Class Project 3 – Text Analysis with R

CSCI 6444 Introduction to big data
and analytics
Professor: Stephen Kaisler

Group 2
Shaival Vora (G37099269)
Nithin Raghava Aitha (G43945136)

Dataset Description

The data set given is a text file called Tarzan of the Apes by Edgar Rice Burroughs. For this project we are going to use chapters I – XV.

Objective

The goal of this project is to conduct text analytics on the extensive text document " Tarzan of the Apes " By employing different analytical techniques, our aim is to comprehend the text, discern its underlying sentiment, and draw insightful observations from it.

Packages

For this project we need some packages and we use “install.packages()” function to install them and “library()” function to load them.

```
# installing the required packages
install.packages('tm')
install.packages('dplyr')
install.packages('tidytext')
install.packages('ggplot2')
install.packages('wordcloud')

# loading the installed packages
library('tm')
library('dplyr')
library('tidytext')
library('ggplot2')
library('wordcloud')
```

These are the packages we initialize at the start but we do need a few other packages that are going to be installed and loaded in the further stage of the project.

Loading the Dataset and Creating a Corpus

We begin by setting up the directory using “ setwd() “ command.

```
> setwd("C:/Users/anith/OneDrive/Desktop/GW/sem 2/Big Daata/project 3")
> getwd()
[1] "C:/Users/anith/OneDrive/Desktop/GW/sem 2/Big Daata/project 3"
```

Then after setting up the directory, we load the dataset.

```
> Book <- readLines('TarzanOfTheApes.txt')
```

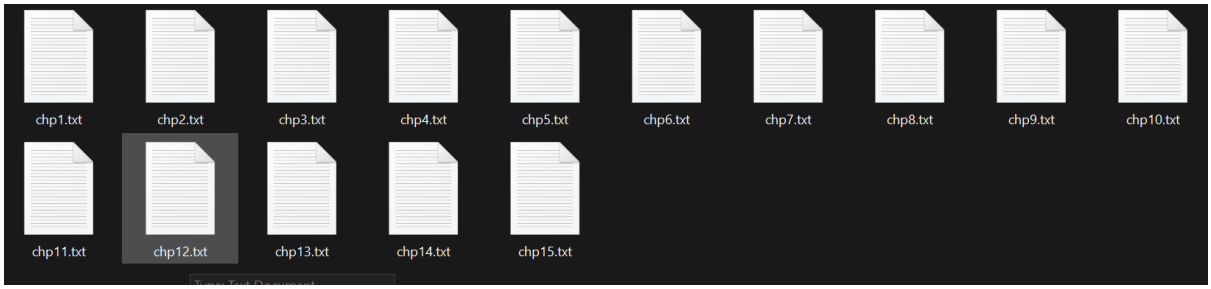
We then divide the chapters into individual documents to create a corpus. Each of the fifteen chapters (Chapter I-XV) is identified using the "which" method as demonstrated below. We segment the chapters based on their titles, such as "Chapter I," and save the content of each chapter in a separate variable. We can then review the output to confirm the accurate separation of chapters.

```
> index_chp1 <- which(Book == "Chapter I", arr.ind=TRUE)
> index_chp2 <- which(Book == "Chapter II", arr.ind=TRUE)
> index_chp3 <- which(Book == "Chapter III", arr.ind=TRUE)
> index_chp4 <- which(Book == "Chapter IV", arr.ind=TRUE)
> index_chp5 <- which(Book == "Chapter V", arr.ind=TRUE)
> index_chp6 <- which(Book == "Chapter VI", arr.ind=TRUE)
> index_chp7 <- which(Book == "Chapter VII", arr.ind=TRUE)
> index_chp8 <- which(Book == "Chapter VIII", arr.ind=TRUE)
> index_chp9 <- which(Book == "Chapter XI", arr.ind=TRUE)
> index_chp10 <- which(Book == "Chapter X", arr.ind=TRUE)
> index_chp11 <- which(Book == "Chapter XI", arr.ind=TRUE)
> index_chp12 <- which(Book == "Chapter XII", arr.ind=TRUE)
> index_chp13 <- which(Book == "Chapter XIII", arr.ind=TRUE)
> index_chp14 <- which(Book == "Chapter XIV", arr.ind=TRUE)
> index_chp15 <- which(Book == "Chapter XV", arr.ind=TRUE)
> index_chp16 <- which(Book == "Chapter XVI", arr.ind=TRUE)

> chp1 <- Book[(index_chp1+1):(index_chp2-1)]
> chp2 <- Book[(index_chp2+1):(index_chp3-1)]
> chp3 <- Book[(index_chp3+1):(index_chp4-1)]
> chp4 <- Book[(index_chp4+1):(index_chp5-1)]
> chp5 <- Book[(index_chp5+1):(index_chp6-1)]
> chp6 <- Book[(index_chp6+1):(index_chp7-1)]
> chp7 <- Book[(index_chp7+1):(index_chp8-1)]
> chp8 <- Book[(index_chp8+1):(index_chp9-1)]
> chp9 <- Book[(index_chp9+1):(index_chp10-1)]
> chp10 <- Book[(index_chp10+1):(index_chp11-1)]
> chp11 <- Book[(index_chp11+1):(index_chp12-1)]
> chp12 <- Book[(index_chp12+1):(index_chp13-1)]
> chp13 <- Book[(index_chp13+1):(index_chp14-1)]
> chp14 <- Book[(index_chp14+1):(index_chp15-1)]
> chp15 <- Book[(index_chp15+1):(index_chp16-1)]
```

We then proceed to create a directory for the chapters and write each chapter to a text file.

```
> dir.create('Chapters')
> # writing each chapter to a text file
> write.table(chp1, file = 'Chapters/chp1.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp2, file = 'Chapters/chp2.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp3, file = 'Chapters/chp3.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp4, file = 'Chapters/chp4.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp5, file = 'Chapters/chp5.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp6, file = 'Chapters/chp6.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp7, file = 'Chapters/chp7.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp8, file = 'Chapters/chp8.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp9, file = 'Chapters/chp9.txt', sep = '\t', row.names = FALSE, col.names = FALSE, q
ote = FALSE)
> write.table(chp10, file = 'Chapters/chp10.txt', sep = '\t', row.names = FALSE, col.names = FALSE,
quote = FALSE)
> write.table(chp11, file = 'Chapters/chp11.txt', sep = '\t', row.names = FALSE, col.names = FALSE,
quote = FALSE)
> write.table(chp12, file = 'Chapters/chp12.txt', sep = '\t', row.names = FALSE, col.names = FALSE,
quote = FALSE)
> write.table(chp13, file = 'Chapters/chp13.txt', sep = '\t', row.names = FALSE, col.names = FALSE,
quote = FALSE)
> write.table(chp14, file = 'Chapters/chp14.txt', sep = '\t', row.names = FALSE, col.names = FALSE,
quote = FALSE)
> write.table(chp15, file = 'Chapters/chp15.txt', sep = '\t', row.names = FALSE, col.names = FALSE,
quote = FALSE)
```



After separating the chapters, the next step is to create a VCorpus. We use the " VCorpus() " method from the "tm" package to accomplish this task and store the corpus in a variable named "CORP". We can then inspect the structure of the corpus using the str() method.

```
> CORP <- VCorpus(DirSource('./Chapters', ignore.case = TRUE, mode = 'text'))
> str(CORP)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 15
  ..$ :List of 2
  .. ..$ content: chr [1:420] "" "Out to Sea" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:215] "" "The Fear-Phantom" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
```

In this overview, we observe that our corpus comprises 15 documents, each representing a chapter, with their content reflecting the text they encompass. The content of each document ranges from approximately 190 to 900 characters. Additionally, we can explore other parameters such as "meta," which holds the metadata associated with each document.

Next, we generate initial Document Term Matrices (DTM). In these matrices, each row corresponds to a document, while each column corresponds to a term within the document. This setup enables us to analyze our data using vector and matrix algebra. Additionally, we create a Term Document Matrix (TDM), where rows represent terms and columns represent documents. These matrices are constructed using the DocumentTermMatrix() and TermDocumentMatrix() methods from the "tm" package.

DocumentTermMatrix:

```
> CORP_DTM <- DocumentTermMatrix(CORP)
> CORP_DTM
<<DocumentTermMatrix (documents: 15, terms: 7557)>>
Non-/sparse entries: 17106/96249
Sparsity           : 85%
Maximal term length: 22
Weighting          : term frequency (tf)
> inspect(CORP_DTM)
<<DocumentTermMatrix (documents: 15, terms: 7557)>>
Non-/sparse entries: 17106/96249
Sparsity           : 85%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
           Terms
Docs        and but for from had his that the was with
  chp1.txt   98  28  40   29  34  54   55 240  56   32
  chp11.txt 125  21  36   25  42  90   23 328  43   45
  chp12.txt  89  21  28   12  44  66   37 199  41   21
  chp13.txt 159  20  20   27  55 105   33 367  55   39
  chp14.txt 104  24  27   23  35  39   39 325  62   26
  chp2.txt   96  29  32   17  27  26   34 211  32   29
  chp4.txt   74  17  20   20  33  49   29 200  45   29
  chp5.txt   95  23  29   12  30  53   31 161  49   24
  chp7.txt  126  26  29   28  37  93   35 373  56   37
  chp8.txt  268  48  62   39  86 158   61 580 114   59
```

```
> str(CORP_DTM)
List of 6
 $ i      : int [1:17106] 1 1 1 1 1 1 1 1 1 1 ...
 $ j      : int [1:17106] 1 2 3 4 5 6 7 8 9 10 ...
 $ v      : num [1:17106] 1 1 1 1 1 1 1 1 1 1 ...
 $ nrow   : int 15
 $ ncol   : int 7557
 $ dimnames:List of 2
  ..$ Docs : chr [1:15] "chp1.txt" "chp10.txt" "chp11.txt" "chp12.txt" ...
  ..$ Terms: chr [1:7557] "'ancient" "'arf" "'e's" "'e;" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

TermDocumentMatrix:

```
> CORP_TDM <- TermDocumentMatrix(CORP)
> CORP_TDM
<<TermDocumentMatrix (terms: 7557, documents: 15)>>
Non-/sparse entries: 17106/96249
Sparsity           : 85%
Maximal term length: 22
Weighting          : term frequency (tf)
> inspect(CORP_TDM)
<<TermDocumentMatrix (terms: 7557, documents: 15)>>
Non-/sparse entries: 17106/96249
Sparsity           : 85%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
       Docs
Terms   chp1.txt chp11.txt chp12.txt chp13.txt chp14.txt chp2.txt chp4.txt chp5.txt chp7.txt
  and         98       125        89       159       104       96       74       95      126
  but         28        21        21        20        24       29       17       23       26
  for         40        36        28        20        27       32       20       29       29
  from        29        25        12        27        23       17       20       12       28
  had         34        42        44        55        35       27       33       30       37
  his         54        90        66       105        39       26       49       53       93
  that        55        23        37        33        39       34       29       31       35
  the        240       328       199       367       325      211      200      161      373
  was         56        43        41        55        62       32       45       49       56
  with        32        45        21        39        26       29       29       24       37
```

```
> str(CORP_TDM)
List of 6
 $ i      : int [1:17106] 1 2 3 4 5 6 7 8 9 10 ...
 $ j      : int [1:17106] 1 1 1 1 1 1 1 1 1 1 ...
 $ v      : num [1:17106] 1 1 1 1 1 1 1 1 1 1 ...
 $ nrow   : int 7557
 $ ncol   : int 15
 $ dimnames:List of 2
  ..$ Terms: chr [1:7557] "'ancient" "'arf" "'e's" "'e;" ...
  ..$ Docs : chr [1:15] "chp1.txt" "chp10.txt" "chp11.txt" "chp12.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

We can see that the DTM and TDM are providing similar results in a different form.

Finding 10 longest words and sentences in each chapter

Before cleaning the corpus let us find 10 longest words and 10 longest sentences from all the chapters and individually from each chapter.
For this we need packages “dplyr”, “tidytext”, and “ggplot2” which we have installed in the beginning.

First we need to convert our corpus into tidy corpus using the tidy method.

```
> tidyCORP <- tidy(CORP)
> tidyCORP
# A tibble: 15 × 8
   author datetimestamp       description heading id        language origin text
   <lgl>  <dttm>              <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
 1 NA     2024-05-04 20:54:09 NA          NA      chp1.txt  en       NA     "\nOut to Sea\n\n\nI…
 2 NA     2024-05-04 20:54:09 NA          NA      chp10.txt en       NA     "\nThe Fear-Phantom\…
 3 NA     2024-05-04 20:54:09 NA          NA      chp11.txt en       NA     "\n\"King of the Ape…
 4 NA     2024-05-04 20:54:09 NA          NA      chp12.txt en       NA     "\nMan's Reason\n\n\…
 5 NA     2024-05-04 20:54:09 NA          NA      chp13.txt en       NA     "\nHis Own Kind\n\n\…
 6 NA     2024-05-04 20:54:09 NA          NA      chp14.txt en       NA     "\nAt the Mercy of t…
 7 NA     2024-05-04 20:54:09 NA          NA      chp15.txt en       NA     "\nThe Forest God\n\…
 8 NA     2024-05-04 20:54:09 NA          NA      chp2.txt  en       NA     "\nThe Savage Home\n…
 9 NA     2024-05-04 20:54:09 NA          NA      chp3.txt  en       NA     "\nLife and Death\n\…
10 NA     2024-05-04 20:54:09 NA          NA      chp4.txt  en       NA     "\nThe Apes\n\n\nIn …
11 NA     2024-05-04 20:54:09 NA          NA      chp5.txt  en       NA     "\nThe White Ape\n\n…
12 NA     2024-05-04 20:54:09 NA          NA      chp6.txt  en       NA     "\nJungle Battles\n\…
13 NA     2024-05-04 20:54:09 NA          NA      chp7.txt  en       NA     "\nThe Light of Know…
14 NA     2024-05-04 20:54:09 NA          NA      chp8.txt  en       NA     "\nThe Tree-top Hunt…
15 NA     2024-05-04 20:54:09 NA          NA      chp9.txt  en       NA     "\nChapter XI\n\n\n\…
```

Then we find the 10 longest words by tokenizing the words.

```
> CORPWords<-tidyCORP %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPWords
# A tibble: 47,187 × 3
   id        word            word_length
   <chr>     <chr>                 <int>
 1 chp11.txt incomprehensible         16
 2 chp3.txt  responsibilities         16
 3 chp10.txt destructiveness          15
 4 chp11.txt experimentation          15
 5 chp13.txt personification          15
 6 chp2.txt  notwithstanding          15
 7 chp4.txt  notwithstanding          15
 8 chp5.txt  resourcefulness          15
 9 chp8.txt  destructiveness          15
10 chp9.txt  destructiveness          15
# ℹ 47,177 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Here we can see the 10 longest words in all the chapters combined with the word “incomprehensible” being the longest with 16 words.
Similarly, by modifying the unnest_tokens method we find the 10 longest sentences.

```
> CORPSentences<-tidyCORP %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences
# A tibble: 1,894 × 3
   id        sentence                                                  sentence_length
   <chr>     <chr>                                                               <int>
 1 chp7.txt  "\n\nfrom this primitive function has arisen, unquestionably, all t…             633
 2 chp3.txt  "\n\nthe last entry in his diary was made the morning following her…             610
 3 chp8.txt  "\n\nmany moons ago, when he had been much smaller, he had desired …             528
 4 chp1.txt  "\n\nfrom the records of the colonial office and from the dead man'…             523
 5 chp7.txt  "\n\nsquatting upon his haunches on the table top in the cabin his …             507
 6 chp3.txt  "\n\nthe brilliant birds and the little monkeys had become accustom…             492
 7 chp12.txt "\nman's reason\n\n\nthere was one of the tribe of tarzan who quest…             468
 8 chp8.txt  "\n\nand then lord greystoke wiped his greasy fingers upon his nake…             450
 9 chp5.txt  "\n\nthe deep waters of the lake he had been taught by his wild mot…             447
10 chp6.txt  "\n\nhe put the book back in the cupboard and closed the door, for …             441
# ℹ 1,884 more rows
# ℹ Use `print(n = ...)` to see more rows
```

The regular expression above uses ‘.’as a delimiter for identifying the longest sentences, assuming the period symbol is not preceded by any ‘r’ ending titles like, Dr., Mr. etc. This regular expression helps correctly split the text into sentences.

We can see that the longest sentence is from Chapter 7 containing a total length of 633 characters.

Now let us compute the 10 longest words and sentences for each chapter. We follow the same process converting each chapter into tidy corpus.

```
> tidyCORP_chp1 <- tidy(CORP[1])
> tidyCORP_chp1
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp1.txt en       NA     "\nOut to Sea\n\n\nI h…
> CORPwords_chp1<-tidyCORP_chp1 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp1
# A tibble: 3,564 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp1.txt responsibility          14
 2 chp1.txt ridiculousness          14
 3 chp1.txt simultaneously          14
 4 chp1.txt painstakingly           13
 5 chp1.txt acknowledging           13
 6 chp1.txt investigation           13
 7 chp1.txt investigation           13
 8 chp1.txt investigation           13
 9 chp1.txt circumstances           13
10 chp1.txt mortification           13
# ℹ 3,554 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp1<-tidyCORP_chp1 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp1
# A tibble: 139 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp1.txt "\n\nfrom the records of the colonial office and from the dead man's…          523
 2 chp1.txt "  without\nwaiting to rise he whipped a revolver from his pocket, f…          385
 3 chp1.txt "\n\nwhen my convivial host discovered that he had told me so much, …          362
 4 chp1.txt "  what reason could he give the officer commanding\nher majesty's s…          357
 5 chp1.txt "\n\nclayton asked no questions--he did not need to--and the followi…          340
 6 chp1.txt "  late in the afternoon he saw her upper works fade\nbelow the far …          337
 7 chp1.txt "  the captain has brought this\ncondition upon himself, so why then…          302
 8 chp1.txt "\n\non the second day after the wounding of black michael, clayton …          283
 9 chp1.txt "\n\nbut, whereas the captain might easily have been brought to regre…          278
10 chp1.txt "\n\ni do not say the story is true, for i did not witness the happe…          274
# ℹ 129 more rows
> tidyCORP_chp3 <- tidy(CORP[9])
> tidyCORP_chp3
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp3.txt en       NA     "\nLife and Death\n\n\nMorning found them but l…
> CORPwords_chp3<-tidyCORP_chp3 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp3
# A tibble: 2,284 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp3.txt responsibilities        16
 2 chp3.txt simultaneously          14
 3 chp3.txt responsibility          14
 4 chp3.txt comparatively           13
 5 chp3.txt uncomfortable           13
 6 chp3.txt acquaintances           13
 7 chp3.txt consciousness           13
 8 chp3.txt horizontally            12
 9 chp3.txt transversely            12
10 chp3.txt sufficiently            12
# ℹ 2,274 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp3<-tidyCORP_chp3 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp3
# A tibble: 75 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp3.txt "\n\nthe last entry in his diary was made the morning following her death,\nand there he reci…          610
 2 chp3.txt "\n\nthe brilliant birds and the little monkeys had become accustomed to\ntheir new acquainta…          492
 3 chp3.txt "  grown careless\nfrom months of continued safety, during which time he had seen no\ndangero…          401
 4 chp3.txt "\n\nthe door he built of pieces of the packing-boxes which had held their\nbelongings, naili…          305
 5 chp3.txt "\n\nwith a low cry she sprang toward the cabin, and, as she entered, gave a\nbackward glance…          302
 6 chp3.txt "\n\nas soon as they had made their meager breakfast of salt pork, coffee\nand biscuit, clayt…          274
 7 chp3.txt "\n\nthe building of a bed, chairs, table, and shelves was a relatively easy\nmatter, so that…          273
 8 chp3.txt "\n\nhe had strengthened the window protections and fitted a unique wooden\nlock to the cabin…          270
 9 chp3.txt "\n\nthat it would have been beset by worries and apprehension had she been\nin full command …          267
10 chp3.txt "\n\nsometimes she would question clayton as to the strange noises of the\nnights; the absenc…          266
# ℹ 65 more rows
> tidyCORP_chp5 <- tidy(CORP[11])
> tidyCORP_chp5
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp5.txt en       NA     "\nThe White Ape\n\n\nTenderly Kala nursed her …
> CORPwords_chp5<-tidyCORP_chp5 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp5
# A tibble: 2,619 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp5.txt resourcefulness         15
 2 chp5.txt uncomfortable           13
 3 chp5.txt experimenting           13
 4 chp5.txt dissatisfied            12
 5 chp5.txt recollection            12
 6 chp5.txt aristocratic            12
 7 chp5.txt appraisement            12
 8 chp5.txt satisfaction            12
 9 chp5.txt unsuspecting            12
10 chp5.txt unbelievable            12
# ℹ 2,609 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp5<-tidyCORP_chp5 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp5
# A tibble: 95 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp5.txt "\n\nthe deep waters of the lake he had been taught by his wild mother to\navoid, and further…          447
 2 chp5.txt "\n\nso intent was he upon this personal appraisement of his features that\nhe did not hear t…          359
 3 chp5.txt "\"\n\nbut when kerchak spoke to her about it kala threatened to run away from\nthe tribe if …          356
 4 chp5.txt "  it\nwas nearly a year from the time the little fellow came into her\npossession before he …          333
 5 chp5.txt "\n\nrapidly he moved his hands and feet in an attempt to scramble upward,\nand, possibly mor…          331
 6 chp5.txt "  to one less wise the wild alarm\nof her fierce cry as she sprang would have seemed a fooli…          323
 7 chp5.txt "\n\ntarzan on his part never lost an opportunity to show that he fully\nreciprocated his fon…          291
 8 chp5.txt "\n\nhad they known that the child had seen thirteen moons before it had\ncome into kala's po…          285
 9 chp5.txt "\n\nin many ways did he differ from them, and they often marveled at his\nsuperior cunning, …          280
10 chp5.txt "\n\nif he could catch his fellow apes with his long arm of many grasses,\nwhy not sabor, the…          257
# ℹ 85 more rows
> tidyCORP_chp7 <- tidy(CORP[13])
> tidyCORP_chp7
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp7.txt en       NA     "\nThe Light of Knowledge\n\n\nAfter what seeme…
> CORPwords_chp7<-tidyCORP_chp7 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp7
# A tibble: 3,942 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp7.txt investigations          14
 2 chp7.txt establishments          14
 3 chp7.txt unquestionably          14
 4 chp7.txt convalescence           13
 5 chp7.txt investigation           13
 6 chp7.txt comparatively           13
 7 chp7.txt possibilities           13
 8 chp7.txt experimenting           13
 9 chp7.txt inexhaustible           13
10 chp7.txt remorselessly           13
# ℹ 3,932 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp7<-tidyCORP_chp7 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp7
# A tibble: 133 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp7.txt "\n\nfrom this primitive function has arisen, unquestionably, all the forms\nand ceremonials …          633
 2 chp7.txt "\n\nsquatting upon his haunches on the table top in the cabin his father\nhad built--his smo…          507
 3 chp7.txt "\n\nhe did not accomplish it in a day, or in a week, or in a month, or in a\nyear; but slowl…          363
 4 chp7.txt "\n\ntublat was his most consistent enemy, but it was through tublat that,\nwhen he was about…          362
 5 chp7.txt "\n\nat last he reached the fast disappearing feast and with his sharp knife\nslashed off a m…          344
 6 chp7.txt "\n\nfor hours absolute quiet reigned in the little clearing, except as it\nwas broken by the…          343
 7 chp7.txt "\n\nhence the younger males as they became adult found it more comfortable\nto take mates fr…          335
 8 chp7.txt "  descended from a\nrace of meat eaters, never in his life, he thought, had he once\nsatisfi…          323
 9 chp7.txt "\n\nmany travelers have seen the drums of the great apes, and some have\nheard the sounds of…          314
10 chp7.txt "\n\nand so he progressed very, very slowly, for it was a hard and laborious\ntask which he h…          306
# ℹ 123 more rows
```

```
> tidyCORP_chp2 <- tidy(CORP[8])
> tidyCORP_chp2
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp2.txt en       NA     "\nThe Savage Home\n\n\nNor did they have long …
> CORPwords_chp2<-tidyCORP_chp2 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp2
# A tibble: 3,177 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp2.txt notwithstanding         15
 2 chp2.txt transportation          14
 3 chp2.txt remorselessly           13
 4 chp2.txt heartlessness           13
 5 chp2.txt threateningly           13
 6 chp2.txt investigation           13
 7 chp2.txt contemplation           13
 8 chp2.txt circumstances           13
 9 chp2.txt forgetfulness           13
10 chp2.txt companionway            12
# ℹ 3,167 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp2<-tidyCORP_chp2 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp2
# A tibble: 120 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp2.txt "\n\nnotwithstanding that both clayton and his wife kept a sharp lookout\nthey saw nothing of…          401
 2 chp2.txt "\n\nto this end clayton selected four trees which formed a rectangle about\neight feet squar…          362
 3 chp2.txt "\n\n\"what they did may we not do?  and even better, for are we not armed\nwith ages of supe…          352
 4 chp2.txt "\n\nno signs of habitation were visible, but that the land might easily\nsupport human life …          320
 5 chp2.txt "\n\nclayton did not believe that black michael had the slightest intention\nof notifying the…          318
 6 chp2.txt "\n\nduring the long hours of darkness they caught but fitful snatches of\nsleep, for the nig…          318
 7 chp2.txt "\"\n\n\"but john, if it were only you and i,\" she sobbed, \"we could endure it\ni know; but…          315
 8 chp2.txt "\n\na hundred yards from the beach was a little level spot, fairly free of\ntrees; here they…          302
 9 chp2.txt "  for\nan hour or more they heard it sniffing and clawing at the trees which\nsupported thei…          263
10 chp2.txt "\n\nall during the day the forest about them had been filled with excited\nbirds of brillian…          260
# ℹ 110 more rows
```

```
> tidyCORP_chp4 <- tidy(CORP[10])
> tidyCORP_chp4
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp4.txt en       NA     "\nThe Apes\n\n\nIn the forest of the table-lan…
> CORPwords_chp4<-tidyCORP_chp4 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp4
# A tibble: 2,516 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp4.txt notwithstanding         15
 2 chp4.txt investigations          14
 3 chp4.txt understanding           13
 4 chp4.txt precipitately           13
 5 chp4.txt uncontrolled            12
 6 chp4.txt successfully            12
 7 chp4.txt intelligence            12
 8 chp4.txt intelligence            12
 9 chp4.txt occasionally            12
10 chp4.txt fearlessness            12
# ℹ 2,506 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp4<-tidyCORP_chp4 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp4
# A tibble: 82 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp4.txt "\n\nwhen kerchak came to a halt a short distance from the cabin and\ndiscovered that he stil…          344
 2 chp4.txt "\n\nhigh up among the branches of a mighty tree she hugged the shrieking\ninfant to her boso…          333
 3 chp4.txt "\n\nwhen the king ape released the limp form which had been john clayton,\nlord greystoke, h…          332
 4 chp4.txt "\n\nhe had seen many of his kind go to their deaths before the loud noise\nmade by the littl…          302
 5 chp4.txt "\n\nusing his long arms as a man uses crutches, and rolling his huge\ncarcass from side to s…          286
 6 chp4.txt "\n\nthe sight that met his eyes must have frozen him with horror, for\nthere, within the doo…          273
 7 chp4.txt "\n\ncautiously he approached the thing, ready to flee precipitately should\nit speak in its …          268
 8 chp4.txt "\n\nbut kerchak was close upon her, so close that he had almost grasped her\nankle had she n…          267
 9 chp4.txt "\n\nshe made the leap successfully, but as she grasped the limb of the\nfurther tree the sud…          262
10 chp4.txt "\n\nthe younger and lighter members of his tribe scampered to the higher\nbranches of the gr…          258
# ℹ 72 more rows
```

```
> tidyCORP_chp6 <- tidy(CORP[12])
> tidyCORP_chp6
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp6.txt en       NA     "\nJungle Battles\n\n\nThe wanderings of the tr…
> CORPwords_chp6<-tidyCORP_chp6 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp6
# A tibble: 2,341 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp6.txt impossibility           13
 2 chp6.txt thousandfold            12
 3 chp6.txt explorations            12
 4 chp6.txt introduction            12
 5 chp6.txt exhilaration            12
 6 chp6.txt accidentally            12
 7 chp6.txt excruciating            12
 8 chp6.txt imagination             11
 9 chp6.txt independent             11
10 chp6.txt combination             11
# ℹ 2,331 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp6<-tidyCORP_chp6 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp6
# A tibble: 79 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp6.txt "\n\nhe put the book back in the cupboard and closed the door, for he did\nnot wish anyone el…          441
 2 chp6.txt "\n\na mile back in the forest the tribe had heard the fierce challenge of\nthe gorilla, and,…          348
 3 chp6.txt "\n\nit was in the next visit to the vicinity, following the adventure with\nold sabor, that,…          345
 4 chp6.txt "\n\nthat her little tarzan could destroy a great bull gorilla she knew to\nbe improbable, an…          344
 5 chp6.txt "\n\nlike some huge phantom, kala swung noiselessly from tree to tree; now\nrunning nimbly al…          333
 6 chp6.txt "\n\nthe boats, and trains, and cows and horses were quite meaningless to\nhim, but not quite…          316
 7 chp6.txt "\n\nhe was alone, as was often the case when he visited the cabin, for the\napes had no love…          297
 8 chp6.txt "\n\nthe language of the apes had so few words that they could talk but\nlittle of what they …          297
 9 chp6.txt "\n\nkala, alone, he was glad to have with him, but now that he was better\nshe was gone long…          292
10 chp6.txt "  more and more weakly the torn and bleeding arm struck home\nwith the long sharp blade, the…          269
# ℹ 69 more rows
```

```
> tidyCORP_chp8 <- tidy(CORP[14])
> tidyCORP_chp8
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp8.txt en       NA     "\nThe Tree-top Hunter\n\n\nThe morning after t…
> CORPwords_chp8<-tidyCORP_chp8 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp8
# A tibble: 7,269 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp8.txt destructiveness         15
 2 chp8.txt sentimentalist          14
 3 chp8.txt uncomfortable           13
 4 chp8.txt possibilities           13
 5 chp8.txt insignificant           13
 6 chp8.txt consternation           13
 7 chp8.txt noiselessness           13
 8 chp8.txt transgressing           13
 9 chp8.txt gesticulating           13
10 chp8.txt superstitious           13
# ℹ 7,259 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp8<-tidyCORP_chp8 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp8
# A tibble: 327 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp8.txt "\n\nmany moons ago, when he had been much smaller, he had desired the skin\nof sabor, the li…          528
 2 chp8.txt "\n\nand then lord greystoke wiped his greasy fingers upon his naked thighs\nand took up the …          450
 3 chp8.txt "  the finding of the still warm body of kulonga--on the very\nverge of their fields and with…          426
 4 chp8.txt "\n\nthat they more greatly feared an attack from the rear than whatever\nunknown enemies lur…          400
 5 chp8.txt "\n\nalso could he write, with printed letters, rapidly and plainly, but\nscript he had not m…          310
 6 chp8.txt "\n\nso it was that as kulonga emerged from the shadow of the jungle a\nslender coil of rope …          310
 7 chp8.txt "\n\nfor several months the tribe hovered near the beach where stood\ntarzan's cabin, and his…          300
 8 chp8.txt "\n\nhow may we judge him, by what standards, this ape-man with the heart\nand head and body …          300
 9 chp8.txt "\n\nthus far his plan had worked to perfection, but when he grasped the\nrope, bracing himse…          295
10 chp8.txt "\n\nat last tarzan tired of the sport, and with a parting roar of challenge\nand a well-aime…          290
# ℹ 317 more rows
```

```
> tidyCORP_chp9 <- tidy(CORP[15])
> tidyCORP_chp9
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp9.txt en       NA     "\nChapter XI\n\n\n\n\nthe leafy fastness of th…
> CORPwords_chp9<-tidyCORP_chp9 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp9
# A tibble: 1,784 × 3
   id        word            word_length
   <chr>     <chr>                 <int>
 1 chp9.txt destructiveness          15
 2 chp9.txt sentimentalist           14
 3 chp9.txt superstitious            13
 4 chp9.txt superstitious            13
 5 chp9.txt gesticulating            13
 6 chp9.txt superstition             12
 7 chp9.txt explanations             12
 8 chp9.txt sufficiently             12
 9 chp9.txt depredations             12
10 chp9.txt occasionally             12
# ℹ 1,774 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp9<-tidyCORP_chp9 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp9
# A tibble: 90 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp9.txt "\ndismay, and conjured in their poor brains only the most frightful of\nvillage, within the …       345
 2 chp9.txt "   then he turned to watch the proceeding\njust as the first of the returning natives entered…       310
 3 chp9.txt "\nvillage to prevent his taking a supply of the arrows which lay below\nthat interested him …       267
 4 chp9.txt "\nsaw her plunge the member into a vessel of water and quickly rub the\nhands, and once when…       251
 5 chp9.txt "\ncreatures to kill senselessly and wantonly for the mere pleasure of\nwhich no other animal…       249
 6 chp9.txt "\nof his own kind, and to learn more of their ways and view the strange\nspot he made his wa…       237
 7 chp9.txt "   he determined to own one of these long, pointed sticks, but\nthen, for he \"saw\" largely …       220
 8 chp9.txt "   he proceeded with unwonted stealth, for kulonga had taught\nhe was quite prepared either t…       217
 9 chp9.txt "\nmen were flying through the little village searching minutely every hut\nhe spoke a few wo…       213
10 chp9.txt "\nthat would be accorded him should he be discovered by these, the first\nsimilarity of form…       210
# ℹ 80 more rows
> tidyCORP_chp11 <- tidy(CORP[3])
> tidyCORP_chp11
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp11.txt en       NA     "\n\"King of the Apes\"\n\n\nIt was not yet da…
> CORPwords_chp11<-tidyCORP_chp11 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp11
# A tibble: 3,526 × 3
   id        word             word_length
   <chr>     <chr>                  <int>
 1 chp11.txt incomprehensible          16
 2 chp11.txt experimentation           15
 3 chp11.txt capriciousness            14
 4 chp11.txt investigation             13
 5 chp11.txt investigation             13
 6 chp11.txt ornamentation             13
 7 chp11.txt determination             13
 8 chp11.txt determination             13
 9 chp11.txt loathsomeness             13
10 chp11.txt reconnoitered             13
# ℹ 3,516 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp11<-tidyCORP_chp11 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp11
# A tibble: 148 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp11.txt "\n\nwhen he had followed kulonga through the forest he had expected to come\nto a city of s…       328
 2 chp11.txt "   he was too shrewd to allow those long hairy arms to encircle\nhim, and just as their bodi…       291
 3 chp11.txt "\n\nthe ape was staggered, and what with the mortal wound in his side had\nalmost collapsed…       272
 4 chp11.txt "\n\"king of the apes\"\n\n\nit was not yet dark when he reached the tribe, though he stoppe…       258
 5 chp11.txt "\nwithdrawing the knife that had so often rendered him master of far\nmightier muscles th…       256
 6 chp11.txt "\n\nlater, when they discovered the overturned cauldron, and that once more\ntheir arrows h…       251
 7 chp11.txt "\n\nhis bow and arrows lay some distance away where he had dropped them\nwhile showing sabo…       251
 8 chp11.txt "\n\ntarzan replaced the box in the cupboard, but always thereafter he\ncarried the features…       249
 9 chp11.txt "\n\nthe greater strength of the ape was slowly prevailing, and the teeth of\nthe straining …       241
10 chp11.txt "   only a few berries and an occasional grub worm rewarded his\nsearch, and he was half fami…       240
# ℹ 138 more rows
> tidyCORP_chp13 <- tidy(CORP[5])
> tidyCORP_chp13
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp13.txt en       NA     "\nHis Own Kind\n\n\nThe following morning, Ta…
> CORPwords_chp13<-tidyCORP_chp13 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp13
# A tibble: 4,731 × 3
   id        word            word_length
   <chr>     <chr>                 <int>
 1 chp13.txt personification          15
 2 chp13.txt distinguishing           14
 3 chp13.txt simultaneously           14
 4 chp13.txt convalescence            13
 5 chp13.txt reverberation            13
 6 chp13.txt gesticulating            13
 7 chp13.txt gesticulating            13
 8 chp13.txt comparatively            13
 9 chp13.txt eventualities            13
10 chp13.txt intelligence             12
# ℹ 4,721 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp13<-tidyCORP_chp13 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp13
# A tibble: 209 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp13.txt "   adjusting his spectacles\nhe looked for a moment at the placard and then, turning away, s…       363
 2 chp13.txt "   we are absolutely in your power, but\nyou'll treat professor porter and miss porter with …       344
 3 chp13.txt "\n\ntarzan had never seen the effects of a firearm before, though his books\nhad taught him…       341
 4 chp13.txt "\"\n\n\"tut, tut, child; tut, tut,\" responded professor porter, in a kindly\nand indulgent…       337
 5 chp13.txt "\n\nfor a moment longer tarzan watched while a number of boxes and bundles\nwere lowered in…       331
 6 chp13.txt "\n\nthen he determined to filch what few garments he could from one of the\nblack men of mb…       301
 7 chp13.txt "\n\nit took a dozen attempts on the part of the thoroughly frightened\nblacks to reenter th…       299
 8 chp13.txt "\n\nas long as they supplied him with arrows and food he would not harm\nthem unless they l…       279
 9 chp13.txt "   with this idea in mind she turned to search for some means of\nputting it into execution;…       268
10 chp13.txt "\n\n\"ah, yes, indeed; yes indeed--most remarkable, most remarkable!\" said\nthe professor;…       264
# ℹ 199 more rows
> tidyCORP_chp15 <- tidy(CORP[7])
> tidyCORP_chp15
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp15.txt en       NA     "\nThe Forest God\n\n\nWhen Clayton heard the …
> CORPwords_chp15<-tidyCORP_chp15 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp15
# A tibble: 1,593 × 3
   id        word           word_length
   <chr>     <chr>                <int>
 1 chp15.txt irretrievably         13
 2 chp15.txt apprehension          12
 3 chp15.txt satisfaction          12
 4 chp15.txt occasionally          12
 5 chp15.txt disappearing          12
 6 chp15.txt overwrought           11
 7 chp15.txt appreciably           11
 8 chp15.txt neighboring           11
 9 chp15.txt comparative           11
10 chp15.txt maintenance           11
# ℹ 1,583 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp15<-tidyCORP_chp15 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp15
# A tibble: 58 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp15.txt "\n\nwhat were the thoughts of his strange captor or guide clayton could\nonly vaguely conje…       387
 2 chp15.txt "\n\nhe was trying to tell the stupid white man to plunge his poisoned\narrows into sabor's …       383
 3 chp15.txt "\n\nfrom the first sensation of chilling fear clayton passed to one of keen\nadmiration and…       306
 4 chp15.txt "\n\nas clayton rounded the corner of the cabin to behold the animal\ndisappearing within, i…       302
 5 chp15.txt "   tarzan, racking his brains\nfor some means to cope single-handed with the infuriated beas…       301
 6 chp15.txt "\n\nfrom one lofty branch the agile creature swung with clayton through a\ndizzy arc to a n…       267
 7 chp15.txt "\n\nfor a moment tarzan looked at the young man closely, as though\nundecided as to just wh…       251
 8 chp15.txt "\n\npawing and tearing at earth and air, sabor rolled and threw herself\nthis way and that …       251
 9 chp15.txt "   he knew that one of the sailors might be the\nauthor of it; but the fact that he had left…       243
10 chp15.txt "\n\nwith the quickness of a striking rattler he launched himself full upon\nsabor's back, h…       243
# ℹ 48 more rows
```

```
> tidyCORP_chp10 <- tidy(CORP[2])
> tidyCORP_chp10
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp10.txt en       NA     "\nThe Fear-Phantom\n\n\nFrom a lofty perch Ta…
> CORPwords_chp10<-tidyCORP_chp10 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp10
# A tibble: 1,780 × 3
   id        word            word_length
   <chr>     <chr>                 <int>
 1 chp10.txt destructiveness          15
 2 chp10.txt sentimentalist           14
 3 chp10.txt gesticulating            13
 4 chp10.txt superstitious            13
 5 chp10.txt superstitious            13
 6 chp10.txt bloodthirsty             12
 7 chp10.txt businesslike             12
 8 chp10.txt impenetrable             12
 9 chp10.txt occasionally             12
10 chp10.txt depredations             12
# ℹ 1,770 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp10<-tidyCORP_chp10 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp10
# A tibble: 90 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp10.txt "   the finding of the still warm body of kulonga--on the very\nverge of their fields and wit…       426
 2 chp10.txt "   he noted the extreme\ncare which the woman took that none of the matter should touch her\…       268
 3 chp10.txt "   he killed for food most\noften, but, being a man, he sometimes killed for pleasure, a thi…       263
 4 chp10.txt "\n\nhe saw that at one point the forest touched the village, and to this\nspot he made his …       236
 5 chp10.txt "\ngathering up all he could carry under one arm, he overturned the\nseething cauldron with …       232
 6 chp10.txt "\n\ntarzan of the apes knew that they had found the body of his victim, but\nthat intereste…       217
 7 chp10.txt "   few were his primitive pleasures, but the\ngreatest of these was to hunt and kill, and so…       214
 8 chp10.txt "   dozing in the shade he saw several men, while at the\nextreme outskirts of the clearing h…       211
 9 chp10.txt "\n\ntarzan knew nothing of poison, but his shrewd reasoning told him that\nit was this dead…       210
10 chp10.txt "   on one side of her lay a\nquantity of wooden arrows the points of which she dipped into t…       190
# ℹ 80 more rows
> tidyCORP_chp12 <- tidy(CORP[4])
> tidyCORP_chp12
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp12.txt en       NA     "\nMan's Reason\n\n\nThere was one of the trib…
> CORPwords_chp12<-tidyCORP_chp12 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp12
# A tibble: 2,761 × 3
   id        word            word_length
   <chr>     <chr>                 <int>
 1 chp12.txt accomplishment          14
 2 chp12.txt manifestation           13
 3 chp12.txt disobediences           13
 4 chp12.txt consternation           13
 5 chp12.txt superstitious           13
 6 chp12.txt disappearance           13
 7 chp12.txt instinctively           13
 8 chp12.txt dismemberment           13
 9 chp12.txt companionship           13
10 chp12.txt circumstances           13
# ℹ 2,751 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp12<-tidyCORP_chp12 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp12
# A tibble: 106 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp12.txt "\nman's reason\n\n\nthere was one of the tribe of tarzan who questioned his authority, and\…       468
 2 chp12.txt "\n\nbut when, finally, he realized that his antagonist was fastened to him\nwhere his teeth…       405
 3 chp12.txt "   but when one day a\nmember of the tribe discovered the blacks in great numbers on the ban…       372
 4 chp12.txt "   the periodic disappearance of their arrows, and the\nstrange pranks perpetrated by unseen…       317
 5 chp12.txt "\n\nhad kala lived, tarzan would have sacrificed all else to remain near\nher, but now that…       312
 6 chp12.txt "   so much had his great strength and agility increased\nin the period following his maturit…       301
 7 chp12.txt "\n\nin the sum total of their points, however, the anthropoid had a shade\nthe better of th…       300
 8 chp12.txt "\nbut this method he realized to be too fraught with danger, and so he\ncommenced picking u…       281
 9 chp12.txt "\n\nduring the rollings and squirmings of the next few minutes, tarzan's\nhold was loosened…       273
10 chp12.txt "\n\nwhen the awe-struck savages saw that the food disappeared overnight\nthey were filled w…       268
# ℹ 96 more rows
> tidyCORP_chp14 <- tidy(CORP[6])
> tidyCORP_chp14
# A tibble: 1 × 8
  author datetimestamp        description heading id        language origin text
  <lgl>  <dttm>                <lgl>       <lgl>   <chr>     <chr>    <lgl>  <chr>
1 NA     2024-05-04 20:54:09 NA          NA      chp14.txt en       NA     "\nAt the Mercy of the Jungle\n\n\nAfter Clayt…
> CORPwords_chp14<-tidyCORP_chp14 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length=nchar(word)) %>% arrange(desc(word_length))
> CORPwords_chp14
# A tibble: 3,300 × 3
   id        word            word_length
   <chr>     <chr>                 <int>
 1 chp14.txt comprehensible          14
 2 chp14.txt simultaneously          14
 3 chp14.txt insignificance          14
 4 chp14.txt bloodcurdling           13
 5 chp14.txt complimenting           13
 6 chp14.txt disparagement           13
 7 chp14.txt instinctively           13
 8 chp14.txt perspiration            12
 9 chp14.txt disheartened            12
10 chp14.txt unaccustomed            12
# ℹ 3,290 more rows
# ℹ Use `print(n = ...)` to see more rows
> CORPSentences_chp14<-tidyCORP_chp14 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, sentence) %>% mutate(sentence_length=nchar(sentence)) %>% arrange(desc(sentence_length))
> CORPSentences_chp14
# A tibble: 143 × 3
   id        sentence                                                    sentence_length
   <chr>     <chr>                                                                <int>
 1 chp14.txt "\n\nfor a moment the lioness remained quiet with intense gaze directed upon\nthe flitting e…       391
 2 chp14.txt "\n\nthe englishman, finally concluding that he was a prisoner, saw no\nalternative open but…       388
 3 chp14.txt "\n\nthe afternoon was drawing to a close, and clayton, disheartened and\ndiscouraged, was i…       384
 4 chp14.txt "\n\nthe strong white teeth sank into the raw and dripping flesh in apparent\nrelish of the …       353
 5 chp14.txt "\n\nclayton spoke to the stranger in english, thanking him for his brave\nrescue and compli…       339
 6 chp14.txt "\n\nthe noise of some great body crashing through the underbrush so close\nbeside him, and …       334
 7 chp14.txt "\n\nclayton, bewildered and confused, hesitated to follow him, for he\nthought he was but b…       320
 8 chp14.txt "\nat the mercy of the jungle\n\nafter clayton had plunged into the jungle, the sailors--a…       314
 9 chp14.txt "\nshe feared not more for herself than for the three men whom she knew to\nbe wandering in …       291
10 chp14.txt "\n\nto tarzan this was scarcely comprehensible; his judgment told him\nthat no man would ve…       263
# ℹ 133 more rows
```

CORPUS Cleansing

Next, the data undergoes a cleaning process to eliminate stop words, numbers, and punctuation marks. Since our text is already in lowercase, we begin by removing quotes from the corpus using the gsub() function.

```
> CORP<-tm_map(CORP, content_transformer(gsub), pattern="'", replacement="")
```

Following the removal of quotes, we define a function called "Remove_Num_Punct()" specifically designed to eliminate numbers and punctuation from the text. This function can be applied to the CORP corpus, resulting in the creation of our clean corpus named "CORP_clean."

```
> Remove_Num_Punct<-function(x) gsub("[^[:alpha:][:space:]]*","",x)
> CORP_clean <- tm::tm_map(CORP, content_transformer(Remove_Num_Punct))
> CORP_clean
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15
> str(CORP_clean)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 15
  ..$ :List of 2
  .. ..$ content: chr [1:420] "" "Out to Sea" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:215] "" "The FearPhantom" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:427] "" "King of the Apes" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp11.txt"
```

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

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

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

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

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

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

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

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

[[8]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 17017

[[9]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 12204

[[10]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13355

[[11]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13961

[[12]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 12301

[[13]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 21204

[[14]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 39060

[[15]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9669
```

We can see that we removed 515, 202, 947, 360, 702, 442, 208, 436, 267, 280, 310, 253, 449, 855, 202 characters from respective documents. Now, let us make sure that we convert our corpus into lowercase using tm_map() function.

```
> CORP_low<-tm_map(CORP_clean, tm::content_transformer(tolower))
> CORP_low
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15
> str(CORP_low)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 15
  ..$ :List of 2
  .. ..$ content: chr [1:420] "" "out to sea" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:215] "" "the fearphantom" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:427] "" "king of the apes" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2024-05-04 20:54:09"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "chp11.txt"
  .. .. ..$ language     : chr "en"
```

The punctuations, numbers, quotes have been removed and everything is in lower case. Now, lets perform the DTM for the cleaned corpus.

```
> CORP_DTM<-DocumentTermMatrix(CORP_low)
> CORP_DTM
<<DocumentTermMatrix (documents: 15, terms: 5425)>>
Non-/sparse entries: 14724/66651
Sparsity           : 82%
Maximal term length: 20
Weighting          : term frequency (tf)
> str(CORP_DTM)
List of 6
 $ i       : int [1:14724] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:14724] 7 8 10 11 25 26 36 37 44 45 ...
 $ v       : num [1:14724] 1 2 3 2 1 1 1 1 1 1 ...
 $ nrow    : int 15
 $ ncol    : int 5425
 $ dimnames:List of 2
  ..$ Docs : chr [1:15] "chp1.txt" "chp10.txt" "chp11.txt" "chp12.txt" ...
  ..$ Terms: chr [1:5425] "abandon" "abandoned" "abandoning" "abashed" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

```
> inspect(CORP_DTM)
<<DocumentTermMatrix (documents: 15, terms: 5425)>>
Non-/sparse entries: 14724/66651
Sparsity           : 82%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
           Terms
Docs        and but for had him his that the was with
  chp1.txt  104  33  40  34  17  54   56 240  57   32
  chp11.txt 130  21  36  42  23  90   24 328  43   45
  chp12.txt  91  21  29  45  21  67   39 199  42   21
  chp13.txt 163  25  21  55  28 105   34 367  55   39
  chp14.txt 107  26  28  35  25  39   40 325  62   26
  chp2.txt  104  31  32  27   5  26   36 213  32   30
  chp4.txt   75  17  20  33  11  49   29 200  45   29
  chp5.txt   96  23  29  30  26  54   31 161  50   24
  chp7.txt  128  27  30  37  22  93   36 373  56   37
  chp8.txt  273  52  62  86  43 159   61 580 115   59
```

We can see that the sparsity has been reduced by 3% from 85 to 82. Although it is still high. So, we need to employ more methods to clean our corpus.

Now, let us remove stop words from the corpus.

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


  sea


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

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

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

 yellow mildewed pages   diary   man long dead
records   colonial office dovetail perfectly   narrative
 convivial host    give   story   painstakingly
pieced     several various agencies

    find  credible  will  least   one
acknowledging    unique remarkable  interesting

  records   colonial office    dead mans diary
 learn   certain young english nobleman   shall call john
clayton lord greystoke  commissioned  make  peculiarly delicate
investigation  conditions   british west coast african colony
whose simple native inhabitants another european power  known
recruiting soldiers   native army   used solely
```

In this step, we have utilized the tm package to remove various stop words from our corpus. By inspecting the first document, we can gain an understanding of the appearance of our corpus after the removal of stop words. As observed, the stop words have been successfully eliminated, enhancing the analyzability of our corpus.

Now let us create the Term Document Matrix again.

```
> CORP_StopWords_TDM<-tm::TermDocumentMatrix(CORP_StopWords)
> CORP_StopWords_TDM
<<TermDocumentMatrix (terms: 5327, documents: 15)>>
Non-/sparse entries: 13583/66322
Sparsity           : 83%
Maximal term length: 20
Weighting          : term frequency (tf)
>
```

We can also find the frequent terms using findFreqTerms with a low freq of 5.

```
> Freq_Terms<-tm::findFreqTerms(CORP_StopWords_TDM, lowfreq = 5)
> Freq_Terms
   [1] "able"          "accompanied"   "accord"        "accorded"      "accustomed"    "across"        "act"
   [8] "added"         "advantage"     "adventure"     "africa"        "african"       "afternoon"     "agility"
  [15] "air"           "alarm"         "alice"         "alive"         "almost"        "alone"         "along"
  [22] "aloud"         "already"       "also"          "always"        "among"         "amphitheater"  "ancient"
  [29] "anger"         "animal"        "animals"       "another"       "answer"        "answered"      "antagonist"
  [36] "anthropoids"   "ape"           "apeman"        "apes"          "apparent"      "apparently"    "appeared"
  [43] "approached"    "approaching"   "archer"        "arm"           "armed"         "arms"          "around"
  [50] "arrow"         "arrows"        "article"       "asked"         "attack"        "attacking"     "attempt"
  [57] "attempted"     "attempting"    "attention"     "aught"         "authority"     "away"          "awful"
  [64] "babe"          "baby"          "back"          "bad"           "balance"       "band"          "bared"
  [71] "battle"        "beach"         "bearing"       "beast"         "beasts"        "beat"          "beating"
  [78] "beautiful"     "became"        "become"        "bed"           "began"         "behind"        "belly"
  [85] "belongings"    "belt"          "beneath"       "bent"          "beside"        "best"          "betokened"
  [92] "better"        "beyond"        "bird"          "birds"         "black"         "blackness"     "blacks"
  [99] "blade"         "bloodthirsty"  "bloody"        "blow"          "blows"         "boar"          "boat"
 [106] "boats"         "bodies"        "body"          "bones"         "book"          "books"         "bore"
 [113] "born"          "bosom"         "bow"           "box"           "boxes"         "boy"           "brain"
 [120] "branch"        "branches"      "breast"        "bright"        "brilliant"     "bring"         "british"
 [127] "broke"         "broken"        "brothers"      "brought"       "brown"         "brute"         "brutes"
 [134] "bugs"          "building"      "built"         "bull"          "bullet"        "cabin"         "call"
 [141] "called"        "came"          "can"           "captain"       "carcass"       "care"          "carried"
 [148] "carry"         "cast"          "caught"        "cauldron"      "caused"        "cautiously"    "ceased"
 [155] "center"        "chain"         "challenge"     "chance"        "change"        "charge"        "charged"
 [162] "chattering"    "chest"         "child"         "children"      "circle"        "clawing"       "clay"
 [169] "clayton"       "claytons"      "clear"         "clearing"      "close"         "closed"        "closely"
```

```
> length(Freq_Terms)
[1] 1114
```

```
> Freq_Terms[15]
[1] "air"
```

```
> nchar(Freq_Terms[15])
[1] 3
```

We can see that there are 1114 words with low freq 5.
We can also check a specific term and the length of that term using nchar() function.

Now, we will unlist tfList and combine all term frequencies.

```
> tfList<-list()
> for(i in seq_along(CORP_StopWords)){
+   tfList[[i]]<-termFreq(CORP_StopWords[[i]])
+ }
> print(tfList[[1]])
         able        aboard    accentuated        accept       accord      accorded   achievement  acknowledging
            1             2              1             1            1             1             1              1
          act       advised       affairs        afford       africa       african           aft      afternoon
            1             1              1             1            4             1             1              1
     agencies   aggregation          agin           ago        alice         almost         alone          along
            1             1              1             1           11             4             1              2
         also       amazed       ambition    ammunition        among        amount        anchor        ancient
            1             1              1             2            1             1             1              1
        angry       another       answered     antagonize       anxious      anything     apartment       appalled
            1             2              1             2            1             1             1              2
     apparent     appointed    appointment     appreciate          arf     arguments           arm          army
            2             1              2             1            1             1             2              3
       around       arrived        aruwimi         asayin          ask        asked           ass      assisting
            1             1              1             1            2             4             1              1
    associate       assumed       atlantic     atmosphere      attempt     attempted     attribute         aught
            1             1              2             1            1             2             1              2
        aunts     authority        average        averted        avoid          away          awful           back
            1             1              1             1            1             3             1              1
    backwards           bad           bags       balance    barkentine  battlefieldsa    battleship           bear
            1             2              1             1            1             1             2              2
      beastly        became           beds     beginning        begun        behind       belaying         belief
            1             1              1             1            1             2             2              1
     believed     belongings       beneath          best     betrayed        better           big       billings
            1             1              1             2            1             1             1              2
          bit         black         blacks         blank      blasted        bloody       bloomin          blow
            2             7              1             1            1             1             1              2
        board          body           bore     borrowing        boxes       brasses        breath         bright
            2             1              1             1            1             1             1              1
      british      brothers       brought        brutal     brutality         brute          bull         bullet
            9             1              3             2            2             2             1              1
     bullied       bullies          bully      business       busted         cabin    calculated           call
            1             1              1             2            1             4             1              1
         came           can         candor       captain     captains       capting          care         career
            3             3              1            22            1             1             3              1
     careful    carelessly       carriage       carried        carry        caused      centered        certain
            1             1              1             2            1             2             1              1
```

Now, let us inspect the TDM again for the cleaned Corpus with no Stop Words.

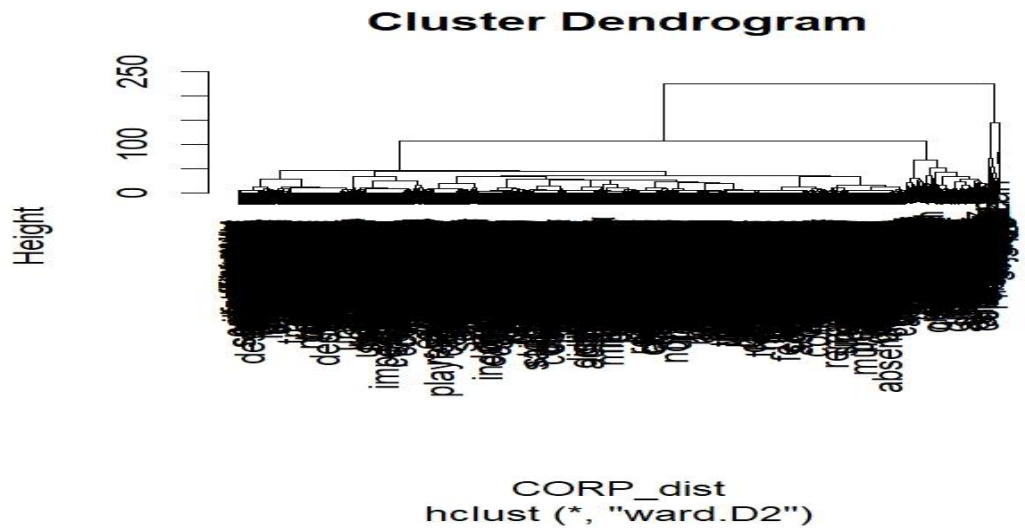
```
> tm::inspect(CORP_StopWords_TDM)
<<TermDocumentMatrix (terms: 5327, documents: 15)>>
Non-/sparse entries: 13583/66322
Sparsity           : 83%
Maximal term length: 20
Weighting          : term frequency (tf)
Sample             :
         Docs
Terms     chp1.txt chp11.txt chp12.txt chp13.txt chp14.txt chp2.txt chp4.txt chp5.txt chp7.txt chp8.txt
  apes           0        18        11         6         2        0        9       12       12       21
  clayton       29         1         0        10        24       18        5        0        0        0
  great          4        18        14         9        17        8       11       12       15       28
  jungle         0         6         5        20        16        3        2        6       14       25
  little        12        12        10        17         3       11       16       24       28       19
  man            8         5         7        31        19       10        3        1        2       13
  now            7         8         5         6        13        5        5        5       11       10
  one           11        15         9        19        10        2       10        8       22       25
  tarzan         0        39        43        33        14        0        0       20       22       67
  upon          11        20        11        31        20       15       14       13       22       42
```

Dendogram

A dendrogram in R is a graphical representation of hierarchical clustering. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.

To create a dendrogram in R, you typically use the hclust() function to perform hierarchical clustering on your data and then visualize the clustering using the plot() function with the hclust object as input.

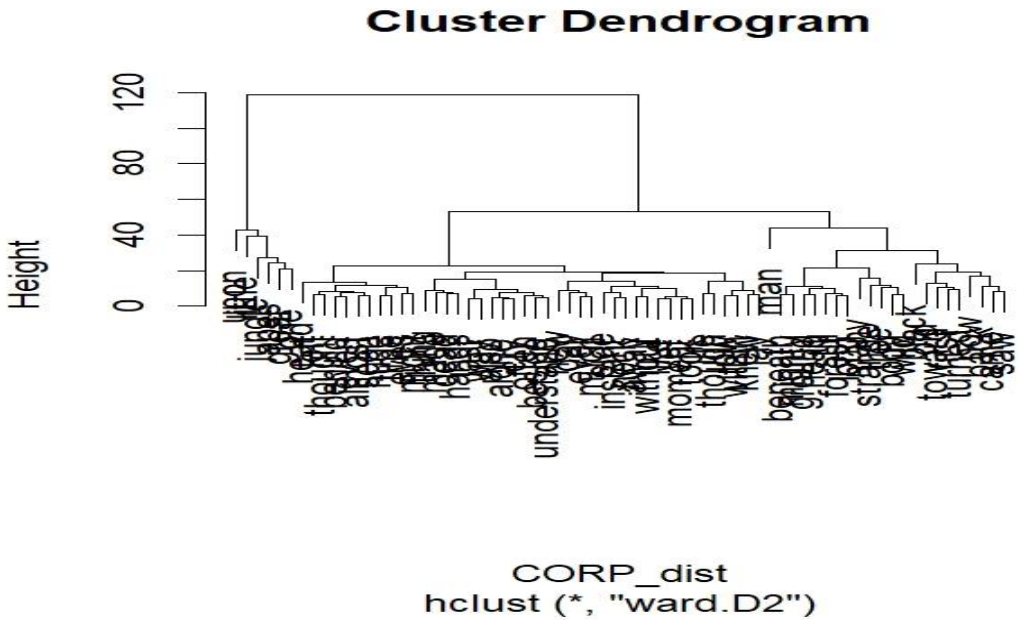
```
> CORP_df <- as.data.frame(as.matrix(CORP_StopWords_TDM))
> CORP_dist <- dist(CORP_df)
> CORP_dg <- hclust(CORP_dist, method ='ward.D2')
> str(CORP_dg)
List of 7
 $ merge      : int [1:5326, 1:2] -1 -1481 -1981 -3107 -3157 -3769 -4347 -4497 -5043 -5157 ...
 $ height     : num [1:5326] 0 0 0 0 0 0 0 0 0 0 ...
 $ order      : int [1:5327] 1950 1960 4532 5279 4149 1078 1645 4871 428 557 ...
 $ labels     : chr [1:5327] "abandon" "abandoned" "abandoning" "abashed" ...
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = CORP_dist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> plot(CORP_dg)
```



We noticed that our dendrogram has many labels, making analysis difficult. To solve this, we applied the removeSparseTerms() function from the tm package to eliminate sparse terms from the term-document matrix. After experimenting with various maximal allowed sparsity values, we found that a value of 0.2 produced a comprehensive dendrogram for analysis, resulting in a sparsity level of 7%.
We do not prefer 0% sparsity as that would lose a great percentage of the terms and we want as many relevant terms as possible for our analysis

```
> CORP_StopWords_TDM <- removeSparseTerms(CORP_StopWords_TDM, sparse = 0.2)
> #generating the new dendogram
> CORP_df <- as.data.frame(as.matrix(CORP_StopWords_TDM))
> CORP_dist <- dist(CORP_df)
> CORP_dg <- hclust(CORP_dist, method ='ward.D2')
> str(CORP_dg)
List of 7
 $ merge      : int [1:69, 1:2] -2 -30 -34 -43 -9 -3 -27 -5 -7 -17 ...
 $ height     : num [1:69] 4.12 4.12 4.12 4.12 4.8 ...
 $ order      : int [1:70] 65 35 28 4 21 50 25 32 58 7 ...
 $ labels     : chr [1:70] "almost" "alone" "also" "apes" ...
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = CORP_dist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> tm::inspect((CORP_StopWords_TDM))
<<TermDocumentMatrix (terms: 70, documents: 15)>>
Non-/sparse entries: 978/72
Sparsity           : 7%
Maximal term length: 10
Weighting          : term frequency (tf)
Sample             :
        Docs
Terms    chp11.txt chp12.txt chp13.txt chp14.txt chp2.txt chp3.txt chp4.txt chp5.txt chp7.txt chp8.txt
  apes          18        11         6         2        0        3        9       12       12       21
  back           7         7        14         7        4        4        4        2        5       18
  great         18        14         9        17        8       13       11       12       15       28
  jungle         6         5        20        16        3        5        2        6       14       25
  little        12        10        17         3       11       20       16       24       28       19
  man            5         7        31        19       10        2        3        1        2       13
  now            8         5         6        13        5        6        5        5       11       10
  one           15         9        19        10        2        9       10        8       22       25
  strange        7         6         4         4        0        5        6        0        7       17
  upon          20        11        31        20       15       14       14       13       22       42
```

```
> plot(CORP_dg)
```



Word Cloud

A word cloud is a graphical representation of text data, where the size of each word corresponds to its frequency or importance. Words are typically displayed in different sizes and colors, with more prominent words indicating higher frequency or importance.

We need wordcloud package for this which we installed in the beginning.

```
> tf_comb <- unlist(tfList)
> tf_summ <- tapply(tf_comb, names(tf_comb), sum)
> Words <- names(tf_summ)
> Words
   [1] "abandon"          "abandoned"        "abandoning"       "abashed"          "abated"           "ability"
   [7] "able"             "aboard"           "abode"            "abruptly"         "absence"          "absent"
  [13] "absentmindedness" "absolute"         "absolutely"       "absorbed"         "abstruse"         "abundance"
  [19] "abundant"         "abuses"           "abysmal"          "accentuate"       "accentuated"      "accept"
  [25] "accession"        "accident"         "accidental"       "accidentally"     "accompanied"      "accompany"
  [31] "accomplish"       "accomplished"     "accomplishment"   "accord"           "accorded"         "account"
  [37] "accounted"        "accuracy"         "accurately"       "accustom"         "accustomed"       "achievement"
  [43] "acknowledging"    "acquaintances"    "acquirement"      "across"           "act"              "action"
  [49] "active"           "add"              "added"            "adding"           "addition"         "additional"
  [55] "addressed"        "adds"             "adequate"         "adjusted"         "adjusting"        "admiration"
  [61] "admired"          "admit"            "admitted"         "adopted"          "adult"            "adults"
  [67] "advance"          "advanced"         "advancing"        "advantage"        "adventure"        "adverbs"
  [73] "adversary"        "advised"          "aerial"           "aerie"            "affairs"          "affected"
  [79] "affection"        "afford"           "affrighted"       "affront"          "afraid"           "africa"
  [85] "african"          "aft"              "afternoon"        "age"              "agencies"         "ages"
  [91] "aggregation"      "agile"            "agility"          "agin"             "ago"              "agonized"
  [97] "agony"            "agree"            "agreeable"        "agreedthat"       "ahead"            "aim"
 [103] "aimed"            "aimlessly"        "aint"             "air"              "airraised"        "alarm"
 [109] "alas"             "alert"            "alice"            "alight"           "alighting"        "alike"
 [115] "alive"            "alla"             "allay"            "allegorical"      "allied"           "allow"
 [121] "allowed"          "almost"           "aloft"            "alone"            "along"            "aloud"
 [127] "alphabet"         "alphabetical"     "already"          "also"             "altardrum"        "alternately"
 [133] "alternative"      "always"           "amazed"           "amazing"          "ambition"         "ambitions"
 [139] "amen"             "american"         "amidst"           "amiss"            "amity"            "ammunition"
 [145] "among"            "amongst"          "amount"           "amphitheater"     "amuck"            "amused"
 [151] "amusement"        "ancestor"         "ancestors"        "ancestry"         "anchor"           "anchored"
 [157] "ancient"          "andoh"            "anger"            "angered"          "angry"            "animal"
 [163] "animals"          "animosity"        "ankle"            "anklet"           "anklets"          "announced"
 [169] "annoy"            "another"          "anothera"         "answer"           "answered"         "answering"
 [175] "answers"          "antagonist"       "antagonize"       "antelope"         "anthropoid"       "anthropoids"
 [181] "anticipation"     "anxiety"          "anxious"          "anyone"           "anything"         "anyway"
 [187] "apart"            "apartment"        "apathy"           "ape"              "apechild"         "apehood"
 [193] "apeman"           "apemans"          "aperture"         "apes"             "apesno"           "aplenty"
```

We gather unique words by using the term frequencies from all 15 documents. By unlisting the tfList variable, we combine all the term frequencies. Then, we sum up the term frequencies for each unique term. Finally, we store the set of words into the words variable.

Next, we utilize the wordcloud() function to create the word cloud. We provide the summed up term frequencies and the unique words as input. However, we encountered warnings indicating that certain words are too lengthy to be accommodated within the word cloud and are consequently omitted.So, we produce another word cloud while setting min.freq = 30, ensuring that only words appearing at least 30 times in the corpus are displayed.

```
> pal <- brewer.pal(9,'BuGn')
> str(pal)
 chr [1:9] "#F7FCFD" "#E5F5F9" "#CCECE6" "#99D8C9" "#66C2A4" "#41AE76" "#238B45" "#006D2C" "#00441B"
> CORP_WC <- wordcloud(Words, tf_summ, colors = pal[-(1:4)])
There were 50 or more warnings (use warnings() to see the first 50)


> CORP_WC <- wordcloud(Words, tf_summ, min.freq = 30, colors = pal[-(1:4)])
```

The second corpus allows us to understand the wordcloud better.

Applying a different pallet

Applying some functions on the Corpus

Following the rubric, we go through several packages like “quanteda”, “syuzhet”, etc. to perform further analysis.

```
install.packages('quanteda')
library('quanteda')
```

We use the CORP_clean copy of the corpus to perform further analysis.
Printing out the first ten lines of the first document:

```
> CORP_text <- CORP_clean[[1]]
> CORP_text$content[1:10]
 [1] ""
 [2] "Out to Sea"
 [3] ""
 [4] ""
 [5] "I had this story from one who had no business to tell it to me or to"
 [6] "any other  I may credit the seductive influence of an old vintage upon"
 [7] "the narrator for the beginning of it and my own skeptical incredulity"
 [8] "during the days that followed for the balance of the strange tale"
 [9] ""
[10] "When my convivial host discovered that he had told me so much and that"
```

Let us apply a few methods from the quanteda package on our corpus.
First, we apply tokenization. We do this by using the lapply() function to apply tokenization to each of the 15 documents of the corpus.

```
> CORP_tokens <-lapply(CORP_clean,function(x) quanteda::tokens(x$content))
> CORP_tokens
$chp1.txt
Tokens consisting of 420 documents.
text1 :
character(0)

text2 :
[1] "Out" "to"  "Sea"

text3 :
character(0)

text4 :
character(0)

text5 :
 [1] "I"       "had"     "this"    "story"   "from"    "one"     "who"     "had"     "no"      "business"
[11] "to"      "tell"
[ ... and 5 more ]

text6 :
 [1] "any"       "other"     "I"         "may"       "credit"    "the"       "seductive" "influence" "of"
[10] "an"        "old"       "vintage"
[ ... and 1 more ]

[ reached max_ndoc ... 414 more documents ]

$chp10.txt
Tokens consisting of 215 documents.
text1 :
character(0)

text2 :
[1] "The"         "FearPhantom"

text3 :
character(0)
```

When we examine the structure of CORP_tokens using str(), we see a huge output for all the 15 documents.

Kwic() helps locate keywords in context from the text.

```
> kwic(CORP_tokens[[1]], pattern = "great")
Keyword-in-context with 4 matches.
 [text144, 13] fierce black mustachios and a | great |
  [text153, 4]                 blank at the | great | mountain of muscle towering before
  [text211, 3]                      as the | great | lines of a British battleship
  [text239, 1]                             | great | battleship The old fellow was

> kwic(CORP_tokens[[2]], pattern = "great")
Keyword-in-context with 4 matches.
  [text45, 2]                          him | great | respect for the little sharp
  [text48, 7]           length he came to a | great | tree heavy laden with thick
 [text199, 7] glances behind them from their | great | rolling eyes
 [text202, 1]                               | great | tree There was much in

> kwic(CORP_tokens[[3]], pattern = "great")
Keyword-in-context with 18 matches.
   [text81, 2]                                the | great | tree and as before he
 [text136, 10]             their poor victim to a | great | post near the center of
  [text212, 5]              of arrows beneath the | great | tree at the end of
  [text216, 6]           Silently he climbed to a | great | height until he found a
  [text245, 5]             they had offended some | great | god by placing their village
  [text247, 7]         was daily placed below the | great | tree from whence the arrows
  [text261, 2]                               The | great | yellow eyes were fixed upon
 [text271, 13]                 one side and as the | great | cat
  [text279, 9]          Apes went down beneath the | great | body of his enemy but
  [text284, 8]        he wriggled from beneath the | great | weight and as he
  [text295, 1]                                   | great | anthropoids
  [text306, 5]              Deftly he removed the | great | pelt for he had practiced
  [text341, 7]           With a frightful roar the | great | beast sprang among the
  [text351, 4]                  Come down Tarzan | great | killer cried Kerchak Come down
  [text366, 7] backdrawn snarling lips exposed his | great | fighting fangs and his
  [text371, 6]              feet of height and his | great | rolling sinews seemed pitifully inadequate
  [text403, 8]            apeman close to him his | great | jaws sought Tarzans
  [text413, 5]            a shuddering tremor the | great | body stiffened for an instant
```

We now use the dfm() function to create sparse document-feature matrices for all the chapters. Here, we also remove the stopwords using dfm_remove().

```
> CORP_DFM_Chp1 <- quanteda::dfm(CORP_tokens[[1]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp2 <- quanteda::dfm(CORP_tokens[[2]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp3 <- quanteda::dfm(CORP_tokens[[3]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp4 <- quanteda::dfm(CORP_tokens[[4]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp5 <- quanteda::dfm(CORP_tokens[[5]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp6 <- quanteda::dfm(CORP_tokens[[6]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp7 <- quanteda::dfm(CORP_tokens[[7]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp8 <- quanteda::dfm(CORP_tokens[[8]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp9 <- quanteda::dfm(CORP_tokens[[9]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp10 <- quanteda::dfm(CORP_tokens[[10]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp11 <- quanteda::dfm(CORP_tokens[[11]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp12 <- quanteda::dfm(CORP_tokens[[12]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp13 <- quanteda::dfm(CORP_tokens[[13]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp14 <- quanteda::dfm(CORP_tokens[[14]]) %>% dfm_remove(StopWords)
> CORP_DFM_Chp15 <- quanteda::dfm(CORP_tokens[[15]]) %>% dfm_remove(StopWords)
```

```
> str(CORP_DFM_Chp1)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    420 obs. of  3 variables:
  .. ..$ docname_: chr [1:420] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 420 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:420] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "anith"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/anith/OneDrive/Desktop/GW/sem 2/Big Daata/project 3"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 5
  .. .. .. ..$ scheme   : chr "unary"
  .. .. .. ..$ base     : NULL
  .. .. .. ..$ c        : NULL
  .. .. .. ..$ smoothing: NULL
  .. .. .. ..$ threshold: NULL
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
  .. ..$ user  : list()
  ..@ i       : int [1:1713] 1 115 305 4 16 23 50 4 26 53 ...
  ..@ p       : int [1:1044] 0 3 7 18 20 21 31 32 33 34 ...
  ..@ Dim     : int [1:2] 420 1043
  ..@ Dimnames:List of 2
```

```
> str(CORP_DFM_Chp6)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    397 obs. of  3 variables:
  .. ..$ docname_: chr [1:397] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 397 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:397] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "anith"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/anith/OneDrive/Desktop/GW/sem 2/Big Daata/project 3"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 5
  .. .. .. ..$ scheme   : chr "unary"
  .. .. .. ..$ base     : NULL
  .. .. .. ..$ c        : NULL
  .. .. .. ..$ smoothing: NULL
  .. .. .. ..$ threshold: NULL
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
```

Next, we get the frequency of terms in the DFM for the chapters.

```
> CORP_Freq_Chp1 <- quanteda::docfreq(CORP_DFM_Chp1)
> CORP_Freq_Chp2 <- quanteda::docfreq(CORP_DFM_Chp2)
> CORP_Freq_Chp3 <- quanteda::docfreq(CORP_DFM_Chp3)
> CORP_Freq_Chp4 <- quanteda::docfreq(CORP_DFM_Chp4)
> CORP_Freq_Chp5 <- quanteda::docfreq(CORP_DFM_Chp5)
> CORP_Freq_Chp6 <- quanteda::docfreq(CORP_DFM_Chp6)
> CORP_Freq_Chp7 <- quanteda::docfreq(CORP_DFM_Chp7)
> CORP_Freq_Chp8 <- quanteda::docfreq(CORP_DFM_Chp8)
> CORP_Freq_Chp9 <- quanteda::docfreq(CORP_DFM_Chp9)
> CORP_Freq_Chp10 <- quanteda::docfreq(CORP_DFM_Chp10)
> CORP_Freq_Chp11 <- quanteda::docfreq(CORP_DFM_Chp11)
> CORP_Freq_Chp12 <- quanteda::docfreq(CORP_DFM_Chp12)
> CORP_Freq_Chp13 <- quanteda::docfreq(CORP_DFM_Chp13)
> CORP_Freq_Chp14 <- quanteda::docfreq(CORP_DFM_Chp14)
> CORP_Freq_Chp15 <- quanteda::docfreq(CORP_DFM_Chp15)
```

```
> str(CORP_Freq_Chp3)
 Named int [1:1000] 1 18 1 1 2 7 2 3 1 1 ...
 - attr(*, "names")= chr [1:1000] "king" "apes" "yet" "dark" ...
> CORP_Freq_Chp3
```

king	apes	yet	dark	reached	tribe	though
1	18	1	1	2	7	2
stopped	exhume	devour	remains	wild	boar	cached
3	1	1	1	4	1	1
preceding	day	take	kulongas	bow	arrows	tree
1	5	1	1	4	10	7
top	hidden	wellladen	tarzan	dropped	branches	midst
1	3	1	37	4	2	1
kerchak	swelling	chest	narrated	glories	adventure	exhibited
12	2	1	1	1	1	2
spoils	conquest	grunted	turned	away	jealous	strange
1	1	1	7	5	1	6
member	band	little	evil	brain	sought	excuse
1	1	12	2	2	6	1
wreak	hatred	upon	next	practicing	first	gleam
1	3	20	2	1	5	2
dawn	lost	nearly	every	bolt	shot	finally
2	1	3	3	1	2	1
learned	guide	shafts	fair	accuracy	ere	month
5	1	1	1	1	5	1
passed	mean	proficiency	cost	entire	supply	continued
2	1	1	1	1	2	1
find	hunting	good	vicinity	beach	varied	archery
3	6	1	2	2	1	1
practice	investigation	fathers	choice	store	books	period
1	2	1	1	1	2	1
young	english	lord	found	back	one	cupboards
7	2	5	5	7	15	1
cabin	small	metal	box	key	lock	moments
2	4	2	4	2	1	1
experimentation	rewarded	successful	opening	receptacle	faded	photograph
1	2	2	2	1	1	3
smooth	faced	man	golden	locket	studded	diamonds
2	2	5	1	2	1	1
linked	gold	chain	letters	book	examined	minutely

Now, let us assign weights to these words.

```
> CORP_Weight_Chp1 <- quanteda::dfm_weight(CORP_DFM_Chp1)
> CORP_Weight_Chp2 <- quanteda::dfm_weight(CORP_DFM_Chp2)
> CORP_Weight_Chp3 <- quanteda::dfm_weight(CORP_DFM_Chp3)
> CORP_Weight_Chp4 <- quanteda::dfm_weight(CORP_DFM_Chp4)
> CORP_Weight_Chp5 <- quanteda::dfm_weight(CORP_DFM_Chp5)
> CORP_Weight_Chp6 <- quanteda::dfm_weight(CORP_DFM_Chp6)
> CORP_Weight_Chp7 <- quanteda::dfm_weight(CORP_DFM_Chp7)
> CORP_Weight_Chp8 <- quanteda::dfm_weight(CORP_DFM_Chp8)
> CORP_Weight_Chp9 <- quanteda::dfm_weight(CORP_DFM_Chp9)
> CORP_Weight_Chp10 <- quanteda::dfm_weight(CORP_DFM_Chp10)
> CORP_Weight_Chp11 <- quanteda::dfm_weight(CORP_DFM_Chp11)
> CORP_Weight_Chp12 <- quanteda::dfm_weight(CORP_DFM_Chp12)
> CORP_Weight_Chp13 <- quanteda::dfm_weight(CORP_DFM_Chp13)
> CORP_Weight_Chp14 <- quanteda::dfm_weight(CORP_DFM_Chp14)
> CORP_Weight_Chp15 <- quanteda::dfm_weight(CORP_DFM_Chp15)
> str(CORP_Weight_Chp3)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    427 obs. of  3 variables:
  .. ..$ docname_: chr [1:427] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 427 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:427] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "anith"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/anith/OneDrive/Desktop/GW/sem 2/Big Daata/project 3"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 5
  .. .. .. ..$ scheme   : chr "unary"
  .. .. .. ..$ base     : NULL
  .. .. .. ..$ c        : NULL
  .. .. .. ..$ smoothing: NULL
  .. .. .. ..$ threshold: NULL
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
```

```
> CORP_Weight_Chp3
Document-feature matrix of: 427 documents, 1,000 features (99.59% sparse) and 0 docvars.
       features
docs    king apes yet dark reached tribe though stopped exhume devour
  text1    0    0   0    0       0     0      0       0      0      0
  text2    1    1   0    0       0     0      0       0      0      0
  text3    0    0   0    0       0     0      0       0      0      0
  text4    0    0   0    0       0     0      0       0      0      0
  text5    0    0   1    1       1     1      1       1      0      0
  text6    0    0   0    0       0     0      0       0      1      1
[ reached max_ndoc ... 421 more documents, reached max_nfeat ... 990 more features ]
```

Now, let us compute the tf-idf score. This helps us weigh a dfm by term frequency-inverse document frequency (tf-idf), with full control over options.

```
> CORP_TFIDF_Chp1 <- quanteda::dfm_tfidf(CORP_DFM_Chp1,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp2 <- quanteda::dfm_tfidf(CORP_DFM_Chp2,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp3 <- quanteda::dfm_tfidf(CORP_DFM_Chp3,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp4 <- quanteda::dfm_tfidf(CORP_DFM_Chp4,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp5 <- quanteda::dfm_tfidf(CORP_DFM_Chp5,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp6 <- quanteda::dfm_tfidf(CORP_DFM_Chp6,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp7 <- quanteda::dfm_tfidf(CORP_DFM_Chp7,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp8 <- quanteda::dfm_tfidf(CORP_DFM_Chp8,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp9 <- quanteda::dfm_tfidf(CORP_DFM_Chp9,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp10 <- quanteda::dfm_tfidf(CORP_DFM_Chp10,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp11 <- quanteda::dfm_tfidf(CORP_DFM_Chp11,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp12 <- quanteda::dfm_tfidf(CORP_DFM_Chp12,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp13 <- quanteda::dfm_tfidf(CORP_DFM_Chp13,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp14 <- quanteda::dfm_tfidf(CORP_DFM_Chp14,scheme_tf = "count", scheme_df="inverse")
> CORP_TFIDF_Chp15 <- quanteda::dfm_tfidf(CORP_DFM_Chp15,scheme_tf = "count", scheme_df="inverse")
```

```
> str(CORP_TFIDF_Chp3)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    427 obs. of  3 variables:
  .. ..$ docname_: chr [1:427] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 427 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:427] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 0 2
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 3 2
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "anith"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "C:/Users/anith/OneDrive/Desktop/GW/sem 2/Big Daata/project 3"
  .. .. ..$ created        : Date[1:1], format: "2024-05-04"
  .. ..$ object:List of 10
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ tokenizer   : chr "tokenize_word4"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 2
  .. .. .. ..$ scheme: chr "inverse"
  .. .. .. ..$ base  : num 10
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
  .. ..$ user  : list()
  ..@ i       : int [1:1748] 1 1 26 87 111 131 143 161 238 250 ...
  ..@ p       : int [1:1001] 0 1 19 20 21 23 30 32 35 36 ...
  ..@ Dim     : int [1:2] 427 1000
  ..@ Dimnames:List of 2
  .. ..$ docs    : chr [1:427] "text1" "text2" "text3" "text4" ...
```

```
> CORP_TFIDF_Chp3
Document-feature matrix of: 427 documents, 1,000 features (99.59% sparse) and 0 docvars.
       features
docs        king     apes      yet     dark  reached   tribe   though  stopped   exhume   devour
  text1 0        0        0        0        0        0       0        0        0        0
  text2 2.630428 1.375155 0        0        0        0       0        0        0        0
  text3 0        0        0        0        0        0       0        0        0        0
  text4 0        0        0        0        0        0       0        0        0        0
  text5 0        0        2.630428 2.630428 2.329398 1.78533 2.329398 2.153307 0        0
  text6 0        0        0        0        0        0       0        0        2.630428 2.630428
[ reached max_ndoc ... 421 more documents, reached max_nfeat ... 990 more features ]
```

Now we install the “syuzhet” package and make it a library.

```
> install.packages('syuzhet')
also installing the dependencies ‘textshape’, ‘zoo’, ‘dtt’

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.3/textshape_1.7.5.tgz'
Content type 'application/x-gzip' length 521774 bytes (509 KB)
==================================================
downloaded 509 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.3/zoo_1.8-12.tgz'
Content type 'application/x-gzip' length 1026435 bytes (1002 KB)
==================================================
downloaded 1002 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.3/dtt_0.1-2.tgz'
Content type 'application/x-gzip' length 19738 bytes (19 KB)
==================================================
downloaded 19 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.3/syuzhet_1.0.7.tgz'
Content type 'application/x-gzip' length 3118163 bytes (3.0 MB)
==================================================
downloaded 3.0 MB


The downloaded binary packages are in
        /var/folders/yn/h35xrlns7ld2b_5gl156brbw0000gn/T//RtmpAMGjzI/downloaded_packages
> library('syuzhet')
>
```

We can extract text as a data frame. Let us apply a few functions from the syuzhet package on our document.

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

Let’s get the sentences in TarzanOfTheApes.txt, we first read the file as one large string as shown with the help of get_text_as_string() function.

```
420
> #read the text document as a string
> CORP_asString <- get_text_as_string("TarzanOfTheApes.txt")
> CORP_asString
 Tarzan of the Apes   By  Edgar Rice Burroughs              CONTENTS        I  Out to Sea      II  The Savage H
e Battles     VII  The Light of Knowledge    VIII  The Tree-top Hunter      IX  Man and Man       X  The Fear-P
XIV  At the Mercy of the Jungle      XV  The Forest God     XVI  "Most Remarkable"    XVII  Burials   XVIII  Th
lage of Torture    XXII  The Search Party   XXIII  Brother Men    XXIV  Lost Treasure     XXV  The Outpost of t
lusion     Chapter I  Out to Sea   I had this story from one who had no business to tell it to me, or to any ot
the beginning of it, and my own skeptical incredulity during the days that followed for the balance of the stra
I was prone to doubtfulness, his foolish pride assumed the task the old vintage had commenced, and so he uneart
he British Colonial Office to support many of the salient features of his remarkable narrative.  I do not say t
act that in the telling of it to you I have taken fictitious names for the principal characters quite sufficien
ewed pages of the diary of a man long dead, and the records of the Colonial Office dovetail perfectly with the
ed it out from these several various agencies.  If you do not find it credible you will at least be as one with
rds of the Colonial Office and from the dead man's diary we learn that a certain young English nobleman, whom w
licate investigation of conditions in a British West Coast African Colony from whose simple native inhabitants
ch it used solely for the forcible collection of rubber and ivory from the savage tribes along the Congo and th
en were enticed away through the medium of fair and glowing promises, but that few if any ever returned to thei
cks were held in virtual slavery, since after their terms of enlistment expired their ignorance was imposed upo
rve.  And so the Colonial Office appointed John Clayton to a new post in British West Africa, but his confident
ack British subjects by the officers of a friendly European power.  Why he was sent, is, however, of little mom
ach his destination.  Clayton was the type of Englishman that one likes best to associate with the noblest monu
rile man--mentally, morally, and physically.  In stature he was above the average height; his eyes were gray, h
ced by his years of army training.  Political ambition had caused him to seek transference from the army to the
```

Many lines have been omitted.

Now we get the sentences using the get_sentences() function from the syuzhet package. It parses the string into individual sentences.

```
> #getting sentences from the string
> CORP_sentences <- get_sentences(CORP_asString)
> CORP_sentences[1:50]
 [1] "Tarzan of the Apes   By  Edgar Rice Burroughs              CONTENTS        I  Out to Sea      II  The Sav
Jungle Battles     VII  The Light of Knowledge    VIII  The Tree-top Hunter      IX  Man and Man       X  The F
nd     XIV  At the Mercy of the Jungle      XV  The Forest God     XVI  \"Most Remarkable\"    XVII  Burials
The Village of Torture    XXII  The Search Party   XXIII  Brother Men    XXIV  Lost Treasure     XXV  The Outpo
I  Conclusion     Chapter I  Out to Sea   I had this story from one who had no business to tell it to me, or to
 [2] "I may credit the seductive influence of an old vintage upon the narrator for the beginning of it, and my
nge tale."
 [3] "When my convivial host discovered that he had told me so much, and that I was prone to doubtfulness, his
ritten evidence in the form of musty manuscript, and dry official records of the British Colonial Office to sup
 [4] "I do not say the story is true, for I did not witness the happenings which it portrays, but the fact that
ters quite sufficiently evidences the sincerity of my own belief that it MAY be true."
 [5] "The yellow, mildewed pages of the diary of a man long dead, and the records of the Colonial Office doveta
s I painstakingly pieced it out from these several various agencies."
 [6] "If you do not find it credible you will at least be as one with me in acknowledging that it is unique, re
 [7] "From the records of the Colonial Office and from the dead man's diary we learn that a certain young Engli
ake a peculiarly delicate investigation of conditions in a British West Coast African Colony from whose simple
ts native army, which it used solely for the forcible collection of rubber and ivory from the savage tribes alo
 [8] "The natives of the British Colony complained that many of their young men were enticed away through the m
lies."
 [9] "The Englishmen in Africa went even further, saying that these poor blacks were held in virtual slavery, s
eir white officers, and they were told that they had yet several years to serve."
[10] "And so the Colonial Office appointed John Clayton to a new post in British West Africa, but his confident
ack British subjects by the officers of a friendly European power."
[11] "Why he was sent, is, however, of little moment to this story, for he never made an investigation, nor, in
[12] "Clayton was the type of Englishman that one likes best to associate with the noblest monuments of histori
ly, morally, and physically."
[13] "In stature he was above the average height; his eyes were gray, his features regular and strong; his carr
```

We now proceed to extract the sentiments from these sentences. To do this, we utilize the get_sentiment() method that takes a vector of strings and one of four methods for sentiment extraction, "syuzhet", "bing", "afinn", and "nrc". We experiment with all of these methods to observe the results

```
[49] "Though the fellow's tone was surly, his words were evidently well meant."
[50] "Ere he had scarce finished his little speech he had turned and was limping off toward the forecastle with
> str(CORP_sentences)
 chr [1:4312] "Tarzan of the Apes   By  Edgar Rice Burroughs              CONTENTS        I  Out to Sea      II
>
```

We first call get_sentiment() with the default value of “syuzhet”

```
> #analyzing the sentiment using syuzhet method
> CORP_syuzhet <- get_sentiment(CORP_sentences,"syuzhet")
> CORP_syuzhet
   [1]  0.90 -0.25  1.10  2.40  0.10  2.60  1.25  1.15 -3.15  0.55  0.00  3.50  2.00  2.95  1.10  2.75
  [17] -0.35 -0.75  1.00  0.00  0.50  0.00  0.00  0.55  0.80  0.00 -2.25 -0.25 -0.25  0.40 -0.50  0.20
  [33] -0.15  0.00 -0.50 -1.00 -0.75 -2.15 -0.50 -0.25 -1.05 -1.50 -1.85 -1.00 -1.20  1.00  1.00 -0.50
  [49]  0.20 -1.05 -0.30  0.10 -0.85 -2.25 -0.50 -0.40 -1.15 -1.75  0.25  1.10 -1.00  1.65  0.00 -0.25
  [65]  0.50  1.05 -0.10  1.15  0.00  0.00  0.00 -0.80  0.80 -1.00  0.00 -1.00  0.40 -1.00 -0.50  0.00
  [81]  0.05  0.00  0.40  0.75 -0.75 -1.00 -0.90  0.15  0.00  0.05  0.50 -1.00 -3.90  1.25  0.00 -0.25
  [97] -0.75  0.75 -0.50 -0.35 -0.25  0.00 -0.55  0.40 -0.60  0.80  0.80 -1.75 -0.50  0.00 -2.65  1.00
 [113]  0.00 -1.65  1.15 -0.10 -1.45  0.25  1.60 -0.10  0.25  0.85  0.50 -0.50 -0.90 -0.90 -0.75 -0.50
 [129]  0.80 -0.55  0.00  0.00  1.15 -2.65  0.80  0.00  0.60  0.50 -0.50  0.00  1.30 -0.50  0.80  0.80
 [145]  0.00 -1.40 -0.75 -3.00  0.50 -0.25 -0.55 -1.00 -0.50  0.40 -0.40 -0.50 -0.20 -0.85 -0.40 -2.80
 [161] -3.20 -0.50 -2.00 -1.00 -0.60 -2.45  0.00  0.00  0.00 -0.75 -0.75  0.50  2.45  0.00  0.00  0.90
 [177] -0.50 -2.10 -0.75  0.55 -1.05  0.00  0.00 -0.75  0.55 -2.15  0.65 -0.75 -0.25  1.40  0.00  0.35
 [193] -1.50 -2.00  0.75  0.50  0.80 -0.10  0.75  2.55 -0.50 -2.25 -4.10  0.50  2.95  0.10  0.40  0.80
 [209]  1.00  0.55  1.35  0.50  0.80 -1.00 -0.85 -0.75 -0.50  1.70 -1.05 -1.65 -1.40  0.00  1.55  0.80
 [225]  0.25  0.25 -0.35  0.65  0.25 -1.35 -1.75 -0.75 -1.30 -0.90 -0.80  2.90 -2.00  0.00  0.00  0.25
 [241]  1.25 -2.00  0.00 -0.50  3.20 -0.25  0.75  0.00  2.20  1.50 -0.70  1.30  2.00  0.65 -0.50  1.85
 [257]  0.55  0.00  0.40  0.40  0.40  1.55  4.15 -3.05  1.30  0.50  0.00  0.00  1.10 -0.25  0.00 -1.60
 [273] -1.25 -0.75  1.50  0.80  0.10 -0.40 -2.25  0.00  3.25 -2.50  1.85  2.10  0.05 -0.05  0.00  0.00
 [289]  2.00  1.35  0.00  2.05 -0.85  0.75  1.25 -1.50  0.00  1.85  3.25 -1.15 -2.50 -0.40  1.25  0.35
 [305] -2.45 -1.05  0.25 -1.25  0.05 -3.35 -1.00  0.40 -1.50  0.50 -2.75 -0.85  0.65  0.00  0.00 -0.45
 [321] -2.75 -0.70 -3.00 -0.65  0.00  0.25 -1.50 -0.45 -1.50  1.35  0.00 -1.25 -1.00 -0.60  0.00 -1.75
 [337]  3.15 -1.80  0.25  1.50  0.10  0.00  0.50 -0.45 -1.00  0.50 -2.65  2.20 -3.30  1.60  0.85 -0.50
 [353]  0.00  0.00 -0.25 -2.85 -8.15 -1.50 -2.05 -1.50 -1.40 -0.75 -1.20 -2.00  0.35 -2.25 -0.40 -3.75
 [369] -1.40 -1.95  0.10 -0.90  0.50  0.85 -2.10  0.50  0.40 -0.05  1.80  0.25  0.50 -2.25  0.40 -0.90
 [385] -0.25  0.00 -0.50 -0.90  1.35 -0.40  0.00 -0.50 -1.85 -2.85  0.00 -0.85 -0.55 -0.60 -1.50 -1.60
 [401] -0.25 -2.00  1.25 -0.05  3.75  0.15 -1.75  0.25 -0.70 -2.30 -2.15 -0.30  0.30  0.35 -2.65 -0.40
 [417] -0.05  1.15 -1.40  0.85 -0.50 -1.00 -1.35 -0.40 -3.00 -0.90 -0.80 -0.90  0.85  0.30  0.00 -0.15
```

If we try the “nrc” method, we get the following results.

```
 [977]  0.40  0.85 -1.25  0.60 -0.25 -1.60  0.10 -2.75 -0.95  0.15  0.00 -0.25 -0.40  0.50  0.50  0.80
 [993]  0.50  0.25 -0.75 -1.50  0.50 -0.65 -1.25  0.00
 [ reached getOption("max.print") -- omitted 3312 entries ]
> #analyzing the sentiment using nrc method
> CORP_nrc <- get_sentiment(CORP_sentences,"nrc")
> CORP_nrc
   [1]  7  2  2  2  1  4  3  3 -3 -2  1  3  2  3  2  3  0 -1  0 -1 -1  0 -1  1  2  0 -1  1 -1  1 -1  1
  [33]  0  1  0 -1 -1 -2 -2 -3  1  0  1  0 -2  0  1 -1 -2  0  1  2  1 -4  0 -1 -2 -2  2  4  0  0 -2  1
  [65]  1  5  1  2  0  0  0 -1  1 -1  0 -1  1 -1  1  0 -1  0  5  1 -1  0  0  1  1  0  1  0 -4  2 -1  1
  [97] -1  1  0  0  1  1 -1  1  0  0  0 -2 -1  1 -2  0  1 -1  2  2  1  2  0  0  0  0  1  0 -1 -1 -1 -1
 [129]  0 -1 -1  0  0  0  0  0  0  0 -1 -1  1 -1  0  3  0  0  1 -3  2 -1 -2  1 -1  3 -1  0  0 -1  0 -3
 [161] -3 -1  0 -2  1 -3  1  0  0 -2 -1  1  4  0  0  2  1  0 -1  0  0  0  1 -1 -1 -3  0  0 -1  1  0  2
 [193] -3 -6  2 -1  0  1  2  2 -1 -3 -8  1  3  1 -2  0  3  0  0  0  0 -1 -3 -2 -1  3 -1 -3 -2 -1  0  0
 [225]  0  1  2  0 -1 -4 -3 -1 -1  0  0  4 -2  0  0 -1  1 -1 -1  0  3  0  1  0  2  1 -4  0  1  0 -1  3
 [257] -2  0 -1  0  0  1  7 -2  0  0 -1  0  2  0 -1  1 -3 -1  1  0  2 -1 -2 -1  1 -2  2  2 -1  0 -1  0
 [289]  1  0 -1  3 -1  0  1 -4  1 -1  1  0 -2 -1  1  2 -3 -2  0 -2  0 -5  0  1 -2  0 -5  0  0  0  0  1
 [321] -4 -2 -2  0  1  0 -3  0 -3  2  0 -1 -2  0  0  0  5  1  1  3  1  0  1  2  1  1 -2  1 -4  2  2  1
 [353]  0  0  1 -3 -7 -4  2 -1 -3 -1  1 -1  1 -2  1 -4  0 -2  1  0  2  2 -3  0  2 -1  5  0  0 -2  1  0
 [385] -1  2  0  1  0  1  0 -1  0 -5  0  1 -3  0  0  0 -2 -3  3  1  7 -1 -1  2  0 -1 -4  1  0 -1 -2 -1
 [417] -1 -1 -2  0  0 -1 -1  0 -3 -1 -2 -3  1 -2  1  2 -1  1  0  0  0  0 -2 -2  4 -1  3  1 -2  1  0  1
 [449]  1  0  0 -1  0  1  3  2  3 -1  1  1  3  4  1  0  0 -2 -2 -3  1  1  1  0 -1  0  1  3 -1  2  1  0
 [481]  0  1 -1 -1 -1  0  2  0  2 -1  1 -3  0 -5  0 -1 -2  0  0  1  1  4  0 -2  0 -1 -1 -4  2  0  1 -3
 [513] -1 -1 -1  0 -2  1  3  0  1  0  0  1 -2 -1  1 -1  2  1  1 -1  3  1  0 -3 -1  2 -1 -1 -3 -1  2  2
 [545]  1  2  0  1  0 -1  2  2  3 -2  0 -2  0 -1  1  3  0 -1 -2 -2  1  1  0  0  0  2  0 -1  1  1  0  1
 [577] -1  0  1 -2 -1  0  0  0 -2 -1 -1 -2 -2  0  2  1 -2 -1 -1 -3  0  0 -3 -3 -4  0  2  1 -2  0  1 -4
 [609] -1 -4 -5 -3 -1 -1  0  0  0 -1  0 -2 -1 -2 -1 -1 -5  1  1  0  1  1 -3  2  0  0  1  1  1  2  2  1
 [641]  2  1 -1  0  0  0  1 -1  2  2  0  2 -1  0  0  2  2  0  1  3  1  1  4 -2  1  0  0  1  6  1 -2  0
 [673]  1 -2  1 -2  0  0  1 -6  1  0  1  2  1  2 -3  1  0 -1  0  1  2 -1  1 -1  1 -3 -3 -3 -1  0 -1 -5
 [705]  0 -1 -1  0 -2 -2 -2  1 -5 -3  0  0 -1 -1 -4  0  7 -2  1 -1  0  4 -2 -6 -2 -1  0 -1  0 -2 -1 -1
 [737] -1  1  2  0 -1  1 -2  1  2 -1  0 -2 -2 -3 -1  0  0  0  1  1 -1 -1  2  1  3  0 -1  0  0 -2 -1 -7
```

For “afinn” we get the following results.

```
 [ reached getOption("max.print") -- omitted 3312 entries ]
> #analyzing the sentiment using afinn method
> CORP_afinn <- get_sentiment(CORP_sentences,"afinn")
> CORP_afinn
   [1] -2 -3  5  4  2  4 -2  1 -8  0  1  9  6  4  1 10  0 -2  1  0  1  0  0  0
  [25]  0 -1 -6  2  0  0  0 -1  0  0 -1 -3 -3 -1  0  4 -1 -3  1 -1 -1  3  1  1
  [49]  0  0 -1  2  4 -5 -2  0 -2  0  1  0  0  1 -2  3  3 -1 -1  3  0  0  0  0
  [73]  1  0  0  0  0  0 -2  0 -4  0  0  2 -2 -3 -3  0  3 -2  2  0 -5  3 -1  0
  [97] -2  2 -2 -3  3  0 -2  0  0  0  0 -2  0  0 -4  2  0 -2  3 -2  0  5  2 -3
 [121] -1  3  0 -1 -5 -2 -2 -2  2 -3  0  0  5 -3  0  0  0  3  1 -1  1 -1  1  0
 [145]  0  0 -1 -10  5  0  0 -3  0 -4  0 -2 -4  0  0 -5 -6  0 -3 -4 -2 -3  0  0
 [169]  0 -3 -3  2  4  0  0  2 -1 -6 -1  0 -3  0  0 -2  0 -2  3 -2  0  0  0  0
 [193]  0 -5  3  1  0  4  4  2 -1 -2 -11  0  4  0  0  2  1  0  3  0  0 -2  0 -1
 [217] -1  2 -1 -5 -9  0  3  0  2  2 -6  2  1 -3 -3 -2  0 -4 -2  6 -2  0  0  0
 [241]  2 -3  0 -3  7 -2  3  0  2  4 -2  2  6  1 -1  5 -1  0  4 -1  0  1 15 -7
 [265]  4  4  0  0  3  0  0  0  0 -2  5  0  2  0 -1  3 13 -6  0  6  2 -1  0  0
 [289]  5 -1  0  7  2  3  3 -4  3  7  3  0 -7  0  5  3 -5  3  3 -1 -3 -6 -2  0
 [313] -5  3 -6  1  0  0 -4  1 -7 -2 -6 -3  2  0 -3 -2 -2  3  0 -2 -5 -2  0 -5
 [337] 12 -5  2  3  0  0  1  0  0  5 -1 -1 -3  0  2 -1  0  0 -1 -2 -19 -1 -2 -2
 [361]  1  1 -2 -2 -7 -4  3 -5 -2 -3  1  2  3 -1 -4  3  0 -3  9  4  5 -4  0 -5
 [385] -3  0  3 -2  2 -3  1  1 -5 -6 -1  0 -5  0 -3 -1  1  3  0 -5  9  1 -1 -1
 [409]  1 -3 -3  3 -2  0  1  0  1  7  0  3 -2  0  0  0 -4 -5  0 -6 -2  0  0 -5
 [433] -6 -2  0  0 -5  1  0 -4  4 -4  4 -2 -1  2  3  1  3 -2  1 -1  0  1 -1  7
 [457]  7  0  2  1  3  2  2  2  3  1  0 -6  1  0 -3 -2 -3  0  0  4 -2  3  2  0
 [481]  3 -1  0 -3  1  1  3  4  8  0  0 -5  0 -10 -1 -4 -5 -7 -2 -2  0  2 -2 -4
 [505] -3 -5 -1  0  0  4  2 -3 -1 -6  1  0 -4  0  3  1  2  0  3 -5  2  2 12 -4
 [529] -1 -3  2  0  1 -2  0 -2  2  0  0 -3 -10  2  1  2  0  3  0 -1  3  3  7  2
 [553]  2 -6  0 -3 -3  0  2  0  0 -3 -7 -7  0 -1  2  0  0  4  0  0  2  0  0  0
```

For “bing” we get the following results.

```
> #analyzing the sentiment using bing method
> CORP_bing <- get_sentiment(CORP_sentences,"bing")
> CORP_bing
   [1] -3 -2  1  1  0  3 -2  3 -3  1  0  5  3  2  0  3 -2 -1  0  1  1  0  0 -1  0  0 -3 -1 -1  0 -1  0
  [33]  0  0 -1 -1 -1 -3 -1  0 -2 -2  0 -2  0  0  0 -2  1 -1  0 -1  0 -3 -1 -1 -2 -2  0  1 -1  0  0  0
  [65]  1  0  0  1  0  0  0  0  0  0  0  0  0  0 -1  0  1  0  0  1  0 -1 -1  1 -1  0  0 -1 -4  1  0 -1
  [97] -1  1 -1 -1  1  0 -1  0  0  1  1 -1 -1  0 -2  0  0 -1  1  1 -3 -1  1 -1  1 -1  1 -1 -2  0 -1  0
 [129]  0 -1  0  0  1 -3  1  0  1  1 -1  0  1  0  0  0  0 -1 -2 -4  2  0 -1 -2 -1 -1  0 -1 -1  0  0 -2
 [161] -1 -1 -3 -1  0 -4  0  0  0 -1 -1  0  0  0  0  1 -1 -2 -2  1 -1  0  0 -1  1 -2  0 -1  0  1  0 -1
 [193] -3 -1  1  2  1  1  1  3 -1 -4 -6  1  2  0  2  0  1  2  3  1  1 -1 -1  0  0  1 -1 -3 -1  0  2  0
 [225]  0  1 -2 -1  1  1 -3 -1 -1  1  0  3  0  0  0  1  1 -2  1  0  1 -2  1  0  2  2 -2  1  2  1  0  2
 [257]  1  0  2  1  1 -1  4 -1  1  1  0  0  1 -1  0 -2 -2 -1  2  1  0  0 -2  1  5 -1  1  2  0  0  0  0
 [289]  0  0  0  2 -1  1  2 -4  1  2 -1 -2 -1 -2  2  0 -4 -2  0 -1  0 -5  0 -1 -2  1 -3  0 -1  0  0 -1
 [321] -1 -1 -3 -2  1  1 -3 -1 -2  2 -1  0 -2  0 -1 -4  4 -3  0  3  0  0  0  0  0  2 -1 -2 -2  0  1 -1
 [353]  0  0  0 -4 -9 -1 -1 -2 -1 -1 -2 -2 -3 -3  0 -5 -1 -2  0 -1  0  1 -3  1  0 -2  4  1  0 -2  0 -2
 [385]  0  0 -1 -1  1 -1  0 -5 -4 -2  0 -2 -2  0 -2 -2  0 -1 -1 -3  4  1 -2  0 -1 -4 -4  0 -1 -1 -1  0
 [417] -1  1 -2  0 -1  0 -2  0 -3 -2  0  0  0  1  0 -1 -4 -1 -2 -1 -2  0  0 -3  3 -1 -1  0 -1 -1  1  1
 [449]  1 -1  1 -1  0  0  0  3  2  0  1 -1  5  2  0 -1  1 -2 -1 -3  1  0 -2 -1  0  0 -1  3  0  1  1  0
 [481]  1 -1 -1 -1  1  0  1  2  2  0  1 -2  1 -7  1 -1 -3 -1  1  1  0  2 -1 -2 -2 -3 -1 -3  2 -1  1 -1
 [513]  0 -2  0  1 -1  0  2  1 -2  0  1 -1  0 -1  2 -2  0 -2  1 -1 -1  1 -1 -1  0  0 -1  0 -3  1  0  2
 [545]  0  2  0  1  0  0  2  1  0 -1  0  1 -3  0  1 -2  0 -1 -3 -3  0 -2  0  1  0 -1  0  0  0  0  0  0
 [577] -5  0  0 -1 -1  1  1 -1 -2  0  0 -1 -2  1  2  1 -1 -1 -2 -3  1  1 -2 -3 -2  0 -1  2 -6  1  3 -1
 [609] -1 -4 -3 -1 -2 -1 -2 -1  0 -3  0 -2 -1 -4  0  0 -2  3  2  2  0  0 -3  4 -2  0  0  0 -1  1  1  3
 [641] -7 -1 -1  0 -1 -1  0 -1 -3 -1  0  3  0 -1 -2  0  0  0 -1  3  0  2  0 -2  1  0  0  2  3  0 -1 -3
 [673]  1 -3  3 -2 -1  0  0 -7  1  2  0  0  0 -2 -2  0  0 -2  0 -2  1  0  0 -1  0 -1 -1 -3 -1 -1 -3 -4
 [705] -1 -1  0 -1 -2 -3 -2  1 -4 -3 -1 -2  1  0 -1  0  1 -2  2 -1 -1  1  0 -6 -1  0  0  1  1 -3 -1 -2
 [737] -1  0  2 -2  0  0  0  0  0  1  0  0 -2 -5 -1  0  0  0  1  1 -2 -3  0 -2  0  0 -2  0  1 -1  0 -6
 [769] -2 -1  1  1 -1 -1  1  1  1 -1 -1  1 -3  0 -1 -1  1  1  0  0 -1 -4  0 -1 -3 -4 -2 -4  0  0  1  0
 [801] -2 -1  0  0 -2 -1  0 -1  1 -1  0  2  0 -1  1  0  2  0  2 -1  0  1  1  0 -1 -1  0 -2  2 -1  1 -1
```

Let us look at the sentiment dictionaries for these four methods

```
> #sentiment dictionary for syuzhet method
> CORP_Dictionary_syuzhet <- get_sentiment_dictionary()
> CORP_Dictionary_syuzhet
                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
```

```
 [ reached 'max' / getOption("max.print") -- omitted 10248 row
> #sentiment dictionary for nrc method
> CORP_Dictionary_nrc <- get_sentiment_dictionary("nrc")
> CORP_Dictionary_nrc
       lang           word sentiment value
1   english           abba  positive     1
2   english        ability  positive     1
3   english abovementioned  positive     1
4   english       absolute  positive     1
5   english     absolution  positive     1
6   english       absorbed  positive     1
7   english      abundance  positive     1
8   english       abundant  positive     1
9   english       academic  positive     1
10  english        academy  positive     1
11  english     acceptable  positive     1
12  english     acceptance  positive     1
13  english     accessible  positive     1
14  english       accolade  positive     1
15  english  accommodation  positive     1
16  english  accompaniment  positive     1
17  english     accomplish  positive     1
18  english   accomplished  positive     1
19  english accomplishment  positive     1
20  english         accord  positive     1
21  english accountability  positive     1
22  english    accountable  positive     1
23  english     accredited  positive     1
24  english        accueil  positive     1
25  english       accurate  positive     1
```

```
> #sentiment dictionary for bing method
> CORP_Dictionary_bing <- get_sentiment_dictionary("bing")
> CORP_Dictionary_bing
                    word value
1                     a+     1
2                 abound     1
3                abounds     1
4              abundance     1
5               abundant     1
6             accessable     1
7             accessible     1
8                acclaim     1
9              acclaimed     1
10           acclamation     1
11              accolade     1
12             accolades     1
13         accommodative     1
14          accomodative     1
15            accomplish     1
16          accomplished     1
17        accomplishment     1
18       accomplishments     1
19               accurate     1
20             accurately     1
21             achievable     1
22            achievement     1
23           achievements     1
24             achievible     1
25                 acumen     1
```

```
 [ reached 'max' / getOption("max.print") -- omitted 6789 rows
> #sentiment dictionary for afinn method
> CORP_Dictionary_afinn <- get_sentiment_dictionary("afinn")
> CORP_Dictionary_afinn
                word value
1            abandon    -2
2          abandoned    -2
3           abandons    -2
4           abducted    -2
5          abduction    -2
6         abductions    -2
7              abhor    -3
8           abhorred    -3
9          abhorrent    -3
10            abhors    -3
11         abilities     2
12           ability     2
13            aboard     1
14           aborted    -1
15            aborts    -1
16          absentee    -1
17         absentees    -1
18           absolve     2
19          absolved     2
20          absolves     2
21         absolving     2
22          absorbed     1
23             abuse    -3
24            abused    -3
25            abuses    -3
```

We can notice that each of these methods have different criteria and rules that assign sentiment to words, and their dictionaries have different lengths.

Let us sum the values of the sentiment vectors in order to get a measure of the overall emotional valence in the text:

```
 [ reached 'max' / getOption("max.print") -- omitted 2882 rows
> #sum of sentiment score for syuzhet method
> CORP_Sum_syuzhet <- sum(CORP_syuzhet)
> CORP_Sum_syuzhet
[1] -469.75
> #sum of sentiment score for nrc method
> CORP_Sum_nrc <- sum(CORP_nrc)
> CORP_Sum_nrc
[1] 0
> #sum of sentiment score for bing method
> CORP_Sum_bing <- sum(CORP_bing)
> CORP_Sum_bing
[1] -1183
> #sum of sentiment score for afinn method
> CORP_Sum_afinn <- sum(CORP_afinn)
> CORP_Sum_afinn
[1] -909
>
```

We can observe that except “nrc”, the other methods say that the content of the text is negative. All the other three hand, gives a very strong negative number for the sentiment. To do sentiment analysis, it is important to understand the basis of the sentiment dictionary we use.

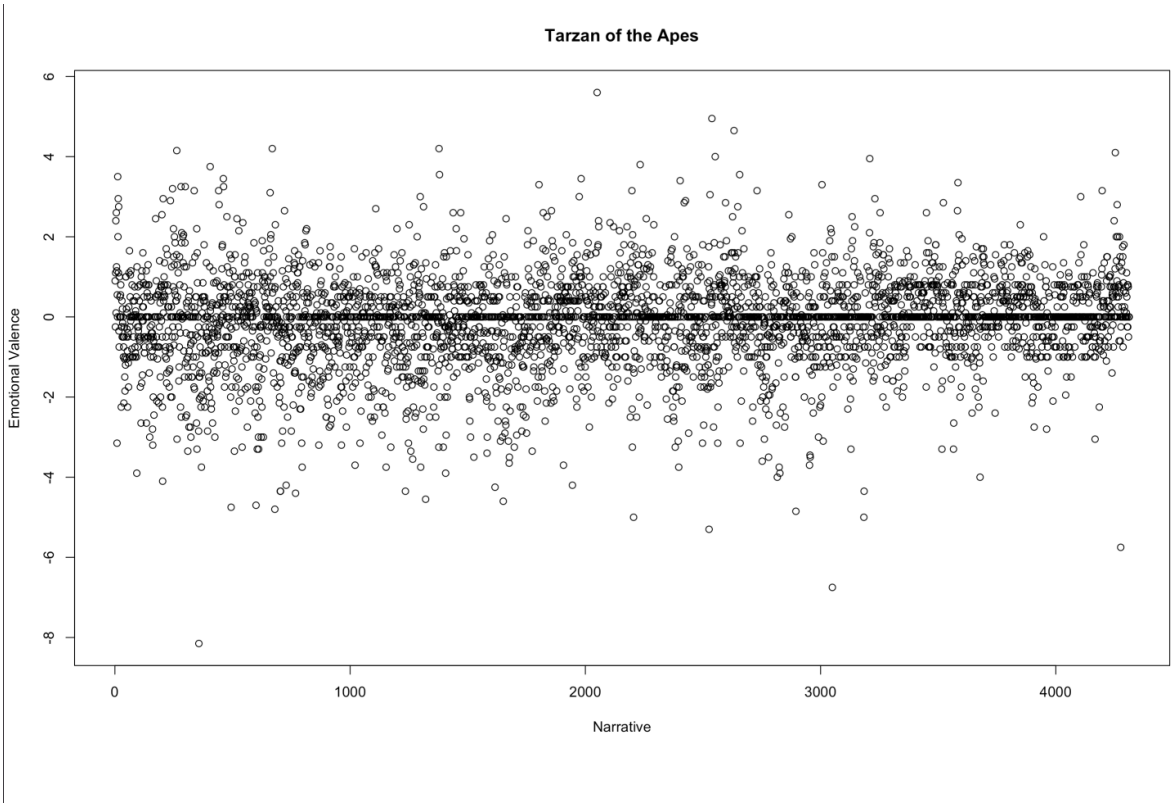
```
[1] -909
> #mean of sentiment score for syuzhet method
> CORP_Mean_syuzhet <- mean(CORP_syuzhet)
> CORP_Mean_syuzhet
[1] -0.1089402
> #mean of sentiment score for nrc method
> CORP_Mean_nrc <- mean(CORP_nrc)
> CORP_Mean_nrc
[1] 0
> #mean of sentiment score for bing method
> CORP_Mean_bing <- mean(CORP_bing)
> CORP_Mean_bing
[1] -0.2743506
> #mean of sentiment score for afinn method
> CORP_Mean_afinn <- mean(CORP_afinn)
> CORP_Mean_afinn
[1] -0.2108071
>
```

We can then examine the summary() results of the sentiments generated from these dictionaries to understand the distribution of sentiment.

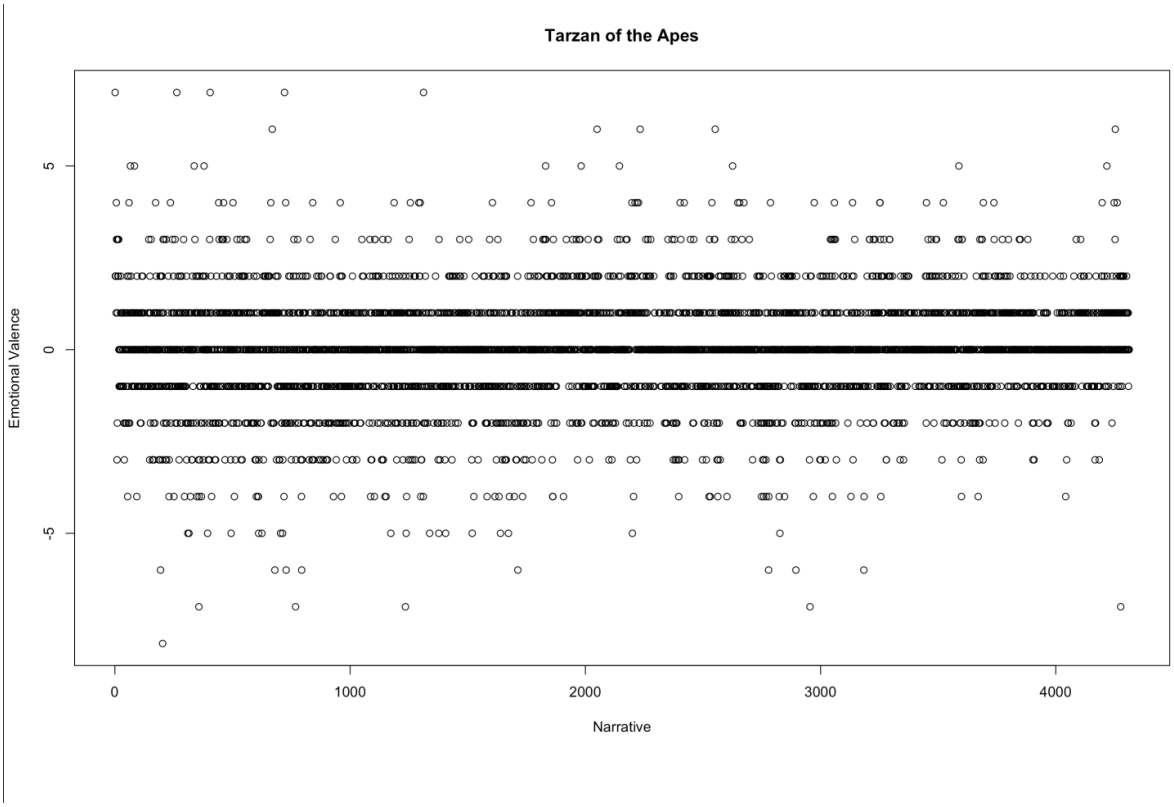
```
> #summaries of the sentiments generated
> summary(CORP_syuzhet)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.1500 -0.6000  0.0000 -0.1089  0.5000  5.6000
> summary(CORP_nrc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     -8      -1       0       0       1       7
> summary(CORP_bing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-9.0000 -1.0000  0.0000 -0.2744  0.0000  6.0000
> summary(CORP_afinn)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-20.0000  -1.0000   0.0000  -0.2108   1.0000  16.0000
>
```

However, to further understand how the sentiment of text changes from the beginning of the text till the end, let us plot the values.

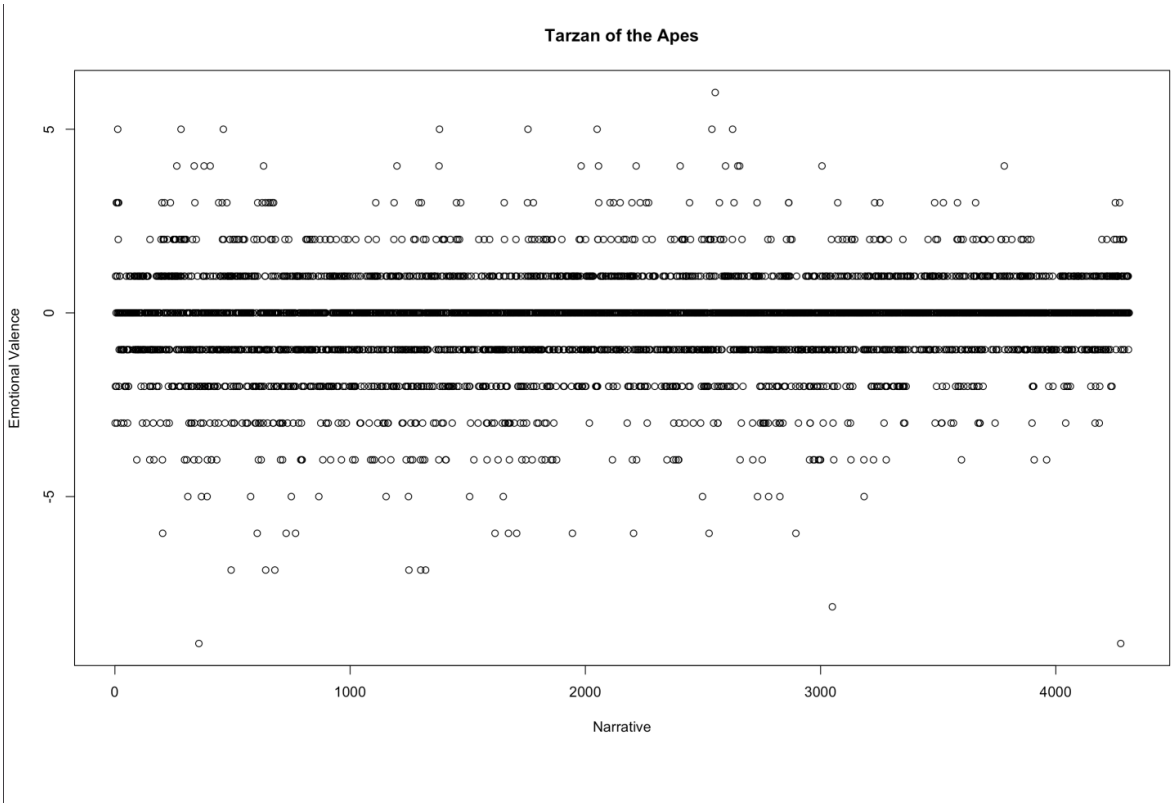
Syuzhet: plot(CORP_syuzhet, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")



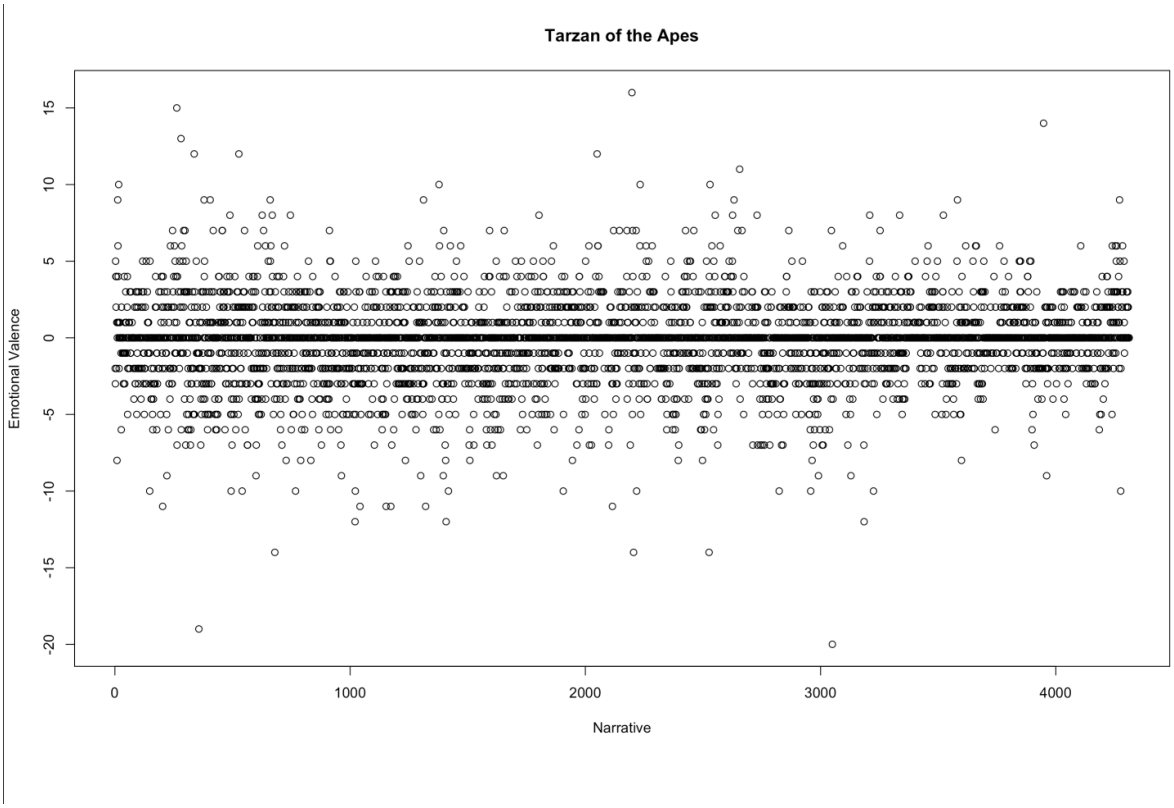
NRC: plot(CORP_nrc, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")



Bing plot(CORP_bing, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")



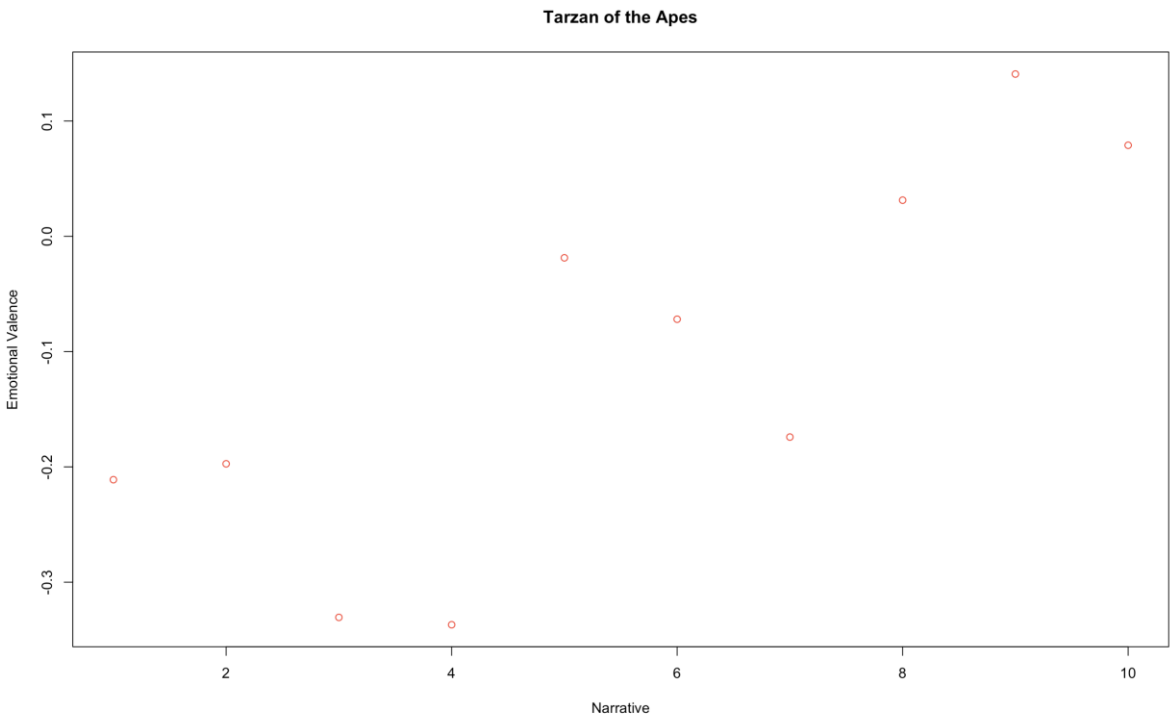
Afinn: plot(CORP_afinn, main="Tarzan of the Apes", xlab="Narrative", ylab="Emotional Valence")



Let us now use the get_percentage_values() function from the syuzhet package to compare the shape of one trajectory to another. We have plotted values for 10, 20 and 50 bins respectively.

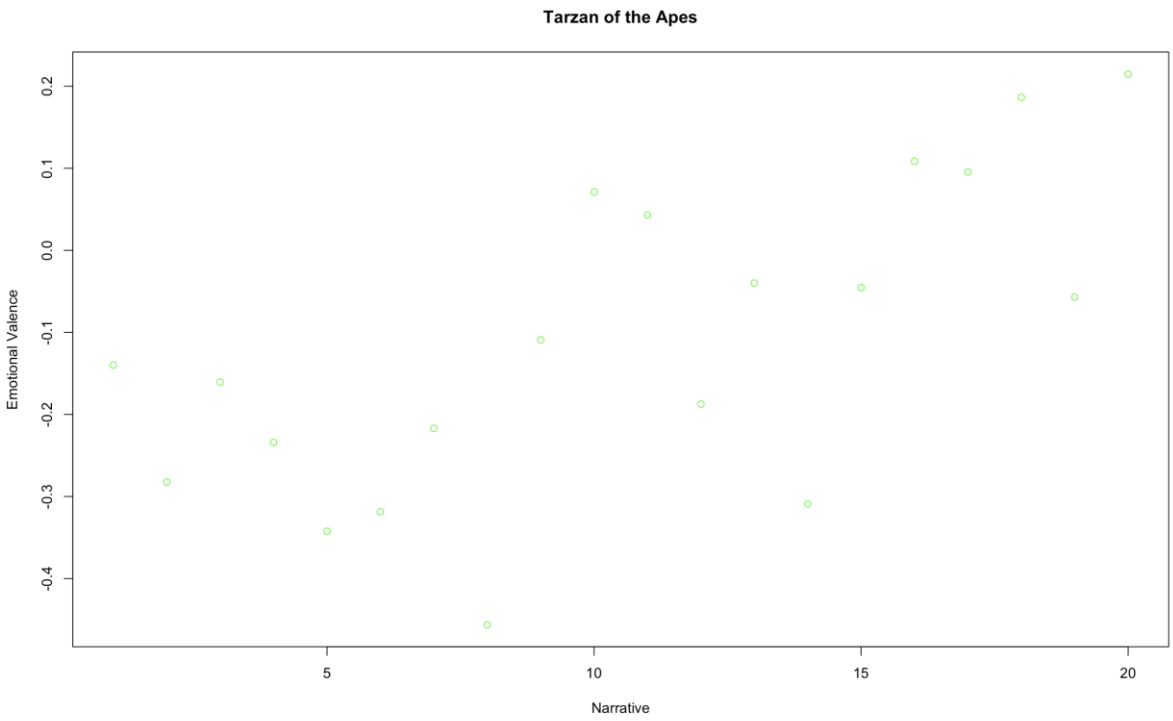
10 bins

```
> #plotting the distribution of emotional valence, grouped by bins
> CORP_SentimentPctValue <- get_percentage_values(CORP_syuzhet , bins=10)
> structure(CORP_SentimentPctValue)
          1           2           3           4           5           6           7           8
-0.21111111 -0.19733179 -0.33051044 -0.33689095 -0.01867749 -0.07192575 -0.17412993  0.03132251
          9          10
 0.14071926  0.07893519
> plot(CORP_SentimentPctValue, main="Tarzan of the Apes" , xlab="Narrative" , ylab="Emotional Valence" , col
="red")
>
```



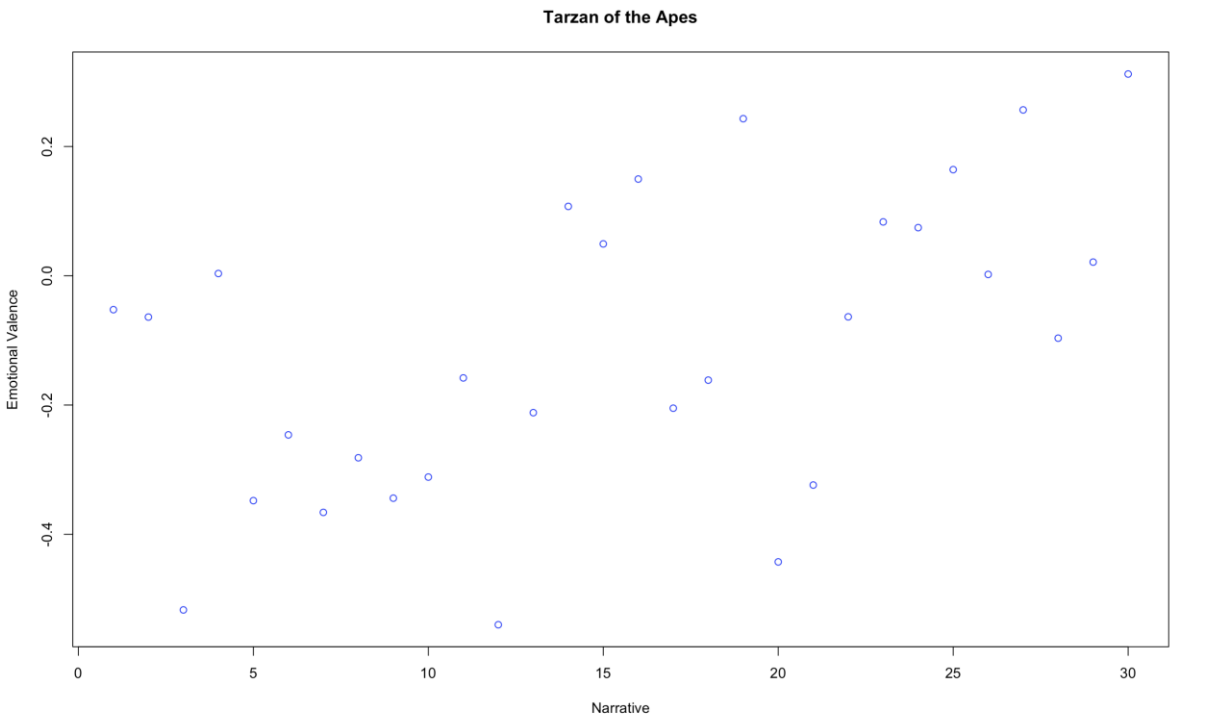
20 bins

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



30 bins

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



Through these plots, we understand that the sentiment shifts a lot throughout the text, from being very high in the beginning, to low in the middle, and high again towards the end. This can imply volatile and more emotional content in the book.

Discussion

This project gave us practical, hands-on experience using Text Analytics. Following the rubric, we had a path to follow for this project, yet the challenges along the way, such as dealing with multiple documents, or accommodating large text, reducing sparsity etc. are all concepts we learned while trying to solve a problem. For the same reason, we now understand why Data Science is an empirical science. We learned about data preprocessing and how important it is to clean the data before performing the analysis, because stopwords, punctuation, quote marks, and other elements do not contribute to the analysis and must be removed as they may skew the analysis we are attempting to perform. We learned various packages for Text Analytics and the interesting capabilities they support, including creating wordclouds, retrieving information about the text, assigning sentiments etc. Through all these bits and pieces of analysis, in the end we are able to gain a somewhat rough image of the text we are looking at.

The Text Analytics project was a fantastic learning experience, and as the third project, it built on the information we obtained from the prior ones. It helps to solidify analytical methodologies, cognitive processes, and troubleshooting. It also demonstrated the ease with which R can be used to execute high-quality, comprehensive data analytics.

To summarize, after completing this project, we are confident in working with R, performing text analytics on large datasets, cleaning data, extracting useful information, generating plots, wordclouds, understanding sentiments, and applying different functions and measures/metrics

to documents that allow us to draw different inferences about the problem domain, i.e. the text, as well as simplifying text for better analysis.