crawling in R

JIIN

2019년 2월 12일

```
#=========================#
# 20190211 R crawling 과제 #
#=========================#

# daum 사이트에서 영화 정보를 크롤링하여 데이터프레임으로 만드시오.
# python, R 아무거나 사용 가능(영화는 극한직업)


daum_movie_info <- function(moviecode){
  movie_url <- paste0("https://movie.daum.net/moviedb/main?movieId=",
moviecode)
  movie_detail <- read_html(movie_url)
# 영화 제목
 Title_kor <-
    movie_detail %>%
    html_nodes(xpath='//*[@id="mArticle"]/div[1]/a/h2') %>%
    html_text()
# 영문 제목
 Title_eng <-
    movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
div[1]/span') %>%
    html_text()
# 영화 장르
Genre <-
  movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[1]/dd[1]') %>%
  html_text()
# 국가
Country <- movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[1]/dd[2]') %>%
  html_text() %>%
  str_replace_all(pattern = "\n|\t", replace="")
# 영화 개봉일
```

```r
Date <- movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[1]/dd[3]') %>%
  html_text() %>%
  str_replace_all(pattern="\n|\t", replace="") %>%
  str_sub(1,10) # str_sub 10번째 문자열까지 뽑아내기
# 러닝 타임
RunningTime <- movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[1]/dd[4]') %>%
  html_text() %>%
  str_sub(1,4)
# 감독
Director <-
  movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[1]/dd[5]/a') %>%
  html_text()
# 주연
MainActors<-
  movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[1]/dd[6]') %>%
  html_text() %>%
  str_replace_all(pattern = "\n|\t", replacement="") %>%
  substring(5)    # 문자열 추출 substring(문자열, 시작, 끝)
# 순위
Rank <-
  movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
dl[2]/dd[1]') %>%
  html_text() %>%
  substring(4)    # substring(시작점)
# 누적관객
# html_nodes(xpath='') 방법으로는 끌어올 수 없었다..
# ".class명" 혹은 "#id명" 도 안됨
# movie_detail %>%
#   html_nodes(xpath='//*[@id="totalAudience"]')%>%
#   html_text()
# 평점
Score <-
  movie_detail %>%
```

```
html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[2]/div/
div[1]/a/em') %>%
  html_text()
Synopsis <-
  movie_detail %>%

html_nodes(xpath='//*[@id="mArticle"]/div[2]/div[2]/div[1]/div[1]/div[3]/p')
%>%
  html_text() %>%
  str_replace_all(pattern="\r|\n|\t", replace="")
daum_movie_info_df <- data.frame(
  Title_kor = Title_kor,
  Title_eng = Title_eng,
  Genre = Genre,
  Country = Country,
  Date = Date,
  RunningTime = RunningTime,
  Director = Director,
  MainActors = MainActors,
  Rank = Rank,
  Score = Score,
  Synopsis = Synopsis
)
 return(daum_movie_info_df)
}

Extreme_Job <- daum_movie_info(119859)
print(Extreme_Job)
```

```
##   Title_kor   Title_eng  Genre Country       Date RunningTime Director
## 1  극한직업 Extreme Job 코미디     한국 2019.01.23       111분    이병헌
##                                MainActors Rank Score
## 1 류승룡, 이하늬, 진선규, 이동휘, 공명  1위   7.8
##
Synopsis
## 1 낮에는 치킨장사! 밤에는 잠복근무!지금까지 이런 수사는 없었다!불철주야 달리고
구르지만 실적은 바닥, 급기야 해체 위기를 맞는 마약반!더 이상 물러설 곳이 없는 팀의
맏형 고반장은 국제 범죄조직의 국내 마약 밀반입 정황을 포착하고 장형사, 마형사,
영호, 재훈까지 4명의 팀원들과 함께 잠복 수사에 나선다.마약반은 24시간 감시를 위해
범죄조직의 아지트 앞 치킨집을 인수해 위장 창업을 하게 되고, 뜻밖의 절대미각을 지닌
마형사의 숨은 재능으로 치킨집은 일약 맛집으로 입소문이 나기 시작한다.수사는 뒷전,
치킨장사로 눈코 뜰 새 없이 바빠진 마약반에게 어느 날 절호의 기회가
찾아오는데…범인을 잡을 것인가, 닭을 잡을 것인가!2019년 새해, 출동!
```