# Homework 5

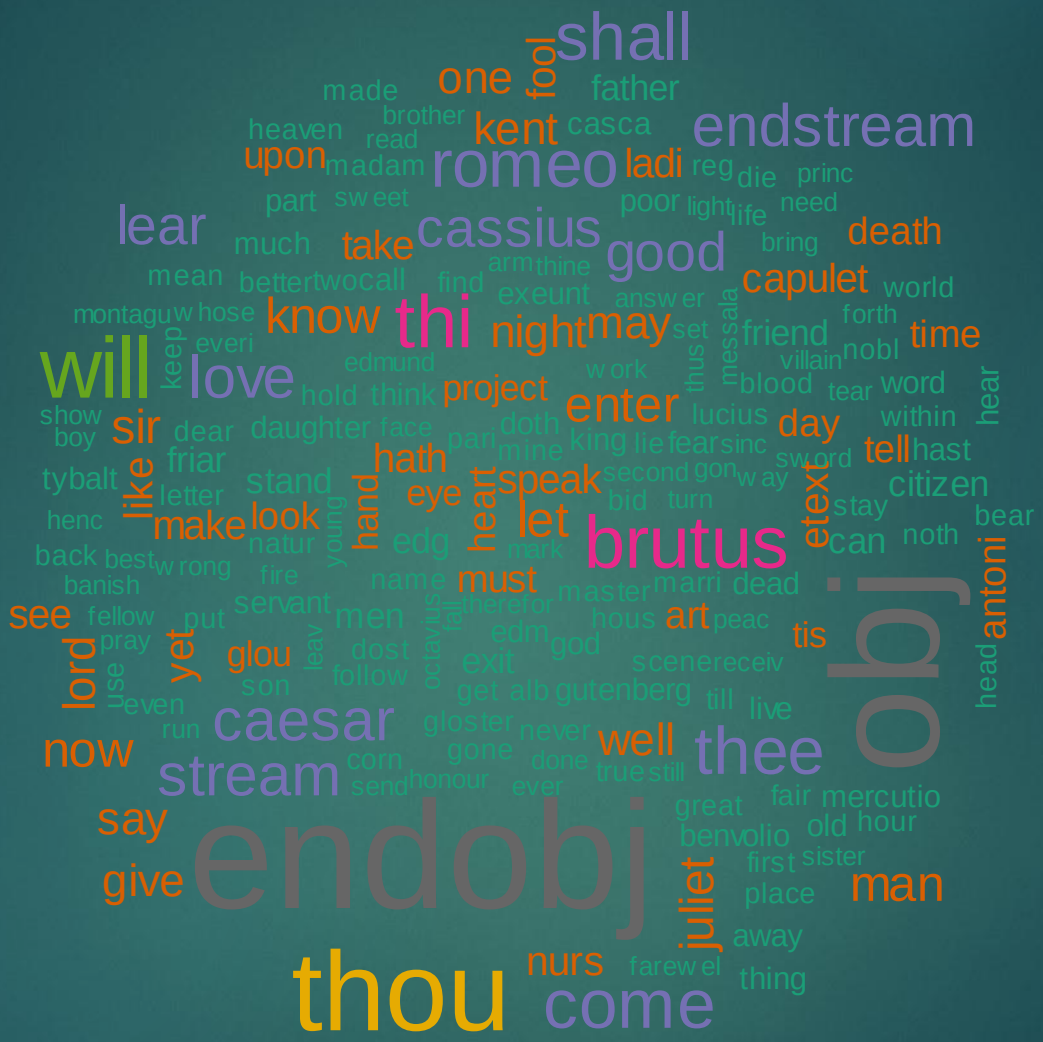
ZHENYU XU

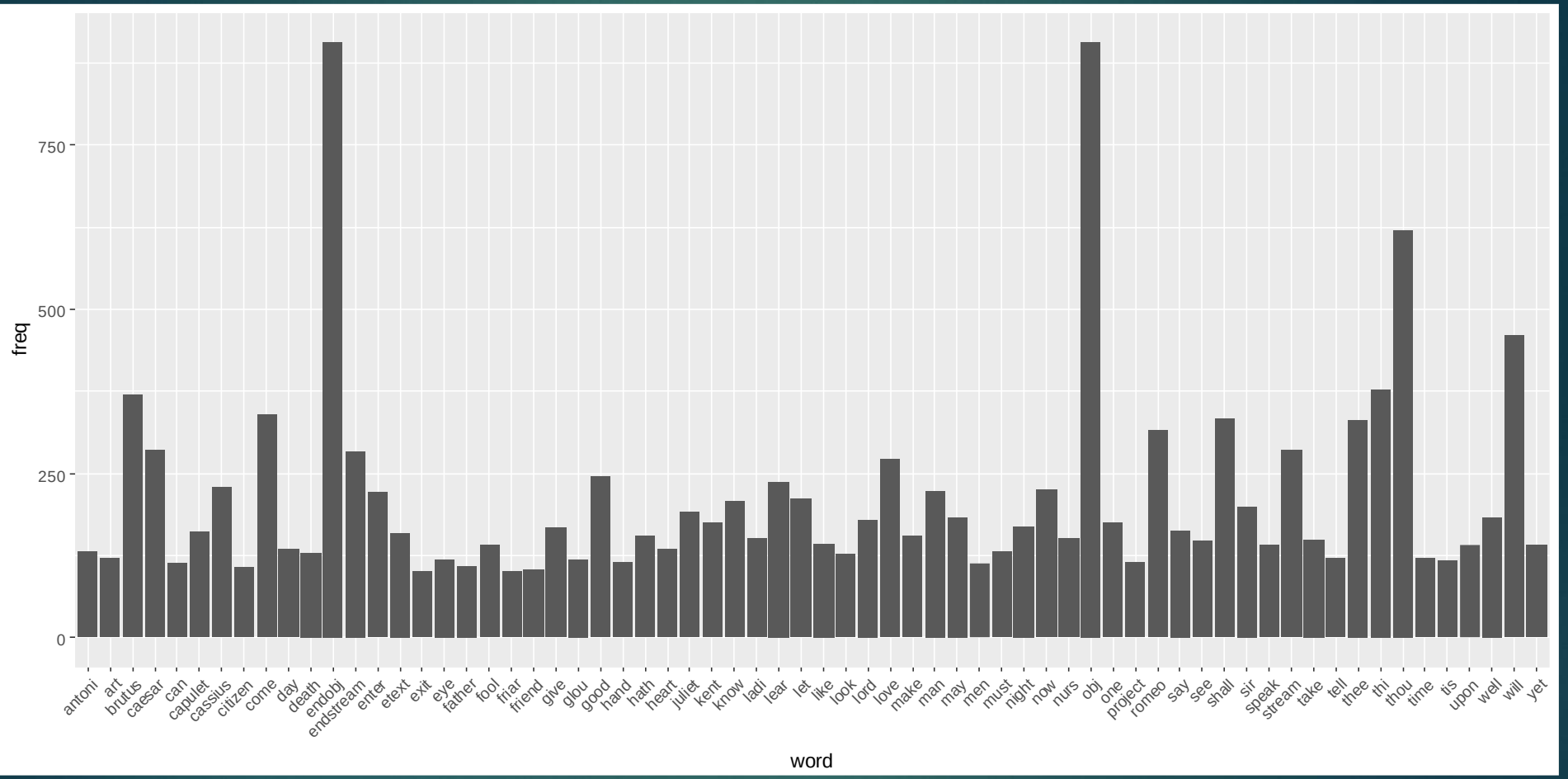
15620161152286

# Q1、2

```
rm(list = ls())
library(RCurl)
library(XML)
library(bitops)
library(stringr)
url=paste(c("http://publicliterature.org/pdf/2ws1610.pdf","http://publicliterature.org/pdf/2ws2410.pd
      f","http://publicliterature.org/pdf/2ws3310.pdf") )
abs=lapply(url, FUN = function(x) htmlParse(x, encoding = "Latin-1"))
clean_txt = function(x) {
  cleantxt = xpathApply(x, "//body//text()
                        [not(ancestor :: script)][ not(ancestor :: style)]
                        [not(ancestor :: noscript)] " ,xmlValue)
  cleantxt = paste(cleantxt, collapse="\n")
  cleantxt = str_replace_all(cleantxt, "\n", " ")
  cleantxt = str_replace_all(cleantxt, "\r", "")
  cleantxt = str_replace_all(cleantxt, "\t", "")
  cleantxt = str_replace_all(cleantxt, "<br>", "")
  return(cleantxt)
}
cleantxt = lapply(abs,clean_txt)
vec_abs = unlist(cleantxt)
vec_abs
```

```
library(tm)
library(SnowballC)
abs      = Corpus(VectorSource(vec_abs))
abs_dtm  = DocumentTermMatrix(abs, control = list(
  stemming = TRUE, stopwords = TRUE, minWordLength = 3,
  removeNumbers = TRUE, removePunctuation = TRUE))
dim(abs_dtm)
inspect(abs_dtm)
#Find the words that occur more than 5 times
findFreqTerms(abs_dtm, 5)
#Remove sparse terms
removeSparseTerms(abs_dtm, 0.5)
inspect(removeSparseTerms(abs_dtm, 0.5))
library(ggplot2)
library(wordcloud)
freq = colSums(as.matrix(abs_dtm))
wf   = data.frame(word=names(freq), freq=freq)
plot = ggplot(subset(wf, freq>100), aes(word, freq))
plot = plot + geom_bar(stat="identity")
plot = plot + theme(axis.text.x=element_text(angle=45, hjust=1))
plot
freq  = colSums(as.matrix(abs_dtm))
dark2 = brewer.pal(8, "Dark2")
wordcloud(names(freq), freq, max.words=200, rot.per=0.1, colors=dark2)
dev.off()
```


shall
fool
one
father
made
brother
heaven
read
kent
casca
endstream
upon
madam
romeo
ladi
reg
die
princ
part
sweet
poor
light
life
need
lear
much
take
cassius
good
bring
death
mean
bettertwocall
find
arm
thine
exeunt
answer
capulet
world
montagu
w
hose
know
thi
night
may
set
friend
forth
time
keep
everi
will
love
edmund
work
thus
messala
villain
nobl
hold
think
project
enter
blood
tear
word
hear
show
sir
dear
daughter
face
pari
doth
king
lie
fear
sinc
lucius
day
within
boy
friar
mine
sword
tell
hast
tybalt
like
stand
hath
second
gon
way
citizen
letter
hand
eye
heart
speak
bid
turn
etext
stay
henc
make
look
let
brutus
can
noth
bear
natur
young
edg
mark
back
best
w
rong
fire
name
must
master
marri
dead
banish
see
fellow
put
servant
men
therefor
hous
art
peac
antoni
pray
yet
glou
dost
leav
edm
god
tis
lord
use
follow
exit
scenereceiv
head
son
get
alb
gutenberg
till
live
even
caesar
run
gloster
never
well
thee
obj
now
stream
corn
gone
done
true
still
send
honour
ever
great
fair
mercutio
say
endobj
benvolio
old
hour
give
first
sister
juliet
man
place
away
nurs
farew
el
thing
thou
come


freq
750
500
250
0
word
antoni
art
brutus
caesar
can
capulet
cassius
citizen
come
day
death
endobj
endstream
enter
etext
exit
eye
father
fool
friar
friend
give
glou
good
hand
hath
heart
juliet
kent
know
ladi
lear
let
like
look
lord
love
make
man
may
men
must
night
now
nurs
obj
one
project
romeo
say
see
shall
sir
speak
stream
take
tell
thee
thi
thou
time
tis
upon
well
will
yet

hist(freq, col = "grey", breaks = 20,ylim = c(0, 5000), xlab = "freq of words")