##패키지 설치 install.packages(‘rJava’)

install.packages(‘memoise’)

install.packages(‘multilinguer’)

1번 방법 : install.packages(‘KoNLP’)

2번 방법 : install.packages(“<https://cran.rproject.org/src/contrib/Archive/KoNLP/KoNLP_0.80.2.tar.gz>”, repos=NULL, type=“source”)

3번 방법 : install.packages(‘remotes’) remotes::install\_github(‘haven-jeon/KoNLP’,upgrade = “never”,INSTALL\_opts=c(“–no-multiarch”))

##패키지 로드

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(multilinguer)

## Warning: package 'multilinguer' was built under R version 3.6.3

## If you have any problem, please issue here.

## https://github.com/mrchypark/multilinguer/issues

library(rJava)

## Warning: package 'rJava' was built under R version 3.6.3

library(KoNLP)

## Checking user defined dictionary!

library(stringr)

## Warning: package 'stringr' was built under R version 3.6.3

useNIADic()

## Backup was just finished!  
## 1213109 words dictionary was built.

##데이터 준비

데이터 불러오기

txt <- readLines("hiphop3.txt")

## Warning in readLines("hiphop3.txt"): 'hiphop3.txt'에서 불완전한 마지막 행이 발견  
## 되었습니다

특수문자 제거

txt <- str\_replace\_all(txt,"\\W"," ")

명사 추출하기

wordtest <- extractNoun("대한민국의 영토는 한반도와 그 부속도서로 한다.")

가사에서 명사추출

nouns <- extractNoun(txt)  
head(nouns)

## [[1]]  
## [1] "싶"  
##   
## [[2]]  
## [1] "말하" "싶"   
##   
## [[3]]  
## [1] "너희" "사진" "도"   
##   
## [[4]]  
## [1] "싶"  
##   
## [[5]]  
## [1] "야속" "한" "시" "간"   
##   
## [[6]]  
## [1] "나" "우리" "밉"

추출한 명사 list를 문자열 벡터로 변환 , 단어별 빈도표 생성

wordcount <- table(unlist(nouns))

데이터 프레임으로 변환

df\_word <- as.data.frame(wordcount , stringsAsFactors = F)

변수명 수정

df\_word <- rename(df\_word,  
 word = Var1 ,   
 freq = Freq)

두 글자 이상 단어 추출

df\_word <- filter(df\_word , freq >= 2 )

빈도수가 제일 높은 20개 추출

top20 <- df\_word %>%   
 arrange(desc(freq)) %>%   
 head(20)  
top20

## word freq  
## 1 나 374  
## 2 내 298  
## 3 것 250  
## 4 너 247  
## 5 I 201  
## 6 지 178  
## 7 y 167  
## 8 e 149  
## 9 거 114  
## 10 t 113  
## 11 한 112  
## 12 날 96  
## 13 m 95  
## 14 말 92  
## 15 you 89  
## 16 도 88  
## 17 my 86  
## 18 수 86  
## 19 위 86  
## 20 a 81

워드 클라우드 패키지 로드

library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.6.3

## Loading required package: RColorBrewer

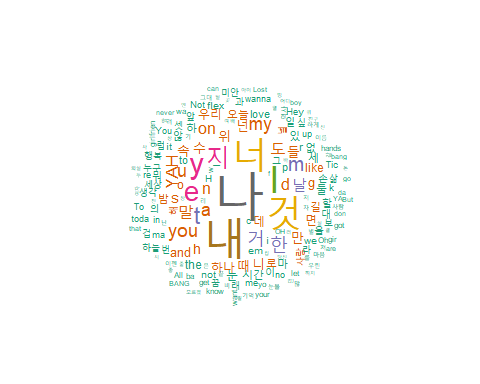
library(RColorBrewer)

단어 색상 목록 만들기

pal <- brewer.pal(8,'Dark2') # Dark2 색상 목록에서 8개 색상 추출

워드 클라우드 생성

set.seed(1234) #난수 고정  
wordcloud(words = df\_word$word, #단어   
 freq = df\_word$freq, #빈도  
 min.freq = 2, # 최소 단어 빈도  
 max.words = 200, #표현 단어 수   
 random.order = F, # 고빈도 단어 중앙 배치  
 rot.per = .1, #회전 단어 비율   
 scale = c(4 , 0.3), #단어 크기 범위  
 colors = pal #색깔 목록  
 )



단어 색상 바꾸기

pal <- brewer.pal(9,'Blues')[5:9] # Dark2 색상 목록에서 8개 색상 추출

워드 클라우드 생성

set.seed(1234) #난수 고정  
wordcloud(words = df\_word$word, #단어   
 freq = df\_word$freq, #빈도  
 min.freq = 2, # 최소 단어 빈도  
 max.words = 200, #표현 단어 수   
 random.order = F, # 고빈도 단어 중앙 배치  
 rot.per = .1, #회전 단어 비율   
 scale = c(4 , 0.3), #단어 크기 범위  
 colors = pal #색깔 목록  
)



##국정원 트윗 텍스트 마이닝

데이터 준비하기

데이터 로드

twitter <- read.csv("twitter.csv",  
 header = T,  
 stringsAsFactors = F,  
 fileEncoding = "UTF-8"  
 )

변수명 수정

twitter <- rename(twitter,  
 no = 번호,  
 id = 계정이름,  
 date = 작성일,  
 tw = 내용  
 )

특수문자 제거

twitter$tw <- str\_replace\_all(twitter$tw , '\\W' , ' ')  
head(twitter$tw)

## [1] "민주당의 ISD관련 주장이 전부 거짓으로 속속 드러나고있다 미국이 ISD를 장악하고 있다고 주장하지만 중재인 123명 가운데 미국인은 10명뿐이라고 한다 "   
## [2] "말로만 미제타도 사실은 미제환장 김정일 운구차가 링컨 컨티넬탈이던데 북한의 독재자나 우리나라 종북들이나 겉으로는 노동자 서민을 대변한다면서 고급 외제차 아이팟에 자식들 미국 유학에 환장하는 위선자들인거죠"   
## [3] "한나라당이 보수를 버린다네요 뭔가착각하는모냥인에 국민들이보수를싫어하는게 아니라뻘짓거리하는분들을싫어하는겁니다야당이진보어쩌고저쩌고한다고해서그들을조아한다고생각하면대착각"   
## [4] "FTA를 대하는 현명한 자세 사실 자유주의 경제의 가장 큰 수해자는 한국이죠 농어업분야 피해를 줄이는 정부대안을 최대한 보완하고 일자리 창출 등 실익을 최대화해 나가는게 현실적인 대처자세일듯 "   
## [5] "곽노현씨 갈수록 가관입니다 뇌물질에 아들 병역 의혹까지 도대체 아이들이 뮐 보고 배우겠습니까 이래도 자리 연연하시겠습니까 "   
## [6] "과거 집권시 한미FTA를 적극 추진하던 세력이 이제 집권하면 폐기하겠다고 주장합니다 어이없어 말도 안 나오네요 표만 얻을 수 있다면 국가 안보나 경제가 어떻게 되든 상관없다는 무책임한 행태들 우리 정치의 후진성을 드러내는 거죠 "

트윗에서 명사 추출

nouns <- extractNoun(twitter$tw)

추출한 명사 list를 문자열 벡터로 변환 , 단어별 빈도표 생성

wordcount <- table(unlist(nouns))

데이터 프레임으로 변환

df\_word <- as.data.frame(wordcount , stringsAsFactors = F)

변수명 수정

df\_word <- rename(df\_word,  
 word = Var1,  
 freq = Freq  
 )

두 글자 이상 단어만 추출

df\_word <- filter(df\_word , nchar(word) >= 2)

상위 20개 추출

top20 <- df\_word %>%   
 arrange(desc(freq)) %>%   
 head(20)  
top20

## word freq  
## 1 종북 2431  
## 2 북한 2216  
## 3 세력 1162  
## 4 좌파 829  
## 5 대한민국 804  
## 6 우리 780  
## 7 들이 566  
## 8 국민 550  
## 9 친북 430  
## 10 단체 394  
## 11 김정일 342  
## 12 진보 335  
## 13 대선 329  
## 14 천안함 319  
## 15 사회 307  
## 16 정부 286  
## 17 전교조 278  
## 18 주장 269  
## 19 정권 265  
## 20 연평도 262

단어 빈도 막대그래프 만들기

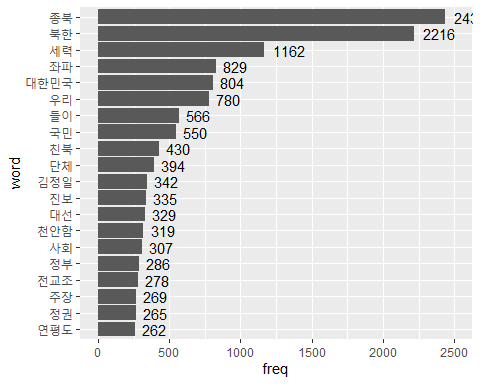
패키지 로드

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

order <- arrange(top20 , freq)$word #빈도 순서 변수 생성

ggplot(data = top20 , aes(x = word , y = freq)) +   
 ylim(0 , 2500) +   
 geom\_col() +  
 coord\_flip() +  
 scale\_x\_discrete(limit = order) + #빈도 순서 변수 기준 막대 정렬  
 geom\_text(aes(label = freq), hjust = -0.3) #빈도 표시



워드 클라우드 만들기

pal <- brewer.pal(8,'Dark2')  
wordcloud(words = df\_word$word,  
 freq = df\_word$freq,  
 min.freq = 10,  
 max.words = 200,  
 random.order = F,  
 rot.per = .1,  
 scale = c(8, 0.2),  
 colors = pal  
 )



색깔바꾸기

pal <- brewer.pal(9,'Blues')[5:9]  
set.seed(1234)  
wordcloud(words = df\_word$word,  
 freq = df\_word$freq,  
 min.freq = 10,  
 max.words = 200,  
 random.order = F,  
 rot.per = .1,  
 scale = c(8, 0.2),  
 colors = pal  
)

## Warning in wordcloud(words = df\_word$word, freq = df\_word$freq, min.freq = 10, :  
## 북한 could not be fit on page. It will not be plotted.

