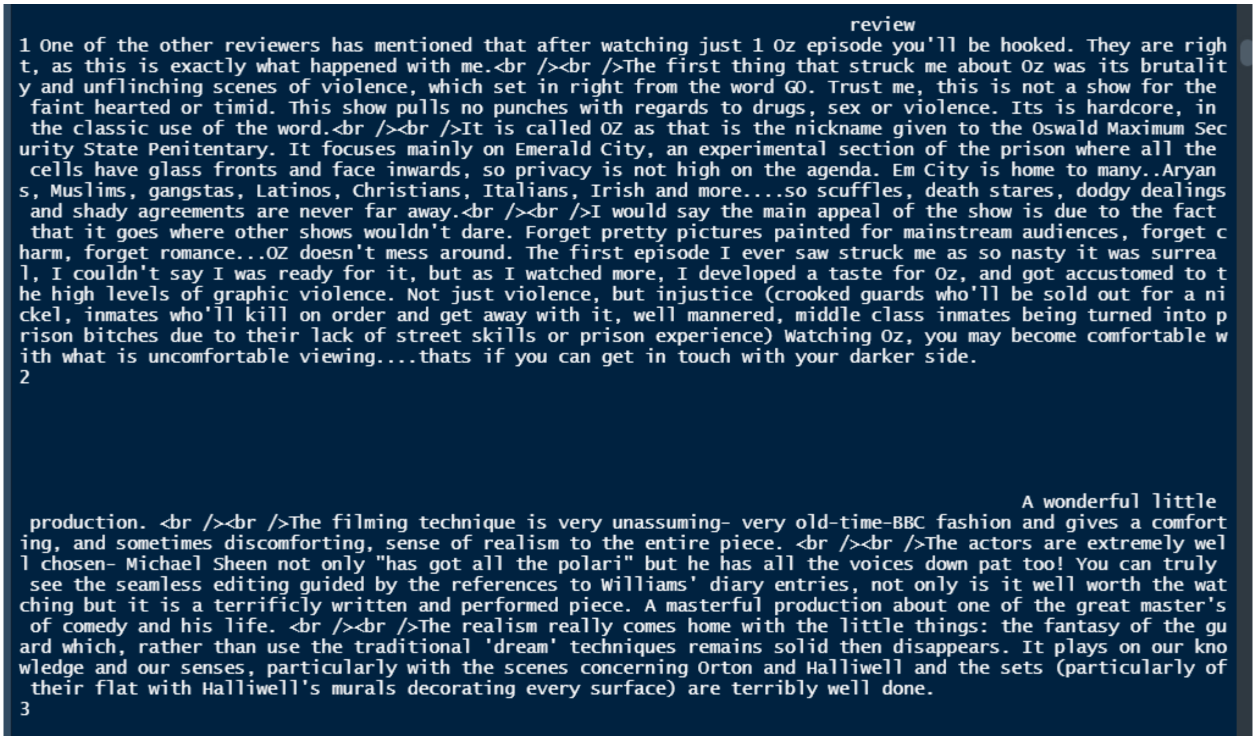
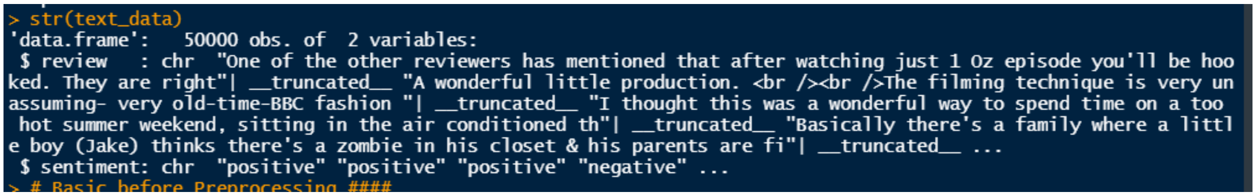
Text mining

```
# install.packages("tm")         # For text mining

# install.packages("textclean")  # For text cleaning

# install.packages("wordcloud")  # For word cloud visualization

# install.packages("SnowballC")  # For stemming

# install.packages("ggplot2")    # For data visualization


library(tm)

library(textclean)

library(wordcloud)

library(SnowballC)

library(ggplot2)


text_data <- read.csv("IMDB Dataset.csv")

head(text_data)  # Displays the first few rows of the dataset
```

Output:



```
                                                                                    review
1 One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are righ
t, as this is exactly what happened with me.<br /><br />The first thing that struck me about Oz was its brutalit
y and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the
 faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in
 the classic use of the word.<br /><br />It is called OZ as that is the nickname given to the Oswald Maximum Sec
urity State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the
 cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryan
s, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings
 and shady agreements are never far away.<br /><br />I would say the main appeal of the show is due to the fact
 that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget c
harm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surrea
l, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to t
he high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a ni
ckel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into p
rison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable w
ith what is uncomfortable viewing....thats if you can get in touch with your darker side.
2





                                                                                                   A wonderful little
 production. <br /><br />The filming technique is very unassuming- very old-time-BBC fashion and gives a comfort
ing, and sometimes discomforting, sense of realism to the entire piece. <br /><br />The actors are extremely wel
l chosen- Michael Sheen not only "has got all the polari" but he has all the voices down pat too! You can truly
 see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the wat
ching but it is a terrificly written and performed piece. A masterful production about one of the great master's
 of comedy and his life. <br /><br />The realism really comes home with the little things: the fantasy of the gu
ard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our kno
wledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of
 their flat with Halliwell's murals decorating every surface) are terribly well done.
3
```



```
> str(text_data)
'data.frame':	50000 obs. of  2 variables:
 $ review   : chr  "One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hoo
ked. They are right"| __truncated__ "A wonderful little production. <br /><br />The filming technique is very un
assuming- very old-time-BBC fashion "| __truncated__ "I thought this was a wonderful way to spend time on a too
 hot summer weekend, sitting in the air conditioned th"| __truncated__ "Basically there's a family where a littl
e boy (Jake) thinks there's a zombie in his closet & his parents are fi"| __truncated__ ...
 $ sentiment: chr  "positive" "positive" "positive" "negative" ...
> # Basic before Preprocessing ####
```

str(text_data)   # Shows the structure of the dataset

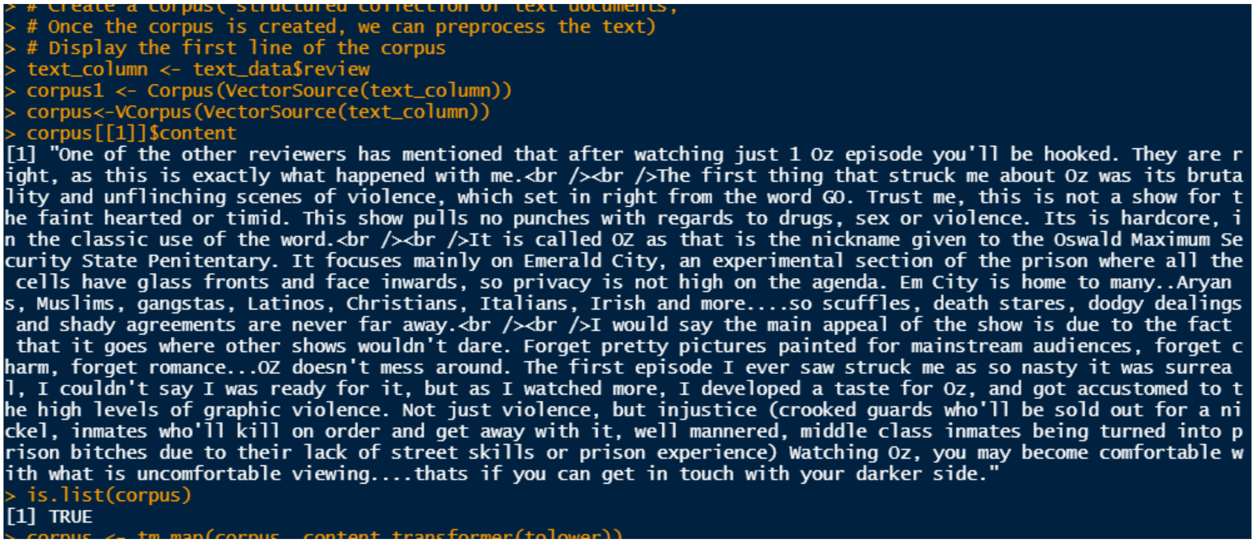
text_column <- text_data$review # Extract the review column

corpus1 <- Corpus(VectorSource(text_column)) # Create a corpus (collection of text docs)

corpus <- VCorpus(VectorSource(text_column)) # Another corpus format

corpus[[1]]$content # Displays the first document in the corpus

is.list(corpus) # Checks if the corpus is stored as a list



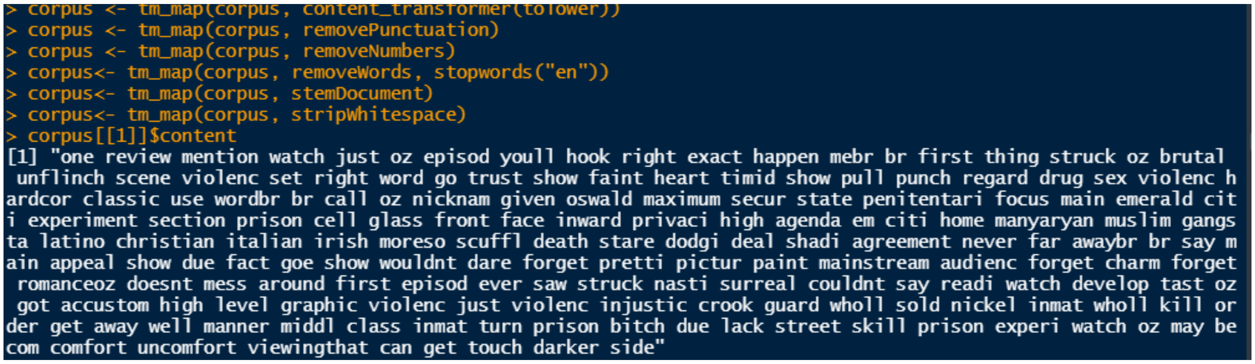
```
> # Create a corpus( structured collection of text documents)
> # Once the corpus is created, we can preprocess the text)
> # Display the first line of the corpus
> text_column <- text_data$review
> corpus1 <- Corpus(VectorSource(text_column))
> corpus<-VCorpus(VectorSource(text_column))
> corpus[[1]]$content
[1] "One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are r
ight, as this is exactly what happened with me.<br /><br />The first thing that struck me about Oz was its bruta
lity and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for t
he faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, i
n the classic use of the word.<br /><br />It is called OZ as that is the nickname given to the Oswald Maximum Se
curity State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the
 cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryan
s, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings
 and shady agreements are never far away.<br /><br />I would say the main appeal of the show is due to the fact
 that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget c
harm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surrea
l, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to t
he high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a ni
ckel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into p
rison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable w
ith what is uncomfortable viewing....thats if you can get in touch with your darker side."
> is.list(corpus)
[1] TRUE
> corpus <- tm_map(corpus, content_transformer(tolower))
```

corpus <- tm_map(corpus, content_transformer(tolower)) # Convert text to lowercase

corpus <- tm_map(corpus, removePunctuation) # Remove punctuation

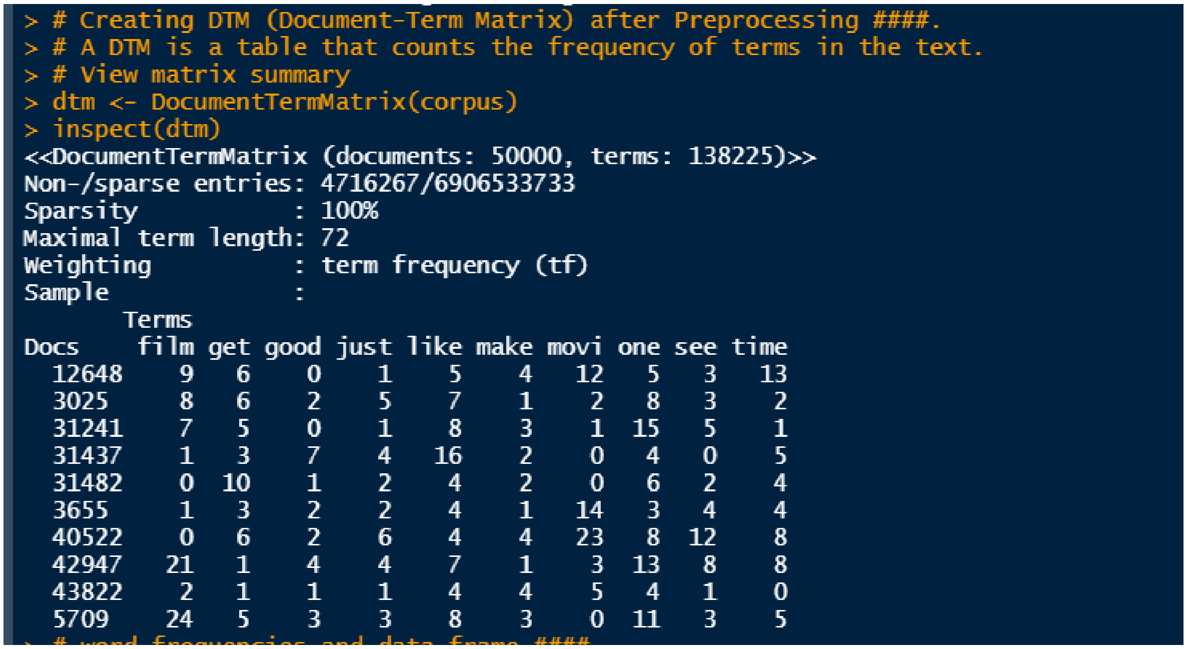
corpus <- tm_map(corpus, removeNumbers) # Remove numbers

```
corpus <- tm_map(corpus, removeWords, stopwords("en"))  # Remove common stopwords

corpus <- tm_map(corpus, stemDocument)  # Apply stemming (reduce words to their base form)

corpus <- tm_map(corpus, stripWhitespace)  # Remove extra spaces

corpus[[1]]$content  # Display the first processed document
```



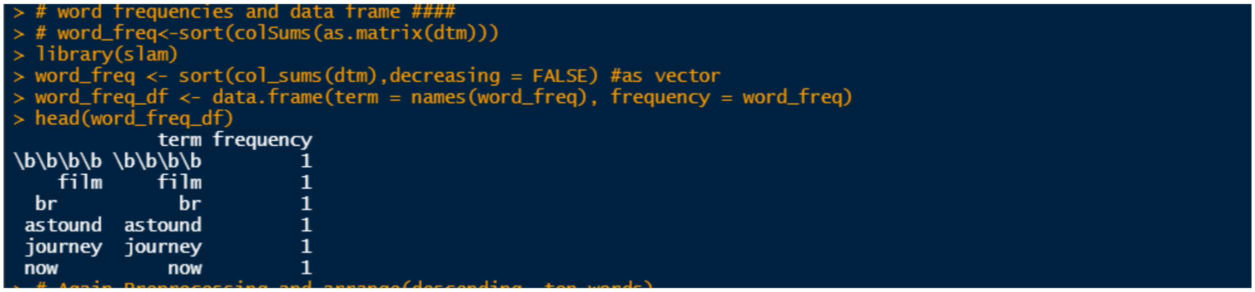
```
> corpus <- tm_map(corpus, content_transformer(tolower))
> corpus <- tm_map(corpus, removePunctuation)
> corpus <- tm_map(corpus, removeNumbers)
> corpus<- tm_map(corpus, removeWords, stopwords("en"))
> corpus<- tm_map(corpus, stemDocument)
> corpus<- tm_map(corpus, stripWhitespace)
> corpus[[1]]$content
[1] "one review mention watch just oz episod youll hook right exact happen mebr br first thing struck oz brutal
 unflinch scene violenc set right word go trust show faint heart timid show pull punch regard drug sex violenc h
ardcor classic use wordbr br call oz nicknam given oswald maximum secur state penitentari focus main emerald cit
i experiment section prison cell glass front face inward privaci high agenda em citi home manyaryan muslim gangs
ta latino christian italian irish moreso scuffl death stare dodgi deal shadi agreement never far awaybr br say m
ain appeal show due fact goe show wouldnt dare forget pretti pictur paint mainstream audienc forget charm forget
 romanceoz doesnt mess around first episod ever saw struck nasti surreal couldnt say readi watch develop tast oz
 got accustom high level graphic violenc just violenc injustic crook guard wholl sold nickel inmat wholl kill or
der get away well manner middl class inmat turn prison bitch due lack street skill prison experi watch oz may be
com comfort uncomfort viewingthat can get touch darker side"
```

```
dtm <- DocumentTermMatrix(corpus)  # Create the DTM

inspect(dtm)  # View summary of the DTM
```



```
> # Creating DTM (Document-Term Matrix) after Preprocessing ####.
> # A DTM is a table that counts the frequency of terms in the text.
> # View matrix summary
> dtm <- DocumentTermMatrix(corpus)
> inspect(dtm)
<<DocumentTermMatrix (documents: 50000, terms: 138225)>>
Non-/sparse entries: 4716267/6906533733
Sparsity           : 100%
Maximal term length: 72
Weighting          : term frequency (tf)
Sample             :
       Terms
Docs    film get good just like make movi one see time
  12648    9   6    0    1    5    4   12   5   3   13
  3025     8   6    2    5    7    1    2   8   3    2
  31241    7   5    0    1    8    3    1  15   5    1
  31437    1   3    7    4   16    2    0   4   0    5
  31482    0  10    1    2    4    2    0   6   2    4
  3655     1   3    2    2    4    1   14   3   4    4
  40522    0   6    2    6    4    4   23   8  12    8
  42947   21   1    4    4    7    1    3  13   8    8
  43822    2   1    1    1    4    4    5   4   1    0
  5709    24   5    3    3    8    3    0  11   3    5
> # word frequencies and data frame ####
```

```
library(slam)

word_freq <- sort(col_sums(dtm), decreasing = FALSE)  # Compute word frequencies

word_freq_df <- data.frame(term = names(word_freq), frequency = word_freq)  # Convert to data
frame

head(word_freq_df)  # Show the first few rows
```

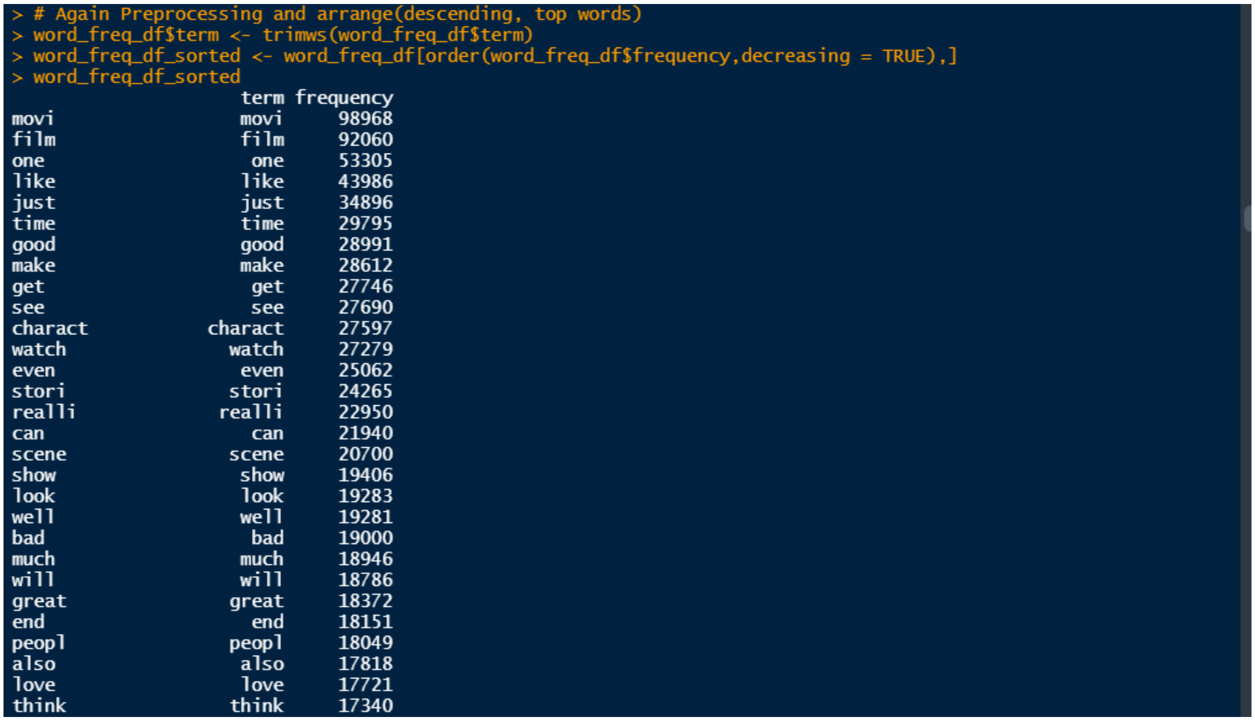


```
> # word frequencies and data frame ####
> # word_freq<-sort(colSums(as.matrix(dtm)))
> library(slam)
> word_freq <- sort(col_sums(dtm),decreasing = FALSE) #as vector
> word_freq_df <- data.frame(term = names(word_freq), frequency = word_freq)
> head(word_freq_df)
                term frequency
\b\b\b\b \b\b\b\b         1
    film     film         1
  br           br         1
 astound  astound         1
 journey  journey         1
 now          now         1
```

word_freq_df$term <- trimws(word_freq_df$term) # Trim whitespace

word_freq_df_sorted <- word_freq_df[order(word_freq_df$frequency, decreasing = TRUE),] # Sort in descending order

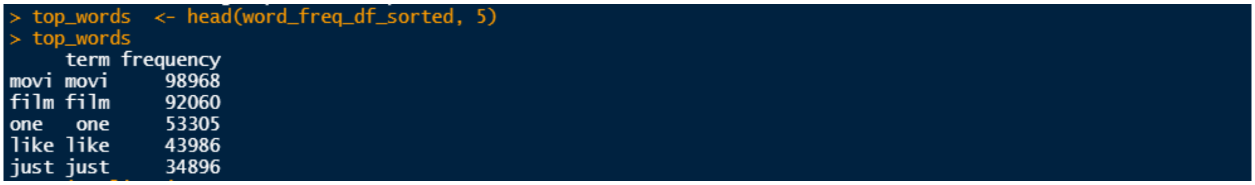
word_freq_df_sorted # Display sorted words



```
> # Again Preprocessing and arrange(descending, top words)
> word_freq_df$term <- trimws(word_freq_df$term)
> word_freq_df_sorted <- word_freq_df[order(word_freq_df$frequency,decreasing = TRUE),]
> word_freq_df_sorted
                          term frequency
movi                      movi     98968
film                      film     92060
one                        one     53305
like                      like     43986
just                      just     34896
time                      time     29795
good                      good     28991
make                      make     28612
get                        get     27746
see                        see     27690
charact                charact     27597
watch                    watch     27279
even                      even     25062
stori                    stori     24265
realli                  realli     22950
can                        can     21940
scene                    scene     20700
show                      show     19406
look                      look     19283
well                      well     19281
bad                        bad     19000
much                      much     18946
will                      will     18786
great                    great     18372
end                        end     18151
peopl                    peopl     18049
also                      also     17818
love                      love     17721
think                    think     17340
```

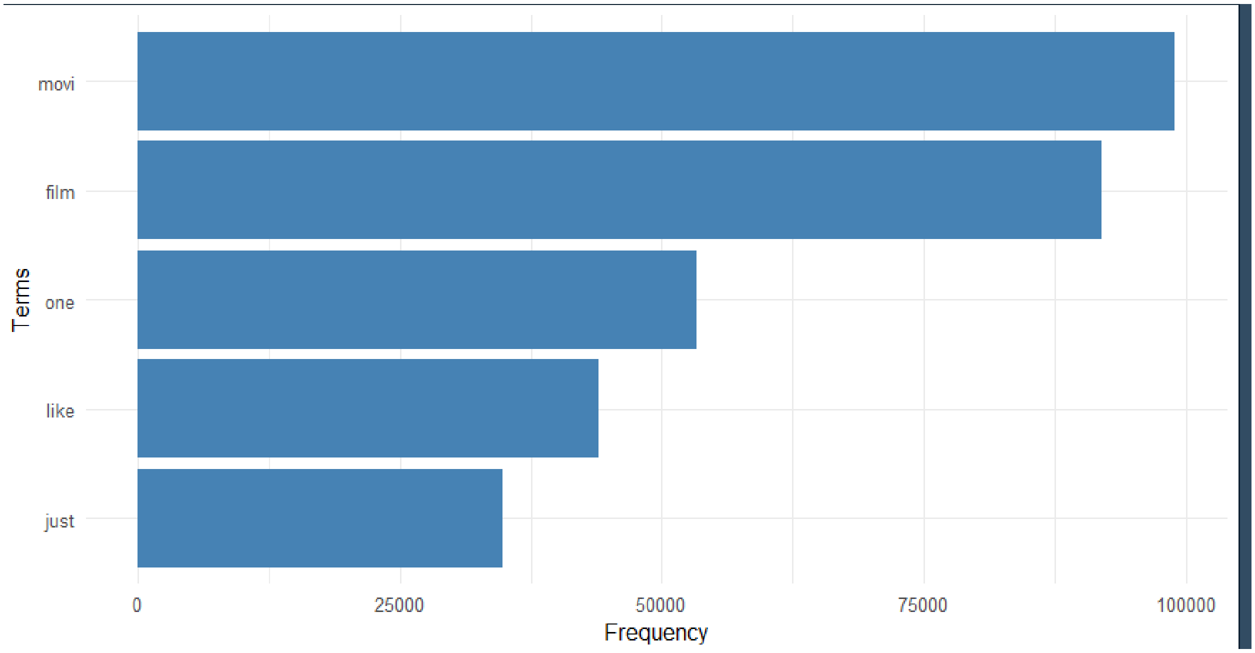
top_words <- head(word_freq_df_sorted, 5) # Select the top 5 most frequent words

top_words

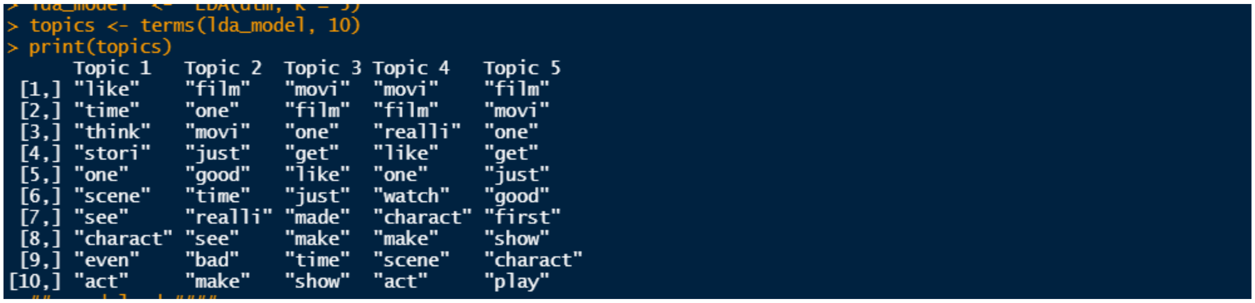


```
> top_words  <- head(word_freq_df_sorted, 5)
> top_words
     term frequency
movi movi     98968
film film     92060
one   one     53305
like like     43986
just just     34896
```

```
ggplot(top_words, aes(x = reorder(term, frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
    coord_flip() +
      theme_minimal() +
        labs(x = "Terms", y = "Frequency")
```



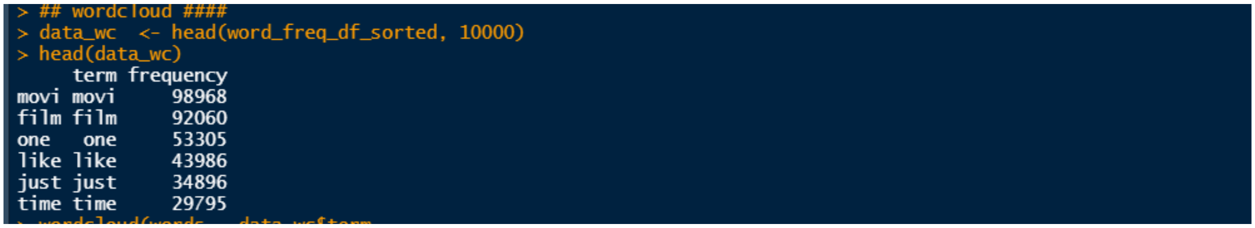
```
library(topicmodels)
lda_model  <-  LDA(dtm, k = 5)  # Apply LDA with 5 topics
topics <- terms(lda_model, 10)  # Extract top 10 terms per topic
print(topics)
```



```
> lda_model  <-  LDA(dtm, k = 5)
> topics <- terms(lda_model, 10)
> print(topics)
      Topic 1   Topic 2  Topic 3 Topic 4   Topic 5
 [1,] "like"    "film"   "movi"  "movi"    "film"
 [2,] "time"    "one"    "film"  "film"    "movi"
 [3,] "think"   "movi"   "one"   "realli"  "one"
 [4,] "stori"   "just"   "get"   "like"    "get"
 [5,] "one"     "good"   "like"  "one"     "just"
 [6,] "scene"   "time"   "just"  "watch"   "good"
 [7,] "see"     "realli" "made"  "charact" "first"
 [8,] "charact" "see"    "make"  "make"    "show"
 [9,] "even"    "bad"    "time"  "scene"   "charact"
[10,] "act"     "make"   "show"  "act"     "play"
```

data_wc <- head(word_freq_df_sorted, 10000) # Use the top 10,000 words

head(data_wc)



```
> ## wordcloud ####
> data_wc  <- head(word_freq_df_sorted, 10000)
> head(data_wc)
     term frequency
movi movi     98968
film film     92060
one   one     53305
like like     43986
just just     34896
time time     29795
```

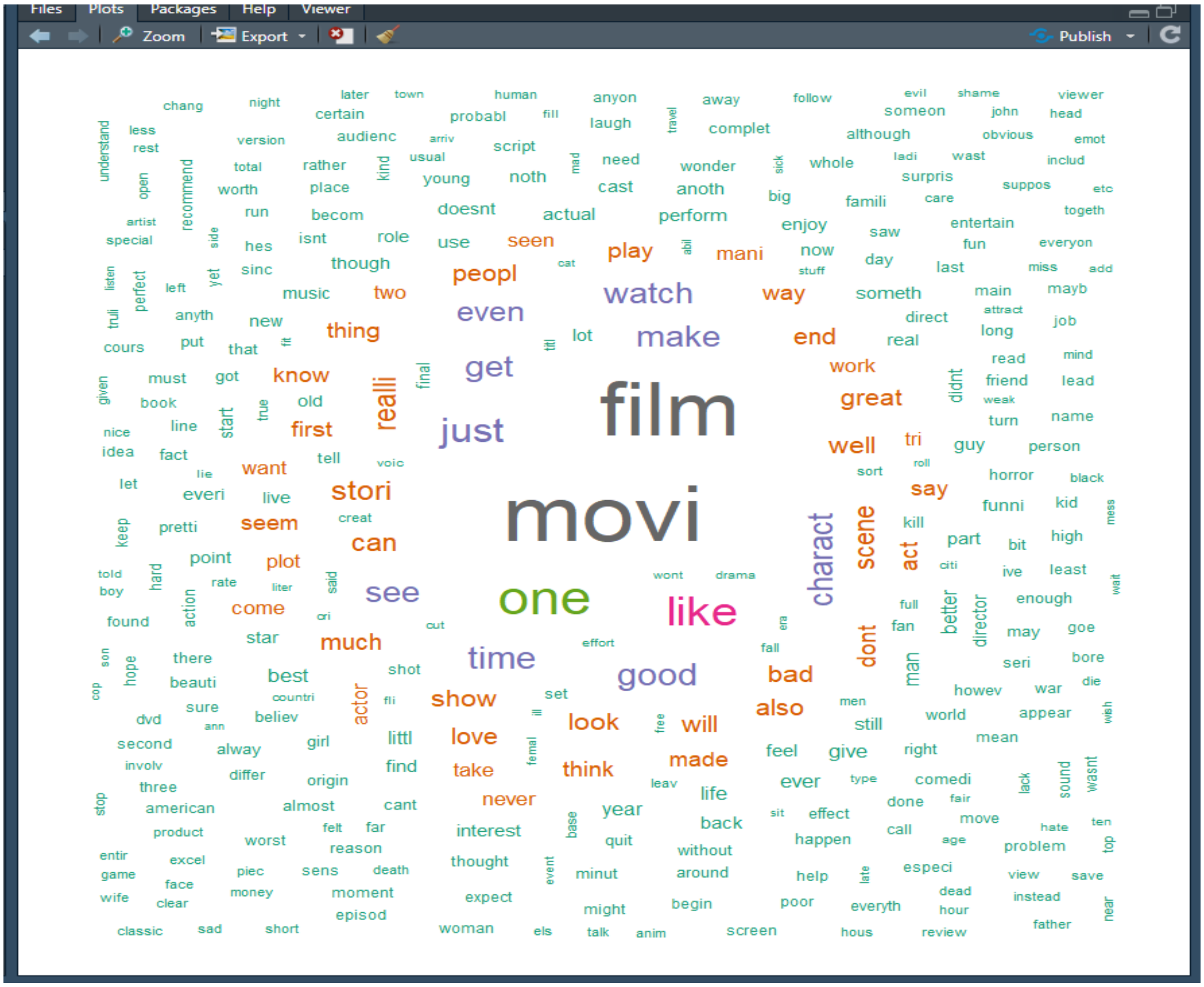
wordcloud(words = data_wc$term,

freq = data_wc$frequency,

max.words = 1000,

random.order = FALSE,

colors = brewer.pal(8, "Dark2"))



Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
chang night later town human anyon away follow evil shame viewer
certain probabl fill laugh travel complet someon john head
understand less rest version audienc arriv script although obvious emot
total rather kind usual mad need wonder sick whole ladi wast includ
open recommend worth place young noth cast anoth surpris suppos etc
run becom doesnt actual perform big famili care togeth
artist special side hes isnt role use seen enjoy saw entertain
listen perfect left yet sinc though peopl cat play abil mani now day fun everyon
truli music two stuff last miss add
anyth new even watch way someth main mayb
cours put that fit thing titl lot make end direct attract long job
given must got know reall final get real read mind
book old work friend lead
nice line start true first just film great didnt weak turn name
idea fact tell voic well tri guy person
let lie want everi live stori sort roll horror black
keep pretti seem creat can movi say funni kid mess
told boy hard point plot rate liter said charact scene act kill part bit high
action see one wont drama citi ive least wait
come ori cut like full better director enough
found star much era fan may goe
son hope there effort time fall dont man seri bore
cop beauti best shot good bad howev war die
sure countri actor fli set show ill men
dvd ann believ look free will also still world appear wish
second alway girl littl love femal mean
involv differ find take think made feel give right
three origin never leav life ever type comedi lack sound wasnt
stop american almost cant year done fair
product felt far interest base back sit effect move hate ten
worst reason quit happen call age problem top
entir excel piec sens death thought event minut without around help late especi view save
game face money moment expect might begin poor dead instead
wife clear episod everyth hour near
classic sad short woman els talk anim screen hous review father