Word Cloud (eng) - sequantial design

Sim

2018-05-21

## 0. Setup

library(tm) # for text mining  
library(SnowballC) # for text stemming  
library(wordcloud) # word-cloud generator   
library(RColorBrewer) # color palettes  
library(ggplot2)  
library(dplyr)  
# activate(c("tm", "SnowballC"))

## 1. Infile and prepare text (docs)

filePath <- paste0("http://www.sthda.com/sthda/RDoc/example-files/",  
 "martin-luther-king-i-have-a-dream-speech.txt")  
text <- readLines(filePath)  
docs <- Corpus(VectorSource(text))  
# inspect(docs)

## 2. Cleanup the text (docs)

# `dplyr` way  
toSpace <- content\_transformer(  
 function (x , pattern) gsub(pattern, " ", x))  
docs <- docs %>%   
 tm\_map(toSpace, "/") %>%  
 tm\_map(toSpace, "@") %>%  
 tm\_map(toSpace, "\\|")

# `base` way  
toSpace <- content\_transformer(  
 function (x , pattern) gsub(pattern, " ", x))  
docs <- tm\_map(docs, toSpace, "/")  
docs <- tm\_map(docs, toSpace, "@")  
docs <- tm\_map(docs, toSpace, "\\|")

docs <- docs %>%   
 tm\_map(content\_transformer(tolower)) %>% # Convert it to lower case  
 tm\_map(removeNumbers) %>% # Remove numbers  
 tm\_map(removeWords, stopwords("english")) %>% # Remove english common stopwords  
 tm\_map(removeWords, c("blabla1", "blabla2")) %>% # Remove your own stop word  
 tm\_map(removePunctuation) %>% # Remove punctuations   
 tm\_map(stripWhitespace) # Eliminate extra white spaces

## 3. Arriving to frequency table! (docs to freqTable)

termMat <- TermDocumentMatrix(docs)  
termTable <- as.matrix(termMat)  
head(termTable, 2)

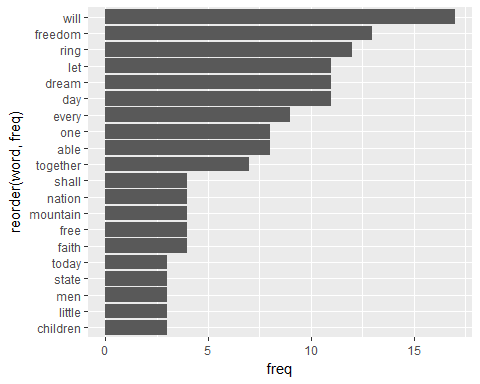
## Docs  
## Terms 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24  
## american 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## deeply 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## Docs  
## Terms 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45  
## american 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## deeply 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## Docs  
## Terms 46  
## american 0  
## deeply 0

freqTable <- data.frame(word = rownames(termTable),  
 freq = rowSums(termTable))  
freqTable$word <- rownames(freqTable)  
freqTable <- freqTable %>% arrange(desc(freq))  
head(freqTable, 2)

## word freq  
## 1 will 17  
## 2 freedom 13

## 4. Render Bar Chart (play with freqTable)

ggplot(data = head(freqTable,20)) +  
 geom\_bar(aes(x=reorder(word, freq), y=freq), stat="identity") +  
 coord\_flip()



## 5. Render Word Cloud (play with freqTable)

wordcloud(words = freqTable$word, freq = freqTable$freq,   
 min.freq = 1, max.words=200, random.order=FALSE, rot.per=0.35,   
 colors=brewer.pal(8, "Dark2"))



## 6. Reference

<http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know>