CS3907/CS6444 Big Data and Analytics
Fall 2022
Class Project #3
By

Mohammed Abdul Irfan(G33655938)
Sumanth Nandeti(G40560437)
Pramukh Nagol(G38478227)

In this project, we will use R to process and analyze the text of the novel A Princess of Mars by Edgar Rice Burroughs. We will first create a VCorpus, a text document collection. We will then use the VCorpus to find the 10 longest words and 10 longest sentences in each chapter of the novel. We will also remove the punctuation from the text and create a data table. Finally, we will use the table to show the items we found.

This project will teach us how to process large text documents and extract valuable insights using R programming language. The goal is to demonstrate the usefulness of text analytics in understanding complex text data.

Data Science is an empirical science because it relies on data to conclude. In this project, I used data to gain insights into the novel A Princess of Mars. I removed the stop words and punctuation from the text, and I used the following packages:

We will be using the following packages:

- tm: tm is a package for text mining in R. It provides functions for reading, cleaning, and analyzing text data.
- stringr: stringr is a package for working with strings in R. It provides functions for extracting, manipulating, and formatting strings.
- tidyverse: tidyverse is a collection of data manipulation and analysis packages in R. It provides a consistent and intuitive interface for working with data.
- tidytext: tidytext is a package for text mining with tidy data. It provides functions for converting text data into tidy format and for performing text analysis on tidy data.
- wordcloud: wordcloud is a package for creating word clouds in R. It provides functions for generating word clouds from text data.
- quanteda: quanteda is a package for quantitative text analysis in R. It provides functions for reading, cleaning, and analyzing text data in a quantitative way.

- syuzhet: syuzhet is a package for sentiment analysis in R. It provides functions for extracting sentiment from text data.
- cluster: cluster is a package for cluster analysis in R. It provides functions for clustering data points into groups.

STEPS FOR INSTALLATION and Creating VCorpus:

1. Install the pacman package and load the pacman library.

```
> install.packages("pacman")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/pacman_0.5.1.tgz'
Content type 'application/x-gzip' length 383212 bytes (374 KB)
==================================================
downloaded 374 KB


The downloaded binary packages are in
        /var/folders/q9/_br0b2x124sc9rvkls9sgl000000gn/T//Rtmp7tpN8V/downloaded_packages
> install.packages('pacman')
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/pacman_0.5.1.tgz'
Content type 'application/x-gzip' length 383212 bytes (374 KB)
==================================================
downloaded 374 KB


The downloaded binary packages are in
        /var/folders/q9/_br0b2x124sc9rvkls9sgl000000gn/T//Rtmp7tpN8V/downloaded_packages
> library(pacman)
>
```

2. Load the following packages: tm, stringr, tidyverse, tidytext, wordcloud, quanteda, syuzhet, and cluster.

```
> library(pacman)
> p_load('tm','stringr','tidyverse','tidytext','wordcloud','quanteda','syuzhet','cluster')
>
```

3. Set and get the working directory to the location of your project files.

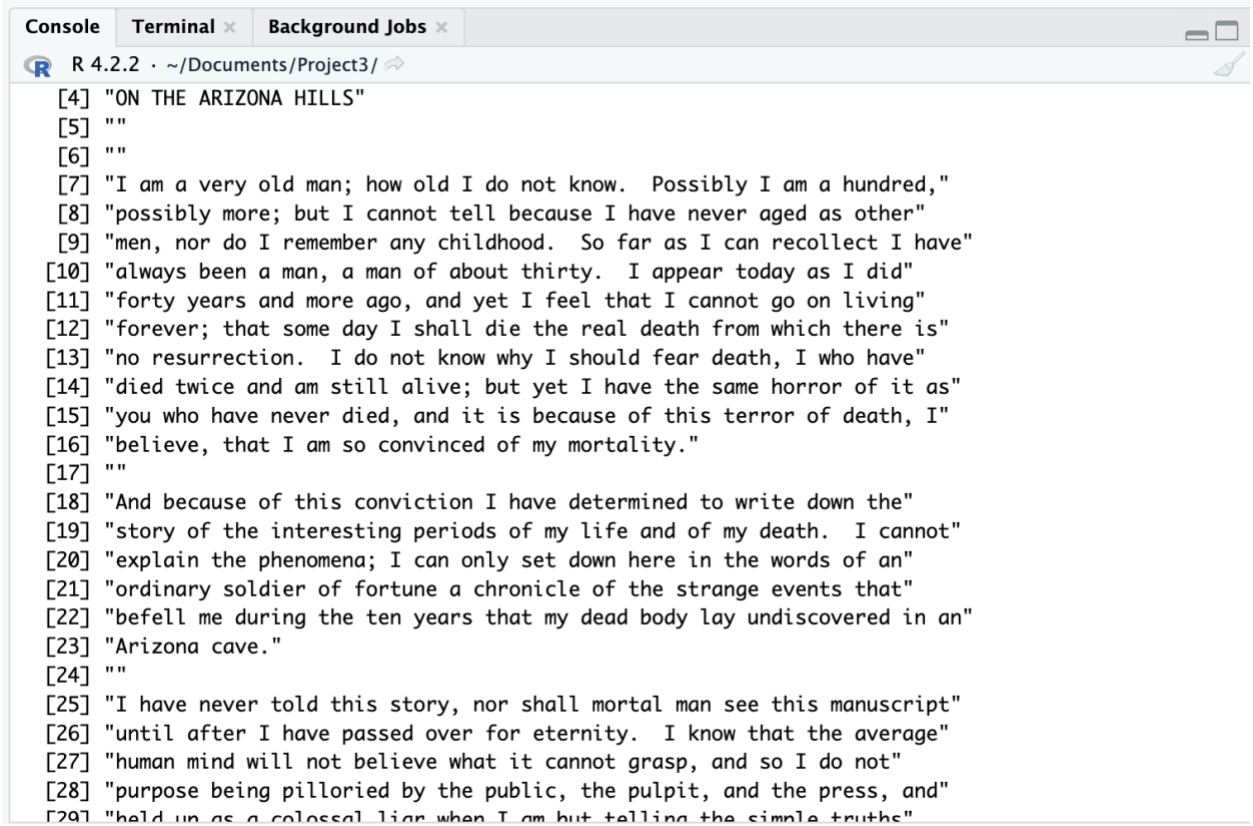
```
> p_load('tm','stringr','tidyverse','tidytext','wordcloud','quanteda','syuzhet','cluster')
> setwd('/Users/sumanthnandeti/Documents/Project3')
> getwd()
[1] "/Users/sumanthnandeti/Documents/Project3"
>
```

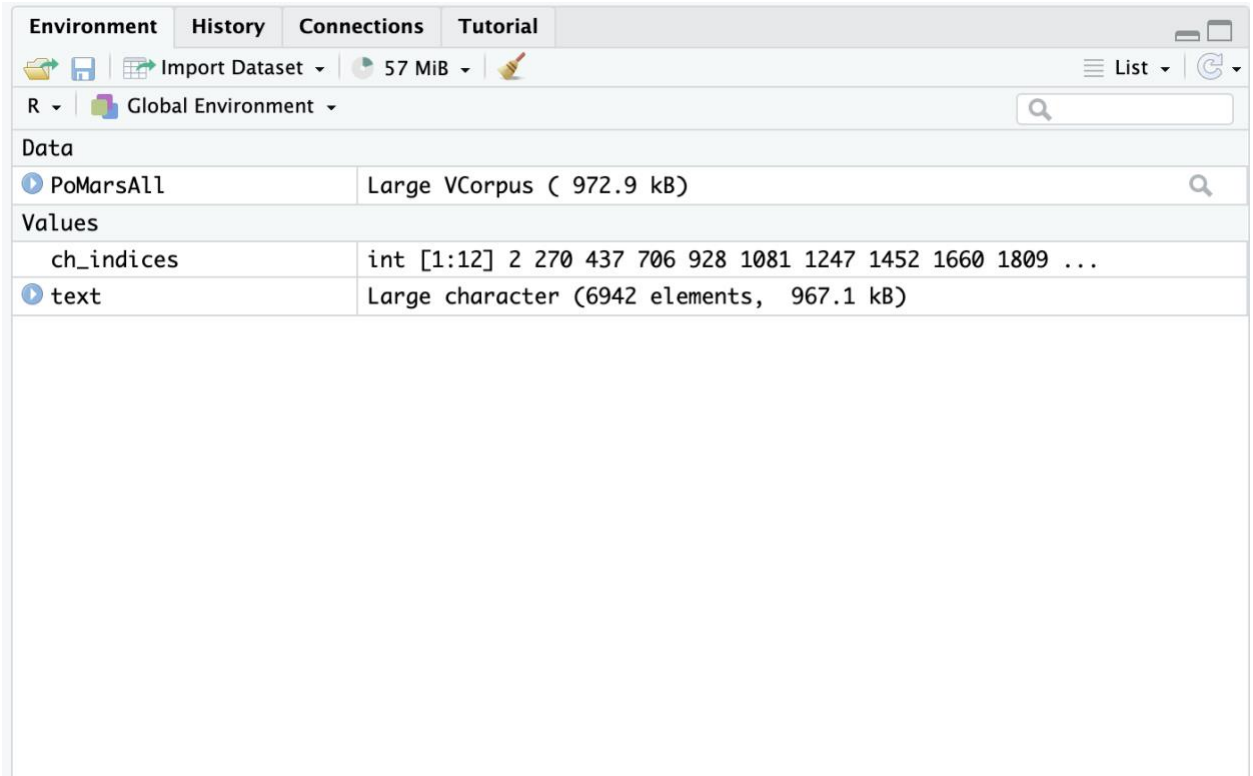
4. Create a VCorpus object from the text files in the GroupProject directory.

Extracting the Chapters of a Princess of Mars

Get the text of the first document in the VCorpus object PoMarsAll and find the indices of all the occurrences of the string "CHAPTER" in the character vector text. The code then restricts the vector ch_indices to the first 12 indices. This is because the novel A Princess of Mars has 12 chapters.

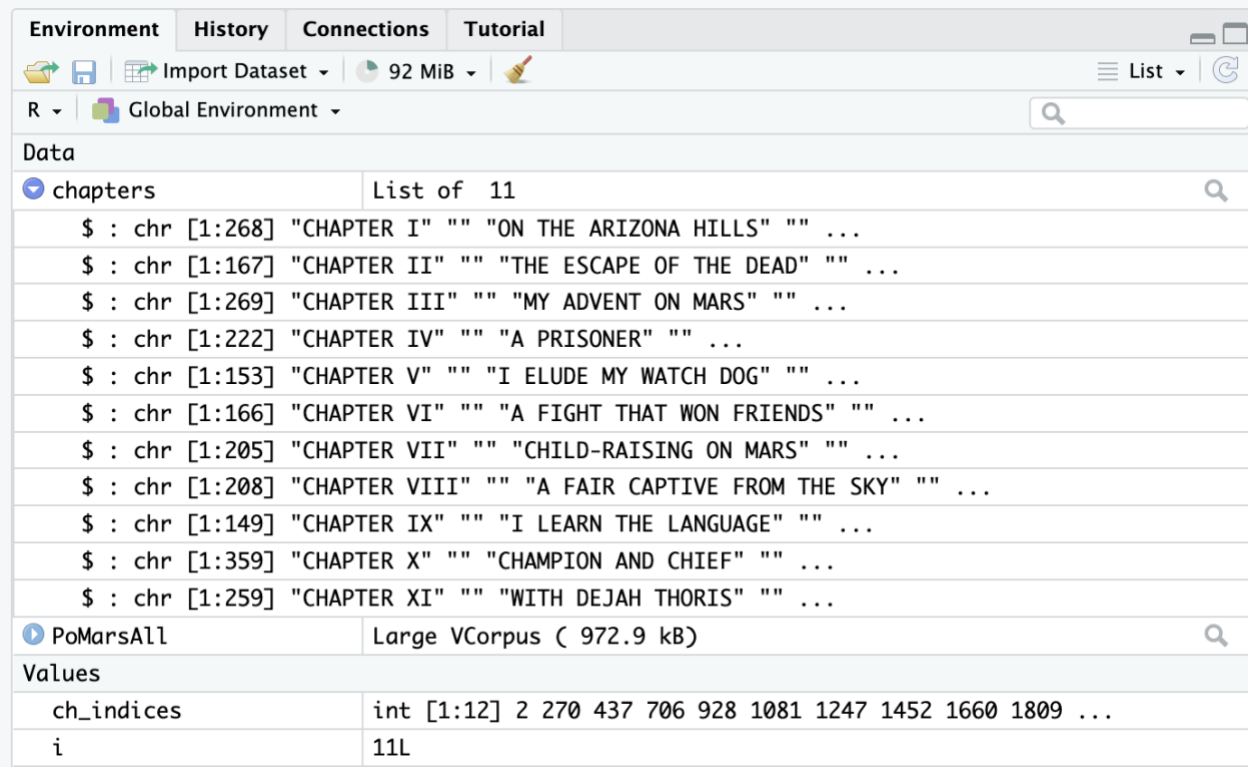
```
text=content(PoMarsAll[[1]])
text
ch_indices = str_which(text, pattern = "^CHAPTER\\s")
ch_indices=ch_indices[1:12]
```


Console
Terminal
Background Jobs
R 4.2.2 · ~/Documents/Project3/
[4] "ON THE ARIZONA HILLS"
[5] ""
[6] ""
[7] "I am a very old man; how old I do not know. Possibly I am a hundred,"
[8] "possibly more; but I cannot tell because I have never aged as other"
[9] "men, nor do I remember any childhood. So far as I can recollect I have"
[10] "always been a man, a man of about thirty. I appear today as I did"
[11] "forty years and more ago, and yet I feel that I cannot go on living"
[12] "forever; that some day I shall die the real death from which there is"
[13] "no resurrection. I do not know why I should fear death, I who have"
[14] "died twice and am still alive; but yet I have the same horror of it as"
[15] "you who have never died, and it is because of this terror of death, I"
[16] "believe, that I am so convinced of my mortality."
[17] ""
[18] "And because of this conviction I have determined to write down the"
[19] "story of the interesting periods of my life and of my death. I cannot"
[20] "explain the phenomena; I can only set down here in the words of an"
[21] "ordinary soldier of fortune a chronicle of the strange events that"
[22] "befell me during the ten years that my dead body lay undiscovered in an"
[23] "Arizona cave."
[24] ""
[25] "I have never told this story, nor shall mortal man see this manuscript"
[26] "until after I have passed over for eternity. I know that the average"
[27] "human mind will not believe what it cannot grasp, and so I do not"
[28] "purpose being pilloried by the public, the pulpit, and the press, and"


Environment
History
Connections
Tutorial
Import Dataset
57 MiB
List
R
Global Environment
Data
PoMarsAll
Large VCorpus (972.9 kB)
Values
ch_indices
int [1:12] 2 270 437 706 928 1081 1247 1452 1660 1809 ...
text
Large character (6942 elements, 967.1 kB)

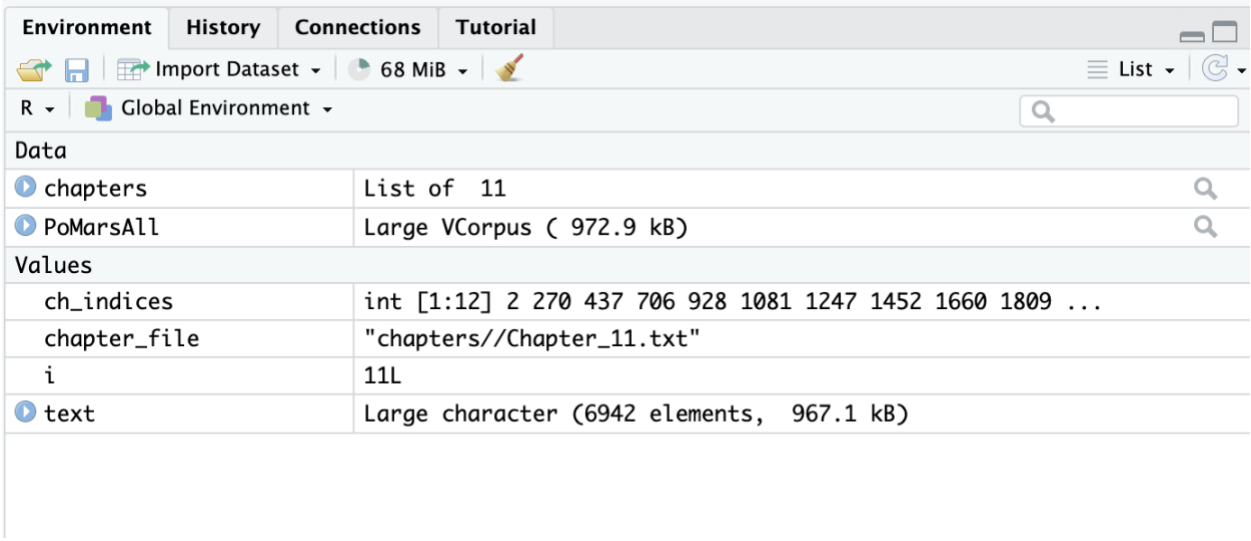
Splitting the Chapters:

```
> ch_indices = str_which(text, pattern = "^CHAPTER\\s")
> ch_indices=ch_indices[1:12]
> chapters = list()
> for (i in 1:(length(ch_indices) - 1)) {
+   chapters[[i]] = text[ch_indices[i]: (ch_indices[i + 1] - 1)]
+ }
>
```



Creating a directory of chapters I-XI:

```
> dir.create("chapters")
> for (i in seq_along(chapters)) {
+   chapter_file <- file.path("chapters/", paste0("Chapter_", i, ".txt"))
+   writeLines(chapters[[i]], chapter_file)
+ }
>
```



Creating a Vcorpus for I-XI chapters:

First, creates a VCorpus object from the text files in the chapter's directory. This object is then converted to a tidy data frame. The text of the first chapter is extracted from the tidy data frame, and the number of the first chapter is removed from the tidy data frame.

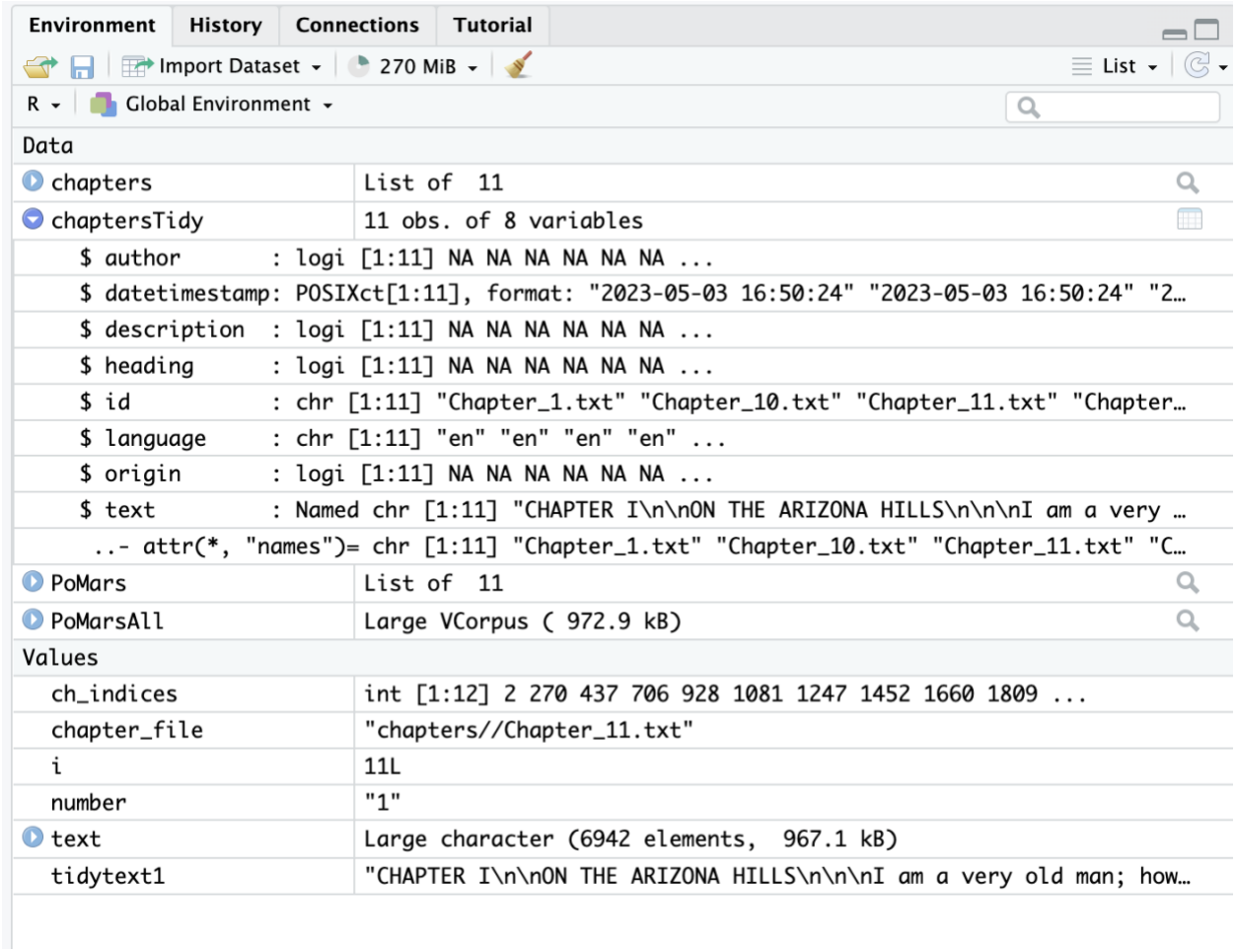
I extracted the text and number of each chapter in the chapter's directory. This information can be used to analyze the text of the chapters or to create a table of the chapters.

```
#creating a vcorpus for I-XI chapters
PoMars=VCorpus(DirSource("/Users/sumanthnandeti/Documents/Project3/chapters",ignore.case = TRUE
str(PoMars)
PoMars
chaptersTidy=tidy(PoMars)
tidytext1=chaptersTidy$text[1]
number <- gsub("[^0-9]", "", chaptersTidy$id[1])
```

Console | Terminal | Background Jobs

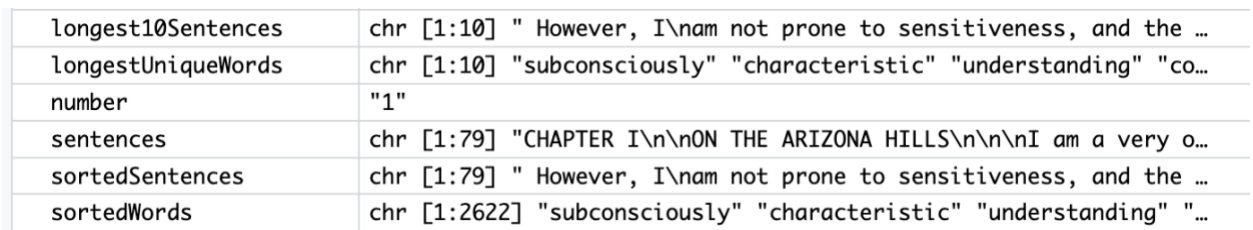
R 4.2.2 · ~/Documents/Project3/

```
> PoMars=VCorpus(DirSource("/Users/sumanthnandeti/Documents/Project3/chapters",ignore.case = TRUE,mode
= "text"))
> str(PoMars)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 11
  ..$ :List of 2
  .. ..$ content: chr [1:268] "CHAPTER I" "" "ON THE ARIZONA HILLS" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-03 16:50:24"
  .. .. ..$ 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 [1:359] "CHAPTER X" "" "CHAMPION AND CHIEF" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-03 16:50:24"
  .. .. ..$ 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 [1:259] "CHAPTER XI" "" "WITH DEJAH THORIS" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-03 16:50:24"
  .. .. ..$ description  : chr(0)
```



Finding 10 largest words and sentences in the first chapter:

```
#finding 10largest words and sentences in first chapter

words <- str_extract_all(tidytext1, "\\w+") %>% unlist()
sortedWords <- words[order(nchar(words), decreasing = TRUE)]
longestUniqueWords <- sortedWords %>% unique() %>% head(10)

sentences <- str_split(tidytext1, "\\.\\s") %>% unlist()
sortedSentences <- sentences[order(nchar(sentences), decreasing = TRUE)]
longest10Sentences <- sortedSentences %>% unique() %>% head(10)
```



defining a function to find 10 largest words and sentences for all XI chapters.

```
find10longWordsSentences <- function(tidytext1,num1) {
  # Find the 10 longest words
  chapter_num <- gsub("[^0-9]", "", chaptersTidy$id[num1])
  words <- str_extract_all(tidytext1, "\\w+") %>% unlist()
  sortedWords <- words[order(nchar(words), decreasing = TRUE)]
  longest10Words <- sortedWords %>% unique() %>% head(10)

  sentences <- str_split(tidytext1, "\\.\\s") %>% unlist()
  sortedSentences <- sentences[order(nchar(sentences), decreasing = TRUE)]
  longest10Sentences <- sortedSentences %>% unique() %>% head(10)

  return(tibble(Chapter = chapter_num,
                ItemType = rep(c("Word", "Sentence"), each = 10),
                Item = c(longest10Words, longest10Sentences),
                Length = c(nchar(longest10Words), nchar(longest10Sentences))))
}
```

Now Use Loop and call the function which has defined above through each text file and read its contents into the vector.

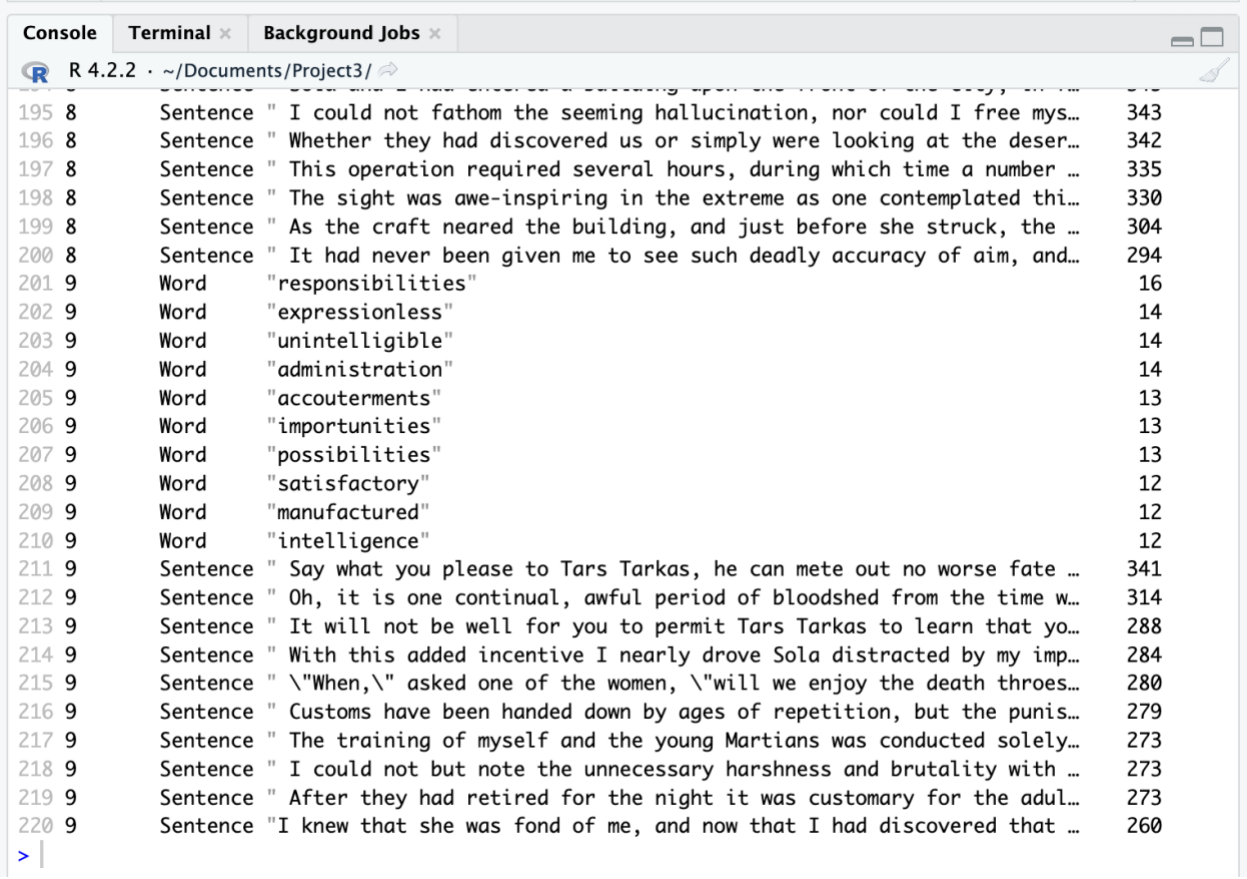
```
for (i in seq_along(chapter_files)) {
  chapter_text[i] <- get_text_as_string(file.path("/Users/sumanthnandeti/Documents/Project3/chap
}
results_table <- map_dfr(seq_along(chapters), ~find10longWordsSentences(chapter_text[.],.))
print(results_table,n = 500)
```

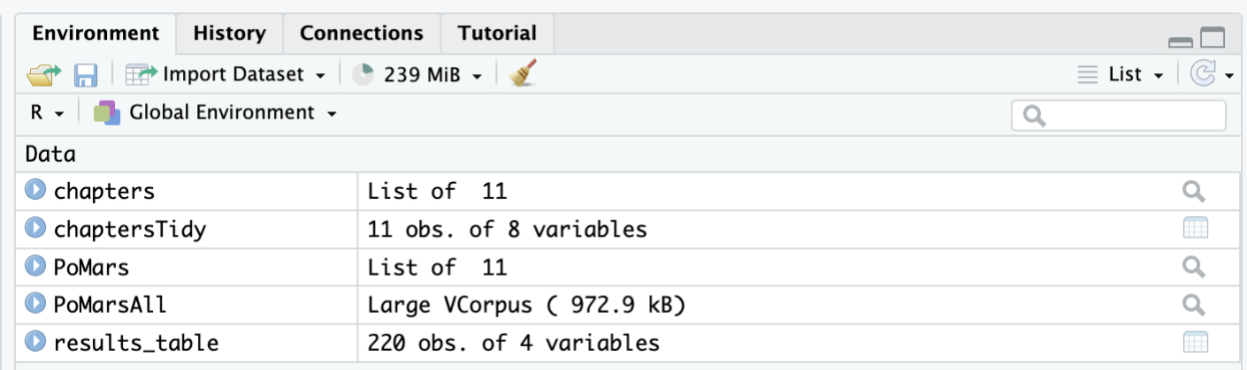
Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
+ }
> results_table <- map_dfr(seq_along(chapters), ~find10longWordsSentences(chapter_text[.],.))
> print(results_table,n = 500)
# A tibble: 220 × 4
```

	Chapter	ItemType	Item	Length
	<chr>	<chr>	<chr>	<int>
1	1	Word	"subconsciously"	14
2	1	Word	"characteristic"	14
3	1	Word	"understanding"	13
4	1	Word	"comparatively"	13
5	1	Word	"sensitiveness"	13
6	1	Word	"consternation"	13
7	1	Word	"perpendicular"	13
8	1	Word	"resuscitation"	13
9	1	Word	"resurrection"	12
10	1	Word	"undiscovered"	12
11	1	Sentence	" However, I am not prone to sensitiveness, and the following of a s…	376
12	1	Sentence	" The fact that it is difficult to aim anything but imprecations acc…	371
13	1	Sentence	" Since we had entered the territory we had not seen a hostile India…	364
14	1	Sentence	" In this instance I was, of course, positive that Powell was the ce…	346
15	1	Sentence	" The morning of Powell's departure was, like nearly all Arizona mor…	325
16	1	Sentence	" My horse was traveling practically unguided as I knew that I had p…	309
17	1	Sentence	" I do not believe that I am made of the stuff which constitutes her…	288
18	1	Sentence	" I was positive now that the trailers were Apaches and that they wi…	278
19	1	Sentence	" I soon became so drowsy that I could scarcely resist the strong de…	272
20	1	Sentence	" I know that the average human mind will not believe what it cannot…	268
21	10	Word	"responsibilities"	16
22	10	Word	"characteristics"	15





Printing All chapter sentences

```
for (chapter in 1:11) {
  cat("Chapter", chapter, "\n")
  results_table %>%
    filter(Chapter == chapter) %>%
    print(n = 20)
}
```

Chapter 1:

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
Chapter 1
# A tibble: 20 × 4
   Chapter ItemType Item                                                  Length
   <chr>   <chr>    <chr>                                                  <int>
 1 1       Word     "subconsciously"                                          14
 2 1       Word     "characteristic"                                          14
 3 1       Word     "understanding"                                           13
 4 1       Word     "comparatively"                                           13
 5 1       Word     "sensitiveness"                                           13
 6 1       Word     "consternation"                                           13
 7 1       Word     "perpendicular"                                           13
 8 1       Word     "resuscitation"                                           13
 9 1       Word     "resurrection"                                            12
10 1       Word     "undiscovered"                                            12
11 1       Sentence " However, I am not prone to sensitiveness, and the following of a se…   376
12 1       Sentence " The fact that it is difficult to aim anything but imprecations accu…   371
13 1       Sentence " Since we had entered the territory we had not seen a hostile Indian…   364
14 1       Sentence " In this instance I was, of course, positive that Powell was the cen…   346
15 1       Sentence " The morning of Powell's departure was, like nearly all Arizona morn…   325
16 1       Sentence " My horse was traveling practically unguided as I knew that I had pr…   309
17 1       Sentence " I do not believe that I am made of the stuff which constitutes hero…   288
18 1       Sentence " I was positive now that the trailers were Apaches and that they wis…   278
19 1       Sentence " I soon became so drowsy that I could scarcely resist the strong des…   272
20 1       Sentence " I know that the average human mind will not believe what it cannot …   268
```

Chapter 2:

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
Chapter 2
# A tibble: 20 × 4
   Chapter ItemType Item                                                  Length
   <chr>   <chr>    <chr>                                                  <int>
 1 2       Word     "contemplation"                                           13
 2 2       Word     "metamorphosis"                                           13
 3 2       Word     "particularly"                                            12
 4 2       Word     "predicaments"                                            12
 5 2       Word     "interruption"                                            12
 6 2       Word     "overstrained"                                            12
 7 2       Word     "bewilderment"                                            12
 8 2       Word     "surroundings"                                            12
 9 2       Word     "unfathomable"                                            12
10 2       Word     "crystallized"                                            12
11 2       Sentence " Few western wonders are more inspiring than the beauties of an Ariz…   455
12 2       Sentence " I reasoned with myself that I had lain helpless for many hours with…   393
13 2       Sentence " To be held paralyzed, with one's back toward some horrible and unkn…   372
14 2       Sentence " Fear is a relative term and so I can only measure my feelings at th…   358
15 2       Sentence " Late in the afternoon my horse, which had been standing with draggi…   351
16 2       Sentence " My first thought was, is this then death!  Have I indeed passed ove…   274
17 2       Sentence " From then until possibly midnight all was silence, the silence of t…   263
18 2       Sentence " My only alternative seemed to lie in flight and my decision was cry…   243
19 2       Sentence " There also came to my nostrils a faintly pungent odor, and I could …   216
20 2       Sentence "CHAPTER II  THE ESCAPE OF THE DEAD   A sense of delicious dreaminess…   213
```

Chapter 3:

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
Chapter 3
# A tibble: 20 × 4
   Chapter ItemType Item                                                  Length
   <chr>   <chr>    <chr>                                                  <int>
 1 3       Word     "characteristics"                                         15
 2 3       Word     "irregularities"                                          14
 3 3       Word     "characteristic"                                          14
 4 3       Word     "consciousness"                                           13
 5 3       Word     "independently"                                           13
 6 3       Word     "accouterments"                                           13
 7 3       Word     "noiselessness"                                           13
 8 3       Word     "gesticulating"                                           13
 9 3       Word     "comparatively"                                           13
10 3       Word     "interminable"                                            12
11 3       Sentence " The throwing down of his weapons and the withdrawing of his troop b…   407
12 3       Sentence " He sat his mount as we sit a horse, grasping the animal's barrel wi…   335
13 3       Sentence " Their eyes were set at the extreme sides of their heads a trifle ab…   334
14 3       Sentence " And his mount!  How can earthly words describe it!  It towered ten …   304
15 3       Sentence " The respite my unexpected agility had given me permitted me to form…   287
16 3       Sentence "The result is that they are infinitely less agile and less powerful,…   282
17 3       Sentence " Coming, as they did, over the soft and soundless moss, which covers…   279
18 3       Sentence " But the little sound caused me to turn, and there upon me, not ten …   262
19 3       Sentence " Instead of progressing in a sane and dignified manner, my attempts …   253
20 3       Sentence " Behind this first charging demon trailed nineteen others, similar i…   250
```

Chapter 4:

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
Chapter 4
# A tibble: 20 × 4
   Chapter ItemType Item                                                  Length
   <chr>   <chr>    <chr>                                                  <int>
 1 4       Word     "circumstances"                                           13
 2 4       Word     "consideration"                                           13
 3 4       Word     "manifestation"                                           13
 4 4       Word     "therapeutics"                                            12
 5 4       Word     "scintillated"                                            12
 6 4       Word     "introduction"                                            12
 7 4       Word     "instructions"                                            12
 8 4       Word     "observation"                                             11
 9 4       Word     "immediately"                                             11
10 4       Word     "surrounding"                                             11
11 4       Sentence " I saw no signs of extreme age among them, nor is there any apprecia…   448
12 4       Sentence " They first repeated the word \"sak\" a number of times, and then Ta…   410
13 4       Sentence " What struck me as most remarkable about this assemblage and the hal…   409
14 4       Sentence " My exhibition had been witnessed by several hundred lesser Martians…   370
15 4       Sentence " The room was well lighted by a number of large windows and was beau…   356
16 4       Sentence " Owing to the waning resources of the planet it evidently became nec…   352
17 4       Sentence "Evidently, then, there were other denizens on Mars than the wild and…   305
18 4       Sentence " As he banged me down upon my feet his face was bent close to mine a…   284
19 4       Sentence " Toward the center of the city was a large plaza, and upon this and …   275
20 4       Sentence " Had the men been strangers, and therefore unable to exchange names,…   270
```

Chapter 5:

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
Chapter 5
# A tibble: 20 × 4
   Chapter ItemType Item                                                  Length
   <chr>   <chr>    <chr>                                                  <int>
 1 5       Word     "characteristics"                                         15
 2 5       Word     "representation"                                          14
 3 5       Word     "ministrations"                                           13
 4 5       Word     "uncomfortable"                                           13
 5 5       Word     "intellectual"                                            12
 6 5       Word     "intelligence"                                            12
 7 5       Word     "straightaway"                                            12
 8 5       Word     "considerable"                                            12
 9 5       Word     "monstrosity"                                             11
10 5       Word     "voluntarily"                                             11
11 5       Sentence " The nights are either brilliantly illumined or very dark, for if ne…   391
12 5       Sentence " This last device produces an intensely brilliant far-reaching white…   366
13 5       Sentence " I could not but wonder what this ferocious-looking monstrosity migh…   347
14 5       Sentence " And it is well that nature has so graciously and abundantly lighted…   345
15 5       Sentence " It came, as I later discovered, not from an animal, as there is onl…   311
16 5       Sentence " The work had evidently been wrought by a master hand, so subtle the…   279
17 5       Sentence " Both of Mars' moons are vastly nearer her than is our moon to Earth…   279
18 5       Sentence "The nearer moon of Mars makes a complete revolution around the plane…   274
19 5       Sentence " Across the threshold lay stretched the sleepless guardian brute, ju…   267
20 5       Sentence " This girl alone, among all the green Martians with whom I came in c…   266
```

Chapter 6:

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
Chapter 6
# A tibble: 20 × 4
   Chapter ItemType Item                                                  Length
   <chr>   <chr>    <chr>                                                  <int>
 1 6       Word     "overwhelmingly"                                          14
 2 6       Word     "accomplishing"                                           13
 3 6       Word     "gesticulated"                                            12
 4 6       Word     "intermediary"                                            12
 5 6       Word     "transcending"                                            12
 6 6       Word     "commencement"                                            12
 7 6       Word     "executioner"                                             11
 8 6       Word     "momentarily"                                             11
 9 6       Word     "perceptibly"                                             11
10 6       Word     "forebodings"                                             11
11 6       Sentence " My beast had an advantage in his first hold, having sunk his mighty…   430
12 6       Sentence " I am ever willing to stand and fight when the odds are not too over…   377
13 6       Sentence " I had at least two friends on Mars; a young woman who watched over …   342
14 6       Sentence " Suddenly I came to myself and, with that strange instinct which see…   337
15 6       Sentence " It is true I held the cudgel, but what could I do with it against h…   296
16 6       Sentence " Evidently devoid of all the finer sentiments of friendship, love, o…   290
17 6       Sentence " I was standing near the window and I knew that once in the street I…   264
18 6       Sentence " With a shriek of fear the ape which held me leaped through the open…   255
19 6       Sentence "CHAPTER VI  A FIGHT THAT WON FRIENDS  The thing, which more nearly …   253
20 6       Sentence " I glimpsed him just before he reached the doorway and the sight of …   252
```

Chapter 7:

```
Chapter 7
# A tibble: 20 × 4
   Chapter ItemType Item                                                    Length
   <chr>   <chr>    <chr>                                                    <int>
 1 7       Word     "representative"                                            14
 2 7       Word     "conversations"                                             13
 3 7       Word     "unnecessarily"                                             13
 4 7       Word     "intentionally"                                             13
 5 7       Word     "circumstances"                                             13
 6 7       Word     "intellectual"                                              12
 7 7       Word     "conversation"                                              12
 8 7       Word     "satisfaction"                                              12
 9 7       Word     "appropriated"                                              12
10 7       Word     "humanitarian"                                              12
11 7       Sentence " Between these walls the little Martians scampered, wild as deer; be…    431
12 7       Sentence "CHAPTER VII  CHILD-RAISING ON MARS  After a breakfast, which was an…    384
13 7       Sentence " As I came to a halt before him, Tars Tarkas pointed over the incuba…    357
14 7       Sentence " I do not mean that the adult Martians are unnecessarily or intentio…    331
15 7       Sentence " Entirely unknown to their mothers, who, in turn, would have difficu…    276
16 7       Sentence " They were not wanted, as their offspring might inherit and transmit…    273
17 7       Sentence " Every one but myself--men, women, and children--were heavily armed,…    270
18 7       Sentence " It is the universal language of Mars, through the medium of which t…    267
19 7       Sentence " As I later learned, they had been to the subterranean vaults in whi…    263
20 7       Sentence " Sola's duties were now doubled, as she was compelled to care for th…    261
```

Chapter 8:

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/Documents/Project3/

Chapter 8
# A tibble: 20 × 4
   Chapter ItemType Item                                                    Length
   <chr>   <chr>    <chr>                                                    <int>
 1 8       Word     "reinforcements"                                            14
 2 8       Word     "simultaneously"                                            14
 3 8       Word     "circumstances"                                             13
 4 8       Word     "southeasterly"                                             13
 5 8       Word     "requisitioned"                                             13
 6 8       Word     "southwesterly"                                             13
 7 8       Word     "unaccountably"                                             13
 8 8       Word     "hallucination"                                             13
 9 8       Word     "majestically"                                              12
10 8       Word     "irresistible"                                              12
11 8       Sentence " For example, a proportion of them, always the best marksmen, direct…    453
12 8       Sentence " Instantly the scene changed as by magic; the foremost vessel swung …    449
13 8       Sentence " As Sola and I entered the plaza a sight met my eyes which filled my…    408
14 8       Sentence " Sola and I had entered a building upon the front of the city, in fa…    345
15 8       Sentence " I could not fathom the seeming hallucination, nor could I free myse…    343
16 8       Sentence " Whether they had discovered us or simply were looking at the desert…    342
17 8       Sentence " This operation required several hours, during which time a number o…    335
18 8       Sentence " The sight was awe-inspiring in the extreme as one contemplated this…    330
19 8       Sentence " As the craft neared the building, and just before she struck, the M…    304
20 8       Sentence " It had never been given me to see such deadly accuracy of aim, and …    294
```

Chapter 9:

```
Chapter 9
# A tibble: 20 × 4
   Chapter ItemType Item                                                    Length
   <chr>   <chr>    <chr>                                                    <int>
 1 9       Word     "responsibilities"                                          16
 2 9       Word     "expressionless"                                            14
 3 9       Word     "unintelligible"                                            14
 4 9       Word     "administration"                                            14
 5 9       Word     "accouterments"                                             13
 6 9       Word     "importunities"                                             13
 7 9       Word     "possibilities"                                             13
 8 9       Word     "satisfactory"                                              12
 9 9       Word     "manufactured"                                              12
10 9       Word     "intelligence"                                              12
11 9       Sentence " Say what you please to Tars Tarkas, he can mete out no worse fate t…   341
12 9       Sentence " Oh, it is one continual, awful period of bloodshed from the time we…   314
13 9       Sentence " It will not be well for you to permit Tars Tarkas to learn that you…   288
14 9       Sentence " With this added incentive I nearly drove Sola distracted by my impo…   284
15 9       Sentence " \"When,\" asked one of the women, \"will we enjoy the death throes …   280
16 9       Sentence " Customs have been handed down by ages of repetition, but the punish…   279
17 9       Sentence " The training of myself and the young Martians was conducted solely …   273
18 9       Sentence " I could not but note the unnecessary harshness and brutality with w…   273
19 9       Sentence " After they had retired for the night it was customary for the adult…   273
20 9       Sentence "I knew that she was fond of me, and now that I had discovered that s…   260
```

Chapter 10:

```
Chapter 10
# A tibble: 20 × 4
   Chapter ItemType Item                                                    Length
   <chr>   <chr>    <chr>                                                    <int>
 1 10      Word     "responsibilities"                                          16
 2 10      Word     "characteristics"                                           15
 3 10      Word     "companionship"                                             13
 4 10      Word     "manifestation"                                             13
 5 10      Word     "ludicrousness"                                             13
 6 10      Word     "precipitately"                                             13
 7 10      Word     "authoritative"                                             13
 8 10      Word     "understanding"                                             13
 9 10      Word     "theoretically"                                             13
10 10      Word     "granddaughter"                                             13
11 10      Sentence " Suffice it, for the present, that I am your friend, and, so far as …   721
12 10      Sentence " I understand that you belittle all sentiments of generosity and kin…   543
13 10      Sentence " What words of moment were to have fallen from his lips were never s…   430
14 10      Sentence " Numerous brilliantly colored and strangely formed wild flowers dott…   415
15 10      Sentence " I saw that the body of my dead antagonist had been stripped, and I …   415
16 10      Sentence " What strange manner of man are you, that you consort with the green…   404
17 10      Sentence " Realizing that I was a somewhat favored character, and also convinc…   386
18 10      Sentence " I was soon successful as her injuries amounted to little more than …   352
19 10      Sentence " Tars Tarkas himself seemed pleased with my reply, but his only comm…   349
20 10      Sentence " The reason for the whole attitude displayed toward me was now appar…   344
```

Chapter 11:

```
Chapter 11
# A tibble: 20 × 4
   Chapter ItemType Item                                                    Length
   <chr>   <chr>    <chr>                                                    <int>
 1 11      Word     "accouterments"                                             13
 2 11      Word     "circumstances"                                             13
 3 11      Word     "questioningly"                                             13
 4 11      Word     "eavesdropping"                                             13
 5 11      Word     "comparatively"                                             13
 6 11      Word     "intermarrying"                                             13
 7 11      Word     "irretrievably"                                             13
 8 11      Word     "architecture"                                              12
 9 11      Word     "compositions"                                              12
10 11      Word     "conversation"                                              12
11 11      Sentence " \"And whereto, then, would your prisoner escape should you leave he…    634
12 11      Sentence "During the ages of hardships and incessant warring between their own…    520
13 11      Sentence " \"Because, John Carter,\" she replied, \"nearly every planet and st…    491
14 11      Sentence " Do not tell me that you have thus returned!  They would kill you ho…    345
15 11      Sentence " A similar wave of feeling seemed to stir her; she drew away from me…    340
16 11      Sentence " \"The fact that you wore no ornaments is a strong proof of your un-…    337
17 11      Sentence " I can readily perceive that you are not of the Barsoom of today; yo…    335
18 11      Sentence " The shores of the ancient seas were dotted with just such cities, a…    327
19 11      Sentence " These three great divisions of the higher Martians had been forced …    307
20 11      Sentence " These ancient Martians had been a highly cultivated and literary ra…    292
```

DocumentTermMatrix:

The DocumentTermMatrix function creates a document-term matrix from a corpus of text. A document-term matrix is a sparse matrix that represents the frequency of terms that occur in each document. The rows of the matrix correspond to the documents, and the columns correspond to the terms.

```
> PoMarsDTM=DocumentTermMatrix(PoMars)
> PoMarsDTM
<<DocumentTermMatrix (documents: 11, terms: 5007)>>
Non-/sparse entries: 9218/45859
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
```

The str function is used to print the structure of the matrix objects, while the inspect function is used to print the contents of the matrix.

```
> inspect(PoMarsDTM)
<<DocumentTermMatrix (documents: 11, terms: 5007)>>
Non-/sparse entries: 9218/45859
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs             and but for had that the upon was which with
  Chapter_1.txt   90  13  20  25   50 190   12  36    20   25
  Chapter_10.txt 113  18  34  41   50 193   18  54    15   26
  Chapter_11.txt  82  14   9  29   37 119   13  18    13   23
  Chapter_2.txt   58  13  17  20   19 137   13  31    12   10
  Chapter_3.txt   92  15  18  19   26 166   20  36    34   20
  Chapter_4.txt   68  21  10  22   19 148   10  26    16   20
  Chapter_5.txt   56  14  13  14   11  96    8  22     9    8
  Chapter_6.txt   53  10  14  28   11 117   13  17    11   24
  Chapter_7.txt   64   9  18  16   12 166    7  17    22   12
  Chapter_8.txt   81   7  15  24   11 184   22  32    17   11
```

```
> str(PoMarsDTM)
List of 6
 $ i       : int [1:9218] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:9218] 36 37 38 41 42 60 68 74 78 81 ...
 $ v       : num [1:9218] 1 1 1 1 7 1 1 4 1 1 ...
 $ nrow    : int 11
 $ ncol    : int 5007
 $ dimnames:List of 2
  ..$ Docs : chr [1:11] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_2.txt" ...
  ..$ Terms: chr [1:5007] "'gentleman'" "\"and" "\"as" "\"because," ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

TermDocumentMatrix:

The TermDocumentMatrix function creates a term-document matrix from a corpus of text. A term-document matrix is a sparse matrix that represents the frequency of documents that contain each term. The rows of the matrix correspond to the terms, and the columns correspond to the documents.

```
> #Term Document Matrix of entire file
> PoMarsTDM=TermDocumentMatrix(PoMars)
> PoMarsTDM
<<TermDocumentMatrix (terms: 5007, documents: 11)>>
Non-/sparse entries: 9218/45859
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
```

The str function is used to print the structure of the matrix objects, while the inspect function is used to print the contents of the matrix.

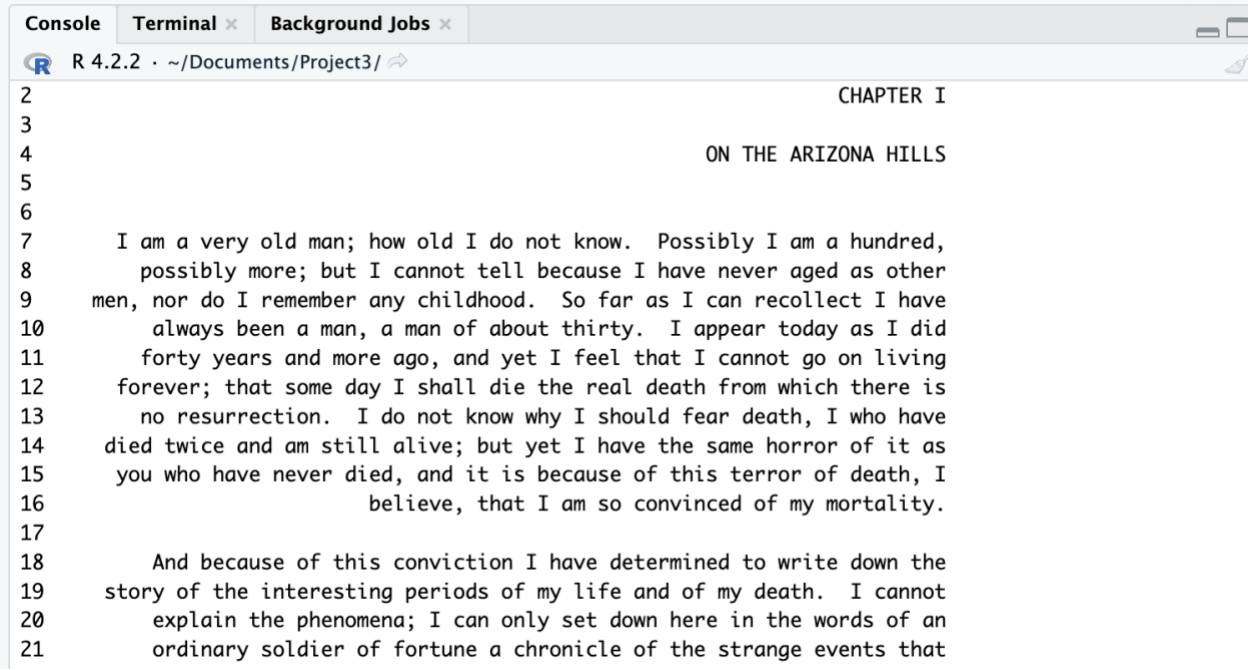
```
> inspect(PoMarsTDM)
<<TermDocumentMatrix (terms: 5007, documents: 11)>>
Non-/sparse entries: 9218/45859
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
       Docs
Terms   Chapter_1.txt Chapter_10.txt Chapter_11.txt Chapter_2.txt Chapter_3.txt Chapter_4.txt
  and              90            113             82            58            92            68
  but              13             18             14            13            15            21
  for              20             34              9            17            18            10
  had              25             41             29            20            19            22
  that             50             50             37            19            26            19
  the             190            193            119           137           166           148
  upon             12             18             13            13            20            10
  was              36             54             18            31            36            26
  which            20             15             13            12            34            16
  with             25             26             23            10            20            20
       Docs
Terms   Chapter_5.txt Chapter_6.txt Chapter_7.txt Chapter_8.txt
  and              56            53             64            81
  but              14            10              9             7
  for              13            14             18            15
  had              14            28             16            24
  that             11            11             12            11
  the              96           117            166           184
  upon              8            13              7            22
  was              22            17             17            32
  which             9            11             22            17
  with              8            24             12            11
```

```
> str(PoMarsTDM)
List of 6
 $ i       : int [1:9218] 36 37 38 41 42 60 68 74 78 81 ...
 $ j       : int [1:9218] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:9218] 1 1 1 1 7 1 1 4 1 1 ...
 $ nrow    : int 5007
 $ ncol    : int 11
 $ dimnames:List of 2
  ..$ Terms: chr [1:5007] "'gentleman'" "\"and" "\"as" "\"because," ...
  ..$ Docs : chr [1:11] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_2.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

Creating a data frame of PoMars text file

PoMarsDF=data.frame(text)

PoMarsDF



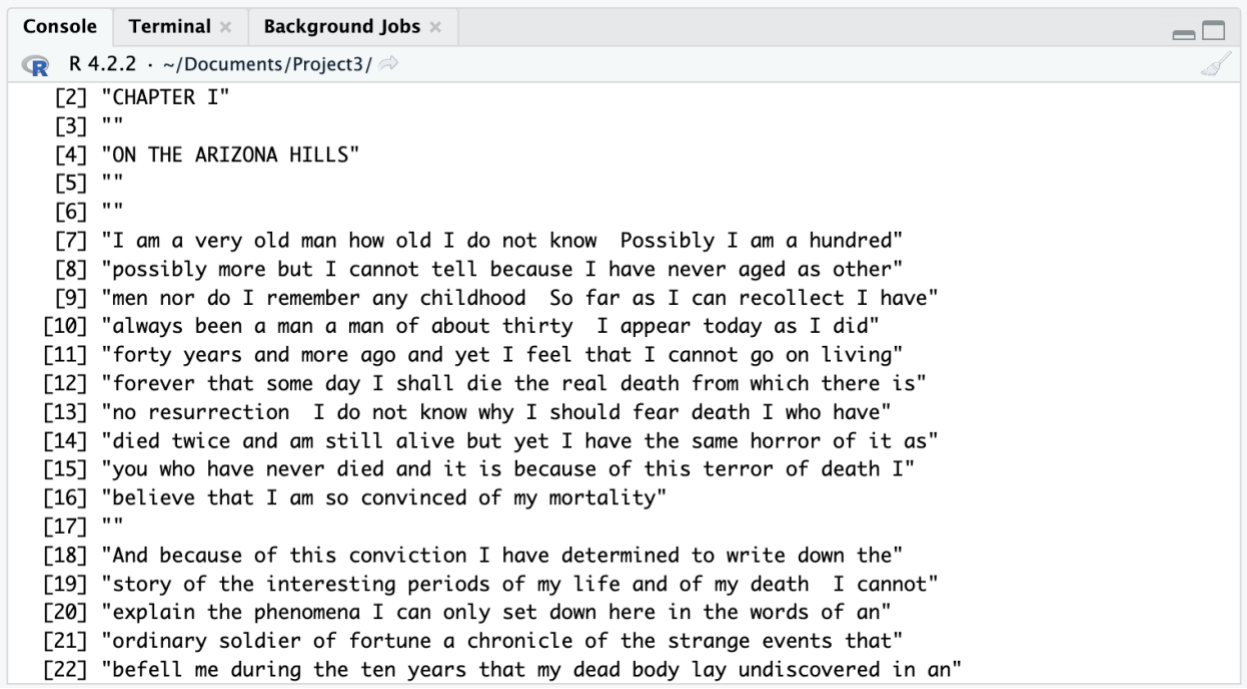
Remove numers and punctuations from the text file.

PoMarsText_noNum=removeNumbers(text)

PoMarsText_noNum

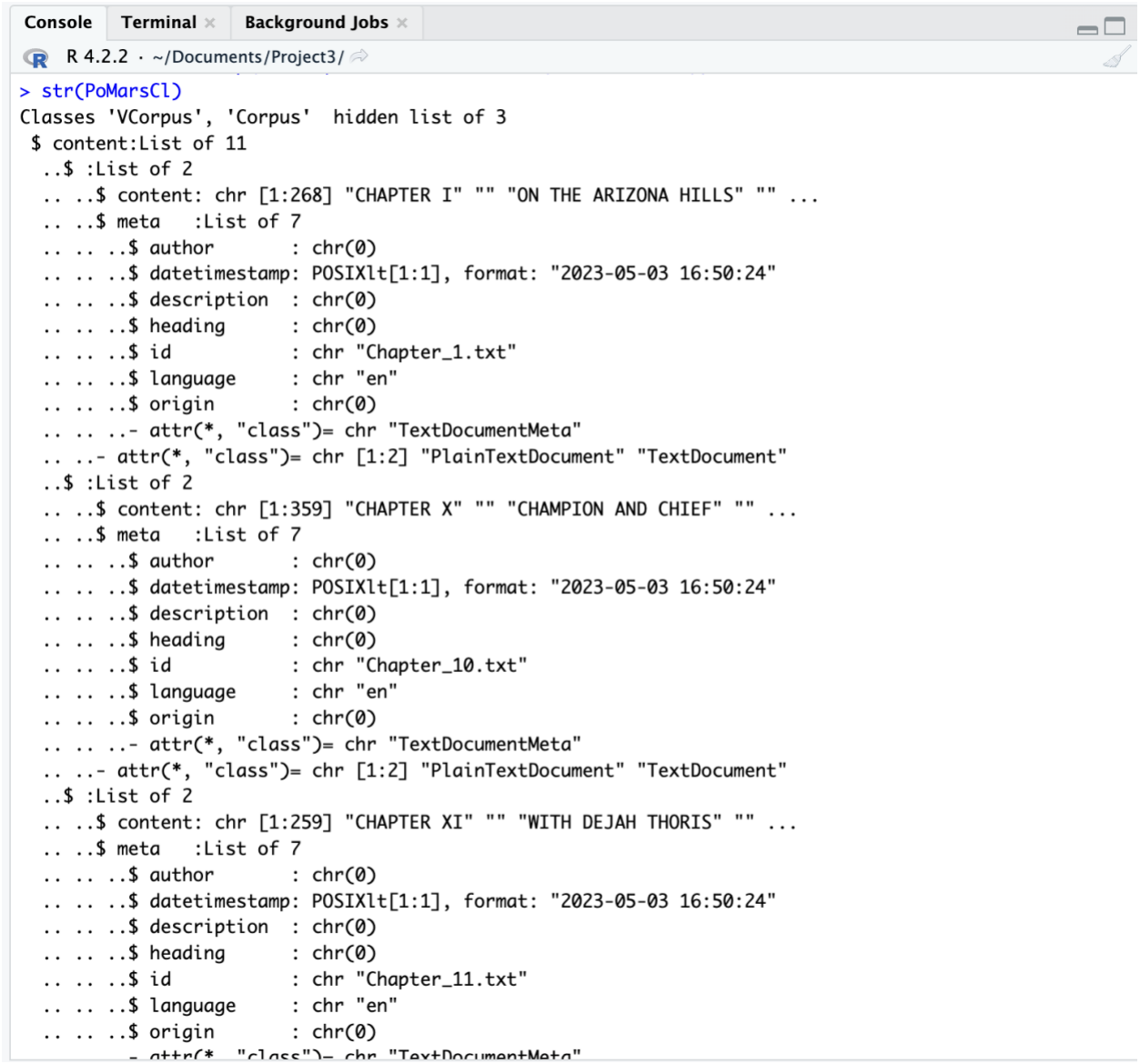
PoMarsText_noNumPunc=removePunctuation (PoMarsText_noNum)

PoMarsText_noNumPunc

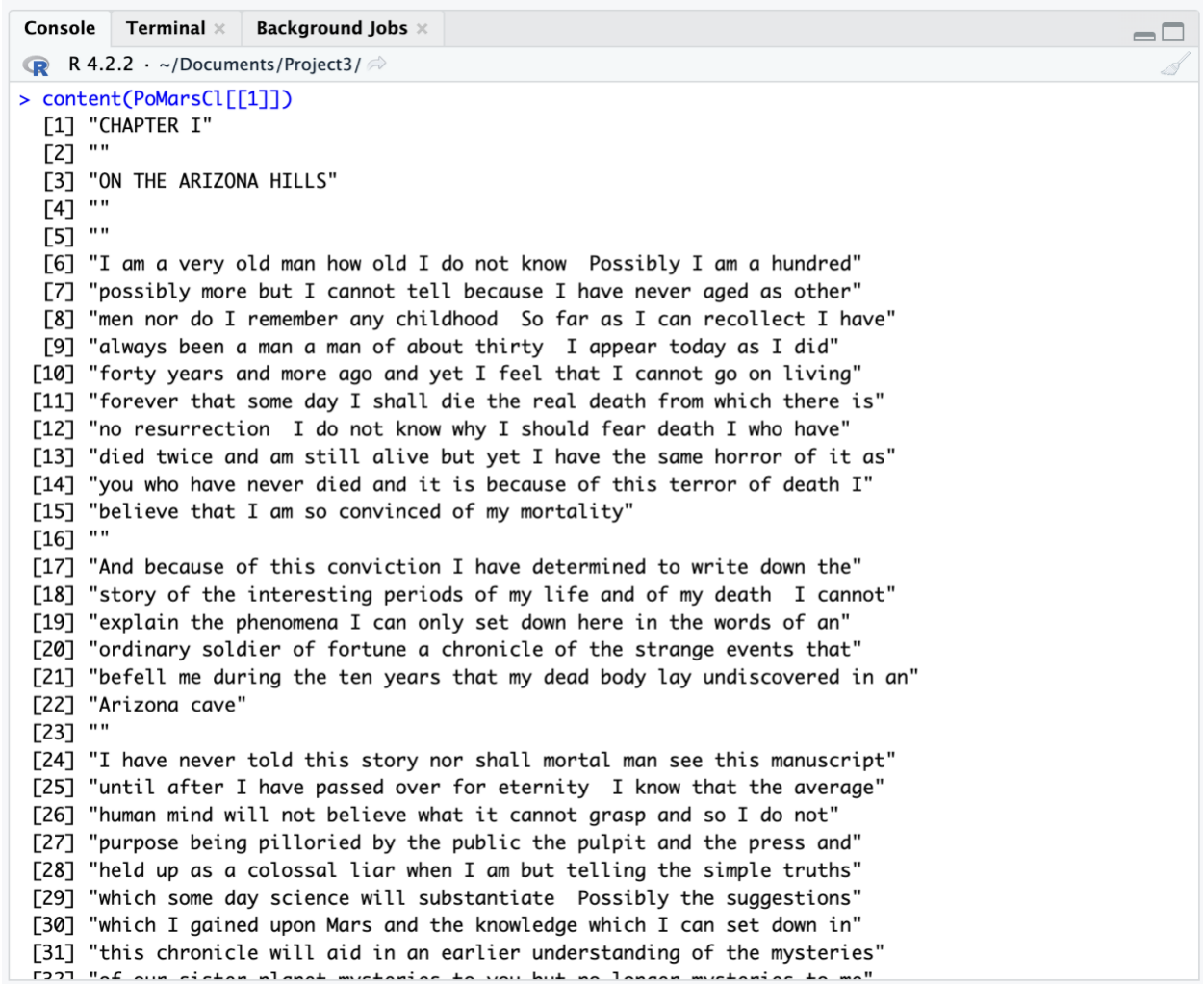


Remove Num and Punc from a vcorpus file.

removeNumPunc<-function(x) gsub ("[^[: alpha:][:space:]]*","",x)
PoMarsCl=tm::tm_map (PoMars, content_transformer(removeNumPunc))
str (PoMarsCl)



content (PoMarsCl [[1]])



inspect (PoMars)

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Project3/
> inspect(PoMars)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13971

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 19122

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13161

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9010

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14201

[[6]]
<<PlainTextDocument>>
```

inspect (PoMarsCl)

```
114:1  (Top Level)                                  R Script
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Project3/
> inspect(PoMarsCl)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13720

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18678

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 12822

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8854

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13932

[[6]]
<<PlainTextDocument>>
```

Converting text to lowercase characters.

PoMarsCl_Low=tm::tm_map (PoMarsCl, content_transformer(tolower))
content (PoMarsCl_Low [[1]])

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/Documents/Project3/
> PoMarsCl_Low=tm::tm_map(PoMarsCl,content_transformer(tolower))
> content(PoMarsCl_Low[[1]])
  [1] "chapter i"
  [2] ""
  [3] "on the arizona hills"
  [4] ""
  [5] ""
  [6] "i am a very old man how old i do not know  possibly i am a hundred"
  [7] "possibly more but i cannot tell because i have never aged as other"
  [8] "men nor do i remember any childhood  so far as i can recollect i have"
  [9] "always been a man a man of about thirty  i appear today as i did"
 [10] "forty years and more ago and yet i feel that i cannot go on living"
 [11] "forever that some day i shall die the real death from which there is"
 [12] "no resurrection  i do not know why i should fear death i who have"
 [13] "died twice and am still alive but yet i have the same horror of it as"
 [14] "you who have never died and it is because of this terror of death i"
 [15] "believe that i am so convinced of my mortality"
 [16] ""
 [17] "and because of this conviction i have determined to write down the"
 [18] "story of the interesting periods of my life and of my death  i cannot"
 [19] "explain the phenomena i can only set down here in the words of an"
 [20] "ordinary soldier of fortune a chronicle of the strange events that"
 [21] "befell me during the ten years that my dead body lay undiscovered in an"
 [22] "arizona cave"
 [23] ""
 [24] "i have never told this story nor shall mortal man see this manuscript"
 [25] "until after i have passed over for eternity  i know that the average"
 [26] "human mind will not believe what it cannot grasp and so i do not"
 [27] "purpose being pilloried by the public the pulpit and the press and"
 [28] "held up as a colossal liar when i am but telling the simple truths"
 [29] "which some day science will substantiate  possibly the suggestions"
 [30] "which i gained upon mars and the knowledge which i can set down in"
```

#Term Document Matrix of entire file
PoMarsTDM=TermDocumentMatrix (PoMarsCl_Low)
PoMarsTDM

```
> PoMarsTDM=TermDocumentMatrix(PoMarsCl_Low)
> PoMarsTDM
<<TermDocumentMatrix (terms: 3870, documents: 11)>>
Non-/sparse entries: 8219/34351
Sparsity           : 81%
Maximal term length: 17
Weighting          : term frequency (tf)
>
```

inspect (PoMarsTDM)

```
> inspect(PoMarsTDM)
<<TermDocumentMatrix (terms: 3870, documents: 11)>>
Non-/sparse entries: 8219/34351
Sparsity           : 81%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
       Docs
Terms   Chapter_1.txt Chapter_10.txt Chapter_11.txt Chapter_2.txt Chapter_3.txt Chapter_4.txt
  and              93            117             89            59            94            70
  but              13             20             14            13            17            21
  for              20             34             10            17            18            10
  had              27             41             29            20            19            22
  that             50             50             38            20            26            19
  the             190            193            122           137           166           148
  upon             12             19             13            13            20            11
  was              38             54             19            33            37            26
  which            20             16             13            13            34            17
  with             25             26             23            10            20            20
       Docs
Terms   Chapter_5.txt Chapter_6.txt Chapter_7.txt Chapter_8.txt
  and              56            57             67            84
  but              14            10              9             7
  for              13            15             18            15
  had              14            29             16            24
  that             11            11             12            11
  the              96           117            166           184
  upon              8            13              7            22
  was              22            18             18            32
  which             9            12             24            17
  with              8            24             12            11
>
```

str (PoMarsTDM)

```
> str(PoMarsTDM)
List of 6
 $ i       : int [1:8219] 2 3 18 25 30 32 35 57 59 66 ...
 $ j       : int [1:8219] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:8219] 1 7 1 1 4 1 1 1 1 3 ...
 $ nrow    : int 3870
 $ ncol    : int 11
 $ dimnames:List of 2
  ..$ Terms: chr [1:3870] "ability" "able" "about" "above" ...
  ..$ Docs : chr [1:11] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_2.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

Document Term Matrix

PoMarsDTM=DocumentTermMatrix (PoMarsCl_Low)

PoMarsDTM

```
> PoMarsDTM=DocumentTermMatrix(PoMarsCl_Low)
> PoMarsDTM
<<DocumentTermMatrix (documents: 11, terms: 3870)>>
Non-/sparse entries: 8219/34351
Sparsity           : 81%
Maximal term length: 17
Weighting          : term frequency (tf)
>
```

inspect (PoMarsDTM)

```
> inspect(PoMarsDTM)
<<DocumentTermMatrix (documents: 11, terms: 3870)>>
Non-/sparse entries: 8219/34351
Sparsity           : 81%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs             and but for had that the upon was which with
  Chapter_1.txt   93  13  20  27   50 190   12  38    20   25
  Chapter_10.txt 117  20  34  41   50 193   19  54    16   26
  Chapter_11.txt  89  14  10  29   38 122   13  19    13   23
  Chapter_2.txt   59  13  17  20   20 137   13  33    13   10
  Chapter_3.txt   94  17  18  19   26 166   20  37    34   20
  Chapter_4.txt   70  21  10  22   19 148   11  26    17   20
  Chapter_5.txt   56  14  13  14   11  96    8  22     9    8
  Chapter_6.txt   57  10  15  29   11 117   13  18    12   24
  Chapter_7.txt   67   9  18  16   12 166    7  18    24   12
  Chapter_8.txt   84   7  15  24   11 184   22  32    17   11
>
```

str (PoMarsDTM)

```
> str(PoMarsDTM)
List of 6
 $ i       : int [1:8219] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:8219] 2 3 18 25 30 32 35 57 59 66 ...
 $ v       : num [1:8219] 1 7 1 1 4 1 1 1 1 3 ...
 $ nrow    : int 11
 $ ncol    : int 3870
 $ dimnames:List of 2
  ..$ Docs : chr [1:11] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_2.txt" ...
  ..$ Terms: chr [1:3870] "ability" "able" "about" "above" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
>
```

#Matrix
PoMarsMatrixTDM=as.matrix(PoMarsTDM)

Console | Terminal | Background Jobs

R 4.2.2 · ~/Documents/Project3/

```
> PoMarsMatrixTDM
                Docs
Terms            Chapter_1.txt Chapter_10.txt Chapter_11.txt Chapter_2.txt Chapter_3.txt
  ability                    0              2              0             0             0
  able                       1              0              0             0             0
  about                      7              0              2             0             8
  above                      0              3              0             0             3
  abreast                    0              0              0             0             0
  abruptly                   0              0              0             0             0
  absence                    0              0              2             0             0
  absolutely                 0              0              0             0             1
  abundantly                 0              0              0             0             0
  accompanied                0              0              1             0             0
  accompany                  0              0              0             0             0
  accomplished               0              0              0             0             0
  accomplishing              0              0              0             0             0
  accord                     0              0              0             0             0
  accordance                 0              1              0             0             0
  accorded                   0              2              0             0             0
  according                  0              1              0             0             0
  account                    1              0              0             0             0
  accounted                  0              1              0             0             0
  accounting                 0              1              0             0             0
  accounts                   0              0              0             0             0
  accouterments              0              0              1             0             1
  accoutrements              0              1              0             0             0
  accuracy                   0              0              0             0             0
  accurately                 1              0              0             0             0
  accustomed                 0              0              0             0             1
  acid                       0              0              0             0             0
  acquainted                 0              0              0             0             0
  acquired                   0              1              0             0             0
  across                     4              2              1             1             0
  act                        0              4              0             0             0
  acted                      1              0              0             1             0
  action                     0              0              0             0             1
  actions                    0              0              0             0             1
  acts                       1              2              0             0             0
  actual                     0              0              0             0             1
  adapted                    0              0              0             0             0
  add                        0              1              0             0             1
  added                      0              1              0             0             0
  adding                     0              0              0             0             0
  addition                   0              0              0             0             1
  additional                 0              0              0             0             0
  address                    0              1              0             0             0
```

PoMarsMatrixDTM=as.matrix(PoMarsDTM)

```
> PoMarsMatrixDTM
               Terms
Docs            ability able about above abreast abruptly absence absolutely abundantly
               Terms
Docs            accompanied accompany accomplished accomplishing accord accordance accorded
               Terms
Docs            according account accounted accounting accounts accouterments accoutrements
               Terms
Docs            accuracy accurately accustomed acid acquainted acquired across act acted action
               Terms
Docs            actions acts actual adapted add added adding addition additional address
               Terms
Docs            addressed addressing adjoining administration ado adoration adult adults
               Terms
Docs            advance advanced advancement advancing advantage advent adventure adventures
               Terms
Docs            adversary advising affect affection affections african after afternoon
               Terms
Docs            afterward afterwards again against age aged ageold ages agile agility ago
               Terms
Docs            agonies agreed ahead aid aim aimed air airtight aisle albinos alien alighted
               Terms
Docs            alighting alike alive all alliance allowed allowing alloy almost alone along
               Terms
Docs            aloud already also alternative although altogether aluminum always america
               Terms
Docs            amity ammunition among amongst amounted amusement anaesthesis ancestor
               Terms
Docs            ancestors ancient and andi angrily animal animals annihilate annihilation
               Terms
Docs            anomalies another answered answering antagonist antecedents antelope antennae
               Terms
Docs            antennaelike antics antiquity anxious anxiously any anything anyway anywhere
```

Remove Stop words.

myStopWords= (tm::stopwords (kind = 'en'))
PoMarsNoStopWords=tm::tm_map (PoMarsCl_Low,tm::removeWords,myStopWords)
str (PoMarsNoStopWords)

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Project3/
> myStopWords=(tm::stopwords(kind = 'en'))
> PoMarsNoStopWords=tm::tm_map(PoMarsCl_Low,tm::removeWords,myStopWords)
> str(PoMarsNoStopWords)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 11
  ..$ :List of 2
  .. ..$ content: chr [1:268] "chapter " "" "  arizona hills" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-03 16:50:24"
  .. .. ..$ 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 [1:359] "chapter x" "" "champion  chief" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-03 16:50:24"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
```

tm::inspect (PoMarsNoStopWords [[1]])

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Project3/
> tm::inspect(PoMarsNoStopWords[[1]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9820

chapter

  arizona hills


    old man  old    know  possibly    hundred
possibly     tell    never aged
men    remember  childhood   far   can recollect
always   man  man   thirty   appear today
forty years   ago  yet  feel    go  living
forever   day  shall die  real death
 resurrection     know    fear death
died twice   still alive  yet     horror
   never died       terror  death
believe     convinced   mortality

    conviction   determined  write
story   interesting periods   life    death
explain  phenomena  can  set     words
```

TDM of PoMars without stop words.

PoMarsNoStopTDM=TermDocumentMatrix (PoMarsNoStopWords)
PoMarsNoStopTDM

```
> PoMarsNoStopTDM=TermDocumentMatrix(PoMarsNoStopWords)
> PoMarsNoStopTDM
<<TermDocumentMatrix (terms: 3772, documents: 11)>>
Non-/sparse entries: 7419/34073
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
>
```

inspect (PoMarsNoStopTDM)

```
> inspect(PoMarsNoStopTDM)
<<TermDocumentMatrix (terms: 3772, documents: 11)>>
Non-/sparse entries: 7419/34073
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
          Docs
Terms      Chapter_1.txt Chapter_10.txt Chapter_11.txt Chapter_2.txt Chapter_3.txt Chapter_4.txt
  feet                 5              5              0             1            15             9
  first                3              6              3             5             5             6
  little               4              4              4             2            11             3
  mars                 1              4              2             1            11             6
  martian              0             13              3             0             6             9
  martians             0              1              4             0             5             8
  one                  3             19              1             2             7             5
  sola                 0              6             11             0             0             2
  toward               3              9              1             3            10             6
  upon                12             19             13            13            20            11
          Docs
Terms      Chapter_5.txt Chapter_6.txt Chapter_7.txt Chapter_8.txt
  feet                 2             4              2             1
  first                1             3              3             5
  little               2             0             10             4
  mars                 9             2              4             1
  martian              5             2             10             7
  martians             1             7             11             3
  one                  6             5              9             8
  sola                 5             4              7             3
  toward               3             3              2             5
  upon                 8            13              7            22
```

str (PoMarsNoStopTDM)

```
> str(PoMarsNoStopTDM)
List of 6
 $ i       : int [1:7419] 2 16 23 28 30 33 55 57 64 68 ...
 $ j       : int [1:7419] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:7419] 1 1 1 4 1 1 1 1 1 1 ...
 $ nrow    : int 3772
 $ ncol    : int 11
 $ dimnames:List of 2
  ..$ Terms: chr [1:3772] "ability" "able" "abreast" "abruptly" ...
  ..$ Docs : chr [1:11] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_2.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

Document Term Matrix of PoMars without stop words.

PoMarsNoStopDTM=DocumentTermMatrix (PoMarsNoStopWords)
PoMarsNoStopDTM

```
> PoMarsNoStopDTM=DocumentTermMatrix(PoMarsNoStopWords)
> PoMarsNoStopDTM
<<DocumentTermMatrix (documents: 11, terms: 3772)>>
Non-/sparse entries: 7419/34073
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
>
```

inspect (PoMarsNoStopDTM)

```
> inspect(PoMarsNoStopDTM)
<<DocumentTermMatrix (documents: 11, terms: 3772)>>
Non-/sparse entries: 7419/34073
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
                Terms
Docs             feet first little mars martian martians one sola toward upon
  Chapter_1.txt     5     3      4    1       0        0   3    0      3   12
  Chapter_10.txt    5     6      4    4      13        1  19    6      9   19
  Chapter_11.txt    0     3      4    2       3        4   1   11      1   13
  Chapter_2.txt     1     5      2    1       0        0   2    0      3   13
  Chapter_3.txt    15     5     11   11       6        5   7    0     10   20
  Chapter_4.txt     9     6      3    6       9        8   5    2      6   11
  Chapter_5.txt     2     1      2    9       5        1   6    5      3    8
  Chapter_6.txt     4     3      0    2       2        7   5    4      3   13
  Chapter_7.txt     2     3     10    4      10       11   9    7      2    7
  Chapter_8.txt     1     5      4    1       7        3   8    3      5   22
```

str (PoMarsNoStopDTM)

```
> str(PoMarsNoStopDTM)
List of 6
 $ i       : int [1:7419] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:7419] 2 16 23 28 30 33 55 57 64 68 ...
 $ v       : num [1:7419] 1 1 1 4 1 1 1 1 1 1 ...
 $ nrow    : int 11
 $ ncol    : int 3772
 $ dimnames:List of 2
  ..$ Docs : chr [1:11] "Chapter_1.txt" "Chapter_10.txt" "Chapter_11.txt" "Chapter_2.txt" ...
  ..$ Terms: chr [1:3772] "ability" "able" "abreast" "abruptly" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

Word Frequency:
wordfreq=colSums (as. matrix (PoMarsNoStopDTM))

```
> wordfreq=colSums(as.matrix(PoMarsNoStopDTM))
> wordfreq
       ability           able        abreast       abruptly         absence      absolutely
             5              3              1              1              3              1
    abundantly     accompanied       accompany    accomplished   accomplishing         accord
             1              3              1              1              1              2
    accordance        accorded       according        account       accounted      accounting
             1              2              1              2              1              1
      accounts   accouterments   accoutrements       accuracy      accurately      accustomed
             1              3              1              2              1              2
          acid      acquainted       acquired         across            act          acted
             1              2              1             13              5              2
        action        actions           acts         actual        adapted            add
             1              4              3              2              1              2
         added         adding       addition     additional        address      addressed
             3              1              1              1              1              4
    addressing       adjoining  administration            ado       adoration          adult
             1              2              1              1              1              4
        adults        advance       advanced    advancement       advancing      advantage
             3              2              7              1              2              4
        advent       adventure     adventures       adversary        advising         affect
             3              3              1              2              1              1
     affection      affections         african      afternoon       afterward      afterwards
             5              1              1              3              6              1
           age           aged         ageold           ages          agile         agility
             5              1              2             12              1              2
           ago        agonies         agreed          ahead            aid            aim
             3              2              1              4              5              3
         aimed            air       airtight          aisle        albinos          alien
```

freqterms=tm::findFreqTerms (PoMarsNoStopTDM)

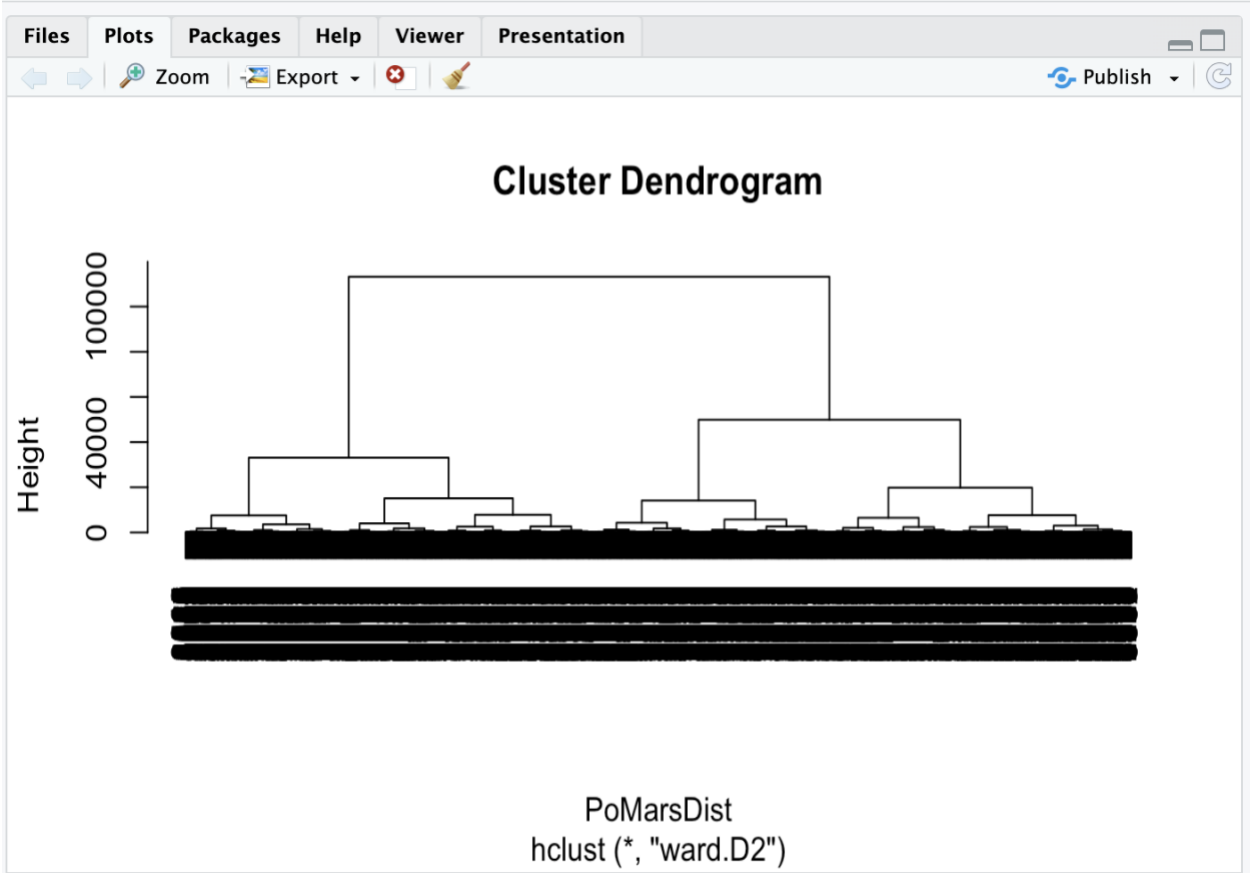
```
> freqterms=tm::findFreqTerms(PoMarsNoStopTDM)
> freqterms
   [1] "ability"         "able"            "abreast"         "abruptly"        "absence"
   [6] "absolutely"      "abundantly"      "accompanied"     "accompany"       "accomplished"
  [11] "accomplishing"   "accord"          "accordance"      "accorded"        "according"
  [16] "account"         "accounted"       "accounting"      "accounts"        "accouterments"
  [21] "accoutrements"   "accuracy"        "accurately"      "accustomed"      "acid"
  [26] "acquainted"      "acquired"        "across"          "act"             "acted"
  [31] "action"          "actions"         "acts"            "actual"          "adapted"
  [36] "add"             "added"           "adding"          "addition"        "additional"
  [41] "address"         "addressed"       "addressing"      "adjoining"       "administration"
  [46] "ado"             "adoration"       "adult"           "adults"          "advance"
  [51] "advanced"        "advancement"     "advancing"       "advantage"       "advent"
  [56] "adventure"       "adventures"      "adversary"       "advising"        "affect"
  [61] "affection"       "affections"      "african"         "afternoon"       "afterward"
  [66] "afterwards"      "age"             "aged"            "ageold"          "ages"
  [71] "agile"           "agility"         "ago"             "agonies"         "agreed"
  [76] "ahead"           "aid"             "aim"             "aimed"           "air"
  [81] "airtight"        "aisle"           "albinos"         "alien"           "alighted"
  [86] "alighting"       "alike"           "alive"           "alliance"        "allowed"
  [91] "allowing"        "alloy"           "almost"          "alone"           "along"
  [96] "aloud"           "already"         "also"            "alternative"     "although"
 [101] "altogether"      "aluminum"        "always"          "america"         "amity"
 [106] "ammunition"      "among"           "amongst"         "amounted"        "amusement"
 [111] "anaesthesis"     "ancestor"        "ancestors"       "ancient"         "andi"
 [116] "angrily"         "animal"          "animals"         "annihilate"      "annihilation"
 [121] "anomalies"       "another"         "answered"        "answering"       "antagonist"
 [126] "antecedents"     "antelope"        "antennae"        "antennaelike"    "antics"
 [131] "antiquity"       "anxious"         "anxiously"       "anything"        "anyway"
```

PoMarsTF=tm::termFreq (PoMarsNoStopWords [[1]])

```
> PoMarsTF=tm::termFreq(PoMarsNoStopWords[[1]])
> PoMarsTF
         able       account     accurately        across         acted          acts
            1             1             1             4             1             1
       advent    adventures     afternoon          aged           ago        agreed
            1             1             1             1             1             1
        ahead           aid           aim         alive        almost         alone
            2             1             1             2             1             1
      already   alternative        always         among       animals       another
            1             1             3             1             2             1
     antelope      anything       apaches    apartments        appear  apprehension
            1             1             3             1             1             1
      arizona           arm         armed        arming          army         arose
            5             1             1             1             3             1
       around        arrows        assure      attacked       attempt     attention
            3             3             1             1             1             1
   attraction    attributed       average         await          back      backward
            1             1             1             1             3             1
        balls        bathed         beast     beautiful        became        become
            1             1             1             1             1             2
       befell       believe          belt         belts          best      bestowed
            2             3             1             1             1             1
       better       bidding          body         borne        bottom          bows
            3             1             5             1             1             1
       braves         brief        bright         brisk     bristling         broad
            1             1             1             1             1             1
        broke       brought        burros          came          camp           can
            1             1             1             3             3             3
      canteen        canter       captain      captains       capture       carbine
```

Dendogram:
PoMarsDF1<-as.data. frame (PoMarsNoStopTDM [[1]])
PoMarsDist<-dist (PoMarsDF1)
PoMarsDG<-hclust (PoMarsDist, method="ward.D2")
str (PoMarsDG)
plot (PoMarsDG)

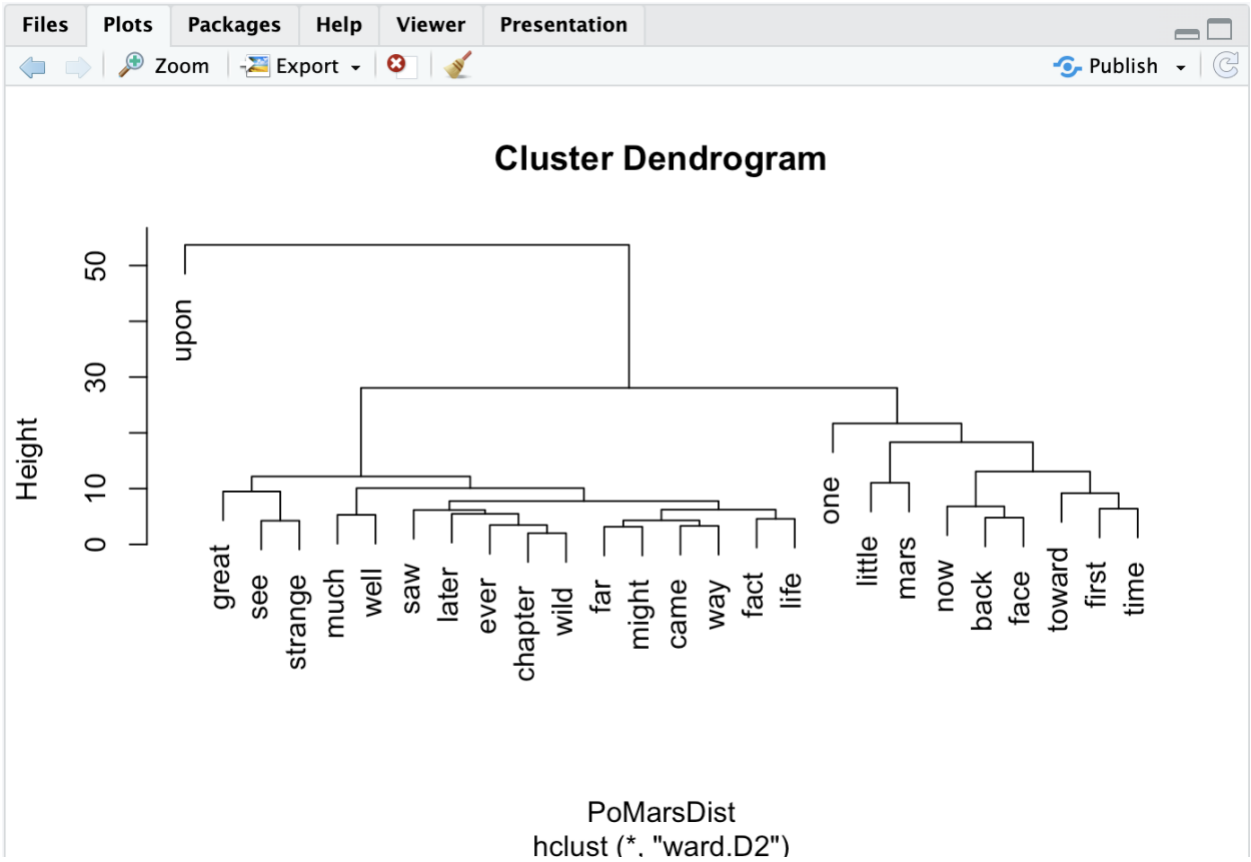
```
> PoMarsDist<-dist(PoMarsDF1)
> PoMarsDG<-hclust(PoMarsDist,method="ward.D2")
> str(PoMarsDG)
List of 7
 $ merge      : int [1:7418, 1:2] -1 -5655 -2 -4 -1817 -2524 -3864 -4557 -5659 -6297 ...
 $ height     : num [1:7418] 0 0 0 0 0 0 0 0 0 0 ...
 $ order      : int [1:7419] 3885 843 3123 2542 30 2541 3886 3124 844 1843 ...
 $ labels     : NULL
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = PoMarsDist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> plot(PoMarsDG)
>
```



We perform hierarchical clustering using Ward's method on a filtered term-document matrix. The matrix is filtered to remove words with high sparsity, i.e., words that appear in a large percentage of documents. A sparsity threshold of 0.85 is set, and words with sparsity above this threshold are removed. The resulting matrix is converted to a data frame and used to calculate the distance matrix. Hierarchical clustering is performed on the distance matrix using the "ward.D2" method. Finally, the dendrogram is plotted. This process allows us to explore the relationships between the documents based on the similarity of the selected words they contain. The resulting dendrogram shows the clusters of documents that are most like each other.

```
#dendogram by removing words with high sparsity
PoMarsNoStopTDM_Mat=as.matrix(PoMarsNoStopTDM)
# Set sparsity threshold (e.g., 0.85 means words that appear in greater than 85% of the document
sparsity_threshold <- 0.85

# Calculate sparsity for each word
sparsity <- rowSums(PoMarsNoStopTDM_Mat > 0) / ncol(PoMarsNoStopTDM_Mat)

# Select words with sparsity below the threshold
selected_words <- which(sparsity > sparsity_threshold)

# Create a new TDM with the selected words
PoMarsFilteredTDM <- PoMarsNoStopTDM_Mat[selected_words,]

# Convert the new TDM to a data frame
PoMarsDF1 <- as.data.frame(PoMarsFilteredTDM)

# Calculate the distance matrix
PoMarsDist <- dist(PoMarsDF1)

# Perform hierarchical clustering using Ward's method
PoMarsDG <- hclust(PoMarsDist, method = "ward.D2")

# Check the structure of the dendrogram
str(PoMarsDG)

# Plot the dendrogram
plot(PoMarsDG)
```

```
> PoMarsNoStopTDM_Mat=as.matrix(PoMarsNoStopTDM)
> sparsity_threshold <- 0.85
> sparsity <- rowSums(PoMarsNoStopTDM_Mat > 0) / ncol(PoMarsNoStopTDM_Mat)
> selected_words <- which(sparsity > sparsity_threshold)
> PoMarsFilteredTDM <- PoMarsNoStopTDM_Mat[selected_words,]
> # Convert the new TDM to a data frame
> PoMarsDF1 <- as.data.frame(PoMarsFilteredTDM)
>
> # Calculate the distance matrix
> PoMarsDist <- dist(PoMarsDF1)
>
> # Perform hierarchical clustering using Ward's method
> PoMarsDG <- hclust(PoMarsDist, method = "ward.D2")
>
> # Check the structure of the dendrogram
> str(PoMarsDG)
List of 7
 $ merge      : int [1:25, 1:2] -3 -7 -2 -4 -19 2 -6 -1 -15 -10 ...
 $ height     : num [1:25] 2 3.16 3.32 3.46 4.24 ...
 $ order      : int [1:26] 23 9 19 20 15 25 18 10 4 3 ...
 $ labels     : chr [1:26] "back" "came" "chapter" "ever" ...
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = PoMarsDist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
>
> # Plot the dendrogram
> plot(PoMarsDG)
```



WORD CLOUD:

We create a word cloud from the term-frequency matrix of the text corpus. The names () function is used to extract the unique words in the corpus, and the wordcloud () function is used to generate the word cloud. The function takes two arguments: the words to be plotted and their frequencies. The colors parameter specifies the color scheme for the word cloud, which is set to a nine-color sequential scheme from the RColorBrewer package. The resulting word cloud visually represents the frequency of the most common words in the text corpus, with larger font sizes indicating higher frequencies. This visualization provides an immediate overview of the most common words in the text, allowing for quick identification of key topics or themes.

```
> words=names(PoMarsTF)
> pal<-brewer.pal(9,"Spectral")
> PoMarsWC=wordcloud(words,PoMarsTF,colors = pal)
```

Quanteda

PoMarstext=PoMarsCl [[1]]
PoMarstext$content [1:10]

Quanteda is an R package for text analysis and natural language processing. It provides a suite of tools for corpus management, text preprocessing, feature selection, and document classification.

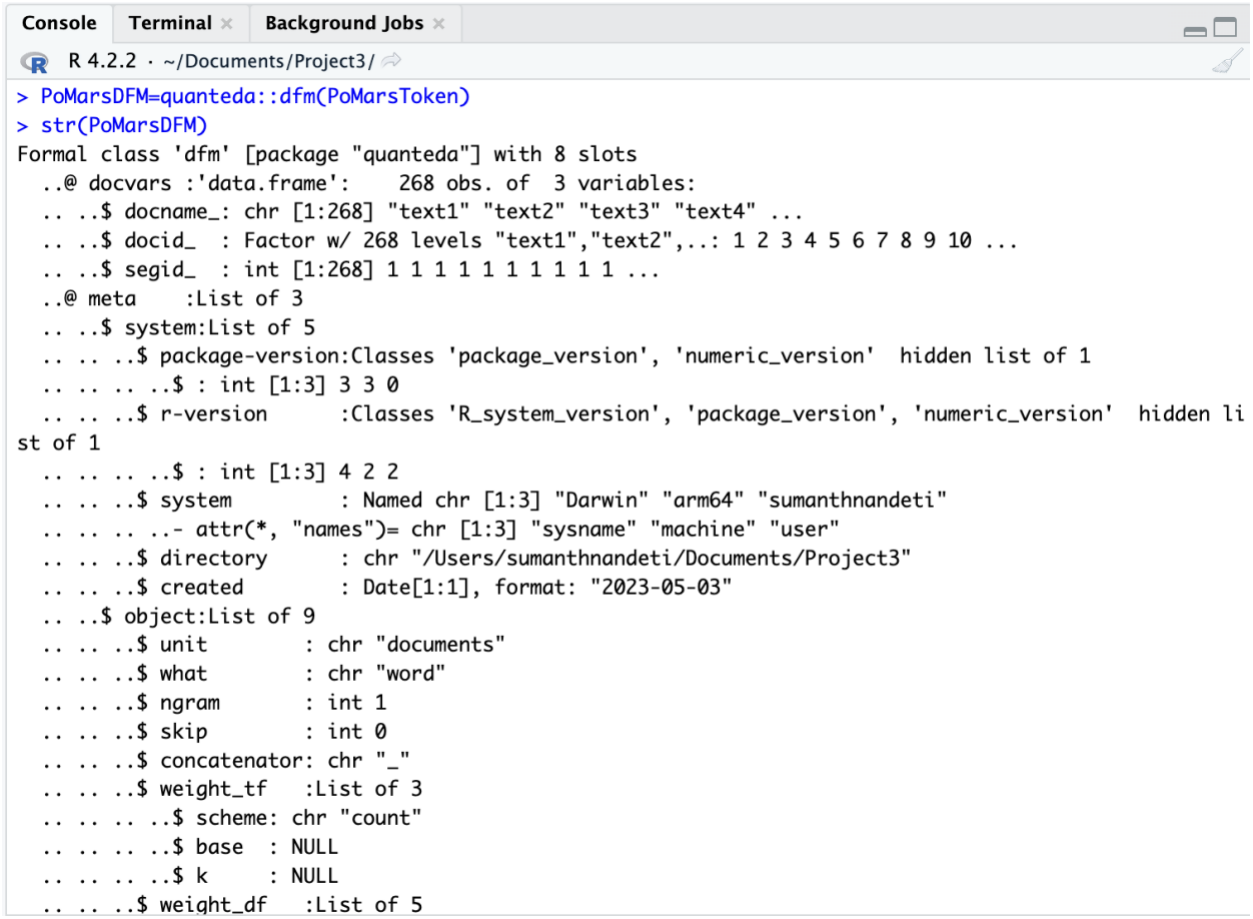
```
> PoMarstext=PoMarsCl[[1]]
> PoMarstext$content[1:10]
 [1] "CHAPTER I"
 [2] ""
 [3] "ON THE ARIZONA HILLS"
 [4] ""
 [5] ""
 [6] "I am a very old man how old I do not know  Possibly I am a hundred"
 [7] "possibly more but I cannot tell because I have never aged as other"
 [8] "men nor do I remember any childhood  So far as I can recollect I have"
 [9] "always been a man a man of about thirty  I appear today as I did"
[10] "forty years and more ago and yet I feel that I cannot go on living"
>
```

Tokens
PoMarsToken<-quanteda::tokens (PoMarstext$content)
PoMarsToken

```
> PoMarsToken<-quanteda::tokens(PoMarstext$content)
> PoMarsToken
Tokens consisting of 268 documents.
text1 :
[1] "CHAPTER" "I"

text2 :
character(0)

text3 :
[1] "ON"      "THE"     "ARIZONA" "HILLS"

text4 :
character(0)

text5 :
character(0)

text6 :
 [1] "I"    "am"   "a"    "very" "old"  "man"  "how"  "old"  "I"    "do"   "not"  "know"
[ ... and 5 more ]

[ reached max_ndoc ... 262 more documents ]
>
```

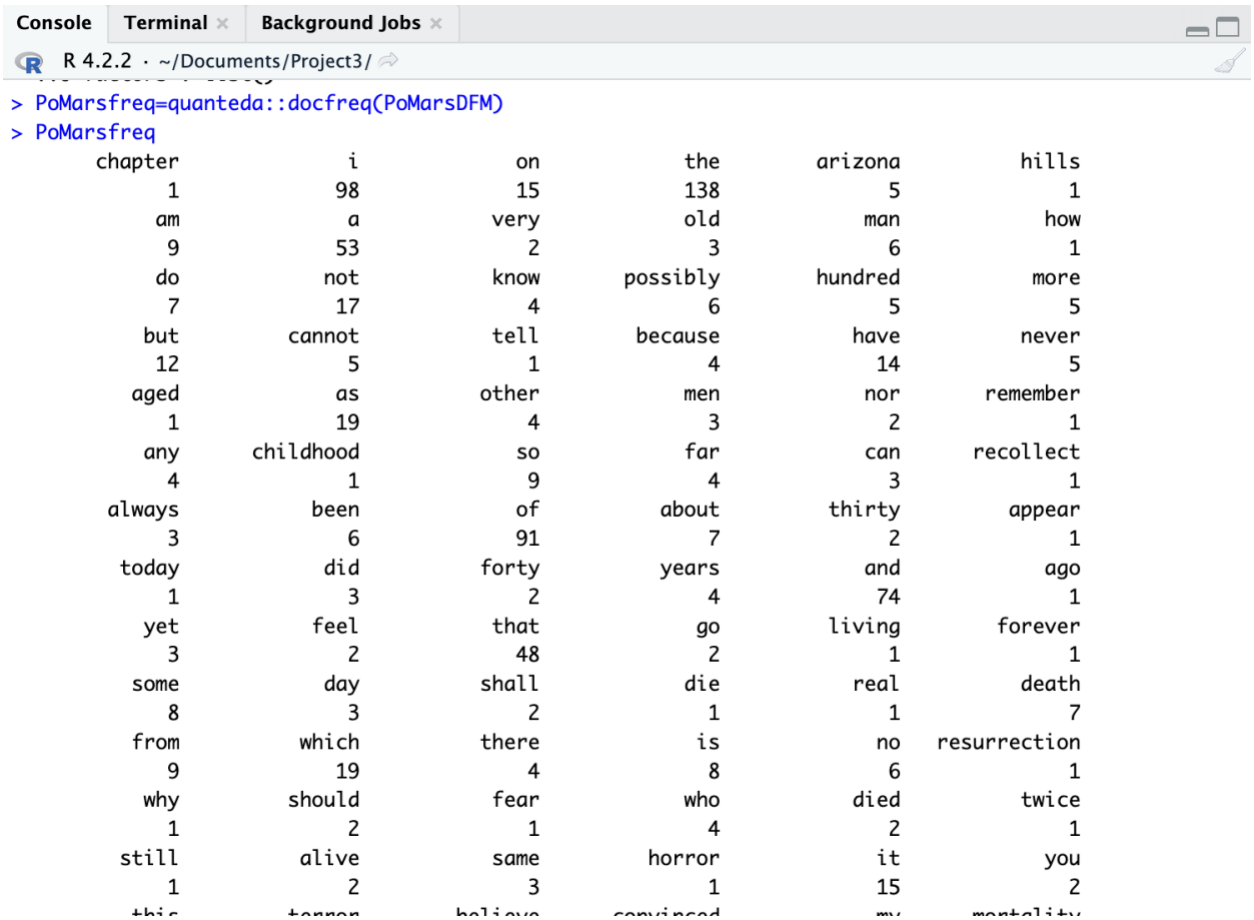
removing lines with no characters

PoMarsDFM=quanteda::dfm (PoMarsToken)
str (PoMarsDFM)



calculate freq of words using quanteda.

PoMarsfreq=quanteda::docfreq (PoMarsDFM)
PoMarsfreq


Console
Terminal
Background Jobs
R 4.2.2 · ~/Documents/Project3/
> PoMarsfreq=quanteda::docfreq(PoMarsDFM)
> PoMarsfreq
chapter i on the arizona hills
1 98 15 138 5 1
am a very old man how
9 53 2 3 6 1
do not know possibly hundred more
7 17 4 6 5 5
but cannot tell because have never
12 5 1 4 14 5
aged as other men nor remember
1 19 4 3 2 1
any childhood so far can recollect
4 1 9 4 3 1
always been of about thirty appear
3 6 91 7 2 1
today did forty years and ago
1 3 2 4 74 1
yet feel that go living forever
3 2 48 2 1 1
some day shall die real death
8 3 2 1 1 7
from which there is no resurrection
9 19 4 8 6 1
why should fear who died twice
1 2 1 4 2 1
still alive same horror it you
1 2 3 1 15 2
> PoMarsWeights=quanteda::dfm_weight(PoMarsDFM)
> PoMarsWeights
Document-feature matrix of: 268 documents, 888 features (98.97% sparse) and 0 docvars.
features
docs chapter i on the arizona hills am a very old
text1 1 1 0 0 0 0 0 0 0 0 0
text2 0 0 0 0 0 0 0 0 0 0 0
text3 0 0 1 1 1 1 0 0 0 0 0
text4 0 0 0 0 0 0 0 0 0 0 0
text5 0 0 0 0 0 0 0 0 0 0 0
text6 0 3 0 0 0 0 2 2 1 2
[reached max_ndoc ... 262 more documents, reached max_nfeat ... 878 more features]

```
> str(PoMarsWeights)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    268 obs. of  3 variables:
  .. ..$ docname_: chr [1:268] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 268 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:268] 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 0
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden li
st of 1
  .. .. .. ..$ : int [1:3] 4 2 2
  .. .. ..$ system         : Named chr [1:3] "Darwin" "arm64" "sumanthnandeti"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/sumanthnandeti/Documents/Project3"
  .. .. ..$ created        : Date[1:1], format: "2023-05-03"
  .. ..$ 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"
```

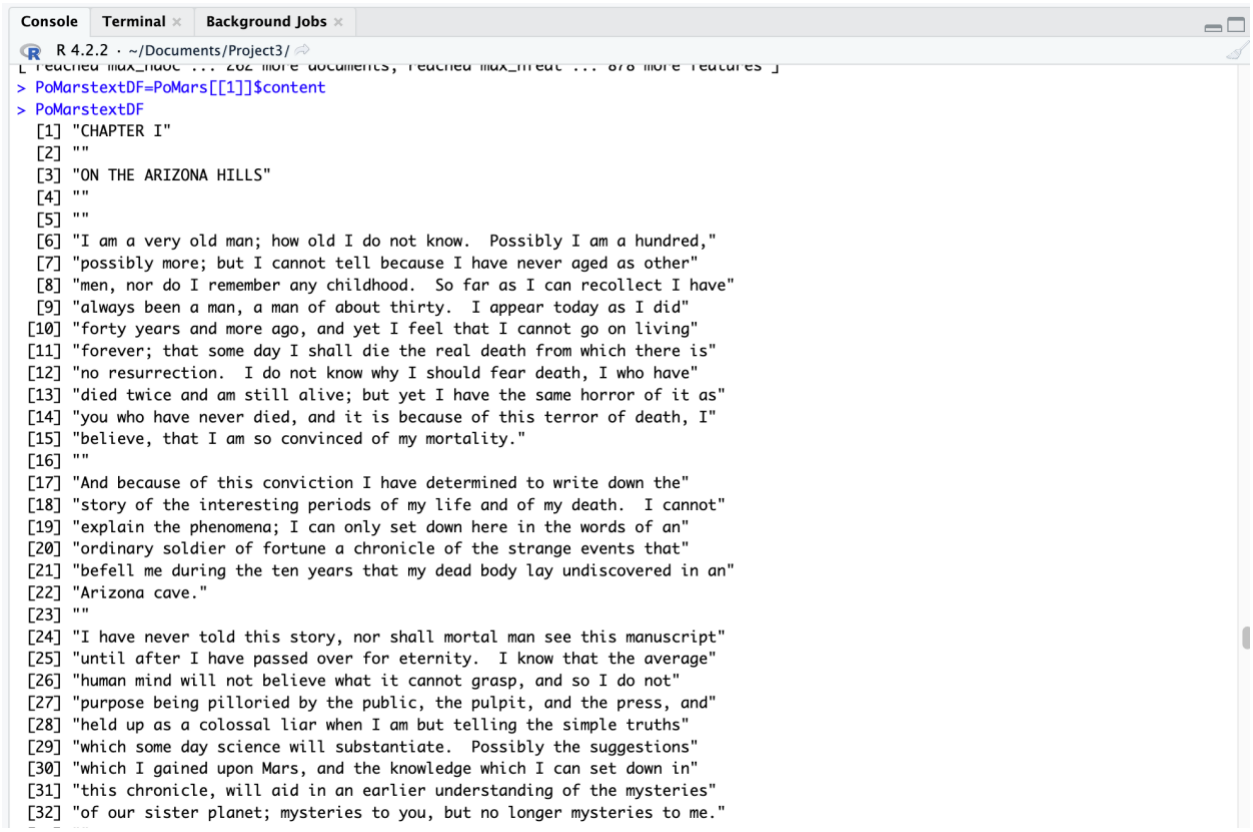
#Term Frequency-InverseDocumentFrequency
PoMarsTFIDF=quanteda::dfm_tfidf (PoMarsDFM, scheme_tf = "count”)
PoMarsTFIDF

```
> PoMarsTFIDF=quanteda::dfm_tfidf(PoMarsDFM, scheme_tf = "count" )
> PoMarsTFIDF
Document-feature matrix of: 268 documents, 888 features (98.97% sparse) and 0 docvars.
       features
docs     chapter         i       on       the  arizona    hills       am        a     very
  text1 2.428135 0.4369087 0        0         0        0        0        0        0
  text2 0        0         0        0         0        0        0        0        0
  text3 0        0         1.252044 0.2882557 1.729165 2.428135 0        0        0
  text4 0        0         0        0         0        0        0        0        0
  text5 0        0         0        0         0        0        0        0        0
  text6 0        1.3107262 0        0         0        0        2.947785 1.407718 2.127105
       features
docs         old
  text1 0
  text2 0
  text3 0
  text4 0
  text5 0
  text6 3.902027
[ reached max_ndoc ... 262 more documents, reached max_nfeat ... 878 more features ]
>
```

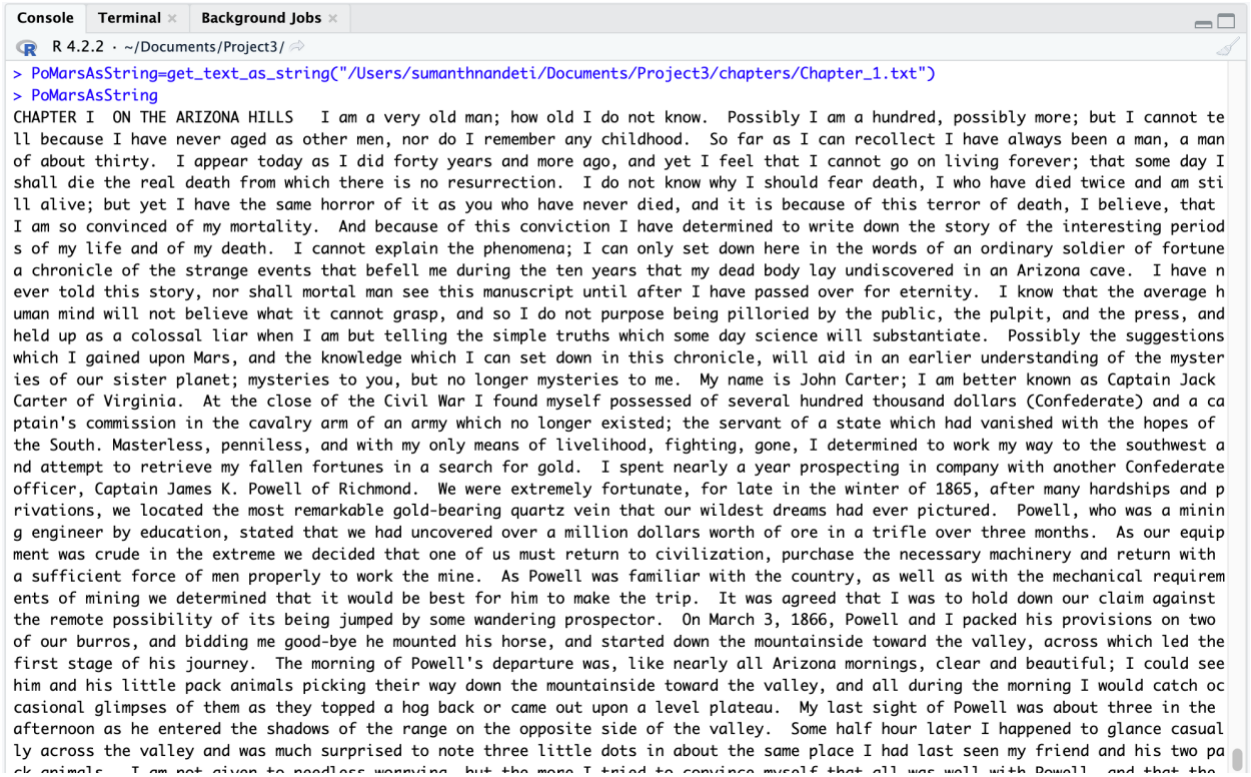
Syuzhet Package

We are using the Syuzhet package to analyze the sentiment of the text in the "Chapter_1" chapter. First, we extract the text from the file and convert it into a string. Then, we use the get_sentences function to break the text into sentences and get_sentiment function to calculate the sentiment score of each sentence using the Syuzhet dictionary.

PoMarstextDF=PoMars [[1]] $content
PoMarstextDF



PoMarsAsString=get_text_as_string("/Users/sumanthnandeti/Documents/Project3/chapters/Chapter_1.txt")
PoMarsAsString
PoMarsSentences=get_sentences (PoMarsAsString)



We then compute the overall sentiment of the text by summing up the sentiment scores of all the sentences. We also calculate the average sentiment score and generate a summary of the sentiment scores using the summary function. Finally, we plot the sentiment scores trajectory using the plot function.
Overall, this code allows us to perform sentiment analysis of the text using sentiment dictionaries and visualize the results.

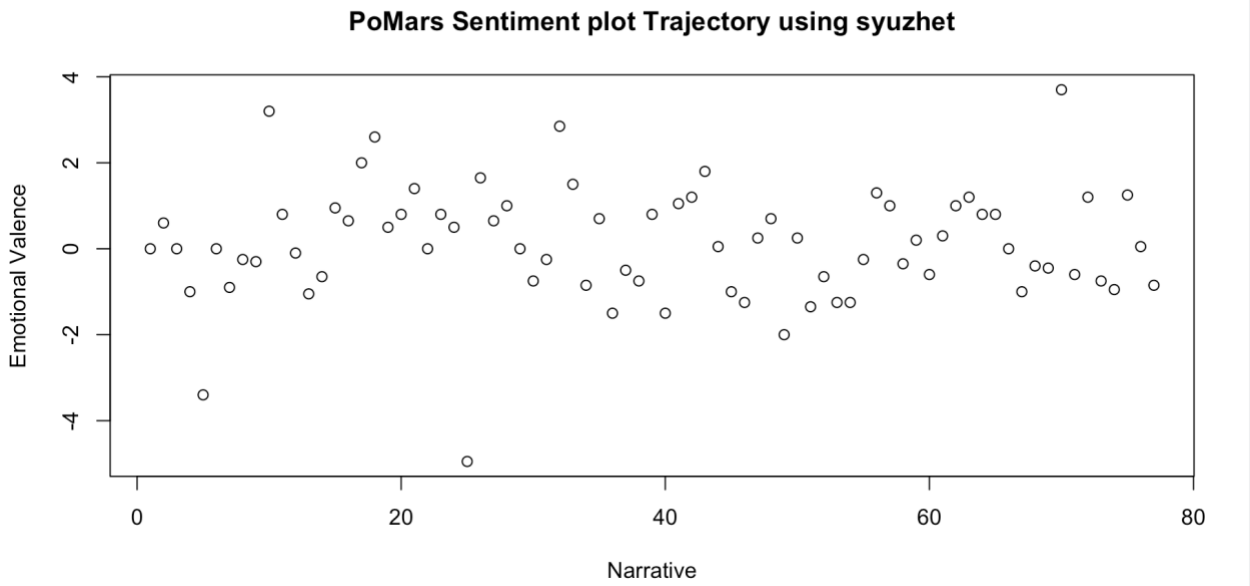
```
#Compute sentiment
PoMarsSentiment=get_sentiment (PoMarsSentences,"syuzhet")
#Sentiment Dictionary using syuzhet
SyuzhetDictionary=get_sentiment_dictionary("syuzhet")
SyuzhetDictionary
```

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Project3/
> PoMarsSentences=get_sentences(PoMarsAsString)
> PoMarsSentiment=get_sentiment(PoMarsSentences,"syuzhet")
> SyuzhetDictionary=get_sentiment_dictionary("syuzhet")
> SyuzhetDictionary
             word value
1          abandon -0.75
2        abandoned -0.50
3        abandoner -0.25
4      abandonment -0.25
5         abandons -1.00
6         abducted -1.00
7        abduction -0.50
8       abductions -1.00
9         aberrant -0.60
10      aberration -0.80
11           abhor -0.50
12        abhorred -1.00
13       abhorrent -0.50
14          abhors -1.00
15       abilities  0.60
16         ability  0.50
17          abject -1.00
18          ablaze -0.25
19        abnormal -0.50
20          aboard  0.25
21         abolish -0.50
22      abominable -0.50
23      abominably -1.00
24       abominate -1.00
25     abomination -0.50
26           abort -0.50
27         aborted -0.80
28        abortion -0.80
29        abortive -1.00
```

finding overall sentiment in the text using syuzhet

```
> SentimentSum=sum(PoMarsSentiment)
> SentimentSum
[1] 8.4
> SentimentMean=mean(PoMarsSentiment)
> SentimentMean
[1] 0.1090909
> summary(PoMarsSentiment)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.9500 -0.7500  0.0000  0.1091  0.8000  3.7000
>
```

```
> plot(PoMarsSentiment, main = "PoMars Sentiment plot Trajectory using syuzhet", xlab = "Narrative", ylab = "Emotional Valence")
>
```



Sentiment Dictionary Bing

We use the Syuzhet package to analyze the sentiment of the text in the "Chapter_1" chapter using the Bing sentiment lexicon. First, the get_sentiment_dictionary () function retrieves the Bing sentiment dictionary. Then, the get_sentiment () function calculates the sentiment scores for each sentence in the text.

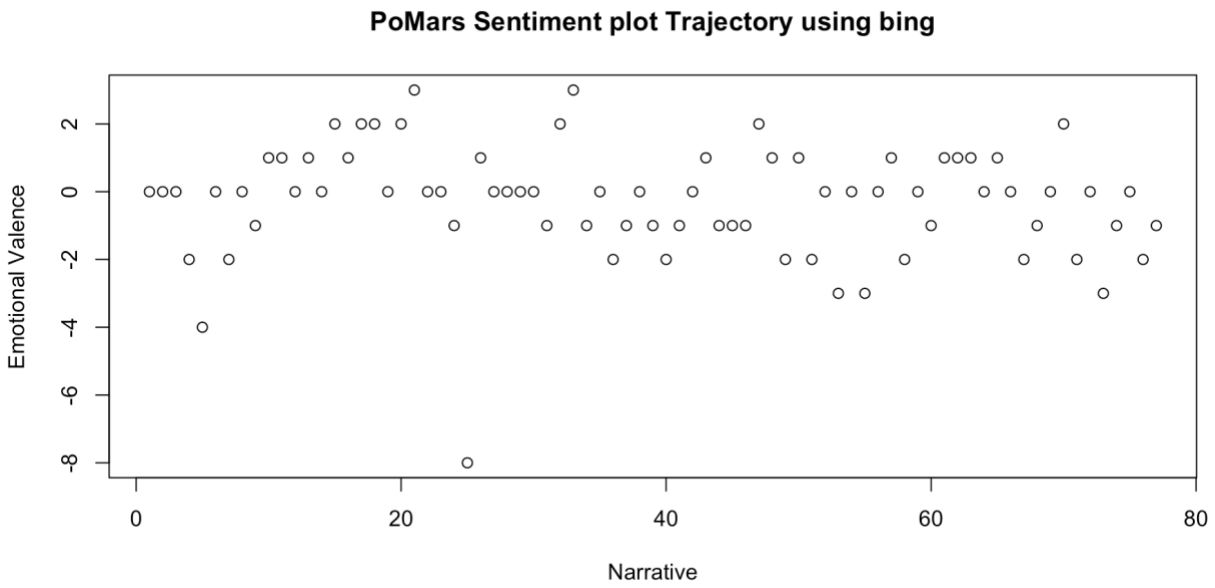
```
BingDictionary=get_sentiment_dictionary("bing")
BingDictionary
PoMarsSentimentBing=get_sentiment (PoMarsSentences,"bing")
```

```
> BingDictionary=get_sentiment_dictionary("bing")
> BingDictionary
                   word value
1                    a+     1
2                abound     1
3               abounds     1
4             abundance     1
5              abundant     1
6            accessable     1
7            accessible     1
8               acclaim     1
9             acclaimed     1
10          acclamation     1
11             accolade     1
12            accolades     1
13        accommodative     1
14         accomodative     1
15           accomplish     1
16         accomplished     1
17       accomplishment     1
18      accomplishments     1
19              accurate     1
20            accurately     1
21            achievable     1
22           achievement     1
23          achievements     1
24            achievible     1
25                acumen     1
26             adaptable     1
27              adaptive     1
28              adequate     1
29            adjustable     1
```

The overall sentiment of the text is then calculated by summing the sentiment scores using the sum () function and finding the mean sentiment score using the mean () function. The summary () function provides a summary of the sentiment scores. Finally, 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. This code provides insight into the overall sentiment of the text and allows for a deeper analysis of the emotions expressed in "Prince of Mars".

```
> SentimentSumBing=sum(PoMarsSentimentBing)
> SentimentSumBing
[1] -22
> SentimentMeanBing=mean(PoMarsSentimentBing)
> SentimentMeanBing
[1] -0.2857143
> summary(PoMarsSentimentBing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.0000 -1.0000  0.0000 -0.2857  1.0000  3.0000
>
```

```
> plot(PoMarsSentimentBing, main = "PoMars Sentiment plot Trajectory using bing", xlab = "Narrative", ylab = "Emotional Valence")
>
```



Sentiment Dictionary NRC of syuzhet:

```
nrcDictionary=get_sentiment_dictionary("nrc")
nrcDictionary
```

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

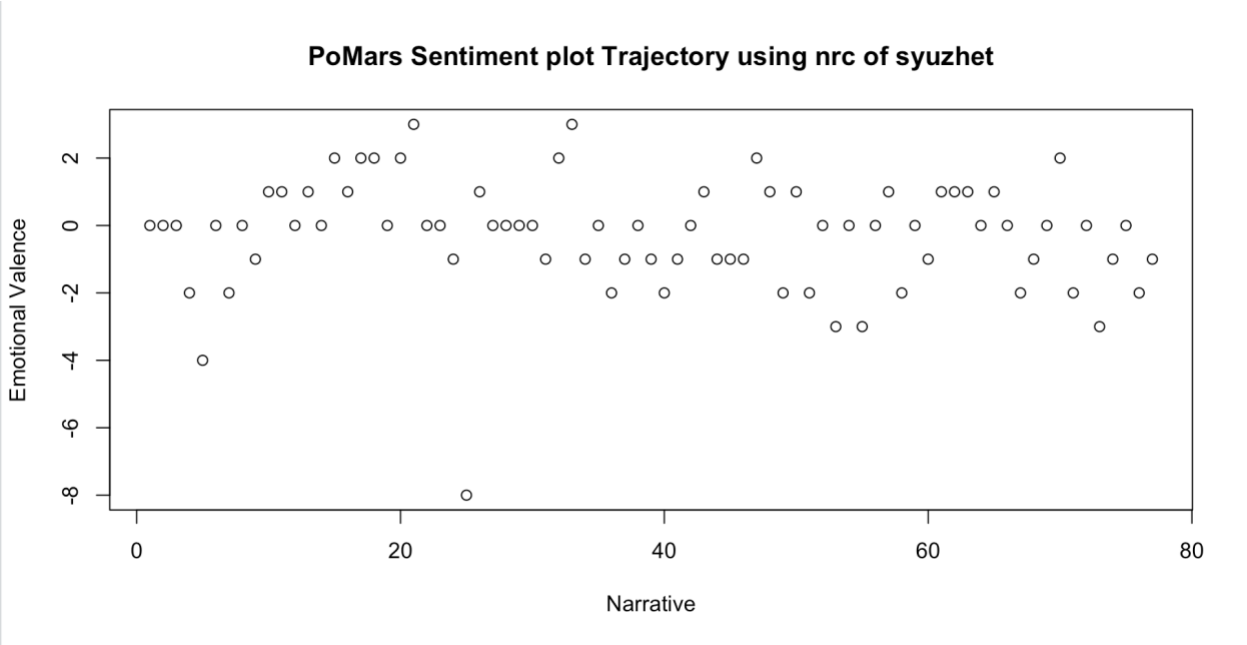
PoMarsSentimentNrc=get_sentiment (PoMarsSentences,"nrc")
PoMarsSentimentNrc

We use the get_sentiment_dictionary () and get_sentiment () functions from the syuzhet package to compute the sentiment of the text using the NRC dictionary. First, we retrieve the NRC sentiment dictionary using get_sentiment_dictionary("nrc") function. Using the NRC dictionary, we then use the get_sentiment () function to calculate the sentiment of each sentence in the text. The results are stored in PoMarsSentimentNrc.

```
> PoMarsSentimentNrc=get_sentiment(PoMarsSentences,"nrc")
> PoMarsSentimentNrc
 [1]  0  1  0 -1 -4  0  3 -1  2  4  0 -1  1  2  2  0 -1  2  1  3  0  0  1 -1 -5  0  2  2  0  1 -1  3 -1 -1  1 -1  0  0 -1 -2  4  2
[43]  1 -1 -2 -1  1  1 -3  0 -2 -2 -1 -1  0  2  1  1  1  0 -1  1  2 -1  0  0  0 -1  0  5  0  0 -1 -2  3 -2  0
>
```

We then calculate the overall sentiment of the text by summing the individual sentence sentiment scores using SentimentSumNRC. We also compute the mean of the sentiment scores using SentimentMeanNrc. We then display the summary statistics of PoMarsSentimentNrc using summary (). Finally, we plot the sentiment scores using plot () with the sentiment scores on the y-axis and the narrative on the x-axis.

```
> SentimentSumNRC=sum(PoMarsSentimentNrc)
> SentimentSumNRC
[1] 14
> SentimentMeanNrc=mean(PoMarsSentimentNrc)
> SentimentMeanNrc
[1] 0.1818182
> summary(PoMarsSentimentNrc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.0000 -1.0000  0.0000  0.1818  1.0000  5.0000
> plot(PoMarsSentimentBing, main = "PoMars Sentiment plot Trajectory using nrc of syuzhet", xlab = "Narrative", ylab = "Emotional Vale
nce")
>
```



Sentiment Dictionary NRC of tidytext:

We use the get_nrc_sentiment () function from the tidytext package to compute the sentiment of each sentence in the PoMarsSentences object, using the NRC sentiment dictionary. The resulting sentiment scores are stored in the PoMarsSentimentNrc1 object.

Percentage value with 10 bins

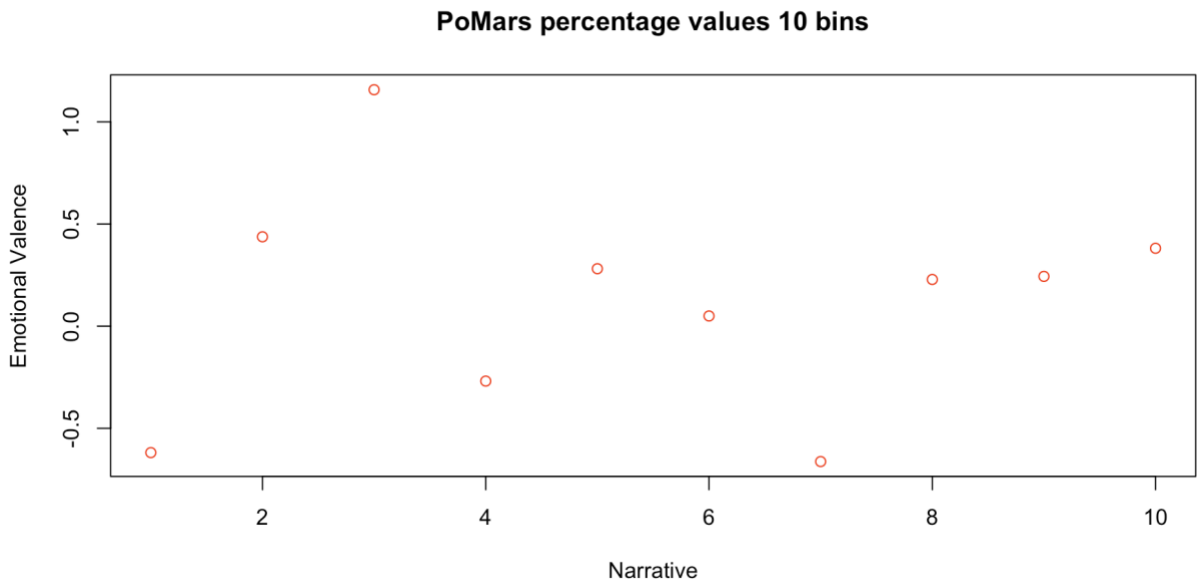
get_percentage_values (): It is a function in R used to compute each bin's observations percentage. It takes in two arguments, the vector of numeric values to be binned and the number of bins to use. The function divides the range of the input values into equally sized bins and counts the number of values that fall into each bin. Then, it calculates the percentage of values that fall into each bin and returns a vector of these percentages.

```
> PoMarsPCTValue=get_percentage_values(PoMarsSentiment, bins=10)
> PoMarsPCTValue
         1          2          3          4          5          6          7          8          9         10
-0.6187500  0.4375000  1.1571429 -0.2687500  0.2812500  0.0500000 -0.6625000  0.2285714  0.2437500  0.3812500
>
```

structure (): A function in R allows you to assign attributes to an object. In the context of the code provided, it assigns attributes to the output of get_percentage_values () function. The attributes assigned by structure () are the number of bins used, the bin range, and the bin width. This can be useful when working with large datasets and when you want to keep track of metadata associated with an object. In addition, assigning attributes with structure () can allow you to manipulate and work with data in a more organized way.

```
> structure(PoMarsPCTValue)
         1          2          3          4          5          6          7          8          9         10
-0.6187500  0.4375000  1.1571429 -0.2687500  0.2812500  0.0500000 -0.6625000  0.2285714  0.2437500  0.3812500
>
```

Plot for percentage values of 10 bins

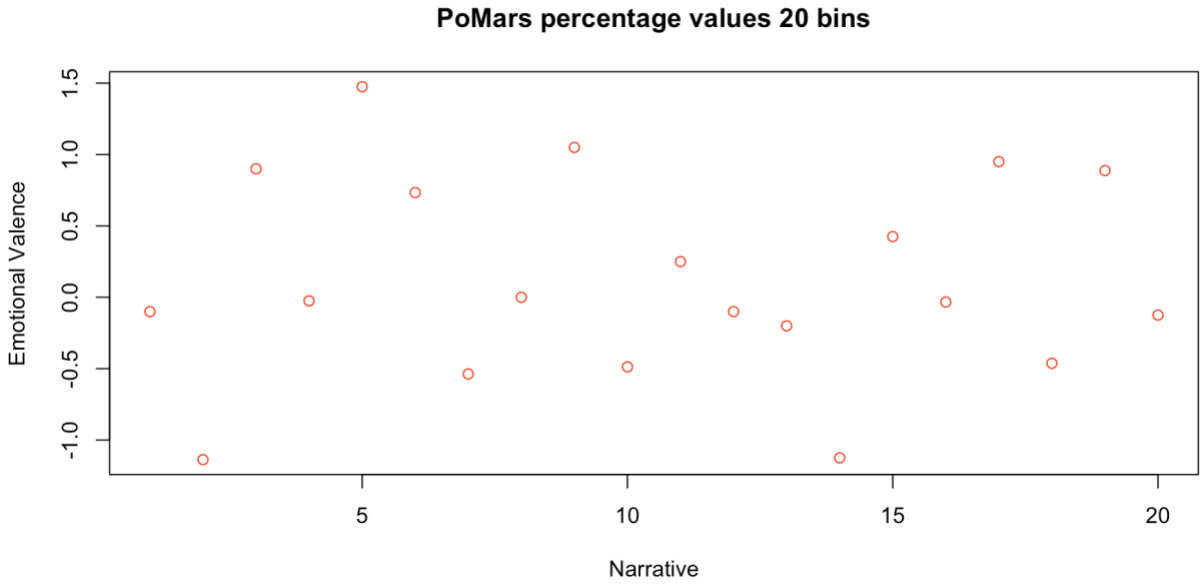


Percentage value with 20 bins

```
> PoMarsPCTValue=get_percentage_values(PoMarsSentiment, bins=20)
> PoMarsPCTValue
         1          2          3          4          5          6          7          8          9         10
-0.10000000 -1.13750000  0.90000000 -0.02500000  1.47500000  0.73333333 -0.53750000  0.00000000  1.05000000 -0.48750000
        11         12         13         14         15         16         17         18         19         20
 0.25000000 -0.10000000 -0.20000000 -1.12500000  0.42500000 -0.03333333  0.95000000 -0.46250000  0.88750000 -0.12500000
>
```

```
> structure(PoMarsPCTValue)
         1          2          3          4          5          6          7          8          9         10
-0.10000000 -1.13750000  0.90000000 -0.02500000  1.47500000  0.73333333 -0.53750000  0.00000000  1.05000000 -0.48750000
        11         12         13         14         15         16         17         18         19         20
 0.25000000 -0.10000000 -0.20000000 -1.12500000  0.42500000 -0.03333333  0.95000000 -0.46250000  0.88750000 -0.12500000
>
```

Plot for percentage values of 20 bins:



Computing sentiment for each chapter:

This Analysis examines the sentiment of many text files in a chapter. Three sentiment lexicons are used in the sentiment analysis: "syuzhet," "bing," and "nrc." The sentiment analysis findings for each chapter are obtained using the computed sentiment () function, which leverages the syuzhet package's get_sentences() and get_sentiment() functions to produce the total sentiment sum and mean for each lexicon. The results are saved in a list, and sentiment analysis is performed on all chapter files using the lapply () function. Finally, the sentiment results are grouped into a table with columns reflecting the chapter title as well as the total sentiment sum and mean for each lexicon. Using the print () function, the resulting table is then written to the console.

```
> compute_sentiment <- function(text) {
+   sentences <- get_sentences(text)
+   sentiment_values <- get_sentiment(sentences, "syuzhet")
+   sentiment_sum <- sum(sentiment_values)
+   sentiment_mean <- mean(sentiment_values)
+   Sentiment_ValuesBing<-get_sentiment(sentences, "bing")
+   sentiment_sum_bing <- sum(Sentiment_ValuesBing)
+   sentiment_mean_bing <- mean(Sentiment_ValuesBing)
+   Sentiment_ValuesNRC<-get_sentiment(sentences, "nrc")
+   sentiment_sum_nrc <- sum(Sentiment_ValuesNRC)
+   sentiment_mean_nrc <- mean(Sentiment_ValuesNRC)
+
+   list(syuzhetSum = sentiment_sum, syuzhetMean = sentiment_mean,bingSum = sentiment_sum_bing, bingMean = sentiment_mean_bing,nrcSum
= sentiment_sum_nrc, nrcMean = sentiment_mean_nrc )
+ }
>
```

```
> sentiment_results <- lapply(chapter_files, function(chapter_file) {
+   chapter_text <- get_text_as_string(chapter_file)
+   compute_sentiment(chapter_text)
+ })
> sentiment_results
[[1]]
[[1]]$syuzhetSum
[1] 8.4

[[1]]$syuzhetMean
[1] 0.1090909

[[1]]$bingSum
[1] -22

[[1]]$bingMean
[1] -0.2857143

[[1]]$nrcSum
[1] 14

[[1]]$nrcMean
[1] 0.1818182
```

```
> chapter_names <- basename(chapter_files)
> sentiment_table <- tibble(
+   Chapter = chapter_names,
+   SentimentSum_syuzhet = sapply(sentiment_results, function(res) res$syuzhetSum),
+   SentimentMean_syuzhet = sapply(sentiment_results, function(res) res$syuzhetMean),
+   SentimentSum_bing = sapply(sentiment_results, function(res) res$bingSum),
+   SentimentMean_bing = sapply(sentiment_results, function(res) res$bingMean),
+   SentimentSum_nrc = sapply(sentiment_results, function(res) res$nrcSum),
+   SentimentMean_nrc = sapply(sentiment_results, function(res) res$nrcMean)
+ )
> print(sentiment_table)
```

```
> print(sentiment_table)
# A tibble: 11 × 7
   Chapter        SentimentSum_syuzhet SentimentMean_syuzhet SentimentSum_bing SentimentMean_bing SentimentSum_nrc SentimentMean…¹
   <chr>                         <dbl>                 <dbl>             <int>              <dbl>            <dbl>            <dbl>
 1 Chapter_1.txt                  8.4                  0.109               -22             -0.286               14            0.182
 2 Chapter_10.txt                 9.85                 0.0801              -33             -0.268               38            0.309
 3 Chapter_11.txt                25.7                  0.292               -5             -0.0568              42            0.477
 4 Chapter_2.txt                -26.8                 -0.505              -60             -1.13               -30           -0.566
 5 Chapter_3.txt                 25.1                  0.289                3              0.0345              52            0.598
 6 Chapter_4.txt                 18                    0.261              -10             -0.145               35            0.507
 7 Chapter_5.txt                 11.6                  0.259                2              0.0444              21            0.467
 8 Chapter_6.txt                -18.6                 -0.371              -32             -0.64                 0            0
 9 Chapter_7.txt                 25.9                  0.425                4              0.0656              48            0.787
10 Chapter_8.txt                -22                   -0.379              -26             -0.448                4            0.0690
11 Chapter_9.txt                 -0.350               -0.00714            -16             -0.327               -2           -0.0408
# … with abbreviated variable name ¹SentimentMean_nrc
>
```

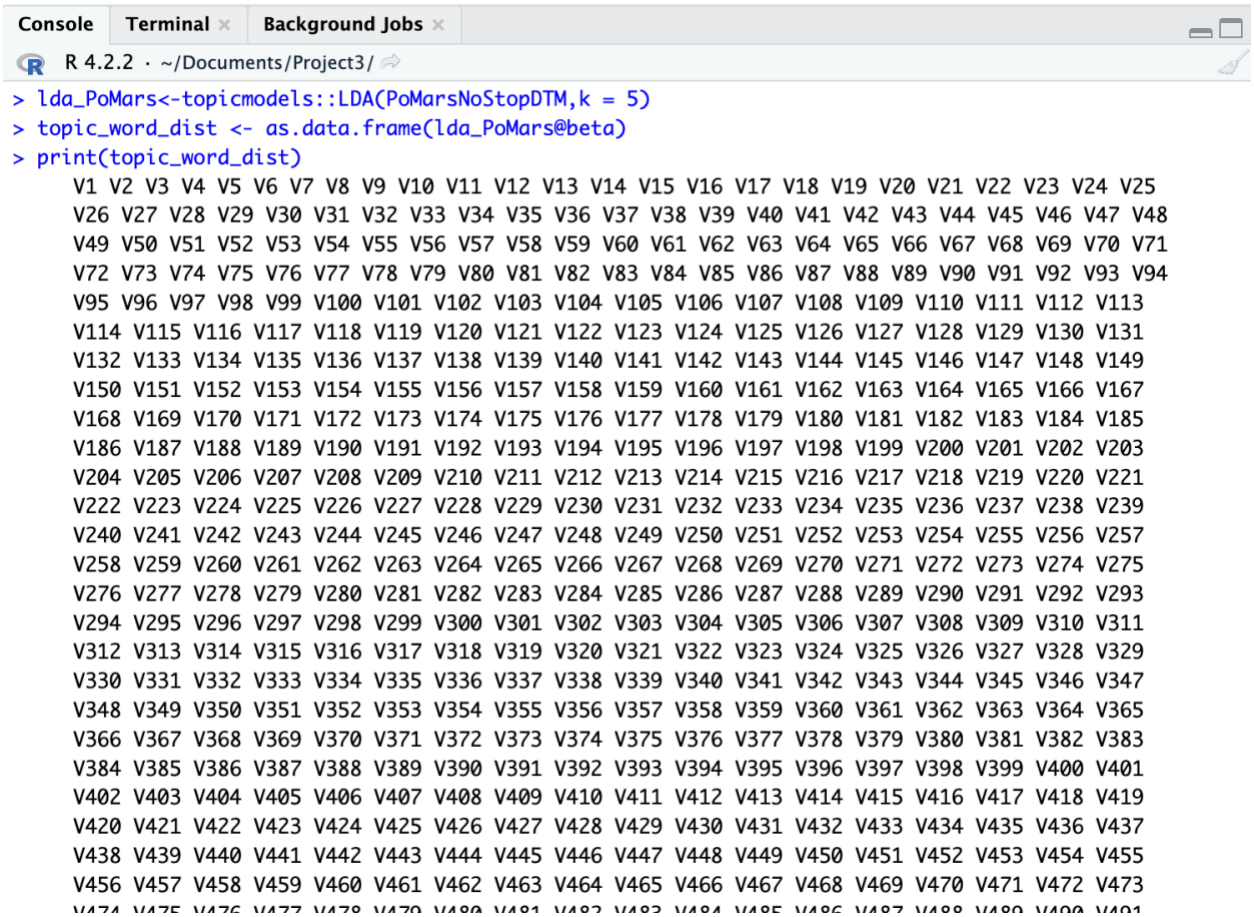
Chapter	Sum_syuzhet	Mean_syuzhet	Sum_bing	Mean_bing	Sum_nrc	Mean_nrc
Chapter_1.txt	8.4	0.109	-22	-0.286	14	0.182
Chapter_10.txt	9.85	0.0801	-33	-0.268	38	0.309
Chapter_11.txt	25.7	0.292	-5	-0.0568	42	0.477
Chapter_2.txt	-26.8	-0.505	-60	-1.13	-30	-0.566
Chapter_3.txt	25.1	0.289	3	0.0345	52	0.598
Chapter_4.txt	18	0.261	-10	-0.145	35	0.507
Chapter_5.txt	11.6	0.259	2	0.0444	21	0.467
Chapter_6.txt	-18.6	-0.371	-32	-0.64	0	0
Chapter_7.txt	25.9	0.425	4	0.0656	48	0.787
Chapter_8.txt	-22	-0.379	-26	-0.448	4	0.0690
Chapter_9.txt	-0.350	-0.00714	-16	-0.327	-2	-0.0408

TOPIC MODELS PACKAGE:

Latent Dirichlet Allocation (LDA) is a generative probabilistic model used for topic modeling in natural language processing and machine learning. The main goal of LDA is to discover the underlying latent topics in a collection of documents. LDA is an unsupervised technique, meaning it does not require labeled data to train the model.

lda_PoMars<-topicmodels::LDA (PoMarsNoStopDTM,k = 5)

Print the topic-word distribution.
topic_word_dist <- as.data. frame(lda_PoMars@beta)
print(topic_word_dist)



Print the document-topic distribution.

```
> document_topic_dist <- as.data.frame(lda_PoMars@gamma)
> print(document_topic_dist)
             V1           V2           V3           V4           V5
1  9.999669e-01 8.266794e-06 8.266794e-06 8.266794e-06 8.266794e-06
2  6.050657e-06 6.050657e-06 9.999758e-01 6.050657e-06 6.050657e-06
3  9.048347e-06 9.999638e-01 9.048347e-06 9.048347e-06 9.048347e-06
4  1.282602e-05 1.282602e-05 9.999487e-01 1.282602e-05 1.282602e-05
5  8.198702e-06 9.999672e-01 8.198702e-06 8.198702e-06 8.198702e-06
6  9.748429e-06 9.748429e-06 9.748429e-06 9.748429e-06 9.999610e-01
7  1.334178e-05 1.334178e-05 1.334178e-05 9.999466e-01 1.334178e-05
8  1.285916e-05 1.285916e-05 9.999486e-01 1.285916e-05 1.285916e-05
9  9.999588e-01 1.030344e-05 1.030344e-05 1.030344e-05 1.030344e-05
10 9.933271e-06 9.933271e-06 9.933271e-06 9.933271e-06 9.999603e-01
11 1.528859e-05 1.528859e-05 1.528859e-05 9.999388e-01 1.528859e-05
>
```

Correlated Topic Model (CTM) is a generative probabilistic model used for topic modeling in natural language processing and machine learning. CTM is an extension of the Latent Dirichlet Allocation (LDA) model, which also aims to discover the underlying latent topics in a collection of documents. However, unlike LDA, CTM allows topics to be correlated, which can lead to a more accurate representation of the underlying structure of the text data.

```
# Estimate the CTM model
ctm_model <- CTM (PoMarsNoStopDTM, k = 5)

# Print the topic-word distribution
topic_word_dist <- as.data. frame(ctm_model@beta)
print(topic_word_dist)
```

```
> topic_word_dist <- as.data.frame(ctm_model@beta)
> print(topic_word_dist)
     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
```

# Print the document-topic distribution
document_topic_dist <- as.data. frame(ctm_model@gamma)
print(document_topic_dist)

```
> document_topic_dist <- as.data.frame(ctm_model@gamma)
> print(document_topic_dist)
              V1           V2           V3           V4          V5
1  4.153385e-13 7.887891e-06 9.956403e-01 4.657702e-06 0.004347187
2  9.928217e-01 4.136640e-06 6.983702e-11 1.787227e-06 0.007172331
3  9.940559e-01 3.894897e-06 5.352244e-11 1.640981e-06 0.005938576
4  4.282260e-13 8.932661e-06 9.956094e-01 5.222060e-06 0.004376421
5  3.099084e-08 6.584824e-09 2.670582e-07 9.948433e-01 0.005156438
6  2.030358e-05 8.896815e-05 1.280460e-05 1.104042e-04 0.999767519
7  3.850755e-08 8.831827e-09 3.416776e-07 9.938793e-01 0.006120321
8  1.352790e-06 9.956244e-01 3.570272e-06 8.023456e-10 0.004370693
9  9.914412e-01 5.321724e-06 9.668530e-11 2.334913e-06 0.008551172
10 1.906430e-05 9.957832e-05 1.431311e-05 9.822699e-05 0.999768817
11 9.942060e-01 4.303840e-06 5.366680e-11 1.744123e-06 0.005787918
```

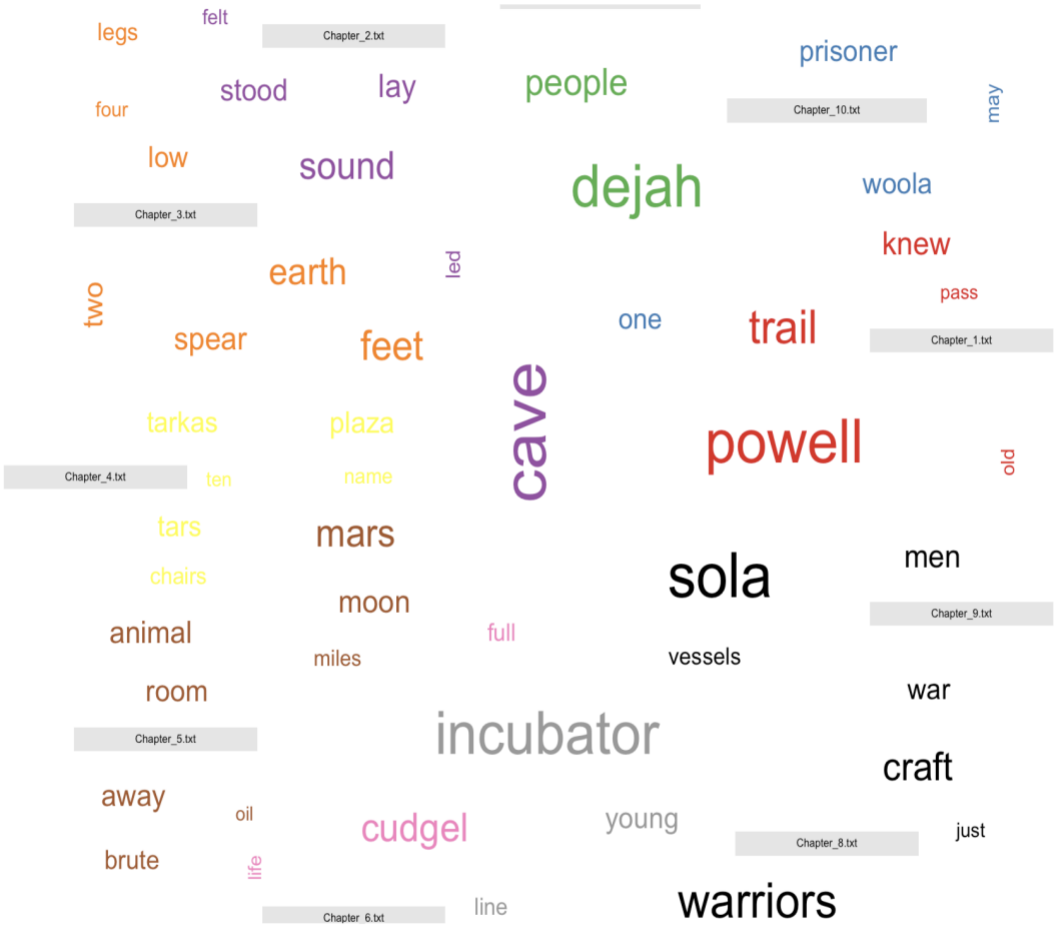
#word cloud package
words=names (PoMarsTF)
pal<-brewer.pal(9,"Spectral")
PoMarsWC=wordcloud (words,PoMarsTF,colors = pal)

```
> words=names(PoMarsTF)
> pal<-brewer.pal(9,"Spectral")
> PoMarsWC=wordcloud(words,PoMarsTF,colors = pal)
```

# Define the colors for each document group
colors <- brewer.pal (9, "Set1")

Create a comparison cloud: Creates a word cloud that compares word frequencies across different texts or groups.
PoMarsNoStopTDM_Mat=as.matrix(PoMarsNoStopTDM)

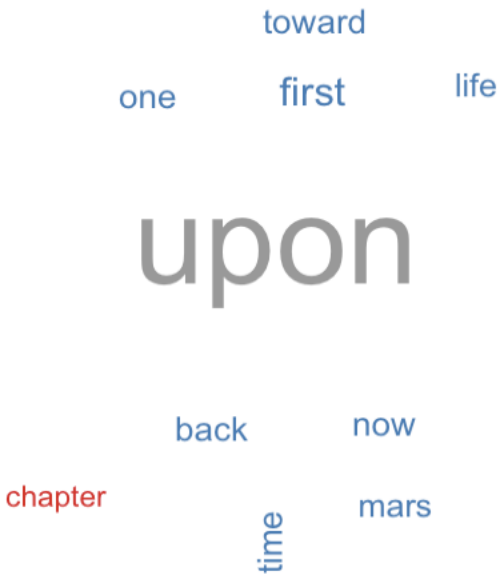
comparison. cloud(PoMarsNoStopTDM_Mat, colors = colors,scale = c(3, 0.5),random.order = FALSE,title.size = 0.5)




```
> colors <- brewer.pal(9, "Set1")
> PoMarsNoStopTDM_Mat=as.matrix(PoMarsNoStopTDM)
> comparison.cloud(PoMarsNoStopTDM_Mat, colors = colors,scale = c(3, 0.5),random.order = FALSE,title.si
ze = 0.5)
There were 50 or more warnings (use warnings() to see the first 50)
> #Commonality Cloud: The commonality.cloud() function is part of the wordcloud package in R and is use
d to create a word cloud that highlights the common words shared by multiple documents
> commonality.cloud(PoMarsNoStopTDM_Mat, colors = colors,scale = c(3, 0.5),random.order = FALSE)
>
```

Commonality Cloud: The commonality. cloud() function is part of the wordcloud package in R and is used to create a word cloud that highlights the common words shared by multiple documents

commonality. cloud (PoMarsNoStopTDM_Mat, colors = colors, scale = c (3, 0.5), random.order = FALSE)



Extra methods for text mining package

weightTf (), weightTfIdf (): Applies term frequency (TF) or term frequency-inverse document frequency (TF-IDF) weighting to a DTM or TDM. These functions can be used to transform the raw term frequencies in a matrix into weighted values that better represent the importance of terms in the documents.

DTM after performing preprocessing is given as the input to the functions.

PoMarsdtm_tf <- weightTf (PoMarsNoStopDTM)
inspect (PoMarsdtm_tf)

```
> PoMarsdtm_tf <- weightTf(PoMarsNoStopDTM)
> inspect(PoMarsdtm_tf)
<<DocumentTermMatrix (documents: 11, terms: 3772)>>
Non-/sparse entries: 7419/34073
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
               Terms
Docs            feet first little mars martian martians one sola toward upon
  Chapter_1.txt    5     3      4    1       0        0   3    0      3   12
  Chapter_10.txt   5     6      4    4      13        1  19    6      9   19
  Chapter_11.txt   0     3      4    2       3        4   1   11      1   13
  Chapter_2.txt    1     5      2    1       0        0   2    0      3   13
  Chapter_3.txt   15     5     11   11       6        5   7    0     10   20
  Chapter_4.txt    9     6      3    6       9        8   5    2      6   11
  Chapter_5.txt    2     1      2    9       5        1   6    5      3    8
  Chapter_6.txt    4     3      0    2       2        7   5    4      3   13
  Chapter_7.txt    2     3     10    4      10       11   9    7      2    7
  Chapter_8.txt    1     5      4    1       7        3   8    3      5   22
```

PoMarsdtm_tfidf <- weightTfIdf (PoMarsNoStopDTM)
inspect (PoMarsdtm_tfidf)

```
> PoMarsdtm_tfidf <- weightTfIdf(PoMarsNoStopDTM)
> inspect(PoMarsdtm_tfidf)
<<DocumentTermMatrix (documents: 11, terms: 3772)>>
Non-/sparse entries: 7309/34183
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
               Terms
Docs                 cave     cudgel       dejah        eggs   incubator      powell
  Chapter_1.txt  0.01634174 0.00000000 0.000000000 0.000000000 0.000000000 0.038811629
  Chapter_10.txt 0.00000000 0.00000000 0.007475476 0.000000000 0.000000000 0.001495095
  Chapter_11.txt 0.00000000 0.00000000 0.038009398 0.000000000 0.000000000 0.000000000
  Chapter_2.txt  0.04437119 0.00000000 0.000000000 0.000000000 0.000000000 0.000000000
  Chapter_3.txt  0.00000000 0.00000000 0.000000000 0.006077673 0.003088088 0.000000000
  Chapter_4.txt  0.00000000 0.00000000 0.000000000 0.000000000 0.000000000 0.000000000
  Chapter_5.txt  0.00000000 0.00000000 0.000000000 0.000000000 0.000000000 0.000000000
  Chapter_6.txt  0.00000000 0.03128685 0.000000000 0.000000000 0.000000000 0.000000000
  Chapter_7.txt  0.00000000 0.00000000 0.000000000 0.022913959 0.029106663 0.000000000
  Chapter_8.txt  0.00000000 0.00000000 0.000000000 0.000000000 0.001870728 0.000000000
               Terms
Docs               tarkas        tars      thoris       trail
  Chapter_1.txt  0.000000000 0.000000000 0.000000000 0.028598042
  Chapter_10.txt 0.004784330 0.004784330 0.007475476 0.000000000
  Chapter_11.txt 0.003179888 0.003179888 0.038009398 0.000000000
  Chapter_2.txt  0.000000000 0.000000000 0.000000000 0.009508112
  Chapter_3.txt  0.000000000 0.000000000 0.000000000 0.000000000
  Chapter_4.txt  0.006851864 0.006851864 0.000000000 0.000000000
  Chapter_5.txt  0.000000000 0.000000000 0.000000000 0.000000000
  Chapter_6.txt  0.005649025 0.005649025 0.000000000 0.000000000
  Chapter_7.txt  0.003620990 0.003620990 0.000000000 0.000000000
  Chapter_8.txt  0.000000000 0.000000000 0.000000000 0.000000000
>
```

stemDocument (): Performs word stemming using the Snowball stemmer.

PoMarsNoStopWords_stemmed <- tm_map (PoMarsNoStopWords, stemDocument)
inspect (PoMarsNoStopWords_stemmed)

```
> PoMarsNoStopWords_stemmed <- tm_map(PoMarsNoStopWords, stemDocument)
> inspect(PoMarsNoStopWords_stemmed)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 7275

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 10058

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 6764
```

lapply (PoMarsNoStopWords_stemmed, as.character)

```
$Chapter_8.txt
  [1] "chapter viii"
  [2] ""
  [3] "fair captiv sky"
  [4] ""
  [5] ""
  [6] "third day incub ceremoni set forth toward home"
  [7] "scarc head process debouch open"
  [8] "ground citi order given immedi"
  [9] "hasti return though train year particular"
 [10] "evolut green martian melt like mist spacious"
 [11] "doorway nearbi build less three minut"
 [12] "entir cavalcad chariot mastodon mount warrior"
 [13] "nowher seen"
 [14] ""
 [15] "sola enter build upon front citi fact"
 [16] "one encount ape"
 [17] "wish see caus sudden retreat mount"
 [18] "upper floor peer window valley"
 [19] "hill beyond saw caus sudden scurri"
 [20] "cover huge craft long low graypaint swung slowli"
```

Discussion on what we learned from this project.

This project has been instrumental in showcasing the various aspects of text analytics, which is a powerful tool for extracting meaningful insights from unstructured textual data. Through the implementation of text pre-processing techniques, term-document matrix creation, and clustering, the project has provided a comprehensive understanding of the building blocks of text analytics.

As demonstrated in this study, the examination of data is an essential initial step in text analytics. We learned how to prepare and clean textual data for subsequent analysis by tokenizing the text, deleting stop words, and converting the text to lowercase. Identifying the longest terms and phrases in each chapter helps us understand the significance of feature extraction and its role in revealing the distinctive qualities of a text. Furthermore, building term-document and document-term matrices allows us to systematically express text data in a structured fashion, making it more accessible for quantitative analysis.

Another essential component of this study is the use of clustering methods to group related terms based on their co-occurrence in chapters. This highlights how unsupervised learning techniques may be used in text analytics to uncover underlying patterns and relationships in data. By displaying the clustering results with a dendrogram, we can better grasp the text's hierarchical structure and identify potential relationships between distinct elements.

Overall, this project has substantially improved our grasp of text analytics by leading us through various steps of data discovery, pre-processing, and analysis. We've learned a lot about the power of text analytics in identifying patterns and relationships in unstructured data through practical application. Furthermore, the research has emphasized the significance of using a systematic and iterative approach to text analysis, which is critical for extracting useful insights from complex textual data.

Methods Used, Results Obtained, and Understanding the Theme of the Book

We used many text analysis techniques in this study to understand the concept of the book by processing its chapters. The methods utilized, the data obtained, and what can be gleaned about the book's theme are described below.

1. Text pre-processing: The code reads the text, divides it into chapters, eliminates numbers and punctuation, and lowercases it. It then generates a VCorpus and eliminates stop words.
2. Finding 10 longest words and sentences: For each chapter, the code finds the 10 longest unique words and the 10 longest sentences.
3. Term-Document Matrix (TDM) and Document-Term Matrix (DTM): TDM and DTM are generated by the code for the pre-processed text corpus. It computes term frequency and detects frequent terms in the corpus.
4. Sparsity-based clustering: The method removes terms with high sparsity (more than 85%) and uses Ward's approach to do hierarchical clustering. To visualize the clustering findings, it generates a dendrogram.
5. Word Cloud: A Word Cloud is created to illustrate the frequency of words in the text. Those that occur frequently are displayed as more significant, whereas those that occur less regularly are shown as smaller. This aids in identifying essential themes and keywords in the work.

6. Quanteda: To tokenize the text, build a document-feature matrix (DFM), and compute word frequencies, the Quanteda package is utilized. This section provides an overview of the text's structure and the significance of specific terms in the book.
7. Syuzhet package: The Syuzhet package extracts sentiment information from text. Various sentiment dictionaries (Syuzhet, Bing, and NRC) are used to assess the text's emotional valence. This improves understanding of the overall tone and vibe of the work, providing insights into the emotions and concepts presented.
8. Sentiment analysis plots: Sentiment analysis is represented visually by plots that depict the sentiment trajectory using the Syuzhet, Bing, and NRC sentiment dictionaries. These storylines aid in identifying the book's emotional highs and lows, showing its mood and atmosphere.
9. Percentage values: The method computes percentage values for sentiment scores in bins of 10 and 20. These values aid in comprehending the distribution of emotions across the narrative, providing a more complete view of the book's emotional landscape.

The results include:

- The longest words and sentences in each chapter, providing insight into the text's vocabulary and complexity.
- TDM and DTM, which provide a summary of the term frequency in the text.
- Clustering results (dendrogram), displaying associations between terms based on their co-occurrence in chapters.