M33-function

LearningSpoonsR

2018년 5월 21일

## I. Dream

## 2. Cleanup the text (`docs`)   
toSpace <- content\_transformer(  
 function (x , pattern) gsub(pattern, " ", x))  
docs <- docs %>%   
 tm\_map(toSpace, "/") %>%  
 tm\_map(toSpace, "@") %>%  
 tm\_map(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)  
freqTable <- data.frame(word = rownames(termTable),  
 freq = rowSums(termTable))  
freqTable$word <- rownames(freqTable)  
freqTable <- freqTable %>% arrange(desc(freq))  
head(freqTable, 2)

## II. Sonaki

## 2. Cleanup the text (`docs`)   
text <- pdf\_text("../script/sonaki.pdf")  
docs <- sapply(text, extractNoun, USE.NAMES = F) # Apply extract Noun  
docs <- unlist(docs)  
docs <- Filter(function(x) {nchar(x) >= 2}, docs) # Character length >= 2  
head(docs, 10)

## 3. Arriving to frequency table! (`docs` to `freqTable`)  
freqTable <- data.frame(table(docs))  
names(freqTable) <- c("word", "freq")  
freqTable <- freqTable %>% arrange(desc(freq))  
head(freqTable)

## III. Function!

cleanDocsGenerateFreqTable <- function(docs, lang) {  
 activate(c("tm", "SnowballC", "wordcloud", "KoNLP", "pdftools"))  
 activate(c("ggplot2", "dplyr", "RColorBrewer"))  
 if (lang == "kr") {  
 docs <- unlist(docs)  
 docs <- Filter(function(x) {nchar(x) >= 2}, docs) # Character length >= 2  
 freqTable <- data.frame(table(docs))  
 names(freqTable) <- c("word", "freq")  
 freqTable <- freqTable %>% arrange(desc(freq))  
 } else { # lang == "en"  
 toSpace <- content\_transformer(  
 function (x , pattern) gsub(pattern, " ", x))  
 docs <- docs %>%   
 tm\_map(toSpace, "/") %>%  
 tm\_map(toSpace, "@") %>%  
 tm\_map(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  
 termMat <- TermDocumentMatrix(docs)  
 termTable <- as.matrix(termMat)  
 freqTable <- data.frame(word = rownames(termTable),  
 freq = rowSums(termTable))  
 freqTable$word <- rownames(freqTable)  
 freqTable <- freqTable %>% arrange(desc(freq))  
 }  
 return(freqTable)  
}