Project 2

Suman Paudel

2024-03-25

```
# Load the packages
library(pdftools)
library(tm)
library(magrittr)
library(wordcloud)
library(Rgraphviz)
library(graph)
library(foreign)
library(gt)
library(tidyverse)
library(jsonlite)
library(httr)
```

Load all of the necessary packages need for this project.

Task 1 Part 1

```
# load the data using Base R read.csv
data <- read.csv("covnep_252days.csv")

summary(data$totalCases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       2     963   13376   19341   77816
```

```
# minimum value is 0 but we need 1 instead
# this can be achieved using multiple ways like ifelse or pmax or subsetting

# using ifelse
totalCases_ifelse <- ifelse(data$totalCases < 1, 1, data$totalCases)
summary(totalCases_ifelse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2     963   13377   19341   77816
```

```r
# using pmax
totalCases_pmax <- pmax(data$totalCases, 1)
summary(totalCases_pmax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2     963   13377   19341   77816
```

```r
# subsetting
totalCases_subsetting <- data$totalCases
totalCases_subsetting[totalCases_subsetting < 1] <- 1
summary(totalCases_subsetting)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2     963   13377   19341   77816
```

Part 2

```r
saq_data <- read.spss("SAQ8.sav",to.data.frame=TRUE)
```

For q01

```r
library(foreign) # for .SAV file we can use foreign package for that as well
library(gt, warn.conflicts = FALSE) #genotype table
library(magrittr) # for using pipes
library(tibble)
library(dplyr)

# read the .sav file using read_sav function from haven
saq_data <- read.spss("SAQ8.sav",to.data.frame=TRUE)


# for q1

q01 <- saq_data$q01

datalevels_q01 <- levels(q01)
freq_q01 <- as.numeric(table(q01))
percent_q01 <- as.numeric(round(prop.table(freq_q01) * 100, 1))
valid_percent_q01 <- as.numeric(round(prop.table(freq_q01) * 100, 1))
cum_percent <- cumsum(percent_q01)

# Create data frame
data <- data.frame(
  Levels = datalevels_q01,
  Freq = freq_q01,
  Percent = percent_q01,
  Val_Percent = valid_percent_q01,
  Cum_Percent = cum_percent
)

# final version of calculated table
```

```r
data <- data %>% add_row(Levels = "Total", Freq = sum(data$Freq),
                 Percent = sum(data$Percent),
                 Val_Percent = sum(data$Val_Percent),
                 Cum_Percent = NULL)

# aethetics table using gt
data %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("Statistics makes me cry")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

Statistics makes me cry

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	270	10.5	10.5	10.5
Agree	1338	52.0	52.0	62.5
Neither	735	28.6	28.6	91.1
Disagree	187	7.3	7.3	98.4
Strongly disagree	41	1.6	1.6	100.0
Total	2571	100.0	100.0	

```r
# for q03

q03 <- saq_data$q03
datalevels_q03 <- levels(q03)
freq_q03 <- as.numeric(table(q03))
percent_q03 <- as.numeric(round(prop.table(freq_q03) * 100, 1))
valid_percent_q03 <- as.numeric(round(prop.table(freq_q03) * 100, 1))
cum_percent_q03 <- cumsum(percent_q03)

data_q03 <- data.frame(
  Levels = datalevels_q03,
  Freq = freq_q03,
  Percent = percent_q03,
  Val_Percent = valid_percent_q03,
  Cum_Percent = cum_percent_q03
)

data_q03 <- data_q03 %>% add_row(Levels = "Total", Freq = sum(data_q03$Freq),
                         Percent = sum(data_q03$Percent),
                         Val_Percent = sum(data_q03$Val_Percent),
                         Cum_Percent = NULL)

# final version of calculated table
data_q03 %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("Statistic makes me cry")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
```

```
            Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

Statistic makes me cry

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	497	19.3	19.3	19.3
Agree	672	26.1	26.1	45.4
Neither	878	34.2	34.2	79.6
Disagree	448	17.4	17.4	97.0
Strongly disagree	76	3.0	3.0	100.0
Total	2571	100.0	100.0	

```
q06 <- saq_data$q06

datalevels_q06 <- levels(q06)
freq_q06 <- as.numeric(table(q06))
percent_q06 <- as.numeric(round(prop.table(freq_q06) * 100, 1))
valid_percent_q06 <- as.numeric(round(prop.table(freq_q06) * 100, 1))
cum_percent_q06 <- cumsum(percent_q06)

data_q06 <- data.frame(
  Levels = datalevels_q06,
  Freq = freq_q06,
  Percent = percent_q06,
  Val_Percent = valid_percent_q06,
  Cum_Percent = cum_percent_q06
)

data_q06 <- data_q06 %>% add_row(Levels = "Total", Freq = sum(data_q06$Freq),
                         Percent = sum(data_q06$Percent),
                         Val_Percent = sum(data_q06$Val_Percent),
                         Cum_Percent = NULL)

# final version of calculated table
data_q06 %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("I have little experience of computer")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

I have little experience of computer

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	702	27.3	27.3	27.3
Agree	1127	43.8	43.8	71.1
Neither	344	13.4	13.4	84.5
Disagree	252	9.8	9.8	94.3

Strongly disagree	146	5.7	5.7	100.0
Total	2571	100.0	100.0	

```
q08 <- saq_data$q08

datalevels_q08 <- levels(q08)
freq_q08 <- as.numeric(table(q08))
percent_q08 <- as.numeric(round(prop.table(freq_q08) * 100, 2))
valid_percent_q08 <- as.numeric(round(prop.table(freq_q08) * 100, 2))
cum_percent_q08 <- cumsum(percent_q08)


data_q08 <- data.frame(
  Levels = datalevels_q08,
  Freq = freq_q08,
  Percent = round(valid_percent_q08,1),
  Val_Percent = round(valid_percent_q08,1),
  Cum_Percent = round(cum_percent_q08,1)
)

data_q08 <- data_q08 %>% add_row(Levels = "Total", Freq = sum(data_q06$Freq),
                         Percent = sum(data_q08$Percent),
                         Val_Percent = sum(data_q08$Val_Percent),
                         Cum_Percent = NULL)


# final version of calculated table
data_q08 %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("Statistics makes me cry")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

Statistics makes me cry

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	383	14.9	14.9	14.9
Agree	1487	57.8	57.8	72.7
Neither	482	18.8	18.8	91.5
Disagree	147	5.7	5.7	97.2
Strongly disagree	72	2.8	2.8	100.0
Total	5142	100.0	100.0	

```
data_1 = 'https://data.covid19india.org/v4/min/timeseries.min.json'
data_2 = 'https://data.covid19india.org/v4/min/data.min.json'
covid_data_1 <- jsonlite::fromJSON(data_1)
covid_data_2 <- jsonlite::fromJSON(data_2)
```

```
covid_1_parsed <-
  covid_data_1 %>% enframe() %>% unnest_wider(value) %>% unnest_wider(dates) %>%
  pivot_longer(cols = !name,
               names_to = 'date',
               values_to = "value") %>% unnest_wider(value) %>%
  mutate(across(c(delta, delta7, total), ~ map(., ~ set_names(
    as_tibble(.x), paste0(cur_column(), "_", names(.))
  )))) %>%
  unnest_wider(c(delta, delta7, total))

covid_1_parsed[150:300, c("delta_confirmed","delta_recovered","delta_tested","delta7_confirmed","delta7_
```

Task 2 Web Scraping

```
## # A tibble: 151 x 6
##    delta_confirmed delta_recovered delta_tested delta7_confirmed
##              <int>           <int>        <int>            <int>
##  1              61             109          315              502
##  2              52             110          390              461
##  3              44             129          235              459
##  4              41             139          569              416
##  5              40              78          564              381
##  6              33              65          184              338
##  7              32              70          270              303
##  8              31              75          307              273
##  9              23              67          267              244
## 10              28              61          811              228
## # i 141 more rows
## # i 2 more variables: delta7_recovered <int>, delta_tested <int>
```

```
covid_2_parsed <- covid_data_2 %>% enframe() %>% unnest_wider(value) %>%
  unnest_wider(c(delta,delta21_14, delta7, total),names_sep = "_") %>% select(-c(districts,meta))
covid_2_parsed[, c("delta_confirmed","delta_recovered","delta_tested","delta7_confirmed","delta7_recove
```

```
## # A tibble: 37 x 6
##    delta_confirmed delta_recovered delta_tested delta7_confirmed
##              <int>           <int>        <int>            <int>
##  1              NA              NA         1376                3
##  2             385             675        39848             2873
##  3               1               9          334               66
##  4             212             236        15060             2056
##  5               8               9       226443               40
##  6               5               3         1403               28
##  7              32              32        11869              205
##  8              45              46        56751              267
##  9              NA               1           NA               NA
## 10              23              53         2361              222
## # i 27 more rows
## # i 2 more variables: delta7_recovered <int>, delta_tested <int>
```

```r
merged_df <- merge(covid_1_parsed, covid_2_parsed, by = "name", all = FALSE)
head(merged_df[7250:8000,])
```

```
##      name       date delta_confirmed.x delta_recovered.x delta_tested.x
## 7250   HP 2021-04-20              1340              1078           9744
## 7251   HP 2021-04-21              1692               908           9291
## 7252   HP 2021-04-22              1774               689           8037
## 7253   HP 2021-04-23              1189               772          10385
## 7254   HP 2021-04-24              2073               877          10534
## 7255   HP 2021-04-25              1363              1161           7164
##      delta_other.x delta_deceased.x delta_vaccinated1.x delta_vaccinated2.x
## 7250            NA               16               40934                9089
## 7251            NA               17               18780                3381
## 7252             4               18               41362                9161
## 7253             4               26               36710                9874
## 7254             7               24               32168                7402
## 7255             4               32                8353                2612
##      delta7_confirmed.x delta7_recovered.x delta7_tested.x delta7_other.x
## 7250               8016               4184           54645            -12
## 7251               8783               4840           56298            -12
## 7252               9523               4940           58520             -9
## 7253               9870               5228           62516             -5
## 7254              10551               5449           62143             11
## 7255              11126               6078           60798             21
##      delta7_deceased.x delta7_vaccinated1.x delta7_vaccinated2.x
## 7250                84               228964                37855
## 7251                88               204933                37985
## 7252                95               215432                43395
## 7253               112               209999                46366
## 7254               124               211167                48268
## 7255               146               210197                48792
##      total_confirmed.x total_recovered.x total_tested.x total_other.x
## 7250             79410             68150        1401986            25
## 7251             81102             69058        1411277            25
## 7252             82876             69747        1419314            29
## 7253             84065             70519        1429699            33
## 7254             86138             71396        1440233            40
## 7255             87501             72557        1447397            44
##      total_deceased.x total_vaccinated1.x total_vaccinated2.x delta_tested.y
## 7250             1206              1225881              152805           3613
## 7251             1223              1244661              156186           3613
## 7252             1241              1286023              165347           3613
## 7253             1267              1322733              175221           3613
## 7254             1291              1354901              182623           3613
## 7255             1323              1363254              185235           3613
##      delta_vaccinated1.y delta_vaccinated2.y delta_confirmed.y delta_deceased.y
## 7250                 371                8192                85                1
## 7251                 371                8192                85                1
## 7252                 371                8192                85                1
## 7253                 371                8192                85                1
## 7254                 371                8192                85                1
## 7255                 371                8192                85                1
##      delta_recovered.y delta_other.y delta21_14_confirmed delta7_confirmed.y
```

```
## 7250               198            NA                  958               1537
## 7251               198            NA                  958               1537
## 7252               198            NA                  958               1537
## 7253               198            NA                  958               1537
## 7254               198            NA                  958               1537
## 7255               198            NA                  958               1537
##      delta7_recovered.y delta7_tested.y delta7_vaccinated1.y
## 7250               1154          64352                13244
## 7251               1154          64352                13244
## 7252               1154          64352                13244
## 7253               1154          64352                13244
## 7254               1154          64352                13244
## 7255               1154          64352                13244
##      delta7_vaccinated2.y delta7_deceased.y delta7_other.y total_confirmed.y
## 7250               234011                20             -1            224106
## 7251               234011                20             -1            224106
## 7252               234011                20             -1            224106
## 7253               234011                20             -1            224106
## 7254               234011                20             -1            224106
## 7255               234011                20             -1            224106
##      total_deceased.y total_recovered.y total_tested.y total_vaccinated1.y
## 7250             3738            218410        3685011             5713695
## 7251             3738            218410        3685011             5713695
## 7252             3738            218410        3685011             5713695
## 7253             3738            218410        3685011             5713695
## 7254             3738            218410        3685011             5713695
## 7255             3738            218410        3685011             5713695
##      total_vaccinated2.y total_other.y
## 7250             3443823            16
## 7251             3443823            16
## 7252             3443823            16
## 7253             3443823            16
## 7254             3443823            16
## 7255             3443823            16
```

```
library(RSelenium)
library(rvest)
library(netstat)
rD <- rsDriver(browser="firefox",verbose = F, port = 14421L)
remDr <- rD[["client"]]
remDr$navigate("https://aqicn.org/forecast/kathmandu/")
aqi_html  <- read_html(remDr$getPageSource() %>% unlist())
aqi_html %>% html_element(".forecast-body-table") %>%  html_nodes("table") %>% html_table() -> forecast_
```

```
aqi_table <- forecast_table %>% .[[1]]
```

```
aqi_table <- aqi_table %>% select(-c('X10','X11','X20','X21','X30','X31','X40','X41','X50','X51','X60',
```

```
aqi_table <- aqi_table %>% filter(X1 != 'UVI')
aqi_table <- aqi_table %>% filter(X1 != 'humidity')
aqi_table <- aqi_table %>% mutate(X1 = replace(X1, 9, "humidity"))
```

```r
aqi_table <- aqi_table %>% mutate(X1 = replace(X1, 1, "Index"))
aqi_table <- aqi_table %>% filter(X1 != '')

headers <- aqi_table[1,]

colnames(aqi_table) <- headers

aqi_table <- aqi_table[-1,]

aqi_table <- aqi_table %>% column_to_rownames(var = 'Index')

library(stringr)

aqi_table[2,] <- floor(as.integer(str_extract(as.character(aqi_table[2,]), "\\d+")) / 1000)
aqi_table[3,] <- floor(as.integer(str_extract(as.character(aqi_table[3,]), "\\d+")) / 100)

lengths <- as.numeric(nchar(aqi_table[4,]))
aqi_table[4,] <- ifelse(lengths == 2, substr(aqi_table[4,], 1, 1), ifelse(lengths %in% 3:4, substr(aqi_

aqi_table
```

```
##                      Monday 25    Monday 25    Monday 25    Monday 25
## hour                         0            3            6            9
## PM2.5                      151          151          148          138
## PM10                        51           51           51           51
## O3                           9            8            9           24
## Wind Speed (m/s)             1            1            1            2
## Temp.                      13°          12°          16°          21°
## humidity         6:03 ~ 18:18 6:03 ~ 18:18 6:03 ~ 18:18 6:03 ~ 18:18
##                      Monday 25    Monday 25    Monday 25    Monday 25
## hour                        12           15           18           21
## PM2.5                      138          138          138          138
## PM10                        50           46           46           51
## O3                          22           20           15            5
## Wind Speed (m/s)             2            1            3            4
## Temp.                      21°          17°          12°          12°
## humidity         6:03 ~ 18:18 6:03 ~ 18:18 6:03 ~ 18:18 6:03 ~ 18:18
##                     Tuesday 26   Tuesday 26   Tuesday 26   Tuesday 26
## hour                         0            3            6            9
## PM2.5                      138          138          137          137
## PM10                        51           51           51           50
## O3                           4            4           10           32
## Wind Speed (m/s)             3            3            1            2
## Temp.                      12°          12°          16°          22°
## humidity         6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19
##                     Tuesday 26   Tuesday 26   Tuesday 26   Tuesday 26
## hour                        12           15           18           21
## PM2.5                      138          138          138          151
## PM10                        46           46           51           51
## O3                          29           22           16            7
## Wind Speed (m/s)             3            2            2            1
## Temp.                      24°          23°          17°          16°
## humidity         6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19
```

```
##                  Wednesday 27 Wednesday 27 Wednesday 27 Wednesday 27
## hour                        0            3            6            9
## PM2.5                     151          148          137          138
## PM10                       51           50           46           46
## O3                          4            5            9           24
## Wind Speed (m/s)            1            3            1            3
## Temp.                     15°          14°          18°          23°
## humidity          6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19
##                  Wednesday 27 Wednesday 27 Wednesday 27 Wednesday 27
## hour                       12           15           18           21
## PM2.5                     138          138          138          138
## PM10                       46           46           46           46
## O3                         23           20           15            6
## Wind Speed (m/s)            3            2            2            2
## Temp.                     24°          21°          17°          17°
## humidity          6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19
##                   Thursday 28  Thursday 28  Thursday 28  Thursday 28
## hour                        0            3            6            9
## PM2.5                     138          138          137          138
## PM10                       46           46           46           46
## O3                          4            4            8           29
## Wind Speed (m/s)            2            1            1            2
## Temp.                     15°          14°          19°          23°
## humidity          5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20
##                   Thursday 28  Thursday 28  Thursday 28  Thursday 28
## hour                       12           15           18           21
## PM2.5                     138          138          138          138
## PM10                       46           46           46           46
## O3                         28           21           16            8
## Wind Speed (m/s)            2            2            1            2
## Temp.                     23°          21°          17°          17°
## humidity          5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20
##                      Friday 29    Friday 29    Friday 29    Friday 29
## hour                        0            3            6            9
## PM2.5                     138          137          137          137
## PM10                       46           46           46           46
## O3                          4            3            8           24
## Wind Speed (m/s)            2            2            1            1
## Temp.                     17°          16°          19°          19°
## humidity          5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20
##                      Friday 29    Friday 29    Friday 29    Friday 29
## hour                       12           15           18           21
## PM2.5                     137          138          138          138
## PM10                       46           46           46           46
## O3                         23           21           15            5
## Wind Speed (m/s)            1            1            2            1
## Temp.                     23°          22°          19°          17°
## humidity          5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20
##                    Saturday 30  Saturday 30  Saturday 30  Saturday 30
## hour                        0            3            6            9
## PM2.5                     138          137          137          137
## PM10                       46           51           51           50
## O3                          6            7
## Wind Speed (m/s)            2            1            1            3
```

```
## Temp.                     16°          16°          21°          26°
## humidity         5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21
##                   Saturday 30  Saturday 30  Saturday 30  Saturday 30
## hour                       12           15           18           21
## PM2.5                     137          138          138          138
## PM10                       46           46           46           46
## O3
## Wind Speed (m/s)            4            3            4            4
## Temp.                     28°          15°          14°          15°
## humidity         5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21
##                      Sunday 31    Sunday 31    Sunday 31    Sunday 31
## hour                        0            3            9           12
## PM2.5                     137          137          137          137
## PM10                       46           46           46           46
## O3
## Wind Speed (m/s)            2            1            5            4
## Temp.                     16°          16°          28°          28°
## humidity         5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21
##                      Sunday 31    Sunday 31
## hour                       15           18
## PM2.5                     138          138
## PM10                       46           46
## O3
## Wind Speed (m/s)            2            3
## Temp.                     24°          18°
## humidity         5:56 ~ 18:21 5:56 ~ 18:21
```

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

```
##  [1] "1.pdf"  "10.pdf" "2.pdf"  "3.pdf"  "4.pdf"  "5.pdf"  "6.pdf"  "7.pdf"
##  [9] "8.pdf"  "9.pdf"
```

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

Task 3

```
##  [1] "1.pdf"  "10.pdf" "2.pdf"  "3.pdf"  "4.pdf"  "5.pdf"  "6.pdf"  "7.pdf"
##  [9] "8.pdf"  "9.pdf"
```

```
# load the pdf files into list
pdf_files <- lapply(files, pdf_text)
```

```
# create a corpus from vector source i.e from list pdf_files
corpus <- Corpus(VectorSource(unlist(pdf_files)))

# make a duplicate of the loaded corpus for future use
corpus_copy <- corpus
```

```r
# convert the all texts in lower
corpus <- tm_map(corpus, tolower)

# remove punctuations
corpus <- tm_map(corpus, removePunctuation)

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

my_stopwords <- c("can","may","used")
# remove stopwords from the corpus
corpus <- tm_map(corpus, removeWords, my_stopwords)

# stem the corpus
corpus <- tm_map(corpus, stemDocument)

# since values and value are same so replaced values and value
remove <- function(x) gsub("values","value",x)
corpus <-  tm_map(corpus, remove)

head(corpus)
```

Preprocessing Corpus

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 6
```

```r
# create Term Document Matrix with word lenght 1 or many
tdm <- TermDocumentMatrix(corpus, control = list((wordLenghts=c(1,Inf))))
head(tdm)
```

Term Document Matrix

```
## <<TermDocumentMatrix (terms: 6, documents: 948)>>
## Non-/sparse entries: 1485/4203
## Sparsity           : 74%
## Maximal term length: 7
## Weighting          : term frequency (tf)
```

```r
remove <- function(x) gsub("values","value",x)
corpus_copy <-  tm_map(corpus_copy, remove)
my_tdm <- TermDocumentMatrix(
  corpus_copy,
  control =
```

```
    list(
      removePunctuation = TRUE,
      stopwords = TRUE,
      tolower = TRUE,
      stemming = FALSE,
      removeNumbers = TRUE,
      bounds = list(global = c(3, Inf)),
      wordLenghts = c(1,Inf),
      removeWords = (c("can","may","used")))
)
```

Best way to create TDM with less code

```
# finding frequency of words which is at least present 10 times
low_frequent_terms <- findFreqTerms(my_tdm, lowfreq = 10)
head(low_frequent_terms)
```

Frequency Terms

```
## [1] "article"   "author"    "authors"   "content"   "data"      "discovery"
```

```
# finding frequency of words which is at max present 10 times
high_frequent_terms <- findFreqTerms(my_tdm, highfreq = 10)
head(high_frequent_terms)
```

```
## [1] "cordoba"           "downloaded"        "interdisciplinary"
## [4] "profile"           "profiles"          "publication"
```

```
findAssocs(my_tdm, "mining", 0.3)
```

Word Association

```
## $mining
##            data       knowledge       databases       discovery         systems
##            0.55            0.54            0.49            0.45            0.45
##        database           kinds        patterns            user           mined
##            0.44            0.42            0.40            0.39            0.39
##     interactive           users        research        analysis     association
##            0.37            0.37            0.36            0.35            0.34
## interestingness          erent       retrieval           rules      multimedia
##            0.33            0.32            0.31            0.31            0.31
##      challenges      techniques
##            0.30            0.30
```

```
findAssocs(my_tdm, "learning", 0.35)
```

```
## $learning
##       machine  intelligence          arti          cial           vol
##          0.74          0.56          0.52          0.50          0.43
##       shavlik        morgan      kaufmann     michalski    statistics
##          0.43          0.42          0.41          0.41          0.40
##        expert      mitchell         ijcai international      learners
##          0.40          0.40          0.39          0.38          0.38
##       quinlan  decisiontree  bibliography     carbonell        kluwer
##          0.38          0.37          0.36          0.36          0.36
##         neter         mateo
##          0.35          0.35
```

```
findAssocs(my_tdm, "classification", 0.4)
```

```
## $classification
## unlabeled
##      0.46
```

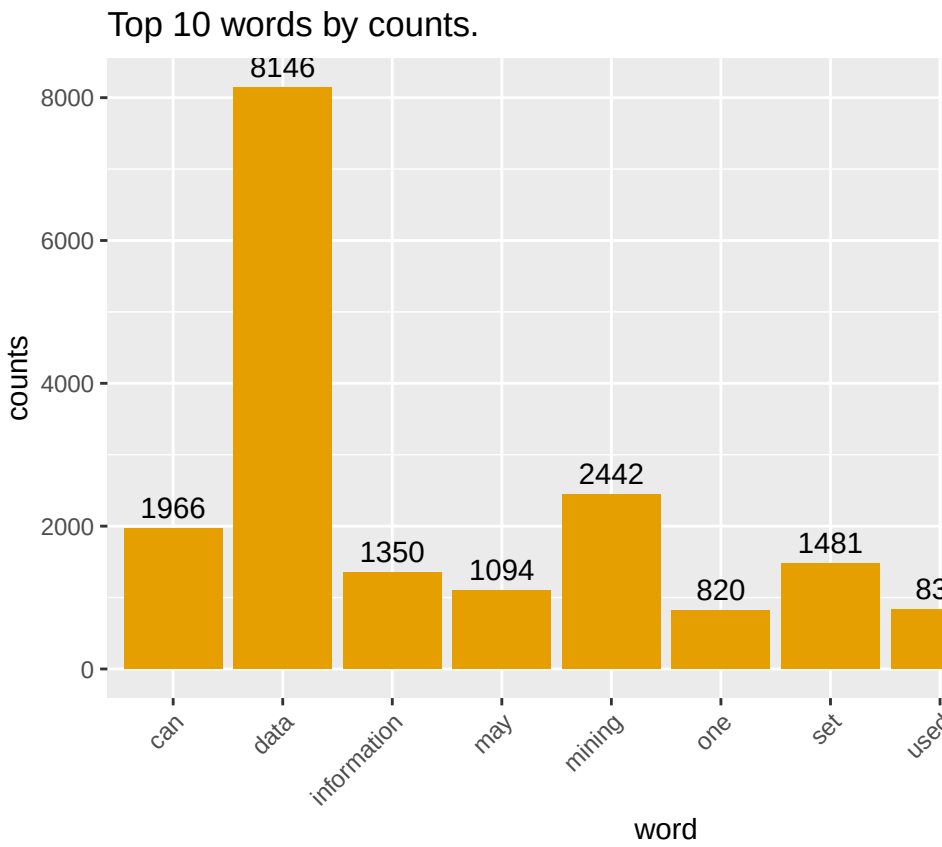
```
# top 10 words and their respective counts
df <-
    my_tdm %>%
    as.matrix() %>%
    rowSums() %>%
    sort(decreasing = TRUE) %>%
    head(10) %>%
    enframe(name = "word", value = "counts")

head(df)
```

Top words in TDM

```
## # A tibble: 6 x 2
##   word        counts
##   <chr>        <dbl>
## 1 data          8146
## 2 mining        2442
## 3 value         2411
## 4 can           1966
## 5 set           1481
## 6 information   1350
```

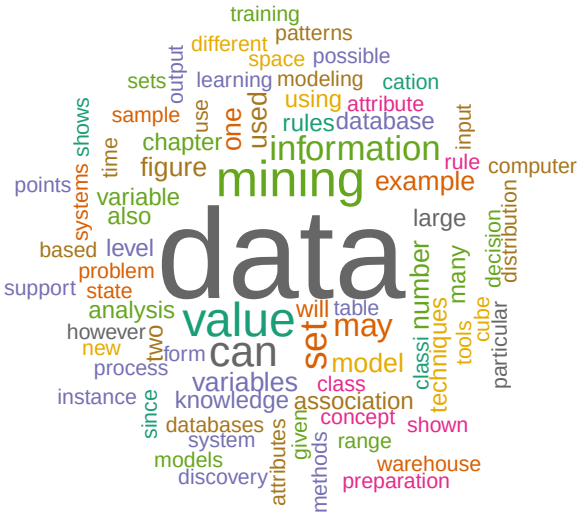
```
# top 10 words and counts using bargraph
bargraph <- ggplot(df, aes(word, counts)) +
  geom_bar(stat = "identity", fill = "#E69F00") +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Top 10 words by counts.") +
  geom_text(aes(label = counts), vjust = -0.5)
bargraph
```

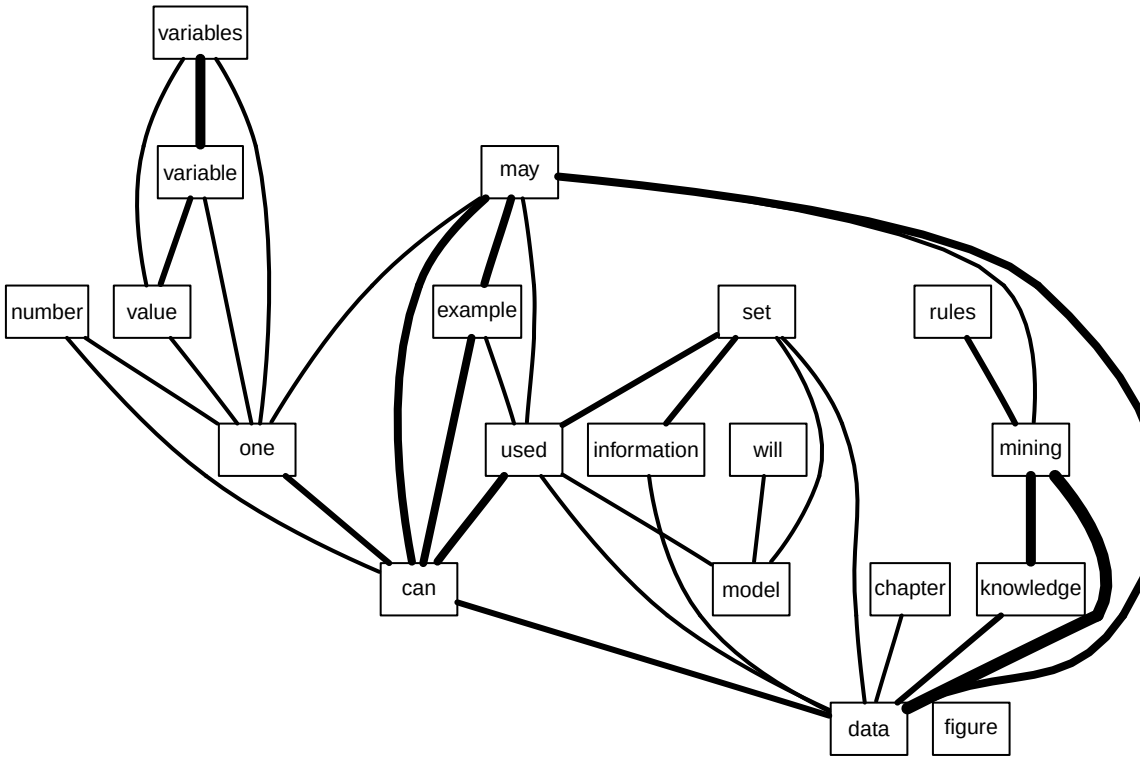


Bar Grahph for Top 10 word counts

```
mat <- as.matrix(my_tdm)
freq <- mat %>% rowSums() %>% sort(decreasing = T)

# plot word cloud
wordcloud(
  words = names(freq),
  freq = freq,
  min.freq = 300,
  max.words = 500,
  random.order = FALSE,
  colors = brewer.pal(8, "Dark2"),

  random.color = TRUE,
  rot.per = 0.35,
  use.r.layout = FALSE
)
```




```
# correlation between top 600 frequent terms
top_600_frequent_tems <- findFreqTerms(my_tdm, lowfreq = 600)
plot(my_tdm, terms = top_600_frequent_tems, corThreshold = 0.2, weighting = T)
```



Word Correlation