|  |
| --- |
| **데이터사이언스기초 6주차 실습과제** |
| **실습 #2 보고서** |

|  |  |
| --- | --- |
| **이름** | 김영식 |
| **학번** | 20175119 |
| **소속**  **학과/대학** | 빅데이터학과 |

## <주의사항>

* 개별 과제 입니다. (팀으로 진행하는 과제가 아니며, 모든 학생이 보고서를 제출해야 함)
* 각각의 문제 바로 아래에 답을 작성 후 제출해 주세요.
  + 소스코드/스크립트 등을 작성 한 경우, 해당 파일의 이름도 적어주세요.
* SMARTLEAD 제출 데드라인: 2022. 04. 04(월요일) ~ 4.10(일요일) 23:59 까지
  + 부정행위 적발 시, 원본(보여준 사람)과 복사본(베낀 사람) 모두 0점 처리함
  + 예외 없음
* SMARTLEAD에 아래의 파일을 제출 해 주세요
  + 보고서 형식으로 제출
  + 보고서 파일명에 이름과 학번을 입력 해 주세요.

## <실습 과제>

|  |
| --- |
| [Q 1] switch(조건문), for/while/repeat(반복문) 사용  <사용되는 데이터 : News\_Final.csv>  News\_Final.csv의 데이터를 이용하여 조건식을 적용하려고 한다. 이 News\_Final.csv 에는 총 4,863 개의 행과 11개의 열로 구성되어 있다. 그중 “Topic”열을 이용하여,각 Topic의 Facebook, GooglePlus, LinkedIn의 인기도의 최대값을 구하고자 한다.  **ex) economy의 Facebook, GooglePlus, LinkedIn의 인기도의 최대값을 구하시오**  **<진행사항>**   1. **Read.csv를 사용하고, header = T, stringsAsFactors = F로 줄 것** 2. **반드시 함수를 이용하여 진행하되 매개 변수의 수는 자유롭게 지정한다.** 3. **데이터에 있는 모든 Topic에 대하여 모두 결과를 출력한다.** 4. **함수의 이름은 임의로 설정한다.** 5. **조건문을 반드시 쓰고, for 또는 while을 사용할 것(apply 사용하지 말 것)** 6. **자신이 작성한 사용자 지정함수에대한 간단한 설명을 포함한다.**   **<데이터 컬럼 설명>**    **<출력 예시>**   1. **사용자 지정함수 코드**      1. **출력 결과 예시** |

답변:

1. **사용자 함수 코드 및 출력 결과(70 점)**

**library(dplyr)**

**library(tibble)**

**user\_topic <- function(dt){**

**all\_topic <- unique(dt$Topic)**

**result <- tibble()**

**for (t in all\_topic) {**

**max\_face <- 0**

**max\_google <- 0**

**max\_linked <- 0**

**dt\_some <- filter(dt,Topic == t)**

**for (i in 1:nrow(dt\_some)) {**

**if (dt\_some$Facebook[i] > max\_face) {**

**max\_face <- dt\_some$Facebook[i]**

**}**

**if (dt\_some$GooglePlus[i] > max\_google) {**

**max\_google <- dt\_some$GooglePlus[i]**

**}**

**if (dt\_some$LinkedIn[i] > max\_linked) {**

**max\_linked <- dt\_some$LinkedIn[i]**

**}**

**}**

**result <- rbind(result, tibble(Facebook = max\_face, GooglePlus = max\_google, LinkedIn = max\_linked))**

**}**

**return(result)**

**}**

**data <- read.csv("News\_Final.csv", header = T, stringsAsFactors = F) %>% as.tibble()**

**user\_topic(data)**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

1. **사용자 함수 코드 설명 (30 점)**

user\_topic <- function(dt)

함수를 생성 하는 부분이다.

그 함수 에는

all\_topic <- unique(dt$Topic)

result <- tibble()

Topic에 있는 두가지 값을 all\_topic에 저장하고, result에는 tibble구조를 사용한다.

for (t in all\_topic) {

여기서는 주제가 총 두개 이므로 두 번 반복한다.

max\_face <- 0

max\_google <- 0

max\_linked <- 0

각 주제의 최대값을 구하기 위해 미리 선언하고 초기화한다.

dt\_some <- filter(dt,Topic == t)

filter함수를 사용하여 원래 존재하던 데이터 들을 추출하여 특정 주제를 가지고 있는 데이터들로 구성한다.

for (i in 1:nrow(dt\_some)) {

if (dt\_some$Facebook[i] > max\_face) {

max\_face <- dt\_some$Facebook[i]

}

if (dt\_some$GooglePlus[i] > max\_google) {

max\_google <- dt\_some$GooglePlus[i]

}

if (dt\_some$LinkedIn[i] > max\_linked) {

max\_linked <- dt\_some$LinkedIn[i]

}

}

for문을 이용해서 각각의 최대값을 구한다.

result <- rbind(result, tibble(Facebook = max\_face, GooglePlus = max\_google, LinkedIn = max\_linked))

rbind와 tibble구조를 사용해서 result에 최대값들을 저장한다.

return(result)

그리고 리턴해준다.

data <- read.csv("News\_Final.csv", header = T, stringsAsFactors = F) %>% as.tibble()

user\_topic(data)

read.csv함수를 사용하여 불러온 데이터를 사용하여 함수를 호출한다.

## <실습 과제>

|  |
| --- |
| [Q 2] **Boxplot 그리기!!**  **<사용되는 데이터 - News\_Final.csv>**  **News\_Final.csv 데이터의 Boxplot을 그리려고 한다. 문제에서 사용되는 열(Column)은 SentimentTitle, SentimentHeadline, Facebook, GooglePlus, LinkedIn로 총 5개이다.**  **<참고사항>**   1. **data.table::fread를 사용하여 News\_Final.csv를 불러온다.**   **(Window환경에서는 encoding = “UTF-8”을 줄 것)**   1. **Facebook, GooglePlus, LinkedIn 열의 값 중 -1을 가지는 값은 결측값이며, boxplot를 그릴 때 결측값은 제거한 후 진행하여야 한다.** 2. **Boxplot 그래프는 SentimentTitle, SentimentHeadline만 포함하는 그래프 하나와 Facebook, GooglePlus, LinkedIn만 포함하는 그래프로 나누어서 2개를 각각 그리도록 한다** |

답변:

1. **코드 첨부(중요하다 생각하는 부분에 주석 달 것)(70 점)**

**library(data.table)**

**#파일을 불러오고 인코딩은 UTF-8을 사용한다.**

**data\_fread <- data.table::fread('News\_Final.csv',encoding = 'UTF-8')**

**#불러온 데이터에서 특정 값만 뽑는다.**

**x <- data\_fread$SentimentTitle**

**y <- data\_fread$SentimentHeadline**

**#boxplot를 사용하여 그래프를 그린다.**

**boxplot(x,y,names=c("Sentimenttitle","SentimentHeadline"))**

**#불러온 데이터에서 특정 값만 뽑는다.**

**a <- data\_fread$Facebook**

**b <- data\_fread$GooglePlus**

**c <- data\_fread$LinkedIn**

**#-1을 결측값이라고 생각하고 -1을 가지고 있는 열을 제거한다.**

**a <- a[!(a==-1)]**

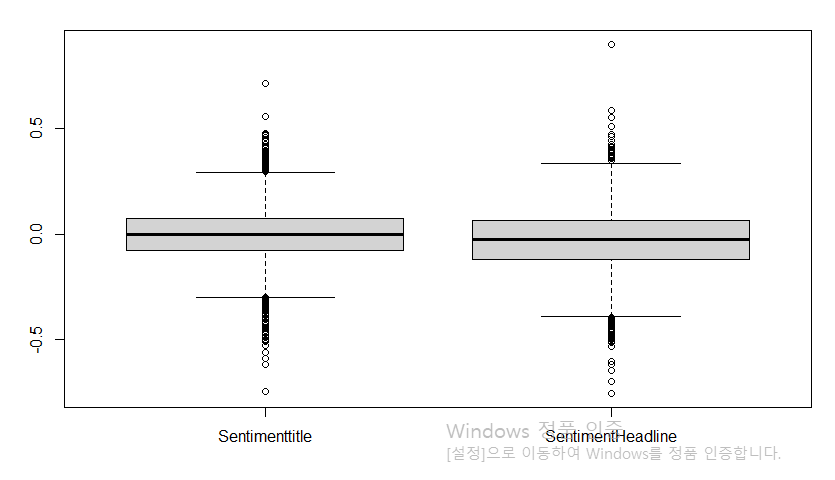
**b <- b[!(b==-1)]**

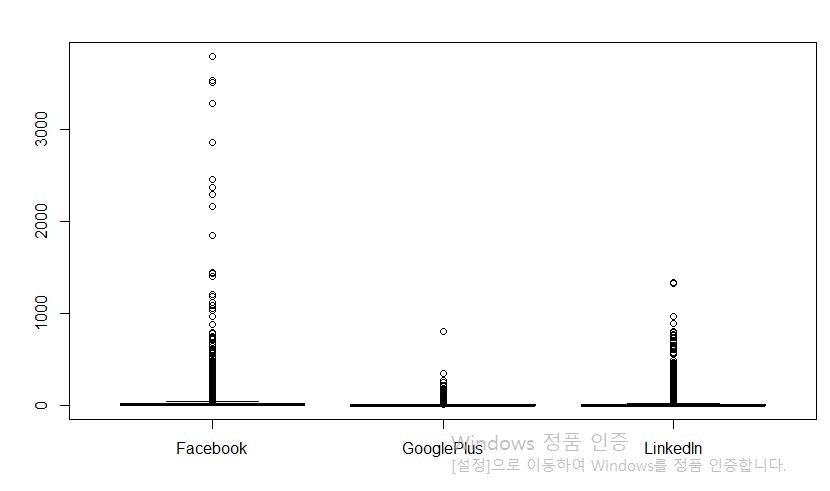
**c <- c[!(c==-1)]**

**#boxplot를 사용하여 그래프를 그린다.**

**boxplot(a,b,c,names=c("Facebook","GooglePlus","Linkedln"))**

1. **News\_Final.csv 의 boxplot 결과 첨부(30 점)**

****

****