeights_Chap14 <- quanteda::dfm_weight(TOTA_DFM_Chap14)
> TOTAWeights_Chap15 <- quanteda::dfm_weight(TOTA_DFM_Chap15)
> str(TOTAWeights_Chap15)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    439 obs. of  3 variables:
  .. ..$ docname_: chr [1:439] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 439 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:439] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "Sravani"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/Sravani/OneDrive/Documents/2024_SPRING/CS6444 - BIG DATA/Projects/Project_3/eBo
k"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 5
  .. .. .. ..$ scheme   : chr "unary"
  .. .. .. ..$ base     : NULL
  .. .. .. ..$ c        : NULL
  .. .. .. ..$ smoothing: NULL
> TOTAWeights_Chap15
Document-feature matrix of: 439 documents, 1,029 features (99.60% sparse) and 0 docvars.
       features
docs    man tarzan apes lived wild jungle existence little change several
  text1   0      0    0     0    0      0         0      0      0       0
  text2   2      0    0     0    0      0         0      0      0       0
  text3   0      0    0     0    0      0         0      0      0       0
  text4   0      0    0     0    0      0         0      0      0       0
  text5   0      1    1     1    1      1         1      1      0       0
  text6   0      0    0     0    0      0         0      0      1       1
[ reached max_ndoc ... 433 more documents, reached max_nfeat ... 1,019 more features ]
>
```

Next, we compute the TF-IDF score, a metric that weighs a Document-Feature Matrix (DFM) by term frequency-inverse document frequency (TF-IDF), offering complete control over various options.

```
> #Computing the tf-idf score.
> TOTA_TFIDF_Chap1 <- quanteda::dfm_tfidf(TOTA_DFM_Chap1,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap2 <- quanteda::dfm_tfidf(TOTA_DFM_Chap2,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap3 <- quanteda::dfm_tfidf(TOTA_DFM_Chap3,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap4 <- quanteda::dfm_tfidf(TOTA_DFM_Chap4,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap5 <- quanteda::dfm_tfidf(TOTA_DFM_Chap5,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap6 <- quanteda::dfm_tfidf(TOTA_DFM_Chap6,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap7 <- quanteda::dfm_tfidf(TOTA_DFM_Chap7,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap8 <- quanteda::dfm_tfidf(TOTA_DFM_Chap8,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap9 <- quanteda::dfm_tfidf(TOTA_DFM_Chap9,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap10 <- quanteda::dfm_tfidf(TOTA_DFM_Chap10,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap11 <- quanteda::dfm_tfidf(TOTA_DFM_Chap11,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap12 <- quanteda::dfm_tfidf(TOTA_DFM_Chap12,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap13 <- quanteda::dfm_tfidf(TOTA_DFM_Chap13,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap14 <- quanteda::dfm_tfidf(TOTA_DFM_Chap14,scheme_tf = "count", scheme_df="inverse")
> TOTA_TFIDF_Chap15 <- quanteda::dfm_tfidf(TOTA_DFM_Chap15,scheme_tf = "count", scheme_df="inverse")
> str(TOTA_TFIDF_Chap7)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    199 obs. of  3 variables:
  .. ..$ docname_: chr [1:199] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 199 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:199] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "Sravani"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/Sravani/OneDrive/Documents/2024_SPRING/CS6444 - BIG DATA/Projects/Project_3/eBo
k"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 2
  .. .. .. ..$ scheme: chr "inverse"
```

```
> TOTA_TFIDF_Chap7
Document-feature matrix of: 199 documents, 516 features (99.23% sparse) and 0 docvars.
       features
docs      forest      god   clayton    heard   report  firearm     fell    agony     fear apprehension
  text1 0        0        0         0        0        0        0        0        0            0
  text2 1.821732 1.696793 0         0        0        0        0        0        0            0
  text3 0        0        0         0        0        0        0        0        0            0
  text4 0        0        0         0        0        0        0        0        0            0
  text5 0        0        0.9186418 1.696793 2.298853 2.298853 2.298853 2.298853 0            0
  text6 0        0        0         0        0        0        0        0        1.997823     2.298853
[ reached max_ndoc ... 193 more documents, reached max_nfeat ... 506 more features ]
>
```

Now, to integrate the "syuzhet" package into our project, we first install it and then load it as a library. Understanding the foundation of the sentiment dictionary is crucial for accurate sentiment analysis. Given that "The Apes Of Tarzan" is a literary work, the syuzhet method is the most suitable choice. Designed specifically for analyzing literary texts, syuzhet provides a nuanced understanding of sentiment dynamics within the context of narrative storytelling, making it an ideal approach for extracting emotions and sentiments from the book.

```
> install.packages("syuzhet")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/Sravani/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/syuzhet_1.0.7.zip'
Content type 'application/zip' length 3109566 bytes (3.0 MB)
downloaded 3.0 MB

package 'syuzhet' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Sravani\AppData\Local\Temp\Rtmp0aCYFc\downloaded_packages
> library(syuzhet)
Warning message:
package 'syuzhet' was built under R version 4.3.3
>
```

We extract the text as a data frame and proceed to apply various functions from the "syuzhet" package to analyze our document.

```
> #Extracting the text as data frame
> TOTA_df <- as.data.frame(TOTA_cl[[1]]$content)
> TOTA_df
                                                                TOTA_cl[[1]]$content
1
2                                                                          Out to Sea
3
4
5       I had this story from one who had no business to tell it to me or to
6     any other  I may credit the seductive influence of an old vintage upon
7      the narrator for the beginning of it and my own skeptical incredulity
8          during the days that followed for the balance of the strange tale
9
10    When my convivial host discovered that he had told me so much and that
11    I was prone to doubtfulness his foolish pride assumed the task the old
12    vintage had commenced and so he unearthed written evidence in the form
13      of musty manuscript and dry official records of the British Colonial
14          Office to support many of the salient features of his remarkable
15                                                                   narrative
16
17       I do not say the story is true for I did not witness the happenings
18    which it portrays but the fact that in the telling of it to you I have
19    taken fictitious names for the principal characters quite sufficiently
20              evidences the sincerity of my own belief that it MAY be true
21
22          The yellow mildewed pages of the diary of a man long dead and the
23   records of the Colonial Office dovetail perfectly with the narrative of
24          my convivial host and so I give you the story as I painstakingly
25                          pieced it out from these several various agencies
26
27     If you do not find it credible you will at least be as one with me in
28                acknowledging that it is unique remarkable and interesting
29
30     From the records of the Colonial Office and from the dead mans diary
```

To retrieve the sentences contained within "TarzanOfTheApes.txt," we employ the get_text_as_string() function, which facilitates the extraction of the file's contents as a unified string. This approach allows us to access the entire text seamlessly, enabling subsequent analysis and manipulation.
By consolidating the text into a single string, we ensure a comprehensive overview of the document, laying the groundwork for further processing using various text analysis techniques.

```
> #Reading the file as one large string
> TOTA_asString <- get_text_as_string("C:\\Users\\Sravani\\OneDrive\\Documents\\2024_SPRING\\CS6444 - BIG DATA\\Projects\\Project_3\\eBool
> TOTA_asString
 Tarzan of the Apes   By  Edgar Rice Burroughs              CONTENTS        I  Out to Sea      II  The Savage Home     III  Life and Deatl
n Kind     XIV  At the Mercy of the Jungle      XV  The Forest God     XVI  "Most Remarkable"    XVII  Burials   XVIII  The Jungle Toll
nt Again  XXVIII  Conclusion     Chapter I  Out to Sea   I had this story from one who had no business to tell it to me, or to any other.
he had told me so much, and that I was prone to doubtfulness, his foolish pride assumed the task the old vintage had commenced, and so he
the happenings which it portrays, but the fact that in the telling of it to you I have taken fictitious names for the principal character:
t, and so I give you the story as I painstakingly pieced it out from these several various agencies.  If you do not find it credible you
Clayton, Lord Greystoke, was commissioned to make a peculiarly delicate investigation of conditions in a British West Coast African Colon
mi.  The natives of the British Colony complained that many of their young men were enticed away through the medium of fair and glowing p
upon by their white officers, and they were told that they had yet several years to serve.  And so the Colonial Office appointed John Cla
of little moment to this story, for he never made an investigation, nor, in fact, did he ever reach his destination.  Clayton was the typ
t; his eyes were gray, his features regular and strong; his carriage that of perfect, robust health influenced by his years of army train
tment he was both elated and appalled.  The preferment seemed to him in the nature of a well-merited reward for painstaking and intellige
into the dangers and isolation of tropical Africa that appalled him.  For her sake he would have refused the appointment, but she would n
ry is silent.  We know only that on a bright May morning in 1888, John, Lord Greystoke, and Lady Alice sailed from Dover on their way to
d from the knowledge of men.  Two months after they weighed anchor and cleared from the port of Freetown a half dozen British war vessels
and hence the search was stopped ere it had scarce begun; though hope lingered in longing hearts for many years.  The Fuwalda, a barkenti
uwalda was no exception to the rule.  Her officers were swarthy bullies, hating and hated by their crew.  The captain, while a competent
lse.  So it was that from the second day out from Freetown John Clayton and his young wife witnessed scenes upon the deck of the Fuwalda
r one then unborn such as has never been paralleled in the history of man.  Two sailors were washing down the decks of the Fuwalda, the f
directly behind the captain.  In another moment he would have passed by and this strange narrative would never have been recorded.  But j
nt the scene was ludicrous; but only for an instant.  With a volley of awful oaths, his face suffused with the scarlet of mortification a
-a huge bear of a man, with fierce black mustachios, and a great bull neck set between massive shoulders.  As he saw his mate go down he
Without waiting to rise he whipped a revolver from his pocket, firing point blank at the great mountain of muscle towering before him; bu
he sun.  Words passed between Clayton and the captain, the former making it plain that he was disgusted with the brutality displayed towa
g, strode aft.  He did not care to antagonize an English official, for the Queen's mighty arm wielded a punitive instrument which he coul
rly, and, finding that it bore his weight, turned to Clayton with a word of gruff thanks.  Though the fellow's tone was surly, his words
ral days, nor did the captain accord them more than the surliest of grunts when he was forced to speak to them.  They took their meals in
crew they bullied, and were only too glad to avoid social intercourse with the polished English noble and his lady, so that the Claytons
lminate so soon in bloody tragedy.  There was in the whole atmosphere of the craft that undefinable something which presages disaster.  O
ter the wounding of Black Michael, Clayton came on deck just in time to see the limp body of one of the crew being carried below by four
grew out of the distant horizon, he half determined to demand that he and Lady Alice be put aboard her, for his fears were steadily incre
he obvious ridiculousness of such a request became suddenly apparent.  What reason could he give the officer commanding her majesty's shi
ing to leave the ship to but one thing--cowardice.  John Clayton, Lord Greystoke, did not ask to be transferred to the British man-of-war
g wife a few short hours before, when safety was within reach--a safety which was now gone forever.  It was mid-afternoon that brought th
he came edging along until close to Clayton he said, in an undertone:  "'Ell's to pay, sir, on this 'ere craft, an' mark my word for it,
veste'day, an' three to-day.  Black Michael's as good as new agin an' 'e's not the bully to stand fer it, not 'e; an' mark my word for it
```

Following the initial step of consolidating the text into a single string, we utilize the get_sentences() function from the syuzhet package to parse this unified string into individual sentences. This process breaks down the text into distinct units, allowing for more granular analysis and examination of each sentence's content and structure.

```
> TOTA_sentence <- get_sentences(TOTA_asString)
> TOTA_sentence[1:10]
 [1] "Tarzan of the Apes   By  Edgar Rice Burroughs              CONTENTS        I  Out to Sea      II  The Savage Home     III  Life and Death      IV  The Apes       V  The White Ap
Battles     VII  The Light of Knowledge    VIII  The Tree-top Hunter      IX  Man and Man       X  The Fear-Phantom      XI  \"King of the Apes\"     XII  Man's Reason    XIII  His Ow
t the Mercy of the Jungle      XV  The Forest God     XVI  \"Most Remarkable\"    XVII  Burials   XVIII  The Jungle Toll     XIX  The Call of the Primitive      XX  Heredity     XXI
rture    XXII  The Search Party   XXIII  Brother Men    XXIV  Lost Treasure     XXV  The Outpost of the World    XXVI  The Height of Civilization   XXVII  The Giant Again  XXVIII  Con
er I  Out to Sea   I had this story from one who had no business to tell it to me, or to any other."
 [2] "I may credit the seductive influence of an old vintage upon the narrator for the beginning of it, and my own skeptical incredulity during the days that followed for the balance
le."
 [3] "When my convivial host discovered that he had told me so much, and that I was prone to doubtfulness, his foolish pride assumed the task the old vintage had commenced, and so he
evidence in the form of musty manuscript, and dry official records of the British Colonial Office to support many of the salient features of his remarkable narrative."
 [4] "I do not say the story is true, for I did not witness the happenings which it portrays, but the fact that in the telling of it to you I have taken fictitious names for the princ
uite sufficiently evidences the sincerity of my own belief that it MAY be true."
 [5] "The yellow, mildewed pages of the diary of a man long dead, and the records of the Colonial Office dovetail perfectly with the narrative of my convivial host, and so I give you
instakingly pieced it out from these several various agencies."
 [6] "If you do not find it credible you will at least be as one with me in acknowledging that it is unique, remarkable, and interesting."
 [7] "From the records of the Colonial Office and from the dead man's diary we learn that a certain young English nobleman, whom we shall call John Clayton, Lord Greystoke, was commis
peculiarly delicate investigation of conditions in a British West Coast African Colony from whose simple native inhabitants another European power was known to be recruiting soldiers
my, which it used solely for the forcible collection of rubber and ivory from the savage tribes along the Congo and the Aruwimi."
 [8] "The natives of the British Colony complained that many of their young men were enticed away through the medium of fair and glowing promises, but that few if any ever returned to
 [9] "The Englishmen in Africa went even further, saying that these poor blacks were held in virtual slavery, since after their terms of enlistment expired their ignorance was imposed
ite officers, and they were told that they had yet several years to serve."
[10] "And so the Colonial Office appointed John Clayton to a new post in British West Africa, but his confidential instructions centered on a thorough investigation of the unfair trea
itish subjects by the officers of a friendly European power."
> str(TOTA_sentence)
 chr [1:4312] "Tarzan of the Apes   By  Edgar Rice Burroughs              CONTENTS        I  Out to Sea      II  The Savage Ho"| __truncated__ ...
```

In the next phase, we extract sentiment from these sentences using the get_sentiment() method, which employs various techniques such as "syuzhet", "bing", "afinn", and "nrc" for sentiment analysis. By applying each method to the TOTA character vector, we aim to explore and compare the sentiment extracted using different approaches. This experimentation allows us to assess the effectiveness and nuances of each sentiment extraction method in capturing the emotional tone and nuances within the text, thus providing valuable insights into the overall sentiment conveyed in the document.

Initially, we invoke the get_sentiment() function with the default parameter set to "syuzhet". This choice enables us to employ the "syuzhet" method for sentiment analysis, offering an initial perspective on the emotional tone and sentiment expressed within the text.

```
> #Different sentiment analysis methods in R to analyze the sentiment of the text stored in TOTA_sentence
> TOTA_Sentiment <- get_sentiment(TOTA_sentence,"syuzhet")
> TOTA_Sentiment
  [1]  0.90 -0.25  1.10  2.40  0.10  2.60  1.25  1.15 -3.15  0.55  0.00  3.50  2.00  2.95  1.10  2.75 -0.35 -0.75  1.00  0.00  0.50  0.00  0.00  0.55  0.80  0.00 -2.25 -0.25
 [29] -0.25  0.40 -0.50  0.20 -0.15  0.00 -0.50 -1.00 -0.75 -2.15 -0.50 -0.25 -1.05 -1.50 -1.85 -1.00 -1.20  1.00  1.00 -0.50  0.20 -1.05 -0.30  0.10 -0.85 -2.25 -0.50 -0.40
 [57] -1.15 -1.75  0.25  1.10 -1.00  1.65  0.00 -0.25  0.50  1.05 -0.10  1.15  0.00  0.00  0.00 -0.80  0.80 -1.00  0.00 -1.00  0.40 -1.00 -0.50  0.00  0.05  0.00  0.40  0.75
 [85] -0.75 -1.00 -0.90  0.15  0.00  0.05  0.50 -1.00 -3.90  1.25  0.00 -0.25 -0.75  0.75 -0.50 -0.35 -0.25  0.00 -0.55  0.40 -0.60  0.80  0.80 -1.75 -0.50  0.00 -2.65  1.00
[113]  0.00 -1.65  1.15 -0.10 -1.45  0.25  1.60 -0.10  0.25  0.85  0.50 -0.50 -0.90 -0.90 -0.75 -0.50  0.80 -0.55  0.00  0.00  1.15 -2.65  0.80  0.00  0.60  0.50 -0.50  0.00
[141]  1.30 -0.50  0.80  0.80  0.00 -1.40 -0.75 -3.00  0.50 -0.25 -0.55 -1.00 -0.50  0.40 -0.40 -0.50 -0.20 -0.85 -0.40 -2.80 -3.20 -0.50 -2.00 -1.00 -0.60 -2.45  0.00  0.00
[169]  0.00 -0.75 -0.75  0.50  2.45  0.00  0.00  0.90 -0.50 -2.10 -0.75  0.55 -1.05  0.00  0.00 -0.75  0.55 -2.15  0.65 -0.75 -0.25  1.40  0.00  0.35 -1.50 -2.00  0.75  0.50
[197]  0.80 -0.10  0.75  2.55 -0.50 -2.25 -4.10  0.50  2.95  0.10  0.40  0.80  1.00  0.55  1.35  0.50  0.80 -1.00 -0.85 -0.75 -0.50  1.70 -1.05 -1.65 -1.40  0.00  1.55  0.80
[225]  0.25  0.25 -0.35  0.65  0.25 -1.35 -1.75 -0.75 -1.30 -0.90 -0.80  2.90 -2.00  0.00  0.00  0.25  1.25 -2.00  0.00 -0.50  3.20 -0.25  0.75  0.00  2.20  1.50 -0.70  1.30
[253]  2.00  0.65 -0.50  1.85  0.55  0.00  0.40  0.40  0.40  1.55  4.15 -3.05  1.30  0.50  0.00  0.00  1.10 -0.25  0.00 -1.60 -1.25 -0.75  1.50  0.80  0.10 -0.40 -2.25  0.00
[281]  3.25 -2.50  1.85  2.10  0.05 -0.05  0.00  0.00  2.00  1.35  0.00  2.05 -0.85  0.75  1.25 -1.50  0.00  1.85  3.25 -1.15 -2.50 -0.40  1.25  0.35 -2.45 -1.05  0.25 -1.25
[309]  0.05 -3.35 -1.00  0.40 -1.50  0.50 -2.75 -0.85  0.65  0.00  0.00 -0.45 -2.75 -0.70 -3.00 -0.65  0.00  0.25 -1.50 -0.45 -1.50  1.35  0.00 -1.25 -1.00 -0.60  0.00 -1.75
[337]  3.15 -1.80  0.25  1.50  0.10  0.00  0.50 -0.45 -1.00  0.50 -2.65  2.20 -3.30  1.60  0.85 -0.50  0.00  0.00 -0.25 -2.85 -8.15 -1.50 -2.05 -1.50 -1.40 -0.75 -1.20 -2.00
[365]  0.35 -2.25 -0.40 -3.75 -1.40 -1.95  0.10 -0.90  0.50  0.85 -2.10  0.50  0.40 -0.05  1.80  0.25  0.50 -2.25  0.40 -0.90 -0.25  0.00 -0.50 -0.90  1.35 -0.40  0.00 -0.50
[393] -1.85 -2.85  0.00 -0.85 -0.55 -0.60 -1.50 -1.60 -0.25 -2.00  1.25 -0.05  3.75  0.15 -1.75  0.25 -0.70 -2.30 -2.15 -0.30  0.30  0.35 -2.65 -0.40 -0.05  1.15 -1.40  0.85
[421] -0.50 -1.00 -1.35 -0.40 -3.00 -0.90 -0.80 -0.90  0.85  0.30  0.00 -0.15 -1.40  0.90 -0.35 -0.50 -0.45  0.30 -0.25 -0.90  3.15 -1.20  2.80 -0.20 -0.15  0.35  0.50  0.75
[449]  0.75 -0.75  0.00 -0.75  0.00 -0.10  1.20  1.75  1.80  0.25  0.35  0.05  3.45  3.25  0.50  0.80  0.50 -1.75 -0.75 -1.20  0.50  0.40 -1.50 -0.50 -1.40  0.00 -0.20  2.50
[477] -0.50  1.25  0.75  0.00  0.75  0.40 -0.50 -0.75 -0.05  0.00  1.10  0.75  1.55 -0.25  0.60 -2.00  0.50 -4.75  1.45 -0.30 -1.75 -1.30  0.80  0.25  0.00  1.15 -0.75 -0.95
[505] -0.75 -1.10 -0.75 -3.35  1.00  1.00  1.55 -2.35  0.25 -1.75 -0.25 -0.75 -1.50  0.00  2.45  0.00 -0.90  0.00  0.50 -0.10 -1.55  0.15  2.15 -2.10  0.60  0.80  0.25 -0.85
[533]  0.25 -0.10 -0.80 -1.90  0.55  1.00 -0.75 -1.25 -3.25  0.00  0.15  2.35  0.40  1.00  0.00 -0.60  0.10 -0.20  1.35  0.50  1.50 -1.80  0.00 -0.25 -0.85 -1.05  1.10  0.15
[561]  0.00 -0.15 -2.75 -1.75  0.25  0.35  0.80  0.20  1.10  0.30  0.60  0.25  0.50  0.60  0.00  0.25 -2.10  0.00  0.10 -1.00 -0.65  0.60  1.00  0.00 -0.35  0.00 -0.25 -1.00
[589] -0.90  0.75  1.85  0.40  0.30 -0.75 -0.25 -2.40  0.35  0.20 -1.75 -4.70 -1.85  0.10  0.15 -0.75 -3.30  0.75  1.10 -2.50 -1.75 -3.00 -3.30 -1.50 -1.25 -1.00 -2.10 -0.50
[617] -0.25 -0.95 -0.25 -0.20 -0.55 -3.00 -0.25 -1.55 -2.25  1.90  1.10  0.90  0.65 -0.20 -3.00  1.65 -1.30  0.00  1.00  0.85  0.50  1.00  0.75  0.35 -1.05 -0.10  0.25  0.00
[645] -0.60  0.00  0.60  0.25 -0.35  0.90  0.00  1.35  0.20  0.00 -0.25  0.25  1.70  0.00  0.00  3.10  1.00  1.95  2.05 -0.90  0.00 -0.25  0.00  0.75  4.20  0.60  0.15 -0.15
[673]  1.10 -1.40  1.50 -1.65 -0.25  0.00  0.70 -4.80  0.75  2.30  0.00  0.25  0.25 -0.40 -0.15 -0.15 -0.10 -1.75  0.00 -0.05 -0.50 -0.25  0.40 -0.25 -0.45 -0.10 -2.15 -2.00
[701] -1.50 -1.10 -4.35 -4.35 -1.65 -1.10 -0.25 -0.15 -3.15 -0.40 -1.00  1.15 -2.85 -0.65 -0.30  0.75  0.40  0.25 -2.25 -0.25  2.65 -1.15 -0.40 -0.45 -0.25  0.95  0.15 -4.20
[729] -1.00  0.00 -0.55  0.85  0.50 -2.50 -0.75 -1.20 -0.75  1.20  1.45 -0.25  0.00 -0.25 -0.40 -0.25  0.25  0.55  0.00 -0.05 -1.10 -2.85 -0.50  0.00 -1.00 -0.50  0.50  0.75
[757] -1.00 -2.20  0.35 -0.90  1.20  1.35 -1.40  1.65  1.30 -1.35 -1.40 -4.40 -0.90  0.80  0.50  0.25 -0.25 -0.80  0.75  1.10  0.85  1.10 -0.85  1.55 -1.90  0.00 -0.75 -0.25
[785]  1.25  1.25 -2.30 -0.30 -0.75 -2.30  0.00 -0.25 -1.75 -3.15 -1.50 -3.75  0.00  1.40  0.60  0.00 -1.50 -0.40  0.00  0.50  0.00 -0.50  1.85  0.15  1.80  0.05 -0.25  2.15
[813]  0.00 -0.75  2.20  0.35  0.60  0.50  0.70 -0.65  0.00 -0.40  0.35 -0.50 -0.50 -1.60  0.00 -1.10  1.30 -0.15  0.50 -1.00 -2.00 -1.00 -1.90 -1.70 -0.75 -1.50  0.55 -1.65
[841]  1.35 -0.25 -1.30 -0.80 -0.90  0.00  0.00  0.00  0.00  1.25 -0.15  0.00  0.00  0.00  0.00  0.00  0.00  0.10  0.30 -0.75 -0.10  0.00  0.00  0.00  0.60  1.35 -3.20 -1.75
[869] -0.45 -0.35 -1.35  1.65 -1.60  0.80 -1.70  0.10  0.90 -0.70  0.60  0.00  0.40 -0.25 -0.75 -1.50 -1.55  0.00 -0.25  0.55  0.00  0.00  0.05 -0.75  0.60  0.10 -0.80 -0.50
[897]  0.00 -2.50 -2.50  0.05  0.35 -0.40 -2.05 -0.25 -0.85 -1.35 -2.20 -0.40 -1.50 -1.00 -2.75  1.25 -1.25  1.25  1.75  1.50 -2.70 -0.25  0.35  0.00 -2.55 -1.15 -0.50  0.35
[925]  0.00  0.00  0.00  0.00 -0.50  0.75 -0.85  0.25  0.80 -0.25 -1.25  0.75  1.00  0.00 -0.60 -0.35  0.30 -0.25 -2.10 -0.10  0.00 -0.65 -0.50  0.50  0.05 -0.75 -0.70 -0.60
[953] -1.40 -0.50 -2.00  1.20 -0.90  0.50  0.50  0.00 -1.75  1.50 -3.20 -1.75  0.00  0.75  0.00 -2.35  0.00  0.00 -1.90  0.50  0.80  0.10 -0.10  0.00  0.40  0.85 -1.25  0.60
[981] -0.25 -1.60  0.10 -2.75 -0.95  0.15  0.00 -0.25 -0.40  0.50  0.50  0.80  0.50  0.25 -0.75 -1.50  0.50 -0.65 -1.25  0.00
[ reached getOption("max.print") -- omitted 3312 entries ]
```

When experimenting with the "bing" method, the results yield a different perspective on sentiment analysis. The sentiment extraction process utilizes the Bing lexicon, potentially offering a distinct interpretation of the emotional content present in the text.

```
> TOTA_Bing <- get_sentiment(TOTA_sentence,"bing")
> TOTA_Bing
  [1] -3 -2  1  1  0  3 -2  3 -3  1  0  5  3  2  0  3 -2 -1  0  1  1  0  0 -1  0  0 -3 -1 -1  0 -1  0  0  0 -1 -1 -1 -3 -1  0 -2 -2  0 -2  0  0  0 -2  1 -1  0 -1  0 -3 -1 -1 -2
 [58] -2  0  1 -1  0  0  0  1  0  0  1  0  0  0  0  0  0  0  0  0  0 -1  0  1  0  0  1  0 -1 -1  1 -1  0  0 -1 -4  1  0 -1 -1  1 -1 -1  1  0 -1  0  0  1  1 -1 -1  0 -2  0  0 -1
[115]  1  1 -3 -1  1 -1  1 -1  1 -1 -2  0 -1  0  0 -1  0  0  1 -3  1  0  1  1 -1  0  1  0  0  0  0 -1 -2 -4  2  0 -1 -2 -1  0 -1 -1  0  0 -2 -1 -1 -3 -1  0 -4  0  0  0 -1 -1
[172]  0  0  0  0  1 -1 -2 -2  1 -1  0  0 -1  1 -2  0 -1  0  1  0 -1 -3 -1  1  2  1  1  1  3 -1 -4 -6  1  2  0  2  0  1  2  3  1  1 -1 -1  0  0  1 -1 -3 -1  0  2  0  0  1 -2 -1
[229]  1  1 -3 -1 -1  1  0  3  0  0  0  1  1 -2  1  0  1 -2  1  0  2  2 -2  1  2  1  0  2  1  0  2  1  1 -1  4 -1  1  1  0  0  1 -1  0 -2 -2 -1  2  1  0  0 -2  1  5 -1  1  2  0
[286]  0  0  0  0  0  0  2 -1  1  2 -4  1  2 -1 -2 -1 -2  2  0 -4 -2  0 -1  0 -5  0 -1 -2  1 -3  0 -1  0  0 -1 -1 -1 -3 -2  1  1 -3 -1 -2  2 -1  0 -2  0 -1 -4  4 -3  0  3  0  0
[343]  0  0  0  2 -1 -2 -2  0  1 -1  0  0  0 -4 -9 -1 -1 -2 -1 -1 -2 -2 -3 -3  0 -5 -1 -2  0 -1  0  1 -3  1  0 -2  4  1  0 -2  0 -2  0  0 -1 -1  1 -1  0 -5 -4 -2  0 -2 -2  0 -2
[400] -2  0 -1 -1 -3  4  1 -2  0 -1 -4 -4  0 -1 -1 -1  0 -1  1 -2  0 -1  0 -2  0 -3 -2  0  0  0  1  0 -1 -4 -1 -2 -1 -2  0  0 -3  3 -1 -1  0 -1 -1  1  1  1 -1  1 -1  0  0  0  3
[457]  2  0  1 -1  5  2  0 -1  1 -2 -1 -3  1  0 -2 -1  0  0 -1  3  0  1  1  0  1 -1 -1 -1  1  0  1  2  2  0  1 -2  1 -7  1 -1 -3 -1  1  1  0  2 -1 -2 -2 -3 -1 -3  2 -1  1 -1  0
[514] -2  0  1 -1  0  2  1 -2  0  1 -1  0 -1  2 -2  0 -2  1 -1 -1  1 -1 -1  0  0 -1  0 -3  1  0  2  0  2  0  1  0  0  2  1  0 -1  0  1 -3  0  1 -2  0 -1 -3 -3  0 -2  0  1  0 -1
[571]  0  0  0  0  0  0 -5  0  0 -1 -1  1  1 -1 -2  0  0 -1 -2  1  2  1 -1 -1 -2 -3  1  1 -2 -3 -2  0 -1  2 -6  1  3 -1 -1 -4 -3 -1 -2 -1 -2 -1  0 -3  0 -2 -1 -4  0  0 -2  3  2
[628]  2  0  0 -3  4 -2  0  0  0 -1  1  1  3 -7 -1 -1  0 -1 -1  0 -1 -3 -1  0  3  0 -1 -2  0  0  0 -1  3  0  2  0 -2  1  0  0  2  3  0 -1 -3  1 -3  3 -2 -1  0  0 -7  1  2  0  0
[685]  0 -2 -2  0  0 -2  0 -2  1  0  0 -1  0 -1 -1 -3 -1 -1 -3 -4 -1 -1  0 -1 -2 -3 -2  1 -4 -3 -1 -2  1  0 -1  0  1 -2  2 -1 -1  1  0 -6 -1  0  0  1  1 -3 -1 -2 -1  0  2 -2  0
[742]  0  0  0  0  1  0  0 -2 -5 -1  0  0  0  1  1 -2 -3  0 -2  0  0 -2  0  1 -1  0 -6 -2 -1  1  1 -1 -1  1  1  1 -1 -1  1 -3  0 -1 -1  1  1  0  0 -1 -4  0 -1 -3 -4 -2 -4  0  0
[799]  1  0 -2 -1  0  0 -2 -1  0 -1  1 -1  0  2  0 -1  1  0  2  0  2 -1  0  1  1  0 -1 -1  0 -2  2 -1  1 -1 -2 -1 -2  0 -1  0  2 -1  1 -1 -2  1  0  0  0  0  0  2 -1  0  0  0 -1
[856]  0 -1  0 -1 -1 -2  0  0  1  1  0 -5 -2  0 -2 -2  2 -2  0 -3 -1  1 -2  0  0  0  0 -1 -4 -2  0  0  0  0  0  1  0  2  2  0 -1  0 -3 -3 -1  1 -1 -2 -2 -1 -1 -3 -1 -2 -1 -2  1
[913] -2  1  2  2 -4  1 -2  0 -2 -2 -1  1  1  0  0  0 -2  1  0 -1  1 -1  0  1  0  0 -1  2  0 -2  1  0 -2 -1  0 -3 -1 -2  0 -1  1  0  0  0 -1  0 -1 -1  1 -4 -3  0  1  0 -1  0
[970] -2 -1  0  1  1  2  0  0  1 -2  0 -1 -1  1 -3  0  0  0  0 -1  1  0  2 -1 -1 -1  1  0 -2  0
[ reached getOption("max.print") -- omitted 3312 entries ]
```

Utilizing the "afinn" method for sentiment extraction produces yet another set of results, providing an alternative viewpoint on the emotional tone conveyed within the text.

```
> TOTA_Afinn <- get_sentiment(TOTA_sentence,"afinn")
> TOTA_Afinn
  [1]  -2  -3   5   4   2   4  -2   1  -8   0   1   9   6   4   1  10   0  -2   1   0   1   0   0   0   0  -1  -6   2   0   0   0  -1   0   0  -1  -3  -3  -1   0   4  -1  -3   1
 [44]  -1  -1   3   1   1   0   0  -1   2   4  -5  -2   0  -2   0   1   0   0   1  -2   3   3  -1  -1   3   0   0   0   0   1   0   0   0   0   0  -2   0  -4   0   0   2  -2  -3
 [87]  -3   0   3  -2   2   0  -5   3  -1   0  -2   2  -2  -3   3   0  -2   0   0   0   0  -2   0   0  -4   2   0  -2   3  -2   0   5   2  -3  -1   3   0  -1  -5  -2  -2  -2   2
[130]  -3   0   0   5  -3   0   0   0   3   1  -1   1  -1   1   0   0   0  -1 -10   5   0   0  -3   0  -4   0  -2  -4   0   0  -5  -6   0  -3  -4  -2  -3   0   0   0  -3  -3   2
[173]   4   0   0   2  -1  -6  -1   0  -3   0   0  -2   0  -2   3  -2   0   0   0   0   0  -5   3   1   0   4   4   2  -1  -2 -11   0   4   0   0   2   1   0   3   0   0  -2   0
[216]  -1  -1   2  -1  -5  -9   0   3   0   2   2  -6   2   1  -3  -3  -2   0  -4  -2   6  -2   0   0   0   2  -3   0  -3   7  -2   3   0   2   4  -2   2   6   1  -1   5  -1   0
[259]   4  -1   0   1  15  -7   4   4   0   0   3   0   0   0   0  -2   5   0   2   0  -1   3  13  -6   0   6   2  -1   0   0   5  -1   0   7   2   3   3  -4   3   7   3   0  -7
[302]   0   5   3  -5   3   3  -1  -3  -6  -2   0  -5   3  -6   1   0   0  -4   1  -7  -2  -6  -3   2   0  -3  -2  -2   3   0  -2  -5  -2   0  -5  12  -5   2   3   0   0   1   0
[345]   0   5  -1  -1  -3   0   2  -1   0   0  -1  -2 -19  -1  -2  -2   1   1  -2  -2  -7  -4   3  -5  -2  -3   1   2   3  -1  -4   3   0  -3   9   4   5  -4   0  -5  -3   0   3
[388]  -2   2  -3   1   1  -5  -6  -1   0  -5   0  -3  -1   1   3   0  -5   9   1  -1  -1   1  -3  -3   3  -2   0   1   0   1   7   0   3  -2   0   0   0  -4  -5   0  -6  -2   0
[431]   0  -5  -6  -2   0   0  -5   1   0  -4   4  -4   4  -2  -1   2   3   1   3  -2   1  -1   0   1  -1   7   7   0   2   1   3   2   2   2   3   1   0  -6   1   0  -3  -2  -3
[474]   0   0   4  -2   3   2   0   3  -1   0  -3   1   1   3   4   8   0   0  -5   0 -10  -1  -4  -5  -7  -2  -2   0   2  -2  -4  -3  -5  -1   0   0   4   2  -3  -1  -6   1   0
[517]  -4   0   3   1   2   0   3  -5   2   2  12  -4  -1  -3   2   0   1  -2   0  -2   2   0   0  -3 -10   2   1   2   0   3   0  -1   3   3   7   2   2  -6   0  -3  -3   0   2
[560]   0   0  -3  -7  -7   0  -1   2   0   0   4   0   0   2   0   0   0  -3   0   0  -4  -2   3   2   1   2   4   1  -2  -2   3   2   4  -2   0  -2  -4  -1   3  -4  -9  -7  -2
[603]   0  -2  -4   4   6  -3  -4  -5  -3  -1  -3  -3  -5  -2   0  -2   0   1  -2  -5  -1  -1  -4   8   1   2  -2  -2  -1   7  -2   0   0   1   1   3   1   6  -4  -1   0   0  -2
[646]   0   1   0  -1   2   0   5   1   0   0   1   3   0   0   9   5   3   6  -3   2   0   0   2   8   4   0  -1   1  -5   1   0   0   0  -4 -14   1   3   0   2  -1  -1  -3  -4
[689]   4  -2  -3  -2  -1   2   0   0   2   4  -2   1   0   1  -4  -6  -5  -2  -1  -1   0  -7   0   0  -3   1   1  -2   4  -3   3   0   6   2  -1  -4   0   1   0  -8  -3   2   0
[732]   2   0  -2  -3   2  -1   1   3  -1   0   0   0   0  -2   8   0   1   2  -6   0   0  -2   3   2   0  -2  -5   2  -5   0  -1  -3   2   0  -2   2 -10   0   1   3   3  -3   0
[775]   0   4  -4  -2   1   2   1   0  -4   0   3   3  -4   0  -3  -8   3   0  -2  -6  -2  -5   0   2   0  -1  -3   0   0   2   3   0  -6  -2   3  -1  -2   1   0  -1   2  -1   3
[818]   1  -2   0   0   0   5  -2  -2  -2   0  -2   3  -1   1   0  -8   0  -2   0  -2  -4   2  -5   4   1   0  -1  -3   0   0   0   0   1   0   0   1  -1   2   0   1   1  -1  -1
[861]  -3   2   0   3   1   2  -6  -4   2   3   0   3  -4   0  -2   0   0  -7   1   0   0  -1  -2  -3  -2   0   0   0   0  -3  -2   1   1   1   0  -2   0  -5  -4  -3   2  -3   0
[904]   1  -4  -1  -2  -3  -5  -4   1   7  -3   5   5   5  -6   0  -1  -1  -2   0  -1  -2   0   0   0   0   1   2   0  -2   0   1  -5   4   0   0   0  -2  -1   0   0  -2   2   0  -3
[947]  -2   2   1   0  -6  -2   0  -2   3   1   2   0   3  -3  -7   1  -9  -3   1   0  -1  -5  -1  -5  -2   3   1   3   1   0   0  -3   1   0   0  -2  -1  -4  -5   2  -2   0   0
[990]   0   3   0  -2   0  -5  -2   0  -2  -2   0
[ reached getOption("max.print") -- omitted 3312 entries ]
>
```

When employing the "nrc" method for sentiment extraction, the resulting sentiments provide additional insights, potentially capturing nuanced emotional nuances within the text.

```
> TOTA_nrc <- get_sentiment(TOTA_sentence,"nrc")
> TOTA_nrc
  [1]  7  2  2  2  1  4  3  3 -3 -2  1  3  2  3  2  3  0 -1  0 -1 -1  0 -1  1  2  0 -1  1 -1  1 -1  1  0  1  0 -1 -1 -2 -2 -3  1  0  1  0 -2  0  1 -1 -2  0  1  2  1 -4  0 -1 -2
 [58] -2  2  4  0  0 -2  1  1  5  1  2  0  0  0 -1  1 -1  0 -1  1 -1  1  0 -1  0  5  1 -1  0  0  1  1  0  1  0 -4  2 -1  1 -1  1  0  0  1  1 -1  1  0  0  0 -2 -1  1 -2  0  1 -1
[115]  2  2  1  2  0  0  0  0  1  0 -1 -1 -1 -1  0 -1 -1  0  0  0  0  0  0  0 -1 -1  1 -1  0  3  0  0  1 -3  2 -1 -2  1 -1  3 -1  0  0 -1  0 -3 -3 -1  0 -2  1 -3  1  0  0 -2 -1
[172]  1  4  0  0  2  1  0 -1  0  0  0  1 -1 -1 -3  0  0 -1  1  0  2 -3 -6  2 -1  0  1  2  2 -1 -3 -8  1  3  1 -2  0  3  0  0  0  0 -1 -3 -2 -1  3 -1 -3 -2 -1  0  0  0  1  2  0
[229] -1 -4 -3 -1 -1  0  0  4 -2  0  0 -1  1 -1 -1  0  3  0  1  0  2  1 -4  0  1  0 -1  3 -2  0 -1  0  0  1  7 -2  0  0 -1  0  2  0 -1  1 -3 -1  1  0  2 -1 -2 -1  1 -2  2  2 -1
[286]  0 -1  0  1  0 -1  3 -1  0  1 -4  1 -1  1  0 -2 -1  1  2 -3 -2  0 -2  0 -5  0  1 -2  0 -5  0  0  0  0  1 -4 -2 -2  0  1  0 -3  0 -3  2  0 -1 -2  0  0  0  5  1  1  3  1  0
[343]  1  2  1  1 -2  1 -4  2  2  1  0  0  1 -3 -7 -4  2 -1 -3 -1  1 -1  1 -2  1 -4  0 -2  1  0  2  2 -3  0  2 -1  5  0  0 -2  1  0 -1  2  0  1  0  1  0 -1  0 -5  0  1 -3  0  0
[400]  0 -2 -3  3  1  7 -1 -1  2  0 -1 -4  1  0 -1 -2 -1 -1 -1 -2  0  0 -1 -1  0 -3 -1 -2 -3  1 -2  1  2 -1  1  0  0  0  0 -2 -2  4 -1  3  1 -2  1  0  1  1  0  0 -1  0  1  3  2
[457]  3 -1  1  1  3  4  1  0  0 -2 -2 -3  1  1  1  0 -1  0  1  3 -1  2  1  0  0  1 -1 -1 -1  0  2  0  2 -1  1 -3  0 -5  0 -1 -2  0  0  1  1  4  0 -2  0 -1 -1 -4  2  0  1 -3 -1
[514] -1 -1  0 -2  1  3  0  1  0  0  1 -2 -1  1 -1  2  1  1 -1  3  1  0 -3 -1  2 -1 -1 -3 -1  2  2  1  2  0  1  0 -1  2  2  3 -2  0 -2  0 -1  1  3  0 -1 -2 -2  1  1  0  0  0  2
[571]  0 -1  1  1  0  1 -1  0  1 -2 -1  0  0  0 -2 -1 -1 -2 -2  0  2  1 -2 -1 -1 -3  0  0 -3 -3 -4  0  2  1 -2  0  1 -4 -1 -4 -5 -3 -1 -1  0  0  0 -1  0 -2 -1 -2 -1 -1 -5  1  1
[628]  0  1  1 -3  2  0  0  1  1  1  2  2  1  2  1 -1  0  0  0  1 -1  2  2  0  2 -1  0  0  2  2  0  1  3  1  1  4 -2  1  0  0  1  6  1 -2  0  1 -2  1 -2  0  0  1 -6  1  0  1  2
[685]  1  2 -3  1  0 -1  0  1  2 -1  1 -1  1 -3 -3 -3 -1  0 -1 -5  0 -1 -1  0 -2 -2 -2  1 -5 -3  0  0 -1 -1 -4  0  7 -2  1 -1  0  4 -2 -6 -2 -1  0 -1  0 -2 -1 -1 -1  1  2  0 -1
[742]  1 -2  1  2 -1  0 -2 -2 -3 -1  0  0  0  1  1 -1 -1  2  1  3  0 -1  0  0 -2 -1 -7 -3  2  0  0  0 -1  0 -1  3  1 -2  1 -2  0  0  0  2  1  0  1 -1 -3 -1 -1 -4 -6 -3 -3  0  1
[799]  2  0 -2  1  0  1 -1 -2  1 -1  1  0  2 -1 -1 -1  0  0 -1  2  0 -1  0 -1 -1 -1  0 -3  0 -2  1  3  1  1 -2 -2 -1 -2 -1 -3  0 -3  4  0 -1 -1 -3  0  0  0  0  1  0  0  0  0  1
[856]  0  0  2 -2 -1  0  0  1 -1  1  2 -3 -2  0 -2 -2  1  0  0 -3  1  2 -1  2  0  1  0 -1  1 -3  0 -1  0  0  0 -1 -2  0  0 -2  0  0 -3 -1  0 -1 -1 -3 -3 -2 -1  0  1 -2 -2 -2  0
[913] -2  2  1  2 -1  0  1  0 -1 -2 -1  0  0  0  0  0 -4  1  1  0  0 -1 -1  0  3  0 -1 -2 -3  1 -2  0  1  1  1  1  0 -2 -1  0 -1  0 -2  2 -1  4  0 -1 -2  2 -4 -1  0  2  1 -2  1
[970] -1 -2  1  0  0 -1  1  1 -1 -2  1  0 -1  1 -3 -2 -1  0 -1 -1  1  0  0  1 -1  0 -2  1 -1 -1  0
[ reached getOption("max.print") -- omitted 3312 entries ]
```

We can examine the sentiment dictionaries used by each method to gain a deeper understanding of how sentiment is interpreted and extracted from the text.

```
> #Sentiment dictionary for "syuzhet"
> TOTA_Dictionary <- get_sentiment_dictionary()
> TOTA_Dictionary
                word value
1            abandon -0.75
2          abandoned -0.50
3          abandoner -0.25
4        abandonment -0.25
5           abandons -1.00
6           abducted -1.00
7          abduction -0.50
8         abductions -1.00
9           aberrant -0.60
10        aberration -0.80
11             abhor -0.50
12           abhorred -1.00
13          abhorrent -0.50
14             abhors -1.00
15          abilities  0.60
16            ability  0.50
17             abject -1.00
18             ablaze -0.25
19           abnormal -0.50
20              aboard  0.25
21             abolish -0.50
22         abominable -0.50
23         abominably -1.00
24          abominate -1.00
25        abomination -0.50
26              abort -0.50
27             aborted -0.80
28            abortion -0.80
29            abortive -1.00
30              aborts -0.60
31              abound  0.25
32             abounds  0.25
33            abrasion -0.60
34            abrasive -0.50
35            abrogate -0.40
36              abrupt -0.25
37            abruptly -0.25
38             abscess -0.50
39             abscond -1.00
40             absence -0.50
```

```
> #First part of the Bing sentiment Dictionary
> TOTA_Dictionary_Bing <- get_sentiment_dictionary("bing")
> TOTA_Dictionary_Bing
                    word value
1                     a+     1
2                 abound     1
3                abounds     1
4              abundance     1
5               abundant     1
6             accessable     1
7             accessible     1
8                 acclaim     1
9               acclaimed     1
10            acclamation     1
11               accolade     1
12              accolades     1
13         accommodative     1
14          accomodative     1
15            accomplish     1
16           accomplished     1
17        accomplishment     1
18       accomplishments     1
19               accurate     1
20             accurately     1
21             achievable     1
22            achievement     1
23           achievements     1
24             achievible     1
25                 acumen     1
26              adaptable     1
27               adaptive     1
28               adequate     1
29             adjustable     1
30              admirable     1
31              admirably     1
32             admiration     1
33                 admire     1
34                admirer     1
35               admiring     1
36             admiringly     1
37                adorable     1
38                   adore     1
39                  adored     1
40                  adorer     1
41                 adoring     1
42               adoringly     1
43                  adroit     1
44                 adroitly     1
45                 adulate     1
```

```
> #First part of the Afinn sentiment Dictionary
> TOTA_Dictionary_Afinn <- get_sentiment_dictionary("afinn")
> TOTA_Dictionary_Afinn
                word value
1            abandon    -2
2          abandoned    -2
3           abandons    -2
4           abducted    -2
5          abduction    -2
6         abductions    -2
7               abhor    -3
8           abhorred    -3
9          abhorrent    -3
10            abhors    -3
11         abilities     2
12           ability     2
13            aboard     1
14           aborted    -1
15            aborts    -1
16           absentee    -1
17          absentees    -1
18            absolve     2
19           absolved     2
20           absolves     2
21          absolving     2
22           absorbed     1
23              abuse    -3
24             abused    -3
25             abuses    -3
26            abusing    -3
27            abusive    -3
28             accept     1
29         acceptable     1
30         acceptance     1
31           accepted     1
32          accepting     1
33            accepts     1
34         accessible     1
35           accident    -2
36          accidental    -2
37       accidentally    -2
38          accidents    -2
39            acclaim     2
```

```
> TOTA_Dictionary_nrc <- get_sentiment_dictionary("nrc")
> TOTA_Dictionary_nrc
       lang           word sentiment value
1   english           abba  positive     1
2   english        ability  positive     1
3   english abovementioned  positive     1
4   english       absolute  positive     1
5   english     absolution  positive     1
6   english       absorbed  positive     1
7   english      abundance  positive     1
8   english       abundant  positive     1
9   english       academic  positive     1
10  english        academy  positive     1
11  english     acceptable  positive     1
12  english     acceptance  positive     1
13  english     accessible  positive     1
14  english       accolade  positive     1
15  english  accommodation  positive     1
16  english  accompaniment  positive     1
17  english     accomplish  positive     1
18  english   accomplished  positive     1
19  english accomplishment  positive     1
20  english         accord  positive     1
21  english accountability  positive     1
22  english    accountable  positive     1
23  english     accredited  positive     1
24  english        accueil  positive     1
25  english       accurate  positive     1
26  english            ace  positive     1
27  english        achieve  positive     1
28  english    achievement  positive     1
29  english acknowledgment  positive     1
30  english        acquire  positive     1
31  english      acquiring  positive     1
32  english        acrobat  positive     1
33  english         action  positive     1
34  english         actual  positive     1
35  english         acuity  positive     1
```

Each of these methods utilizes different criteria and rules to assign sentiment to words, resulting in dictionaries of varying lengths. This diversity in approach contributes to the differences observed in sentiment analysis results across the methods.

In order to capture the overarching emotional tone of the text, we compute the cumulative sentiment scores by summing the values of the sentiment vectors generated by the various sentiment analysis methods. This summation provides us with a quantitative measure of the collective emotional valence embedded within the text.

```
> #Finding the sum of sentiment scores for sentiment analysis results obtained from the 'syuzhet' package to get a measure of the overall emotional valence in the text
> TOTA_Sum <- sum(TOTA_Sentiment)
> TOTA_Sum
[1] -469.75
> #Sum of sentiment scores obtained from the Bing method.
> TOTA_BingSum <- sum(TOTA_Bing)
> TOTA_BingSum
[1] -1183
> #Sum of sentiment scores obtained using Afinn method
> TOTA_AfinnSum <- sum(TOTA_Afinn)
> TOTA_AfinnSum
[1] -909
> #Sum of sentiment scores obtained using nrc method
> TOTA_nrcSum <- sum(TOTA_nrc)
> TOTA_nrcSum
[1] 0
```

- The analysis of sentiment scores generated from multiple methods provides insight into the emotional valence portrayed in the text. The cumulative sentiment score from the 'syuzhet' package indicates an overall negative emotional tone with a score of -469.75.
- Sentiment scores obtained through the Bing method reveal a notably negative sentiment, evidenced by a sum of -1183.
- Sentiment scores derived from the Afinn method also suggest a predominantly negative emotional sentiment, albeit less intense, with a sum of -909.

- The sentiment analysis results from the NRC method yield a neutral overall sentiment, as indicated by a sum of 0.

These findings collectively offer a nuanced understanding of the emotional dynamics embedded within the text, highlighting variations in sentiment across different analytical approaches.

Similarly, we can find the mean of sentiment scores.

```
> #Finding the mean of sentiment scores for sentiment analysis results obtained from the 'syuzhet' package
> TOTA_Mean <- mean(TOTA_Sentiment)
> TOTA_Mean
[1] -0.1089402
> #Mean of sentiment scores obtained using the Bing method
> TOTA_BingMean <- mean(TOTA_Bing)
> TOTA_BingMean
[1] -0.2743506
> #Mean of sentiment scores obtained using the Afinn method
> TOTA_AfinnMean <- mean(TOTA_Afinn)
> TOTA_AfinnMean
[1] -0.2108071
> #Mean of sentiment scores obtained using the nrc method
> TOTA_nrcMean <- mean(TOTA_nrc)
> TOTA_nrcMean
[1] 0
```

By examining the summary() results of the sentiments derived from these dictionaries, we can gain insights into the distribution of sentiment throughout the text. This analysis helps us understand the overall emotional tone of the book as captured by each sentiment dictionary, enabling a deeper understanding of the book's thematic elements and narrative structure.

```
> #Summary results of the sentiments generated from these dictionaries to understand the distribution of sentiment
> summary(TOTA_Sentiment)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.1500 -0.6000  0.0000 -0.1089  0.5000  5.6000
> summary(TOTA_Bing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-9.0000 -1.0000  0.0000 -0.2744  0.0000  6.0000
> summary(TOTA_Afinn)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-20.0000  -1.0000   0.0000  -0.2108   1.0000  16.0000
> summary(TOTA_nrc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     -8      -1       0       0       1       7
```

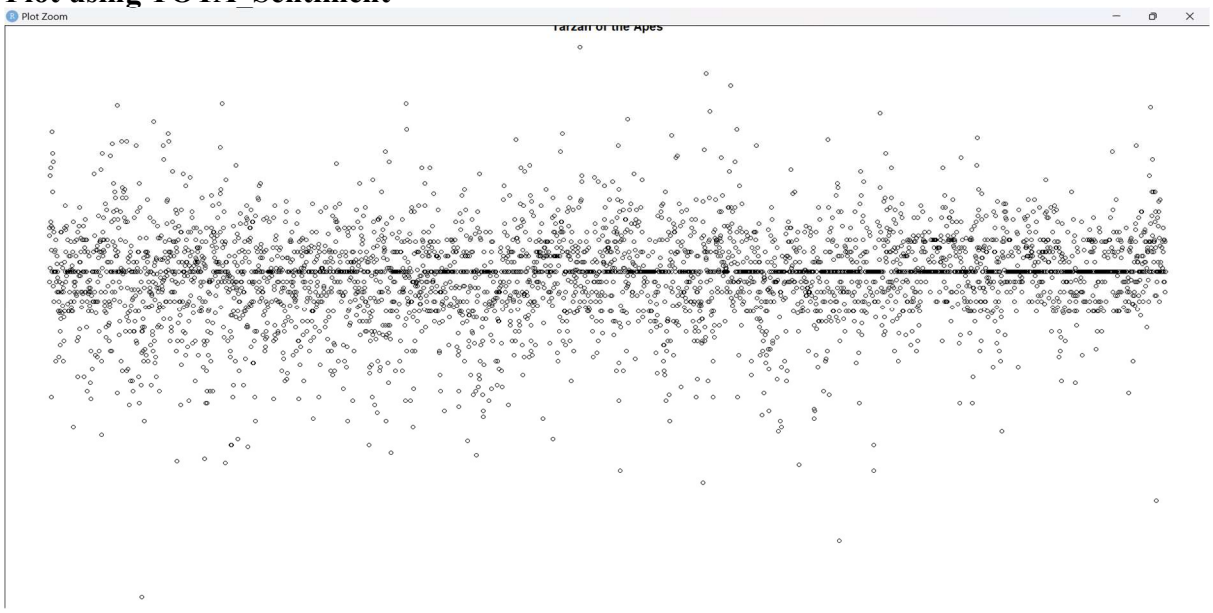
- For the sentiment scores generated by the 'syuzhet' package, the distribution ranges from -8.15 to 5.6, with a mean sentiment score of -0.1089. This indicates a relatively balanced distribution of positive and negative sentiment throughout the text.
- When using the Bing method, the sentiment scores range from -9 to 6, with a mean score of -0.2744. The distribution suggests a slightly more negative sentiment compared to the 'syuzhet' analysis.
- Similarly, the Afinn method yields sentiment scores ranging from -20 to 16, with a mean of -0.2108. This method shows a wider range of sentiment, including more extreme positive and negative scores compared to the other methods.
- Lastly, the NRC method produces sentiment scores ranging from -8 to 7, with a mean score of 0. This suggests a relatively neutral sentiment distribution, with sentiments balanced around zero.

## PLOTTING TO UNDERSTAND HOW THE SENTIMENT OF TEXT CHANGES FROM THE BEGINNING OF THE TEXT TILL THE END

These plots visualize the change in sentiment throughout "Tarzan of the Apes" using different sentiment analysis methods. Each plot represents the emotional valence over the course of the narrative.
The x-axis indicates the progression of the narrative, while the y-axis represents the emotional valence as determined by the respective sentiment analysis method.

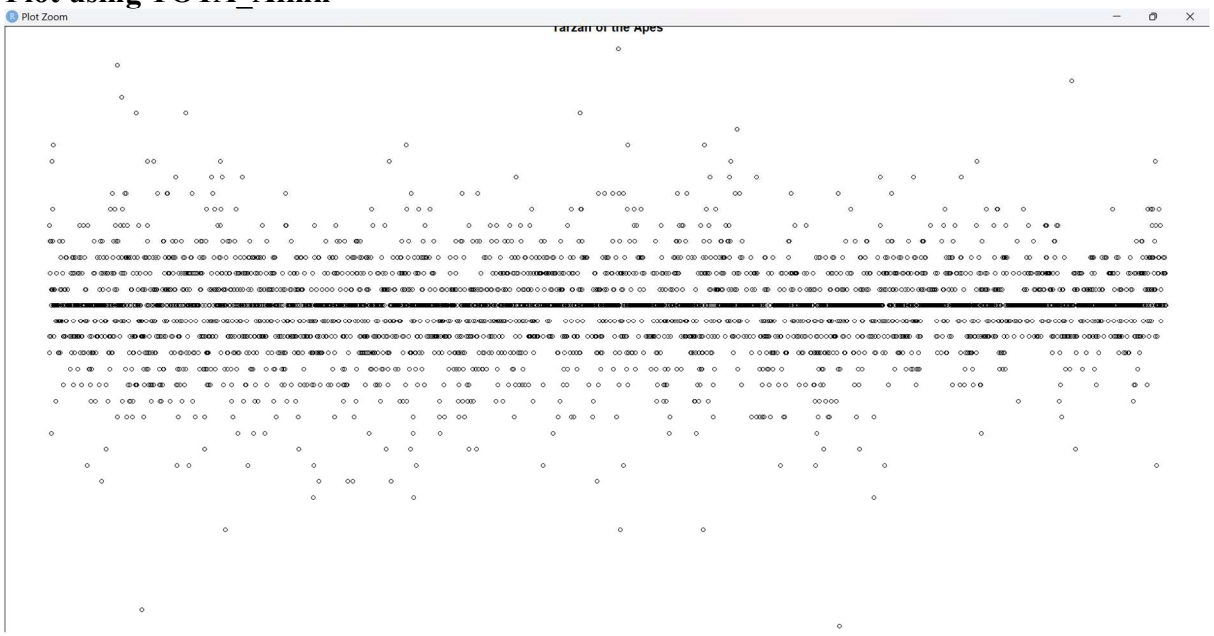
```
#Plotting to understand how the sentiment of text changes from the beginning of the text till the end
plot(TOTA_Sentiment, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")
plot(TOTA_Bing, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")
plot(TOTA_Afinn, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")
plot(TOTA_nrc, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")
```

**Plot using TOTA_Sentiment**

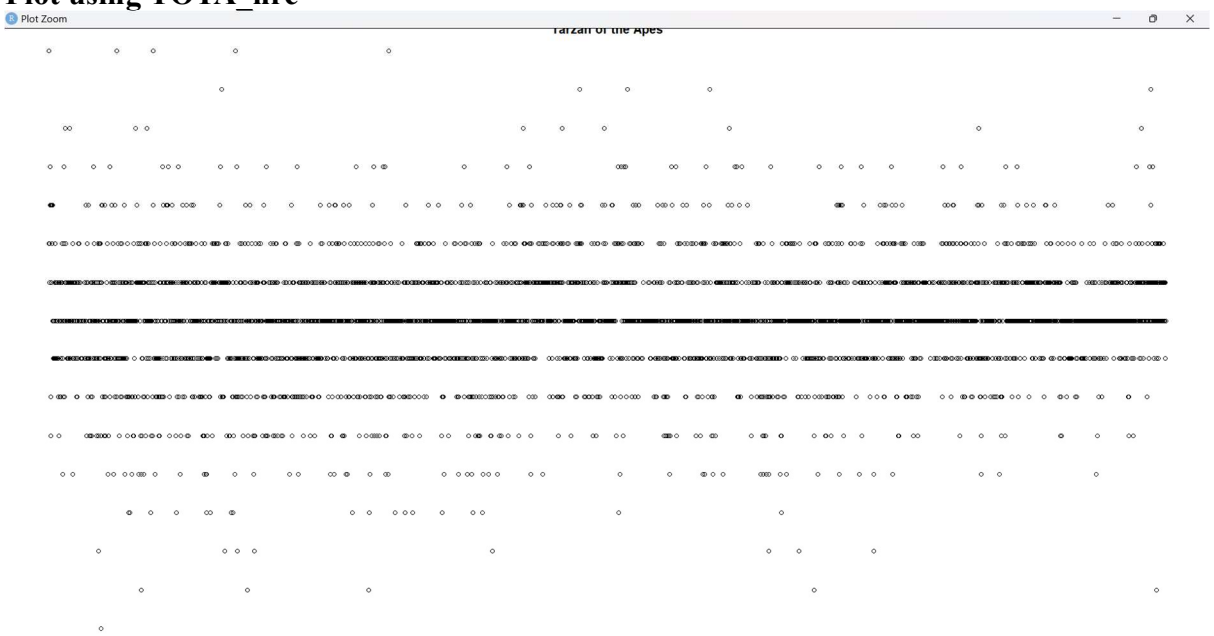



**Plot using TOTA_Bing**

**Plot using TOTA_Afinn**




**Plot using TOTA_nrc**




By comparing the plots generated using different sentiment dictionaries - 'syuzhet', Bing, Afinn, and NRC - readers can gain insights into how sentiment fluctuates throughout the text according to different lexicons and methodologies.

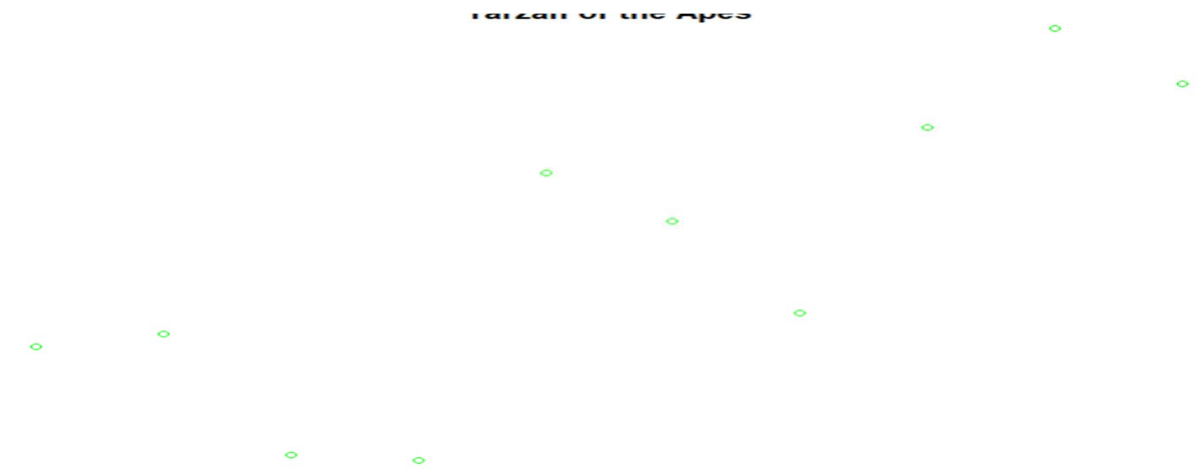
## CREATING PLOTS TO VISUALIZE THE DISTRIBUTION OF THE EMOTIONAL VALENCE OF "TARZAN OF THE APES" BASED ON SENTIMENT SCORES, GROUPED INTO BINS

Now, we can visualize the distribution of emotional valence in "Tarzan of the Apes" based on sentiment scores grouped into bins. First, it calculates the percentage values of sentiment scores using the `get_percentage_values()` function, dividing the scores into 10 bins. The resulting structure of `TOTA_SentimentPctValue` is then plotted, with the main title "Tarzan of the Apes" and axes labeled for clarity. The plot is colored green.

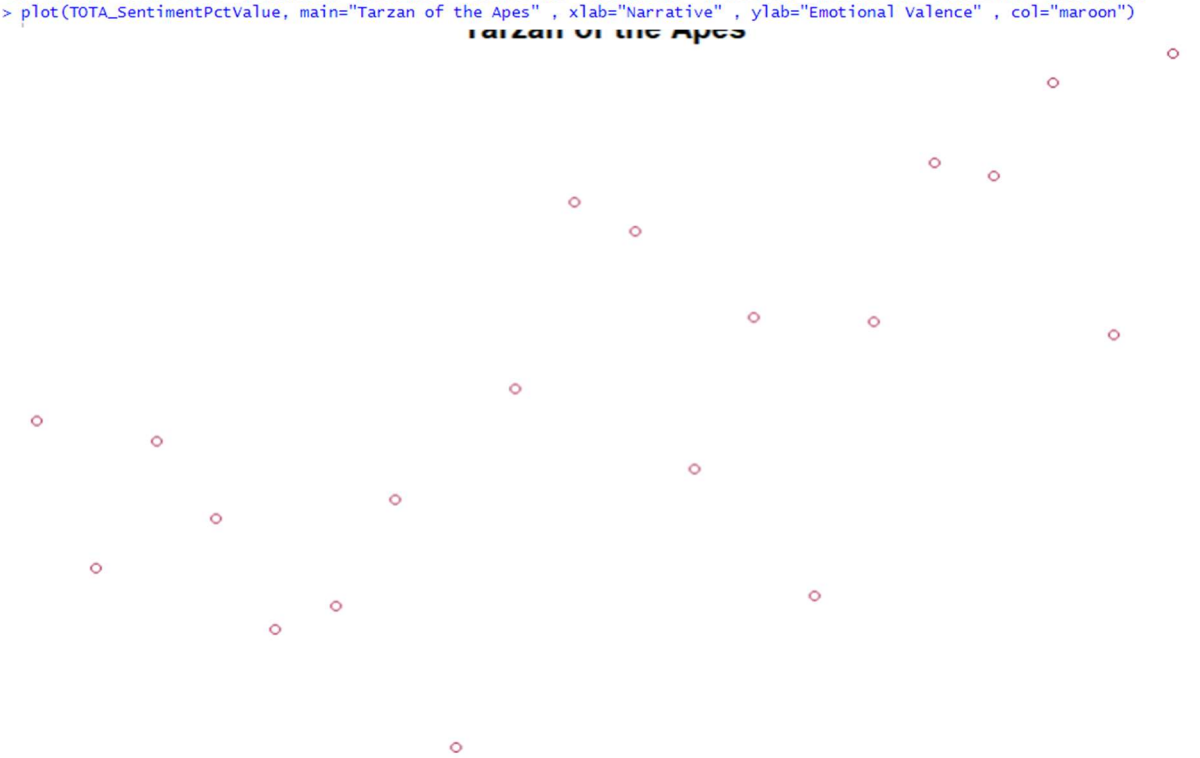
Subsequently, the process is repeated for 20 bins and 30 bins, generating separate plots for each. The color of the plots varies to aid differentiation: blue for 10 bins, maroon for 20 bins, and black for 30 bins.

Through these plots, the distribution of emotional valence across the narrative can be visually analyzed, providing insights into the text's overall sentiment composition & how sentiment evolves throughout the text and identifying any notable patterns or trends.

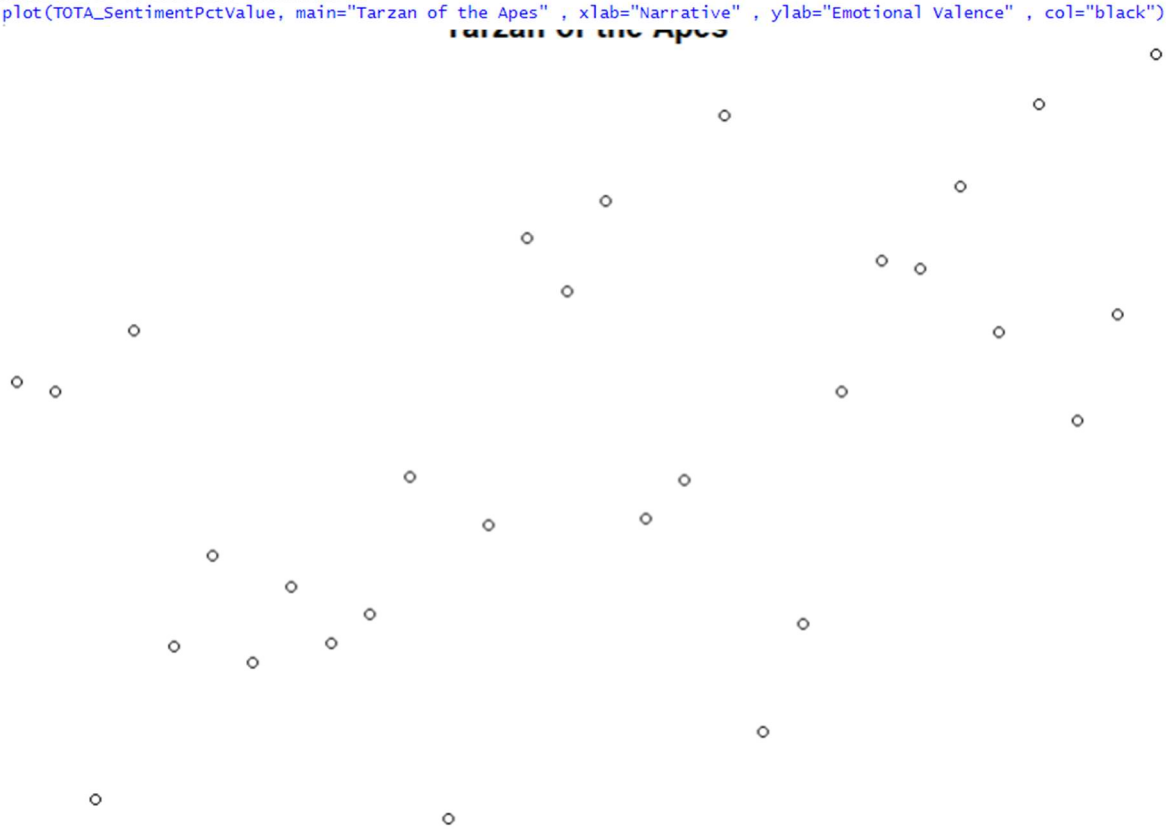
```
> #Creating plots to visualize the distribution of the emotional valence of "Tarzan of the Apes" based on sentiment scores, grouped into bins
> TOTA_SentimentPctValue <- get_percentage_values(TOTA_Sentiment , bins=10)
> structure(TOTA_SentimentPctValue)
          1           2           3           4           5           6           7           8           9          10
-0.21111111 -0.19733179 -0.33051044 -0.33689095 -0.01867749 -0.07192575 -0.17412993  0.03132251  0.14071926  0.07893519
> plot(TOTA_SentimentPctValue, main="Tarzan of the Apes" , xlab="Narrative" , ylab="Emotional Valence" , col="green")
```



```
> TOTA_SentimentPctValue <- get_percentage_values(TOTA_Sentiment , bins=20)
> structure(TOTA_SentimentPctValue)
          1           2           3           4           5           6           7           8           9          10
-0.13981481 -0.28240741 -0.16046512 -0.23402778 -0.34232558 -0.31875000 -0.21697674 -0.45625000 -0.10906977  0.07129630
         11          12          13          14          15          16          17          18          19          20
 0.04305556 -0.18744186 -0.03981481 -0.30906977 -0.04537037  0.10837209  0.09537037  0.18627907 -0.05694444  0.21481481
> plot(TOTA_SentimentPctValue, main="Tarzan of the Apes" , xlab="Narrative" , ylab="Emotional Valence" , col="maroon")
```



```
> TOTA_SentimentPctValue <- get_percentage_values(TOTA_Sentiment , bins=30)
> structure(TOTA_SentimentPctValue)
           1            2            3            4            5            6            7            8            9
-0.052430556 -0.063888889 -0.517013889  0.003496503 -0.347916667 -0.246180556 -0.366083916 -0.281597222 -0.344097222
          10           11           12           13           14           15           16           17           18
-0.311458333 -0.158041958 -0.539930556 -0.211805556  0.107342657  0.049305556  0.149652778 -0.204895105 -0.161458333
          19           20           21           22           23           24           25           26           27
 0.243055556 -0.442708333 -0.323776224 -0.063541667  0.083333333  0.074475524  0.164236111  0.002083333  0.256643357
          28           29           30
-0.096527778  0.021180556  0.312152778
> plot(TOTA_SentimentPctValue, main="Tarzan of the Apes" , xlab="Narrative" , ylab="Emotional Valence" , col="black")
```



The plots generated by the analysis reveal that the sentiment in the text fluctuates significantly. This suggests that the ebook encompasses a diverse range of emotions and experiences, with sentiment shifting frequently. Based on the given examples from the document, it can be concluded that the ebook conveys a mix of emotions. For instance, words like "joyful," "excited," and "happy" suggest positive emotions, indicating moments of happiness or excitement within the text. Conversely, words like "sorrowful," "lonely," and "gloomy" imply feelings of sadness or melancholy. Therefore, the ebook seems to evoke a range of emotions, from joy to sorrow, reflecting the complexity of human experiences and emotions depicted within its pages. While the initial analysis provides insight into the general sentiment fluctuations within the text, delving deeper may reveal more nuanced emotions and themes.

## DISCUSSION

This project has provided invaluable hands-on experience with text analytics, offering a comprehensive journey through the exploration of textual data. While following a structured path outlined in the rubric, we encountered numerous challenges that enriched our understanding of text analytics. Dealing with multiple documents, accommodating large text, and mitigating sparsity were just a few of the obstacles we encountered. These challenges illuminated the empirical nature of data science, emphasizing the importance of practical experience in navigating real-world data complexities. Through the exploration of data in this project, we've gained valuable insights into the intricacies of natural language processing (NLP) and the diverse challenges it presents. One key aspect we've learned is the importance of preprocessing techniques such as tokenization, stemming, and lemmatization in preparing textual data for analysis. These techniques not only help standardize text but also enable more accurate analysis by reducing noise and capturing the essence of the text.

One of the fundamental lessons gleaned from this project is the critical role of data preprocessing in text analytics. By cleaning the data, and removing stop words, punctuation, and quotation marks, we recognized the significance of preparing the data before analysis. By eliminating extraneous elements that do not contribute to the analysis, we ensured the integrity and accuracy of our insights. Moreover, we explored various packages for text analytics within R, uncovering their diverse capabilities such as generating word clouds, extracting information, and assigning sentiments. Through these analyses, we pieced together a coherent understanding of the textual data, despite its initial complexity.

The Text Analytics project served as a significant learning opportunity, consolidating the knowledge acquired from previous projects and enhancing our proficiency in various aspects of analytics. It played a pivotal role in refining our approaches, thought processes, and troubleshooting skills, further solidifying our foundation in analytics. Moreover, the project underscored the convenience and effectiveness of using R for conducting comprehensive data analytics, reaffirming its position as a powerful tool in this domain.

In summary, completing this project has left us feeling confident and competent in utilizing R for a multitude of tasks, including text analytics on large datasets, data cleaning, information extraction, and visualization generation. We now possess the capability to interpret sentiments, apply diverse functions and metrics to analyze documents and draw insightful conclusions about the underlying problem domain. Additionally, the project has equipped us with the skills to streamline and simplify text for more efficient analysis, further enhancing our ability to extract meaningful insights from textual data.