Source code

##url 페이지 url지정

url <-"https://www.rottentomatoes.com/top/bestofrt/?year=2019"

request\_url <- url

page\_html <- read\_html(request\_url, encoding="UTF-8")

## html\_nodes()함수 이용하여 url경로와 xpath이용하여 불로올 데이터 경로 설정

nodes = html\_nodes(page\_html, xpath='//\*[@id="top\_movies\_main"]/div/table/tr/td/a')

nodes

## 토마토 Top데이터가 1-100개로 나타나있고 user Rating, 장르 경로를 설정하려면 top100 데이터 a Tag link를 타고 들어가야 하기때문에 html\_attr(nodes,"href") 를 통해 link접속

type(html\_text(nodes))

titleLink <- html\_attr(nodes,"href")

titleLink

# 100개의 데이터를 담기위해 데이터 담을 변수의 길이를 설정

ratingData = vector(mode="character", length = length(titleLink))

titleData = vector(mode="character", length = length(titleLink))

genreData = vector(mode="character", length = length(titleLink))

#반복문으로 각 index데이터를 저장

for(i in 1:length(titleLink)){

# 100개 리스트 페지 경로 =>https://www.rottentomatoes.com/m/parasite\_2019 로 되어있어 titleLink[i]를 통해 각 순위별 데이터값 index를 활용하여 호출

enterLink = str\_c("https://www.rottentomatoes.com",titleLink[i])

moviePage\_html <- read\_html(enterLink, encoding = "UTF-8")

#xpath로 받아온 데이터 값을 titleData[]에 담는다

titleNode <- html\_nodes(moviePage\_html, xpath='//\*[@id="topSection"]/div[2]/div[1]/h1')

titleData[i] <- html\_text(titleNode)

ratingnNode <-html\_nodes(moviePage\_html, xpath = '//\*[@id="topSection"]/div[2]/div[1]/section/section/div[2]/h2/a/span[2]')

ratingData[i] =ifelse(length(ratingnNode)!=0,html\_text(ratingnNode) ,"0")

ratingData[i] = str\_remove\_all(ratingData[i], "[\n\\s]")

genreNode <-html\_nodes(moviePage\_html,xpath = '//\*[@class="meta-value"]')[2]

genreData[i] <-html\_text(genreNode)

genreData[i] <- str\_trim(genreData[i])

}

View(titleData)

View(ratingData)

View(genreData)

df = data.frame(Title=c(titleData), Raiting=c(ratingData), Genre=c(genreData))

View(df)

install.packages("writexl")

library(writexl)

writexl::write\_xlsx(df, "C:/R\_Workspce/R\_Lecture/data.xlsx")

구현화면

