X101

JDW

2019-12-19

Contents

Chapter 1


PRODUCE 101
잠시 후 저녁 8시 <프로듀스 X 101> 최종회
당신의 소년에게 투표하라 !
최종 데뷔 멤버는?
파이널 생방송 데뷔평가 본/방/사/수

X 101 . k-pop
. X 101 101 ( )
. 11
.
.

.

1.
2.
3.

, d
.
, 1 ,
.
.
.

Chapter 2

.

.

X101 .

879400	진짜거짓말이노	ㅇㅇ(223.39) [SKT]	06.01	9	0
879399	김요한 한민관 닮음	ㅇㅇ(223.38) [SKT]	06.01	14	0
879398	이한결 이세진 강민희 어디갔어	ㅇㅇ(59.26) [KT]	06.01	21	0
879396	아 제발 내새끼 분량좀 줘라ㅜㅜ	ㅇㅇ(223.62) [SKT]	06.01	10	0
879395	해남이 시발ㅋㅋㅋㅋㅋㅋㅋㅋ개귀여어	ㅇㅇ(182.215) [LGU+]	06.01	15	0
879394	요한 뽑은 진우 이제 픽에서 제외	ㅇㅇ(39.7) [KT모바일]	06.01	38	3
879393	아 이진우 존나웃겨 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ	물결이	06.01	36	0
879392	이진우 원픽될것 같다 [1]	ㅇㅇ(118.45) [KT]	06.01	24	2
879390	이진혁 씹인싸야 존나 웃김 [1]	ㅇㅇ(223.62) [SKT]	06.01	127	2
879389	강약약강 본성들보소	ㅇㅇ(223.33) [SKT]	06.01	14	0
879388	아 진우 ㅈㄴ 커영 ㅓ	ㅇㅇ(114.204) [SKB]	06.01	12	0

Figure 2.1: x101

4 . 06/01 07/19 .

R rvest .

.

2.1 code

```
#
setwd('C:/Users/JDW/Desktop/PROJECT/PRODUCEX/CRAWLING DATA')

library(rvest)
```

```
library(dplyr)
library(lubridate)
library(stringr)
```

.

```
basic_url <- 'https://gall.dcinside.com/board/lists/?id=producex&page='

#    ----

urls <- NULL
for(x in 0:599){
    urls[x + 1] <- paste0(basic_url, x + 1)
}
```

. GET POST . GET URL . ,
URL ( ) . URL
.

R rvest .

for .

.

```
## [1] "https://gall.dcinside.com/board/lists/?id=producex&page=1"
## [2] "https://gall.dcinside.com/board/lists/?id=producex&page=2"
## [3] "https://gall.dcinside.com/board/lists/?id=producex&page=3"
## [4] "https://gall.dcinside.com/board/lists/?id=producex&page=4"
## [5] "https://gall.dcinside.com/board/lists/?id=producex&page=5"
## [6] "https://gall.dcinside.com/board/lists/?id=producex&page=6"
```

```
## [1] "https://gall.dcinside.com/board/lists/?id=producex&page=595"
## [2] "https://gall.dcinside.com/board/lists/?id=producex&page=596"
## [3] "https://gall.dcinside.com/board/lists/?id=producex&page=597"
## [4] "https://gall.dcinside.com/board/lists/?id=producex&page=598"
## [5] "https://gall.dcinside.com/board/lists/?id=producex&page=599"
## [6] "https://gall.dcinside.com/board/lists/?id=producex&page=600"
```

for 1 600 600 .

. .

```
#    ----

html1 <- NULL
p_title <- NULL
p_writer <- NULL
p_writer_id <- NULL
p_time <- NULL
p_count <- NULL
p_recommend <- NULL
dc <- NULL
```

for .

```
for(url in urls){
            html1 <- read_html(url)
            p_title <- c(p_title, html1 %>%
                             html_nodes('.ub-content.us-post') %>%
                             html_nodes('.gall_tit.ub-word') %>%
                             html_node('a') %>%
                             html_text())
            p_writer <- c(p_writer, html1 %>%
                              html_nodes('.ub-content.us-post') %>%
                              html_nodes('.gall_writer.ub-writer') %>%
                              html_attr('data-nick'))
            p_writer_id <- c(p_writer_id, html1 %>%
                                 html_nodes('.ub-content.us-post') %>%
                                 html_nodes('.gall_writer.ub-writer') %>%
                                 html_attr('data-uid'))
            p_time <- c(p_time, html1 %>%
                            html_nodes('.ub-content.us-post') %>%
                            html_nodes('.gall_date') %>%
                            html_attr('title'))
            p_count <- c(p_count, html1 %>%
                             html_nodes('.ub-content.us-post') %>%
                             html_nodes('.gall_count') %>%
                             html_text())
            p_recommend <- c(p_recommend, html1 %>%
                                 html_nodes('.ub-content.us-post') %>%
                                 html_nodes('.gall_recommend') %>%
                                 html_text())

}
```

for R , urls . for ,
(p_title, p_time ..) . for .

```
#
dc <- data.frame(p_title, p_writer, p_writer_id, p_time, p_count, p_recommend)

#    ----
dc$p_time <- ymd_hms(dc$p_time)
dc <- dc[day(dc$p_time) == day(today()-1), ]
```

for 'dc' . lubridate ymd_hms() p_time .

taskscheduleR ,
dc <- dc[day(dc$p_time) == day(today()-1),] .
dc$p_time today() - 1 dc .

V1	p_title	p_writer	p_writer_id	p_time	p_count	p_recommend
42448	???			2019-06-01 00:00:02	5	0
42449	4			2019-06-01 00:00:01	11	1
42450				2019-06-01 00:00:01	4	0
42451			yuri1228	2019-06-01 00:00:00	40	0
42452			adkljf	2019-06-01 00:00:00	36	0
42453				2019-06-01 00:00:00	27	1

```
#    ----
write.csv(dc, paste0(ymd(today()-1), ".csv"))
```

csv .

Chapter 3

3.1

. . .
.
. .

R .

```
# ----      ----

library(dplyr)
library(ggplot2)
library(lubridate)
library(stringr)
library(gtools)

# ---- for              ----

dir <- 'C:/Users/JDW/Desktop/PROJECT/PRODUCEX/CRAWLING DATA/data'
filelist <- list.files(dir)
filelist <- mixedsort(filelist, decreasing = T)
pxdata <- data.frame()
temp <- NULL

for(file in filelist){
    if(str_sub(file, -3, -1) == 'csv'){
        temp <- fread(paste(dir, file, sep = '/'), header = T, stringsAsFactors = F)
        temp <- temp %>% select(p_title, p_time)
        pxdata <- rbind(pxdata, temp)
```

```
        rm(temp)
    }else{
        next
    }

}
```

. dir . r list.files()
, filelist . gtools::mixedsort() r sort()
. list.files(dir) filelist .
.

3.1.1 mixedsort()

```
t1 <- c(1:20) %>% as.character()
t2 <- c(1:20) %>% as.numeric()
```

```
str(t1)
```

```
##  chr [1:20] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" ...
```

```
str(t2)
```

```
##  num [1:20] 1 2 3 4 5 6 7 8 9 10 ...
```

t1 t2 1~20 . t1 , t2 . sort()
.

```
sort(t1)
```

```
##  [1] "1"  "10" "11" "12" "13" "14" "15" "16" "17" "18" "19" "2"  "20" "3"  "4"
## [16] "5"  "6"  "7"  "8"  "9"
```

```
sort(t2)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

t2 t1 1 10 11 ... . gtools::mixedsort()
.

```
library(gtools)

mixedsort(t1)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
## [16] "16" "17" "18" "19" "20"
```

```
mixedsort(t2)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

t2 . t1 . mixedsort()
.

3.2

p_title , p_time .

(1) - .

```
pxdata <- pxdata %>% select(p_title, p_time)
```

p_time ‘ ’, ‘ ’ .

```
pxdata <- pxdata %>% mutate(p_ymd = as.Date(p_time),
                            p_wday = lubridate::wday(as.Date(p_time), label = T),
                            )
```

. 11 12 30
. , .

05 31 X101 5 06 01 5 (p_weeknum = 5) .

```
pxdata$p_weeknum <- week(as.Date(pxdata$p_time) - 5) - 17
```

p_title	p_time	p_ymd	p_wday	p_weeknum
?	2019-06-01 23:59:50	2019-06-01		5
	2019-06-01 23:59:48	2019-06-01		5
	2019-06-01 23:59:43	2019-06-01		5
	2019-06-01 23:59:42	2019-06-01		5
	2019-06-01 23:59:31	2019-06-01		5
	2019-06-01 23:59:25	2019-06-01		5
1 <U+270B>	2019-06-01 23:59:05	2019-06-01		5
??	2019-06-01 23:58:54	2019-06-01		5
	2019-06-01 23:58:52	2019-06-01		5
?	2019-06-01 23:58:51	2019-06-01		5

, (`p_ymd`, `p_wday`, `p_weeknum`) .
.

Chapter 4

EDA

EDA .

EDA , .

.

4.1 dataset

```
DT::datatable(head(pxdata, 10))
```

Show 10 entries Search:

	p_title	p_time	p_ymd	p_wday	p_weeknum
1	비율감 연습생이 있다?	2019-06-01 23:59:50	2019-06-01	토	4
2	디모데 떨어진거 너무 아깝다	2019-06-01 23:59:48	2019-06-01	토	4
3	픽 남는데 추천해 줘	2019-06-01 23:59:43	2019-06-01	토	4
4	투표할때	2019-06-01 23:59:42	2019-06-01	토	4
5	남자 새끼들이 병신처럼 쳐울거나	2019-06-01 23:59:31	2019-06-01	토	4
6	근데 김요한은 운 하나는 존나 좋은거같아서 부럽긔	2019-06-01 23:59:25	2019-06-01	토	4
7	김성현 꼰대기 1도 없고 하찮다<U+270B>	2019-06-01 23:59:05	2019-06-01	토	4
8	랜디 롤 찍는 거냐?? 외로워도 슬퍼도 나는 안울어	2019-06-01 23:58:54	2019-06-01	토	4
9	남도현 옹짤 프추좀 눌러주시긔	2019-06-01 23:58:52	2019-06-01	토	4
10	어제 상위권애들 왤케 거만했냐? 좆같게	2019-06-01 23:58:51	2019-06-01	토	4

Showing 1 to 10 of 10 entries Previous 1 Next