

[R] R을 활용한 크롤링 - 로또 1등 당첨 배출점 크롤링 하기

노트북: [TIL-MY]

만든 날짜: 2020-08-10 오후 5:36

URL: <https://continuous-development.tistory.com/57>

나무늘보의 개발 블로그

홈

태그

방명록

R

[R] R을 활용한 크롤링 - 로또 1등 당첨 배출점 크롤링 하기

· by 꾸까꾸 · 2020. 8. 7. · 수정 · 삭제

해당 사이트의 배출점을 크롤링하겠다. 단순 페이지 크롤링에서 스크립트 기능까지 사용하는 크롤링 까지 하겠다.

분류 전체보기 

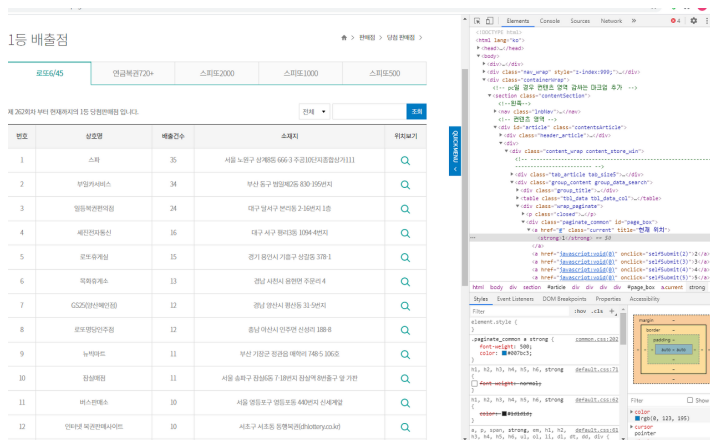
Python 

Database

ASP.NET

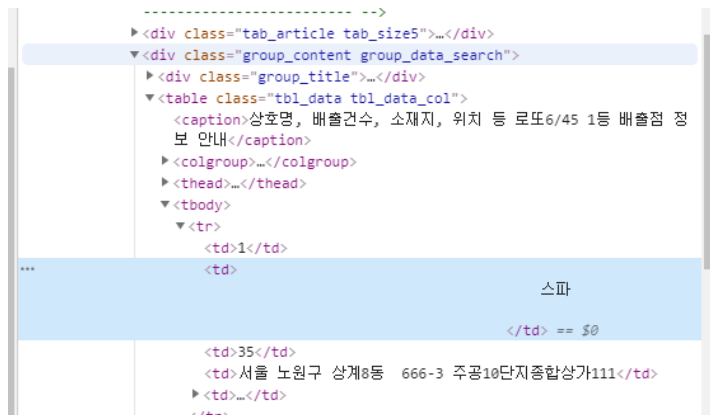
Algorithm

Machine learning
| Deep lear..

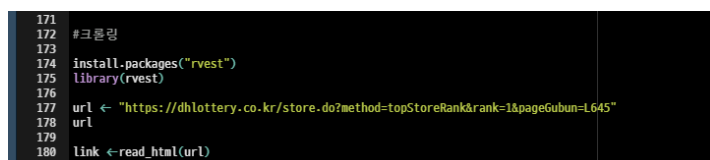


크롤링을 하기 위해서는 첫번째로 내가 원하는 데이터의 위치를 알아야 한다.

저 매출점의 데이터 얻기 위해서는 개발자 도구(F12)에서 해당 부분을 클릭하면 아래와 같이 나온다. 이제 이 부분을 크롤링하기 위해 준비하자.



#html 소스 가져오기



AWS

ETC..

R

공지사항

글 보실 때 주의사항

: 최근글 : 인기글

[Python] 파이썬 기초 3 ----

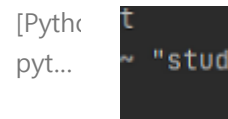
2020.08.10

[Python] 파이썬 기초 2 ----

2020.08.10

[Python] 파이썬 기초 1 ----

2020.08.10



2020.08.10



2020.08.10

최근댓글

태그

python print,

파이썬 기본 출력

문,

python 튜플,

rvest는 R의 웹 스크래핑(Web Scraping)을 위한 패키지로 해당 패키지를 인스톨하고 라이브러리로 넣어 준다.

url은 해당 사이트의 주소이다.

link 값에 html을 read 하고 확인하면 아래와 같이 해당 html 소스를 가져온 것을 볼 수 있다.

[illegible]

※ class는 .으로 찾아들어가고 id는 #으로 찾아들어간다.

#해당 값 가져오기

여기서부터는 해당 태그 부분에 들어가는 작업이다.

1. 로드워킹

```
181
182 #로드워킹(처음부터 순차적으로 찾아 들어가는 방법)
183 link %>%
184   html_nodes('body') %>% #테그 이름으로 찾는것을 의미한다 (중복되지 않는경우에서만 허용된다.)
185   html_nodes('.containerWrap') %>%
186   html_nodes('.contentSection') %>%
187   html_nodes('#article')
188
```

```

198
199 lotto15 <- Link %>%
200     html_nodes('tbody tr td') %>%
201     html_text()
202
203

```

2. 디렉트

ggplot, Oracle,
파이썬 join,
파이썬 형변환,
AWS, 사용법,
날짜함수, rbind,
Python,
Oracle SQL,
파이썬,
python variable,
substr, DDL,
R로 하는 크롤링,
파이썬 함수,
R을 통한 크롤링,
파이썬 tuple,
SQL, 인스턴스,
파이썬 타입,
anaconda 가상환
경,
cbind,
테이블 생성,
행렬,
anaconda 가상환
경 설정,
설정

전체 방문자

162

Today : 5

Yesterday : 0

```

189 #아이디값으로 바로 찾는 방법
190 link %>%
191   html_nodes('#article')

```

가져온 값을 lotto15라는 값에 넣는다.

#전처리 작업

필요 없는 값을 지워주기 위해 전처리를 한다.

```

203
204 library(stringr)
205
206 lotto15 <- str_replace_all(lotto15, "\\t|\\n|\\r", "" )
207 lotto15 <- str_replace_all(lotto15, "[[:space:]]", "" )
208

```

데이터 전처리 후 출력 값이다.

```

[4] "서울노원구신계8동666-3주공10단지공립실가111" "스리지도보기" "72"
[7] "부일가서비스" "34" "부산동구명암제2동838-195번지"
[10] "부일가서비스지도보기" "3" "영등포구천원리삼"
[13] "74" "대구달서구문리동2-16번지1층" "영등포구천원리삼지도보기"
[16] "4" "세관전자동선" "16"
[19] "대구서구평리3동1094-4번지" "세관전자동선지도보기" "5"
[22] "로또유게삼" "15" "경기용인시기흥구성갈동378-1"
[25] "로또유게삼지도보기" "6" "목와유게삼"
[28] "13" "경남사원시용원면주문리4" "목와유게삼지도보기"
[31] "7" "6525(양산해안검)지도보기" "12"
[34] "경남양산시평산동31-5번지" "12" "충남아산시인주면신성리188-8"
[37] "로또영등포구영" "9" "동원리"
[40] "로또영등포구영지도보기" "부산기장공경관매각리748-5186호" "동원리지도보기"
[43] "11" "잠실매점" "11"
[46] "18" "잠실매점지도보기" "11"
[49] "서울송파구잠실6동7-18번지점설역8면송구알가관" "잠실매점지도보기" "11"
[52] "버스원매삼" "10" "서울영등포구영등포동448번지신세계엘"
[55] "버스원매삼지도보기" "12" "인터넷북권매삼사이트"
[58] "18" "서울구시송동동행복권(dhlottery.co.kr)" "인터넷북권매삼지도보기"
[61] "13" "제이북권매삼" "13"
[64] "서울중로구중로5-6가동58번지행행발당1층183호" "제이북권매삼지도보기" "14"
[67] "검정분식관식" "9" "서울중랑구영우문동498-13번지"
[70] "검정분식관식지도보기" "15" "검이프매삼"
[73] "9" "인천중구연안동58-98번지5호" "라이프매삼지도보기"

```

#데이터를 정형화

이제 가지고 데이터를 정형화하는 방법

지금 현재 값이 vector로 순서대로 들어있다. 두 번째 값에는 이름 세 번째 값에는 번호 네 번째 값에는 address 가 들어있다. 그래서 이걸

5로 나눠서 나머지 값이 2,3,4 일 때 저장하는 조건으로 for문을 만들었다. 이걸 다 만든 후 data.frame으로 출력하면

```
210
211 storeName <- NULL
212 cnt <- NULL
213 address <- NULL
214
215 for(idx in 1:length(lotto15)){
216   if(idx %5 == 2){
217     storeName <- c(storeName, lotto15[idx])
218   } else if(idx %5 == 3){
219     cnt <- c(cnt, lotto15[idx])
220   } else if(idx %5 == 4){
221     address <- c(address, lotto15[idx])
222   }
223 }
224
225 lottoDF <- data.frame(storeName, cnt, address)
226 lottoDF
227
```

아래와 같은 값이 나온다.

| | storeName | cnt | address |
|----|-------------|-----|------------------------------|
| 1 | 스파 | 35 | 서울노원구상계8동666-3주공10단지종합상가111 |
| 2 | 부일카서비스 | 34 | 부산동구범일제2동830-195번지 |
| 3 | 일등복권편의점 | 24 | 대구달서구본리동2-16번지1층 |
| 4 | 세진전자통신 | 16 | 대구서구평리3동1094-4번지 |
| 5 | 로또휴게실 | 15 | 경기용인시기흥구상갈동378-1 |
| 6 | 목화휴게소 | 13 | 경남사천시용현면주문리4 |
| 7 | GS25(양산해인점) | 12 | 경남양산시평산동31-5번지 |
| 8 | 로또명당인주점 | 12 | 충남아산시인주면선성리188-8 |
| 9 | 뉴빅마트 | 11 | 부산기장군정관읍매학리748-5106호 |
| 10 | 잠실매점 | 11 | 서울송파구잠실6동7-18번지잠실역8번출구앞가판 |
| 11 | 비스판매소 | 10 | 서울영등포구영등포동440번지신세계앞 |
| 12 | 인터넷복권판매사이트 | 10 | 서초구서초동동행복권(dhlotttery.co.kr) |
| 13 | 제이복권방 | 10 | 서울중로구중로5.6가동58번지평창빌딩1층103호 |
| 14 | 갈렙분식환식 | 9 | 서울중랑구망우본동490-13번지 |
| 15 | 라이프마트 | 9 | 인천중구연안동58-98번지5호 |

정리돼서 나오는 것을 볼 수 있다.

#셀레니움을 이용한 페이지 자동 화 크롤링

첫 번째 페이지는 잘 나왔지만 2페이지 3페이지~마지막 페이지까지의 값을 갖고 싶다.

| | | | | |
|----|--------------|----|------------------------------|--|
| 8 | 로보캣데이인주식 | 12 | 홍남기(이시)인원만 산정액 1888 | |
| 9 | 뉴비테크 | 11 | 부산 거점공공기관출역제역 7485 3000 | |
| 10 | 제일컴퍼니 | 11 | 서울 송파구 가락동 73번지(1)관세청출구 및 가락 | |
| 11 | 비즈니스엔젤 | 30 | 서울 영등포구 가산동 442번지 (1)44만평 | |
| 12 | 인원수 적정현황조사인원 | 30 | 서울구 서초구 동원동 73번지(1)44만평 | |
| 13 | 제일인원수 | 30 | 서울 영등포구 가산동 442번지(1)44만평 | |
| 14 | 인원수적정현황조사인원 | 9 | 서울 송파구 가산동 442번지(1)44만평 | |
| 15 | 제일인원수 | 9 | 인원 적정현황조사인원 500 | |

제삼의 판매업자

1 2 3 4 5 6 7 8 9 10 > >

[illegible]

여기서 다음 순번의 로또 배출점을 구하려 한다.

페이지를 넘기는 형태를 보니 selfSubmit이라는 onclick을 통해 페이지가 넘어가고 있었다.

```
</a>
<a href="javascript:void(0)" onclick="selfSubmit(2)">2</a> ==
<a href="javascript:void(0)" onclick="selfSubmit(3)">3</a>
<a href="javascript:void(0)" onclick="selfSubmit(4)">4</a>
<a href="javascript:void(0)" onclick="selfSubmit(5)">5</a>
<a href="javascript:void(0)" onclick="selfSubmit(6)">6</a>
<a href="javascript:void(0)" onclick="selfSubmit(7)">7</a>
<a href="javascript:void(0)" onclick="selfSubmit(8)">8</a>
<a href="javascript:void(0)" onclick="selfSubmit(9)">9</a>
<a href="javascript:void(0)" onclick="selfSubmit(10)">10</a>
<a class="go next" href="javascript:void(0)" onclick=
"selfSubmit(11)">다음 페이지</a>
<a class="go end" href="javascript:void(0)" onclick=
"selfSubmit(198)">끝 페이지</a>
```

사전에 이거 먼저 깔아주셔야 됩니다.

| 내 PC > 로컬 디스크 (C:) > Rselenium | | | | |
|---------------------------------------|--------------------|---------------------|----------|--|
| 이름 | 수정된 날짜 | 유형 | 크기 | |
| chromedriver.exe | 2020-08-07 오후 1:40 | 응용 프로그램 | 8,825KB | |
| geckodriver.exe | 2020-08-07 오후 1:40 | 응용 프로그램 | 5,999KB | |
| selenium-server-standalone-3.11.0.jar | 2020-08-07 오후 1:40 | Executable Jar File | 22,874KB | |

cmd 창을 통해 셀레니움을 깐 폴더로 이동해서 아래의 명령어를 입력해준다.

```
java -Dwebdriver.gecko.driver="geckodriver.exe" -jar selenium-server-standalone-3.11.0.jar -port 4445
```

```
C:\Users\user> cd Rselenium
C:\Users\user\Rselenium> java -Dwebdriver.gecko.driver="geckodriver.exe" -jar selenium-server-standalone-3.11.0.jar -port 4445
```

이제 R에서 셀레니움을 install을 해준다.

셀레니움은 동적 웹페이지 크롤링을 가능케 하는 함수이다. 동적이라는 건 어떠한 동작을 하면서 크롤링하는 작업을 말한다.

```
228
229 install.packages("Rselenium")
230 library(Rselenium)
231
232
```

아래 작업은 페이지의 마지막 값을 알기 위해 사용하였다.

```
232
233 last <- link %>%
234   html_nodes('.paginate_common') %>%
235   html_nodes('a') %>%
236   html_attr('onclick') %>% tail(1)
237
```

```
>
> last
[1] "selfSubmit(198)"
```

```
> end <- regmatches(last,gregexpr('[0-9]',last))
> end <- as.numeric(end[[1]])
> end <- as.numeric(paste(end[1],end[2],end[3], sep=""))
> end
[1] 198
```

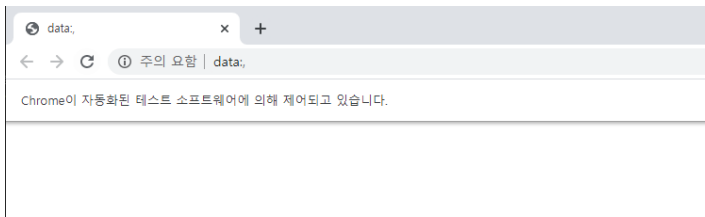
이 가져온 값을 전처리를 통해 숫자의 값만 가져온다.

remoteDriver를 세팅해준다.

remoteDriver 클래스는 JsonWireProtocol을 사용하여 Selenium 서버와 통신하게끔 한다.

```
243
244 remDr <- remoteDriver(remoteServerAddr = "localhost", #내 아이피에서
245   port=4445L, # 포트는 4445를 사용하고
246   browserName = 'chrome') #브라우저는 크롬을 사용한다.
247 remDr$open() #자동화할 것을 제외한, 사이트들을 연다.
```

open을 할경우 아래와 같이 창이 뜬다. 저렇게 자동화된 테스트 소프트웨어~ 라고 써있다.



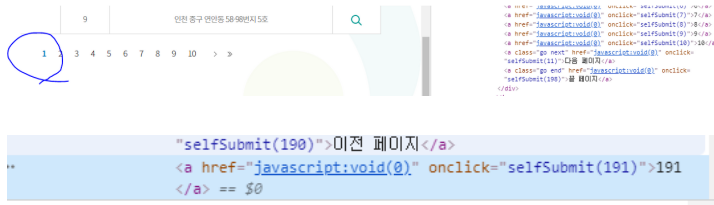
그 다음 url을 넣어서 navigate를 넣어주면 해당 url로 이동한다.

```
remDr$navigate("https://dhlottery.co.kr/store.do?method=topStoreRank&rank=1&pageGubun=L645") #해당 url로 이동하게 한다.
```



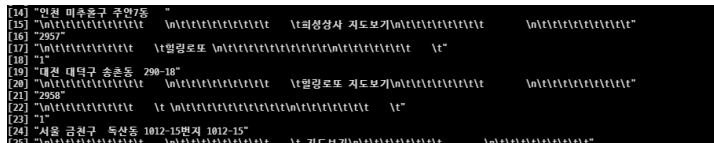
크롤링하는 소스이다.

```
251 lottoStore = c() #초기화 해준다.
252
253 #1등이 배출된 곳 내내지 부지의 정보를 가져오기 위한 for문
254 #현재 자동화 되고있는 사이트를 여러개의 for문을 통해 submit을 불러오는 명령어다.
255
256 for(idx in 1:100){
257   front <- "selfSubmit"
258   back <- ""
259   script <- paste(front, idx, back, sep="") # 이 3문장으로 script에 링크 스크립트를 실행 시킬 단어를 만든다. 여기서 script는 selfSubmit(수치) 값이다.
260   pagemove <- remDr$executeScript(script, args=1:2) # 이 script에 들어있는 텍스트를 받아 javascript를 실행 시킨다.
261
262   #소스 받아오기
263   source <- remDr$getPageSource()[[1]] # 현재 자동화로 개진 사이트의 소스를 가져온다.
264   js_html <- read_html(source) # 소스를 읽어온다.
265
266   js_link <- html_nodes(js_html, "tbody") #tbody라는 태그를 가져온다.
267   js_link
268
269   stores <- js_link %>%
270     html_nodes("tr") %>%
271     html_nodes("td") %>%
272     html_text()
273
274   lottoStore = c(lottoStore, stores) # lottoStore에 백터 형태로 추가해서 넣어주는 작업을 한다. 이곳에 for문에서 나온 데이터를 lottoStore에 계속 넣는다.
275 }
276
277 }
```



이 selfSubmit에 따라 자바스크립트가 실행되면 페이지가 1->198까지 순차적으로 이동한다.

위에 for문을 돌리게 되면 페이지가 자동으로 넘어가고 넘어가는 페이지를 lottoStore에 저장한다.



값을 출력해 보면 위같은 형식으로 나온다.

내가 가져온 값은

| 번호 | 상품명 | 매출단위 | 소재지 | 회사명/기 |
|----|-----|------|----------------------------------|-------|
| 1 | 스파 | 35 | 서울 노원구 상계8동 666-3 주공10단지종합상가1111 | Q |

이 부분에 해당하는 값을 전부 가져왔다.

그다음 가져온 값을 전처리하는 작업을 거친다.

```
280 #리미트 값임
281 lottoStore <- str_replace_all(lottoStore, "\\t\\n\\r", "") # 필요없는 \t\n\r text 지우기
282 lottoStore <- str_replace_all(lottoStore, "[[:space:]]", "" )
283
284 storeName <- NULL
285 cnt <- NULL
286 address <- NULL
287
288 for(idx in 1:length(lottoStore)){ #두번째 값이 storeName 이고 세번째 값이 cnt(번호) 이고 4번째 값이 address 이다.
289   if(idx %%% 2 == 2){ #리미트 값은 필요없어서 이 순서대로 값을 가져와서 저장한다.
290     storeName <- c(storeName, lottoStore[idx])
291   }else if(idx %%% 5 == 3){
292     cnt <- c(cnt, lottoStore[idx])
293   }else if(idx %%% 5 == 4){
294     address <- c(address, lottoStore[idx])
295   }
296 }
297
298 lottoDF <- data.frame(storeName, cnt, address) #만든걸 데이터프레임으로 만든다.
299 lottoDF
300
```

전처리 후 결과이다.

```
[985] "4" "영등포구동원동115-2번지우마슈퍼점" "185"
[986] "4" "4"
[987] "우밀" "4"
[988] "서울송파구마천2동25-2번지" "우밀저도보기"
[989] "183" "우정식당"
[990] "4" "부산동래구온천제1동185-93번지"
[991] "우정식당" "184"
[992] "우수대동" "4"
[993] "경기수원시권선구호매실동87-2" "우수대동저도보기"
[994] "185" "원당역북권방"
[995] "4" "경기고양시석왕구성사동410-7"
[996] "원당역북권방저도보기" "186"
[997] "안산대여점" "4"
```

마지막 만든 데이터 프레임을 csv파일로 생성한다.

```
297
298 lottoDF <- data.frame(storeName, cnt, address) #만든걸 데이터프레임으로 만든다.
299 lottoDF
300
301 write.csv(lottoDF, "lotto_store.csv", row.names = F) # csv 파일로 읽는다.
302
```

csv 파일을 읽으면 아래와 같이 나온다.

| | A | B | C | D | E | F | G | H |
|----|-----------|-----|-----------------------------|---|---|---|---|---|
| 1 | storeName | cnt | address | | | | | |
| 2 | 스파 | 35 | 서울노원구상계8동666-3주공10단지중합상가111 | | | | | |
| 3 | 부일카서브 | 34 | 부산동구범일제2동830-195번지 | | | | | |
| 4 | 일등복권 | 24 | 대구달서구분리동2-16번지1층 | | | | | |
| 5 | 세진전자 | 16 | 대구서구평리3동1094-4번지 | | | | | |
| 6 | 로도휴게소 | 15 | 경기용인시기흥구상갈동378-1 | | | | | |
| 7 | 목화휴게소 | 13 | 경남사천시용현면주문리4 | | | | | |
| 8 | GS25(양산) | 12 | 경남양산시평산동31-5번지 | | | | | |
| 9 | 로도명당 | 12 | 충남아산시인주면신성리188-8 | | | | | |
| 10 | 뉴빅마트 | 11 | 부산기장군정관읍매학리748-5106호 | | | | | |
| 11 | 잠실매점 | 11 | 서울송파구잠실6동7-18번지잠실역8번출구앞가판 | | | | | |
| 12 | 버스판매소 | 10 | 서울영등포구영등포동440번지신세계앞 | | | | | |
| 13 | 인터넷복권 | 10 | 서초구서초동동행복권(dhlottery.co.kr) | | | | | |
| 14 | 제이복권 | 10 | 서울중로구중로5.6가동58번지평창빌딩1층103호 | | | | | |
| 15 | 갈림분식 | 9 | 서울중랑구망우본동490-13번지 | | | | | |
| 16 | 라이프마트 | 9 | 인천중구연안동58-98번지5호 | | | | | |
| 17 | 북마산복권 | 9 | 경남창원시마산합포구39-4번지 | | | | | |



'R' 카테고리의 다른 글

- [R]R을 활용한 상관분석과 회귀분석
- 2 (0) 02:42:26
- [R] R을 활용한 크롤링 - 로또 1등 당첨 배출점 크롤링 하기 (0) 2020.08.07
- [R] R에서 교차검증을 위한 데이터 셋 분리방법 3가지 (0) 2020.08.07
- [R] R을 활용한 상관분석과 회귀분석
- 1 (0) 2020.08.06
- [R] R을 통한 텍스트마이닝에서 워드 클라우드 까지 (0) 2020.08.05
- [R] R로 하는 비정형 데이터 처리 (facebook 데이터를 통한 긍정/부정 나누기) (0) 2020.08.03

태그

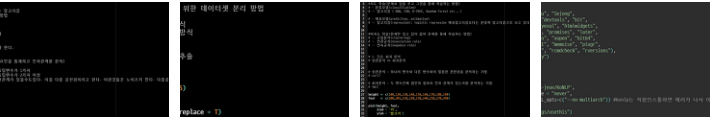
R 크롤링

R로 하는 크롤링

R을 통한 크롤링

크롤링

관련글



[R]R을 활... [R] R에서 ... [R] R을 활... [R] R을 통...

댓글 0

