options(warn=-1)  
#install.packages("readxl")  
#install.packages("stringi")  
#install.packages("tidyverse")  
#install.packages("tokenizers")  
  
library(stringi)  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.3  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 1.0.0 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tokenizers)  
  
# Loading  
library("readxl")  
# xls files  
x = read\_excel("W32 - Frequency.xlsx")

x = tolower(x) ## chuyen data ve lower case

x = stri\_trans\_general(x, "Latin-ASCII") #xoa dau tieng Viet

x = gsub(" ", "", x) #gop ten va ho lai voi nhau

x = str\_replace\_all(x, "[[:punct:]]", " ") #remove tat ca cac dau va ky tu dac biet  
x = str\_replace\_all(x, "[[:digit:]]+", " ") #remove tat ca cac so

x = gsub("buitiendung", " tiendung ",x)   
  
x = gsub("thumonbuitiendung", " tiendung ",x)   
#x = gsub("buibuitiendung", " buitiendung ",x)  
  
x = gsub("cr", " CR7 ",x)   
x = gsub("ronaldo", " CR7 ",x)   
x = gsub("ronando", " CR7 ",x)   
x = gsub("ronado", " CR7 ",x)   
x = gsub("rolnado", " CR7 ",x)  
x = gsub("Cầu thủ ronaldo", " CR7 ",x)  
x = gsub("cr7istianocr7", " CR7 ",x)  
  
  
x = gsub("baothy", " baothi ",x)   
x = gsub("sontungmpt", " SONTUNGMTP ",x)  
x = gsub("sontungmp", " SONTUNGMTP ",x)   
x = gsub("nguyenthanhtung", " SONTUNGMTP ",x)   
x = gsub("st", " SONTUNGMTP ",x)   
  
x = gsub("huonggiangidol", " huonggiang ",x)   
x = gsub("casihuonggiang", " huonggiang ",x)

x = tokenize\_words(x) #tach tung ky tu ra   
x = table(unlist(x)) #chuyen du lieu ve factor  
final\_list = sort(x, decreasing = TRUE) #sort du lieu theo thu tu giam dan  
  
head(final\_list,10)

##   
## cr7 tiendung baoanh huonggiang   
## 169 79 68 57   
## hoailinh ninhduonglanngoc quanghai cauthu   
## 17 13 13 12   
## sontungmtp i   
## 11 10

summary(final\_list) ##tom tat du lieu

## Number of cases in table: 605   
## Number of factors: 1