Word Cloud (kor) - sonaki

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)  
library(KoNLP) # Korean Natural Language Processing  
library(pdftools) # Extract text from pdf

## 1. Infile and prepare text (docs)

## 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)

## [1] "소나기" "황순원"   
## [3] "소년" "개울가"   
## [5] "소녀" "증손녀(曾孫女)딸이라는"  
## [7] "소녀" "개울"   
## [9] "물장난" "서울서는"

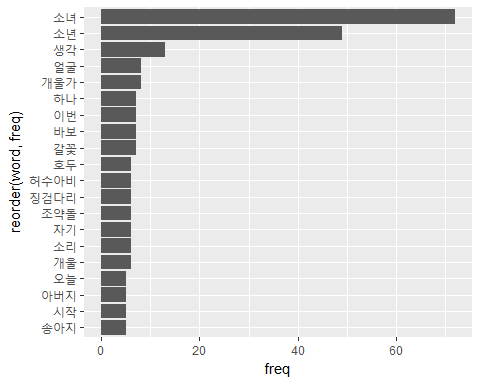
## 3. Arriving to frequency table! (docs to freqTable)

freqTable <- data.frame(table(docs))  
names(freqTable) <- c("word", "freq")  
freqTable <- freqTable %>% arrange(desc(freq))  
head(freqTable)

## word freq  
## 1 소녀 72  
## 2 소년 49  
## 3 생각 13  
## 4 개울가 8  
## 5 얼굴 8  
## 6 갈꽃 7

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

ggplot(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=100, random.order=FALSE, rot.per=0.35,   
 colors=brewer.pal(8, "Dark2"))

