P01

Manish Shah

2023-04-21

```
library(pdftools)
library(tm)
library(wordcloud)
library(graph)
library(Rgraphviz)
library(igraph)
library(topicmodels)
```

Setting working directory to MDS503P01

```
knitr::opts_knit$set(root.dir = "D:/R programming runs/Assisgnments/Project 1/MDS503P01")
```

Getting all the pdf files using regex from working directory

```
pdf_files <- list.files(pattern = "pdf$")
```

Creating corpus with help of tm package

```
pdf_corp <- Corpus(URISource(pdf_files), readerControl=list(reader = readPDF()))
```

Removing punctuation, stopwords, & converting words to lowercase and then creating *term document matrix*

```
pdf_tdm <- TermDocumentMatrix(pdf_corp, control = list(removePunctuation=T,
                                                       stopwords=T,
                                                       tolower=T,
                                                       removeNumbers=T))
inspect(pdf_tdm)
```

```
## <<TermDocumentMatrix (terms: 3542, documents: 5)>>
## Non-/sparse entries: 4884/12826
## Sparsity           : 72%
## Maximal term length: 78
```

```
## Weighting          : term frequency (tf)
## Sample             :
##             Docs
## Terms         advantages-and-disadvantages-for-nurses-of-using-social-media.pdf
##   can                                                                        27
##   facebook                                                                    4
##   information                                                                23
##   marketing                                                                   0
##   media                                                                      63
##   network                                                                     0
##   online                                                                     10
##   sites                                                                       3
##   social                                                                     66
##   use                                                                        11
##             Docs
## Terms         Social-Media-Spread-During-Covid-19-The-Pros-and-Cons-of-Likes-and-Shares.pdf
##   can                                                                                    8
##   facebook                                                                               2
##   information                                                                           13
##   marketing                                                                              0
##   media                                                                                 23
##   network                                                                                0
##   online                                                                                 0
##   sites                                                                                  1
##   social                                                                                21
##   use                                                                                    2
##             Docs
## Terms         Social Network Sites and Well-Being The Role of Social Connection.pdf
##   can                                                                            11
##   facebook                                                                       43
##   information                                                                     3
##   marketing                                                                       0
##   media                                                                           5
##   network                                                                        57
##   online                                                                         10
##   sites                                                                          62
##   social                                                                        136
##   use                                                                            57
##             Docs
## Terms         Social_Media_Marketing_SOCIAL_MEDIA_MARK.pdf
##   can                                                 36
##   facebook                                            17
##   information                                         25
##   marketing                                           94
##   media                                              137
##   network                                             10
##   online                                              42
##   sites                                               13
##   social                                             156
##   use                                                 16
##             Docs
## Terms         The Impact of Social Media on Children, Adolescents, and Families.pdf
##   can                                                                            20
```

```
##   facebook                                                                11
##   information                                                              9
##   marketing                                                                0
##   media                                                                   42
##   network                                                                  2
##   online                                                                  38
##   sites                                                                   33
##   social                                                                  44
##   use                                                                     14
```

Getting most frequent used words where the words will be repeated atleast 25 times

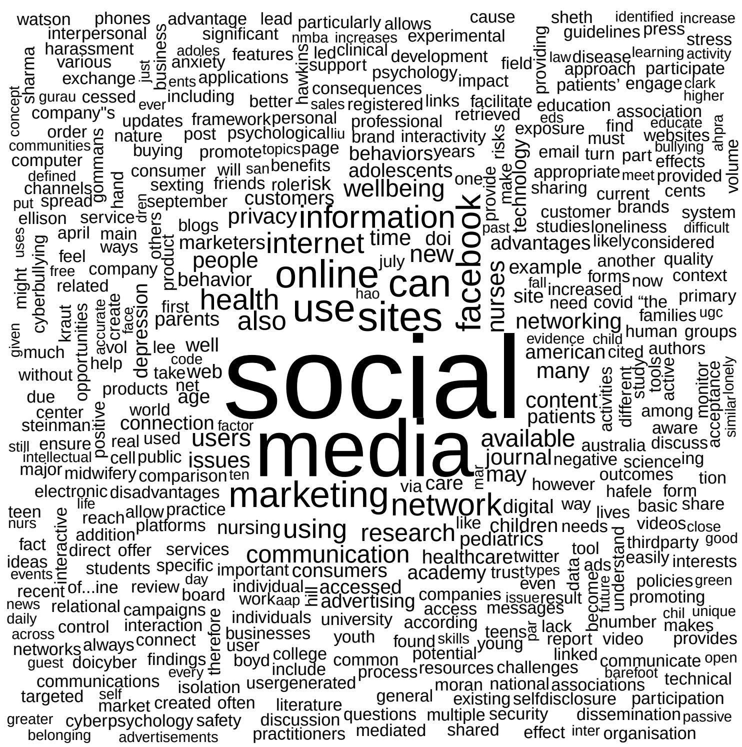
```
pdf.freq <- findFreqTerms(pdf_tdm, lowfreq=25, highfreq=Inf)
pdf.freq
```

```
##  [1] "also"          "available"     "can"           "communication"
##  [5] "content"       "customers"     "facebook"      "health"
##  [9] "information"   "internet"      "issues"        "journal"
## [13] "many"          "marketing"     "may"           "media"
## [17] "network"       "networking"    "new"           "nurses"
## [21] "online"        "people"        "privacy"       "research"
## [25] "sites"         "social"        "time"          "use"
## [29] "users"         "using"         "wellbeing"
```

Plot word cloud without any colors

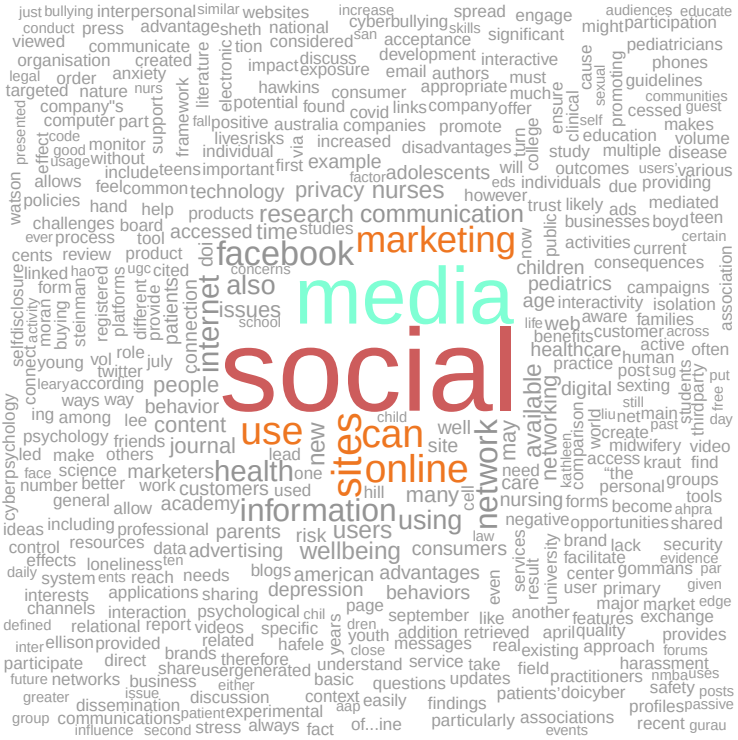
We convert term-document matrix(sparse matrix) to regular matrix and then with help of ***word-cloud*** package, we create word cloud

```
m <- as.matrix(pdf_tdm)
freq <- sort(rowSums(m), decreasing=T)
wordcloud(words=names(freq), freq=freq, min.freq=4, random.order=F)
```



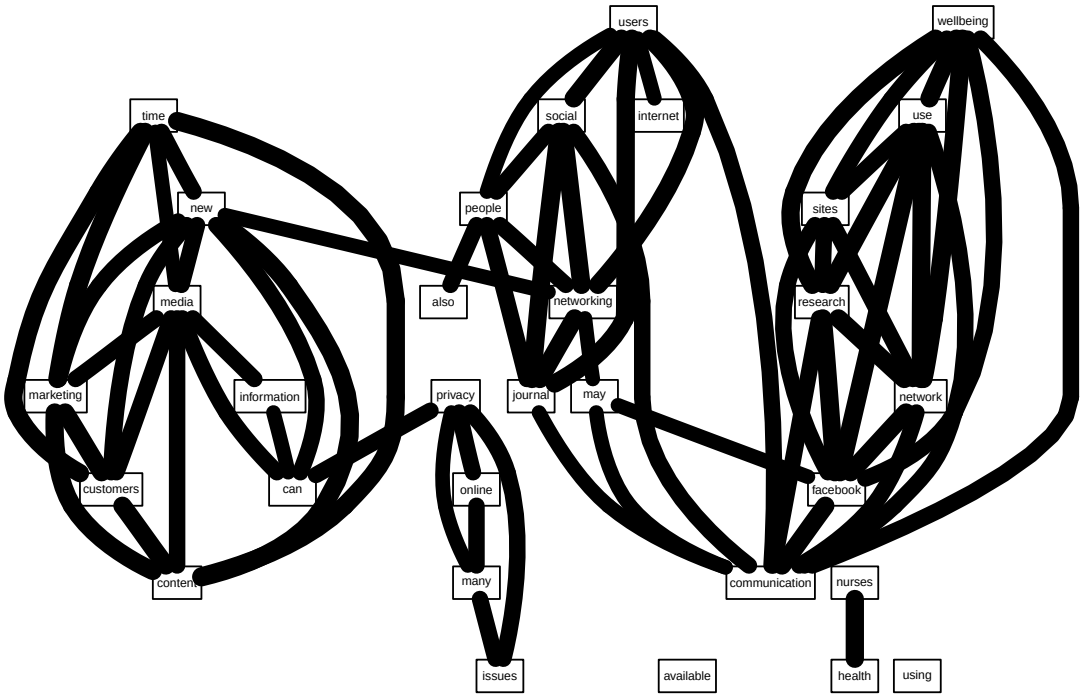
Plot word cloud with color

```
my_colors <- sample(colors(), 10)
wordcloud(words=names(freq), freq=freq, min.freq=4, random.order=F, colors = my_colors)
```



Plot network graph from term-document matrix

```
plot(pdf_tdm, terms=pdf.freq, corThreshold=0.8, weighting=T)
```



Generating topic modeling using topicmodel package

```
set.seed(16)
myLda <- LDA(as.DocumentTermMatrix(pdf_tdm), k=5)
terms(myLda, 10)
```

```
##       Topic 1       Topic 2       Topic 3       Topic 4         Topic 5
##  [1,] "social"      "media"       "social"      "social"        "social"
##  [2,] "media"       "online"      "sites"       "sites"         "media"
##  [3,] "health"      "parents"     "available"   "network"       "marketing"
##  [4,] "nurses"      "pediatrics"  "facebook"    "use"           "online"
##  [5,] "information" "accessed"    "using"       "facebook"      "can"
##  [6,] "can"         "children"    "american"    "wellbeing"     "information"
##  [7,] "also"        "can"         "use"         "research"      "customers"
##  [8,] "healthcare"  "social"      "adolescents" "internet"      "content"
##  [9,] "nursing"     "july"        "sexting"     "communication" "consumers"
## [10,] "care"        "adolescents" "online"      "users"         "new"
```

Here, k=5 denotes 5 topic & 10 in terms() denotes 10 terms in each of the 5 topics
We can observe that the 'social', 'media', 'marketing', 'online', 'network', etc. are most used terms in all of the topics.