Data Analytics
Assignment-IV: Text Analytics

Name: - Ashish
Roll no.: - 17BCS006

==========

```
> install.packages('SnowballC')
> install.packages('tm')
> library(tm)
Loading required package: NLP
>
> tw <- read.csv("C:/Users/Ashish/Downloads/Ashish - Classes/D.A/LAB/Assig-5/
Trump_Data.csv", ,stringsAsFactors = FALSE)
>
>
> #tw1 <- tw$text
> tw1 <- tw["text"]
> head(tw1)

text
1 Judge Kavanaugh showed America exactly why I nominated him. His testimony was power
ful honest and riveting. DemocratsÃ• search and destroy strategy is disgraceful and
this process has been a total sham and effort to delay obstruct and resist. The
Senate must vote!
2
https://t.co/9o5gZlJiTd
3
Join me this Saturday in Wheeling West Virginia at 7pmE! Tickets: https://t.co/
JyRaBpsOeR https://t.co/hiruLIxa7w
4
Congressman Lee Zeldin is doing a fantastic job in D.C. Tough and smart he loves our
Country and will always be there to do the right thing. He has my Complete and Total
Endorsement!
5                  China is actually placing propaganda ads in the Des Moines Register
other papers made to look like news. ThatÃ•s because we are beating them on Trade
opening markets and the farmers will make a fortune when this is over!
https://t.co/ppdvTX7oz16  Avenatti is a third rate lawyer who is good at making false
usations like he did on me and like he is now doing on Judge Brett Kavanaugh.
He is just looking for attention and doesnÃ•t want people to look at his past record
relationships- a total low-life!
>
>
> # Create Corpus
> docs <- VCorpus(VectorSource(tw1))
> summary(docs)
  Length Class             Mode
1 2      PlainTextDocument list
> inspect(docs)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

$text
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 1544659

>
> # inspect a particular document
> writeLines(as.character(docs[[1]]))
@BillLester651: @DanScavino @realDonaldTrump Latinos for Trump believes in Trump
```

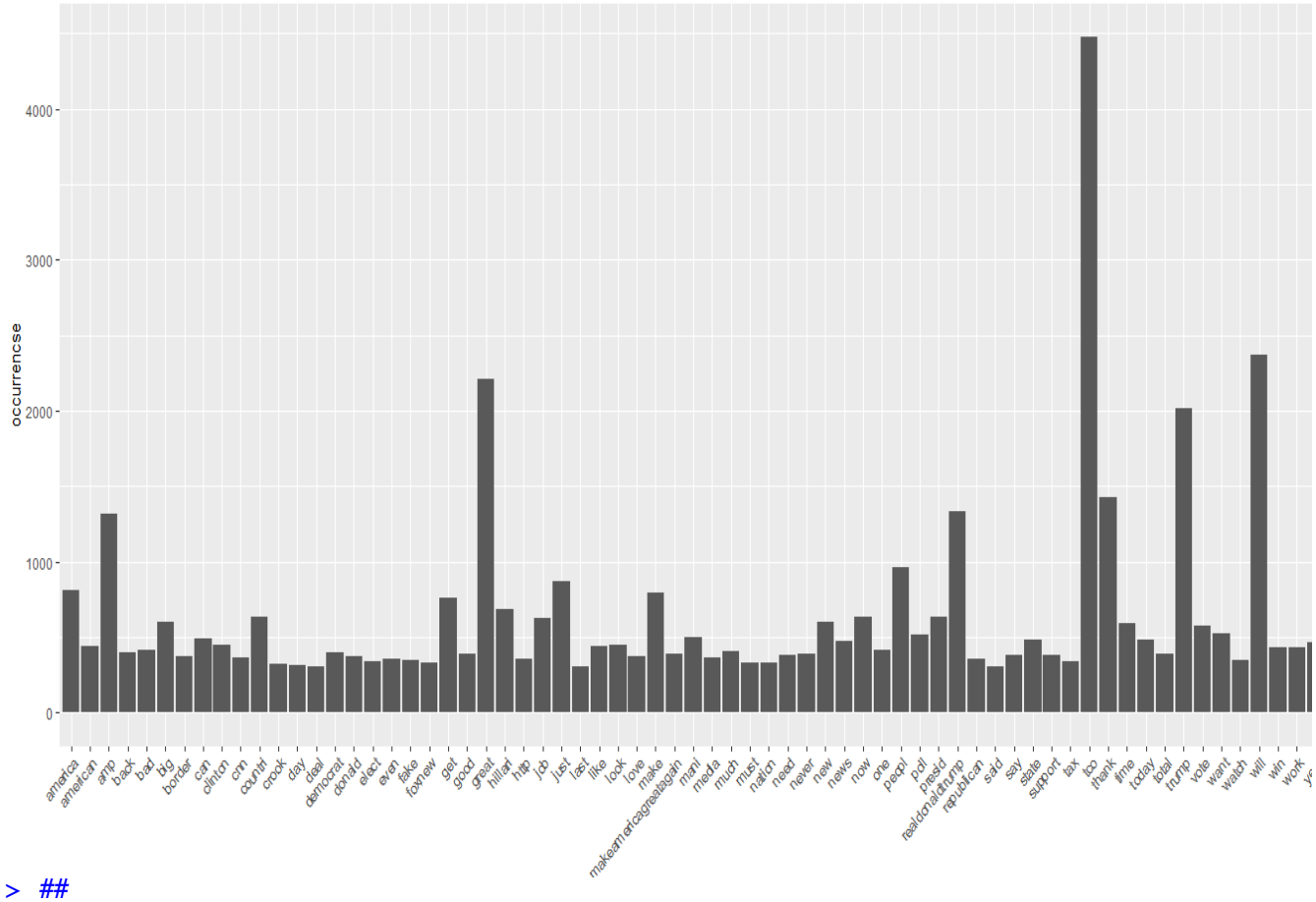
```
Getting ready to leave for my GREAT resort Turnberry in Scotland. Hosting The
 Women's British Open (biggest tournament). Will be back Sat.
I really like the Koch Brothers (members of my P.B. Club) but I don't want their
 money or anything else from them. Cannot influence Trump!
Thank you. https://t.co/oRATmdIzqE
People like lawyer Elizabeth Beck and failed writer Harry Hurt &amp;amp; others talk
about me but know nothing about meÃ‘crazy!
How can a dummy dope like Harry Hurt who wrote a failed book about me but doesnÃ•t
know me or anything about me be on TV discussing Trump?
Trump Will Make America GREAT!!!!  #ChangeTheWorldIn5Words
... while Tom Brady is guilty because he REPLACED his LEGAL cellphone?
Per @rushlimbaugh: Why does Hillary Clinton get the benefit of the doubt (after she
 DESTROYS her illegal email server) ...
Via @SaintPetersblog by @MitchEPerry: Ã’Shock poll: Donald Trump leads Jeb Bush
26-20% Ã‰ in FloridaÃ“ http://t.co/49IAe1Nlm7
.
.
.
.
.
.
.
@ZStr8Up: After the liberal 60's and 70's how did we end up with Reagan? Hollywood.
 The Donald may have a serious chance. @realDonaldTrump
@fyrftr211: @realDonaldTrump #Trump2016. Time for a true leader to lead our great
country!
@MediciMario: @realDonaldTrump Pls run u would be great!
Tomorrow will be a really big day for America. MAKE AMERICA GREAT AGAIN!
> ##
> ## ----- Start Preprocessing ------- ##
> ##
> toSpace <- content_transformer(function(X,pattern) { return(gsub(pattern, " ",X))})
> docs <- tm_map(docs, toSpace, "-")
> docs <- tm_map(docs, toSpace, ":")
> docs <- tm_map(docs, toSpace, "'")
> docs <- tm_map(docs, toSpace, " _")
> docs <- tm_map(docs, toSpace, "-")
> docs <- tm_map(docs, toSpace, "-")
> docs <- tm_map(docs, toSpace, "/")
> docs <- tm_map(docs, toSpace, "@")
> docs <- tm_map(docs, toSpace, "\\|")
> docs <- tm_map(docs, toSpace, "https")
>
>
> ##
> # Remove punctuation
> docs <- tm_map(docs, removePunctuation)
> ##
> # Treansform to lower case
> docs <- tm_map(docs, content_transformer(tolower))
> ##
> # strip digits
> docs <- tm_map(docs, removeNumbers)
> ##
> # Remove stopword from standard stopword list (How to chek this? How to add your
 own?)
> docs <- tm_map(docs, removeWords, stopwords("english"))
> ##
> # Strip whitespace (cosmetic?)
> docs <- tm_map(docs, stripWhitespace)
> ##
> # inspect output
> writeLines(as.character(docs[[1]]))
billlester danscavino realdonaldtrump latinos trump believes trump
getting ready leave great resort turnberry scotland hosting women s british open
biggest tournament will back sat
 really like koch brothers members pb club don t want money anything else influence
trump
thank tco oratmdizqe
```

```
people like lawyer elizabeth beck failed writer harry hurt amp others talk know
nothing meã‘crazy
 can dummy dope like harry hurt wrote failed book doesnã•t know anything tv
discussing trump
trump will make america great changetheworldinwords
 tom brady guilty replaced legal cellphone
per rushlimbaugh hillary clinton get benefit doubt destroys illegal email server
via saintpetersblog mitcheperry ã’shock poll donald trump leads jeb bush ã‰
floridaã“ http tco iaenlm
 jereegeestavich realdonaldtrump blowing everyone else away just ahead doubling wow
 outrage bias free language guide claims word american problematic http tco vkzaicgs
 tomnocera danscavino dallasmavs mcuban trump s work ethic gives hope s tireless
 smart worker winner whiner
 truly love millions people sticking despite many media lies great silent majority
loomin
.
.
.
.
.
tomorrow makeamericagreatagain tco pdmndtovvc
 looneytunes politicians going get us mess talk simpletruth makeamericagreatagain
 zstrup liberal s s end reagan hollywood donald may serious chance realdonaldtrump
 fyrftr realdonaldtrump trump time true leader lead great country
 medicimario realdonaldtrump pls run u great
tomorrow will really big day america make america great
> ## Need Snowballc library for stemming
> library(SnowballC)
> # Stem document
> docs <- tm_map(docs, stemDocument)
> ##
> ##
> ## ------- some clean up -----------------  ##
> ##
> ##
> docs <- tm_map(docs, content_transformer(gsub), pattern = "organiz",
 replacement = "organ")
> docs <- tm_map(docs, content_transformer(gsub), pattern = "organis",
 replacement = "organ")
> docs <- tm_map(docs, content_transformer(gsub), pattern = "andgovern",
replacement = "govern")
> docs <- tm_map(docs, content_transformer(gsub), pattern = "inenterpris",
replacement = "enterp")
> docs <- tm_map(docs, content_transformer(gsub), pattern = "team-",
replacement = "team")
> #inspect
> writeLines(as.character(docs[[1]]))
jaketapp donaldtrump iowa say gucci store s worth money romney fact check true http
tco xvdvhmntvr
davidsbaldwin realdonaldtrump today s day trump good luck
desheay realdonaldtrump twitter banner awesom cant wait tomorrowtrump
premus realdonaldtrump let s take countri back trumpl
jarodpitmon full support trump realdonaldtrump guy will help countri
nickyflash trumppresid trump realdonaldtrump
rhumeey realli want see america look likewith realdonaldtrump helm affair tco
zfvjpyjvoh
snurk realdonaldtrumpãšlov alway respect fighter overr loser polititian inspir great
trump
knight realdonaldtrump realjoemurray america can great trump helm confid count
isabelsimon realdonaldtrump brentcfritz donald good chanc anyon el peopl like himno
 politician s
donjubb realdonaldtrump shock worldtrumppresid
insuraid realdonaldtrump go make hillari s head spin tomorrow morn presidenti debat
yes plea trump
brentcfritz today day america becom great realdonaldtrump huge news will shock world
.
.
.
```

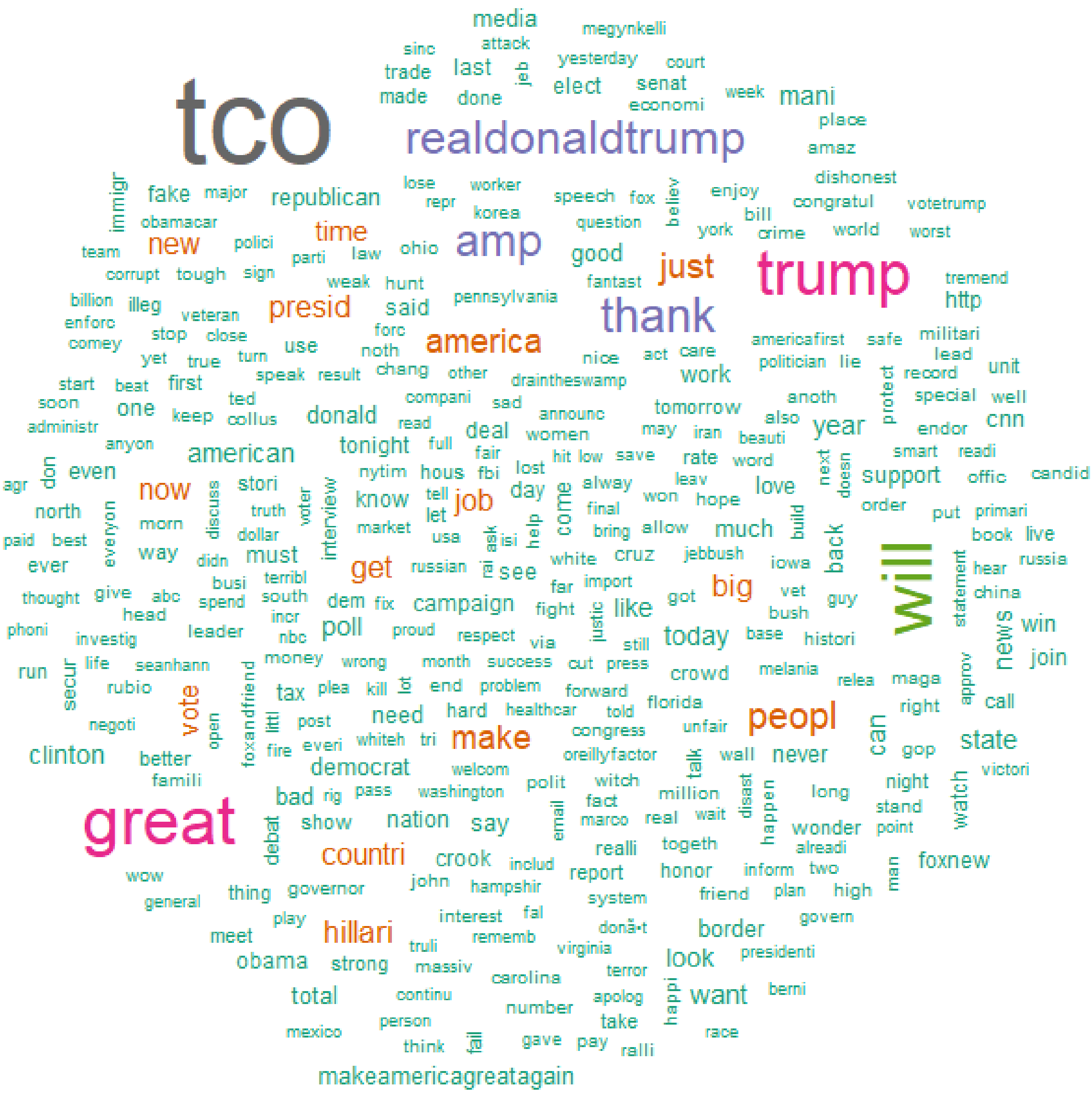
```
.
.
.
.
great tco oetkjzrd
thank tco qhtxizplx
live periscop tco hlrtzbgeb
tomorrow makeamericagreatagain tco pdmndtovvc
looneytun politician go get us mess talk simpletruth makeamericagreatagain
zstrup liber s s end reagan hollywood donald may serious chanc realdonaldtrump
fyrftr realdonaldtrump trump time true leader lead great countri
medicimario realdonaldtrump pls run u great
tomorrow will realli big day america make america gre
> ##
> ##
> ## ---- Create document-term matrix
> ##
> ##
> dtm <- TermDocumentMatrix(docs)
> m <- as.matrix(dtm)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)
> head(d, 10)
                           word freq
tco                         tco 4474
will                       will 2375
great                     great 2207
trump                     trump 2016
thank                     thank 1429
realdonaldtrump realdonaldtrump 1335
amp                         amp 1316
peopl                     peopl  963
just                       just  873
america                 america  816
> ##
> ##
> ##
> # - collapse matrix by summing over columns - this gets total counts
(over all docs) for each text
> freq <- rowSums(as.matrix(dtm))
> ##
> ##
> # - length should be total number of terms
> length(freq)
[1] 16322
> ##
> ##
> # - create sort order (asc)
> ord <- order(freq, decreasing =  TRUE)
> ##
> ##
> # - inspect most frequently occurring terms
> freq[head(ord)]
tco           will          great         trump          thank realdonaldtrump
4474           2375           2207           2016           1429           1335
> ##
> ##
> # - inspect least frequently occurring terms
> freq[tail(ord)]
 zzjbgdzpe  zzkmwuyqi  zzlegvnjn    zzlscdn    zzudwaf zzycnodcqj
         1          1          1          1          1          1
> ##
> ##
> # remove  words
> ##
> ##
> dtmr <- DocumentTermMatrix(docs, control = list(wordLengths = c(2,20),
bounds = list(global = c(1,27))))
>
```

```
> dtmr <- TermDocumentMatrix(docs)
> m <- as.matrix(dtmr)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)
> head(d, 10)
                           word freq
tco                         tco 4474
will                       will 2375
great                     great 2207
trump                     trump 2016
thank                     thank 1429
realdonaldtrump realdonaldtrump 1335
amp                         amp 1316
peopl                     peopl  963
just                       just  873
america                 america  816
>#collapse matrix by summing over columns - this gets total counts (over all docs)
for each text
> freqr <- rowSums(as.matrix(dtmr))
> # - length should be total number of terms
> length(freqr)
[1] 16322
> # - create sort order (asc)
> ord <- order(freqr, decreasing =  TRUE)
> # - inspect most frequently occurring terms
> freq[head(ord)]
tco           will          great          trump          thank realdonaldtrump
4474           2375           2207           2016           1429           1335
> ##
> ##
> # - inspect least frequently occurring terms
> freq[tail(ord)]
 zzjbgdzpe  zzkmwuyqi  zzlegvnjn    zzlscdn    zzudwaf zzycnodcqj
         1          1          1          1          1          1
>
>
> # list most frequent terms. Lower bound Specified as second argument
> findFreqTerms(dtmr, lowfreq = 100)
  [1] "administr"             "allow"                 "also"
  [4] "alway"                 "amaz"                  "america"
  [7] "american"              "amp"                   "announc"
 [10] "anoth"                 "ask"                   "attack"
 [13] "back"                  "bad"                   "beat"
 [16] "believ"                "best"                  "better"
 [19] "big"                   "bill"                  "billion"
 [22] "book"                  "border"                "bring"
 [25] "bush"                  "busi"                  "call"
 [28] "campaign"              "can"                   "candid"
 [31] "care"                  "carolina"              "chang"
 [34] "china"                 "clinton"               "cnn"
 [37] "collus"                "come"                  "compani"
 [40] "congratul"             "congress"              "continu"
 [43] "countri"               "crime"                 "crook"
 [46] "crowd"                 "cruz"                  "cut"
.
.
.
.
.
.
 [941] "late"            "later"           "latest"          "laugh"         "launch"
 [946] "law"             "lawsuit"         "lawyer"          "lead"          "leader"
 [951] "leadership"      "leagu"           "leak"            "leaker"        "learn"
 [956] "least"           "leav"            "led"             "left"          "legal"
 [961] "legendari"       "legisl"          "lesm"            "less"          "let"
 [966] "letã•"           "letter"          "level"           "liar"          "liber"
 [971] "liberti"         "lie"             "life"            "lift"          "light"
 [976] "lightweight"     "like"            "likewi"          "limbaugh"      "limit"
```

```
 [981] "lindsey"         "lindseygrahamsc" "line"            "link"            "lisa"
 [986] "list"            "listens"         "littl"           "live"            "loan"
 [991] "lobbyist"        "local"           "locat"           "london"          "long"
 [996] "longer"          "look"            "lose"            "loser"           "loss"
 [ reached getOption("max.print") -- omitted 864 entries]
> ##
> ##
> ##
> # -------histogram plot--------------
> ##
> ##
> ##
> wf = data.frame(term = names(freqr), occurrencse = freqr)
> library(ggplot2)
> p <- ggplot(subset(wf, freqr>200), aes(term, occurrencse))
> p <- p + geom_bar(stat = "identity")
> p <- p + theme(axis.text.x = element_text(angle = 45, hjust = 1))
> p
```




```
> ##
> ##
> # - Wordcloud --------- #
> ##
> ##
> #install.packages('wordcloud')
> library(wordcloud)
> set.seed(42)
> # limit words by spcifying min frequency
> wordcloud(names(freqr), freqr, min.freq = 70, colors = brewer.pal(6, "Dark2"))
```



media
megynkelli
sinc
attack
trade
last
jeb
yesterday
court
made
done
elect
senat
week
mani
economi
tco
realdonaldtrump
place
amaz
immigr
fake
major
republican
lose
worker
speech
fox
believ
enjoy
dishonest
repr
korea
bill
congratul
votetrump
obamacar
question
york
crime
world
worst
team
new
polici
time
ohio
amp
good
parti
law
trump
corrupt
tough
sign
just
tremend
weak
hunt
fantast
http
billion
illeg
veteran
presid
said
pennsylvania
thank
enforc
forc
americafirst
safe
militari
comey
stop
close
use
america
noth
nice
act
care
lead
politician
lie
yet
true
turn
chang
other
work
protect
record
unit
start
beat
first
speak
result
draintheswamp
special
well
soon
ted
compani
sad
tomorrow
anoth
one
keep
collus
donald
announc
also
cnn
administr
read
deal
may
iran
year
endor
anyon
tonight
full
women
beauti
smart
readi
american
fair
hit
low
save
rate
word
agr
don
even
nytim
hous
fbi
lost
alway
next
doesn
support
offic
candid
stori
leav
now
discuss
voter
interview
tell
job
day
come
love
north
truth
know
let
final
won
hope
order
put
primari
morn
dollar
market
help
bring
allow
much
build
book
live
paid
best
usa
ask
isi
everyon
back
way
didn
must
white
cruz
jebbush
iowa
will
ever
russian
see
russia
get
terribl
hear
busi
far
import
big
vet
statement
thought
give
abc
spend
south
dem
fix
campaign
fight
like
got
guy
china
head
incr
bush
news
win
phoni
investig
leader
nbc
poll
proud
respect
via
justic
still
today
base
histori
money
wrong
month
success
cut
press
join
life
seanhann
crowd
melania
run
secur
foxandfriend
relea
maga
approv
rubio
vote
tax
plea
kill
lot
end
problem
forward
florida
peopl
right
call
negoti
open
post
need
hard
healthcar
told
unfair
can
littl
fire
everi
whiteh
tri
make
congress
state
clinton
better
gop
democrat
oreillyfactor
wall
never
welcom
talk
famili
night
victori
polit
witch
bad
rig
pass
washington
million
disast
long
stand
email
fact
happen
great
show
nation
say
marco
real
wait
wonder
point
debat
realli
togeth
alreadi
countri
watch
crook
includ
inform
two
foxnew
report
honor
wow
john
hampshir
man
thing
governor
system
friend
plan
high
general
play
interest
fal
govern
hillari
border
meet
remember
donât
presidenti
truli
virginia
look
obama
strong
massiv
terror
carolina
happi
want
berni
total
continu
apolog
number
person
take
race
mexico
fail
gave
pay
think
ralli
makeamericagreatagain