INTRO TO BIG DATA AND ANALYTICS

CLASS PROJECT – 3

TEXT ANALYTICS in R

By

Group - 13

Rithik Reddy P V (G23879158)

Vivek Kommareddy (G43709632)

Professor: Stephen Kaisler

In this project, Edgar Rice Burroughs' "Tarzan of the Apes" will be analysed and studied using R. We will begin by creating a VCorpus to process the text of the book. Next, we will determine which 10 words and sentences are the longest in each chapter, remove all punctuation, and aggregate the results into a data table. This exercise will demonstrate how text analytics can be used to use R programming to extract insightful information from large amounts of text data.

As an empirical science, data science makes inferences from data analysis, as our project showed when we used different R packages to analyse the novel and eliminate punctuation and stop words.

Theme of the book:

"Tarzan of the Apes" examines the nature vs. nurture controversy by contrasting Tarzan's aristocratic British ancestry with his untamed African jungle upbringing. In the end, the novel raises the question of whether ancestry or environment has a greater influence on identity, arguing that moral decisions rather than social conventions are what define genuine humanity.

The following R packages are used in the project:

1. **tm**: The "tm" package stands for Text Mining, and it's used for managing text data, enabling preprocessing, transformation, metadata management, and document-term matrices among other text mining functionalities.

2. **stringr**: This package provides a set of simple and consistent tools for working with strings, i.e., character data, in R. It's built on top of stringi, which handles string operations. The package simplifies and optimizes common string operations like substring detection, replacement, and manipulation.

3. **tidyverse**: The tidyverse is a collection of R packages designed for data science. It includes several packages like ggplot2, dplyr, tidyr, readr,

and purrr that help in data manipulation, exploration, and visualization in a consistent and coherent manner.

4. **tidytext**: This package allows for text processing and analysis in a tidy data framework. It simplifies the process of turning text into tidy data structures for analysis and visualization, working seamlessly with other tidyverse tools.

5. **wordcloud**: The wordcloud package is used for creating aesthetic and customizable word clouds that help in visualizing the frequency or importance of words within a text corpus.
6. **quanteda**: Quanteda provides tools for quantitative text analysis in R. It allows for the construction of document-feature matrices and implements efficient management of text data, enabling complex text analytical techniques.

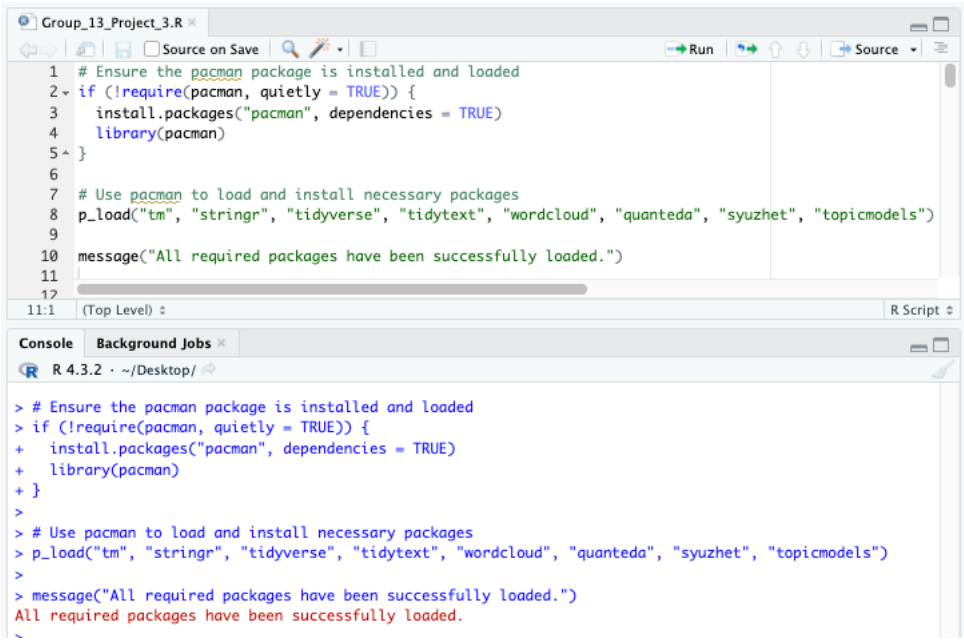
7. **syuzhet**: This package extracts sentiment and sentiment-derived plot arcs from text. It uses various sentiment dictionaries to provide insights into the emotional trajectory of a narrative text.

8. **topicmodels**: This package facilitates topic modelling with statistical methods like Latent Dirichlet Allocation (LDA) and the Correlated Topic Model (CTM). It's used to discover abstract topics within a collection of documents.

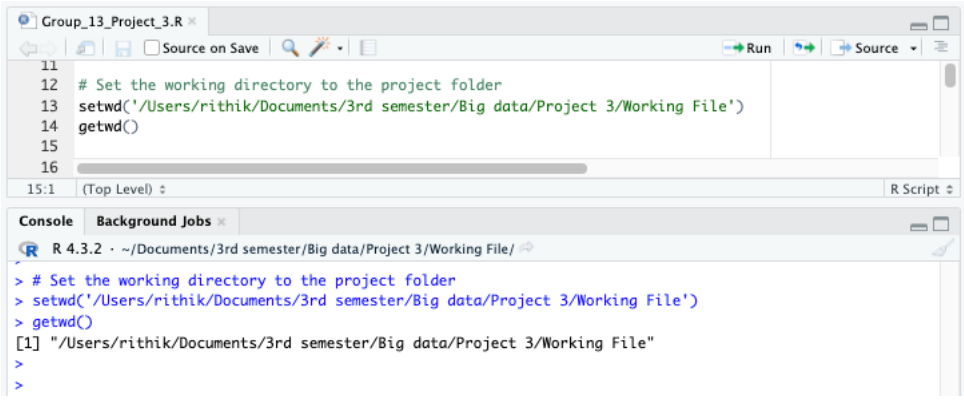
Each of these packages plays a specific role in text data manipulation, analysis, or visualization, making them invaluable for text analytics in R.

Steps for Setup and Installation:

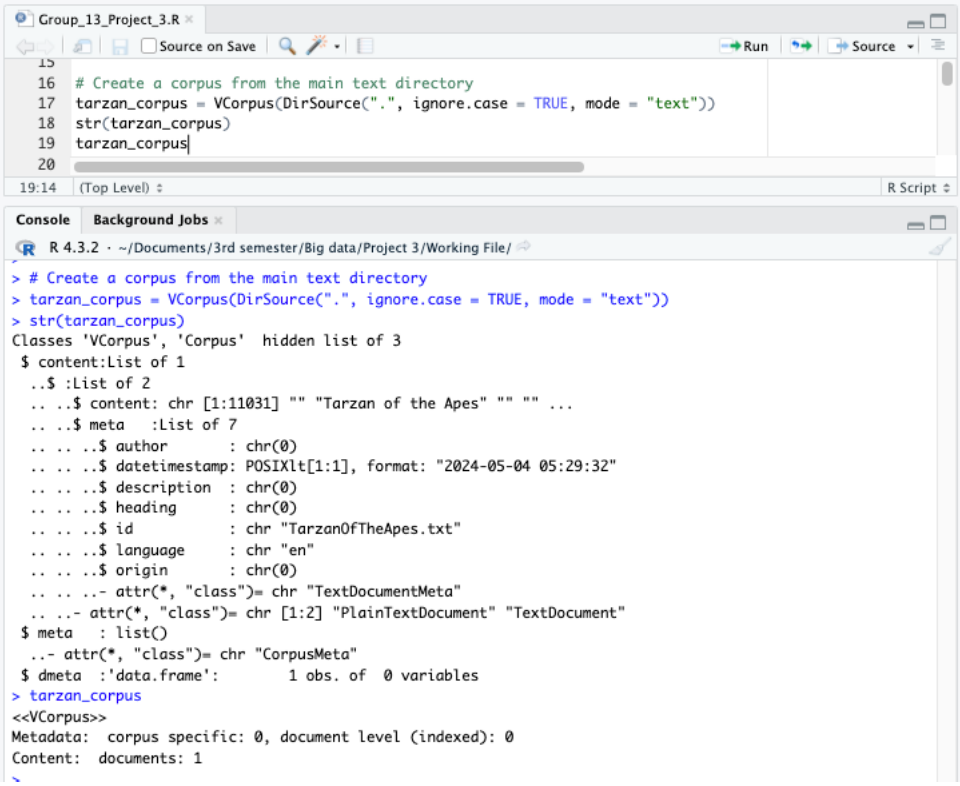
1. Ensure the pacman package is installed and use it to load and install necessary packages.



2. Set the working directory for the project.

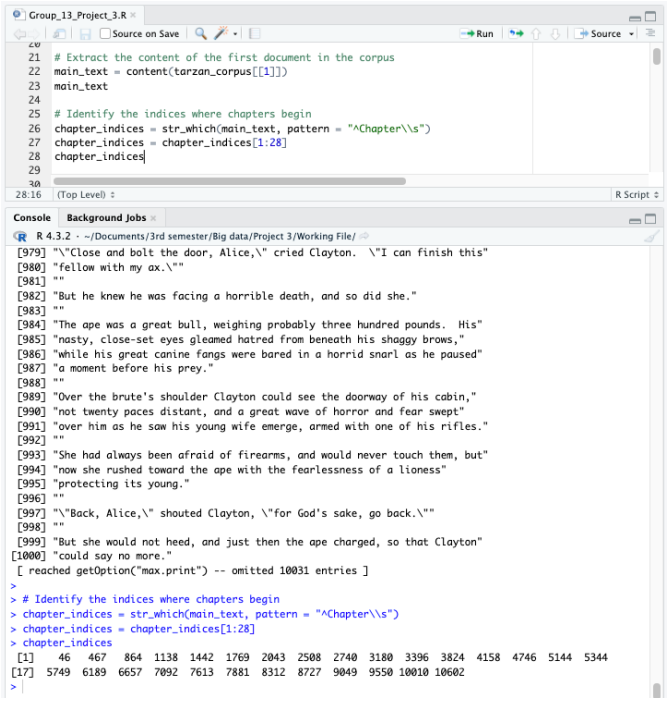


3. Create a corpus from the main text directory.

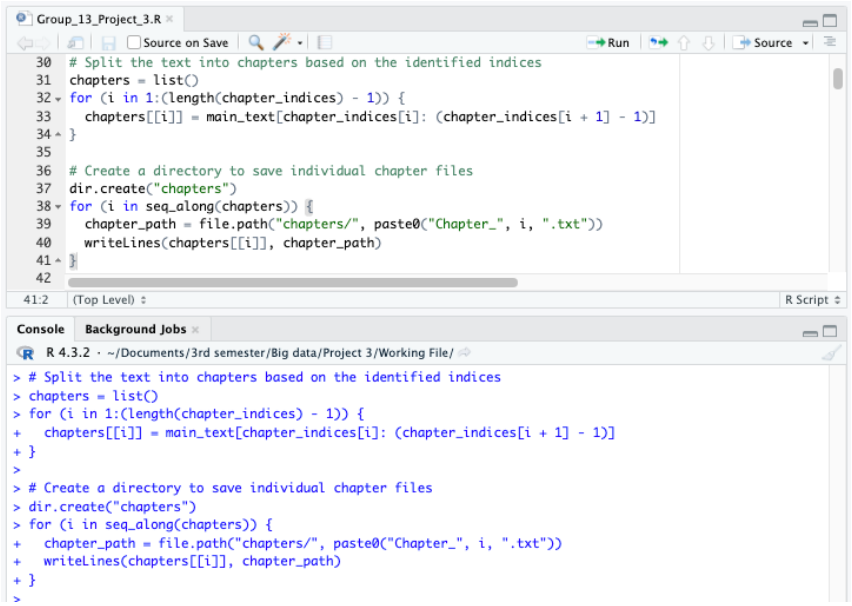


Extracting the chapters from the Text document

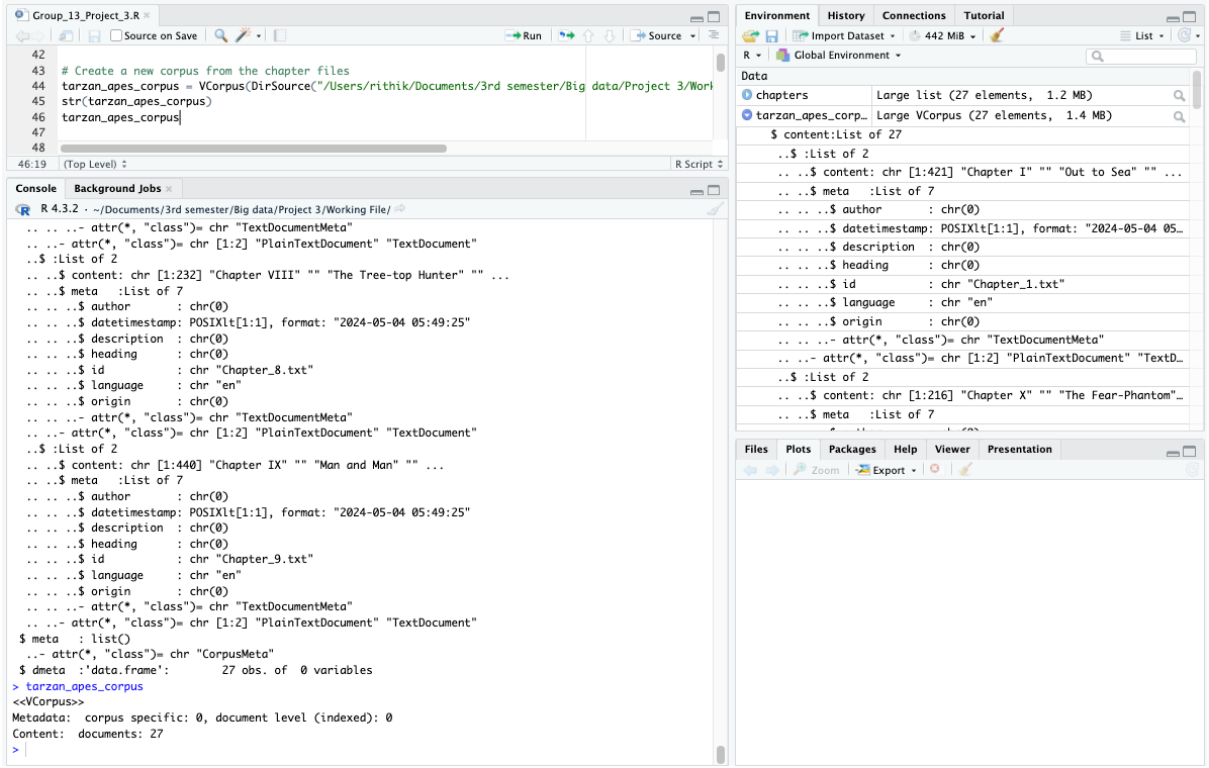
1. Identify the indices where chapters begin.



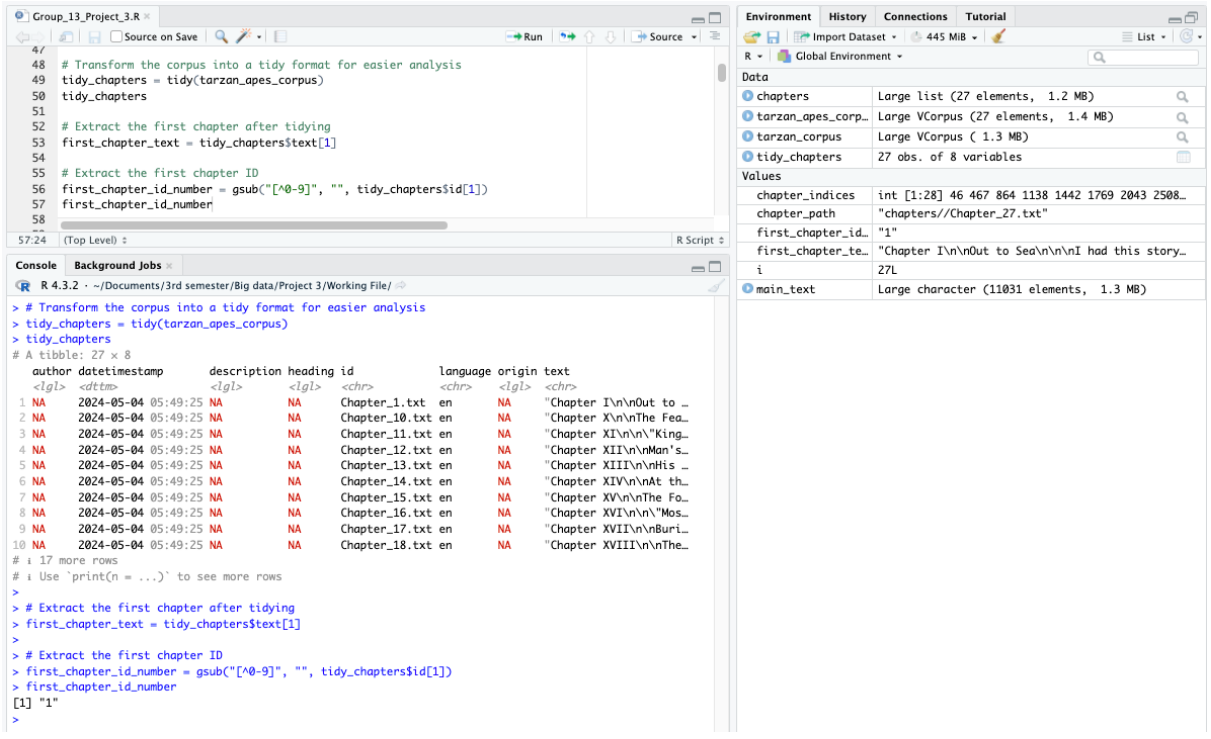
2. Split the text into chapters based on the indices and create a directory to save individual chapter files.



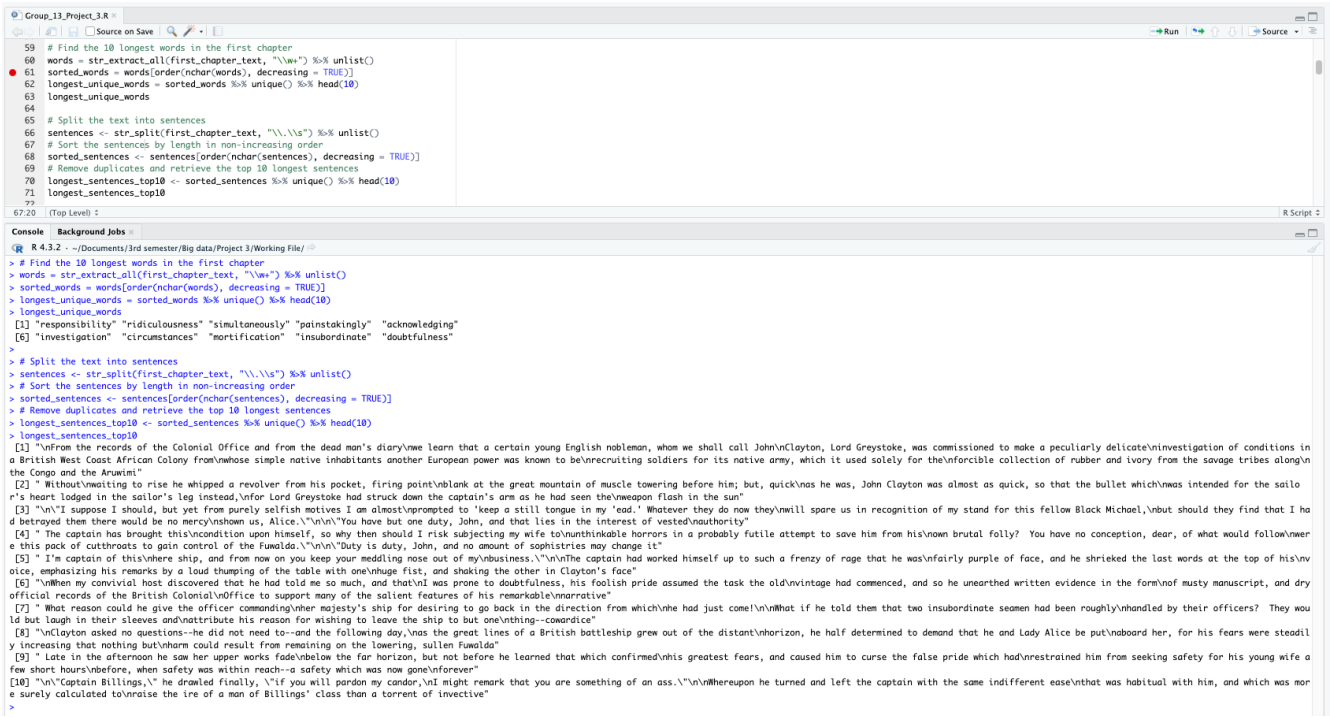
Create Vcorpus for the 27 chapters.



Transform the corpus into tidy format for easier analysis and extract the first chapter after tidying and then remove the chapter number from the text.



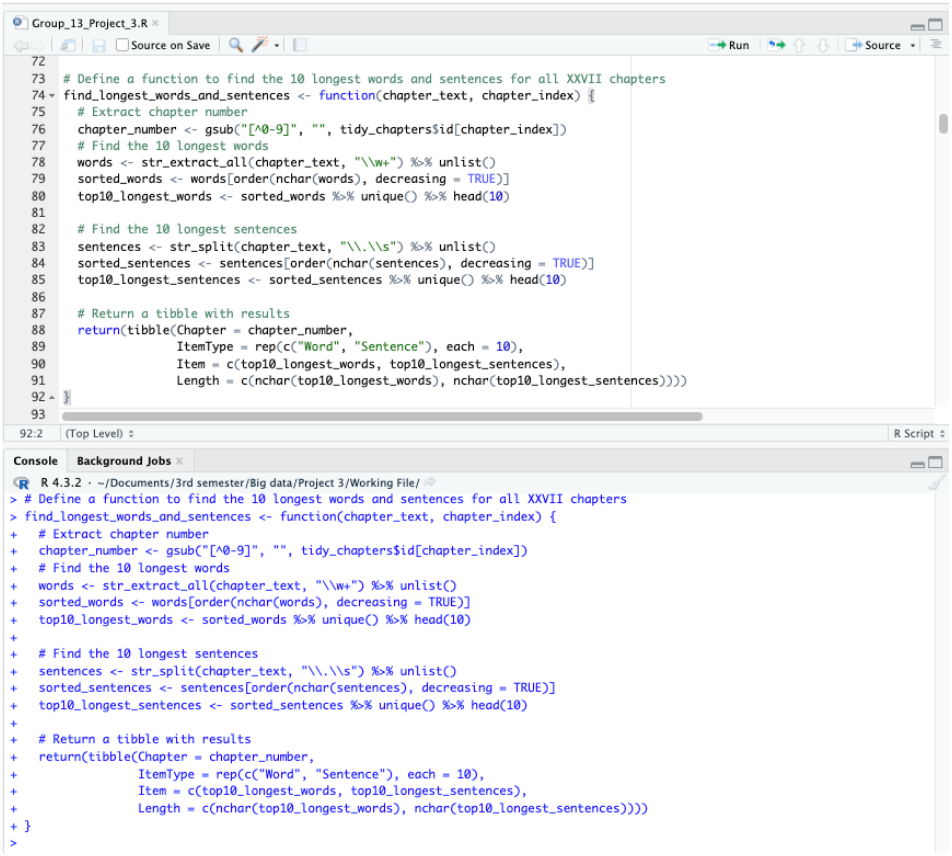
Finding 10 largest words and sentences in the first chapter



```
# Find the 10 longest words in the first chapter
words = str_extract_all(first_chapter_text, "\\w+") %>% unlist()
sorted_words = words[order(nchar(words), decreasing = TRUE)]
longest_unique_words = sorted_words %>% unique() %>% head(10)
longest_unique_words

# Split the text into sentences
sentences <- str_split(first_chapter_text, "\\.\\s") %>% unlist()
# Sort the sentences by length in non-increasing order
sorted_sentences <- sentences[order(nchar(sentences), decreasing = TRUE)]
# Remove duplicates and retrieve the top 10 longest sentences
longest_sentences_top10 <- sorted_sentences %>% unique() %>% head(10)
longest_sentences_top10
```

```
> # Find the 10 longest words in the first chapter
> words = str_extract_all(first_chapter_text, "\\w+") %>% unlist()
> sorted_words = words[order(nchar(words), decreasing = TRUE)]
> longest_unique_words = sorted_words %>% unique() %>% head(10)
> longest_unique_words
 [1] "responsibility" "ridiculousness" "simultaneously" "painstakingly"  "acknowledging"
 [6] "investigation"  "circumstances"  "mortification"  "insubordinate"  "doubtfulness"
>
> # Split the text into sentences
> sentences <- str_split(first_chapter_text, "\\.\\s") %>% unlist()
> # Sort the sentences by length in non-increasing order
> sorted_sentences <- sentences[order(nchar(sentences), decreasing = TRUE)]
> # Remove duplicates and retrieve the top 10 longest sentences
> longest_sentences_top10 <- sorted_sentences %>% unique() %>% head(10)
> longest_sentences_top10
 [1] "\nFrom the records of the Colonial Office and from the dead man's diary\nwe learn that a certain young English nobleman, whom we shall call John\nClayton, Lord Greystoke, was commissioned to make a peculiarly delicate\ninvestigation of conditions in a British West Coast African Colony from\nwhose simple native inhabitants another European power was known to be\nrecruiting soldiers for its native army, which it used solely for the\nforcible collection of rubber and ivory from the savage tribes along\n the Congo and the Aruwimi"
 [2] " Without\nwaiting to rise he whipped a revolver from his pocket, firing point\nblank at the great mountain of muscle towering before him; but, quick\nas he was, John Clayton was almost as quick, so that the bullet which\nwas intended for the sailo r's heart lodged in the sailor's leg instead,\nfor Lord Greystoke had struck down the captain's arm as he had seen the\nweapon flash in the sun"
 [3] "\n\"I suppose I should, but yet from purely selfish motives I am almost\nprompted to 'keep a still tongue in my 'ead.' Whatever they do now they'nwill spare us in recognition of my stand for this fellow Black Michael,\nbut should they find that I ha d betrayed them there would be no mercy\nshown us, Alice.\"\n\n\"You have but one duty, John, and that lies in the interest of vested\nauthority"
 [4] " The captain has brought this\ncondition upon himself, so why then should I risk subjecting my wife to\nunthinkable horrors in a probably futile attempt to save him from his\nown brutal folly?  You have no conception, dear, of what would follow\nwer e this pack of cutthroats to gain control of the Fuwalda.\"\n\n\"Duty is duty, John, and no amount of sophistries may change it"
 [5] " I'm captain of this\nhere ship, and from now on you keep your meddling nose out of my\nbusiness.\"\n\nThe captain had worked himself up to such a frenzy of rage that he was\nfairly purple of face, and he shrieked the last words at the top of his\nv oice, emphasizing his remarks by a loud thumping of the table with one\nhuge fist, and shaking the other in Clayton's face"
 [6] "\nWhen my convivial host discovered that he had told me so much, and that\nI was prone to doubtfulness, his foolish pride assumed the task the old\nvintage had commenced, and so he unearthed written evidence in the form\nof musty manuscript, and dry official records of the British Colonial\nOffice to support many of the salient features of his remarkable\nnarrative"
 [7] " What reason could he give the officer commanding\nher majesty's ship for desiring to go back in the direction from which\nhe had just come!\n\nWhat if he told them that two insubordinate seamen had been roughly\nhandled by their officers?  They wou ld but laugh in their sleeves and\nattribute his reason for wishing to leave the ship to but one\nthing--cowardice"
 [8] "\nClayton asked no questions--he did not need to--and the following day,\nas the great lines of a British battleship grew out of the distant\nhorizon, he half determined to demand that he and Lady Alice be put\naboard her, for his fears were steadil y increasing that nothing but\nharm could result from remaining on the lowering, sullen Fuwalda"
 [9] " Late in the afternoon he saw her upper works fade\nbelow the far horizon, but not before he learned that which confirmed\nhis greatest fears, and caused him to curse the false pride which had\nrestrained him from seeking safety for his young wife a few short hours\nbefore, when safety was within reach--a safety which was now gone\nforever"
[10] "\n\"Captain Billings,\" he drawled finally, \"if you will pardon my candor,\nI might remark that you are something of an ass.\"\n\nWhereupon he turned and left the captain with the same indifferent ease\nthat was habitual with him, and which was mor e surely calculated to\nraise the ire of a man of Billings' class than a torrent of invective"
>
```

Defining a function to find top 10 longest words and sentences for all 27 chapters.



```
# Define a function to find the 10 longest words and sentences for all XXVII chapters
find_longest_words_and_sentences <- function(chapter_text, chapter_index) {
  # Extract chapter number
  chapter_number <- gsub("[^0-9]", "", tidy_chapters$id[chapter_index])
  # Find the 10 longest words
  words <- str_extract_all(chapter_text, "\\w+") %>% unlist()
  sorted_words <- words[order(nchar(words), decreasing = TRUE)]
  top10_longest_words <- sorted_words %>% unique() %>% head(10)

  # Find the 10 longest sentences
  sentences <- str_split(chapter_text, "\\.\\s") %>% unlist()
  sorted_sentences <- sentences[order(nchar(sentences), decreasing = TRUE)]
  top10_longest_sentences <- sorted_sentences %>% unique() %>% head(10)

  # Return a tibble with results
  return(tibble(Chapter = chapter_number,
                ItemType = rep(c("Word", "Sentence"), each = 10),
                Item = c(top10_longest_words, top10_longest_sentences),
                Length = c(nchar(top10_longest_words), nchar(top10_longest_sentences))))
}
```

```
> # Define a function to find the 10 longest words and sentences for all XXVII chapters
> find_longest_words_and_sentences <- function(chapter_text, chapter_index) {
+   # Extract chapter number
+   chapter_number <- gsub("[^0-9]", "", tidy_chapters$id[chapter_index])
+   # Find the 10 longest words
+   words <- str_extract_all(chapter_text, "\\w+") %>% unlist()
+   sorted_words <- words[order(nchar(words), decreasing = TRUE)]
+   top10_longest_words <- sorted_words %>% unique() %>% head(10)
+
+   # Find the 10 longest sentences
+   sentences <- str_split(chapter_text, "\\.\\s") %>% unlist()
+   sorted_sentences <- sentences[order(nchar(sentences), decreasing = TRUE)]
+   top10_longest_sentences <- sorted_sentences %>% unique() %>% head(10)
+
+   # Return a tibble with results
+   return(tibble(Chapter = chapter_number,
+                 ItemType = rep(c("Word", "Sentence"), each = 10),
+                 Item = c(top10_longest_words, top10_longest_sentences),
+                 Length = c(nchar(top10_longest_words), nchar(top10_longest_sentences))))
+ }
>
```

Retrieving all the chapter files and loading them into chapter_texts vector:

```
Group_13_Project_3.R
Source on Save                                                    Run

# Retrieve and process text from each chapter file to generate a results table
chapter_files <- list.files(path = "/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters", pattern = "\\.txt$")
chapter_texts <- vector("character", length(chapter_files))

# Read the contents of each text file
for (i in seq_along(chapter_files)) {
  chapter_texts[i] <- readr::read_file(file.path("/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters", chapter_files[i]))
}

103:29  (Top Level)
```

```
Console   Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
>
>
>
> # Retrieve and process text from each chapter file to generate a results table
> chapter_files <- list.files(path = "/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters", pattern = "\\.txt$")
> chapter_texts <- vector("character", length(chapter_files))
>
> # Read the contents of each text file
> for (i in seq_along(chapter_files)) {
+   chapter_texts[i] <- readr::read_file(file.path("/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters", chapter_files[i]))
+ }
>
```

Results table:

```
Group_13_Project_3.R
Source on Save                                   Run          Source

results_table <- map_dfr(seq_along(chapter_texts), ~find_longest_words_and_sentences(chapter_texts[.], .))
print(results_table, n = 600)

105:1  (Top Level)                                                                          R Script
```

```
Console   Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
487 7     Word     "possibilities"                                                                                        13
488 7     Word     "experimenting"                                                                                        13
489 7     Word     "inexhaustible"                                                                                        13
490 7     Word     "remorselessly"                                                                                        13
491 7     Sentence "\nFrom this primitive function has arisen, unquestionably, all the forms\nand ceremonials of modern church and state, for through all the\nc…   632
492 7     Sentence "\nSquatting upon his haunches on the table top in the cabin his father\nhad built--his smooth, brown, naked little body bent over the book\n…   506
493 7     Sentence "\nHe did not accomplish it in a day, or in a week, or in a month, or in a\nyear; but slowly, very slowly, he learned after he had grasped th…   362
494 7     Sentence "\nTublat was his most consistent enemy, but it was through Tublat that,\nwhen he was about thirteen, the persecution of his enemies suddenly…   361
495 7     Sentence "\nAt last he reached the fast disappearing feast and with his sharp knife\nslashed off a more generous portion than he had hoped for, an ent…   343
496 7     Sentence "\nFor hours absolute quiet reigned in the little clearing, except as it\nwas broken by the discordant notes of brilliantly feathered parrots…   342
497 7     Sentence "\nHence the younger males as they became adult found it more comfortable\nto take mates from their own tribe, or if they captured one of ano…   334
498 7     Sentence " Descended from a\nrace of meat eaters, never in his life, he thought, had he once\nsatisfied his appetite for animal food; and so now his a…   322
499 7     Sentence "\nMany travelers have seen the drums of the great apes, and some have\nheard the sounds of their beating and the noise of the wild, weird\nr…   313
500 7     Sentence "\nAnd so he progressed very, very slowly, for it was a hard and laborious\ntask which he had set himself without knowing it--a task which mi…   305
501 8     Word     "uncomfortable"                                                                                        13
502 8     Word     "possibilities"                                                                                        13
503 8     Word     "insignificant"                                                                                        13
504 8     Word     "occasionally"                                                                                         12
505 8     Word     "considerable"                                                                                         12
506 8     Word     "overhanging"                                                                                          11
507 8     Word     "superiority"                                                                                          11
508 8     Word     "conflicting"                                                                                          11
509 8     Word     "surrounding"                                                                                          11
510 8     Word     "destruction"                                                                                          11
511 8     Sentence "\nMany moons ago, when he had been much smaller, he had desired the skin\nof Sabor, the lioness, or Numa, the lion, or Sheeta, the leopard t…   527
512 8     Sentence "\nFor several months the tribe hovered near the beach where stood\nTarzan's cabin, and his studies took up the greater portion of his\ntime,…   299
513 8     Sentence "\nThus far his plan had worked to perfection, but when he grasped the\nrope, bracing himself behind a crotch of two mighty branches, he foun…   294
514 8     Sentence "\nAt last Tarzan tired of the sport, and with a parting roar of challenge\nand a well-aimed ripe fruit that spread soft and sticky over the\…   289
515 8     Sentence "\nPresently the beast dropped to the earth again and Tarzan came quickly\nto seize the rope, but Sabor had now found that it was only a slen…   268
516 8     Sentence "\nNow and again some ancient patriarch of the woods, rent by a flashing\nbolt, would crash in a thousand pieces among the surrounding trees,…   254
517 8     Sentence "\nAt the bottom of his little English heart beat the great desire to\ncover his nakedness with CLOTHES for he had learned from his picture\n…   235
518 8     Sentence "\nOnce old Sabor, crossing their path, sent them scurrying to the safety\nof the higher branches, for if she respected their number and thei…   222
519 8     Sentence "\nWith an angry lash of her tail she bared her yellow fangs, curling her\ngreat lips in a hideous snarl that wrinkled her bristling snout in…   219
520 8     Sentence "\nHe lost a long rope on this occasion, but he knew that had it been\nSabor who had thus dragged him from his perch the outcome might have\n…   218
521 9     Word     "consternation"                                                                                        13
522 9     Word     "noiselessness"                                                                                        13
523 9     Word     "transgressing"                                                                                        13
524 9     Word     "bloodthirsty"                                                                                         12
525 9     Word     "convulsively"                                                                                         12
526 9     Word     "impenetrable"                                                                                         12
527 9     Word     "appropriated"                                                                                         12
528 9     Word     "investigated"                                                                                         12
529 9     Word     "laboriously"                                                                                          11
530 9     Word     "brutishness"                                                                                          11
531 9     Sentence "\nAnd then Lord Greystoke wiped his greasy fingers upon his naked thighs\nand took up the trail of Kulonga, the son of Mbonga, the king; whi…   449
532 9     Sentence "\nThat they more greatly feared an attack from the rear than whatever\nunknown enemies lurked in their advance was evidenced by the formatio…   399
533 9     Sentence "\nAlso could he write, with printed letters, rapidly and plainly, but\nscript he had not mastered, for though there were several copy books\…   309
534 9     Sentence "\nSo it was that as Kulonga emerged from the shadow of the jungle a\nslender coil of rope sped sinuously above him from the lowest branch of…   309
535 9     Sentence "\nHow may we judge him, by what standards, this ape-man with the heart\nand head and body of an English gentleman, and the training of a wil…   299
536 9     Sentence "\nFor three days the little cavalcade marched slowly through the heart of\nthis unknown and untracked forest, until finally, early in the fo…   286
537 9     Sentence "\nBut who was Kulonga that he might not be eaten as fairly as Horta, the\nboar, or Bara, the deer?  Was he not simply another of the countle…   272
538 9     Sentence "Chapter IX\n\nMan and Man\n\n\nTarzan of the Apes lived on in his wild, jungle existence with little\nchange for several years, only that he…   270
539 9     Sentence " That any creature of the jungle could produce the\nred-and-yellow fangs which devoured wood and left nothing but fine dust\nsurprised Tarza…   266
540 9     Sentence "\nAs the man stood there with taut drawn bow Tarzan recognized him not so\nmuch the NEGRO as the ARCHER of his picture book--\n\n          …   252
>
```

Writing a for-loop to print individual results of all chapters:

Group_13_Project_3.R

```r
for (chapter in 1:28){
  cat("Chapter", chapter, "\n")
  results_table %>%
    filter(Chapter == chapter)%>%
    print(n=20)
}
```

Console — R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

```
> for (chapter in 1:28){
+   cat("Chapter", chapter, "\n")
+   results_table %>%
+     filter(Chapter == chapter)%>%
+     print(n=20)
+ }
Chapter 1
# A tibble: 20 × 4
   Chapter ItemType Item                                                                                   Length
   <chr>   <chr>    <chr>                                                                                   <int>
 1 1       Word     "responsibility"                                                                           14
 2 1       Word     "ridiculousness"                                                                           14
 3 1       Word     "simultaneously"                                                                           14
 4 1       Word     "painstakingly"                                                                            13
 5 1       Word     "acknowledging"                                                                            13
 6 1       Word     "investigation"                                                                            13
 7 1       Word     "circumstances"                                                                            13
 8 1       Word     "mortification"                                                                            13
 9 1       Word     "insubordinate"                                                                            13
10 1       Word     "doubtfulness"                                                                             12
11 1       Sentence "\nFrom the records of the Colonial Office and from the dead man's diary\nwe learn that a…    522
12 1       Sentence " Without\nwaiting to rise he whipped a revolver from his pocket, firing point\nblank at …    384
13 1       Sentence "\n\"I suppose I should, but yet from purely selfish motives I am almost\nprompted to 'ke…    382
14 1       Sentence " The captain has brought this\ncondition upon himself, so why then should I risk subject…    368
15 1       Sentence " I'm captain of this\nhere ship, and from now on you keep your meddling nose out of my\n…    363
16 1       Sentence "\nWhen my convivial host discovered that he had told me so much, and that\nI was prone t…    361
17 1       Sentence " What reason could he give the officer commanding\nher majesty's ship for desiring to go…    356
18 1       Sentence "\nClayton asked no questions--he did not need to--and the following day,\nas the great l…    339
19 1       Sentence " Late in the afternoon he saw her upper works fade\nbelow the far horizon, but not befor…    336
20 1       Sentence "\n\"Captain Billings,\" he drawled finally, \"if you will pardon my candor,\nI might rem…    332
Chapter 2
# A tibble: 20 × 4
   Chapter ItemType Item                                                                                   Length
   <chr>   <chr>    <chr>                                                                                   <int>
 1 2       Word     "Notwithstanding"                                                                          15
 2 2       Word     "transportation"                                                                           14
 3 2       Word     "remorselessly"                                                                            13
 4 2       Word     "heartlessness"                                                                            13
 5 2       Word     "threateningly"                                                                            13
 6 2       Word     "investigation"                                                                            13
 7 2       Word     "contemplation"                                                                            13
```

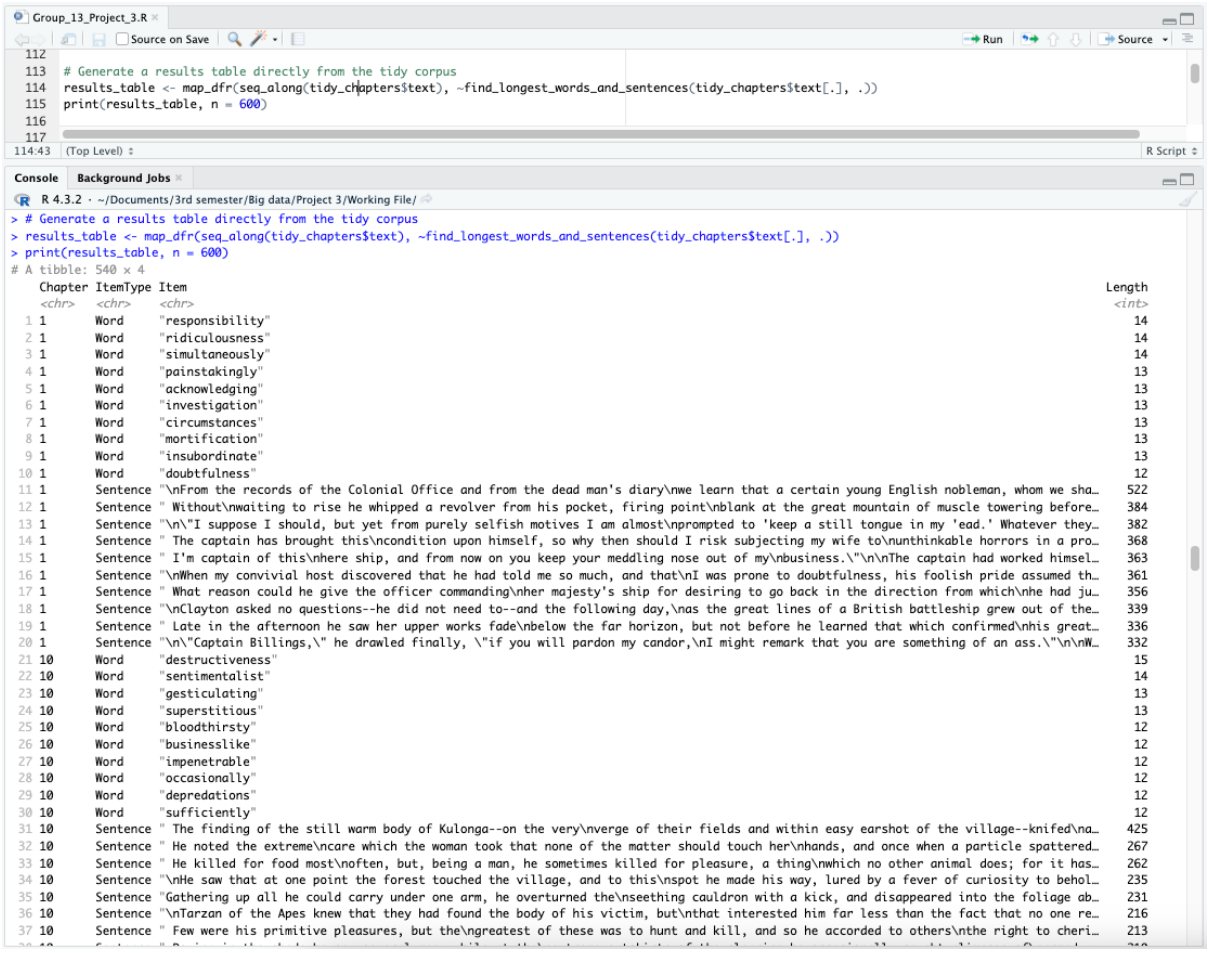

Group_13_Project_3.R

```r
for (chapter in 1:28){
  cat("Chapter", chapter, "\n")
  results_table %>%
    filter(Chapter == chapter)%>%
    print(n=20)
}
```


Console — R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

```
Chapter 3
# A tibble: 20 × 4
   Chapter ItemType Item                                                                                   Length
   <chr>   <chr>    <chr>                                                                                   <int>
 1 3       Word     "responsibilities"                                                                         16
 2 3       Word     "simultaneously"                                                                           14
 3 3       Word     "responsibility"                                                                           14
 4 3       Word     "comparatively"                                                                            13
 5 3       Word     "uncomfortable"                                                                            13
 6 3       Word     "acquaintances"                                                                            13
 7 3       Word     "consciousness"                                                                            13
 8 3       Word     "horizontally"                                                                             12
 9 3       Word     "transversely"                                                                             12
10 3       Word     "sufficiently"                                                                             12
11 3       Sentence "\nThe last entry in his diary was made the morning following her death,\nand there he re…    609
12 3       Sentence "\nThe brilliant birds and the little monkeys had become accustomed to\ntheir new acquain…    491
13 3       Sentence " I thought we were no longer in London, but in some horrible\nplace where great beasts a…    410
14 3       Sentence " Grown careless\nfrom months of continued safety, during which time he had seen no\ndang…    400
15 3       Sentence "\nThe door he built of pieces of the packing-boxes which had held their\nbelongings, nai…    304
16 3       Sentence "\nWith a low cry she sprang toward the cabin, and, as she entered, gave a\nbackward glan…    301
17 3       Sentence "\nAs soon as they had made their meager breakfast of salt pork, coffee\nand biscuit, Cla…    273
18 3       Sentence "\nThe building of a bed, chairs, table, and shelves was a relatively easy\nmatter, so th…    272
19 3       Sentence "\nHe had strengthened the window protections and fitted a unique wooden\nlock to the cab…    269
20 3       Sentence "\nThat it would have been beset by worries and apprehension had she been\nin full comman…    266
Chapter 4
# A tibble: 20 × 4
   Chapter ItemType Item                                                                                   Length
   <chr>   <chr>    <chr>                                                                                   <int>
 1 4       Word     "Notwithstanding"                                                                          15
 2 4       Word     "investigations"                                                                           14
 3 4       Word     "understanding"                                                                            13
 4 4       Word     "precipitately"                                                                            13
 5 4       Word     "uncontrolled"                                                                             12
 6 4       Word     "successfully"                                                                             12
 7 4       Word     "intelligence"                                                                             12
 8 4       Word     "occasionally"                                                                             12
 9 4       Word     "fearlessness"                                                                             12
10 4       Word     "alternative"                                                                              11
11 4       Sentence "\nWhen Kerchak came to a halt a short distance from the cabin and\ndiscovered that he st…    343
12 4       Sentence "\nHigh up among the branches of a mighty tree she hugged the shrieking\ninfant to her bo…    332
13 4       Sentence "\nWhen the king ape released the limp form which had been John Clayton,\nLord Greystoke,…    331
```

Generating results table directly from tidy corpus:

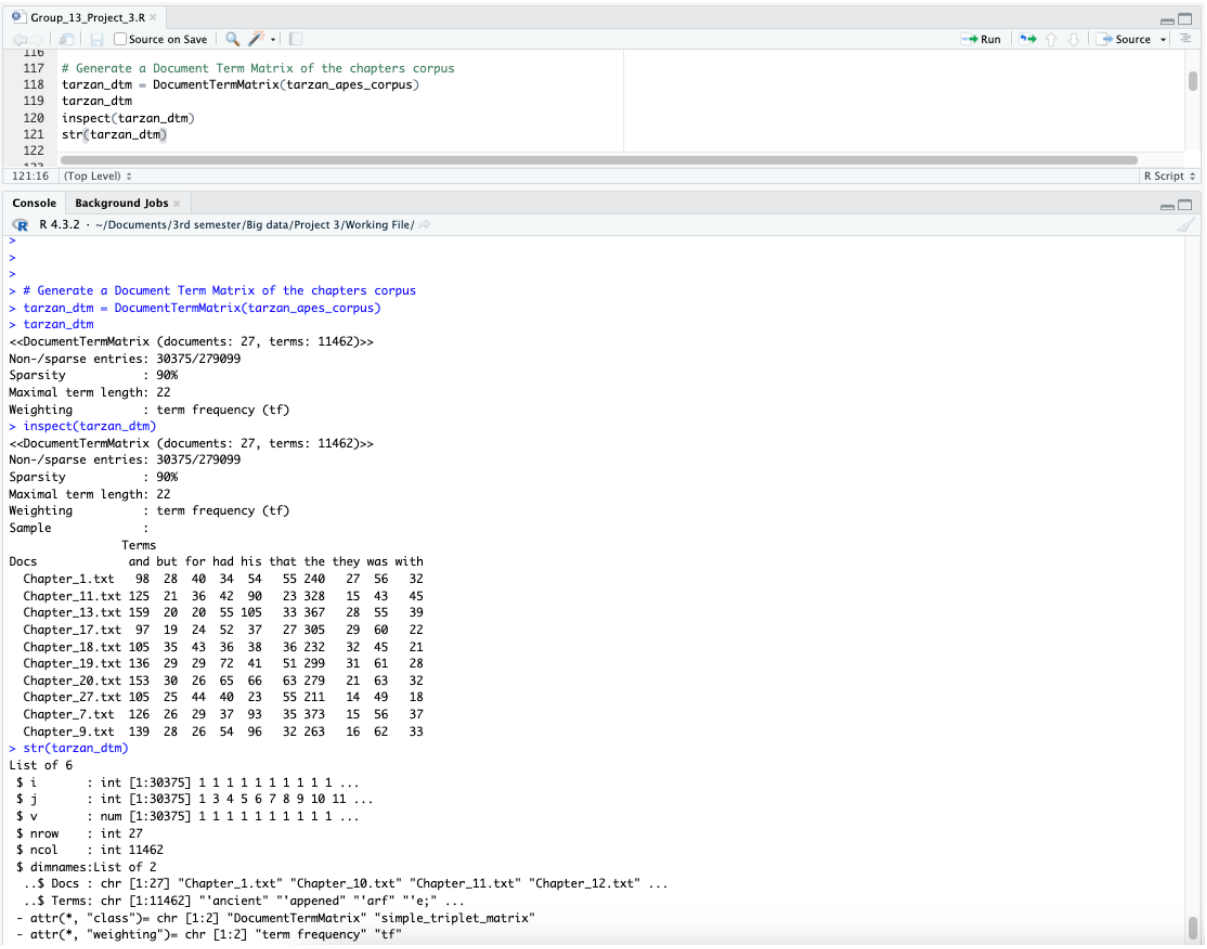


```
Group_13_Project_3.R
Source on Save | Run | Source

112
113 # Generate a results table directly from the tidy corpus
114 results_table <- map_dfr(seq_along(tidy_chapters$text), ~find_longest_words_and_sentences(tidy_chapters$text[.], .))
115 print(results_table, n = 600)
116
117

114:43 (Top Level) R Script

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

> # Generate a results table directly from the tidy corpus
> results_table <- map_dfr(seq_along(tidy_chapters$text), ~find_longest_words_and_sentences(tidy_chapters$text[.], .))
> print(results_table, n = 600)
# A tibble: 540 × 4
   Chapter ItemType Item                                                                                                Length
   <chr>   <chr>    <chr>                                                                                               <int>
 1 1       Word     "responsibility"                                                                                       14
 2 1       Word     "ridiculousness"                                                                                       14
 3 1       Word     "simultaneously"                                                                                       14
 4 1       Word     "painstakingly"                                                                                        13
 5 1       Word     "acknowledging"                                                                                        13
 6 1       Word     "investigation"                                                                                        13
 7 1       Word     "circumstances"                                                                                        13
 8 1       Word     "mortification"                                                                                        13
 9 1       Word     "insubordinate"                                                                                        13
10 1       Word     "doubtfulness"                                                                                         12
11 1       Sentence "\nFrom the records of the Colonial Office and from the dead man's diary\nwe learn that a certain young English nobleman, whom we sha…  522
12 1       Sentence " Without\nwaiting to rise he whipped a revolver from his pocket, firing point\nblank at the great mountain of muscle towering before…  384
13 1       Sentence "\n\"I suppose I should, but yet from purely selfish motives I am almost\nprompted to 'keep a still tongue in my 'ead.' Whatever they…  382
14 1       Sentence " The captain has brought this\ncondition upon himself, so why then should I risk subjecting my wife to\nunthinkable horrors in a pro…  368
15 1       Sentence " I'm captain of this\nhere ship, and from now on you keep your meddling nose out of my\nbusiness.\"\n\nThe captain had worked himsel…  363
16 1       Sentence "\nWhen my convivial host discovered that he had told me so much, and that\nI was prone to doubtfulness, his foolish pride assumed th…  361
17 1       Sentence " What reason could he give the officer commanding\nher majesty's ship for desiring to go back in the direction from which\nhe had ju…  356
18 1       Sentence "\nClayton asked no questions--he did not need to--and the following day,\nas the great lines of a British battleship grew out of the…  339
19 1       Sentence " Late in the afternoon he saw her upper works fade\nbelow the far horizon, but not before he learned that which confirmed\nhis great…  336
20 1       Sentence "\n\"Captain Billings,\" he drawled finally, \"if you will pardon my candor,\nI might remark that you are something of an ass.\"\n\nW…  332
21 10      Word     "destructiveness"                                                                                      15
22 10      Word     "sentimentalist"                                                                                       14
23 10      Word     "gesticulating"                                                                                        13
24 10      Word     "superstitious"                                                                                        13
25 10      Word     "bloodthirsty"                                                                                         12
26 10      Word     "businesslike"                                                                                         12
27 10      Word     "impenetrable"                                                                                         12
28 10      Word     "occasionally"                                                                                         12
29 10      Word     "depredations"                                                                                         12
30 10      Word     "sufficiently"                                                                                         12
31 10      Sentence " The finding of the still warm body of Kulonga--on the very\nverge of their fields and within easy earshot of the village--knifed\na…  425
32 10      Sentence " He noted the extreme\ncare which the woman took that none of the matter should touch her\nhands, and once when a particle spattered…  267
33 10      Sentence " He killed for food most\noften, but, being a man, he sometimes killed for pleasure, a thing\nwhich no other animal does; for it has…  262
34 10      Sentence "\nHe saw that at one point the forest touched the village, and to this\nspot he made his way, lured by a fever of curiosity to behol…  235
35 10      Sentence "Gathering up all he could carry under one arm, he overturned the\nseething cauldron with a kick, and disappeared into the foliage ab…  231
36 10      Sentence "\nTarzan of the Apes knew that they had found the body of his victim, but\nthat interested him far less than the fact that no one re…  216
37 10      Sentence " Few were his primitive pleasures, but the\ngreatest of these was to hunt and kill, and so he accorded to others\nthe right to cheri…  213
```

Document Term Matrix:

A Document Term Matrix (DTM) is a mathematical matrix used in the field of text mining and natural language processing to represent the frequency of terms in a corpus of text documents. It's a critical data structure where each row represents a document in the corpus, and each column represents a unique term from all the documents. The cells in the matrix contain the frequency count of terms as they appear in each document.

DTM for Chapters' Corpus

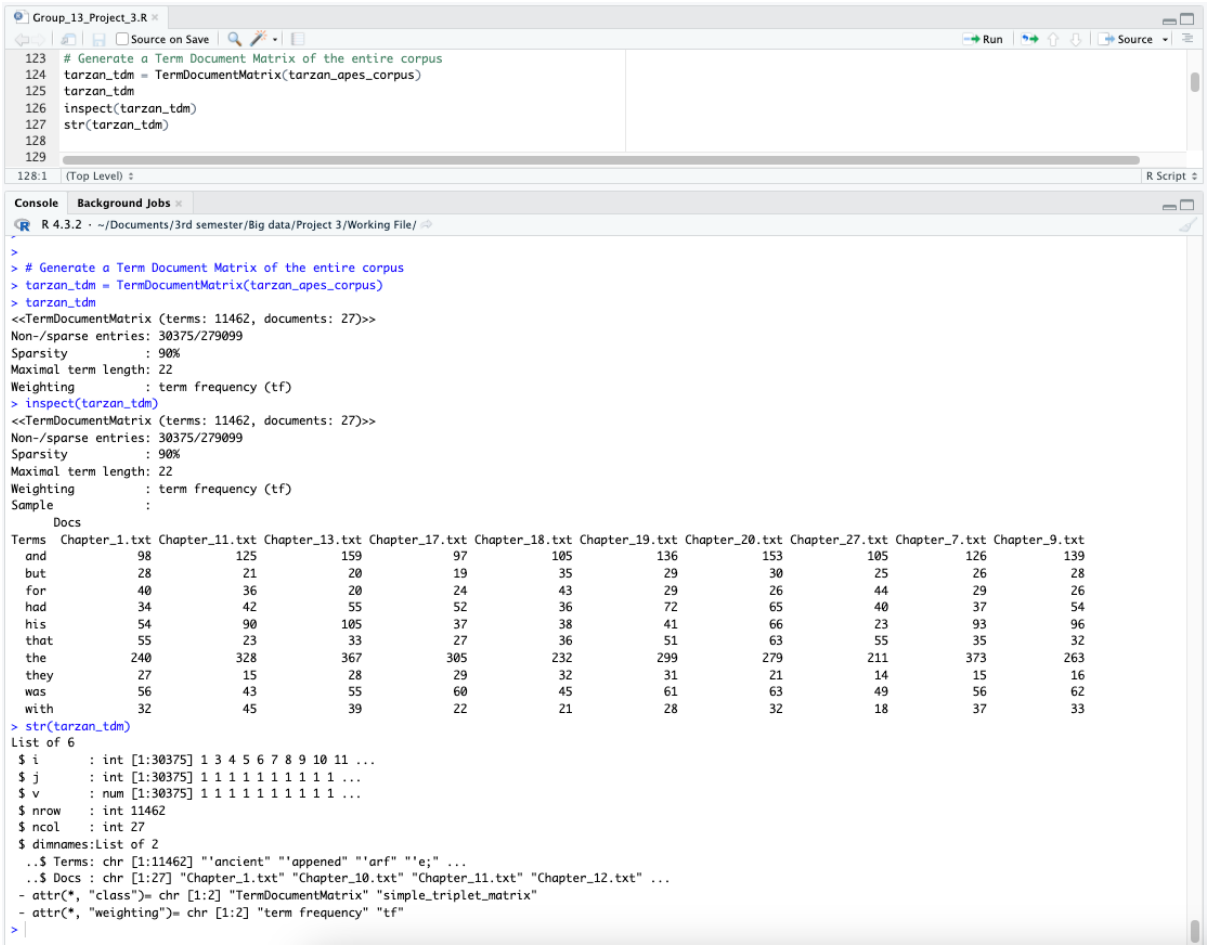


```
Group_13_Project_3.R
Source on Save | Run | Source

116
117  # Generate a Document Term Matrix of the chapters corpus
118  tarzan_dtm = DocumentTermMatrix(tarzan_apes_corpus)
119  tarzan_dtm
120  inspect(tarzan_dtm)
121  str(tarzan_dtm)
122

121:16  (Top Level)                                                R Script

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

>
>
>
> # Generate a Document Term Matrix of the chapters corpus
> tarzan_dtm = DocumentTermMatrix(tarzan_apes_corpus)
> tarzan_dtm
<<DocumentTermMatrix (documents: 27, terms: 11462)>>
Non-/sparse entries: 30375/279099
Sparsity           : 90%
Maximal term length: 22
Weighting          : term frequency (tf)
> inspect(tarzan_dtm)
<<DocumentTermMatrix (documents: 27, terms: 11462)>>
Non-/sparse entries: 30375/279099
Sparsity           : 90%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
                Terms
Docs             and but for had his that the they was with
  Chapter_1.txt   98  28  40  34  54   55 240   27  56   32
  Chapter_11.txt 125  21  36  42  90   23 328   15  43   45
  Chapter_13.txt 159  20  20  55 105   33 367   28  55   39
  Chapter_17.txt  97  19  24  52  37   27 305   29  60   22
  Chapter_18.txt 105  35  43  36  38   36 232   32  45   21
  Chapter_19.txt 136  29  29  72  41   51 299   31  61   28
  Chapter_20.txt 153  30  26  65  66   63 279   21  63   32
  Chapter_27.txt 105  25  44  40  23   55 211   14  49   18
  Chapter_7.txt  126  26  29  37  93   35 373   15  56   37
  Chapter_9.txt  139  28  26  54  96   32 263   16  62   33
> str(tarzan_dtm)
List of 6
 $ i       : int [1:30375] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:30375] 1 3 4 5 6 7 8 9 10 11 ...
 $ v       : num [1:30375] 1 1 1 1 1 1 1 1 1 1 ...
 $ nrow    : int 27
 $ ncol    : int 11462
 $ dimnames:List of 2
  ..$ Docs : chr [1:27] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_12.txt" ...
  ..$ Terms: chr [1:11462] "'ancient" "'appened" "'arf" "'e;" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

Term Document Matrix:

A Term Document Matrix (TDM) is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a TDM, rows represent terms and columns represent documents, which is the transpose of a Document Term Matrix (DTM). Each cell in the matrix contains the frequency of a term in a document.

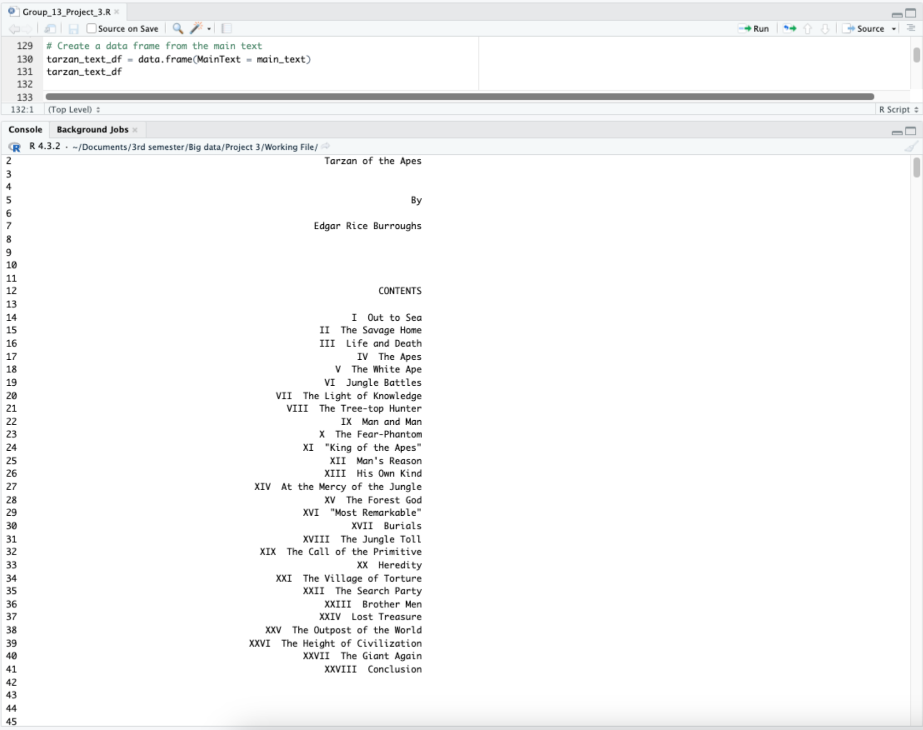
Generating TDM for the entire corpus:



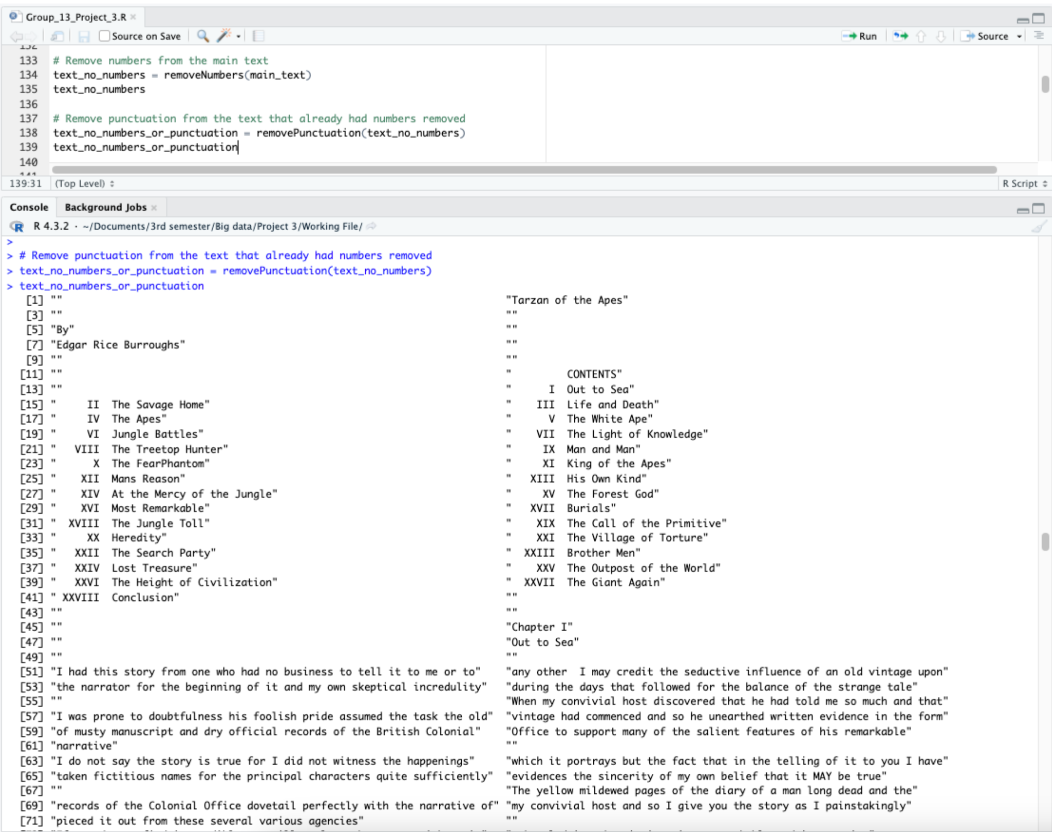
```
Group_13_Project_3.R
Source on Save | Run | Source
123  # Generate a Term Document Matrix of the entire corpus
124  tarzan_tdm = TermDocumentMatrix(tarzan_apes_corpus)
125  tarzan_tdm
126  inspect(tarzan_tdm)
127  str(tarzan_tdm)
128
129
128:1 (Top Level) R Script

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

>
> # Generate a Term Document Matrix of the entire corpus
> tarzan_tdm = TermDocumentMatrix(tarzan_apes_corpus)
> tarzan_tdm
<<TermDocumentMatrix (terms: 11462, documents: 27)>>
Non-/sparse entries: 30375/279099
Sparsity           : 90%
Maximal term length: 22
Weighting          : term frequency (tf)
> inspect(tarzan_tdm)
<<TermDocumentMatrix (terms: 11462, documents: 27)>>
Non-/sparse entries: 30375/279099
Sparsity           : 90%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
      Docs
Terms  Chapter_1.txt Chapter_11.txt Chapter_13.txt Chapter_17.txt Chapter_18.txt Chapter_19.txt Chapter_20.txt Chapter_27.txt Chapter_7.txt Chapter_9.txt
  and            98            125            159             97            105            136            153            105            126            139
  but            28             21             20             19             35             29             30             25             26             28
  for            40             36             20             24             43             29             26             44             29             26
  had            34             42             55             52             36             72             65             40             37             54
  his            54             90            105             37             38             41             66             23             93             96
  that           55             23             33             27             36             51             63             55             35             32
  the           240            328            367            305            232            299            279            211            373            263
  they           27             15             28             29             32             31             21             14             15             16
  was            56             43             55             60             45             61             63             49             56             62
  with           32             45             39             22             21             28             32             18             37             33
> str(tarzan_tdm)
List of 6
 $ i       : int [1:30375] 1 3 4 5 6 7 8 9 10 11 ...
 $ j       : int [1:30375] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:30375] 1 1 1 1 1 1 1 1 1 1 ...
 $ nrow    : int 11462
 $ ncol    : int 27
 $ dimnames:List of 2
  ..$ Terms: chr [1:11462] "'ancient" "'appened" "'arf" "'e;" ...
  ..$ Docs : chr [1:27] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_12.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

The choice between using a TDM and a DTM often depends on the specific needs of the application, such as the orientation that is more intuitive or efficient for processing with particular algorithms or data structures.

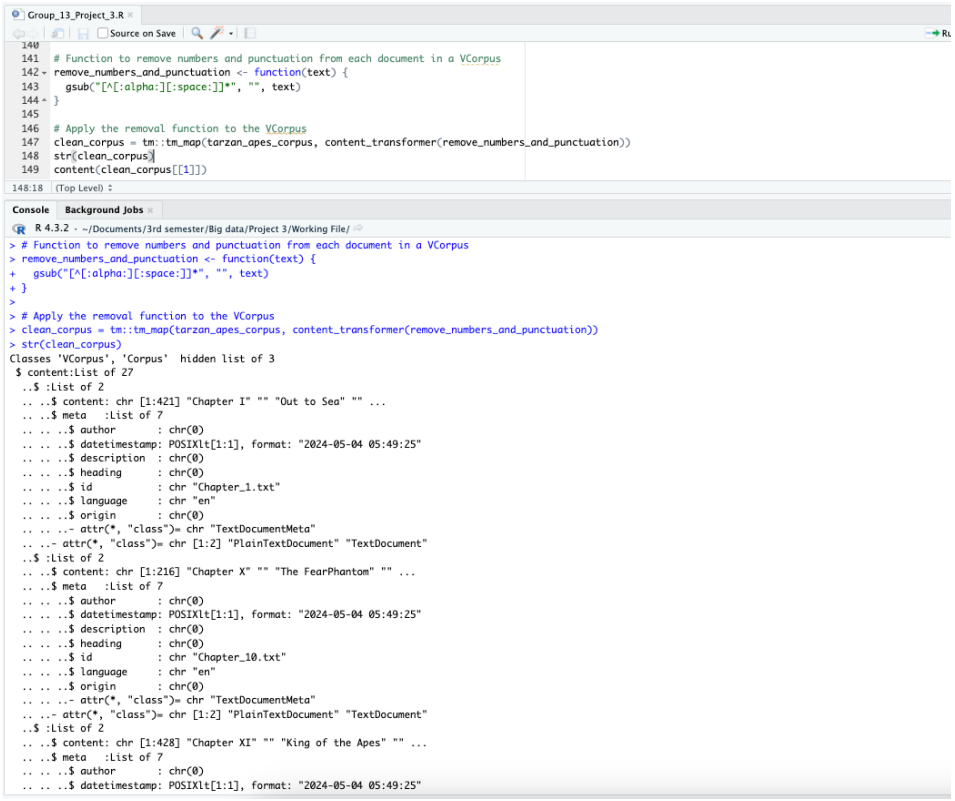
Creating a data frame for the Tarzan of the Apes text file:



Removing numbers and punctuation from the main text:

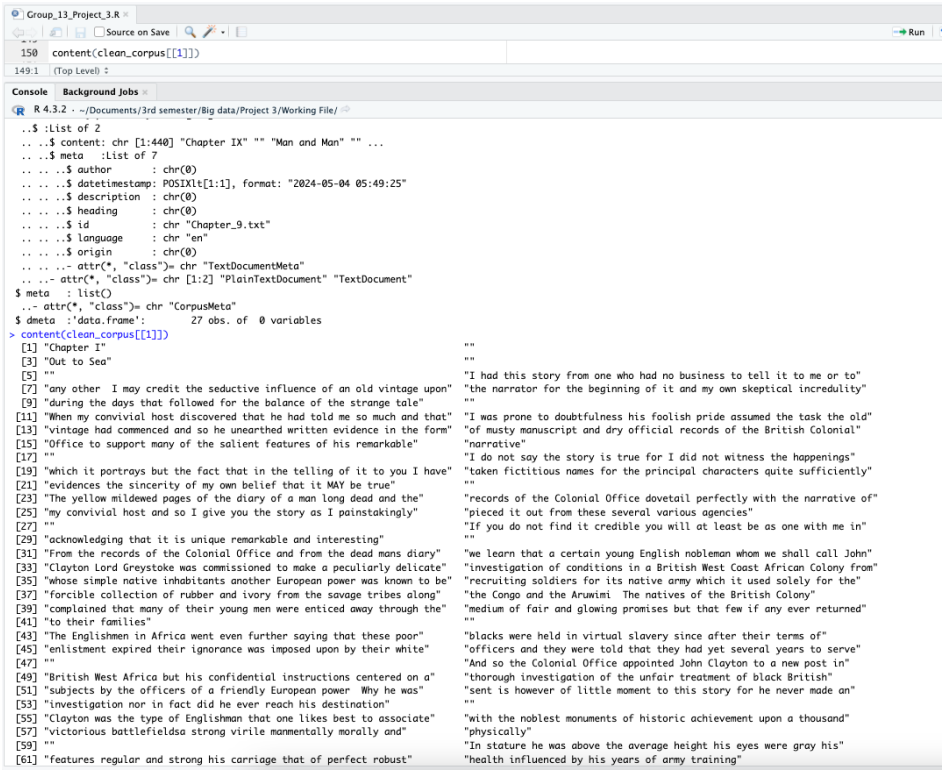


Writing a function to remove punctuation and numbers from each VCorpus file:



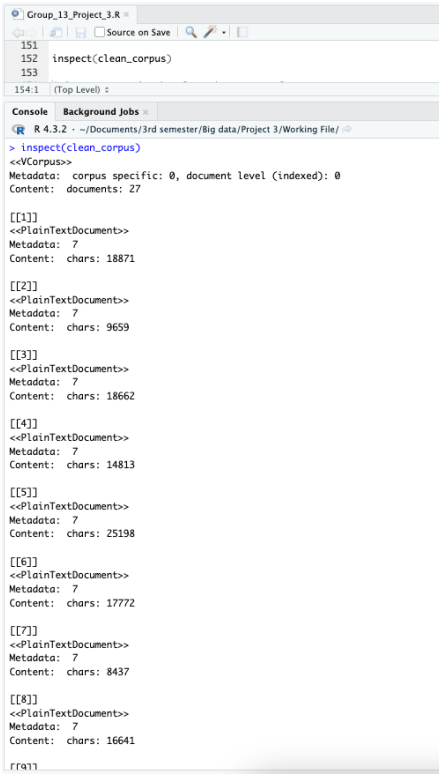
```
Group_13_Project_3.R
Source on Save
140
141 # Function to remove numbers and punctuation from each document in a VCorpus
142 remove_numbers_and_punctuation <- function(text) {
143   gsub("[^[:alpha:][:space:]]*", "", text)
144 }
145
146 # Apply the removal function to the VCorpus
147 clean_corpus = tm::tm_map(tarzan_apes_corpus, content_transformer(remove_numbers_and_punctuation))
148 str(clean_corpus)
149 content(clean_corpus[[1]])
148:18 (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Function to remove numbers and punctuation from each document in a VCorpus
> remove_numbers_and_punctuation <- function(text) {
+   gsub("[^[:alpha:][:space:]]*", "", text)
+ }
>
> # Apply the removal function to the VCorpus
> clean_corpus = tm::tm_map(tarzan_apes_corpus, content_transformer(remove_numbers_and_punctuation))
> str(clean_corpus)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 27
  ..$ :List of 2
  .. ..$ content: chr [1:421] "Chapter I" "" "Out to Sea" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 05:49:25"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "Chapter_1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:216] "Chapter X" "" "The FearPhantom" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 05:49:25"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "Chapter_10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:428] "Chapter XI" "" "King of the Apes" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 05:49:25"
```

Content of cleaned Chapter-1:



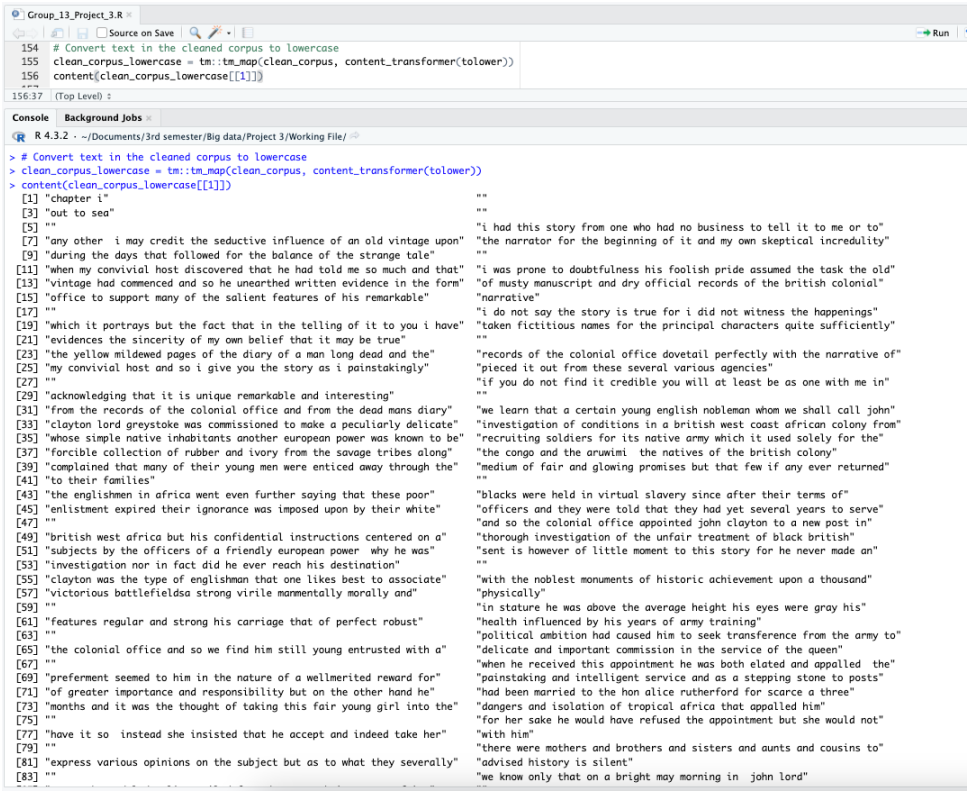
```
Group_13_Project_3.R
Source on Save
Run
150 content(clean_corpus[[1]])
149:1 (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
  ..$ :List of 2
  .. ..$ content: chr [1:440] "Chapter IX" "" "Man and Man" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 05:49:25"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "Chapter_9.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 $ meta   : list()
  ..- attr(*, "class")= chr "CorpusMeta"
 $ dmeta  :'data.frame':        27 obs. of  0 variables
> content(clean_corpus[[1]])
  [1] "Chapter I"                                                    ""
  [3] "Out to Sea"                                                   ""
  [5] ""                                                             "I had this story from one who had no business to tell it to me or to"
  [7] "any other  I may credit the seductive influence of an old vintage upon"  "the narrator for the beginning of it and my own skeptical incredulity"
  [9] "during the days that followed for the balance of the strange tale"        ""
 [11] "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"
 [13] "vintage had commenced and so he unearthed written evidence in the form"  "of musty manuscript and dry official records of the British Colonial"
 [15] "Office to support many of the salient features of his remarkable"        "narrative"
 [17] ""                                                                        "I do not say the story is true for I did not witness the happenings"
 [19] "which it portrays but the fact that in the telling of it to you I have"  "taken fictitious names for the principal characters quite sufficiently"
 [21] "evidences the sincerity of my own belief that it MAY be true"            ""
 [23] "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"
 [25] "my convivial host and so I give you the story as I painstakingly"        "pieced it out from these several various agencies"
 [27] ""                                                                        "If you do not find it credible you will at least be as one with me in"
 [29] "acknowledging that it is unique remarkable and interesting"              ""
 [31] "From the records of the Colonial Office and from the dead mans diary"    "we learn that a certain young English nobleman whom we shall call John"
 [33] "Clayton Lord Greystoke was commissioned to make a peculiarly delicate"   "investigation of conditions in a British West Coast African Colony from"
 [35] "whose simple native inhabitants another European power was known to be"  "recruiting soldiers for its native army which it used solely for the"
 [37] "forcible collection of rubber and ivory from the savage tribes along"    "the Congo and the Aruwimi  The natives of the British Colony"
 [39] "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"
 [41] "to their families"                                                       ""
 [43] "The Englishmen in Africa went even further saying that these poor"       "blacks were held in virtual slavery since after their terms of"
 [45] "enlistment expired their ignorance was imposed upon by their white"      "officers and they were told that they had yet several years to serve"
 [47] ""                                                                        "And so the Colonial Office appointed John Clayton to a new post in"
 [49] "British West Africa but his confidential instructions centered on a"     "thorough investigation of the unfair treatment of black British"
 [51] "subjects by the officers of a friendly European power  Why he was"       "sent is however of little moment to this story for he never made an"
 [53] "investigation nor in fact did he ever reach his destination"             ""
 [55] "Clayton was the type of Englishman that one likes best to associate"     "with the noblest monuments of historic achievement upon a thousand"
 [57] "victorious battlefieldsa strong virile manmentally morally and"          "physically"
 [59] ""                                                                        "In stature he was above the average height his eyes were gray his"
 [61] "features regular and strong his carriage that of perfect robust"         "health influenced by his years of army training"
```

Inspecting the Clean Corpus:



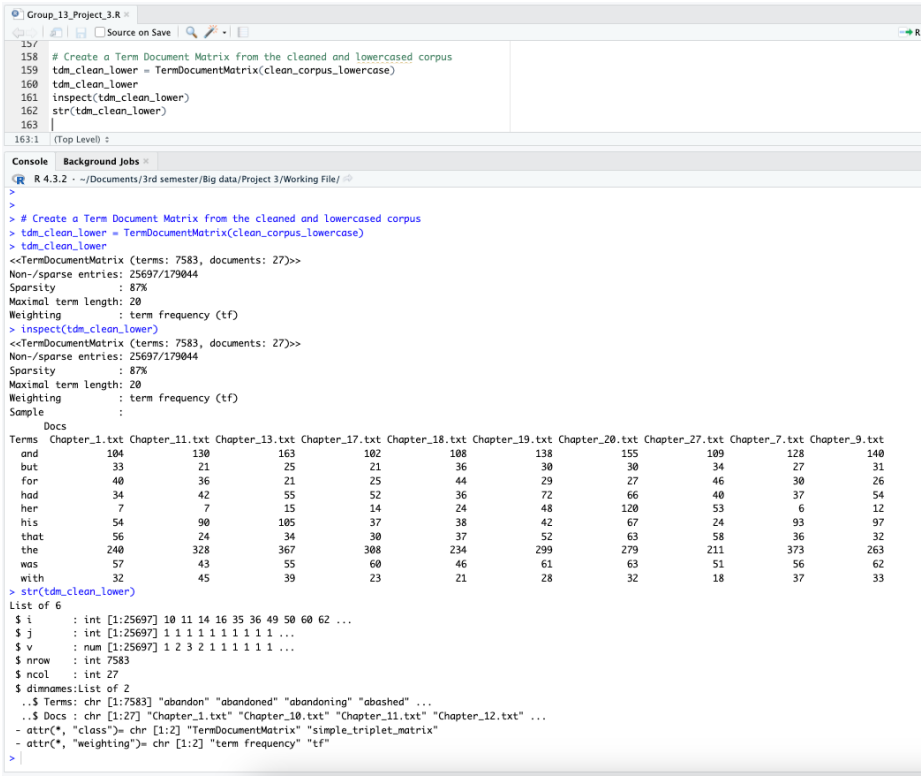
```
Group_13_Project_3.R
Source on Save
151
152  inspect(clean_corpus)
153
154:1  (Top Level)
Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> inspect(clean_corpus)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 27

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18871

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9659

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18662

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14813

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 25198

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 17772

[[7]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8437

[[8]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 16641

[[9]]
```

Converting the Corpus into lower case and viewing the first chapter:



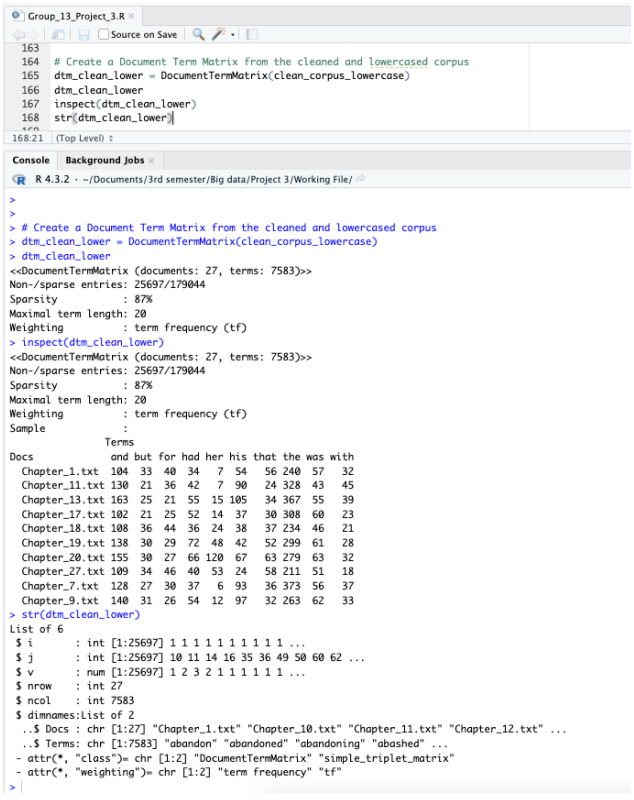
```
Group_13_Project_3.R
Source on Save
Run
154  # Convert text in the cleaned corpus to lowercase
155  clean_corpus_lowercase = tm::tm_map(clean_corpus, content_transformer(tolower))
156  content(clean_corpus_lowercase[[1]])
156:37  (Top Level)
Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Convert text in the cleaned corpus to lowercase
> clean_corpus_lowercase = tm::tm_map(clean_corpus, content_transformer(tolower))
> content(clean_corpus_lowercase[[1]])
  [1] "chapter i"                                                    ""
  [3] "out to sea"                                                   ""
  [5] ""                                                             "i had this story from one who had no business to tell it to me or to"
  [7] "any other  i may credit the seductive influence of an old vintage upon"  "the narrator for the beginning of it and my own skeptical incredulity"
  [9] "during the days that followed for the balance of the strange tale"       ""
 [11] "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"
 [13] "vintage had commenced and so he unearthed written evidence in the form"  "of musty manuscript and dry official records of the british colonial"
 [15] "office to support many of the salient features of his remarkable"        "narrative"
 [17] ""                                                            "i do not say the story is true for i did not witness the happenings"
 [19] "which it portrays but the fact that in the telling of it to you i have"  "taken fictitious names for the principal characters quite sufficiently"
 [21] "evidences the sincerity of my own belief that it may be true"            ""
 [23] "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"
 [25] "my convivial host and so i give you the story as i painstakingly"        "pieced it out from these several various agencies"
 [27] ""                                                            "if you do not find it credible you will at least be as one with me in"
 [29] "acknowledging that it is unique remarkable and interesting"              ""
 [31] "from the records of the colonial office and from the dead mans diary"    "we learn that a certain young english nobleman whom we shall call john"
 [33] "clayton lord greystoke was commissioned to make a peculiarly delicate"   "investigation of conditions in a british west coast african colony from"
 [35] "whose simple native inhabitants another european power was known to be"  "recruiting soldiers for its native army which it used solely for the"
 [37] "forcible collection of rubber and ivory from the savage tribes along"    "the congo and the aruwimi  the natives of the british colony"
 [39] "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"
 [41] "to their families"                                           ""
 [43] "the englishmen in africa went even further saying that these poor"       "blacks were held in virtual slavery since after their terms of"
 [45] "enlistment expired their ignorance was imposed upon by their white"      "officers and they were told that they had yet several years to serve"
 [47] ""                                                            "and so the colonial office appointed john clayton to a new post in"
 [49] "british west africa but his confidential instructions centered on a"     "thorough investigation of the unfair treatment of black british"
 [51] "subjects by the officers of a friendly european power  why he was"       "sent is however of little moment to this story for he never made an"
 [53] "investigation nor in fact did he ever reach his destination"             ""
 [55] "clayton was the type of englishman that one likes best to associate"     "with the noblest monuments of historic achievement upon a thousand"
 [57] "victorious battlefieldsa strong virile manmentally morally and"          "physically"
 [59] ""                                                            "in stature he was above the average height his eyes were gray his"
 [61] "features regular and strong his carriage that of perfect robust"         "health influenced by his years of army training"
 [63] ""                                                            "political ambition had caused him to seek transference from the army to"
 [65] "the colonial office and so we find him still young entrusted with a"     "delicate and important commission in the service of the queen"
 [67] ""                                                            "when he received this appointment he was both elated and appalled  the"
 [69] "preferment seemed to him in the nature of a wellmerited reward for"      "painstaking and intelligent service and as a stepping stone to posts"
 [71] "of greater importance and responsibility but on the other hand he"       "had been married to the hon alice rutherford for scarce a three"
 [73] "months and it was the thought of taking this fair young girl into the"   "dangers and isolation of tropical africa that appalled him"
 [75] ""                                                            "for her sake he would have refused the appointment but she would not"
 [77] "have it so  instead she insisted that he accept and indeed take her"     "with him"
 [79] ""                                                            "there were mothers and brothers and sisters and aunts and cousins to"
 [81] "express various opinions on the subject but as to what they severally"   "advised history is silent"
 [83] ""                                                            "we know only that on a bright may morning in  john lord"
```

Creating a TDM from the cleaned and lower-cased corpus:



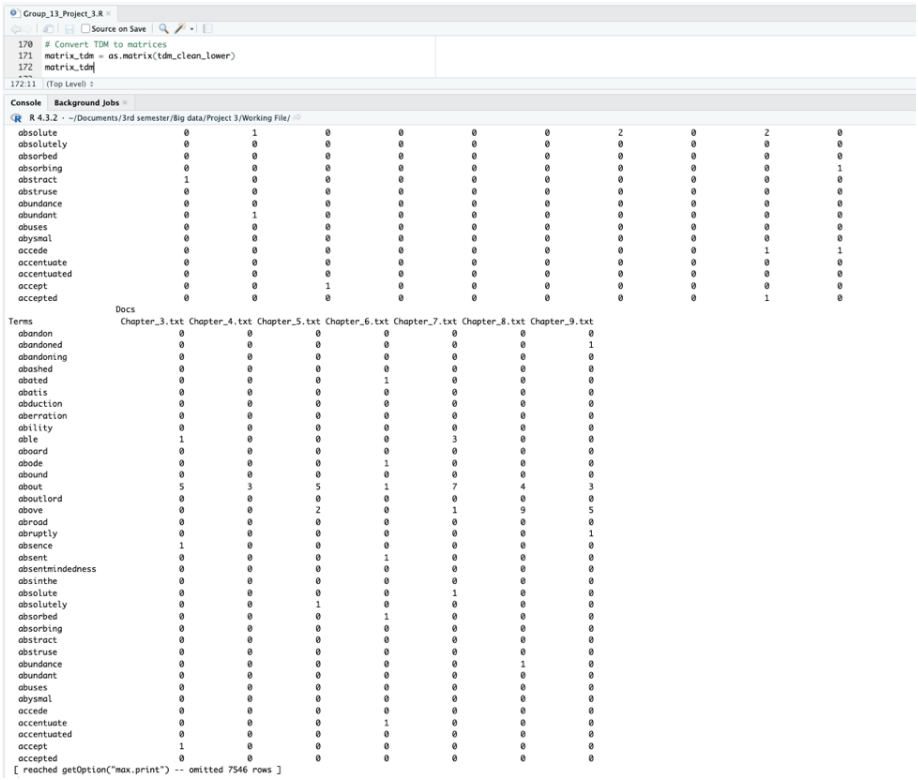
```
Group_13_Project_3.R
Source on Save
157
158  # Create a Term Document Matrix from the cleaned and lowercased corpus
159  tdm_clean_lower = TermDocumentMatrix(clean_corpus_lowercase)
160  tdm_clean_lower
161  inspect(tdm_clean_lower)
162  str(tdm_clean_lower)
163
163:1  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
>
>
> # Create a Term Document Matrix from the cleaned and lowercased corpus
> tdm_clean_lower = TermDocumentMatrix(clean_corpus_lowercase)
> tdm_clean_lower
<<TermDocumentMatrix (terms: 7583, documents: 27)>>
Non-/sparse entries: 25697/179044
Sparsity           : 87%
Maximal term length: 20
Weighting          : term frequency (tf)
> inspect(tdm_clean_lower)
<<TermDocumentMatrix (terms: 7583, documents: 27)>>
Non-/sparse entries: 25697/179044
Sparsity           : 87%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
      Docs
Terms  Chapter_1.txt Chapter_11.txt Chapter_13.txt Chapter_17.txt Chapter_18.txt Chapter_19.txt Chapter_20.txt Chapter_27.txt Chapter_7.txt Chapter_9.txt
  and            104            130            163            102            108            138            155            109            128            140
  but             33             21             25             21             36             30             30             34             27             31
  for             40             36             21             25             44             29             27             46             30             26
  had             34             42             55             52             36             72             66             40             37             54
  her              7              7             15             14             24             48            120             53              6             12
  his             54             90            105             37             38             42             67             24             93             97
  that            56             24             34             30             37             52             63             58             36             32
  the            240            328            367            308            234            299            279            211            373            263
  was             57             43             55             60             46             61             63             51             56             62
  with            32             45             39             23             21             28             32             18             37             33
> str(tdm_clean_lower)
List of 6
 $ i       : int [1:25697] 10 11 14 16 35 36 49 50 60 62 ...
 $ j       : int [1:25697] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:25697] 1 2 3 2 1 1 1 1 1 1 ...
 $ nrow    : int 7583
 $ ncol    : int 27
 $ dimnames:List of 2
  ..$ Terms: chr [1:7583] "abandon" "abandoned" "abandoning" "abashed" ...
  ..$ Docs : chr [1:27] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_12.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

Creating a DTM from the cleaned and lower-cased corpus:



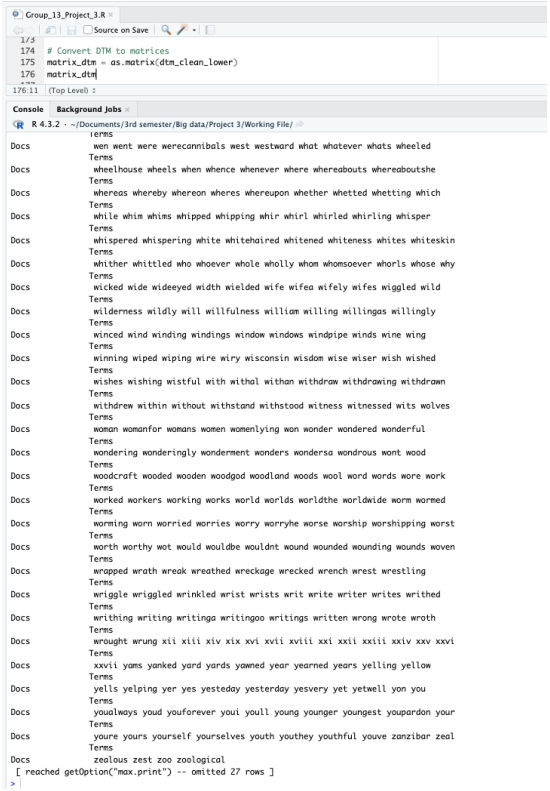
```
Group_13_Project_3.R
Source on Save
163
164  # Create a Document Term Matrix from the cleaned and lowercased corpus
165  dtm_clean_lower = DocumentTermMatrix(clean_corpus_lowercase)
166  dtm_clean_lower
167  inspect(dtm_clean_lower)
168  str(dtm_clean_lower)
168:21  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
>
>
> # Create a Document Term Matrix from the cleaned and lowercased corpus
> dtm_clean_lower = DocumentTermMatrix(clean_corpus_lowercase)
> dtm_clean_lower
<<DocumentTermMatrix (documents: 27, terms: 7583)>>
Non-/sparse entries: 25697/179044
Sparsity           : 87%
Maximal term length: 20
Weighting          : term frequency (tf)
> inspect(dtm_clean_lower)
<<DocumentTermMatrix (documents: 27, terms: 7583)>>
Non-/sparse entries: 25697/179044
Sparsity           : 87%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs            and but for had her his that the was with
  Chapter_1.txt  104  33  40  34   7  54   56 240  57   32
  Chapter_11.txt 130  21  36  42   7  90   24 328  43   45
  Chapter_13.txt 163  25  21  55  15 105   34 367  55   39
  Chapter_17.txt 102  21  25  52  14  37   30 308  60   23
  Chapter_18.txt 108  36  44  36  24  38   37 234  46   21
  Chapter_19.txt 138  30  29  72  48  42   52 299  61   28
  Chapter_20.txt 155  30  27  66 120  67   63 279  63   32
  Chapter_27.txt 109  34  46  40  53  24   58 211  51   18
  Chapter_7.txt  128  27  30  37   6  93   36 373  56   37
  Chapter_9.txt  140  31  26  54  12  97   32 263  62   33
> str(dtm_clean_lower)
List of 6
 $ i       : int [1:25697] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:25697] 10 11 14 16 35 36 49 50 60 62 ...
 $ v       : num [1:25697] 1 2 3 2 1 1 1 1 1 1 ...
 $ nrow    : int 27
 $ ncol    : int 7583
 $ dimnames:List of 2
  ..$ Docs : chr [1:27] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_12.txt" ...
  ..$ Terms: chr [1:7583] "abandon" "abandoned" "abandoning" "abashed" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

Converting TDM to matrix:



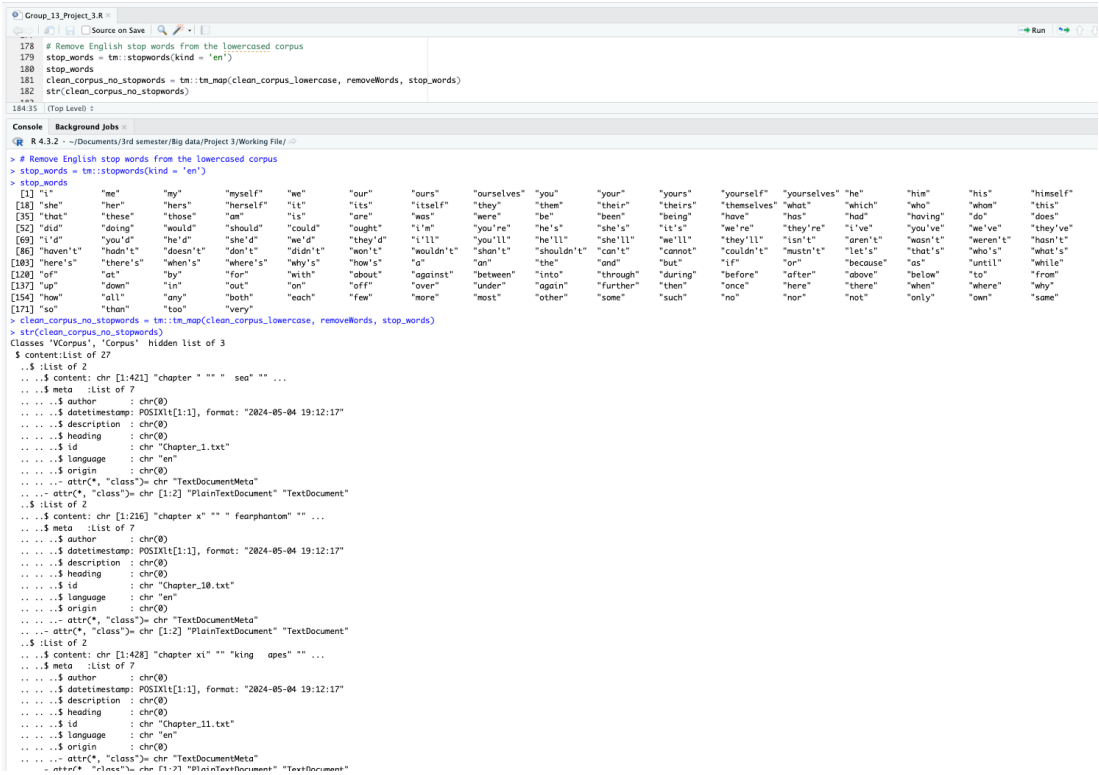
```
Group_13_Project_3.R
Source on Save
170  # Convert TDM to matrices
171  matrix_tdm = as.matrix(tdm_clean_lower)
172  matrix_tdm
172:11  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

  absolute                0          1          0          0          0          0          2          0          2          0
  absolutely              0          0          0          0          0          0          0          0          0          0
  absorbed                0          0          0          0          0          0          0          0          0          0
  absorbing               0          0          0          0          0          0          0          0          0          1
  abstract                1          0          0          0          0          0          0          0          0          0
  abstruse                0          0          0          0          0          0          0          0          0          0
  abundance               0          0          0          0          0          0          0          0          0          0
  abundant                0          1          0          0          0          0          0          0          0          0
  abuses                  0          0          0          0          0          0          0          0          0          0
  abysmal                 0          0          0          0          0          0          0          0          0          0
  accede                  0          0          0          0          0          0          0          0          1          1
  accentuate              0          0          0          0          0          0          0          0          0          0
  accentuated             0          0          0          0          0          0          0          0          0          0
  accept                  0          0          1          0          0          0          0          0          0          0
  accepted                0          0          0          0          0          0          0          0          1          0
                    Docs
Terms                Chapter_3.txt Chapter_4.txt Chapter_5.txt Chapter_6.txt Chapter_7.txt Chapter_8.txt Chapter_9.txt
  abandon                        0          0          0          0          0          0          0
  abandoned                      0          0          0          0          0          0          1
  abandoning                     0          0          0          0          0          0          0
  abashed                        0          0          0          0          0          0          0
  abated                         0          0          0          1          0          0          0
  abatis                         0          0          0          0          0          0          0
  abduction                      0          0          0          0          0          0          0
  aberration                     0          0          0          0          0          0          0
  ability                        0          0          0          0          0          0          0
  able                           1          0          0          0          3          0          0
  aboard                         0          0          0          0          0          0          0
  abode                          0          0          0          1          0          0          0
  abound                         0          0          0          0          0          0          0
  about                          5          3          5          1          7          4          3
  aboutlord                      0          0          0          0          0          0          0
  above                          0          0          2          0          1          9          5
  abroad                         0          0          0          0          0          0          0
  abruptly                       0          0          0          0          0          0          1
  absence                        1          0          0          0          0          0          0
  absent                         0          0          0          1          0          0          0
  absentmindedness               0          0          0          0          0          0          0
  absinthe                       0          0          0          0          0          0          0
  absolute                       0          0          0          0          1          0          0
  absolutely                     0          0          1          0          0          0          0
  absorbed                       0          0          0          1          0          0          0
  absorbing                      0          0          0          0          0          0          0
  abstract                       0          0          0          0          0          0          0
  abstruse                       0          0          0          0          0          0          0
  abundance                      0          0          0          0          0          1          0
  abundant                       0          0          0          0          0          0          0
  abuses                         0          0          0          0          0          0          0
  abysmal                        0          0          0          0          0          0          0
  accede                         0          0          0          0          0          0          0
  accentuate                     0          0          0          1          0          0          0
  accentuated                    0          0          0          0          0          0          0
  accept                         1          0          0          0          0          0          0
  accepted                       0          0          0          0          0          0          0
[ reached getOption("max.print") -- omitted 7546 rows ]
```

Converting DTM to matrix:



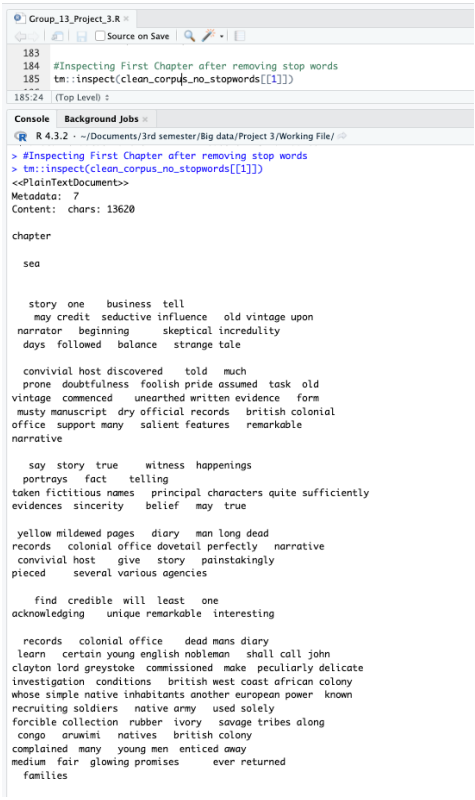
```
Group_13_Project_3.R
Source on Save
173
174  # Convert DTM to matrices
175  matrix_dtm = as.matrix(dtm_clean_lower)
176  matrix_dtm
176:11  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

                    Terms
Docs                 wen went were werecannibals west westward what whatever whats wheeled
                    Terms
Docs                 wheelhouse wheels when whence whenever where whereabouts whereaboutshe
                    Terms
Docs                 whereas whereby whereon wheres whereupon whether whetted whetting which
                    Terms
Docs                 while whim whims whipped whipping whir whirl whirled whirling whisper
                    Terms
Docs                 whispered whispering white whitehaired whitened whiteness whites whiteskin
                    Terms
Docs                 whither whittled who whoever whole wholly whom whomsoever whorls whose why
                    Terms
Docs                 wicked wide wideeyed width wielded wife wifea wifely wifes wiggled wild
                    Terms
Docs                 wilderness wildly will willfulness william willing willingas willingly
                    Terms
Docs                 winced wind winding windings window windows windpipe winds wine wing
                    Terms
Docs                 winning wiped wiping wire wiry wisconsin wisdom wise wiser wish wished
                    Terms
Docs                 wishes wishing wistful with withal withan withdraw withdrawing withdrawn
                    Terms
Docs                 withdrew within without withstand withstood witness witnessed wits wolves
                    Terms
Docs                 woman womanfor womans women womenlying won wonder wondered wonderful
                    Terms
Docs                 wondering wonderingly wonderment wonders wondersa wondrous wont wood
                    Terms
Docs                 woodcraft wooded wooden woodgod woodland woods wool word words wore work
                    Terms
Docs                 worked workers working works world worlds worldthe worldwide worm wormed
                    Terms
Docs                 worming worn worried worries worry worryhe worse worship worshipping worst
                    Terms
Docs                 worth worthy wot would wouldbe wouldnt wound wounded wounding wounds woven
                    Terms
Docs                 wrapped wrath wreak wreathed wreckage wrecked wrench wrest wrestling
                    Terms
Docs                 wriggle wriggled wrinkled wrist wrists writ write writer writes writhed
                    Terms
Docs                 writhing writing writinga writingoo writings written wrong wrote wroth
                    Terms
Docs                 wrought wrung xii xiii xiv xix xvi xvii xviii xxi xxii xxiii xxiv xxv xxvi
                    Terms
Docs                 xxvii yams yanked yard yards yawned year yearned years yelling yellow
                    Terms
Docs                 yells yelping yer yes yesteday yesterday yesvery yet yetwell yon you
                    Terms
Docs                 youalways youd youforever youi youll young younger youngest youpardon your
                    Terms
Docs                 youre yours yourself yourselves youth youthey youthful youve zanzibar zeal
                    Terms
Docs                 zealous zest zoo zoological
 [ reached getOption("max.print") -- omitted 27 rows ]
>
```

Removing Stop Words:



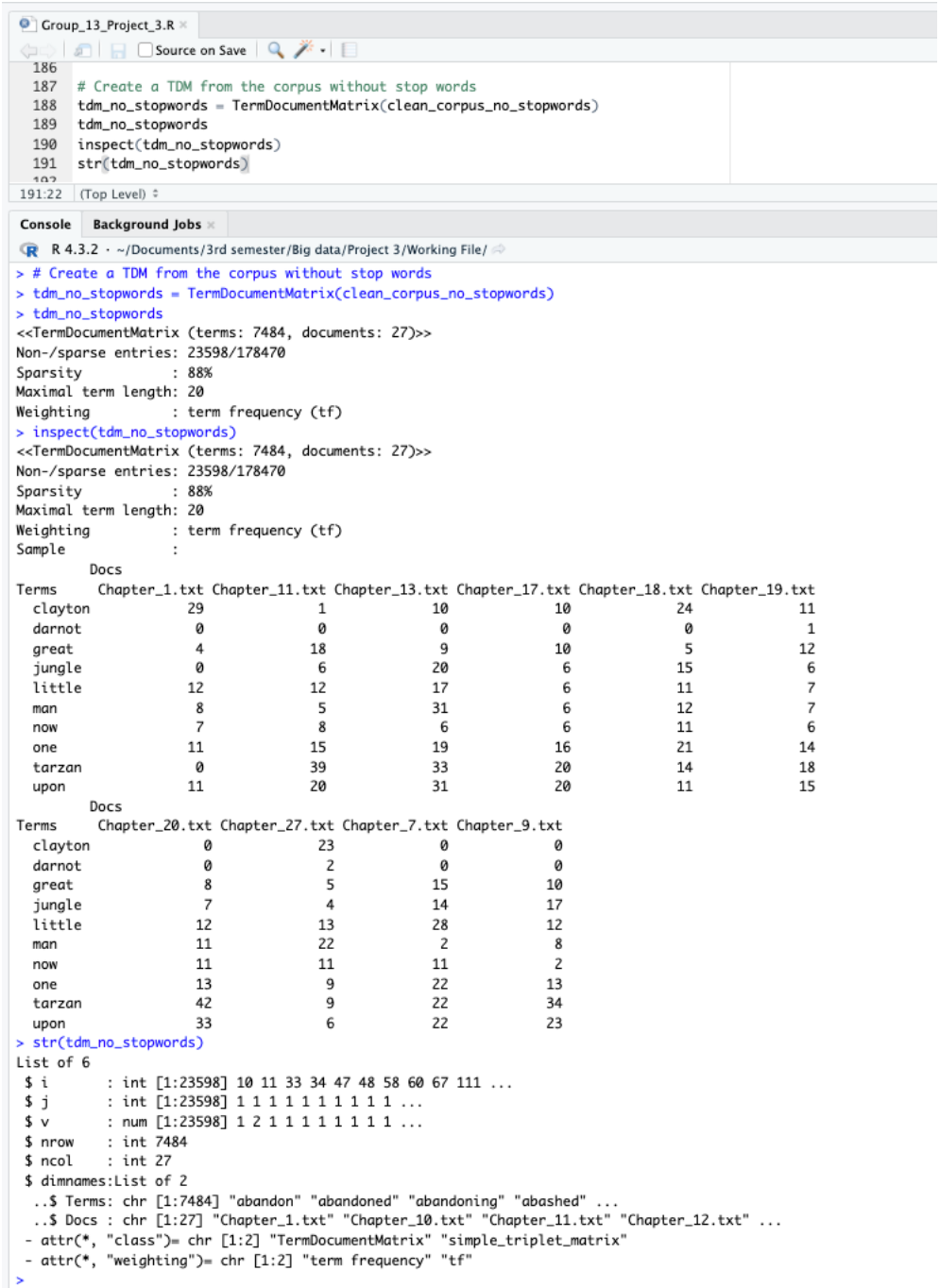
```
Group_13_Project_3.R
Source on Save                                                                    Run
178 # Remove English stop words from the lowercased corpus
179 stop_words = tm::stopwords(kind = 'en')
180 stop_words
181 clean_corpus_no_stopwords = tm::tm_map(clean_corpus_lowercase, removeWords, stop_words)
182 str(clean_corpus_no_stopwords)
184:35  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

> # Remove English stop words from the lowercased corpus
> stop_words = tm::stopwords(kind = 'en')
> stop_words
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"        "your"       "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"
 [18] "she"        "her"        "hers"       "herself"    "it"         "its"        "itself"     "they"       "them"       "their"      "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"
 [35] "that"       "these"      "those"      "am"         "is"         "are"        "was"        "were"       "be"         "been"       "being"      "have"       "has"        "had"        "having"     "do"         "does"
 [52] "did"        "doing"      "would"      "should"     "could"      "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"      "they're"    "i've"       "you've"     "we've"      "they've"
 [69] "i'd"        "you'd"      "he'd"       "she'd"      "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"    "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
 [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"     "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"      "that's"     "who's"      "what's"
[103] "here's"     "there's"    "when's"     "where's"    "why's"      "how's"      "a"          "an"         "the"        "and"        "but"        "if"         "or"         "because"    "as"         "until"      "while"
[120] "of"         "at"         "by"         "for"        "with"       "about"      "against"    "between"    "into"       "through"    "during"     "before"     "after"      "above"      "below"      "to"         "from"
[137] "up"         "down"       "in"         "out"        "on"         "off"        "over"       "under"      "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"      "why"
[154] "how"        "all"        "any"        "both"       "each"       "few"        "more"       "most"       "other"      "some"       "such"       "no"         "nor"        "not"        "only"       "own"        "same"
[171] "so"         "than"       "too"        "very"
> clean_corpus_no_stopwords = tm::tm_map(clean_corpus_lowercase, removeWords, stop_words)
> str(clean_corpus_no_stopwords)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 27
  ..$ :List of 2
  .. ..$ content: chr [1:421] "chapter " "" "  sea" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 19:12:17"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "Chapter_1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:216] "chapter x" "" " fearphantom" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 19:12:17"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "Chapter_10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:428] "chapter xi" "" "king   apes" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 19:12:17"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "Chapter_11.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
```

Inspecting First Chapter After removing Stop Words:



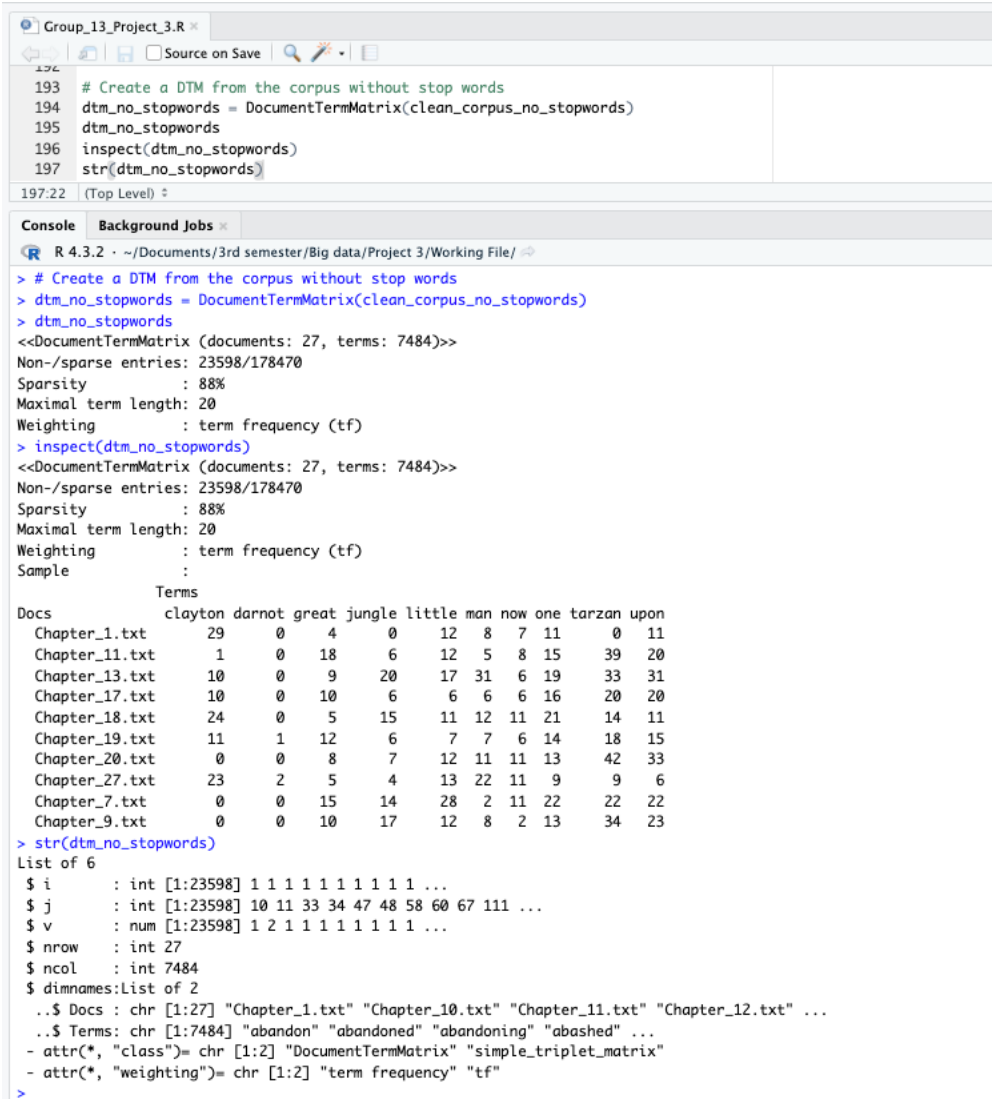
```
Group_13_Project_3.R
Source on Save
183
184 #Inspecting First Chapter after removing stop words
185 tm::inspect(clean_corpus_no_stopwords[[1]])
185:24  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

> #Inspecting First Chapter after removing stop words
> tm::inspect(clean_corpus_no_stopwords[[1]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13620

chapter

  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
 convivial host    give   story   painstakingly
pieced     several various agencies

    find  credible  will  least   one
acknowledging    unique remarkable  interesting

  records   colonial office    dead mans diary
 learn   certain young english nobleman   shall call john
clayton lord greystoke  commissioned  make  peculiarly delicate
investigation  conditions   british west coast african colony
whose simple native inhabitants another european power  known
recruiting soldiers   native army   used solely
forcible collection  rubber  ivory   savage tribes along
 congo   aruwimi   natives   british colony
complained  many   young men  enticed away
medium  fair  glowing promises      ever returned
  families
```

TDM from the corpus with stop words removed:



```
Group_13_Project_3.R
Source on Save
186
187  # Create a TDM from the corpus without stop words
188  tdm_no_stopwords = TermDocumentMatrix(clean_corpus_no_stopwords)
189  tdm_no_stopwords
190  inspect(tdm_no_stopwords)
191  str(tdm_no_stopwords)
192
191:22  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

> # Create a TDM from the corpus without stop words
> tdm_no_stopwords = TermDocumentMatrix(clean_corpus_no_stopwords)
> tdm_no_stopwords
<<TermDocumentMatrix (terms: 7484, documents: 27)>>
Non-/sparse entries: 23598/178470
Sparsity           : 88%
Maximal term length: 20
Weighting          : term frequency (tf)
> inspect(tdm_no_stopwords)
<<TermDocumentMatrix (terms: 7484, documents: 27)>>
Non-/sparse entries: 23598/178470
Sparsity           : 88%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
         Docs
Terms     Chapter_1.txt Chapter_11.txt Chapter_13.txt Chapter_17.txt Chapter_18.txt Chapter_19.txt
  clayton            29             1            10            10            24            11
  darnot              0             0             0             0             0             1
  great               4            18             9            10             5            12
  jungle              0             6            20             6            15             6
  little             12            12            17             6            11             7
  man                 8             5            31             6            12             7
  now                 7             8             6             6            11             6
  one                11            15            19            16            21            14
  tarzan              0            39            33            20            14            18
  upon               11            20            31            20            11            15
         Docs
Terms     Chapter_20.txt Chapter_27.txt Chapter_7.txt Chapter_9.txt
  clayton              0             23            0             0
  darnot               0              2            0             0
  great                8              5           15            10
  jungle               7              4           14            17
  little              12             13           28            12
  man                 11             22            2             8
  now                 11             11           11             2
  one                 13              9           22            13
  tarzan              42              9           22            34
  upon                33              6           22            23
> str(tdm_no_stopwords)
List of 6
 $ i       : int [1:23598] 10 11 33 34 47 48 58 60 67 111 ...
 $ j       : int [1:23598] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:23598] 1 2 1 1 1 1 1 1 1 1 ...
 $ nrow    : int 7484
 $ ncol    : int 27
 $ dimnames:List of 2
  ..$ Terms: chr [1:7484] "abandon" "abandoned" "abandoning" "abashed" ...
  ..$ Docs : chr [1:27] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_12.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

DTM of the corpus after stop words removed:



Group_13_Project_3.R

Source on Save

```
193  # Create a DTM from the corpus without stop words
194  dtm_no_stopwords = DocumentTermMatrix(clean_corpus_no_stopwords)
195  dtm_no_stopwords
196  inspect(dtm_no_stopwords)
197  str(dtm_no_stopwords)
```

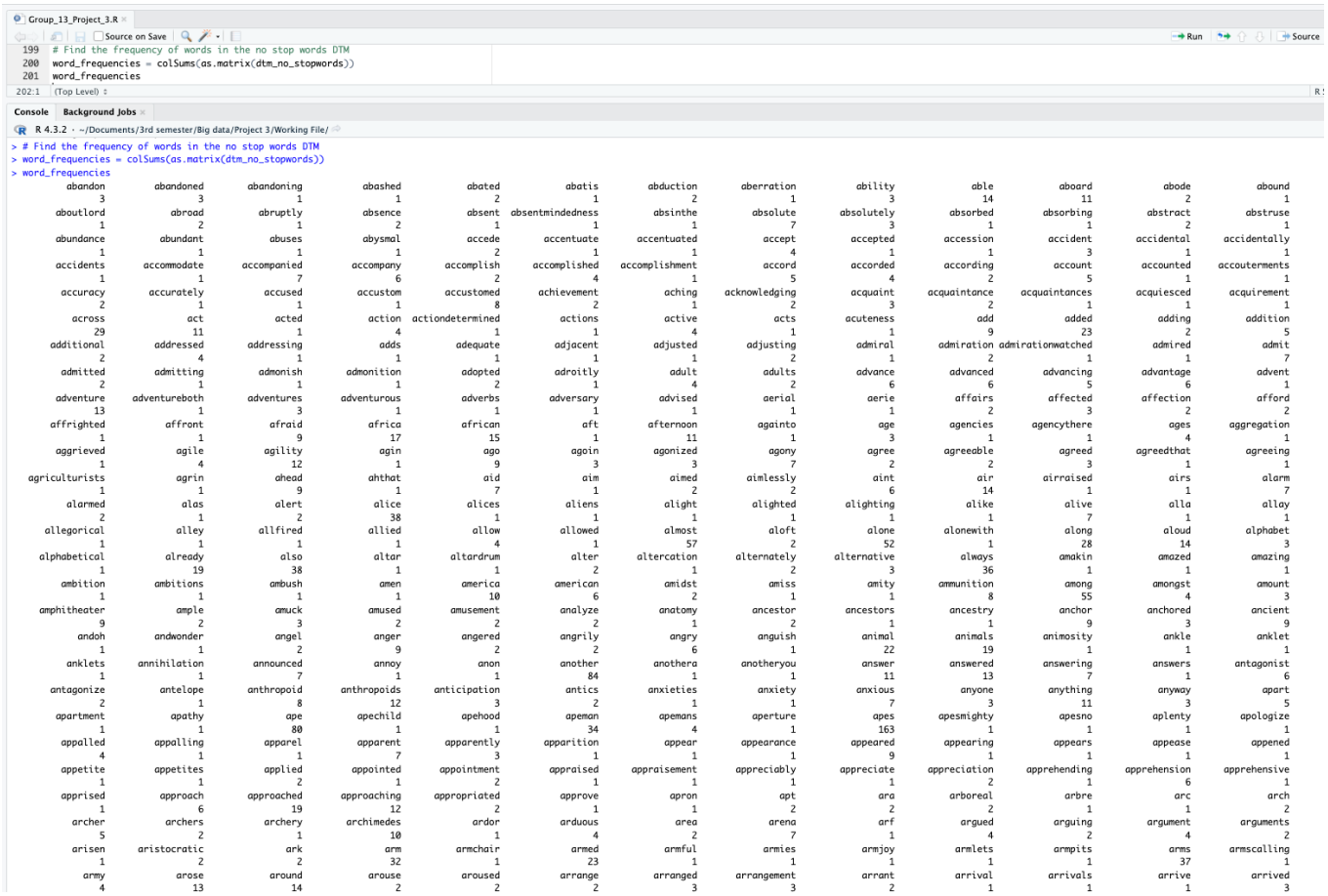
197:22 (Top Level)

Console Background Jobs

R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

```
> # Create a DTM from the corpus without stop words
> dtm_no_stopwords = DocumentTermMatrix(clean_corpus_no_stopwords)
> dtm_no_stopwords
<<DocumentTermMatrix (documents: 27, terms: 7484)>>
Non-/sparse entries: 23598/178470
Sparsity           : 88%
Maximal term length: 20
Weighting          : term frequency (tf)
> inspect(dtm_no_stopwords)
<<DocumentTermMatrix (documents: 27, terms: 7484)>>
Non-/sparse entries: 23598/178470
Sparsity           : 88%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs            clayton darnot great jungle little man now one tarzan upon
  Chapter_1.txt      29      0     4      0     12   8   7  11      0   11
  Chapter_11.txt      1      0    18      6     12   5   8  15     39   20
  Chapter_13.txt     10      0     9     20     17  31   6  19     33   31
  Chapter_17.txt     10      0    10      6      6   6   6  16     20   20
  Chapter_18.txt     24      0     5     15     11  12  11  21     14   11
  Chapter_19.txt     11      1    12      6      7   7   6  14     18   15
  Chapter_20.txt      0      0     8      7     12  11  11  13     42   33
  Chapter_27.txt     23      2     5      4     13  22  11   9      9    6
  Chapter_7.txt       0      0    15     14     28   2  11  22     22   22
  Chapter_9.txt       0      0    10     17     12   8   2  13     34   23
> str(dtm_no_stopwords)
List of 6
 $ i       : int [1:23598] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:23598] 10 11 33 34 47 48 58 60 67 111 ...
 $ v       : num [1:23598] 1 2 1 1 1 1 1 1 1 1 ...
 $ nrow    : int 27
 $ ncol    : int 7484
 $ dimnames:List of 2
  ..$ Docs : chr [1:27] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_12.txt" ...
  ..$ Terms: chr [1:7484] "abandon" "abandoned" "abandoning" "abashed" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

Frequency of the words in the corpus after stop words removed in DTM:



Group_13_Project_3.R

```
199  # Find the frequency of words in the no stop words DTM
200  word_frequencies = colSums(as.matrix(dtm_no_stopwords))
201  word_frequencies
```

Console | Background Jobs

R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

```
> # Find the frequency of words in the no stop words DTM
> word_frequencies = colSums(as.matrix(dtm_no_stopwords))
> word_frequencies
```

abandon	abandoned	abandoning	abashed	abated	abatis	abduction	aberration	ability	able	aboard	abode	abound
3	3	1	1	2	1	2	1	3	14	11	2	1
aboutlord	abroad	abruptly	absence	absent	absentmindedness	absinthe	absolute	absolutely	absorbed	absorbing	abstract	abstruse
1	2	1	2	1	1	1	7	3	1	1	2	1
abundance	abundant	abuses	abysmal	accede	accentuate	accentuated	accept	accepted	accession	accident	accidental	accidentally
1	1	1	1	2	1	1	4	1	1	3	1	1
accidents	accommodate	accompanied	accompany	accomplish	accomplished	accomplishment	accord	accorded	according	account	accounted	accouterments
1	1	7	6	2	4	1	5	4	2	5	1	1
accuracy	accurately	accused	accustom	accustomed	achievement	aching	acknowledging	acquaint	acquaintance	acquaintances	acquiesced	acquirement
2	1	1	1	8	2	1	2	3	2	1	1	1
across	act	acted	action	actiondetermined	actions	active	acts	acuteness	add	added	adding	addition
29	11	1	4	1	1	4	1	1	9	23	2	5
additional	addressed	addressing	adds	adequate	adjacent	adjusted	adjusting	admiral	admiration	admirationwatched	admired	admit
2	4	1	1	1	1	1	2	1	2	1	1	7
admitted	admitting	admonish	admonition	adopted	adroitly	adult	adults	advance	advanced	advancing	advantage	advent
2	1	1	1	2	1	4	2	6	6	5	6	1
adventure	adventureboth	adventures	adventurous	adverbs	adversary	advised	aerial	aerie	affairs	affected	affection	afford
13	1	3	1	1	1	1	1	1	2	3	2	2
affrighted	affront	afraid	africa	african	aft	afternoon	againto	age	agencies	agencythere	ages	aggregation
1	1	9	17	15	1	11	1	3	1	1	4	1
aggrieved	agile	agility	agin	ago	agoin	agonized	agony	agree	agreeable	agreed	agreedthat	agreeing
1	4	12	1	9	3	3	7	2	2	3	1	1
agriculturists	agrin	ahead	ahthat	aid	aim	aimed	aimlessly	aint	air	airraised	airs	alarm
1	1	9	1	7	1	2	2	6	14	1	1	7
alarmed	alas	alert	alice	alices	aliens	alight	alighted	alighting	alike	alive	alla	allay
2	1	2	38	1	1	1	1	1	1	7	1	1
allegorical	alley	allfired	allied	allow	allowed	almost	aloft	alone	alonewith	along	aloud	alphabet
1	1	1	1	4	1	57	2	52	1	28	14	3
alphabetical	already	also	altar	altardrum	alter	altercation	alternately	alternative	always	amakin	amazed	amazing
1	19	38	1	1	2	1	2	3	36	1	1	1
ambition	ambitions	ambush	amen	america	american	amidst	amiss	amity	ammunition	among	amongst	amount
1	1	1	1	10	6	2	1	1	8	55	4	3
amphitheater	ample	amuck	amused	amusement	analyze	anatomy	ancestor	ancestors	ancestry	anchor	anchored	ancient
9	2	3	2	2	2	1	2	1	1	9	3	9
andoh	andwonder	angel	anger	angered	angrily	angry	anguish	animal	animals	animosity	ankle	anklet
1	1	2	9	2	2	6	1	22	19	1	1	1
anklets	annihilation	announced	annoy	anon	another	anothera	anotheryou	answer	answered	answering	answers	antagonist
1	1	7	1	1	84	1	1	11	13	7	1	6
antagonize	antelope	anthropoid	anthropoids	anticipation	antics	anxieties	anxiety	anxious	anyone	anything	anyway	apart
2	1	8	12	3	2	1	1	7	3	11	3	5
apartment	apathy	ape	apechild	apehood	apeman	apemans	aperture	apes	apesmighty	apesno	aplenty	apologize
1	1	80	1	1	34	4	1	163	1	1	1	1
appalled	appalling	apparel	apparent	apparently	apparition	appear	appearance	appeared	appearing	appears	appease	appened
4	1	1	7	3	1	1	1	9	1	1	1	1
appetite	appetites	applied	appointed	appointment	appraised	appraisement	appreciably	appreciate	appreciation	apprehending	apprehension	apprehensive
1	1	2	1	2	1	1	1	1	2	1	6	1
apprised	approach	approached	approaching	appropriated	approve	apron	apt	ara	arboreal	arbre	arc	arch
1	6	19	12	2	1	1	2	2	2	1	1	2
archer	archers	archery	archimedes	ardor	arduous	area	arena	arf	argued	arguing	argument	arguments
5	2	1	10	1	4	2	7	1	4	2	4	2
arisen	aristocratic	ark	arm	armchair	armed	armful	armies	armjoy	armlets	armpits	arms	armscalling
1	2	2	32	1	23	1	1	1	1	1	37	1
army	arose	around	arouse	aroused	arrange	arranged	arrangement	arrant	arrival	arrivals	arrive	arrived
4	13	14	2	2	2	3	3	2	1	1	1	3

Frequency of the words in the corpus after stop words removed in TDM:

Group_13_Project_3.R

Source on Save

```
203 # Find the frequency of words in the no stop words TDM
204 frequent_terms = tm::findFreqTerms(tdm_no_stopwords)
205 frequent_terms
```

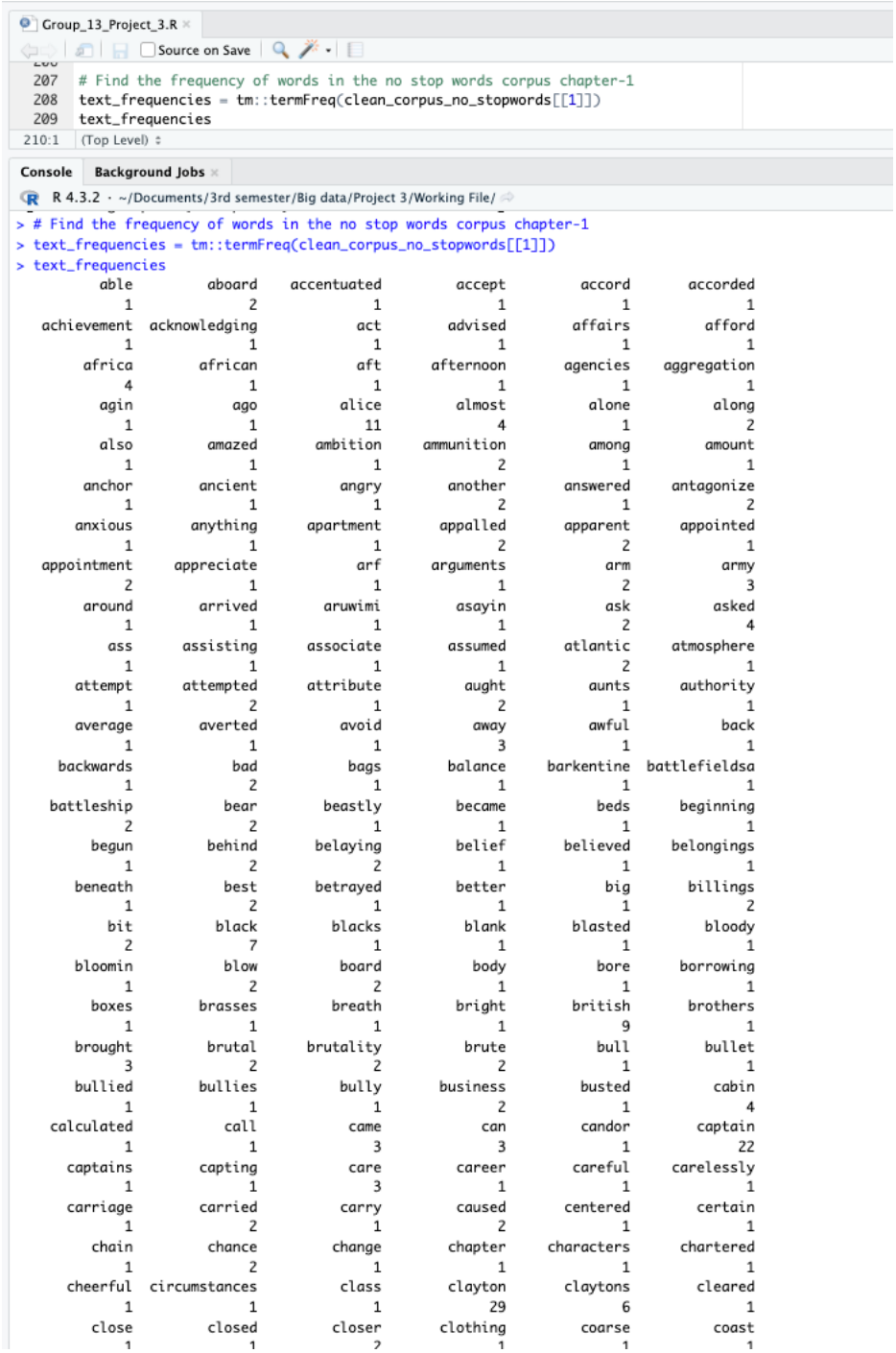
205:15 (Top Level)

Console | Background Jobs

R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

```
 [785] "breakfast"          "breaking"           "breaks"             "breast"
 [789] "breasts"            "breath"             "breathe"            "breathed"
 [793] "breathes"           "breathing"          "breathless"         "breathlessly"
 [797] "breaths"            "bred"               "breech"             "breechcloth"
 [801] "breeding"           "breeze"             "brief"              "bright"
 [805] "brightened"         "brightly"           "brilliancy"         "brilliant"
 [809] "brilliantly"        "bring"              "bringing"           "bristle"
 [813] "bristling"          "british"            "broad"              "broadminded"
 [817] "broke"              "broken"             "bronze"             "brooding"
 [821] "brook"              "brother"            "brotherhood"        "brothers"
 [825] "brought"            "brow"               "brown"              "brows"
 [829] "bruises"            "brush"              "brushed"            "brushing"
 [833] "brutal"             "brutality"          "brute"              "brutereason"
 [837] "brutes"             "brutish"            "brutishness"        "bubbled"
 [841] "bubbling"           "budged"             "bug"                "bugs"
 [845] "build"              "building"           "buildings"          "built"
 [849] "builthis"           "bulged"             "bulk"               "bull"
 [853] "bullet"             "bullets"            "bullied"            "bullies"
 [857] "bulls"              "bully"              "bullying"           "bunched"
 [861] "bundle"             "bundles"            "burden"             "burdens"
 [865] "burial"             "burials"            "buried"             "burly"
 [869] "burn"               "burned"             "burning"            "burnished"
 [873] "burst"              "bursting"           "bury"               "burying"
 [877] "bush"               "bushes"             "busily"             "business"
 [881] "businesslike"       "businessman"        "busted"             "busy"
 [885] "butcher"            "butchering"         "butshe"             "butt"
 [889] "butts"              "buying"             "buzz"               "bygone"
 [893] "cabbage"            "cabin"              "cable"              "cabled"
 [897] "cached"             "caches"             "cage"               "cakes"
 [901] "calamity"           "calculated"         "calculations"       "call"
 [905] "called"             "calling"            "calls"              "calm"
 [909] "calmly"             "came"               "camea"              "cameall"
 [913] "camp"               "can"                "candor"             "canine"
 [917] "canines"            "canler"             "canlers"            "canned"
 [921] "cannibal"           "cannibalism"        "cannibals"          "cannon"
 [925] "cannonade"          "cant"               "canvas"             "capacity"
 [929] "cape"               "capn"               "caprice"            "capricious"
 [933] "capriciousness"     "captain"            "captains"           "capting"
 [937] "captive"            "captor"             "captors"            "capture"
 [941] "captured"           "captures"           "car"                "carbines"
 [945] "carcass"            "card"               "cards"              "care"
 [949] "careening"          "career"             "careful"            "carefully"
 [953] "careless"           "carelessly"         "cares"              "caressing"
 [957] "carnivable"         "carnival"           "carnivora"          "carpenters"
 [961] "carpeted"           "carriage"           "carried"            "carriers"
 [965] "carrioneating"      "carry"              "carrying"           "cartridge"
 [969] "cartridges"         "carved"             "carven"             "case"
 [973] "casks"              "cassava"            "cast"               "castaway"
 [977] "castaways"          "casting"            "cat"                "catch"
 [981] "catching"           "caterpillar"        "catlike"            "cats"
 [985] "caucasian"          "caught"             "cauldron"           "cause"
 [989] "caused"             "caution"            "cautioned"          "cautiously"
 [993] "cavalcade"          "cease"              "ceased"             "ceaseless"
 [997] "ceaselessly"        "cecil"              "cecilmr"            "ceiling"
 [ reached getOption("max.print") -- omitted 6484 entries ]
>
```

Frequency of the words in the corpus after stop words removed in the corpus Chapter 1:



Group_13_Project_3.R

Source on Save

```
207  # Find the frequency of words in the no stop words corpus chapter-1
208  text_frequencies = tm::termFreq(clean_corpus_no_stopwords[[1]])
209  text_frequencies
```

210:1 (Top Level)

Console | Background Jobs

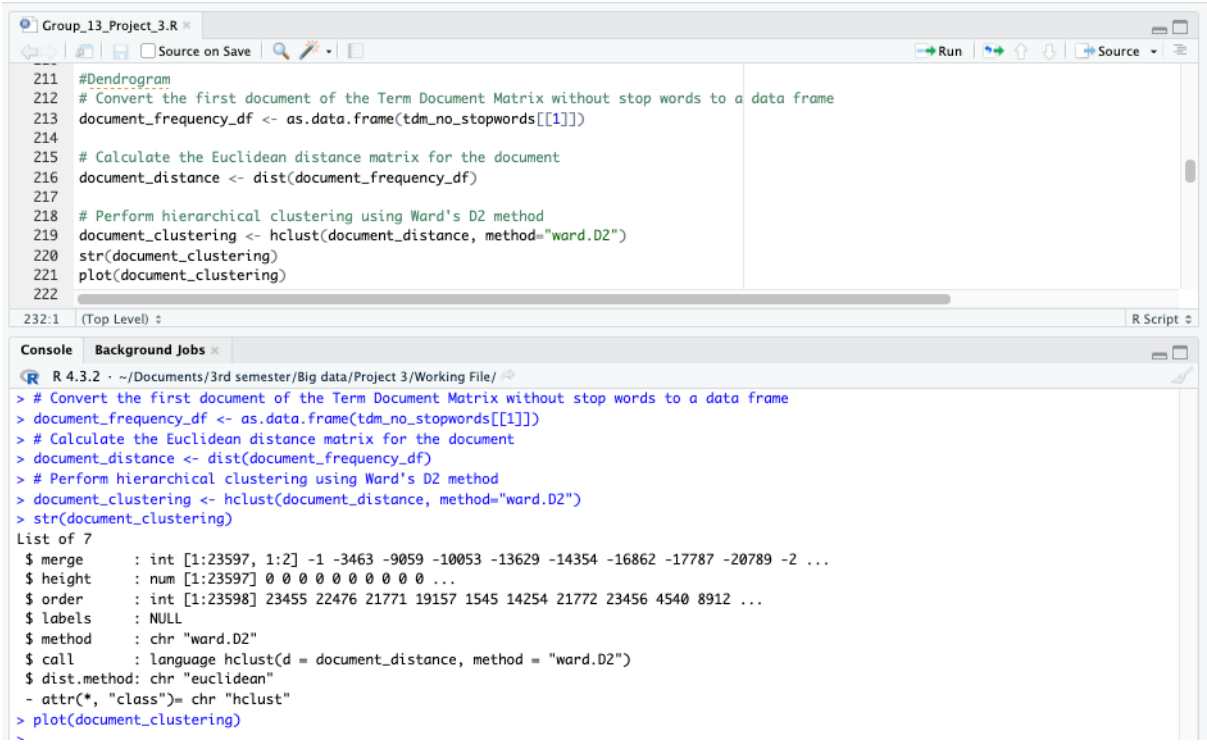
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

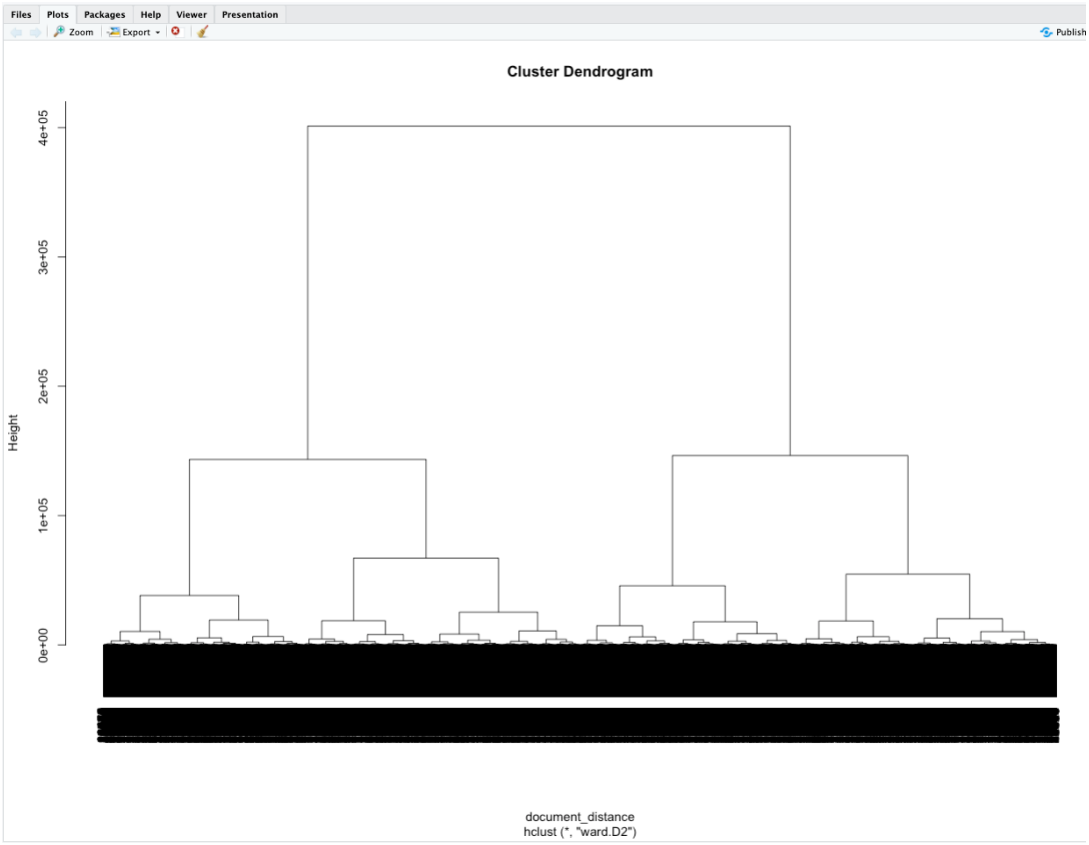
```
> # Find the frequency of words in the no stop words corpus chapter-1
> text_frequencies = tm::termFreq(clean_corpus_no_stopwords[[1]])
> text_frequencies
```

able	aboard	accentuated	accept	accord	accorded
1	2	1	1	1	1
achievement	acknowledging	act	advised	affairs	afford
1	1	1	1	1	1
africa	african	aft	afternoon	agencies	aggregation
4	1	1	1	1	1
agin	ago	alice	almost	alone	along
1	1	11	4	1	2
also	amazed	ambition	ammunition	among	amount
1	1	1	2	1	1
anchor	ancient	angry	another	answered	antagonize
1	1	1	2	1	2
anxious	anything	apartment	appalled	apparent	appointed
1	1	1	2	2	1
appointment	appreciate	arf	arguments	arm	army
2	1	1	1	2	3
around	arrived	aruwimi	asayin	ask	asked
1	1	1	1	2	4
ass	assisting	associate	assumed	atlantic	atmosphere
1	1	1	1	2	1
attempt	attempted	attribute	aught	aunts	authority
1	2	1	2	1	1
average	averted	avoid	away	awful	back
1	1	1	3	1	1
backwards	bad	bags	balance	barkentine	battlefieldsa
1	2	1	1	1	1
battleship	bear	beastly	became	beds	beginning
2	2	1	1	1	1
begun	behind	belaying	belief	believed	belongings
1	2	2	1	1	1
beneath	best	betrayed	better	big	billings
1	2	1	1	1	2
bit	black	blacks	blank	blasted	bloody
2	7	1	1	1	1
bloomin	blow	board	body	bore	borrowing
1	2	2	1	1	1
boxes	brasses	breath	bright	british	brothers
1	1	1	1	9	1
brought	brutal	brutality	brute	bull	bullet
3	2	2	2	1	1
bullied	bullies	bully	business	busted	cabin
1	1	1	2	1	4
calculated	call	came	can	candor	captain
1	1	3	3	1	22
captains	capting	care	career	careful	carelessly
1	1	3	1	1	1
carriage	carried	carry	caused	centered	certain
1	2	1	2	1	1
chain	chance	change	chapter	characters	chartered
1	2	1	1	1	1
cheerful	circumstances	class	clayton	claytons	cleared
1	1	1	29	6	1
close	closed	closer	clothing	coarse	coast
1	1	2	1	1	1

Create a Dendrogram:

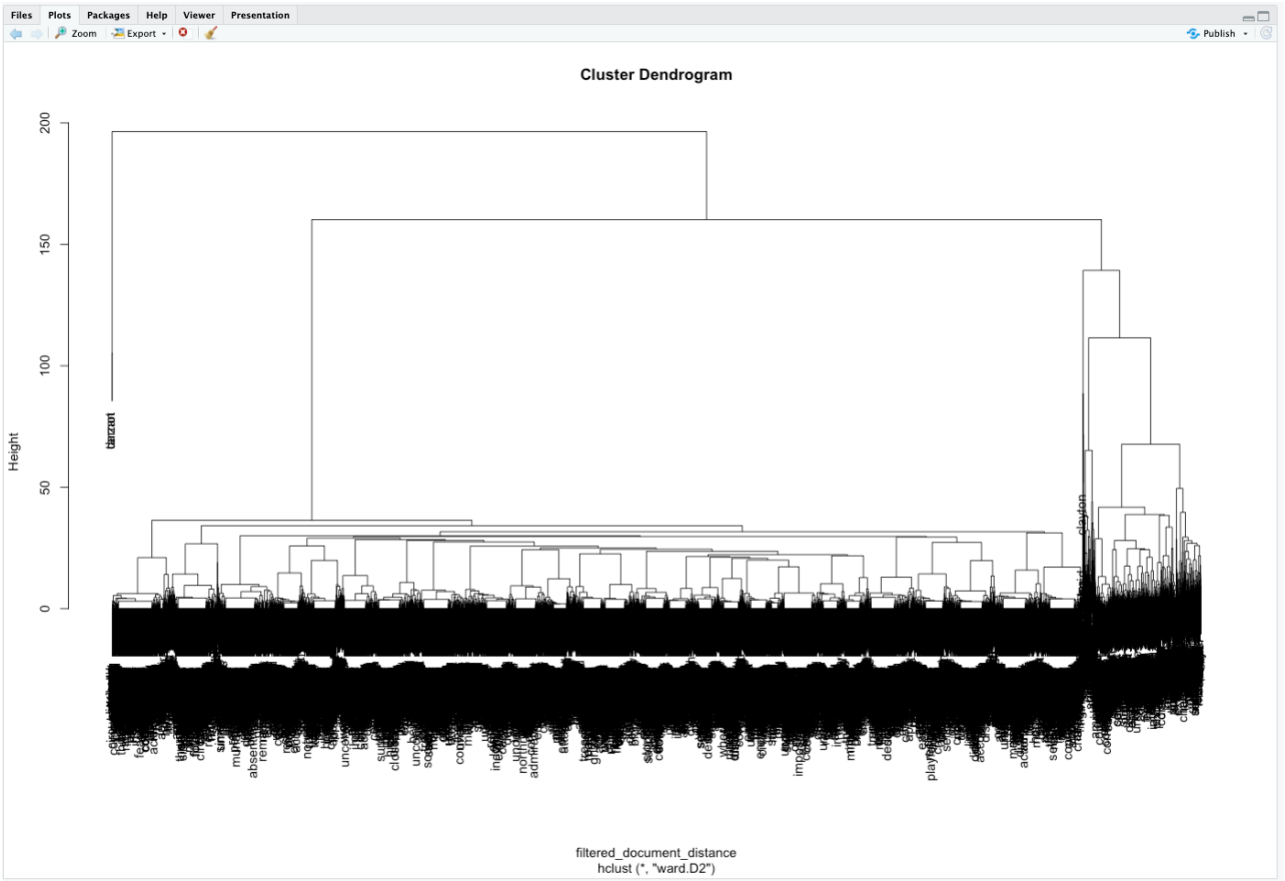
- Convert the first chapter of TDM without stop words to data frame.
- Calculate Euclidean distance matrix for the data frame.
- Perform hierarchical clustering using Ward's D2 method.





The TDM is converted into a matrix format, which allows for numerical manipulation. We then calculate the sparsity for each word across all documents—sparsity being the proportion of documents that contain the term. We set a threshold at 0.9, targeting words that appear in less than 90% of the documents to avoid overly common terms and to focus on those that might provide unique insights into the dataset. Words meeting this criterion are retained in a new filtered TDM. Following this, a Euclidean distance matrix is computed from the filtered TDM, which quantifies the differences between documents based on their term frequencies. This distance matrix serves as the basis for hierarchical clustering using Ward's method, effectively grouping documents into clusters that share similar content. The process culminates with the plotting of a dendrogram, a visual representation of the clustering results that illustrates the relationships and hierarchical grouping among the documents based on their textual content. This technique is particularly useful for identifying natural groupings in text data, aiding in tasks such as document classification or thematic organization.


Group_13_Project_3.R
Source on Save
Run
Source
# Dendrogram after removing words with high sparsity
# Convert the TDM to a matrix
tdm_matrix <- as.matrix(tdm_no_stopwords)
# Set sparsity threshold (e.g., 0.9 means words that appear in less than 99% of the documents are kept)
sparsity_threshold <- 0.9
# Calculate sparsity for each word
sparsity_values <- rowSums(tdm_matrix > 0) / ncol(tdm_matrix)
# Select words with sparsity below the threshold
low_sparsity_words <- which(sparsity_values < sparsity_threshold)
# Create a new TDM with the selected low sparsity words
filtered_tdm <- tdm_matrix[low_sparsity_words,]
# Convert the filtered TDM to a data frame
filtered_document_df <- as.data.frame(filtered_tdm)
# Calculate the distance matrix
filtered_document_distance <- dist(filtered_document_df)
# Perform hierarchical clustering using Ward's method on the filtered data
filtered_document_clustering <- hclust(filtered_document_distance, method = "ward.D2")
str(filtered_document_clustering)
plot(filtered_document_clustering)
250:1
(Top Level)
R Script
Console
Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Dendrogram after removing words with high sparsity
> # Convert the TDM to a matrix
> tdm_matrix <- as.matrix(tdm_no_stopwords)
> # Set sparsity threshold (e.g., 0.9 means words that appear in less than 99% of the documents are kept)
> sparsity_threshold <- 0.9
> # Calculate sparsity for each word
> sparsity_values <- rowSums(tdm_matrix > 0) / ncol(tdm_matrix)
> # Select words with sparsity below the threshold
> low_sparsity_words <- which(sparsity_values < sparsity_threshold)
> # Create a new TDM with the selected low sparsity words
> filtered_tdm <- tdm_matrix[low_sparsity_words,]
> # Convert the filtered TDM to a data frame
> filtered_document_df <- as.data.frame(filtered_tdm)
> # Calculate the distance matrix
> filtered_document_distance <- dist(filtered_document_df)
> # Perform hierarchical clustering using Ward's method on the filtered data
> filtered_document_clustering <- hclust(filtered_document_distance, method = "ward.D2")
> str(filtered_document_clustering)
List of 7
 $ merge   : int [1:7447, 1:2] -3 -38 -46 -119 -166 -175 -197 -380 -389 -489 ...
 $ height   : num [1:7447] 0 0 0 0 0 0 0 0 0 0 ...
 $ order   : int [1:7448] 1548 6451 3739 430 853 5487 1089 2769 580 1782 ...
 $ labels   : chr [1:7448] "abandon" "abandoned" "abandoning" "abashed" ...
 $ method   : chr "ward.D2"
 $ call   : language hclust(d = filtered_document_distance, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> plot(filtered_document_clustering)
>



Word Cloud:

Word Cloud is a visual representation of word frequency data. It begins by setting word_frequencies to the term frequencies extracted from a previously processed text, presumably "Tarzan of the Apes." A color palette named color_palette is defined using the brewer.pal function, which selects nine colors from the "Spectral" scheme. The wordcloud function then generates the word cloud, using the words and their frequencies as input. Each word's size in the cloud is proportional to its frequency in the text, and the words are colored according to the specified palette. This visualization helps in quickly perceiving the most prominent words within the text, providing insights into the themes and focus areas of "Tarzan of the Apes.

Group_13_Project_3.R

Source on Save

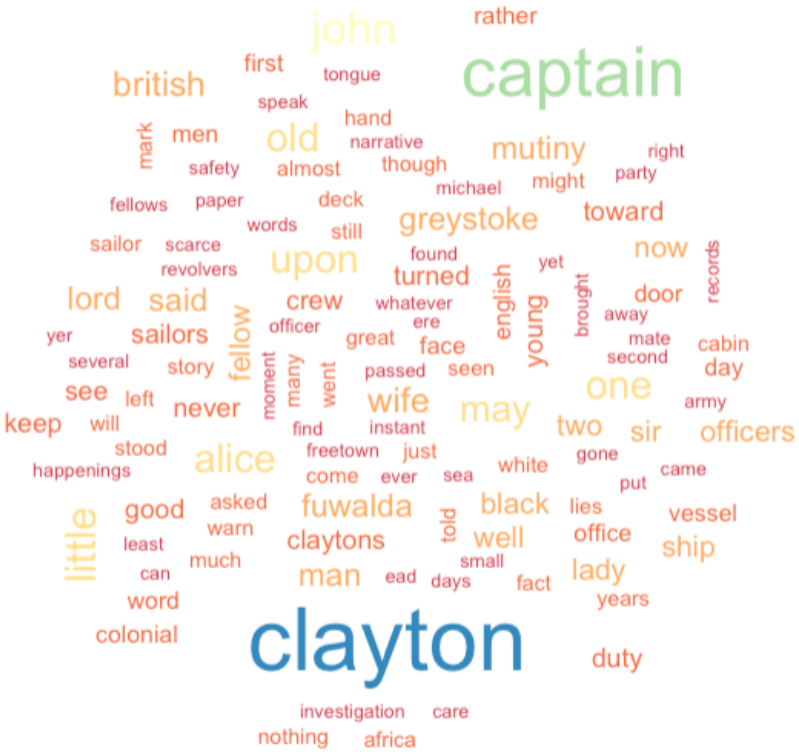
```
249
250 # Word Cloud
251 word_frequencies <- word_frequencies
252 color_palette <- brewer.pal(9, "Spectral")  # Define a color palette
253 word_cloud <- wordcloud(names(word_frequencies), word_frequencies, colors = color_palette)
```

255:1 (Top Level)

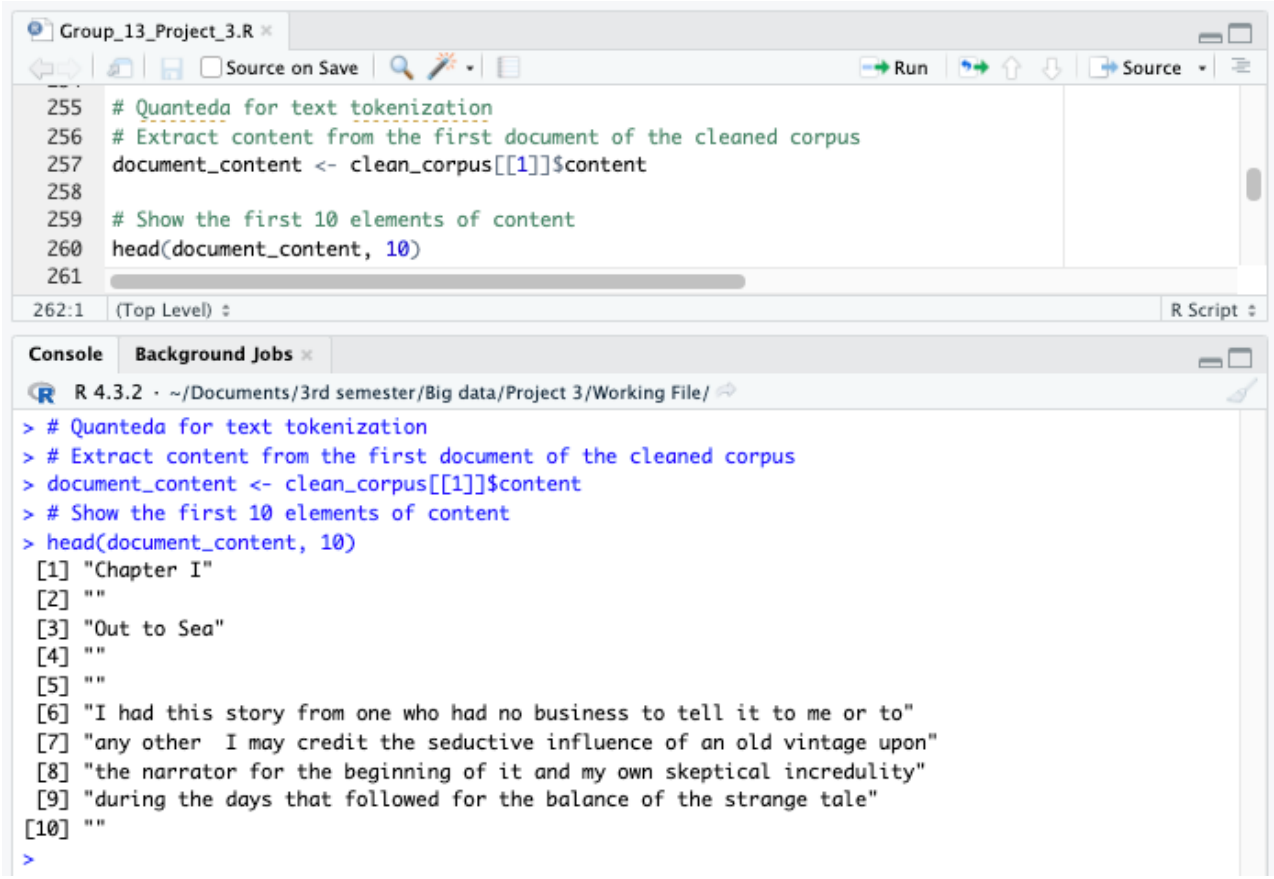
Console | Background Jobs

R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

```
> # Word Cloud
> word_frequencies <- word_frequencies
> color_palette <- brewer.pal(9, "Spectral")  # Define a color palette
> word_cloud <- wordcloud(names(word_frequencies), word_frequencies, colors = color_palette)
>
>
```



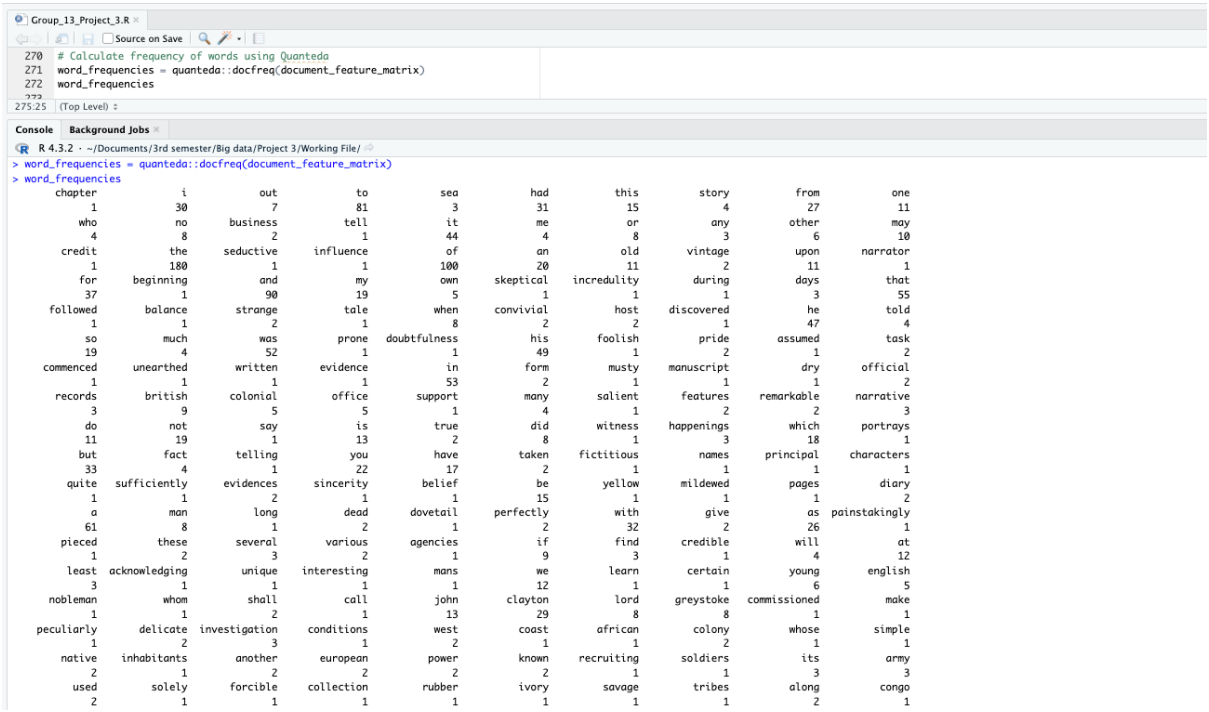
Quanteda package specifically focuses on extracting and displaying content from the first document of a cleaned text corpus. The variable document_content retrieves the content of the first document from a cleaned corpus. The function head(document_content, 10) is then used to display the first 10 elements of this content, which typically include the initial sections or fragments of the text. This can be helpful for a quick inspection or analysis of the starting portion of the document to understand the context or content structure before proceeding with deeper text analysis tasks.



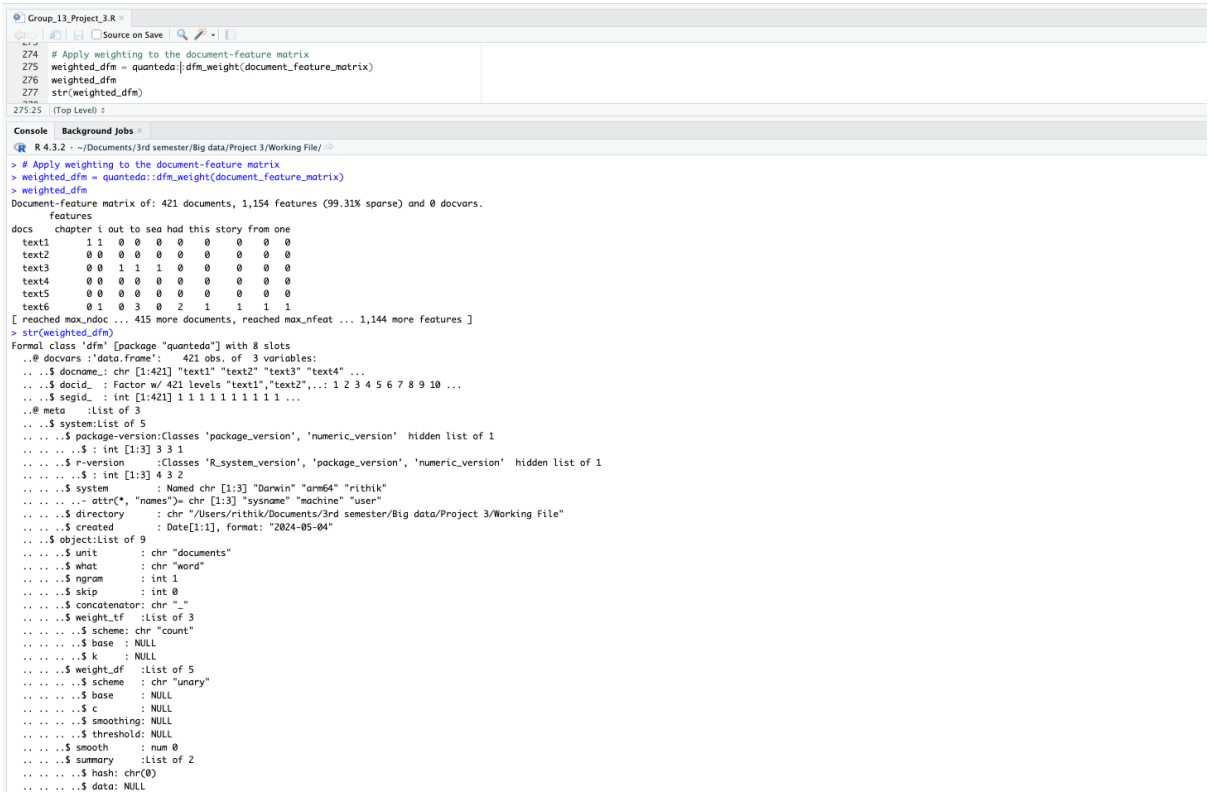
Tokenize the content and convert tokens to matrix using Quanteda:

```
Group_13_Project_3.R
Source on Save
262  # Tokenize the content
263  TarzanApesToken <- quanteda::tokens(document_content)
264  TarzanApesToken
265
266  # Convert tokens to a document-feature matrix using Quanteda
267  document_feature_matrix = quanteda::dfm(TarzanApesToken)
268  str(document_feature_matrix)
270:1  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Tokenize the content
> TarzanApesToken <- quanteda::tokens(document_content)
> TarzanApesToken
Tokens consisting of 421 documents.
text1 :
[1] "Chapter" "I"

text2 :
character(0)

text3 :
[1] "Out" "to"  "Sea"

text4 :
character(0)

text5 :
character(0)

text6 :
 [1] "I"        "had"      "this"     "story"    "from"     "one"      "who"      "had"
 [9] "no"       "business" "to"       "tell"
[ ... and 5 more ]

[ reached max_ndoc ... 415 more documents ]
> # Convert tokens to a document-feature matrix using Quanteda
> document_feature_matrix = quanteda::dfm(TarzanApesToken)
> str(document_feature_matrix)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    421 obs. of  3 variables:
  .. ..$ docname_: chr [1:421] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 421 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:421] 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] 3 3 1
  .. .. ..$ 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] "Darwin" "arm64" "rithik"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 9
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
            $ k     : NULL
```

Calculating frequency of the words using Quanteda and applying to the document-feature matrix:

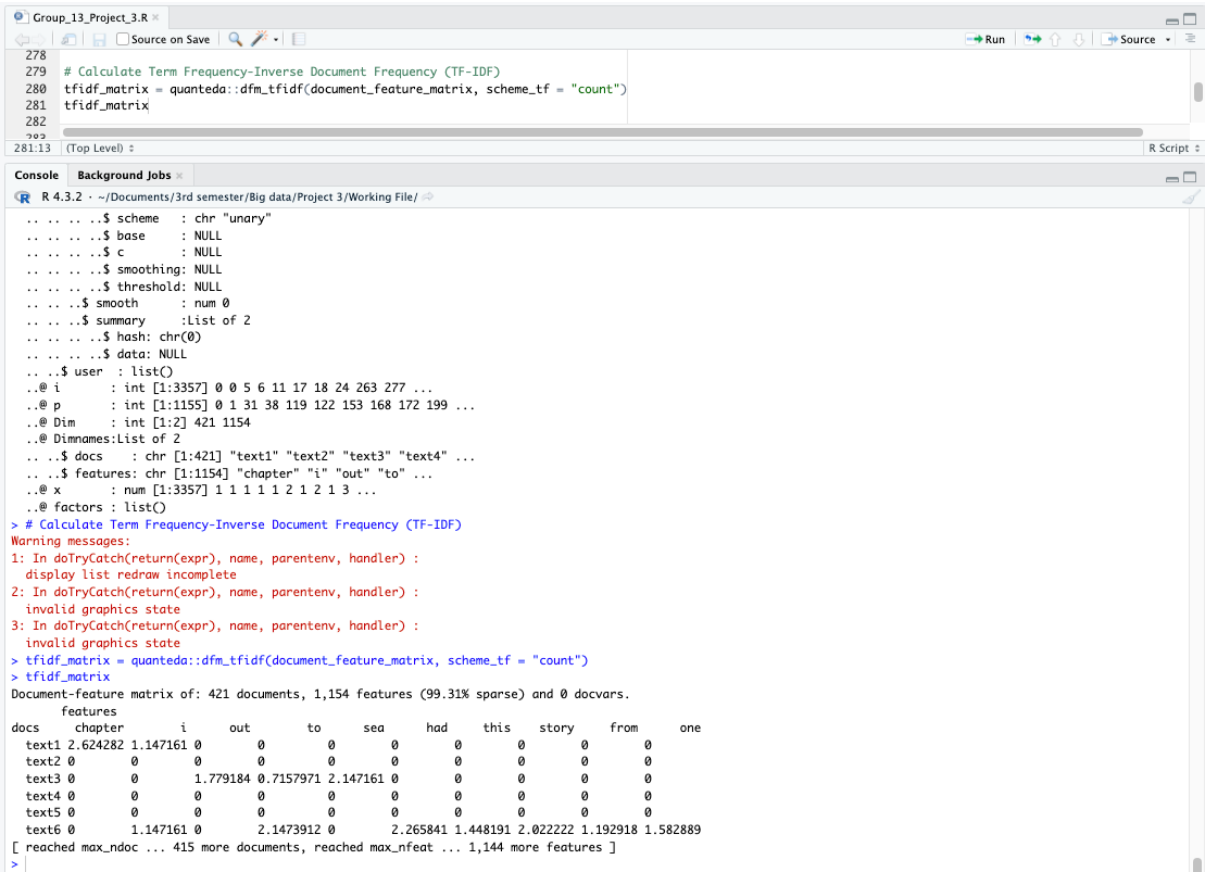


```
Group_13_Project_3.R
Source on Save
270  # Calculate frequency of words using Quanteda
271  word_frequencies = quanteda::docfreq(document_feature_matrix)
272  word_frequencies
275:25  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> word_frequencies = quanteda::docfreq(document_feature_matrix)
> word_frequencies
       chapter            i          out           to          sea          had         this        story         from          one
             1           30            7           81            3           31           15            4           27           11
           who           no     business         tell           it           me           or          any        other          may
             4            8            2            1           44            4            8            3            6           10
        credit          the    seductive    influence           of           an          old      vintage         upon     narrator
             1          180            1            1          100           20           11            2           11            1
           for    beginning          and           my          own     skeptical   incredulity       during         days         that
            37            1           90           19            5            1            1            1            3           55
      followed      balance      strange         tale         when     convivial         host   discovered           he         told
             1            1            2            1            8            2            2            1           47            4
            so         much          was        prone   doubtfulness          his      foolish        pride      assumed         task
            19            4           52            1            1           49            1            2            1            2
    commenced    unearthed      written     evidence           in         form        musty   manuscript          dry     official
             1            1            1            1           53            2            1            1            1            2
       records      british     colonial       office      support         many      salient     features    remarkable     narrative
             3            9            5            5            1            4            1            2            2            3
            do          not          say           is         true          did      witness   happenings        which     portrays
            11           19            1           13            2            8            1            3           18            1
           but         fact      telling          you         have        taken    fictitious        names    principal   characters
            33            4            1           22           17            2            1            1            1            1
         quite  sufficiently    evidences    sincerity     belief           be       yellow     mildewed        pages        diary
             1            1            2            1            1           15            1            1            1            2
             a          man         long         dead     dovetail    perfectly         with         give           as  painstakingly
            61            8            1            2            1            2           32            2           26            1
        pieced        these      several      various     agencies           if         find     credible         will           at
             1            2            3            2            1            9            3            1            4           12
         least  acknowledging       unique   interesting        mans           we        learn      certain        young      english
             3            1            1            1            1           12            1            1            6            5
      nobleman         whom        shall         call         john      clayton         lord     greystoke   commissioned         make
             1            1            2            1           13           29            8            8            1            1
    peculiarly     delicate  investigation   conditions         west        coast      african       colony        whose       simple
             1            2            3            1            2            1            1            2            1            1
        native   inhabitants      another     european        power        known   recruiting     soldiers          its        army
             2            1            2            2            2            1            1            1            3            3
          used      solely      forcible   collection       rubber        ivory      savage       tribes        along        congo
             2            1            1            1            1            1            1            1            2            1
```



```
Group_13_Project_3.R
Source on Save
274  # Apply weighting to the document-feature matrix
275  weighted_dfm = quanteda::dfm_weight(document_feature_matrix)
276  weighted_dfm
277  str(weighted_dfm)
275:25  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Apply weighting to the document-feature matrix
> weighted_dfm = quanteda::dfm_weight(document_feature_matrix)
> weighted_dfm
Document-feature matrix of: 421 documents, 1,154 features (99.31% sparse) and 0 docvars.
       features
docs    chapter i out to sea had this story from one
  text1       1 1   0  0   0   0    0     0    0   0
  text2       0 0   0  0   0   0    0     0    0   0
  text3       0 0   1  1   1   0    0     0    0   0
  text4       0 0   0  0   0   0    0     0    0   0
  text5       0 0   0  0   0   0    0     0    0   0
  text6       0 1   0  3   0   2    1     1    1   1
[ reached max_ndoc ... 415 more documents, reached max_nfeat ... 1,144 more features ]
> str(weighted_dfm)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    421 obs. of  3 variables:
  .. ..$ docname_: chr [1:421] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 421 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:421] 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] 3 3 1
  .. .. ..$ 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] "Darwin" "arm64" "rithik"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 9
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ 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
  .. .. .. ..$ threshold: NULL
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
```

Calculate Term Frequency-Inverse Document Frequency (TF_IDF):

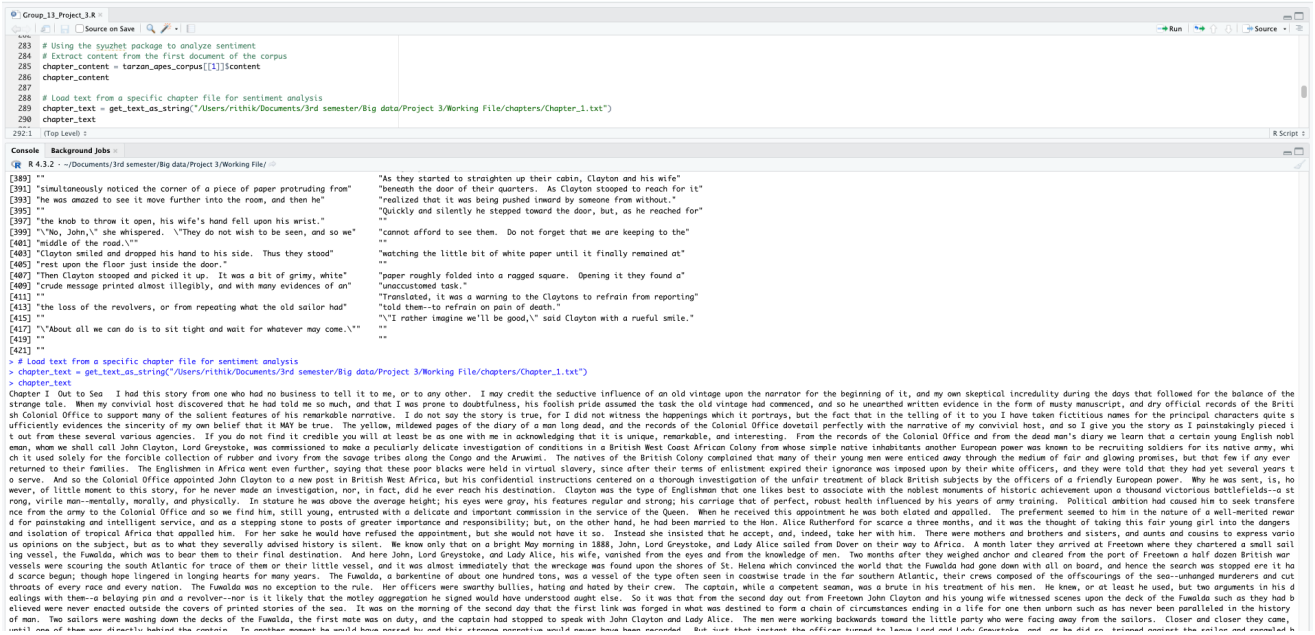


```
Group_13_Project_3.R
Source on Save | Run | Source
278
279  # Calculate Term Frequency-Inverse Document Frequency (TF-IDF)
280  tfidf_matrix = quanteda::dfm_tfidf(document_feature_matrix, scheme_tf = "count")
281  tfidf_matrix
282
281:13 (Top Level) R Script

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
  .. .. .. ..$ scheme   : chr "unary"
  .. .. .. ..$ base     : NULL
  .. .. .. ..$ c        : NULL
  .. .. .. ..$ smoothing: NULL
  .. .. .. ..$ threshold: NULL
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
  .. ..$ user  : list()
  ..@ i       : int [1:3357] 0 0 5 6 11 17 18 24 263 277 ...
  ..@ p       : int [1:1155] 0 1 31 38 119 122 153 168 172 199 ...
  ..@ Dim     : int [1:2] 421 1154
  ..@ Dimnames:List of 2
  .. ..$ docs    : chr [1:421] "text1" "text2" "text3" "text4" ...
  .. ..$ features: chr [1:1154] "chapter" "i" "out" "to" ...
  ..@ x       : num [1:3357] 1 1 1 1 1 2 1 2 1 3 ...
  ..@ factors : list()
> # Calculate Term Frequency-Inverse Document Frequency (TF-IDF)
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> tfidf_matrix = quanteda::dfm_tfidf(document_feature_matrix, scheme_tf = "count")
> tfidf_matrix
Document-feature matrix of: 421 documents, 1,154 features (99.31% sparse) and 0 docvars.
       features
docs     chapter        i      out       to      sea      had     this    story     from      one
  text1 2.624282 1.147161 0        0        0        0        0        0        0        0
  text2 0        0        0        0        0        0        0        0        0        0
  text3 0        0        1.779184 0.7157971 2.147161 0        0        0        0        0
  text4 0        0        0        0        0        0        0        0        0        0
  text5 0        0        0        0        0        0        0        0        0        0
  text6 0        1.147161 0        2.1473912 0        2.265841 1.448191 2.022222 1.192918 1.582889
[ reached max_ndoc ... 415 more documents, reached max_nfeat ... 1,144 more features ]
>
```

Using the Syuzhet Package to analyze sentiment:

The Syuzhet package in R is designed for sentiment analysis and narrative extraction from textual data, primarily focusing on the emotional arc of a story. This package implements several sentiment extraction techniques and relies on sentiment dictionaries to assign emotional values to words, allowing researchers and analysts to plot the trajectory of sentiment over the course of a narrative. By using the Syuzhet package, users can extract the underlying emotional progression of a story, identify pivotal points, and analyze how narrative elements influence reader or audience perceptions. This makes it particularly valuable in fields such as literary studies, content analysis, and any context where understanding the emotional flow of a text is important.

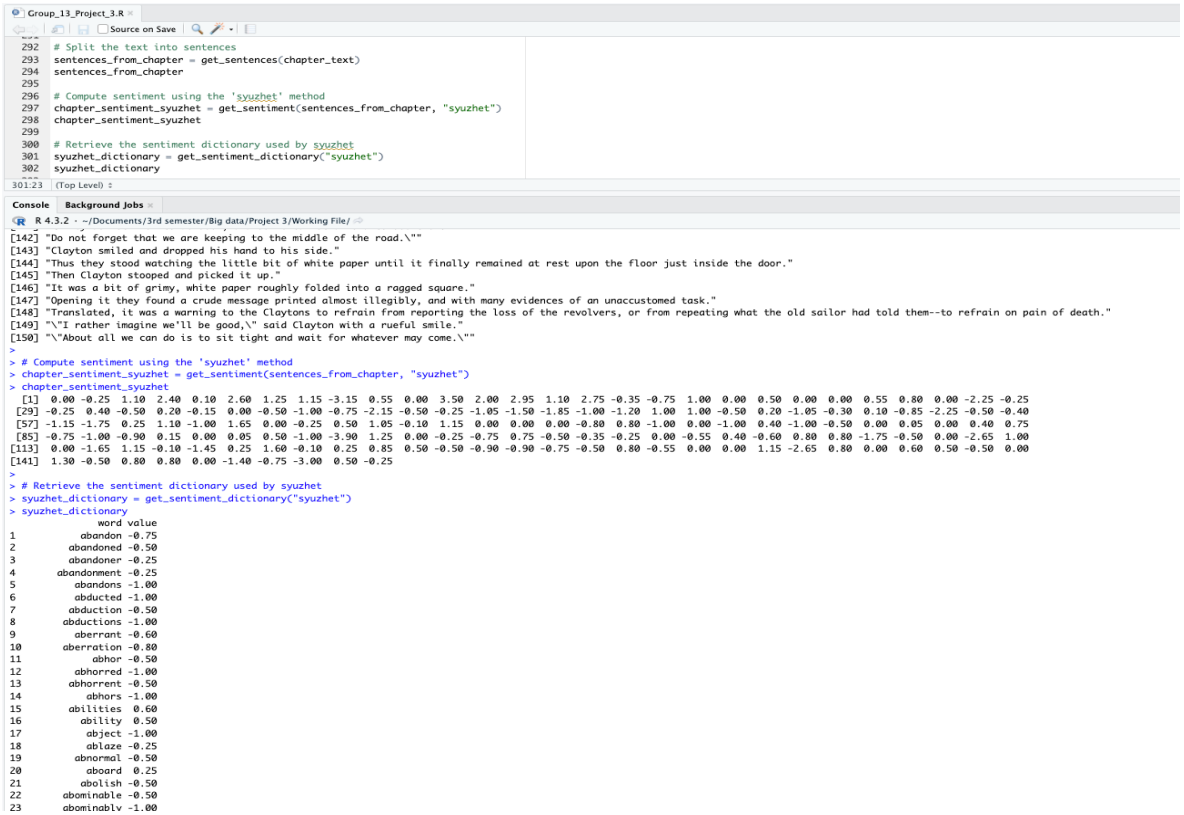
Extracting content from the first document of the corpus of all chapters and loading the text as string:



```
# Using the syuzhet package to analyze sentiment
# Extract content from the first document of the corpus
chapter_content = tarzan_apes_corpus[[1]]$content
chapter_content

# Load text from a specific chapter file for sentiment analysis
chapter_text = get_text_as_string("/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters/Chapter_1.txt")
chapter_text
```

```
[389] ""                                                                    "As they started to straighten up their cabin, Clayton and his wife"
[391] "simultaneously noticed the corner of a piece of paper protruding from"  "beneath the door of their quarters.  As Clayton stooped to reach for it"
[393] "he was amazed to see it move further into the room, and then he"     "realized that it was being pushed inward by someone from without."
[395] ""                                                                    "Quickly and silently he stepped toward the door, but, as he reached for"
[397] "the knob to throw it open, his wife's hand fell upon his wrist."     ""
[399] "\"No, John,\" she whispered.  \"They do not wish to be seen, and so we"  "cannot afford to see them.  Do not forget that we are keeping to the"
[401] "middle of the road.\""                                               ""
[403] "Clayton smiled and dropped his hand to his side.  Thus they stood"   "watching the little bit of white paper until it finally remained at"
[405] "rest upon the floor just inside the door."                           ""
[407] "Then Clayton stooped and picked it up.  It was a bit of grimy, white"  "paper roughly folded into a ragged square.  Opening it they found a"
[409] "crude message printed almost illegibly, and with many evidences of an"  "unaccustomed task."
[411] ""                                                                    "Translated, it was a warning to the Claytons to refrain from reporting"
[413] "the loss of the revolvers, or from repeating what the old sailor had"  "told them--to refrain on pain of death."
[415] ""                                                                    "\"I rather imagine we'll be good,\" said Clayton with a rueful smile."
[417] "\"About all we can do is to sit tight and wait for whatever may come.\""  ""
[419] ""                                                                    ""
[421] ""
> # Load text from a specific chapter file for sentiment analysis
> chapter_text = get_text_as_string("/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters/Chapter_1.txt")
> chapter_text
Chapter I  Out to Sea   I had this story from one who had no business to tell it to me, or to any other.  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 of the
strange tale.  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 unearthed written evidence in the form of musty manuscript, and dry official records of the Briti
sh Colonial Office to support many of the salient features of his remarkable narrative.  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 principal characters quite s
ufficiently evidences the sincerity of my own belief that it MAY be true.  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 the story as I painstakingly pieced i
t out from these several various agencies.  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.  From the records of the Colonial Office and from the dead man's diary we learn that a certain young English nobl
eman, whom we shall call John Clayton, Lord Greystoke, was commissioned to make a 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 for its native army, whi
ch it used solely for the forcible collection of rubber and ivory from the savage tribes along the Congo and the Aruwimi.  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 their families.  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 upon by their white officers, and they were told that they had yet several years t
o serve.  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 treatment of black British subjects by the officers of a friendly European power.  Why he was sent, is, ho
wever, of little moment to this story, for he never made an investigation, nor, in fact, did he ever reach his destination.  Clayton was the type of Englishman that one likes best to associate with the noblest monuments of historic achievement upon a thousand victorious battlefields--a st
rong, virile man--mentally, morally, and physically.  In stature he was above the average height; his eyes were gray, his features regular and strong; his carriage that of perfect, robust health influenced by his years of army training.  Political ambition had caused him to seek transfere
nce from the army to the Colonial Office and so we find him, still young, entrusted with a delicate and important commission in the service of the Queen.  When he received this appointment he was both elated and appalled.  The preferment seemed to him in the nature of a well-merited rewar
d for painstaking and intelligent service, and as a stepping stone to posts of greater importance and responsibility; but, on the other hand, he had been married to the Hon. Alice Rutherford for scarce a three months, and it was the thought of taking this fair young girl into the dangers
and isolation of tropical Africa that appalled him.  For her sake he would have refused the appointment, but she would not have it so.  Instead she insisted that he accept, and, indeed, take her with him.  There were mothers and brothers and sisters, and aunts and cousins to express vario
us opinions on the subject, but as to what they severally advised history 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 Africa.  A month later they arrived at Freetown where they chartered a small sail
ing vessel, the Fuwalda, which was to bear them to their final destination.  And here John, Lord Greystoke, and Lady Alice, his wife, vanished from the eyes and from the knowledge of men.  Two months after they weighed anchor and cleared from the port of Freetown a half dozen British war
vessels were scouring the south Atlantic for trace of them or their little vessel, and it was almost immediately that the wreckage was found upon the shores of St. Helena which convinced the world that the Fuwalda had gone down with all on board, and hence the search was stopped ere it ha
d scarce begun; though hope lingered in longing hearts for many years.  The Fuwalda, a barkentine of about one hundred tons, was a vessel of the type often seen in coastwise trade in the far southern Atlantic, their crews composed of the offscourings of the sea--unhanged murderers and cut
throats of every race and every nation.  The Fuwalda was no exception to the rule.  Her officers were swarthy bullies, hating and hated by their crew.  The captain, while a competent seaman, was a brute in his treatment of his men.  He knew, or at least he used, but two arguments in his d
ealings with them--a belaying pin and a revolver--nor is it likely that the motley aggregation he signed would have understood aught else.  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 such as they had b
elieved were never enacted outside the covers of printed stories of the sea.  It was on the morning of the second day that the first link was forged in what was destined to form a chain of circumstances ending in a life for 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 first mate was on duty, and the captain had stopped to speak with John Clayton and Lady Alice.  The men were working backwards toward the little party who were facing away from the sailors.  Closer and closer they came,
```

Split the text into sentences, compute sentiment using get_sentiment method and retrieve the sentiment dictionary:

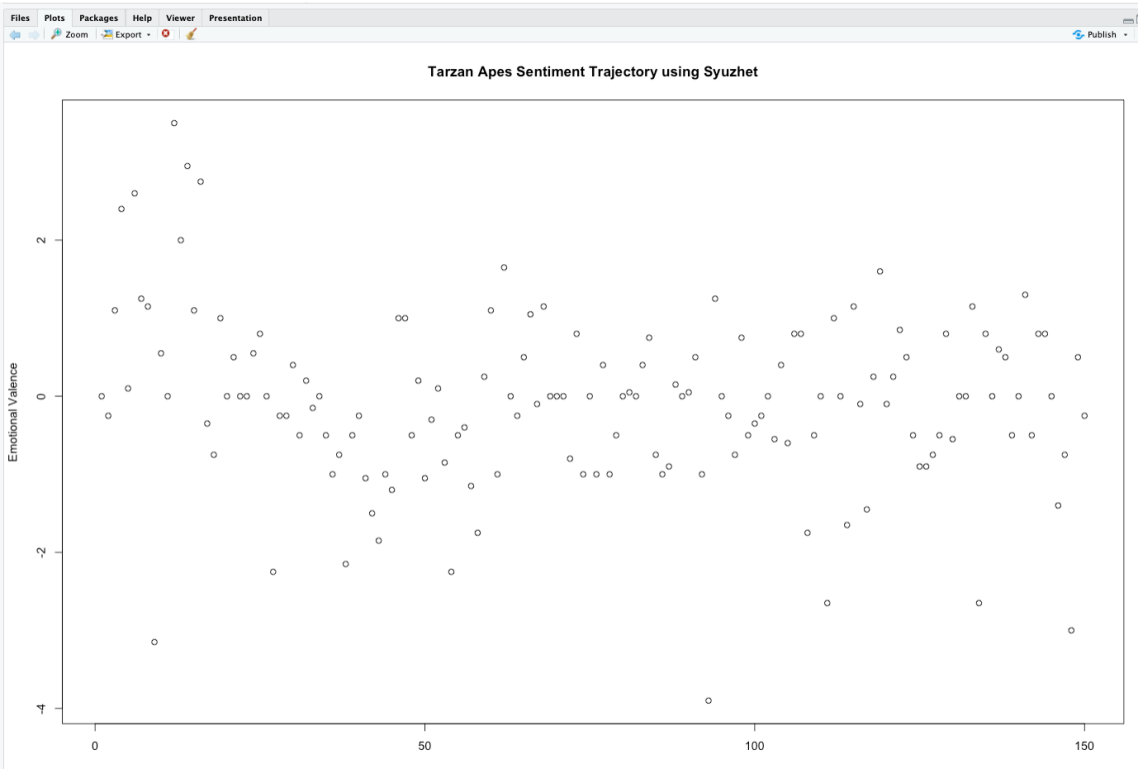


```
# Split the text into sentences
sentences_from_chapter = get_sentences(chapter_text)
sentences_from_chapter

# Compute sentiment using the 'syuzhet' method
chapter_sentiment_syuzhet = get_sentiment(sentences_from_chapter, "syuzhet")
chapter_sentiment_syuzhet

# Retrieve the sentiment dictionary used by syuzhet
syuzhet_dictionary = get_sentiment_dictionary("syuzhet")
syuzhet_dictionary
```

```
[142] "Do not forget that we are keeping to the middle of the road.\""
[143] "Clayton smiled and dropped his hand to his side."
[144] "Thus they stood watching the little bit of white paper until it finally remained at rest upon the floor just inside the door."
[145] "Then Clayton stooped and picked it up."
[146] "It was a bit of grimy, white paper roughly folded into a ragged square."
[147] "Opening it they found a crude message printed almost illegibly, and with many evidences of an unaccustomed task."
[148] "Translated, it was a warning to the Claytons to refrain from reporting the loss of the revolvers, or from repeating what the old sailor had told them--to refrain on pain of death."
[149] "\"I rather imagine we'll be good,\" said Clayton with a rueful smile."
[150] "\"About all we can do is to sit tight and wait for whatever may come.\""
>
> # Compute sentiment using the 'syuzhet' method
> chapter_sentiment_syuzhet = get_sentiment(sentences_from_chapter, "syuzhet")
> chapter_sentiment_syuzhet
  [1]  0.00 -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
>
> # Retrieve the sentiment dictionary used by syuzhet
> syuzhet_dictionary = get_sentiment_dictionary("syuzhet")
> syuzhet_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
```

Calculate the sentiment sum, mean and summarize overall sentiment:

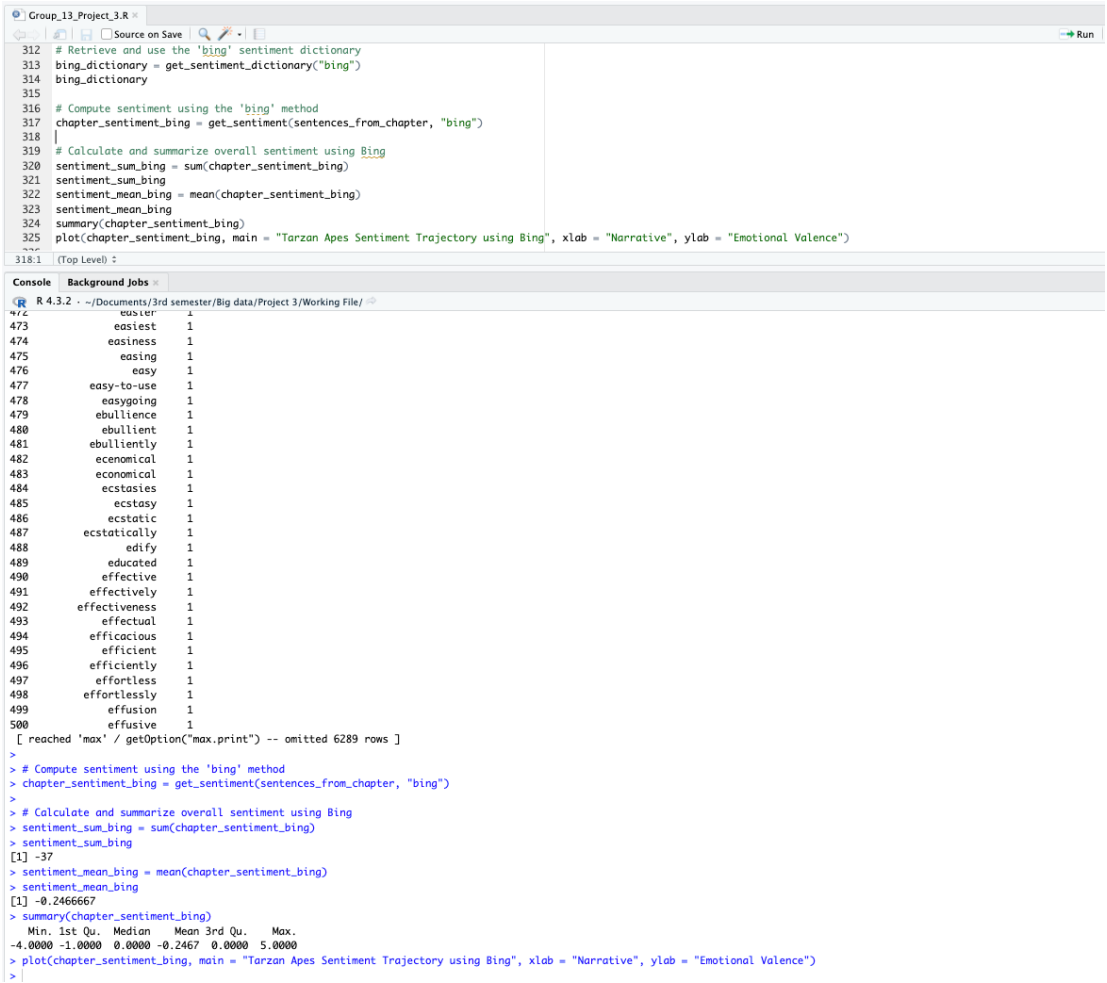
```
Group_13_Project_3.R
Source on Save
303
304 # Calculate and summarize overall sentiment using syuzhet
305 sentiment_sum_syuzhet = sum(chapter_sentiment_syuzhet)
306 sentiment_sum_syuzhet
307 sentiment_mean_syuzhet = mean(chapter_sentiment_syuzhet)
308 sentiment_mean_syuzhet
309 summary(chapter_sentiment_syuzhet)
310 plot(chapter_sentiment_syuzhet, main = "Tarzan Apes Sentiment Trajectory using Syuzhet", xlab = "Narrative", ylab = "Emotional Valence")
311
307:56 (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Calculate and summarize overall sentiment using syuzhet
> sentiment_sum_syuzhet = sum(chapter_sentiment_syuzhet)
> sentiment_sum_syuzhet
[1] -13.35
> sentiment_mean_syuzhet = mean(chapter_sentiment_syuzhet)
> sentiment_mean_syuzhet
[1] -0.089
> summary(chapter_sentiment_syuzhet)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.9000 -0.7125  0.0000 -0.0890  0.5000  3.5000
> plot(chapter_sentiment_syuzhet, main = "Tarzan Apes Sentiment Trajectory using Syuzhet", xlab = "Narrative", ylab = "Emotional Valence")
>
>
```

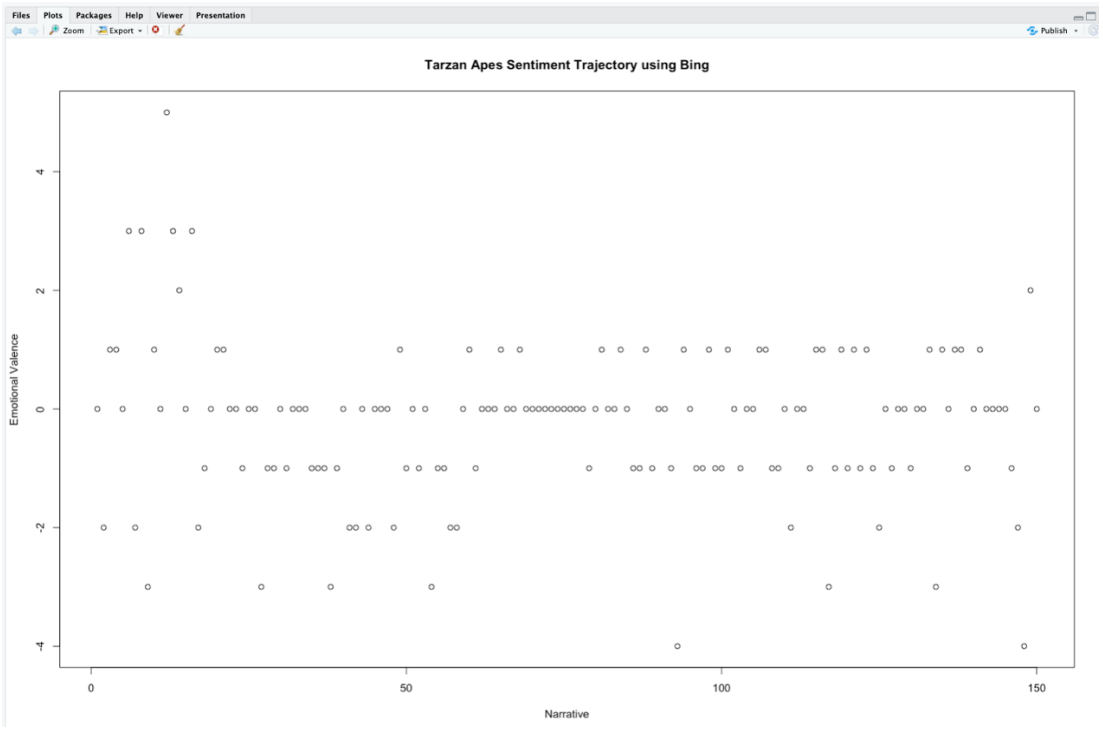


Bing Dictionary:

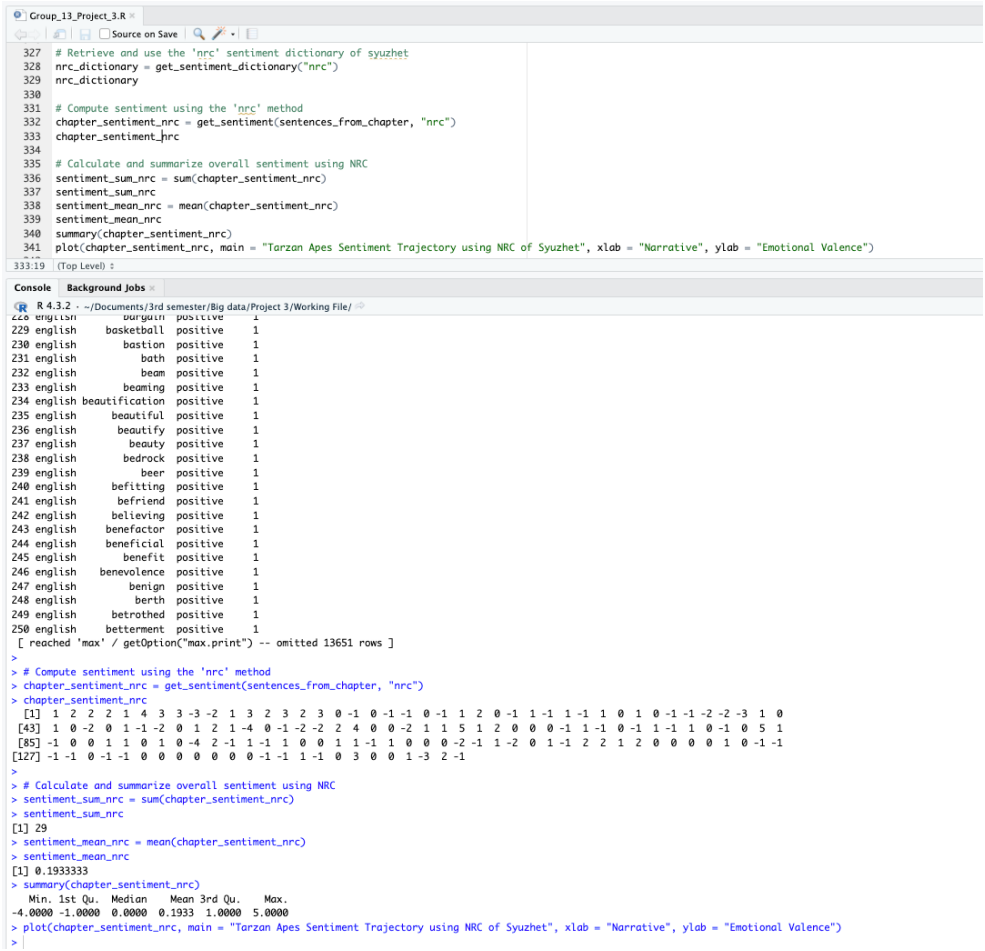
The Bing dictionary in sentiment analysis is a predefined lexicon used to classify words as either positive or negative. It is commonly utilized in tools like the Syuzhet package in R for straightforward sentiment scoring of text, helping to quickly evaluate the overall emotional tone conveyed by the words used. A plot is generated using the plot () function, which shows the sentiment trajectory of the text, with the x-axis representing the narrative and the y-axis representing the emotional valence.

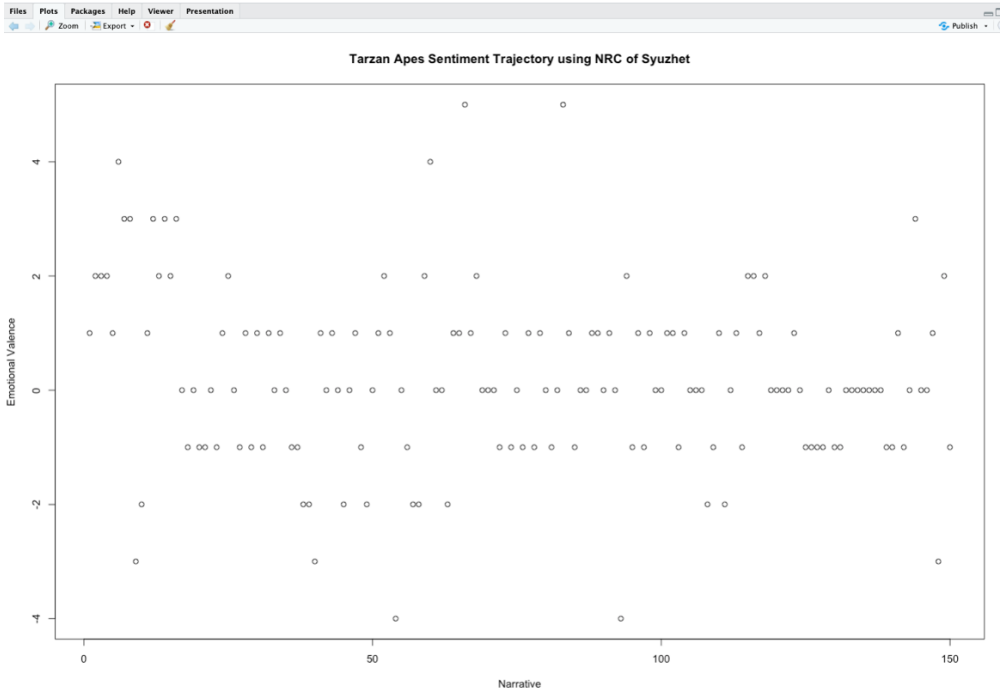
Retrieving bing dictionary, computing sentiment using bing method, and calculating the overall sentiment using bing:





Retrieving nrc dictionary, computing sentiment using nrc method, and calculating the overall sentiment using nrc:

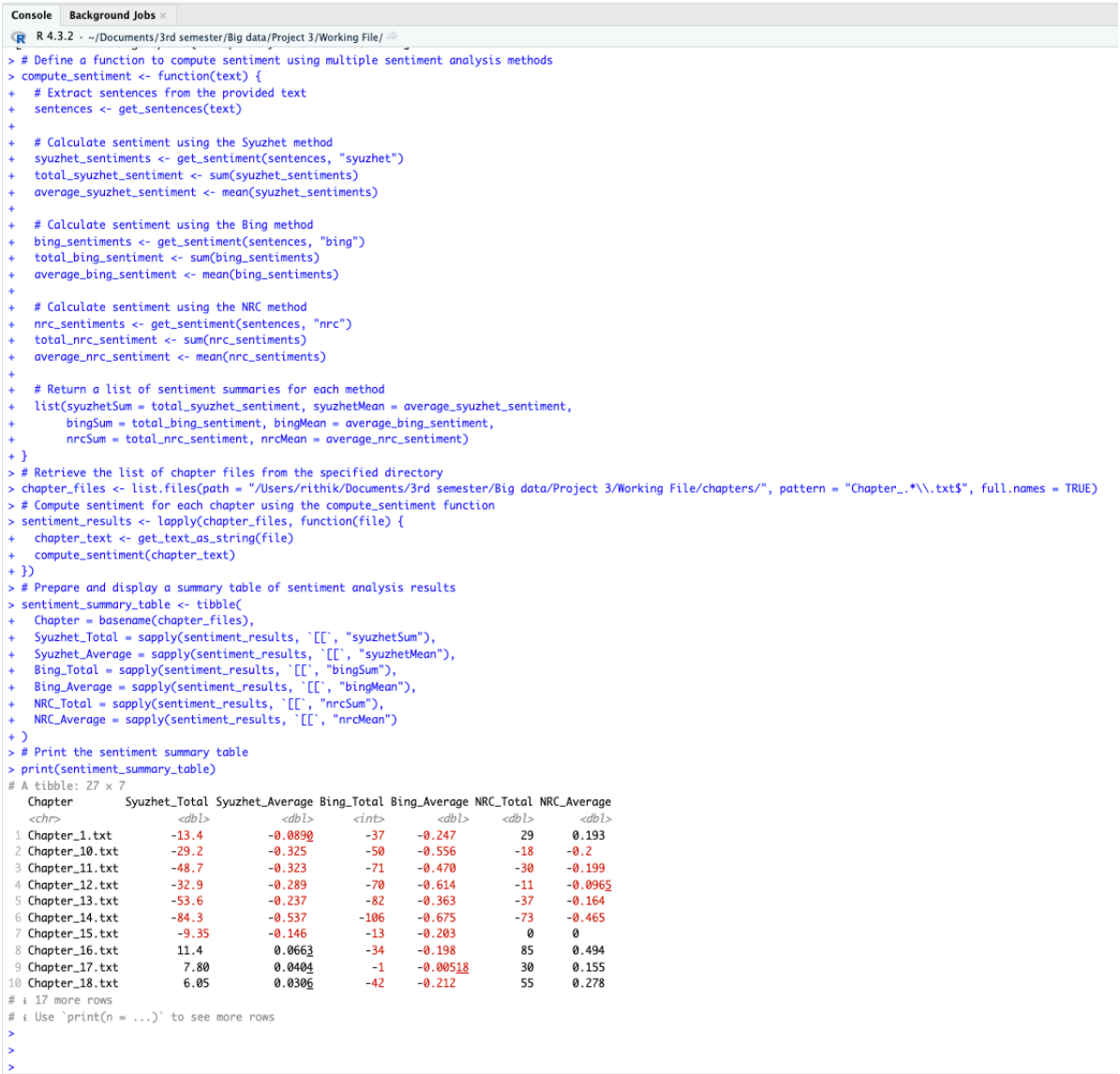




Computing NRC sentiment using tidytext for the chapter sentences:

```
Group_13_Project_3.R
Source on Save
343 # Compute NRC sentiment using tidytext for provided sentences
344 tarzan_apes_nrc_sentiment <- get_nrc_sentiment(sentences_from_chapter)
345 tarzan_apes_nrc_sentiment
346
347:1 (Top Level)
Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Compute NRC sentiment using tidytext for provided sentences
> tarzan_apes_nrc_sentiment <- get_nrc_sentiment(sentences_from_chapter)
> tarzan_apes_nrc_sentiment
    anger anticipation disgust fear joy sadness surprise trust negative positive
1       0            0       0    0   0       0        0     0        0        1
2       0            1       0    0   0       0        0     1        2        4
3       0            0       1    0   2       0        1     2        2        4
4       0            0       0    0   0       0        0     3        1        2
5       0            1       0    0   1       0        0     1        0        1
6       0            0       0    0   1       0        2     2        0        4
7       1            2       2    1   2       0        1     3        3        6
8       0            1       0    0   1       0        1     0        0        3
9       1            1       1    1   1       1        0     2        4        1
10      1            3       2    0   1       2        0     3        3        1
11      0            2       0    1   1       1        1     1        0        1
12      0            1       0    0   2       0        0     2        0        3
13      0            1       1    0   1       1        0     1        0        2
14      0            3       0    0   2       0        1     3        0        3
15      0            0       0    0   1       0        0     0        0        2
16      1            2       0    0   1       0        1     2        1        4
17      0            2       0    1   1       2        1     0        2        2
18      0            0       0    0   0       1        0     0        1        0
19      0            0       0    0   0       0        0     0        0        0
20      0            0       0    0   0       0        0     1        1        0
21      0            0       2    0   0       0        0     1        2        1
22      1            1       0    2   1       1        1     0        1        1
23      0            0       2    1   0       1        1     1        3        2
24      0            5       0    2   2       2        1     3        3        4
25      0            0       0    0   0       0        0     2        0        2
26      0            0       0    1   0       0        0     1        0        0
27      1            0       0    0   0       0        0     1        1        0
28      1            0       0    1   0       1        0     1        1        2
29      2            0       0    1   0       1        0     1        1        0
30      0            1       1    0   1       0        1     1        1        2
31      0            1       0    0   0       0        0     0        1        0
32      0            0       1    0   0       0        0     1        1        2
33      0            0       1    0   0       0        0     0        1        1
34      0            0       0    0   0       0        0     0        0        1
35      0            0       0    0   0       0        0     0        0        0
36      0            0       2    0   0       1        1     2        3        2
37      0            0       0    0   0       0        0     0        1        0
38      2            1       2    2   0       3        0     0        3        1
39      1            0       0    1   0       0        0     0        2        0
40      2            0       1    2   0       1        0     0        3        0
41      3            0       2    2   1       0        0     2        2        3
42      2            2       1    2   1       0        1     1        2        2
43      1            4       2    2   1       1        1     2        3        4
44      2            0       0    1   1       0        0     2        2        2
45      1            0       1    0   0       1        0     0        3        1
46      2            0       0    2   1       1        0     2        1        1
47      0            0       0    0   0       1        0     1        0        1
48      1            1       2    1   1       2        1     3        4        3
49      2            0       1    0   0       0        0     0        2        0
50      0            0       0    1   0       1        0     0        1        1
51      0            0       0    1   0       0        0     1        1        2
52      0            1       0    0   0       1        0     0        1        3
```

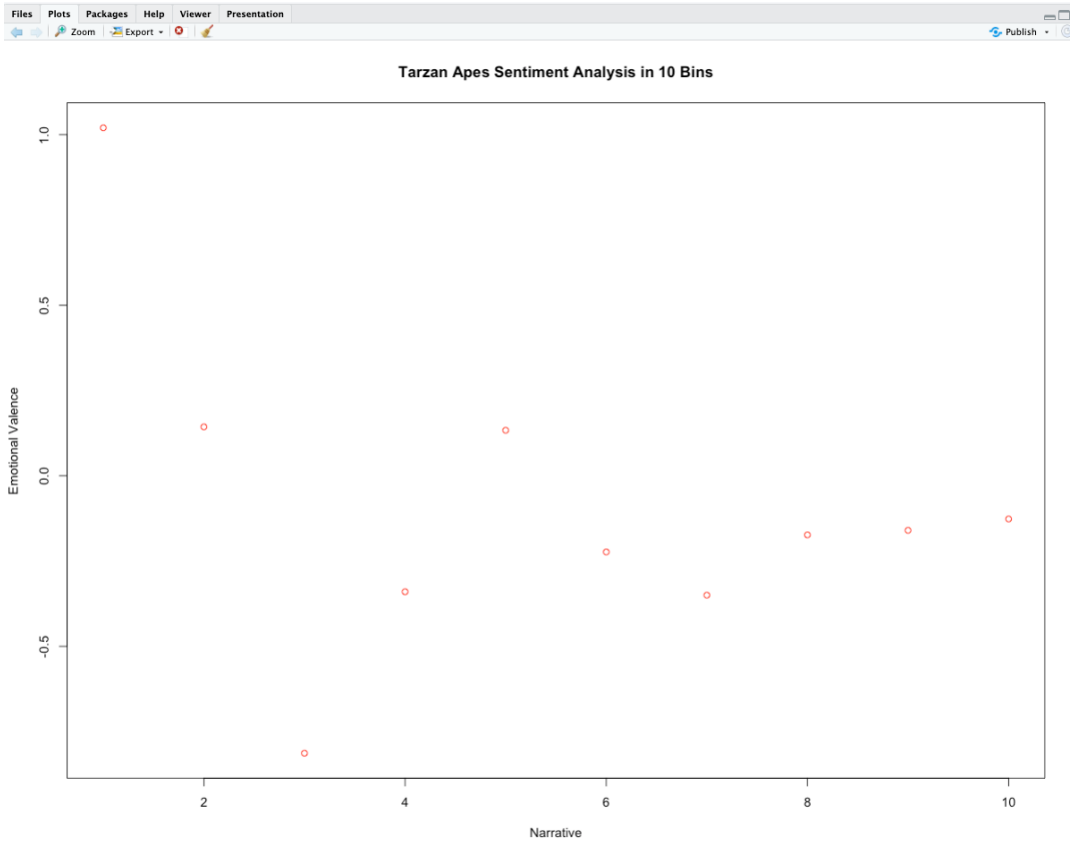
We now define a function compute_sentiment that analyzes the sentiment of given text using three methods: Syuzhet, Bing, and NRC, by extracting sentences and computing their sentiment scores to derive total and average sentiment values for each method. It then iterates over text files in a specified directory, each representing a chapter, using lapply to apply the sentiment analysis function to each chapter's text. The results are aggregated into a summary table using tibble, which organizes sentiment data by chapter, showing both total and average scores for each sentiment method. Finally, the summary table is displayed, providing a comprehensive view of sentiment trends across the chapters.



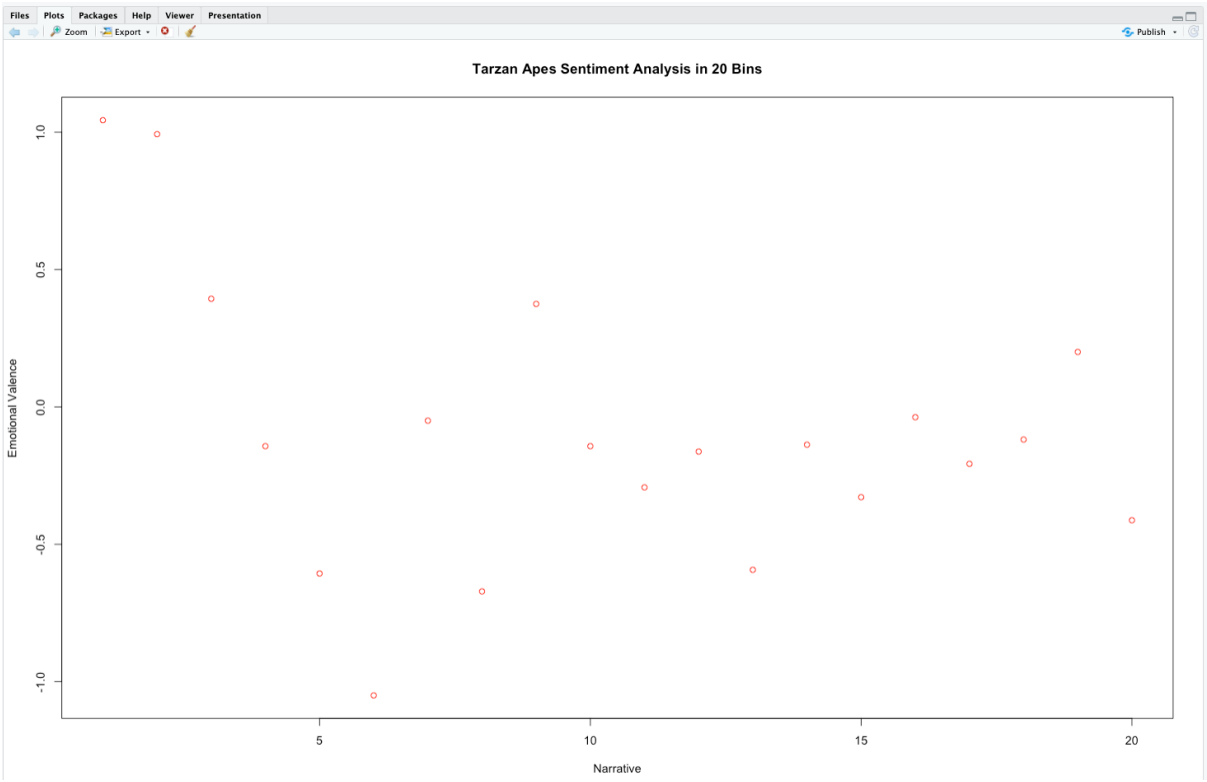
```
Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Define a function to compute sentiment using multiple sentiment analysis methods
> compute_sentiment <- function(text) {
+   # Extract sentences from the provided text
+   sentences <- get_sentences(text)
+
+   # Calculate sentiment using the Syuzhet method
+   syuzhet_sentiments <- get_sentiment(sentences, "syuzhet")
+   total_syuzhet_sentiment <- sum(syuzhet_sentiments)
+   average_syuzhet_sentiment <- mean(syuzhet_sentiments)
+
+   # Calculate sentiment using the Bing method
+   bing_sentiments <- get_sentiment(sentences, "bing")
+   total_bing_sentiment <- sum(bing_sentiments)
+   average_bing_sentiment <- mean(bing_sentiments)
+
+   # Calculate sentiment using the NRC method
+   nrc_sentiments <- get_sentiment(sentences, "nrc")
+   total_nrc_sentiment <- sum(nrc_sentiments)
+   average_nrc_sentiment <- mean(nrc_sentiments)
+
+   # Return a list of sentiment summaries for each method
+   list(syuzhetSum = total_syuzhet_sentiment, syuzhetMean = average_syuzhet_sentiment,
+        bingSum = total_bing_sentiment, bingMean = average_bing_sentiment,
+        nrcSum = total_nrc_sentiment, nrcMean = average_nrc_sentiment)
+ }
> # Retrieve the list of chapter files from the specified directory
> chapter_files <- list.files(path = "/Users/rithik/Documents/3rd semester/Big data/Project 3/Working File/chapters/", pattern = "Chapter_.*\\.txt$", full.names = TRUE)
> # Compute sentiment for each chapter using the compute_sentiment function
> sentiment_results <- lapply(chapter_files, function(file) {
+   chapter_text <- get_text_as_string(file)
+   compute_sentiment(chapter_text)
+ })
> # Prepare and display a summary table of sentiment analysis results
> sentiment_summary_table <- tibble(
+   Chapter = basename(chapter_files),
+   Syuzhet_Total = sapply(sentiment_results, `[[`, "syuzhetSum"),
+   Syuzhet_Average = sapply(sentiment_results, `[[`, "syuzhetMean"),
+   Bing_Total = sapply(sentiment_results, `[[`, "bingSum"),
+   Bing_Average = sapply(sentiment_results, `[[`, "bingMean"),
+   NRC_Total = sapply(sentiment_results, `[[`, "nrcSum"),
+   NRC_Average = sapply(sentiment_results, `[[`, "nrcMean")
+ )
> # Print the sentiment summary table
> print(sentiment_summary_table)
# A tibble: 27 × 7
   Chapter        Syuzhet_Total Syuzhet_Average Bing_Total Bing_Average NRC_Total NRC_Average
   <chr>                  <dbl>           <dbl>      <int>        <dbl>     <dbl>       <dbl>
 1 Chapter_1.txt          -13.4         -0.0890        -37       -0.247        29       0.193
 2 Chapter_10.txt         -29.2         -0.325         -50       -0.556       -18      -0.2
 3 Chapter_11.txt         -48.7         -0.323         -71       -0.470       -30      -0.199
 4 Chapter_12.txt         -32.9         -0.289         -70       -0.614       -11      -0.0965
 5 Chapter_13.txt         -53.6         -0.237         -82       -0.363       -37      -0.164
 6 Chapter_14.txt         -84.3         -0.537        -106       -0.675       -73      -0.465
 7 Chapter_15.txt          -9.35        -0.146         -13       -0.203         0       0
 8 Chapter_16.txt          11.4          0.0663        -34       -0.198        85       0.494
 9 Chapter_17.txt           7.80         0.0404         -1       -0.00518      30       0.155
10 Chapter_18.txt           6.05         0.0306        -42       -0.212        55       0.278
# ℹ 17 more rows
# ℹ Use `print(n = ...)` to see more rows
>
>
>
```

We now perform sentiment analysis on the "Tarzan of the Apes" text, dividing the results into two separate histograms with different levels of granularity. First, we compute the sentiment distribution across 10 bins, generating a summary and plotting it to visually represent the sentiment variations in a mid-level detail. And then repeat the process for 20 bins, offering a finer resolution of the sentiment distribution. Both plots are colored red and labelled with axes indicating narrative progression and emotional valence. This dual-level analysis helps in understanding how sentiment changes more subtly or broadly throughout the text.

```
Group_13_Project_3.R
Source on Save                                                                    Run
396 # Analyze sentiment distribution in 10 bins
397 tarzan_apes_sentiment_10_bins <- get_percentage_values(chapter_sentiment_syuzhet, bins=10)
398 tarzan_apes_sentiment_10_bins
399 structure(tarzan_apes_sentiment_10_bins)
400 plot(tarzan_apes_sentiment_10_bins, main = "Tarzan Apes Sentiment Analysis in 10 Bins", xlab = "Narrative", ylab = "Emotional Valence", col='red')
398:30  (Top Level)

Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Analyze sentiment distribution in 10 bins
> tarzan_apes_sentiment_10_bins <- get_percentage_values(chapter_sentiment_syuzhet, bins=10)
> tarzan_apes_sentiment_10_bins
        1         2         3         4         5         6         7         8         9        10
 1.0200000  0.1433333 -0.8133333 -0.3400000  0.1333333 -0.2233333 -0.3500000 -0.1733333 -0.1600000 -0.1266667
> structure(tarzan_apes_sentiment_10_bins)
        1         2         3         4         5         6         7         8         9        10
 1.0200000  0.1433333 -0.8133333 -0.3400000  0.1333333 -0.2233333 -0.3500000 -0.1733333 -0.1600000 -0.1266667
> plot(tarzan_apes_sentiment_10_bins, main = "Tarzan Apes Sentiment Analysis in 10 Bins", xlab = "Narrative", ylab = "Emotional Valence", col='red')
>
```



```
Group_13_Project_3.R
Source on Save | Run | Source
402 # Analyze sentiment distribution in 20 bins
403 tarzan_apes_sentiment_20_bins <- get_percentage_values(chapter_sentiment_syuzhet, bins=20)
404 tarzan_apes_sentiment_20_bins
405 structure(tarzan_apes_sentiment_20_bins)
406 plot(tarzan_apes_sentiment_20_bins, main = "Tarzan Apes Sentiment Analysis in 20 Bins", xlab = "Narrative", ylab = "Emotional Valence", col='red')
408:19 (Top Level)

Console | Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Analyze sentiment distribution in 20 bins
> tarzan_apes_sentiment_20_bins <- get_percentage_values(chapter_sentiment_syuzhet, bins=20)
> tarzan_apes_sentiment_20_bins
         1          2          3          4          5          6          7          8          9         10         11         12         13         14         15
 1.0437500  0.9928571  0.3937500 -0.1428571 -0.6062500 -1.0500000 -0.0500000 -0.6714286  0.3750000 -0.1428571 -0.2928571 -0.1625000 -0.5928571 -0.1375000 -0.3285714
        16         17         18         19         20
-0.0375000 -0.2071429 -0.1187500  0.2000000 -0.4125000
> structure(tarzan_apes_sentiment_20_bins)
         1          2          3          4          5          6          7          8          9         10         11         12         13         14         15
 1.0437500  0.9928571  0.3937500 -0.1428571 -0.6062500 -1.0500000 -0.0500000 -0.6714286  0.3750000 -0.1428571 -0.2928571 -0.1625000 -0.5928571 -0.1375000 -0.3285714
        16         17         18         19         20
-0.0375000 -0.2071429 -0.1187500  0.2000000 -0.4125000
> plot(tarzan_apes_sentiment_20_bins, main = "Tarzan Apes Sentiment Analysis in 20 Bins", xlab = "Narrative", ylab = "Emotional Valence", col='red')
>
>
```



Topic modelling using two statistical methods:

Latent Dirichlet Allocation (LDA):

LDA Model: The LDA function from the topicmodels package is used to fit an LDA model to the document-term matrix (DTM) named dtm_no_stopwords.

This matrix excludes stop words, improving the quality of topics identified. The parameter k = 10 specifies that ten latent topics should be identified.

Topic-Word Distribution: The resulting topic-word distribution (beta matrix) is converted to a data frame and printed. This matrix shows the probabilities of each word belonging to a particular topic.

Document-Topic Distribution: The document-topic distribution (gamma matrix) is converted to a data frame and printed. This matrix shows the probabilities of each document belonging to each of the identified topics.

```
> # Topic Modeling: Latent Dirichlet Allocation (LDA)
> # Conduct LDA to identify latent topics within the corpus
> lda_tarzan_apes <- topicmodels::LDA(dtm_no_stopwords, k = 10)
> # Extract and print the topic-word distribution
> topic_word_distribution_lda <- as.data.frame(lda_tarzan_apes@beta)
> print(topic_word_distribution_lda)
     V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41 V42 V43 V44
     V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57 V58 V59 V60 V61 V62 V63 V64 V65 V66 V67 V68 V69 V70 V71 V72 V73 V74 V75 V76 V77 V78 V79 V80 V81 V82 V83 V84 V85 V86
     V87 V88 V89 V90 V91 V92 V93 V94 V95 V96 V97 V98 V99 V100 V101 V102 V103 V104 V105 V106 V107 V108 V109 V110 V111 V112 V113 V114 V115 V116 V117 V118 V119 V120 V121 V122
     V123 V124 V125 V126 V127 V128 V129 V130 V131 V132 V133 V134 V135 V136 V137 V138 V139 V140 V141 V142 V143 V144 V145 V146 V147 V148 V149 V150 V151 V152 V153 V154 V155 V156
     V157 V158 V159 V160 V161 V162 V163 V164 V165 V166 V167 V168 V169 V170 V171 V172 V173 V174 V175 V176 V177 V178 V179 V180 V181 V182 V183 V184 V185 V186 V187 V188 V189 V190
     V191 V192 V193 V194 V195 V196 V197 V198 V199 V200 V201 V202 V203 V204 V205 V206 V207 V208 V209 V210 V211 V212 V213 V214 V215 V216 V217 V218 V219 V220 V221 V222 V223 V224
     V225 V226 V227 V228 V229 V230 V231 V232 V233 V234 V235 V236 V237 V238 V239 V240 V241 V242 V243 V244 V245 V246 V247 V248 V249 V250 V251 V252 V253 V254 V255 V256 V257 V258
     V259 V260 V261 V262 V263 V264 V265 V266 V267 V268 V269 V270 V271 V272 V273 V274 V275 V276 V277 V278 V279 V280 V281 V282 V283 V284 V285 V286 V287 V288 V289 V290 V291 V292
     V293 V294 V295 V296 V297 V298 V299 V300 V301 V302 V303 V304 V305 V306 V307 V308 V309 V310 V311 V312 V313 V314 V315 V316 V317 V318 V319 V320 V321 V322 V323 V324 V325 V326
     V327 V328 V329 V330 V331 V332 V333 V334 V335 V336 V337 V338 V339 V340 V341 V342 V343 V344 V345 V346 V347 V348 V349 V350 V351 V352 V353 V354 V355 V356 V357 V358 V359 V360
     V361 V362 V363 V364 V365 V366 V367 V368 V369 V370 V371 V372 V373 V374 V375 V376 V377 V378 V379 V380 V381 V382 V383 V384 V385 V386 V387 V388 V389 V390 V391 V392 V393 V394
     V395 V396 V397 V398 V399 V400 V401 V402 V403 V404 V405 V406 V407 V408 V409 V410 V411 V412 V413 V414 V415 V416 V417 V418 V419 V420 V421 V422 V423 V424 V425 V426 V427 V428
     V429 V430 V431 V432 V433 V434 V435 V436 V437 V438 V439 V440 V441 V442 V443 V444 V445 V446 V447 V448 V449 V450 V451 V452 V453 V454 V455 V456 V457 V458 V459 V460 V461 V462
     V463 V464 V465 V466 V467 V468 V469 V470 V471 V472 V473 V474 V475 V476 V477 V478 V479 V480 V481 V482 V483 V484 V485 V486 V487 V488 V489 V490 V491 V492 V493 V494 V495 V496
     V497 V498 V499 V500 V501 V502 V503 V504 V505 V506 V507 V508 V509 V510 V511 V512 V513 V514 V515 V516 V517 V518 V519 V520 V521 V522 V523 V524 V525 V526 V527 V528 V529 V530
     V531 V532 V533 V534 V535 V536 V537 V538 V539 V540 V541 V542 V543 V544 V545 V546 V547 V548 V549 V550 V551 V552 V553 V554 V555 V556 V557 V558 V559 V560 V561 V562 V563 V564
     V565 V566 V567 V568 V569 V570 V571 V572 V573 V574 V575 V576 V577 V578 V579 V580 V581 V582 V583 V584 V585 V586 V587 V588 V589 V590 V591 V592 V593 V594 V595 V596 V597 V598
     V599 V600 V601 V602 V603 V604 V605 V606 V607 V608 V609 V610 V611 V612 V613 V614 V615 V616 V617 V618 V619 V620 V621 V622 V623 V624 V625 V626 V627 V628 V629 V630 V631 V632
     V633 V634 V635 V636 V637 V638 V639 V640 V641 V642 V643 V644 V645 V646 V647 V648 V649 V650 V651 V652 V653 V654 V655 V656 V657 V658 V659 V660 V661 V662 V663 V664 V665 V666
     V667 V668 V669 V670 V671 V672 V673 V674 V675 V676 V677 V678 V679 V680 V681 V682 V683 V684 V685 V686 V687 V688 V689 V690 V691 V692 V693 V694 V695 V696 V697 V698 V699 V700
     V701 V702 V703 V704 V705 V706 V707 V708 V709 V710 V711 V712 V713 V714 V715 V716 V717 V718 V719 V720 V721 V722 V723 V724 V725 V726 V727 V728 V729 V730 V731 V732 V733 V734
     V735 V736 V737 V738 V739 V740 V741 V742 V743 V744 V745 V746 V747 V748 V749 V750 V751 V752 V753 V754 V755 V756 V757 V758 V759 V760 V761 V762 V763 V764 V765 V766 V767 V768
     V769 V770 V771 V772 V773 V774 V775 V776 V777 V778 V779 V780 V781 V782 V783 V784 V785 V786 V787 V788 V789 V790 V791 V792 V793 V794 V795 V796 V797 V798 V799 V800 V801 V802
     V803 V804 V805 V806 V807 V808 V809 V810 V811 V812 V813 V814 V815 V816 V817 V818 V819 V820 V821 V822 V823 V824 V825 V826 V827 V828 V829 V830 V831 V832 V833 V834 V835 V836
     V837 V838 V839 V840 V841 V842 V843 V844 V845 V846 V847 V848 V849 V850 V851 V852 V853 V854 V855 V856 V857 V858 V859 V860 V861 V862 V863 V864 V865 V866 V867 V868 V869 V870
     V871 V872 V873 V874 V875 V876 V877 V878 V879 V880 V881 V882 V883 V884 V885 V886 V887 V888 V889 V890 V891 V892 V893 V894 V895 V896 V897 V898 V899 V900 V901 V902 V903 V904
     V905 V906 V907 V908 V909 V910 V911 V912 V913 V914 V915 V916 V917 V918 V919 V920 V921 V922 V923 V924 V925 V926 V927 V928 V929 V930 V931 V932 V933 V934 V935 V936 V937 V938
```

```
> # Extract and print the document-topic distribution
> document_topic_distribution_lda <- as.data.frame(lda_tarzan_apes@gamma)
> print(document_topic_distribution_lda)
             V1           V2           V3           V4           V5           V6           V7           V8           V9          V10
1  1.102068e-05 1.102068e-05 1.102068e-05 1.102068e-05 1.102068e-05 9.999008e-01 1.102068e-05 1.102068e-05 1.102068e-05 1.102068e-05
2  2.116003e-05 2.116003e-05 2.116003e-05 2.116003e-05 2.116003e-05 2.116003e-05 2.116003e-05 9.998096e-01 2.116003e-05 2.116003e-05
3  1.072573e-05 1.072573e-05 1.072573e-05 1.072573e-05 1.072573e-05 1.072573e-05 9.999035e-01 1.072573e-05 1.072573e-05 1.072573e-05
4  1.392002e-05 1.392002e-05 1.392002e-05 1.392002e-05 1.392002e-05 1.392002e-05 9.998747e-01 1.392002e-05 1.392002e-05 1.392002e-05
5  7.913832e-06 7.913832e-06 7.913832e-06 7.913832e-06 9.999288e-01 7.913832e-06 7.913832e-06 7.913832e-06 7.913832e-06 7.913832e-06
6  1.145636e-05 1.145636e-05 1.145636e-05 1.145636e-05 1.145636e-05 1.145636e-05 1.145636e-05 1.145636e-05 1.145636e-05 9.998969e-01
7  2.351059e-05 2.351059e-05 2.351059e-05 9.997884e-01 2.351059e-05 2.351059e-05 2.351059e-05 2.351059e-05 2.351059e-05 2.351059e-05
8  1.215105e-05 1.215105e-05 1.215105e-05 1.215105e-05 1.215105e-05 1.215105e-05 1.215105e-05 9.998906e-01 1.215105e-05 1.215105e-05
9  9.998998e-01 1.113139e-05 1.113139e-05 1.113139e-05 1.113139e-05 1.113139e-05 1.113139e-05 1.113139e-05 1.113139e-05 1.113139e-05
10 1.105301e-05 1.105301e-05 1.105301e-05 9.999005e-01 1.105301e-05 1.105301e-05 1.105301e-05 1.105301e-05 1.105301e-05 1.105301e-05
11 1.107250e-05 9.999003e-01 1.107250e-05 1.107250e-05 1.107250e-05 1.107250e-05 1.107250e-05 1.107250e-05 1.107250e-05 1.107250e-05
12 1.208866e-05 1.208866e-05 9.998912e-01 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05
13 9.897292e-06 9.897292e-06 9.897292e-06 9.999109e-01 9.897292e-06 9.897292e-06 9.897292e-06 9.897292e-06 9.897292e-06 9.897292e-06
14 1.786816e-05 9.998392e-01 1.786816e-05 1.786816e-05 1.786816e-05 1.786816e-05 1.786816e-05 1.786816e-05 1.786816e-05 1.786816e-05
15 9.946408e-01 1.208866e-05 1.208866e-05 5.262503e-03 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05 1.208866e-05
16 1.407605e-05 1.407605e-05 1.407605e-05 9.998733e-01 1.407605e-05 1.407605e-05 1.407605e-05 1.407605e-05 1.407605e-05 1.407605e-05
17 9.998568e-01 1.590660e-05 1.590660e-05 1.590660e-05 1.590660e-05 1.590660e-05 1.590660e-05 1.590660e-05 1.590660e-05 1.590660e-05
18 1.154766e-05 1.154766e-05 1.154766e-05 1.154766e-05 1.154766e-05 1.154766e-05 1.154766e-05 1.154766e-05 9.998961e-01 1.154766e-05
19 1.119091e-05 1.119091e-05 1.119091e-05 1.119091e-05 1.119091e-05 1.119091e-05 1.119091e-05 1.119091e-05 9.998993e-01 1.119091e-05
20 1.016433e-05 1.016433e-05 1.016433e-05 1.016433e-05 9.999085e-01 1.016433e-05 1.016433e-05 1.016433e-05 1.016433e-05 1.016433e-05
21 1.637681e-05 1.637681e-05 1.637681e-05 1.637681e-05 1.637681e-05 9.998526e-01 1.637681e-05 1.637681e-05 1.637681e-05 1.637681e-05
22 9.809357e-01 1.512736e-05 1.512736e-05 1.512736e-05 1.512736e-05 1.512736e-05 1.512736e-05 1.512736e-05 1.512736e-05 1.894330e-02
23 1.448745e-05 1.448745e-05 1.448745e-05 1.448745e-05 1.448745e-05 1.448745e-05 1.448745e-05 1.448745e-05 1.448745e-05 9.998696e-01
24 1.696686e-05 1.696686e-05 1.696686e-05 1.696686e-05 1.696686e-05 1.696686e-05 1.696686e-05 1.696686e-05 1.696686e-05 9.998473e-01
25 9.624220e-06 9.624220e-06 9.624220e-06 9.624220e-06 9.624220e-06 9.624220e-06 9.999134e-01 9.624220e-06 9.624220e-06 9.624220e-06
26 1.898471e-05 1.898471e-05 1.898471e-05 1.898471e-05 1.898471e-05 9.998291e-01 1.898471e-05 1.898471e-05 1.898471e-05 1.898471e-05
27 1.023057e-05 1.023057e-05 1.023057e-05 1.023057e-05 1.023057e-05 1.023057e-05 9.999079e-01 1.023057e-05 1.023057e-05 1.023057e-05
>
```

Correlated Topic Model (CTM):

CTM Model: A CTM model is fitted to the same document-term matrix using the CTM function. The parameter k = 2 specifies that two correlated topics should be identified.

Topic-Word Distribution: The topic-word distribution (beta matrix) is converted to a data frame and printed, revealing the probability of each word being associated with specific topics.

Document-Topic Distribution: The document-topic distribution (gamma matrix) is converted to a data frame and printed, showing the probability of each document being linked to a particular topic.

```
> # Topic Modeling: Correlated Topic Model (CTM)
> # Conduct CTM to identify correlated topics within the corpus
> ctm_tarzan_apes <- CTM(dtm_no_stopwords, k = 2)
> # Extract and print the topic-word distribution
> topic_word_distribution_ctm <- as.data.frame(ctm_tarzan_apes@beta)
> print(topic_word_distribution_ctm)
     V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41 V42 V43 V44
     V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57 V58 V59 V60 V61 V62 V63 V64 V65 V66 V67 V68 V69 V70 V71 V72 V73 V74 V75 V76 V77 V78 V79 V80 V81 V82 V83 V84 V85 V86
     V87 V88 V89 V90 V91 V92 V93 V94 V95 V96 V97 V98 V99 V100 V101 V102 V103 V104 V105 V106 V107 V108 V109 V110 V111 V112 V113 V114 V115 V116 V117 V118 V119 V120 V121 V122
     V123 V124 V125 V126 V127 V128 V129 V130 V131 V132 V133 V134 V135 V136 V137 V138 V139 V140 V141 V142 V143 V144 V145 V146 V147 V148 V149 V150 V151 V152 V153 V154 V155 V156
     V157 V158 V159 V160 V161 V162 V163 V164 V165 V166 V167 V168 V169 V170 V171 V172 V173 V174 V175 V176 V177 V178 V179 V180 V181 V182 V183 V184 V185 V186 V187 V188 V189 V190
     V191 V192 V193 V194 V195 V196 V197 V198 V199 V200 V201 V202 V203 V204 V205 V206 V207 V208 V209 V210 V211 V212 V213 V214 V215 V216 V217 V218 V219 V220 V221 V222 V223 V224
     V225 V226 V227 V228 V229 V230 V231 V232 V233 V234 V235 V236 V237 V238 V239 V240 V241 V242 V243 V244 V245 V246 V247 V248 V249 V250 V251 V252 V253 V254 V255 V256 V257 V258
     V259 V260 V261 V262 V263 V264 V265 V266 V267 V268 V269 V270 V271 V272 V273 V274 V275 V276 V277 V278 V279 V280 V281 V282 V283 V284 V285 V286 V287 V288 V289 V290 V291 V292
     V293 V294 V295 V296 V297 V298 V299 V300 V301 V302 V303 V304 V305 V306 V307 V308 V309 V310 V311 V312 V313 V314 V315 V316 V317 V318 V319 V320 V321 V322 V323 V324 V325 V326
     V327 V328 V329 V330 V331 V332 V333 V334 V335 V336 V337 V338 V339 V340 V341 V342 V343 V344 V345 V346 V347 V348 V349 V350 V351 V352 V353 V354 V355 V356 V357 V358 V359 V360
     V361 V362 V363 V364 V365 V366 V367 V368 V369 V370 V371 V372 V373 V374 V375 V376 V377 V378 V379 V380 V381 V382 V383 V384 V385 V386 V387 V388 V389 V390 V391 V392 V393 V394
     V395 V396 V397 V398 V399 V400 V401 V402 V403 V404 V405 V406 V407 V408 V409 V410 V411 V412 V413 V414 V415 V416 V417 V418 V419 V420 V421 V422 V423 V424 V425 V426 V427 V428
     V429 V430 V431 V432 V433 V434 V435 V436 V437 V438 V439 V440 V441 V442 V443 V444 V445 V446 V447 V448 V449 V450 V451 V452 V453 V454 V455 V456 V457 V458 V459 V460 V461 V462
     V463 V464 V465 V466 V467 V468 V469 V470 V471 V472 V473 V474 V475 V476 V477 V478 V479 V480 V481 V482 V483 V484 V485 V486 V487 V488 V489 V490 V491 V492 V493 V494 V495 V496
     V497 V498 V499 V500 V501 V502 V503 V504 V505 V506 V507 V508 V509 V510 V511 V512 V513 V514 V515 V516 V517 V518 V519 V520 V521 V522 V523 V524 V525 V526 V527 V528 V529 V530
     V531 V532 V533 V534 V535 V536 V537 V538 V539 V540 V541 V542 V543 V544 V545 V546 V547 V548 V549 V550 V551 V552 V553 V554 V555 V556 V557 V558 V559 V560 V561 V562 V563 V564
     V565 V566 V567 V568 V569 V570 V571 V572 V573 V574 V575 V576 V577 V578 V579 V580 V581 V582 V583 V584 V585 V586 V587 V588 V589 V590 V591 V592 V593 V594 V595 V596 V597 V598
     V599 V600 V601 V602 V603 V604 V605 V606 V607 V608 V609 V610 V611 V612 V613 V614 V615 V616 V617 V618 V619 V620 V621 V622 V623 V624 V625 V626 V627 V628 V629 V630 V631 V632
     V633 V634 V635 V636 V637 V638 V639 V640 V641 V642 V643 V644 V645 V646 V647 V648 V649 V650 V651 V652 V653 V654 V655 V656 V657 V658 V659 V660 V661 V662 V663 V664 V665 V666
     V667 V668 V669 V670 V671 V672 V673 V674 V675 V676 V677 V678 V679 V680 V681 V682 V683 V684 V685 V686 V687 V688 V689 V690 V691 V692 V693 V694 V695 V696 V697 V698 V699 V700
     V701 V702 V703 V704 V705 V706 V707 V708 V709 V710 V711 V712 V713 V714 V715 V716 V717 V718 V719 V720 V721 V722 V723 V724 V725 V726 V727 V728 V729 V730 V731 V732 V733 V734
     V735 V736 V737 V738 V739 V740 V741 V742 V743 V744 V745 V746 V747 V748 V749 V750 V751 V752 V753 V754 V755 V756 V757 V758 V759 V760 V761 V762 V763 V764 V765 V766 V767 V768
     V769 V770 V771 V772 V773 V774 V775 V776 V777 V778 V779 V780 V781 V782 V783 V784 V785 V786 V787 V788 V789 V790 V791 V792 V793 V794 V795 V796 V797 V798 V799 V800 V801 V802
     V803 V804 V805 V806 V807 V808 V809 V810 V811 V812 V813 V814 V815 V816 V817 V818 V819 V820 V821 V822 V823 V824 V825 V826 V827 V828 V829 V830 V831 V832 V833 V834 V835 V836
     V837 V838 V839 V840 V841 V842 V843 V844 V845 V846 V847 V848 V849 V850 V851 V852 V853 V854 V855 V856 V857 V858 V859 V860 V861 V862 V863 V864 V865 V866 V867 V868 V869 V870
     V871 V872 V873 V874 V875 V876 V877 V878 V879 V880 V881 V882 V883 V884 V885 V886 V887 V888 V889 V890 V891 V892 V893 V894 V895 V896 V897 V898 V899 V900 V901 V902 V903 V904
```

```
> # Extract and print the document-topic distribution
> document_topic_distribution_ctm <- as.data.frame(ctm_tarzan_apes@gamma)
> print(document_topic_distribution_ctm)
             V1          V2
1  9.902018e-01 0.009798197
2  7.376646e-06 0.999992623
3  5.512601e-06 0.999994487
4  6.187591e-06 0.999993812
5  4.800476e-06 0.999995200
6  5.862511e-06 0.999994137
7  8.452258e-06 0.999991548
8  5.546850e-06 0.999994453
9  9.104017e-01 0.089598308
10 6.508217e-06 0.999993492
11 8.841272e-01 0.115872808
12 9.877120e-01 0.012288032
13 1.331065e-01 0.866893547
14 7.526370e-06 0.999992474
15 6.307900e-06 0.999993692
16 6.094495e-06 0.999993906
17 7.224091e-06 0.999992776
18 2.243581e-02 0.977564193
19 5.673311e-06 0.999994327
20 6.223580e-01 0.377641956
21 9.804808e-01 0.019519184
22 9.904181e-01 0.009581890
23 9.877778e-01 0.012222220
24 9.866276e-01 0.013372402
25 9.899078e-01 0.010092158
26 7.063104e-06 0.999992937
27 9.308732e-01 0.069126813
>
```

The goal of both analyses is to identify and understand the hidden themes or topics within the text corpus and explore their prevalence across documents. The LDA assumes topics are independent, whereas CTM allows for correlation between topics.

Word Cloud Visualization:

Palette Definition: A color palette called word_cloud_palette is defined using 11 colors from the "Spectral" scheme via the brewer.pal function.

Word Cloud Creation: The wordcloud function generates a standard word cloud visualization, using the word names and their corresponding frequencies (word_frequencies) as inputs. Each word is colored according to the defined palette, and its size is proportional to its frequency.

```
> # Create a word cloud to visualize word frequencies
> word_cloud_palette <- brewer.pal(11, "Spectral")
> tarzan_apes_word_cloud <- wordcloud(names(word_frequencies), word_frequencies, colors = word_cloud_palette)
>
```

Comparison and Commonality Clouds:

Comparison Colors: A color palette called comparison_colors is defined using 6 colors from the "Set1" palette via brewer.pal.

Matrix Creation: A term-document matrix (tdm_no_stopwords) is converted to a matrix for easy manipulation in subsequent visualizations.

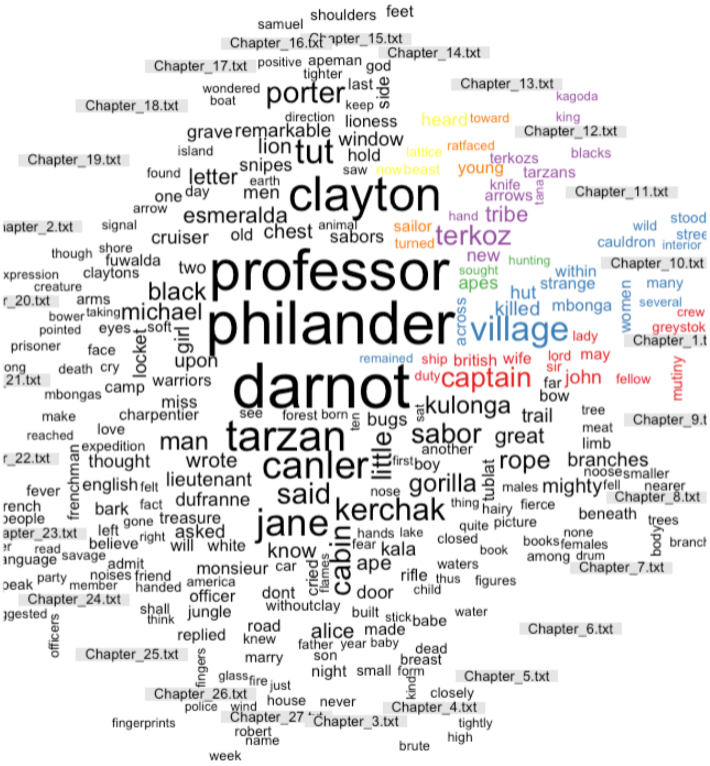
```
Console  Background Jobs
R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/
> # Define colors for document groups in comparison and commonality clouds
> comparison_colors <- brewer.pal(6, "Set1")
> # Create a matrix from the Term Document Matrix for visualization
> tarzan_apes_tdm_matrix <- as.matrix(tdm_no_stopwords)
> tarzan_apes_tdm_matrix
                    Docs
Terms                Chapter_1.txt Chapter_10.txt Chapter_11.txt Chapter_12.txt Chapter_13.txt Chapter_14.txt Chapter_15.txt
  abandon                        0             0             1             0             1             0             0
  abandoned                      0             0             0             1             0             0             0
  abandoning                     0             0             0             1             0             0             0
  abashed                        0             0             0             0             1             0             0
  abated                         0             0             0             0             0             0             0
  abatis                         0             0             0             0             0             0             0
  abduction                      0             0             0             0             0             0             0
  aberration                     0             0             0             0             0             0             0
  ability                        0             0             0             1             0             0             0
  able                           1             0             0             1             1             0             0
  aboard                         2             0             0             0             1             0             0
  abode                          0             0             0             0             0             0             0
  abound                         0             0             0             0             0             0             0
  aboutlord                      0             0             0             0             0             0             0
  abroad                         0             0             0             0             0             0             0
  abruptly                       0             0             0             0             0             0             0
  absence                        0             0             0             0             0             0             0
  absent                         0             0             0             0             0             0             0
  absentmindedness               0             0             0             0             1             0             0
  absinthe                       0             0             0             0             0             0             0
  absolute                       0             0             0             0             0             0             0
  absolutely                     0             0             0             1             1             0             0
  absorbed                       0             0             0             0             0             0             0
  absorbing                      0             0             0             0             0             0             0
  abstract                       0             0             0             0             0             0             0
  abstruse                       0             0             0             0             1             0             0
  abundance                      0             0             0             0             0             0             0
  abundant                       0             0             0             0             0             0             0
  abuses                         0             0             0             1             0             0             0
  abysmal                        0             0             0             0             0             1             0
  accede                         0             0             0             0             0             0             0
  accentuate                     0             0             0             0             0             0             0
  accentuated                    1             0             0             0             0             0             0
  accept                         1             0             0             1             0             0             0
  accepted                       0             0             0             0             0             0             0
  accession                      0             0             0             0             0             0             0
  accident                       0             0             0             1             0             0             0
                    Docs
Terms                Chapter_16.txt Chapter_17.txt Chapter_18.txt Chapter_19.txt Chapter_2.txt Chapter_20.txt Chapter_21.txt
  abandon                        0             0             0             0             0             0             0
  abandoned                      0             0             0             0             0             0             0
  abandoning                     0             0             0             0             0             0             0
  abashed                        0             0             0             0             0             0             0
  abated                         0             0             0             1             0             0             0
  abatis                         0             0             0             0             0             0             1
  abduction                      0             0             0             2             0             0             0
  aberration                     1             0             0             0             0             0             0
```

Comparison Cloud:

This visualizes word frequency differences across documents by highlighting the most frequent words unique to individual documents.

It uses the comparison.cloud function with the predefined color palette. Words are colored and scaled according to their importance across multiple documents.

```
> # Comparison Cloud: Visualize word frequency differences across documents
> comparison.cloud(tarzan_apes_tdm_matrix, colors = comparison_colors, scale = c(4, 0.5), random.order = FALSE, title.size = 0.9)
```

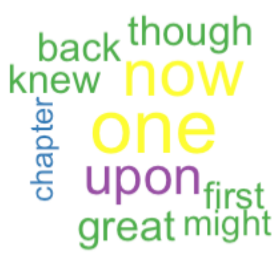


Commonality Cloud:

This visualization highlights common words shared among multiple documents.

The commonality.cloud function generates the plot, coloring and scaling the words to show shared or overlapping terms.

```
> # Commonality Cloud: Visualize common words across multiple documents
> commonality.cloud(tarzan_apes_tdm_matrix, colors = comparison_colors, scale = c(4, 0.5), random.order = FALSE)
>
```



Each of these visualizations provides different perspectives on the text corpus, helping uncover both unique and common themes across documents.

Applying various transformations to text data for better analysis and interpretation.

Term Frequency (TF) Weighting:

weightTf Function: The weightTf function is applied to the Document Term Matrix (DTM) without stop words (dtm_no_stopwords). It modifies the DTM to include raw term frequencies, representing how many times each word occurs in each document.
Inspect: The inspect function prints the weighted DTM's contents, allowing you to view and verify the term frequencies.

```
> # Apply Term Frequency weighting to the Document Term Matrix
> tarzan_apes_dtm_tf <- weightTf(dtm_no_stopwords)
> inspect(tarzan_apes_dtm_tf)
<<DocumentTermMatrix (documents: 27, terms: 7484)>>
Non-/sparse entries: 23598/178470
Sparsity           : 88%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs            clayton darnot great jungle little man now one tarzan upon
  Chapter_1.txt      29      0     4      0     12   8   7  11      0   11
  Chapter_11.txt      1      0    18      6     12   5   8  15     39   20
  Chapter_13.txt     10      0     9     20     17  31   6  19     33   31
  Chapter_17.txt     10      0    10      6      6   6   6  16     20   20
  Chapter_18.txt     24      0     5     15     11  12  11  21     14   11
  Chapter_19.txt     11      1    12      6      7   7   6  14     18   15
  Chapter_20.txt      0      0     8      7     12  11  11  13     42   33
  Chapter_27.txt     23      2     5      4     13  22  11   9      9    6
  Chapter_7.txt       0      0    15     14     28   2  11  22     22   22
  Chapter_9.txt       0      0    10     17     12   8   2  13     34   23
>
```

Term Frequency-Inverse Document Frequency (TF-IDF) Weighting:
weightTfIdf Function: This function applies TF-IDF weighting to the DTM without stop words. TF-IDF adjusts the term frequencies by accounting for how common a word is across all documents, giving higher importance to terms that appear less frequently across the whole corpus.

Inspect: The inspect function displays the TF-IDF-weighted DTM's contents, showing how the weights have been adjusted.

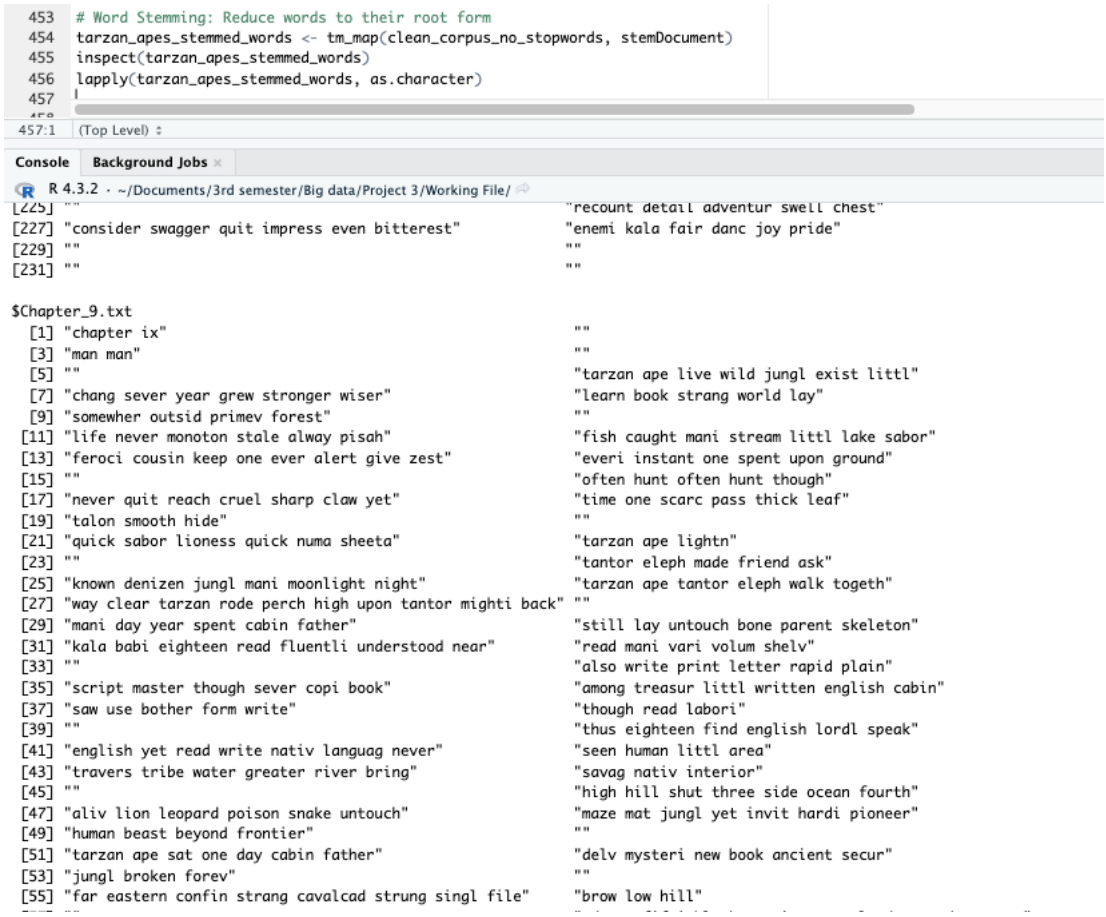
```
> # Apply TF-IDF weighting to the Document Term Matrix
> tarzan_apes_dtm_tfidf <- weightTfIdf(dtm_no_stopwords)
> inspect(tarzan_apes_dtm_tfidf)
<<DocumentTermMatrix (documents: 27, terms: 7484)>>
Non-/sparse entries: 23328/178740
Sparsity           : 88%
Maximal term length: 20
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
                Terms
Docs               canler     clayton      darnot        jane   philander      porter   professor        said       tarzan
  Chapter_1.txt  0.0000000 0.008602598 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.003969558 0.0000000000
  Chapter_12.txt 0.0000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.002507012 0.0073518097
  Chapter_16.txt 0.0000000 0.000327071 0.000000000 0.001222623 0.047148987 0.015282784 0.040953121 0.005470948 0.0014924229
  Chapter_17.txt 0.0000000 0.002996217 0.000000000 0.003920052 0.006775222 0.006160082 0.006125087 0.006014162 0.0027343445
  Chapter_2.txt  0.0000000 0.005857047 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.0000000000
  Chapter_24.txt 0.0000000 0.005566284 0.011857348 0.009603371 0.006051349 0.009603371 0.014223755 0.011459418 0.0003907526
  Chapter_26.txt 0.0000000 0.000602448 0.037537701 0.000000000 0.000000000 0.000000000 0.000000000 0.010077206 0.0072847617
  Chapter_27.txt 0.0709234 0.006292542 0.001894104 0.019942672 0.005413230 0.009204310 0.012584029 0.013729038 0.0011235456
  Chapter_7.txt  0.0000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.0026004916
  Chapter_8.txt  0.0000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.0048970126
                Terms
Docs               village
  Chapter_1.txt  0.0000000000
  Chapter_12.txt 0.0057448154
  Chapter_16.txt 0.0000000000
  Chapter_17.txt 0.0000000000
  Chapter_2.txt  0.0000000000
  Chapter_24.txt 0.0032824051
  Chapter_26.txt 0.0007697302
  Chapter_27.txt 0.0000000000
  Chapter_7.txt  0.0000000000
  Chapter_8.txt  0.0000000000
>
```

Word Stemming:

tm_map with stemDocument: The tm_map function is used to apply stemDocument to every document within clean_corpus_no_stopwords. Stemming reduces words to their root forms (e.g., "running" to "run"), enabling the grouping of different word forms.

Inspect: The inspect function displays the cleaned and stemmed documents, showing the new forms of words after stemming.

Convert to Characters: The lapply function converts the stemmed words into a character vector, providing easier access to the stemmed content for further processing or analysis.



```
453 # Word Stemming: Reduce words to their root form
454 tarzan_apes_stemmed_words <- tm_map(clean_corpus_no_stopwords, stemDocument)
455 inspect(tarzan_apes_stemmed_words)
456 lapply(tarzan_apes_stemmed_words, as.character)
457
457:1 (Top Level)

Console  Background Jobs

R 4.3.2 · ~/Documents/3rd semester/Big data/Project 3/Working File/

[225] ""                                                    "recount detail adventur swell chest"
[227] "consider swagger quit impress even bitterest"        "enemi kala fair danc joy pride"
[229] ""                                                    ""
[231] ""                                                    ""

$Chapter_9.txt
  [1] "chapter ix"                                          ""
  [3] "man man"                                             ""
  [5] ""                                                    "tarzan ape live wild jungl exist littl"
  [7] "chang sever year grew stronger wiser"                "learn book strang world lay"
  [9] "somewher outsid primev forest"                       ""
 [11] "life never monoton stale alway pisah"                "fish caught mani stream littl lake sabor"
 [13] "feroci cousin keep one ever alert give zest"         "everi instant one spent upon ground"
 [15] ""                                                    "often hunt often hunt though"
 [17] "never quit reach cruel sharp claw yet"               "time one scarc pass thick leaf"
 [19] "talon smooth hide"                                   ""
 [21] "quick sabor lioness quick numa sheeta"               "tarzan ape lightn"
 [23] ""                                                    "tantor eleph made friend ask"
 [25] "known denizen jungl mani moonlight night"            "tarzan ape tantor eleph walk togeth"
 [27] "way clear tarzan rode perch high upon tantor mighti back" ""
 [29] "mani day year spent cabin father"                    "still lay untouch bone parent skeleton"
 [31] "kala babi eighteen read fluentli understood near"    "read mani vari volum shelv"
 [33] ""                                                    "also write print letter rapid plain"
 [35] "script master though sever copi book"                "among treasur littl written english cabin"
 [37] "saw use bother form write"                           "though read labori"
 [39] ""                                                    "thus eighteen find english lordl speak"
 [41] "english yet read write nativ languag never"          "seen human littl area"
 [43] "travers tribe water greater river bring"             "savag nativ interior"
 [45] ""                                                    "high hill shut three side ocean fourth"
 [47] "aliv lion leopard poison snake untouch"              "maze mat jungl yet invit hardi pioneer"
 [49] "human beast beyond frontier"                         ""
 [51] "tarzan ape sat one day cabin father"                 "delv mysteri new book ancient secur"
 [53] "jungl broken forev"                                  ""
 [55] "far eastern confin strang cavalcad strung singl file" "brow low hill"
```

Overall, these transformations aim to standardize and enhance text data for more accurate and meaningful insights in subsequent text mining tasks.

Studying text data analysis using R from a book like "Tarzan of the Apes" can teach you a lot about the useful applications of text analytics. First, this project highlights the significance of data pre-processing, which involves cleaning the data to remove punctuation and stop words and normalizing the text using lemmatization or stemming to make it easier to analyze. We can represent textual data in a structured manner through tokenization and the creation of Document Term Matrices or Term Document Matrices. This provides a strong basis for additional exploratory analyses, like sentiment analysis and the identification of frequently occurring words. These preliminary measures emphasize the importance of careful data preparation, which is necessary to get accurate and significant results.

Important lessons about the interpretive potential of word clouds, comparison clouds, and sentiment arcs are also imparted by the project. Through their ability to assist users in perceiving unique word usage patterns, emotional trajectories, and common themes among documents, these tools offer instant insights into text data. Latent Dirichlet Allocation (LDA)-based topic modeling serves as another example of how statistical approaches can reveal hidden themes or topics in the text. Examining sentiment analysis techniques such as Syuzhet, Bing, or NRC helps readers understand how the author's stylistic and thematic decisions are influenced by the shifting emotional tones throughout the story. All things considered, this project highlights the enormous potential of text analytics in giving otherwise unstructured data clarity and structure, highlighting its relevance to a variety of fields, from business intelligence to literature.

THANK YOU