安装包裹

# Install

install.packages("tm") # for text mining

install.packages("SnowballC") # for text stemming

install.packages("wordcloud") # word-cloud generator

install.packages("RColorBrewer") # color palettes

载入包裹

# Load

library("tm")

library("SnowballC")

library("wordcloud")

library("RColorBrewer")

读取文件

text <- readLines(file.choose())

docs <- Corpus(VectorSource(text))

inspect(docs)

toSpace <- content\_transformer(function (x , pattern ) gsub(pattern, " ", x))

docs <- tm\_map(docs, toSpace, "/")

docs <- tm\_map(docs, toSpace, "@")

docs <- tm\_map(docs, toSpace, "\\|")

# Remove numbers去除数字

docs <- tm\_map(docs, removeNumbers)

# Remove punctuations 去除标点符号

docs <- tm\_map(docs, removePunctuation)

# Eliminate extra white spaces去除空格，

docs <- tm\_map(docs, stripWhitespace)

# Remove your own stop word

# specify your stopwords as a character vector去除指定词

docs <- tm\_map(docs, removeWords, c("ulnone", "pardpardeftabrislslmultqjpartightenfactor", "pardpardeftabfirislslmultqjpartightenfactor", "pardpardeftabfirislslmultqjpartightenfactor"))

Document matrix is a table containing the frequency of the words. Column names are words and row names are documents.

用document matrix 指令查看频率最高的前13个词汇

dtm <- TermDocumentMatrix(docs)

m <- as.matrix(dtm)

v <- sort(rowSums(m),decreasing=TRUE)

d <- data.frame(word = names(v),freq=v)

head(d, 13)

> head(d, 13)

word freq

the the 360

you you 185

and and 130

for for 88

was was 66

have have 57

with with 51

are are 50

she she 50

can can 47

her her 46

when when 44

his his 43

|  |  |  |
| --- | --- | --- |
|  | 词汇 | 频率 |
| 1 | The | 360 |
| 2 | You | 185 |
| 3 | And | 130 |
| 4 | For | 88 |
| 5 | Was | 66 |
| 6 | Have | 57 |
| 7 | With | 51 |
| 8 | Are | 50 |
| 9 | She | 50 |
| 10 | Can | 47 |
| 11 | Her | 46 |
| 12 | When | 44 |
| 13 | His | 43 |

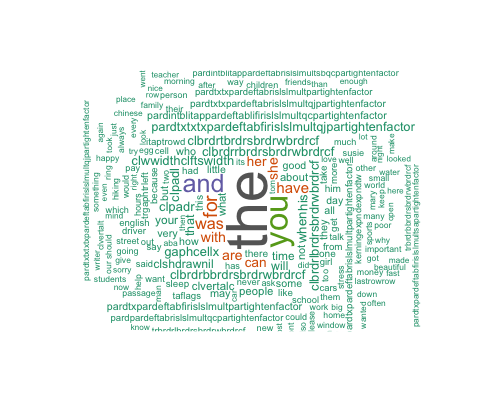
画出Cloud Word图

set.seed(1234)

wordcloud(words = d$word, freq = d$freq, min.freq = 1,

max.words=200, random.order=FALSE, rot.per=0.35,

colors=brewer.pal(8, "Dark2"))



histogram:

barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,

col ="lightblue", main ="Most frequent words",

ylab = "Word frequencies")

