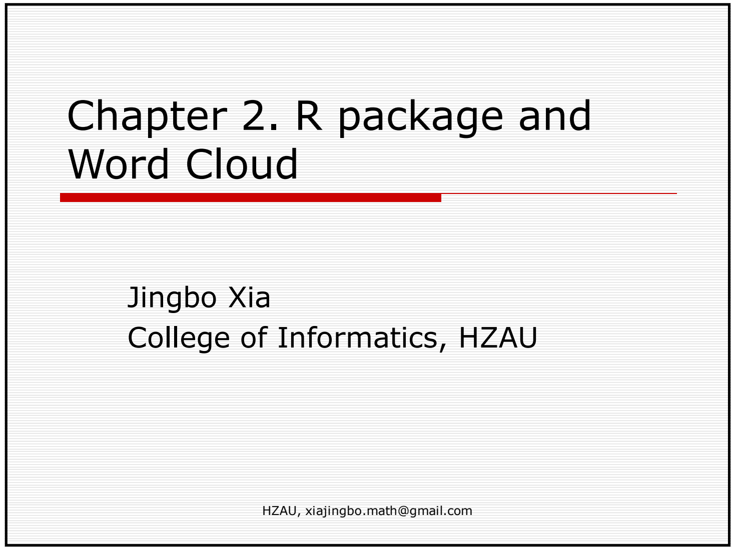
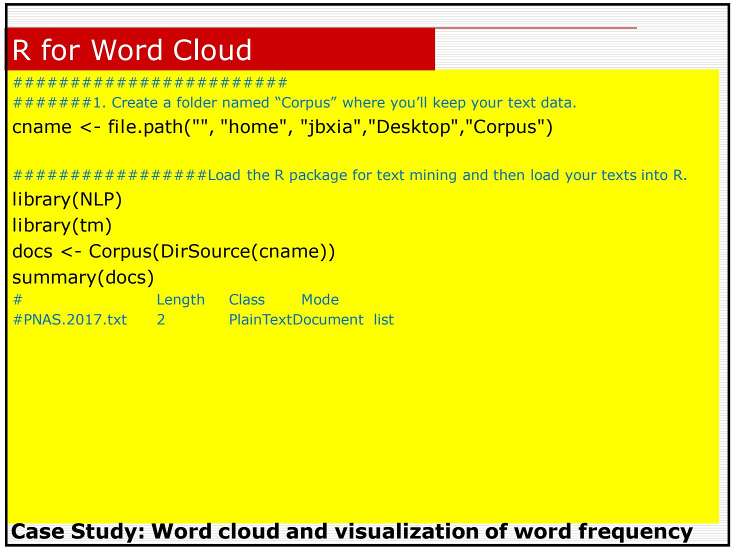


Chapter 2. R package and Word Cloud

Jingbo Xia
College of Informatics, HZAU

HZAU, xiajingbo.math@gmail.com

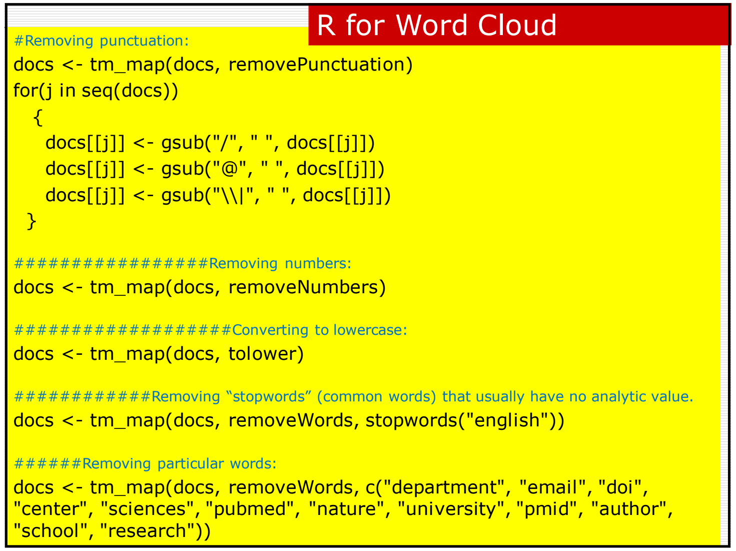


R for Word Cloud

```
########################
#######1. Create a folder named “Corpus” where you’ll keep your text data.
cname <- file.path("", "home", "jbxia","Desktop","Corpus")

#################Load the R package for text mining and then load your texts into R.
library(NLP)
library(tm)
docs <- Corpus(DirSource(cname))
summary(docs)
#                  Length     Class            Mode
#PNAS.2017.txt     2          PlainTextDocument  list
```

Case Study: Word cloud and visualization of word frequency



R for Word Cloud

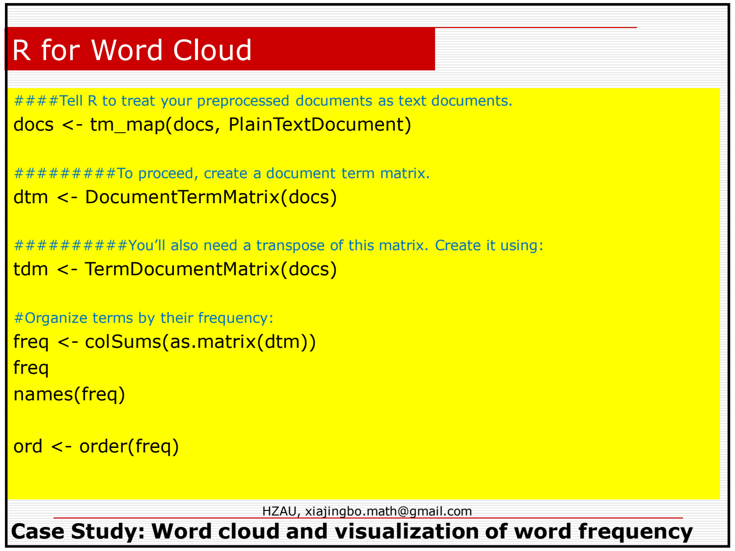
```
#Removing punctuation:
docs <- tm_map(docs, removePunctuation)
for(j in seq(docs))
  {
    docs[[j]] <- gsub("/", " ", docs[[j]])
    docs[[j]] <- gsub("@", " ", docs[[j]])
    docs[[j]] <- gsub("\\|", " ", docs[[j]])
 }

#################Removing numbers:
docs <- tm_map(docs, removeNumbers)

###################Converting to lowercase:
docs <- tm_map(docs, tolower)

############Removing “stopwords” (common words) that usually have no analytic value.
docs <- tm_map(docs, removeWords, stopwords("english"))

######Removing particular words:
docs <- tm_map(docs, removeWords, c("department", "email", "doi",
"center", "sciences", "pubmed", "nature", "university", "pmid", "author",
"school", "research"))
```



R for Word Cloud

```
####Tell R to treat your preprocessed documents as text documents.
docs <- tm_map(docs, PlainTextDocument)

#########To proceed, create a document term matrix.
dtm <- DocumentTermMatrix(docs)

##########You’ll also need a transpose of this matrix. Create it using:
tdm <- TermDocumentMatrix(docs)

#Organize terms by their frequency:
freq <- colSums(as.matrix(dtm))
freq
names(freq)

ord <- order(freq)
```

HZAU, xiajingbo.math@gmail.com

Case Study: Word cloud and visualization of word frequency

R for Word Cloud

```
###If you prefer to export the matrix to Excel:
m <- as.matrix(dtm)
write.csv(m, file="dtm.csv")


###############Word Frequency
###There are lots of terms, just check some of the most and least frequently occurring words.
freq[head(ord, 10)]
freq[tail(ord, 50)]


wf <- data.fram(word = names(freq), freq=freq)
Head(wf)
```

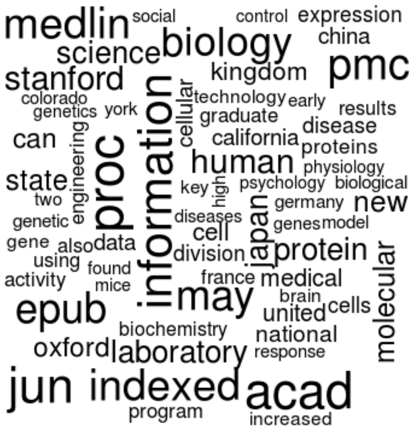
Case Study: Word cloud and visualization of word frequency

R for Word Cloud

```
###########Plot words that appear at least 50 times.
library(ggplot2)
p <- ggplot(subset(wf, freq>50), aes(word, freq))
p <- p + geom_bar(stat="identity")
p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))
p


#########word cloud
set.seed(142)
wordcloud(names(freq), freq, min.freq=25)
```

Case Study: Word cloud and visualization of word frequency



Assignment:

Currently, you know how to visualize the word frequency in texts via certain built package.

Please download interested Pubmed texts and find something novel and prepare a 10 minutes talk in next class. Two members win the points.

Chapter 2.
R package and Word Cloud