library(data.table)

data<- fread("1F601\_2.csv",encoding='UTF-8')

colnames(data) <- c("V1","V2")

delete <- function(data) {

data <- data[ grep("카카오톡", data$V2, invert = TRUE) , ]

data <- data[ grep("카톡", data$V2, invert = TRUE) , ]

data <- data[ grep("상담", data$V2, invert = TRUE) , ]

data <- data[ grep("주소", data$V2, invert = TRUE) , ]

data <- data[ grep("예약", data$V2, invert = TRUE) , ]

data <- data[ grep("문의", data$V2, invert = TRUE) , ]

data <- data[ grep("환불", data$V2, invert = TRUE) , ]

data <- data[ grep("링크", data$V2, invert = TRUE) , ]

data <- data[ grep("추천", data$V2, invert = TRUE) , ]

data <- data[ grep("노리터", data$V2, invert = TRUE) , ]

data <- data[ grep("ㅂㅐ팅", data$V2, invert = TRUE) , ]

data <- data[ grep("디엠", data$V2, invert = TRUE) , ]

data <- data[ grep("양도", data$V2, invert = TRUE) , ]

data <- data[ grep("뎀주세요", data$V2, invert = TRUE) , ]

data <- data[ grep("굿즈", data$V2, invert = TRUE) , ]

data <- data[ grep("첫콘", data$V2, invert = TRUE) , ]

data <- data[ grep("막콘", data$V2, invert = TRUE) , ]

data <- data[ grep("앙스타", data$V2, invert = TRUE) , ]

data <- data[ grep("힝구",data$V2, invert = TRUE) , ]

data <- data[ grep("콘서트", data$V2, invert = TRUE) , ]

data <- data[ grep("입금", data$V2, invert = TRUE) , ]

data <- data[ grep("탄님", data$V2, invert = TRUE) , ]

data <- data[ grep("밍탄", data$V2, invert = TRUE) , ]

data <- data[ grep("후탄", data$V2, invert = TRUE) , ]

data <- data[ grep("티탄", data$V2, invert = TRUE) , ]

data <- data[ grep("반모", data$V2, invert = TRUE) , ]

data <- data[ grep("찔러봐주세", data$V2, invert = TRUE) , ]

data <- data[ grep("멜론", data$V2, invert = TRUE) , ]

data <- data[ grep("나눔", data$V2, invert = TRUE) , ]

data <- data[ grep("전력주제", data$V2, invert = TRUE) , ]

data <- data[ grep("본무대", data$V2, invert = TRUE) , ]

data <- data[ grep("뮤직뱅크", data$V2, invert = TRUE) , ]

data <- data[ grep("라빗츠", data$V2, invert = TRUE) , ]

data <- data[ grep("본계", data$V2, invert = TRUE) , ]

data <- data[ grep("본계정", data$V2, invert = TRUE) , ]

data <- data[ grep("홈마", data$V2, invert = TRUE) , ]

data <- data[ grep("연성", data$V2, invert = TRUE) , ]

data <- data[ grep("라잇이",data$V2, invert = TRUE) , ]

data <- data[ grep("직거래", data$V2, invert = TRUE) , ]

data <- data[ grep("쿨거", data$V2, invert = TRUE) , ]

data <- data[ grep("찔러봐", data$V2, invert = TRUE) , ]

data <- data[ grep("아무구역", data$V2, invert = TRUE) , ]

data <- data[ grep("키르아", data$V2, invert = TRUE) , ]

data <- data[ grep("투표해", data$V2, invert = TRUE) , ]

data <- data[ grep("선팔", data$V2, invert = TRUE) , ]

data <- data[ grep("종주님", data$V2, invert = TRUE) , ]

data <- data[ grep("샐리", data$V2, invert = TRUE) , ]

data <- data[ grep("네코마", data$V2, invert = TRUE) , ]

data <- data[ grep("우티교환", data$V2, invert = TRUE) , ]

data <- data[ grep("양도표", data$V2, invert = TRUE) , ]

data <- data[ grep("초동", data$V2, invert = TRUE) , ]

data <- data[ grep("추금", data$V2, invert = TRUE) , ]

data <- data[ grep("테오", data$V2, invert = TRUE) , ]

}

tmp <- delete(data[,2])

df = as.data.frame(x = tmp)

rmCelb<-function(tweetData){ #tweetData <- type Vector

user\_name<-regmatches(tweetData,regexpr("@[[:graph:]]\*",tweetData)) #user\_name 추출

freq\_name<-data.frame(table(user\_name)) #빈도표 작성

freq\_name<-freq\_name[c(order(-freq\_name$Freq)),] #빈도순으로 정렬

}

# rowNum : 빈도수가 100 이상인 user\_name이 포함된 모든 행번호 출력하는 함수

freqName1 <- rmCelb(df$V2) # rowNum 함수의 input data

rowNum <- function(freqNameData,df){ #freqNameData : input 데이터는 rmCelb함수 사용한 결과

rowNumber <- list() #list 생성

n <- nrow(freqNameData)

n <- trunc(n\*0.002)

rmName <- as.vector(freqNameData[1:n,]$user\_name) #빈도수가 100 이상인 username만 추출

for(i in 1:length(rmName)){ # 해당 username이 포함된 모든 행번호 rowNumber 리스트에 저장

rowNumber[[i]] <- grep(rmName[i],df$V2)

print(i)

}

unlist(rowNumber) # list to vector

}

newdf <- df$V2[-rowNum(freqName1,df)]

##########

remove <- function(x) {

x <- gsub("[<].+[>]", "", x)

x<- gsub("@[[:graph:]]\*", "", x)

x<- gsub("#[[:graph:]]\*","",x)

x <- gsub("http://[[:graph:]]\*", "", x)

x<- gsub("https://[[:graph:]]\*", "", x)

x <- gsub("pic.twitter.com/[[:graph:]]\*", "", x)

x<- gsub("RT","",x)

x<- gsub("\n"," ",x)

x <- gsub("\_"," ",x)

x <- gsub("[[:punct:][:blank:]]+", " ",x)

x <- gsub('[0-9]+', '', x)

x <- gsub('[a-zA-Z]','',x)

x <- gsub("\r"," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

x <- gsub(" "," ",x)

}

newnewdf <- remove(newdf)

write.csv(newnewdf,"1F601\_1\_1차.csv",fileEncoding='UTF-8')