

```
# 1a. Install the required packages
install.packages("tm") # for text mining
install.packages("SnowballC") # for text stemming
install.packages("wordcloud") # word-cloud generator
install.packages("RColorBrewer") # color palettes
```

```
# 1b. Load the required packages
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

```
# 2. Text mining
# load the text
# getpath and setpath
> getwd()
> list.files()
> filePath = paste(getwd(), "/dream-speech.txt", sep="")
> help.search("concatenate")
> str(filePath)
> text <- readLines(filePath)
> docs <- Corpus(VectorSource(text))
> ls()
> list.files()
> names(docs)
> str(docs)
> inspect(docs)
```

```
# Text transformation
# tm_map() is used to remove unnecessary white space, to convert the text to lower case, to
remove common stopwords like 'the', "we".
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
```

```
# Cleaning the text
# Also remove numbers and punctuation with removeNumbers and removePunctuation
arguments using tm_map()
```

```
# Convert the text to lower case
```

```

docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)

# 3. Build a term-document 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)

# 4. Generate the different Wordcloud
# seed number you choose is the starting point used in the generation of a sequence of random
numbers
set.seed(100)
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'))

wordcloud(docs, scale=c(5,0.5), min.freq = 1, max.words=200, random.order=FALSE,
rot.per=0.35, colors=brewer.pal(8, 'Dark2'))

wordcloud(docs, scale=c(5,0.5), max.words=100, random.order=FALSE, rot.per=0.35,
use.r.layout=FALSE, colors=brewer.pal(8, 'Dark2'))

# Arguments of the word cloud generator function :

# words : the words to be plotted
# freq : their frequencies
# min.freq : words with frequency below min.freq will not be plotted
# max.words : maximum number of words to be plotted
# random.order : plot words in random order. If false, they will be plotted in decreasing

```

frequency

rot.per : proportion words with 90 degree rotation (vertical text)

colors : color words from least to most frequent. Use, for example, colors ="black" for single color.