Practical -8

Write a program for measuring similarity among documents and detecting passages which have been reused.

Installation of required packages before executing program:-

```
install.packages("tm")
require("tm")
install.packages("ggplot2")
install.packages("textreuse")
install.packages("devtools")
```

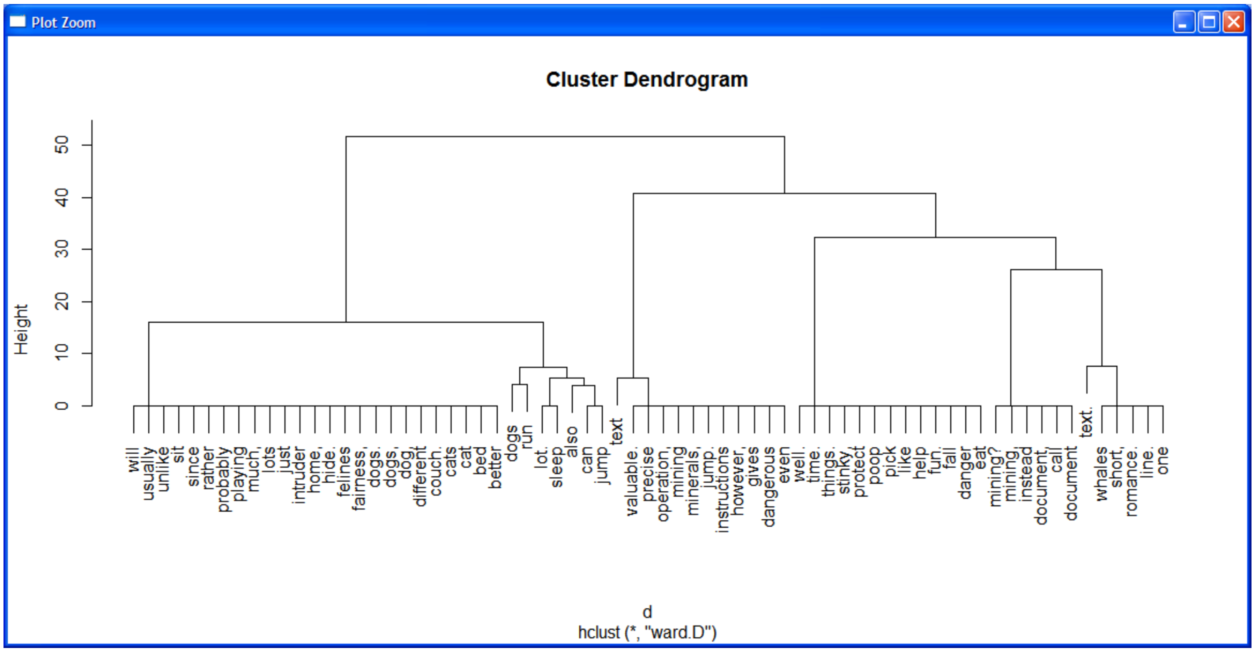
Source Code 1:-

```
my.corpus <- Corpus(DirSource("c:/msc/r-corpus"))

my.corpus <- tm_map(my.corpus, removeWords, stopwords("english"))

my.tdm <- TermDocumentMatrix(my.corpus)

#inspect(my.tdm)

my.dtm <- DocumentTermMatrix(my.corpus, control = list(weighting =
weightTfIdf, stopwords = TRUE))

#inspect(my.dtm)

my.df <- as.data.frame(inspect(my.tdm))

my.df.scale <- scale(my.df)

d <- dist(my.df.scale,method="euclidean")

fit <- hclust(d, method="ward")

plot(fit)
```

OutPut:-

```
<<TermDocumentMatrix (terms: 69, documents: 6)>>
Non-/sparse entries: 97/317
Sparsity           : 77%
Maximal term length: 12
Weighting          : term frequency (tf)

              Docs
Terms          File1.txt File2.txt File3.txt File4.txt File5.txt File6.txt
  also                 0         1         1         1         0         0
  bed                  0         0         0         1         0         0
  better               0         0         0         1         0         0
  call                 0         1         0         0         0         0
  can                  0         0         1         1         0         0
  cat                  0         0         0         1         0         0
  cats                 0         0         0         1         0         0
  couch.               0         0         0         1         0         0
  danger               0         0         1         0         0         0
  dangerous            0         0         0         0         1         1
  different            0         0         0         1         0         0
  document             0         1         0         0         0         0
  document,            0         1         0         0         0         0
  dog,                 0         0         0         1         0         0
  dogs                 0         0         3         1         0         0
  dogs,                0         0         0         1         0         0
  dogs.                0         0         0         1         0         0
  eat                  0         0         1         0         0         0
  even                 0         0         0         0         1         1
  fairness,            0         0         0         1         0         0
  fall                 0         0         1         0         0         0
  felines              0         0         0         1         0         0
  fun.                 0         0         1         0         0         0
  gives                0         0         0         0         1         1
  help                 0         0         1         0         0         0
  hide.                0         0         0         1         0         0
  home,                0         0         0         1         0         0
  however,             0         0         0         0         1         1
  instead              0         1         0         0         0         0
  instructions         0         0         0         0         1         1
  intruder             0         0         0         1         0         0
  jump                 0         0         1         1         0         0
  jump.                0         0         0         0         1         1
  just                 0         0         0         1         0         0
  like                 0         0         1         0         0         0
  line.                1         0         0         0         0         1
  lot.                 0         0         0         2         0         0
  lots                 0         0         0         1         0         0
  minerals,            0         0         0         0         1         1
  mining               0         0         0         0         1         1
  mining,              0         1         0         0         0         0
  mining?              0         1         0         0         0         0
  much,                0         0         0         1         0         0
  one                  1         0         0         0         0         1
  operation,           0         0         0         0         1         1
  pick                 0         0         1         0         0         0
  playing         0         0         0         1         0         0
  poop            0         0         1         0         0         0
  precise         0         0         0         0         1         1
  probably        0         0         0         1         0         0
  protect         0         0         1         0         0         0
  rather          0         0         0         1         0         0
  romance.        1         0         0         0         0         1
  run             0         0         2         2         1         1
  short,          1         0         0         0         0         1
  since           0         0         0         1         0         0
  sit             0         0         0         1         0         0
  sleep           0         0         0         2         0         0
  stinky,         0         0         1         0         0         0
  text            0         1         0         0         1         1
  text.           2         1         0         0         0         2
  things.         0         0         1         0         0         0
  time.           0         0         1         0         0         0
  unlike          0         0         0         1         0         0
  usually         0         0         0         1         0         0
  valuable.       0         0         0         0         1         1
  well.           0         0         1         0         0         0
  whales          1         0         0         0         0         1
  will            0         0         0         1         0         0
> barplot(as.matrix(my.tdm))
```

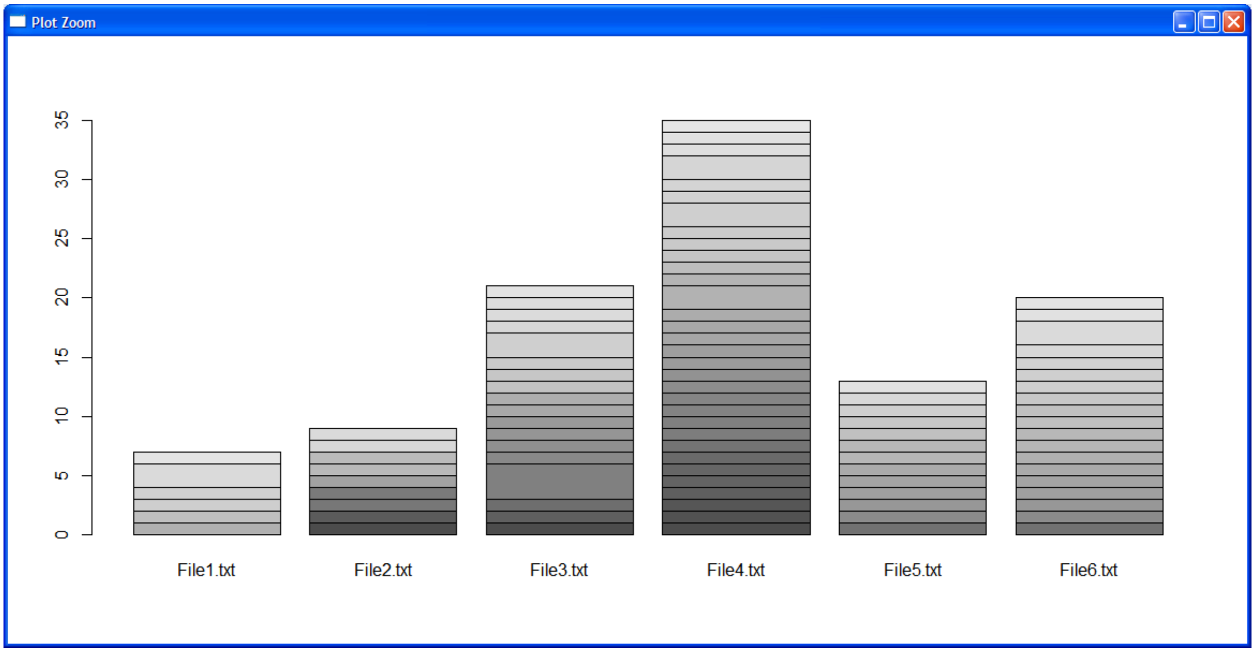
```
> my.df.scale <- scale(my.df)
> d <- dist(my.df.scale,method="euclidean")
> fit <- hclust(d, method="ward")
The "ward" method has been renamed to "ward.D"; note new "ward.D2"

> plot(fit)
```


Plot Zoom
Cluster Dendrogram
Height
0
10
20
30
40
50
will
usually
unlike
sit
since
rather
probably
playing
much,
lots
just
intruder
home,
hide.
felines
fairness,
dogs.
dogs,
dog,
different
couch.
cats
cat
bed
better
dogs
run
lot.
sleep
also
can
jump
text
valuable.
precise
operation,
mining
minerals,
jump.
instructions
however,
gives
dangerous
even
well.
time.
things.
stinky,
protect
poop
pick
like
help
fun.
fall
danger
eat
mining?
mining,
instead
document,
call
document
text.
whales
short,
romance.
line.
one
d
hclust (*, "ward.D")

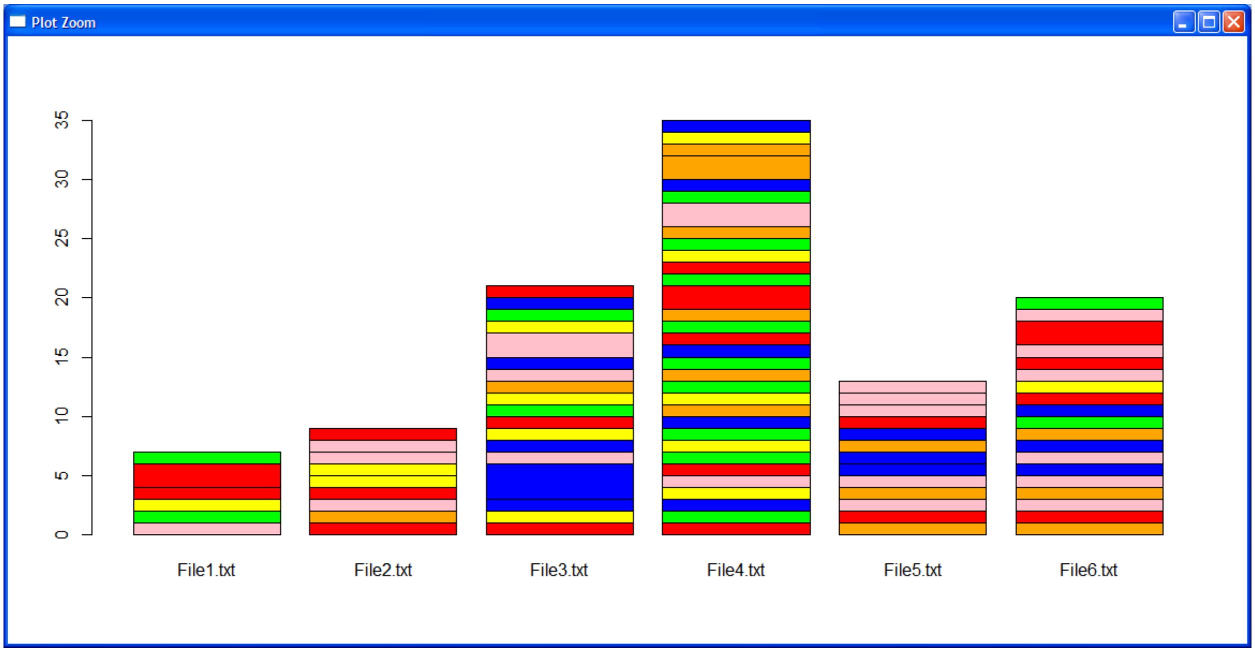
Sourec code 2 (using bar plot with and without color):-

```
my.corpus <- Corpus(DirSource("c:/msc/r-corpus"))

my.corpus <- tm_map(my.corpus, removeWords, stopwords("english"))

my.tdm <- TermDocumentMatrix(my.corpus)

inspect(my.tdm)

my.df <- as.data.frame(inspect(my.tdm))

barplot(as.matrix(my.tdm))

#barplot(as.matrix(my.tdm),col = color)
```

OutPut:-



barplot(as.matrix(my.tdm),col = color)



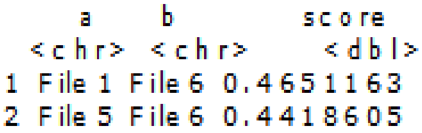
Sourec code 3 (using minhash and jaccard similarity):-

library(textreuse)

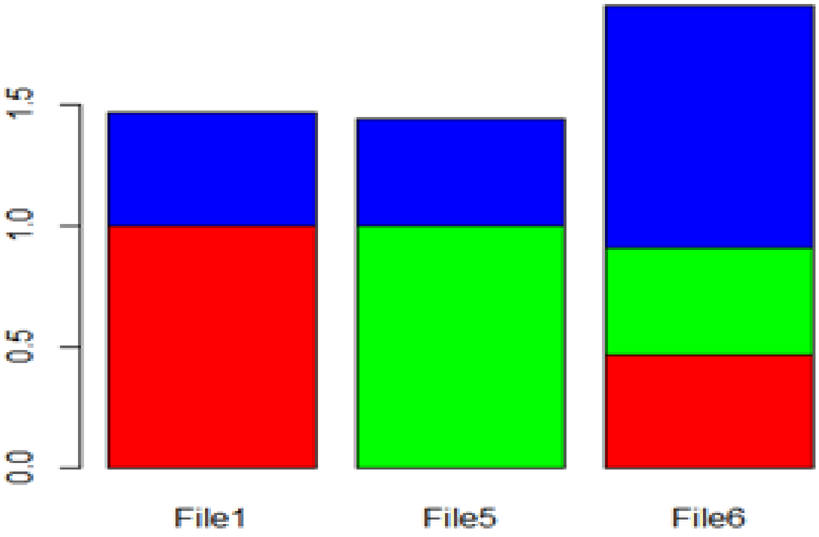
Source Code:-

```
minhash <- minhash_generator(200, seed = 235)
ats <- TextReuseCorpus(dir = "c:/msc/r-corpus", tokenizer = tokenize_ngrams, n =
5, minhash_func = minhash)
buckets <- lsh(ats, bands = 50, progress = interactive())
candidates <- lsh_candidates(buckets)
scores <- lsh_compare(candidates, ats, jaccard_similarity, progress = FALSE)
scores
color <- c("red","green","blue","orange","yellow","pink")
barplot(as.matrix(scores),col = color)
```

Output:



```
      a     b         score
  <chr> <chr>         <dbl>
1 File1 File6 0.4651163
2 File5 File6 0.4418605
```


File1
File5
File6