# Natural language processing and network analysis

## BY: LIM YU JIN

install necessary libraries

```
library(factoextra)
library(slam)
library(tm)
library(SnowballC)
library(igraph)
```

```
# get the file path and summary of the documents
cname = file.path(".","corpus")
docs = Corpus(DirSource((cname)))
summary(docs)
```

```
##           Length Class             Mode
## Doc1.txt  2      PlainTextDocument  list
## Doc10.txt 2      PlainTextDocument  list
## Doc11.txt 2      PlainTextDocument  list
## Doc12.txt 2      PlainTextDocument  list
## Doc13.txt 2      PlainTextDocument  list
## Doc14.txt 2      PlainTextDocument  list
## Doc15.txt 2      PlainTextDocument  list
## Doc2.txt  2      PlainTextDocument  list
## Doc3.txt  2      PlainTextDocument  list
## Doc4.txt  2      PlainTextDocument  list
## Doc5.txt  2      PlainTextDocument  list
## Doc6.txt  2      PlainTextDocument  list
## Doc7.txt  2      PlainTextDocument  list
## Doc8.txt  2      PlainTextDocument  list
## Doc9.txt  2      PlainTextDocument  list
```

Reference source of documents:

doc1 - Ro, C. (2022b, March 11). Why gig work is so hard to regulate. *BBC Worklife*. https://www.bbc.com/worklife/article/20220308-why-gig-work-is-so-hard-to-regulate (https://www.bbc.com/worklife/article/20220308-why-gig-work-is-so-hard-to-regulate)

doc2 - Khadka, B. N. S. (2023, May 29). Why Everest base camp won’t be moving anytime soon. *BBC News*. https://www.bbc.com/news/world-asia-65723447 (https://www.bbc.com/news/world-asia-65723447)

doc3 - McGrath, B. M. (2023, May 17). Global warming set to break key 1.5C limit for first time. *BBC News*. https://www.bbc.com/news/science-environment-65602293 (https://www.bbc.com/news/science-environment-65602293)

doc4 - Smale, B. D. G. a. W. (2023, May 14). Should social media face-altering filters be regulated? *BBC News*. https://www.bbc.com/news/business-65544054 (https://www.bbc.com/news/business-65544054)

doc5 - Schutz, B. E. (2023, May 1). The people turning time into a currency. *BBC News*. https://www.bbc.com/news/business-65397192 (https://www.bbc.com/news/business-65397192)

doc6 - Hern, A. (2020b, January 29). Gig economy traps workers in precarious existence, says report. *The Guardian*. https://www.theguardian.com/business/2020/jan/29/gig-economy-traps-workers-in-precarious-existence-says-report (https://www.theguardian.com/business/2020/jan/29/gig-economy-traps-workers-in-precarious-existence-says-report)

doc7 - Erway, C. (2023, May 26). Maryland Crab Sammy: a singular American sandwich. *BBC Travel*. https://www.bbc.com/travel/article/20230525-maryland-crab-sammy-a-singular-american-sandwich (https://www.bbc.com/travel/article/20230525-maryland-crab-sammy-a-singular-american-sandwich)

doc8 - Ramadurai, C. (2023, May 11). India's disappearing Chinese community. *BBC*. https://www.bbc.com/travel/article/20230511-indias-disappearing-chinese-community (https://www.bbc.com/travel/article/20230511-indias-disappearing-chinese-community)

doc9 - Galloway, L. (2022, March 30). The 333 islands opening to the world. *BBC Travel*. https://www.bbc.com/travel/article/20211214-the-333-islands-opening-to-the-world (https://www.bbc.com/travel/article/20211214-the-333-islands-opening-to-the-world)

doc10 - Robson, D. (2023, May 26). The languages that make maths easier. *BBC Future*. https://www.bbc.com/future/article/20230511-whats-the-best-language-for-learning-maths (https://www.bbc.com/future/article/20230511-whats-the-best-language-for-learning-maths)

doc11 - Hardach, S. (2023, April 10). Why do some people 'mirror-write'? *BBC*. Retrieved April 10, 2023, from https://www.bbc.com/future/article/20230405-why-do-some-people-mirror-write (https://www.bbc.com/future/article/20230405-why-do-some-people-mirror-write)

doc12 - Hardach, S. (2023b, May 4). How dyslexia changes in other languages. *BBC Future*. https://www.bbc.com/future/article/20230302-can-dyslexia-change-in-other-languages (https://www.bbc.com/future/article/20230302-can-dyslexia-change-in-other-languages)

doc13 – Renwick, D. (2023, May 26). Female electricians: a climate solution? *BBC Future*. https://www.bbc.com/future/article/20230525-how-more-us-female-electricians-helps-climate-change (https://www.bbc.com/future/article/20230525-how-more-us-female-electricians-helps-climate-change)

doc14 - Elster, N. (2023, April 19). How an objective measure of pain could counter bias in medicine. *BBC Future*. https://www.bbc.com/future/article/20230414-the-search-for-an-objective-measure-of-pain (https://www.bbc.com/future/article/20230414-the-search-for-an-objective-measure-of-pain)

doc15 - Kraus, N. (2023, May 26). How rhythm shapes our lives. *BBC Future*. https://www.bbc.com/future/article/20230526-how-rhythm-shapes-our-lives (https://www.bbc.com/future/article/20230526-how-rhythm-shapes-our-lives)

For the all documents i found on the web, I copy the text and paste into an empty text file .

So there are a total of 15 text files. Then I put all of the text files into a folder, named corpus.

Then for the 15 documents, i named them as doc1, doc2, doc3, doc4, doc5, doc6, doc7, doc8, doc9, doc10, doc11, doc12, doc13, doc14 and doc15.

## Tokenisation

Convert Hyphen to space

```
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "-")
```

Remove numbers

```
docs <- tm_map(docs,removeNumbers)
```

remove punctuation

```
docs <- tm_map(docs,removePunctuation)
```

change letter to lower case

```
docs <- tm_map(docs,content_transformer(tolower))
```

remove stopwords ( ex. a, is, the, …)

```
docs <- tm_map(docs,removeWords,stopwords("english"))
```

remove white space

```
docs <- tm_map(docs,stripWhitespace)
```

stemming

```
docs <- tm_map(docs,stemDocument,language = "english")
```

Create Document-Term Matrix ( DTM )

```
dtm <- DocumentTermMatrix(docs)
```

inspect DTM

we can see that the terms are very sparse because the sparsity is 69%

```
inspect(dtm[1:15,1:5])
```

```
## <<DocumentTermMatrix (documents: 15, terms: 5)>>
## Non-/sparse entries: 23/52
## Sparsity           : 69%
## Maximal term length: 6
## Weighting          : term frequency (tf)
## Sample             :
##            Terms
## Docs        access accid accord actual add
##   Doc1.txt       2     1      1      1   1
##   Doc10.txt      0     0      1      0   0
##   Doc11.txt      0     0      1      0   0
##   Doc12.txt      0     0      2      3   0
##   Doc13.txt      2     0      4      1   1
##   Doc14.txt      0     0      0      0   2
##   Doc15.txt      0     0      0      1   0
##   Doc4.txt       0     0      1      0   3
##   Doc7.txt       0     0      2      0   1
##   Doc8.txt       0     0      3      0   0
```

word frequencies

length of the frequency is 3103

```
freq = colSums(as.matrix(dtm))
length(freq)
```

```
## [1] 3103
```

frequency of the token:

head:

```
ord = order(freq)
freq[head(freq)]
```

```
##   actual   access    agenc    adequ advisori      add
##        7        4        3        3        1        9
```

tail:

```
freq[tail(freq)]
```

```
##    add access access access access accord
##      9      4      4      4      4     16
```

frequency of frequencies

Top 10 most frequent frequencies:

```
head(table(freq),10)
```

```
## freq
##    1    2    3    4    5    6    7    8    9   10
## 1475  507  288  168  137   89   76   61   40   40
```

Top 10 least frequent frequencies:

```
tail(table(freq),10)
```

```
## freq
## 62 63 65 74 75 76 79 93 96 99
##  1  2  1  2  1  1  1  1  1  1
```

size of original DTM

15 x 3103

```
dim(dtm)
```

```
## [1]   15 3103
```

remove non frequent / sparse terms

After remove those sparse terms, size of DTM is 15 x 21

```
dtms <- removeSparseTerms(dtm,0.26)
dtm.matrix <- as.matrix(dtms)
dim(dtm.matrix)
```

```
## [1] 15 21
```

21 tokens left

```
inspect(dtms)
```

```
## <<DocumentTermMatrix (documents: 15, terms: 21)>>
## Non-/sparse entries: 269/46
## Sparsity           : 15%
## Maximal term length: 5
## Weighting          : term frequency (tf)
## Sample             :
##             Terms
## Docs         also can like make one say time use will year
##   Doc1.txt     4   2    4    1   3   8    0   0    5    3
##   Doc10.txt    1   9    6    5  15   3    0   5    1    5
##   Doc11.txt    2   3    8    6   3   5    1   8    1    0
##   Doc12.txt    9  23    4    8   5  17    1   7    0    2
##   Doc13.txt    7   2   11    5   1  14    6   0    3    5
##   Doc14.txt    2   5    8    2  10  14    1   7    7    1
##   Doc15.txt    6  15    4    8  16   2   13  12    8    3
##   Doc3.txt     0   0    8    1   5   6    6   1   11   17
##   Doc4.txt     1   3    0    4   3   7    2   4    4    3
##   Doc5.txt     4   8    0    2   0  11   14   3    1    3
```

save the DTM matrix into a csv file

```
write.csv(dtm.matrix,"dtms.csv")
```

distance matrix

```
dismatrix = dist(scale(dtms))
```

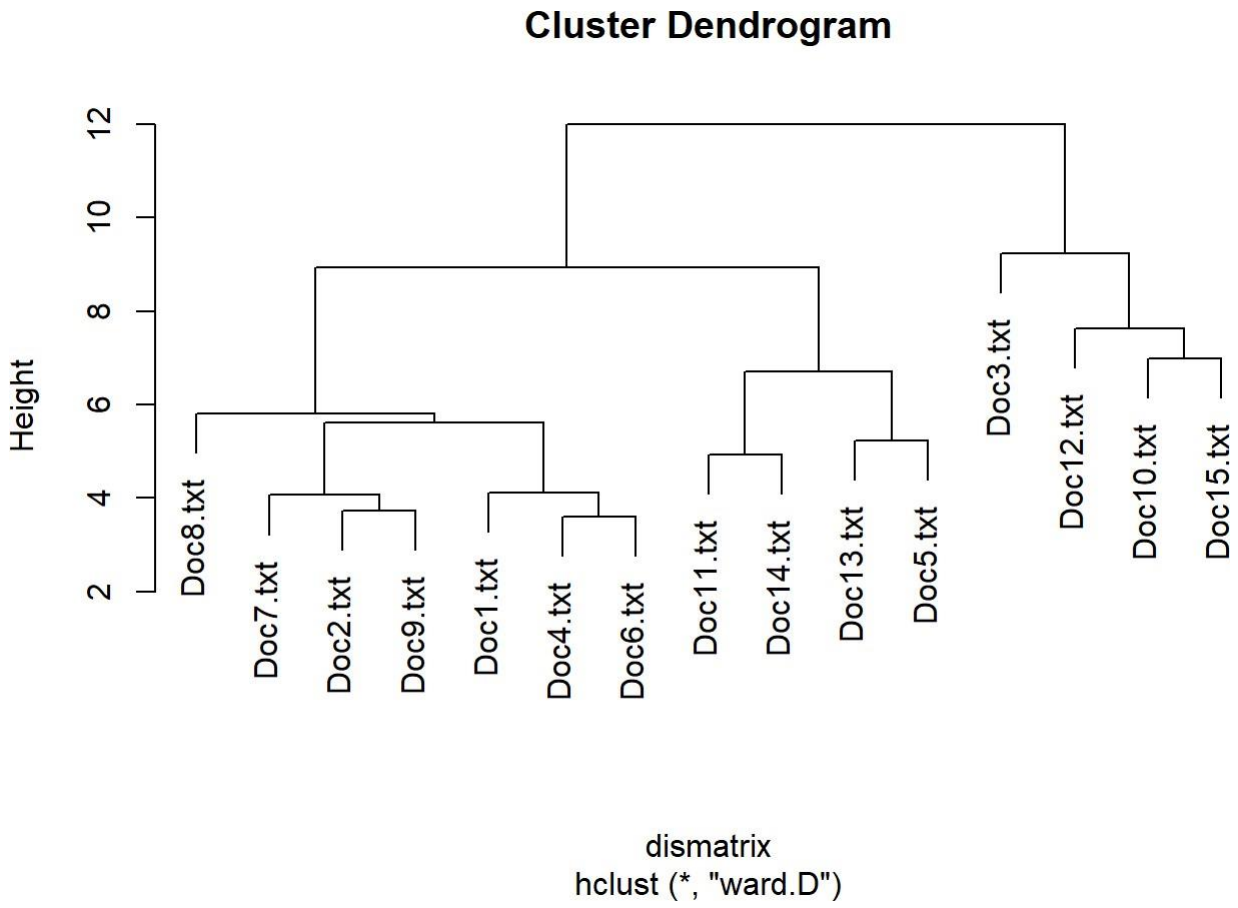
# clustering

use Euclidean distance

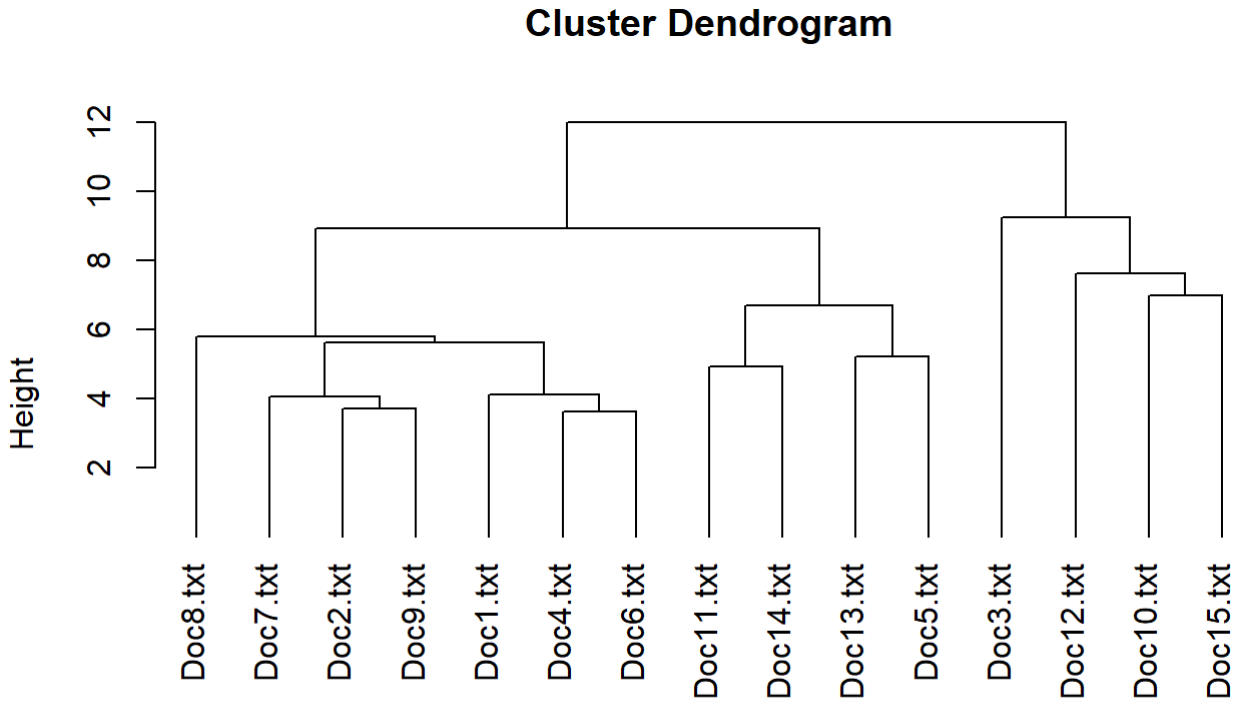
```
fit = hclust(dismatrix, method = "ward.D")
```

plot Dendogram

```
plot(fit)
```

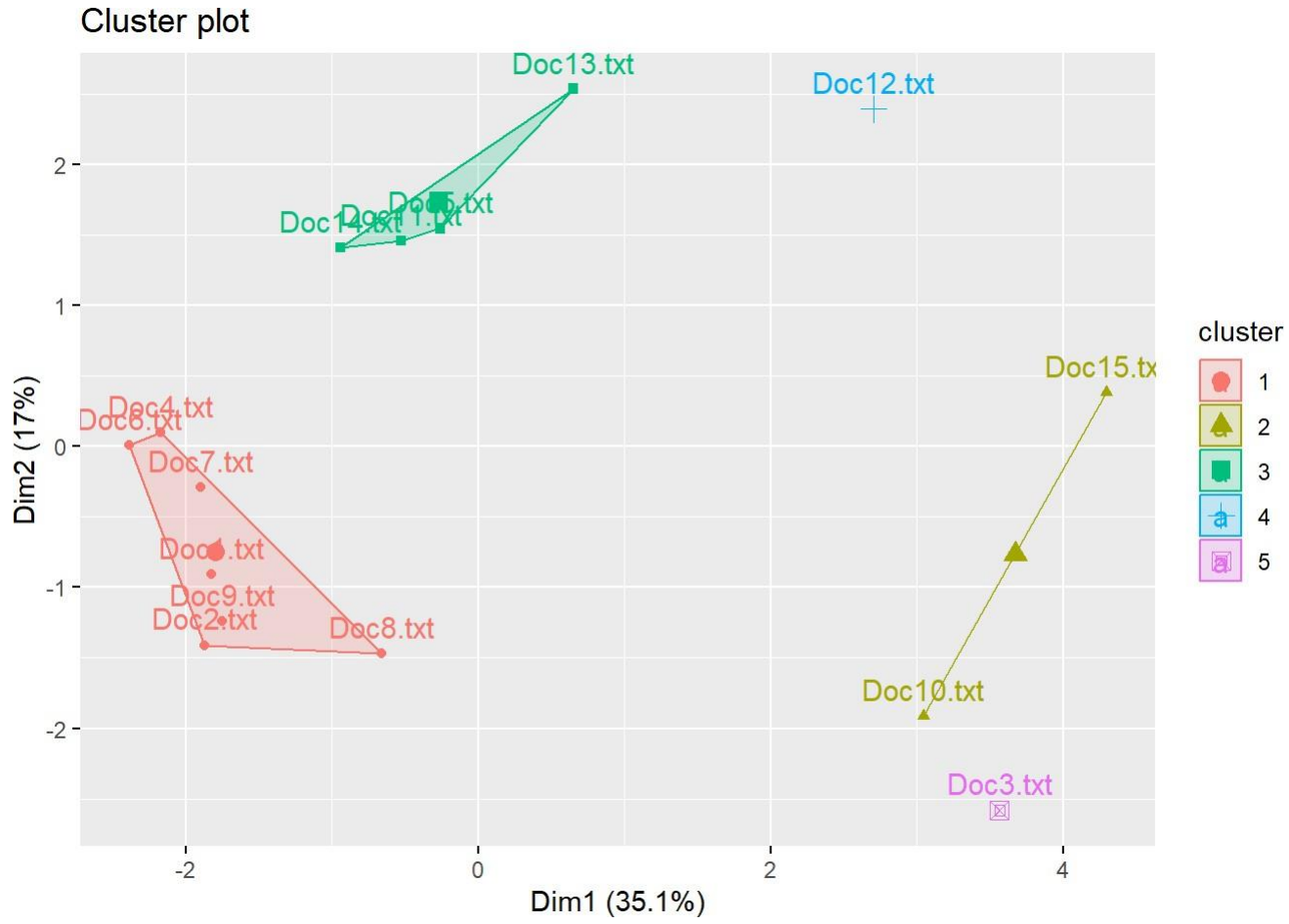



```
plot(fit,hang = -1)
```




dismatrix
hclust (*, "ward.D")

```
sub_grp <- cutree(fit, k = 5)
fviz_cluster(list(data = dismatrix, cluster = sub_grp))
```



convert to binary matrix

```
dtm.matrix1 <- as.matrix((dtm.matrix>0)+0)
```

multiply binary matrix by its transpose

```
doc.Matrix <- dtm.matrix1 %*% t(dtm.matrix1)
```

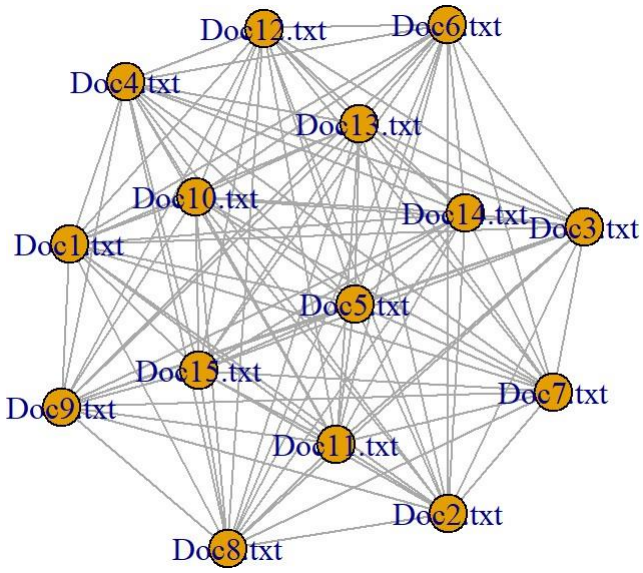
make leading diagonal 0

```
diag(doc.Matrix) = 0
```

create graph object

```
doc.Matrix.graph <- graph_from_adjacency_matrix(doc.Matrix,mode = "undirected",weighted = TRU
E)

plot(doc.Matrix.graph)
```




we can see that doc12 and doc15 are more closer to each other means that they are more related.

Whereas doc2 and doc5 are far from each other, so they are less related.

calculate stats from the graph object

```
format(closeness(doc.Matrix.graph),digits = 2)
```

```
##  Doc1.txt Doc10.txt Doc11.txt Doc12.txt Doc13.txt Doc14.txt Doc15.txt  Doc2.txt
##  "0.0044"  "0.0049"  "0.0045"  "0.0042"  "0.0044"  "0.0042"  "0.0042"  "0.0052"
##  Doc3.txt  Doc4.txt  Doc5.txt  Doc6.txt  Doc7.txt  Doc8.txt  Doc9.txt
##  "0.0049"  "0.0044"  "0.0047"  "0.0049"  "0.0053"  "0.0053"  "0.0052"
```

multiply transpose binary matrix by binary matrix

```
Token.Matrix <- t(dtm.matrix1) %*% dtm.matrix1
```

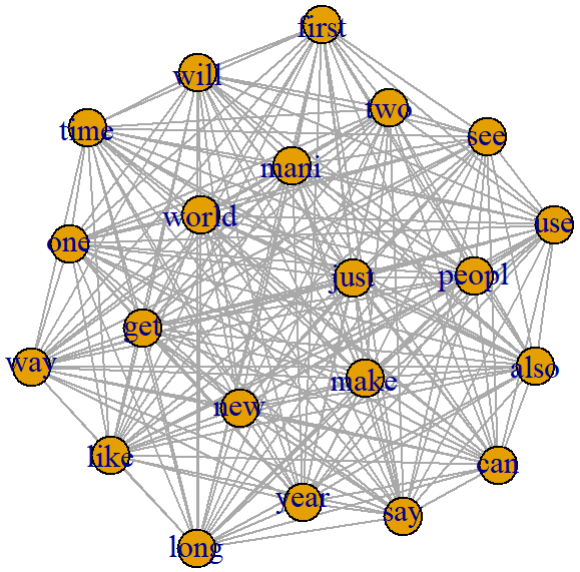
make leading diagonal 0

```
diag(Token.Matrix) = 0
```

create graph object

```
Token.Matrix.graph <- graph_from_adjacency_matrix(Token.Matrix,mode = "undirected",weighted = TRUE)

plot(Token.Matrix.graph)
```




For the tokens, we can see that token “use” and token “also” are very related, whereas token like “mani” and token “new” are far to each other, so they are less related.

calculate stats from the graph object

```
format(closeness(Token.Matrix.graph),digits = 2)
```

```
##     also      can    first      get     just     like     long     make
## "0.0045" "0.0048" "0.0049" "0.0048" "0.0045" "0.0049" "0.0049" "0.0039"
##     mani      new      one    peopl      say      see      two      way
## "0.0050" "0.0045" "0.0042" "0.0049" "0.0045" "0.0045" "0.0048" "0.0048"
##     will    world     year      use     time
## "0.0045" "0.0039" "0.0042" "0.0046" "0.0049"
```

```
dtmsa <- as.data.frame(dtm.matrix)

dtmsa$ABS = rownames(dtmsa)

dtmsb = data.frame()

for(i in 1:nrow(dtmsa)){

  for(j in 1:(ncol(dtmsa)-1)){

    touse = cbind(dtmsa[i,j],dtmsa[i,ncol(dtmsa)],colnames(dtmsa[j]))

    dtmsb = rbind(dtmsb,touse)}}

colnames(dtmsb) = c("weight","abs","token")
```

delete 0 weights

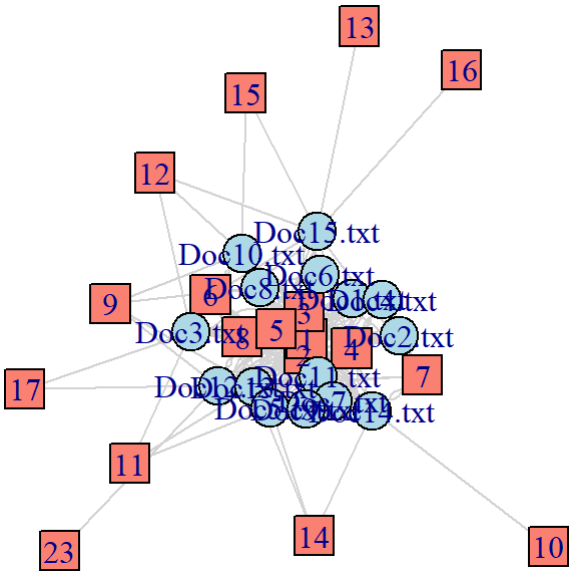
```
dtmsc = dtmsb[dtmsb$weight != 0,]
```

create graph object and declare bipartite

```
g <- graph.data.frame(dtmsc,directed = FALSE)
bipartite.mapping(g)
```

```
## $res
## [1] TRUE
##
## $type
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
V(g)$type <- bipartite_mapping(g)$type
V(g)$color <- ifelse(V(g)$type,"lightblue","salmon")
V(g)$shape <- ifelse(V(g)$type,"circle","square")
E(g)$color <- "lightgray"
plot(g)
```




token 1,2,3,4,5,6,8 are more related to all the documents

token 7 is more related to doc 2, doc4, doc14

other token are far away from the documents, hence they are less related to the documents.