

R 기반 웹 크롤링

한국환경정책·평가연구원 (KEI)

부연구위원 진대용

dyjin@kei.re.kr

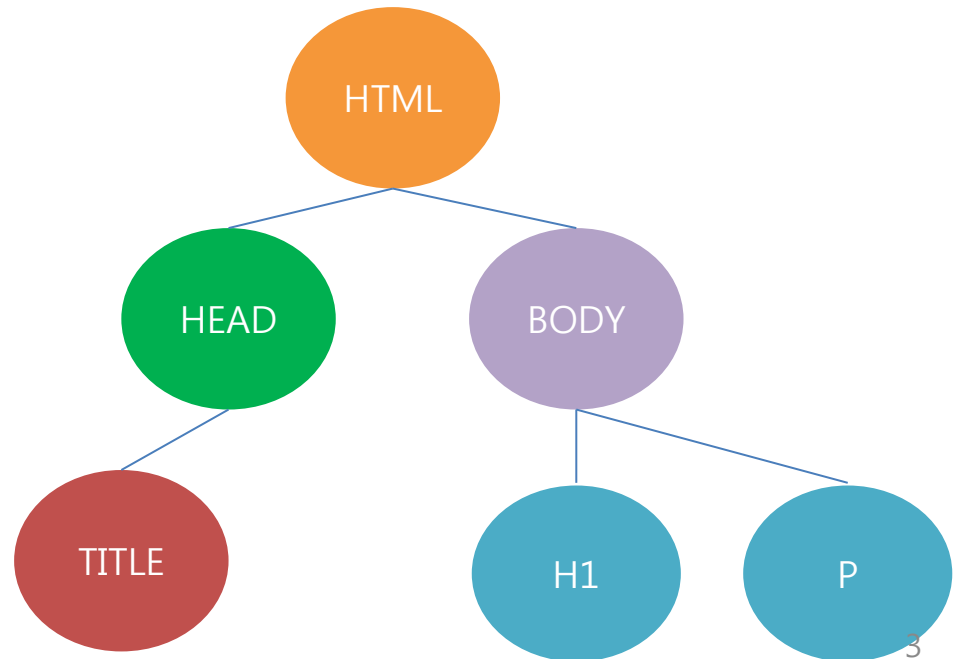
웹의 기본구조

- HTML + CSS + 자바스크립트로 동작
 - HTML : 구조
 - CSS : 스타일
 - 자바스크립트 : 동작

HTML의 구조

- Hyper Text Markup Language
 - 문서의 구조를 나타내는 마크업 언어
- <tag>내용</tag> 형태
- 태그 내 태그 구조

```
<!DOCTYPE html>
<html>
  <head>
    <title>Page Title</title>
  </head>
  <body>
    <h1>This is a Heading</h1>
    <p>This is a paragraph.</p>
  </body>
</html>
```

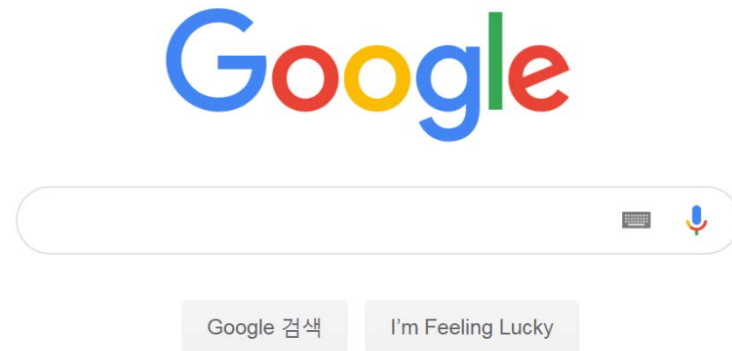


HTML Element

- <태그이름 (속성=속성값)> 내용 </태그이름>

구글

구글



주요 HTML 태그 정리

- `<h1>`글자크기 조정 태그`<h1>`
 - h1,h2,h3...
- `<div>`레이아웃 태그`</div>`
- `<p>` 문단 태그 `</p>`
- ``구글``
- `` 다른 태그 내에서 글자 일부 선택``
- 속성으로는 id, class, href, src 등이 주로 쓰임

CSS : Cascading Style Sheet

- CSS를 통해 스타일 지정 : 글자색, 글자 크기 등
- CSS Selector
 - HTML Element들을 선택

```
<html>
  <head>
    <style>
      p.bluetext {
        text-align: center;
        color: blue;
      }
    </style>
  </head>
  <body>
    <p class="bluetext">Hello World!</p>
  </body>
</html>
```

“p로 된 태그에 class 속성이
bluetext인 부분에 스타일을 적용”

CSS : Cascading Style Sheet

- CSS를 통해 스타일 지정 : 글자색, 글자 크기 등
- CSS Selector
 - HTML Element들을 선택

```
<html>
  <head>
    <style>
      p.bluetext {
        text-align: center;
        color: blue;
      }
    </style>
  </head>
  <body>
    <p class="bluetext">Hello World!</p>
  </body>
</html>
```


“p로 된 태그에 class 속성이
bluetext인 부분에 스타일을 적용”

CSS Selector

| Selector | Example | Example description |
|--|------------------|--|
| <u>.class</u> | .intro | Selects all elements with class="intro" |
| <u>#id</u> | #firstname | Selects the element with id="firstname" |
| <u>element</u> | p | Selects all <p> elements |
| <u>element,element</u> | div, p | Selects all <div> elements and all <p> elements |
| <u>element> element</u> | div > p | Selects all <p> elements where the parent is a <div> element |
| <u>element element</u> | div p | Selects all <p> elements inside <div> elements |
| <u>[attribute]</u> | [target] | Selects all elements with a target attribute |
| <u>[attribute= value]</u> | [target=_blank] | Selects all elements with target="_blank" |
| <u>:nth-child(n)</u> | p:nth-child(2) | Selects every <p> element that is the second child of its parent |
| <u>:nth-of-type(n)</u> | p:nth-of-type(2) | Selects every <p> element that is the second <p> element of its parent |


https://www.w3schools.com/cssref/css_selectors.asp

CSS Selector 연습

 Creative Cloud

크리에이티브의 실현
새로운 앱, 새로운 기능, 지금 바로
시작하세요! 월 11,000원부터(부가세 포함)

지금 가입



Click a selector to see which element(s) that gets selected in the result:

Selector:
li:nth-child(1)

All elements that are the first child of their parent.

.intro
#LastName
.intro, #LastName
h1
h1, p
div p
div > p
ul + p
ul ~ table
*
[id]
[id=my-Address]
[id\$=ess]
[id=my]
[id^=L]
[title~=beautiful]
[id*=s]
:checked
:disabled
:enabled
:empty
:focus
p:first-child
p::first-letter
p::first-line
p:first-of-type
h1:hover
input:in-range
input:out-of-range
input:invalid
input:valid
p:lang(it)
p:last-child
p:last-of-type
tr:nth-child(even)
tr:nth-child(odd)
li:nth-child(1)
li:nth-last-child(1)
li:nth-of-type(2)

Result:

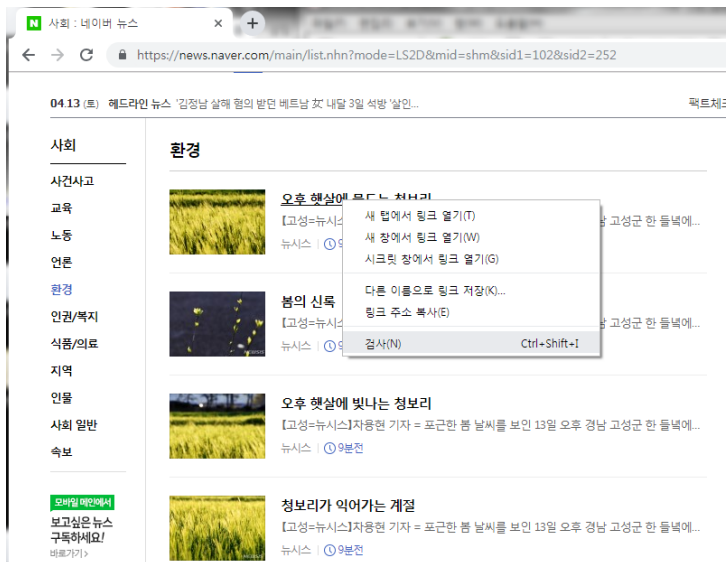
```
<h1>Welcome to My Homepage</h1>
<div class="intro">
<p>My name is Donald <span id="LastName"> Duck.</span> </p>
<p id="my-Address"> I live in Duckburg</p>
<p> I have many friends:</p>
</div>
<ul id="Listfriends">
  • <li>Goofy</li>
  • <li>Mickey</li>
  • <li>Daisy</li>
  • <li>Pluto</li>
</ul>
<p>All my friends are great! <br>
But I really like Daisy!!</p>
<p lang="it" title="Hello beautiful">Ciao bella</p>
<h3>We are all animals!</h3>
<p> <b>My latest discoveries have led me to believe that we are all animals:</b> </p>
<table>
  Name  Type of Animal
  Mickey Mouse
  Goofey Dog
  Daisy Duck
  Pluto Dog
</table>
```

<https://www.w3schools.com/cssref/tryel.asp>

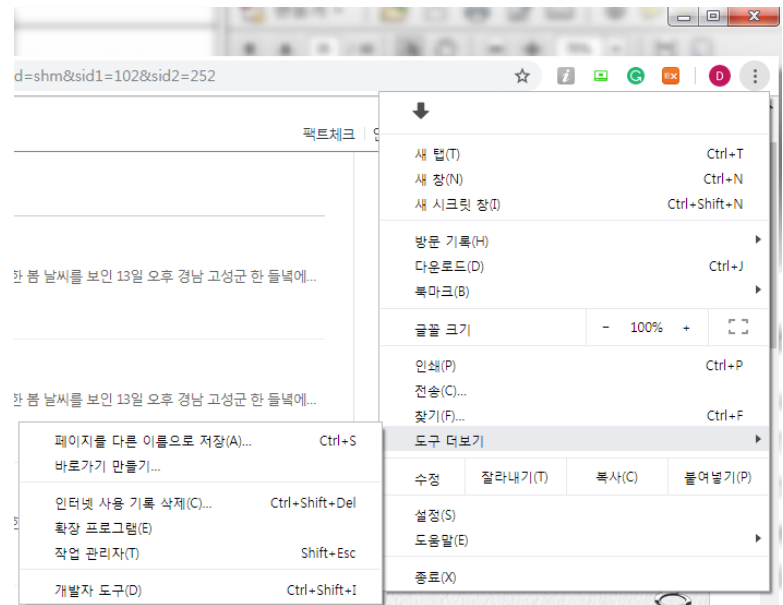
9

크롬 개발자 도구

- 네이버 환경뉴스 제목의 CSS selector 찾기
 - 네이버 뉴스 > 사회 > 환경



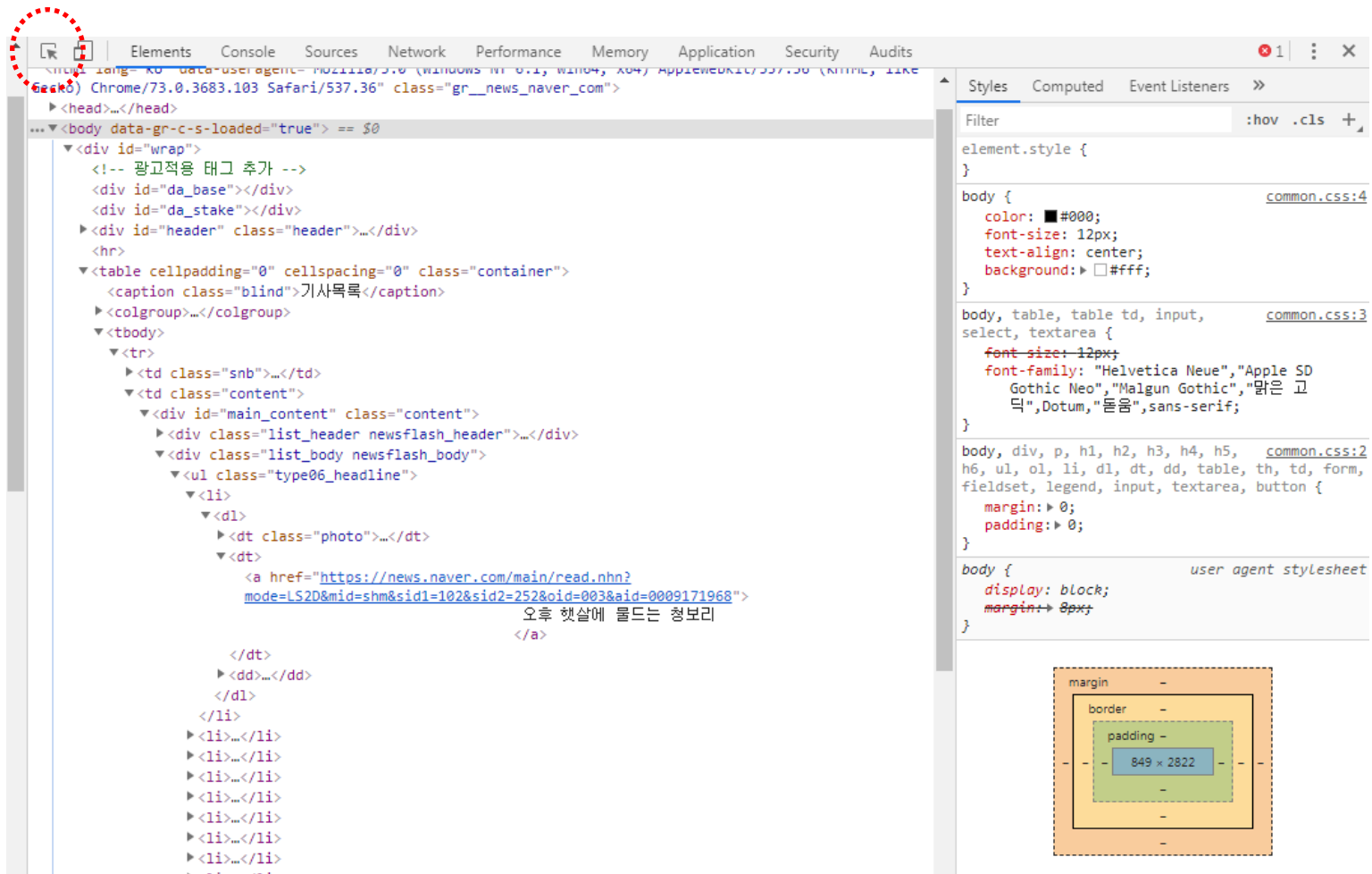
오른쪽 마우스 클릭 > 검사



F12 또는

도구 더보기 > 개발자 도구

크롬 개발자 도구

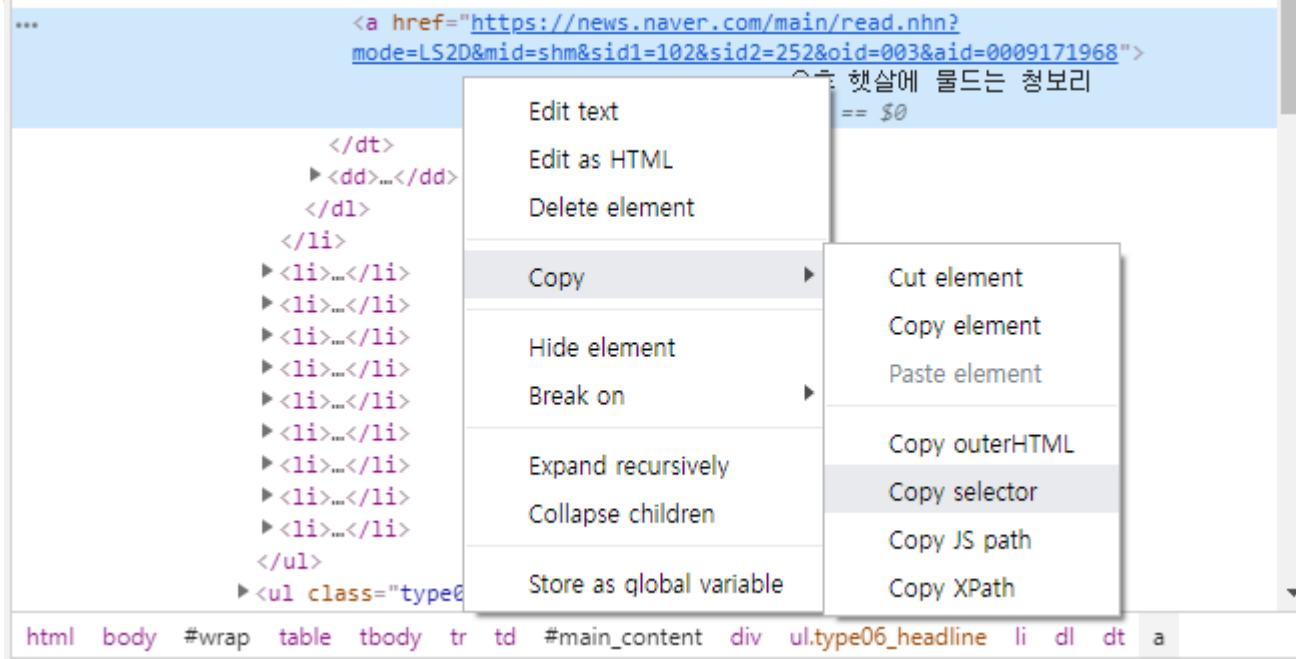


환경



【고성=뉴시스】차용현 기자 = 포근한 봄 날씨를 보인 13일 오후 경남 고성군 한 들녘에...

뉴스 | ㉔ 19부전



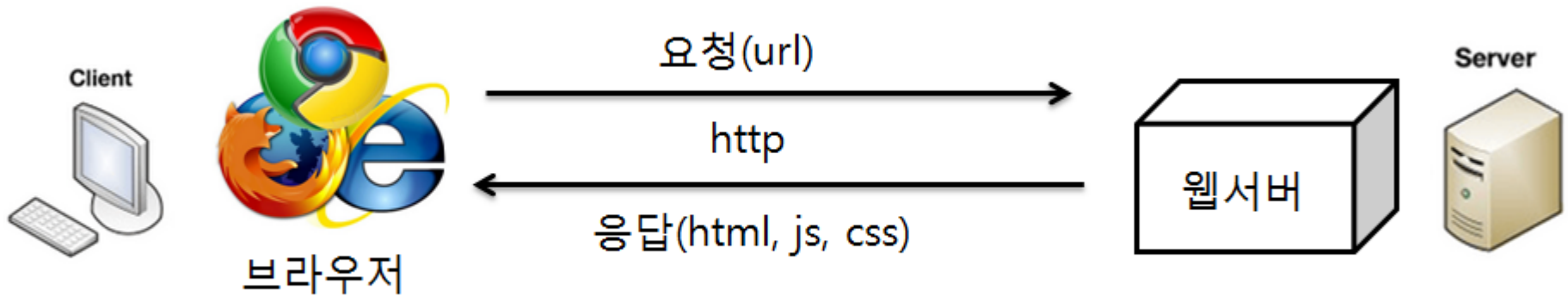
```
#main_content > div.list_body.newsflash_body > ul.type06_headline > li:nth-child(1) > dl > dt:nth-child(2) > a
```

크롬 개발자 도구

```
<!doctype html>
<html lang="ko" data-useragent="Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.3683.103 Safari/537.36" class=
"gr_news_naver_com">
  <head>...</head>
  <body data-gr-c-s-loaded="true">
    <div id="wrap">
      <!-- 광고적용 태그 추가 -->
      <div id="da_base"></div>
      <div id="da_stake"></div>
      <div id="header" class="header">...</div>
      <hr>
      <table cellpadding="0" cellspacing="0" class="container">
        <caption class="blind">기사목록</caption>
        <colgroup>...</colgroup>
        <tbody>
          <tr>
            <td class="snb">...</td>
            <td class="content">
              <div id="main_content" class="content">
                <div class="list_header newsflash_header">...</div>
                <div class="list_body newsflash_body">
                  <ul class="type06_headline">
                    <li>
                      <dl>
                        <dt class="photo">...</dt>
                        <dt>
                          <a href="https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=102&sid2=252&oid=003&aid=0009171968">
                            오후 햇살에 물드는 청보리
                          </a> == $0
                        </dt>
                      <dd>...</dd>
                    </dl>
                  </li>
                </ul>
              </div>
            </td>
          </tr>
        </tbody>
      </table>
    </div>
  </body>
</html>
```

#main_content > div.list_body.newsflash_body > ul.type06_headline > li:nth-child(1) > dl > dt:nth-child(2) > a

서버-클라이언트 구조



HTTP : Hyper Text Transfer Protocol : 요청을 보내면 응답을 보내줌



브라우저
URL에 정보 표시



서버에 정보 전달
URL에 정보표시 X
파일전송



크롤링 및 과정

- 크롤링이란?
 - 웹페이지 가져오기
- 크롤링과정 : 요청 - 추출 - 저장
 1. HTML을 R 구조의 객체 생성 및 페이지 크롤링
 2. CSS Selector로 원하는 HTML 요소 추출
 3. 출력

R 기반 크롤링

- 주요 패키지
 - httr
 - rvest
 - Rselenium
 - 동적 웹페이지 크롤링

httr 패키지 사용법

- 표준 HTTP 요청 및 응답에 활용
- GET 방식 페이지 요청 방법
 - `d <- GET(URL, query=list(인자1="값1", 인자2="값2"))`
- POST 방식 페이지 요청 방법
 - `d <- POST (URL, body=list(인자1="값1", 인자2="값2"), encode)`

httr 패키지 사용법

- 상태코드 확인
 - 200(성공), 404(Not Found, 찾을 수 없음)
 - https://ko.wikipedia.org/wiki/HTTP_상태 코드
 - `status_code(d)`, `d$status_code`
- HTTP 응답 결과 확인
 - `print(d)`
- HTTP 응답 BODY를 HTML 형태로 반환
 - `content(d)`
 - `cat(content(d, "text"), "\n")`

httr 패키지 사용법

- <http://www.naver.com> 크롤링

```
1 #install.packages("httr",dependencies = T)
2 library(httr)
3
4 d <- GET("http://www.naver.com")
5 result <- cat(content(d, "text"), "\n")
6
```

httr 패키지 사용법

- <http://comic.naver.com/webtoon/list.nhn?titleId=183559&weekday=mon>

```
1 #install.packages("httr",dependencies = T)
2 library(httr)
3
4 d1 <- GET("http://comic.naver.com/webtoon/list.nhn?titleId=183559&weekday=mon")
5 d2 <- GET("http://comic.naver.com/webtoon/list.nhn",query=list(titleId=183559,weekday="mon"))
6
7 d1_result <- content(d1, "text")
8 cat(d1_result,"\n")
9
10 d2_result <- content(d2, "text")
11 cat(d2_result,"\n")
```

httr 패키지 사용법

- 교보문고 '기계학습' 검색 결과

```
1 library(httr)
2
3 param <- list(vPstrCategory='TOT',vPstrKeyword='&#44592;&#44228
; &#54617;&#49845;',vPplace='top',searchCategory='TOT')
4 d3 <- POST("http://www.kyobobook.co.kr/search/SearchCommonMain
.jsp",body=param,encode='form')
5 d3_result <- content(d3, "text")
6 cat(d3_result,"\n")
7 write.table(d3_result,"d3_result.txt")
```

URL 한글 변환

- 참고
 - 퍼센트 인코딩 방식
 - 퍼센트(URL) 인코딩 / 디코딩 변환
 - <http://www.hipenpal.com/tool/url-encode-and-decode-in-korean.php>

URL 한글 변환

```
temp = URLdecode("%26%2347672%3B%26%2349888%3B%26%2347084%3B%26%2345789%3B")
# [1] "&#47672;&#49888;&#47084;&#45789;"
s[1] <- intToUtf8(47672)
s[2] <- intToUtf8(49888)
s[3] <- intToUtf8(47084)
s[4] <- intToUtf8(45789)
```

- 질문) 백터를 합치는 방법?

Rvest 패키지 사용법

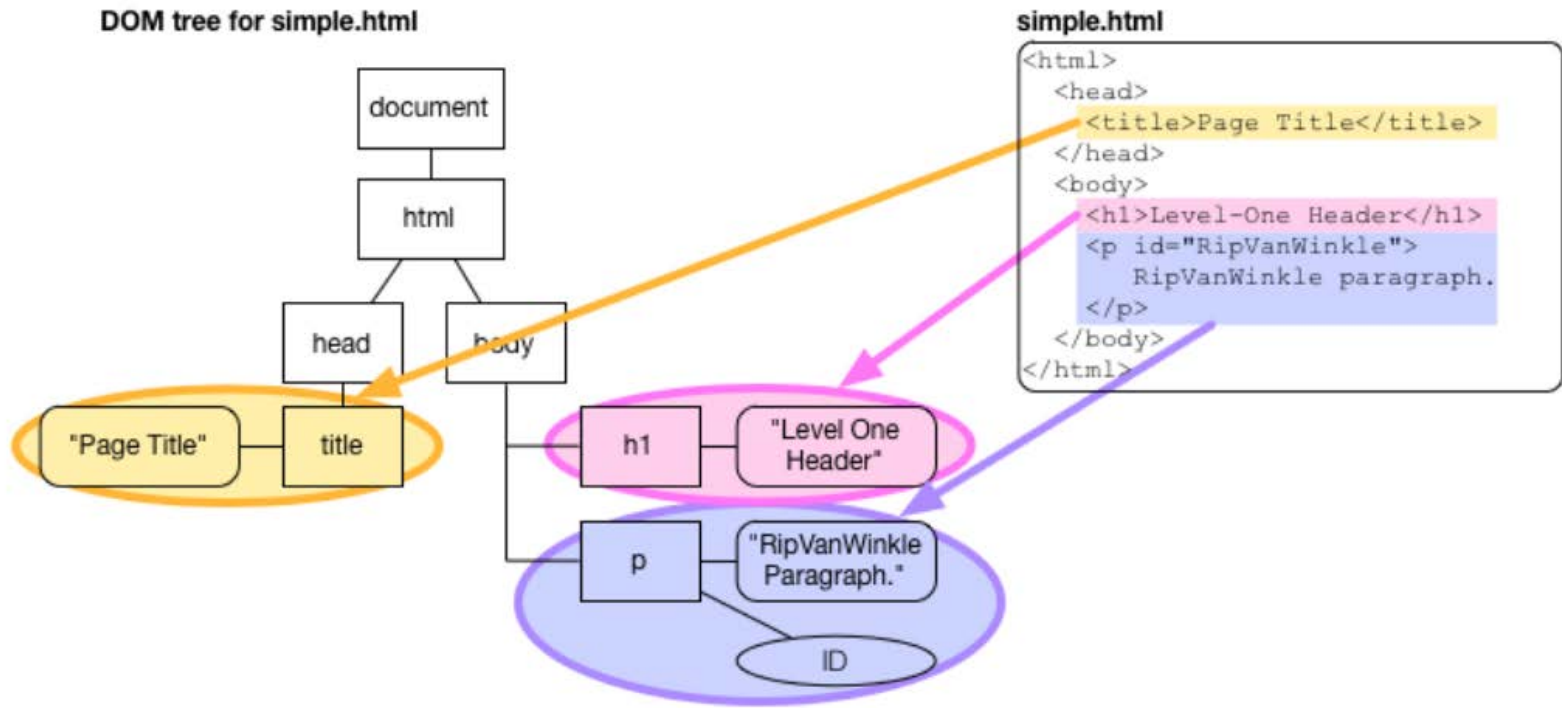
- Rvest
 - 웹페이지 수집 및 가공

Rvest 패키지 주요함수

- `read_html()` : 응답객체를 HTML로 변환
- `html_node()`, `html_nodes()` : HTML 요소 추출
- `html_attr()` : HTML 속성 기반 추출
- `html_text()` : 텍스트 출력
- `html_table()` : 테이블 출력
- 변환 > 요소 추출 > 출력

Node?

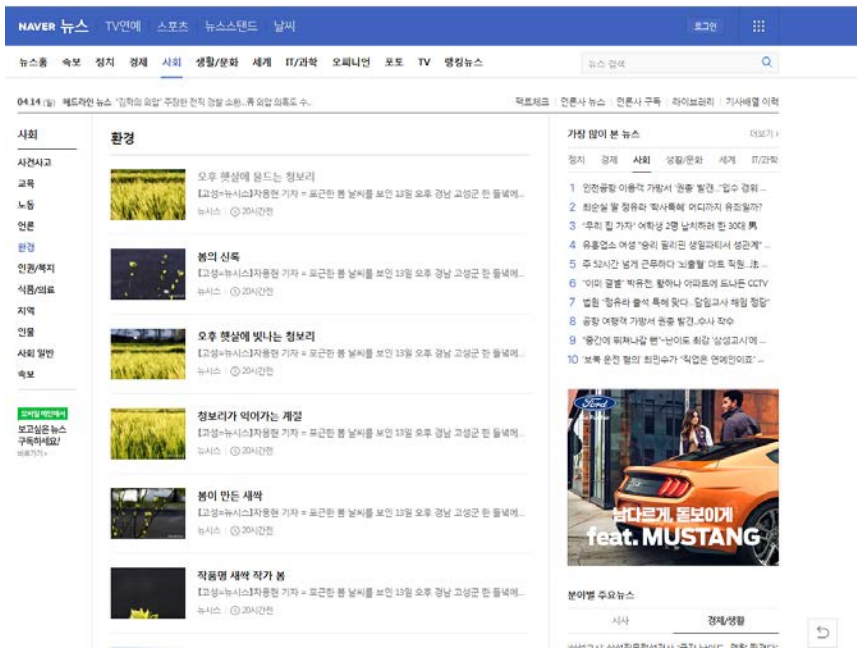
- Node = HTML tag



참고자료 : <https://mrchypark.github.io/getWebR/#26>

네이버 환경뉴스 제목 수집

- 네이버 환경뉴스
 - 네이버 뉴스 > 사회 > 환경



2019년 4월 13일

<https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid2=252&sid1=102&date=20190413>

네이버 환경뉴스 제목 수집

- #main_content > div.list_body.newsflash_body > ul.type06_headline > li:nth-child(1) > dl > dt:nth-child(2) > a
- #main_content > div.list_body.newsflash_body > ul.type06_headline > li:nth-child(2) > dl > dt:nth-child(2) > a
- ...

네이버 환경뉴스 제목 수집

- 4월 12 첫 페이지 제목 수집

```
1 library(httr)
2 library(rvest)
```

HTML 변환

```
3
4 URL <- 'https://news.naver.com/main/list.nhn'
5 param <- list(mode='LS2D',mid='shm',sid1=102,sid2=252,date=20190412)
6 d <- GET(URL,query=param)
7 html <- read_html(d,encoding="EUC-KR")
```

```
8
9 title_nodes <- html_nodes(html,"#main_content > div.list_body.newsflash_body
> ul.type06_headline > li > dl > dt:nth-child(2) > a")
```

```
10
11 titles <- html_text(title_nodes)
```

텍스트 출력

요소 추출

```
12
13 # 정리
14 titles <- gsub("[\t\r\n]", "", titles)
15 titles <- trimws(titles, which = "left")
16 titles <- trimws(titles, which = "right")
```

정리

URL 구조 분석

- 네이버 환경 뉴스 페이지 변경



2페이지 클릭:

<https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid2=252&sid1=102&date=20190412&page=2>

3페이지 클릭:

<https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid2=252&sid1=102&date=20190412&page=3>

네이버 환경뉴스 수집

- 4월 12일 1-5페이지 제목,날짜,내용(주소) 수집
- 제목(HTML 내용), 제목의 링크(HTML속성)
 - #main_content > div.list_body.newsflash_body > ul.type06_headline > li > dl > dt:nth-child(2) > a
- 날짜 : 4월 12일

네이버 환경뉴스 수집

```
1 library(httr)
2 library(rvest)
3
4 title_list = NULL
5 date_list = NULL
6 content_url_list = NULL
7
8 for(i in 1:5){
9   URL <- 'https://news.naver.com/main/list.nhn'
10  param <- list(mode='LS2D',mid='shm',sid1=102,sid2=252,date
    =20190412,page=i)
11
12
13  d <- GET(URL,query=param)
14  html <- read_html(d,encoding="EUC-KR")
```


네이버 환경뉴스 수집

```
15 title_nodes <- html_nodes(html, "#main_content > div.list_body  
  .newsflash_body > ul.type06_headline > li > dl > dt:nth-child(2)  
  > a")  
16  
17 dates <- "20190412"  
18 date_list <- c(date_list, dates)  
19  
20 titles <- html_text(title_nodes)  
21 contents <- html_attr(title_nodes, 'href')  
22  
23 # 정리  
24 titles <- gsub("[\t\r\n]", "", titles)  
25 titles <- trimws(titles, which = "left")  
26 titles <- trimws(titles, which = "right")  
27 title_list <- c(title_list, titles)  
28  
29 content_url_list <- c(content_url_list, contents)  
30  
31 }  
32  
33 result_mat <- cbind(date_list, title_list, content_url_list)
```

네이버 환경뉴스 수집

| | date_list | title_list | content_url_list |
|----|-----------|--------------------------------------|--|
| 1 | 20190412 | [날씨] '갑갑한 주말'...구름·안개에 미세먼지도 '나쁨' | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 2 | 20190412 | BBC·디스커버리 참여한 다큐 '세행계티' KBS서 세계 첫방송 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 3 | 20190412 | '환경오염 논란' 영풍석포제련소, 기존지 넘긴 수질오염 물질... | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 4 | 20190412 | 장기미집행 대전 매봉공원 내 아파트 건설계획 부결 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 5 | 20190412 | 대전 매봉공원 특례사업 제동...도시계획법 "생태 양호" 부결 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 6 | 20190412 | 보꽃 편 들레방아 공원 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 7 | 20190412 | [환경뉴스] 16분 동안 잠수하는 '스쿠버 다이버' 도마뱀의 비밀 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 8 | 20190412 | 제주해경, 차귀도 해상서 선저 폐수 유출한 중국어선 적발 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 9 | 20190412 | 해경, 제주 해상에 기름 유출 중국 어선 나포 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 10 | 20190412 | "유네스코 세계자연유산 마을 파괴하는 동물테마파크 중단..." | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 11 | 20190412 | 수소차 충전소 설명 듣는 김현미 장관 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 12 | 20190412 | 박천규 차관, 싱가포르 환경수자원부 선임국무장관 면담 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 13 | 20190412 | 수소충전소 개소식 참석한 김현미 장관 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 14 | 20190412 | 수소충전소 개소식서 기념사하는 김현미 장관 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 15 | 20190412 | 수소차 충전 시연하는 김현미 장관 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 16 | 20190412 | 이시종 지사 "청주 토끼리 소각장 환경평가에 주민 뜻 반영..." | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 17 | 20190412 | 불화 영풍석포제련소 방류수에 또 오염물질 검출 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 18 | 20190412 | 제주동물테마파크사업 중단 요구 기자회견 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 19 | 20190412 | '제주도에 사파리 공원은 절대 반대' | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 20 | 20190412 | '람사르 습지드시에 동물원은 필요없다' | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 21 | 20190412 | 노라조 원훈, 기후변화 홍보대사 위촉 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 22 | 20190412 | 위촉패 전달 받는 노라조 조빈 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |
| 23 | 20190412 | 박지훈, 기후변화 홍보대사 위촉 | https://news.naver.com/main/read.nhn?mode=LS2D&mid=... |

네이버 실시간 검색어 수집

NAVER <https://www.naver.com>

네이버를 시작페이지로 > | [유니버세이버](#) [해피빈](#)

NAVER

메일 카페 블로그 지식iN 쇼핑 Pay TV 사전 뉴스 증권 부동산 지도 영화 뮤직 책 웹툰 더보기

현대애상 다이렉트
2019 올해는 내 차 보험료 얼마까지 할인될까?
손해보험협회 심의결 제3096호(2018.4.25)

32% 주행거리특약 (연간 3천km 이하)
13% 자녀할인특약 (태아, 부부 1인 한정 기준)
11.3% 무사고 (3년 이상)
4.7% 교통법규준수 (무사고기준)

지금 확인

연합뉴스 > '이미션 거취' 與野 정면 격돌... "불법 없다" vs "검찰 고발"
네이버뉴스 연예 스포츠 경제

뉴스탠드 > 전체 언론사 MY 뉴스

KJD MBN sportalkorea ZNet Korea 전자신문 스포츠서울

급상승 검색어 DataLab. 급상승 트래킹 >

1~10위 11~20위

- 1 에스아이빌리지
- 2 인스타 오류
- 3 정밀세 차가격
- 4 자동차 래핑 가격
- 5 페이스북 오류
- 6 안현모
- 7 명계남
- 8 시카리오
- 9 국한직업 정밀세차
- 10 원피스 880화 애니

2019. 04. 14. 21:44:00 기준

네이버 실시간 검색어 수집

- CSS Selector

- #PM_ID_ct > div.header > div.section_navbar > div.area_hotkeyword.PM_CL_realtimeKeyword_base > div.ah_roll.PM_CL_realtimeKeyword_rolling_base > div > ul > **li:nth-child(1)** > a > span.ah_k
- #PM_ID_ct > div.header > div.section_navbar > div.area_hotkeyword.PM_CL_realtimeKeyword_base > div.ah_roll.PM_CL_realtimeKeyword_rolling_base > div > ul > **li:nth-child(2)** > a > span.ah_k
- ...

네이버 실시간 검색어 수집

```
1 library(httr)
2 library(rvest)
3
4 URL <- 'http://www.naver.com'
5
6 d <- GET(URL)
7 html <- read_html(d,encoding="EUC-KR")
8 rt_nodes <- html_nodes(html,"#PM_ID_ct > div.header > div
  .section_navbar > div.area_hotkeyword.PM_CL_realtimekeyword_base
  > div.ah_roll.PM_CL_realtimekeyword_rolling_base > div > ul > li
  > a > span.ah_k")
9 rt_texts <-html_text(rt_nodes)
```

| | V1 |
|----|----------------------|
| 1 | 김응명 |
| 2 | 정밀세차가격 |
| 3 | 인스타 오류 |
| 4 | 에스아이빌리지 |
| 5 | 자동차 래핑 가격 |
| 6 | 페이스북 오류 |
| 7 | 안현모 |
| 8 | 시카리오 |
| 9 | 명계남 |
| 10 | 원피스 880화 애니 |
| 11 | 극한직업 정밀세차 |
| 12 | 세상에서 제일 예쁜 내 딸 인물관계도 |
| 13 | 경수진 |
| 14 | 박지빈 |
| 15 | 김소연 이상우 |
| 16 | 유선 |
| 17 | 하트 오브 더 씨 |
| 18 | 우원재 |
| 19 | 한재아 |
| 20 | 주광덕 |

서울시 응답소 민원 수집

원순씨에게 바랍니다<민원신청>

https://eungdapso.seoul.go.kr/Shr/Shr01/Shr01_lis.jsp

인권침해구제신청 >

민생침해신고(눈물그만) >

노동조사신고 >

지방세 이의신청 >

공직비리 익명신고

지방공무원, 지방공기업 대상

▶ 신청하기

▶ 결과보기

▶ 이전 시장

공개일

2011-10-27 ~ 2019-04-14 (입력예 : 2013-10-01)

검색어

검색어 구분

Q

번호

제목

신청일

조회수

11832

가로수 가지치기와 미세먼지

2019-03-26

69

11831

버스 민원입니다.

2019-03-21

51

11830

안 타는 쓰레기통투 5L 만들어주세요.

2019-03-20

15

11829

현릉로 지하화 요청

2019-03-20

10

11828

노후경유차 과태료 유예는 언제까지?

2019-03-19

76

11827

"서울시 제3화장장" 건립을 건의 드.

2019-03-19

8

11826

시각장애인 톨택시 확대 운영부탁합니다

2019-03-16

30

11825

구로구에 한국산업박물관을 유치해 주세요.

2019-03-14

25

11824

청년스타트업을 위한 마케팅 지원 관련.

2019-03-14

32

11823

덜컹한차량 배출5등급차량 선정에 불만.

2019-03-12

16

1

2

3

4

5

>

· '원순씨에게 바랍니다(시장에게 바란다)'는 이용자 의견에 답변이 필요하다고 판단되는 경우 민원사무에 준하여 처리됩니다.
(서울특별시 인터넷 홈페이지 운영 활성화에 관한 조례 제6조)

38

서울시 응답소 민원 수집

- CSS Selector

- #content_cont > div.info_wrap > div > form > div.pclist_table.mt20 > div:nth-child(2) > ul > li.pclist_list_tit42
- #content_cont > div.info_wrap > div > form > div.pclist_table.mt20 > div:nth-child(3) > ul > li.pclist_list_tit42
- ...

서울시 응답소 민원 수집

```
1 library(httr)
2 library(rvest)
3
4 URL <- 'https://eungdapso.seoul.go.kr/Shr/Shr01/Shr01_lis.jsp'
5
6 d <- GET(URL)
7 html <- read_html(d,encoding="UTF-8")
8 title_nodes <- html_nodes(html,"#content_cont > div.info_wrap > div >
  form > div.pclist_table.mt20 > div > ul > li.pclist_list_tit42")
9 titles <-html_text(title_nodes)
10
11 titles <- gsub("[\t\r\n]", "",titles)
12 titles <- trimws(titles, which = "left")
13 titles <- trimws(titles, which = "right")
```

| | v1 |
|----|-------------------------|
| 1 | 가로수 가지치기와 미세먼지 |
| 2 | 버스 민원입니다. |
| 3 | 안 타는 쓰레기봉투 5L 만들어주세요... |
| 4 | 현릉로 지하철 요청 |
| 5 | 노후경유차 과태료 유예는 언제까지? |
| 6 | "서울시 제3화장장" 건립을 건의 드... |
| 7 | 시각장애인 콜택시 확대 운영부탁합니다 |
| 8 | 구로구에 한국산업박물관을 유치해 주세... |
| 9 | 청년스타트업을 위한 마케팅 지원 관련... |
| 10 | 덜컹한차량 배출5등급차량 선정에 불만... |

웹 크롤링 시 에러

- User-agent 추가

- 클라이언트 정보 : 웹 브라우저, 운영체제 등
- `d <- GET(url, query=list(params), user_agent(agent="agent"))`

- Referer 추가

- 이전 페이지의 URI 정보
- `d <- GET(url, add_headers(referer="referer"))`

- 참고 자료 : 네이버 부동산 크롤링

- https://statklee.github.io/yonsei/code/Naver_Land_APT_List.R

추가 연습자료

- 나무위키 크롤링
 - <https://stat-and-news-by-daragon9.tistory.com/109>
- 네이버 증권 삼성전자 주가 크롤링
 - <https://stat-and-news-by-daragon9.tistory.com/107>
- 네이버 영화 리뷰 크롤링
 - <https://stat-and-news-by-daragon9.tistory.com/104>
- 다음 영화 리뷰 크롤링
 - <https://stat-and-news-by-daragon9.tistory.com/103>

참고 내용

- R 파이프 연산 (체인 연산)
 - `a %>% c(5) %>% c(7) %>% sum`

Rselenium

- RSelenium + PhantomJS



<https://semaphoreci.com/blog/2018/03/27/phantomjs-is-dead-use-chrome-headless-in-continuous-integration.html>

Rselenium 설치

- Java (JDK)
- Selenium Server Standalone
 - <https://selenium-release.storage.googleapis.com/index.html>
- 크롬 드라이버 설치
 - <https://sites.google.com/a/chromium.org/chromedriver/>
 - `java -Dwebdriver.chrome.driver=chromedriver.exe -jar selenium-server-standalone-3.9.1.jar`
- Rselenium 라이브러리 설치

서버 실행화면

```
20:22:17.167 INFO - Selenium build info: version: '3.9.1', revision: '63f7b50'
20:22:17.168 INFO - Launching a standalone Selenium Server on port 4444
2019-05-04 20:22:17.243:INFO::main: Logging initialized @364ms to org.seleniumhq
.jetty9.util.log.StdErrLog
2019-05-04 20:22:17.351:INFO:osjs.Server:main: jetty-9.4.7.v20170914, build time
stamp: 2017-11-22T06:27:37+09:00, git hash: 82b8fb23f757335bb3329d540ce37a2a2615
f0a8
2019-05-04 20:22:17.373:WARN:osjs.SecurityHandler:main: ServletContext@o.s.j.s.S
ervletContextHandler@731a74c{/,null,STARTING} has uncovered http methods for pat
h: /
2019-05-04 20:22:17.378:INFO:osjs.ContextHandler:main: Started o.s.j.s.ServletC
ontextHandler@731a74c{/,null,AVAILABLE}
2019-05-04 20:22:17.727:INFO:osjs.AbstractConnector:main: Started ServerConnecto
r@4d41cee{HTTP/1.1,[http/1.1]}{0.0.0.0:4444}
2019-05-04 20:22:17.728:INFO:osjs.Server:main: Started @849ms
20:22:17.728 INFO - Selenium Server is up and running on port 4444
```

Rselenium 크롬 드라이버 설정

```
library(httr)
library(rvest)
library(RSelenium)

remDrv <- remoteDriver(remoteServerAddr = "localhost",
  port = 4444L, browserName = "chrome")

remDrv$open()
```

```
> remDrv$open()
[1] "Connecting to remote server"
$mobileEmulationEnabled
[1] FALSE

$timeouts
$timeouts$implicit
[1] 0

$timeouts$pageLoad
[1] 300000

$timeouts$script
[1] 30000

$hasTouchScreen
[1] FALSE

$platform
[1] "windows NT"
```

Rselenium 주요 함수

- `remDrv$open()` : 웹 브라우저 열기
- `remDrv$close()` : 웹 브라우저 닫기
- `remDrv$closeWindow()` : 현재 창 닫기
- `remDrv$navigate(url)` : 웹사이트 접속
- `d <- remDrv$getPageSource()` : 페이지 내용 가져오기

Rselenium 주요 함수

- `e <- findElement(using='xpath', value="값")` : Xpath 요소 얻기
- `e <- findElement(using='css' ,value="값")` : CSS 요소 얻기
- `e$sendKeysToElement(sendKeys="값")` : 입력하기
- `btn$clickElement()` : 버튼 클릭

교보문고 장바구니 책목록 출력

- 웹사이트

- 로그인

- <https://www.kyobobook.co.kr/login/login.laf>

- 책 목록

- <http://order.kyobobook.co.kr/cart/cartListMain>

교보문고 장바구니 책목록 출력

```
library(httr)
library(rvest)
library(RSelenium)

remDrv <- remoteDriver(remoteServerAddr = "localhost",
  port = 4444L, browserName = "chrome")

remDrv$open()

remDrv$navigate(url = 'https://www.kyobobook.co.kr/login
/login.laf')

# 아이디, 비밀번호 입력창 찾기
id <- remDrv$findElement(using = 'css', value = '#memid')
pw <- remDrv$findElement(using = 'css', value = '#pw')

# 아이디, 비밀번호 입력
id$sendKeysToElement(sendKeys = list('eodyds1a'))
pw$sendKeysToElement(sendKeys = list('eodyds1a'))
```

교보문고 장바구니 책목록 출력

```
# 버튼
btn <- remDrv$findElement(using = 'css', value = '#login_zone1 > p:nth
-child(2) > input')

# 버튼 클릭
btn$clickElement()

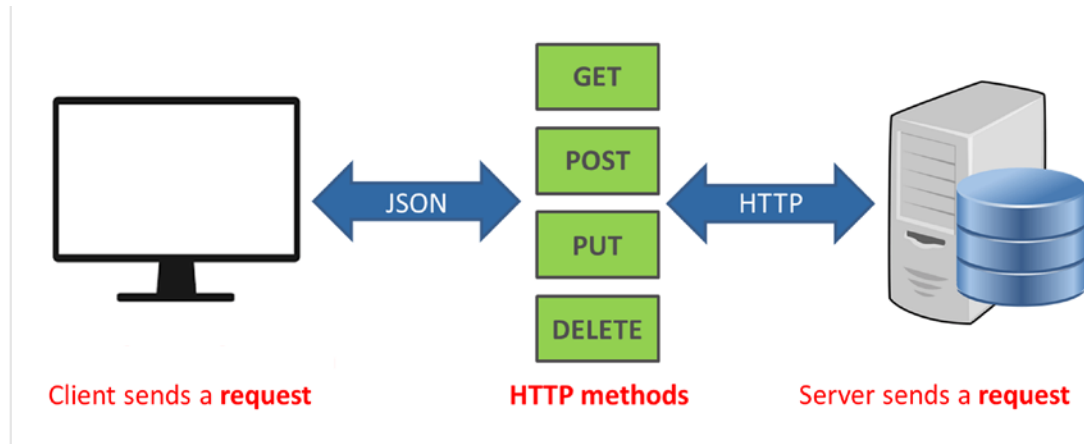
# 장바구니 페이지로 이동

remDrv$navigate(url = 'http://order.kyobobook.co.kr/cart/cartListMain'
)

d <- remDrv$getPageSource()[[1]]
html <- read_html(d,encoding="UTF-8")
book_nodes <- html_nodes(html,css = '#cartFrm > table > tbody > tr >
td.align_left.ver_top > div > a:nth-child(2) > span')
book_texts <- html_text(book_nodes)
```

OPENAPI 활용 크롤링 예시

- REST API(Representational State Transfer)



- 업비트 개발자 센터
 - <https://docs.upbit.com/>

참고 : JSON 포맷

예제 [편집]

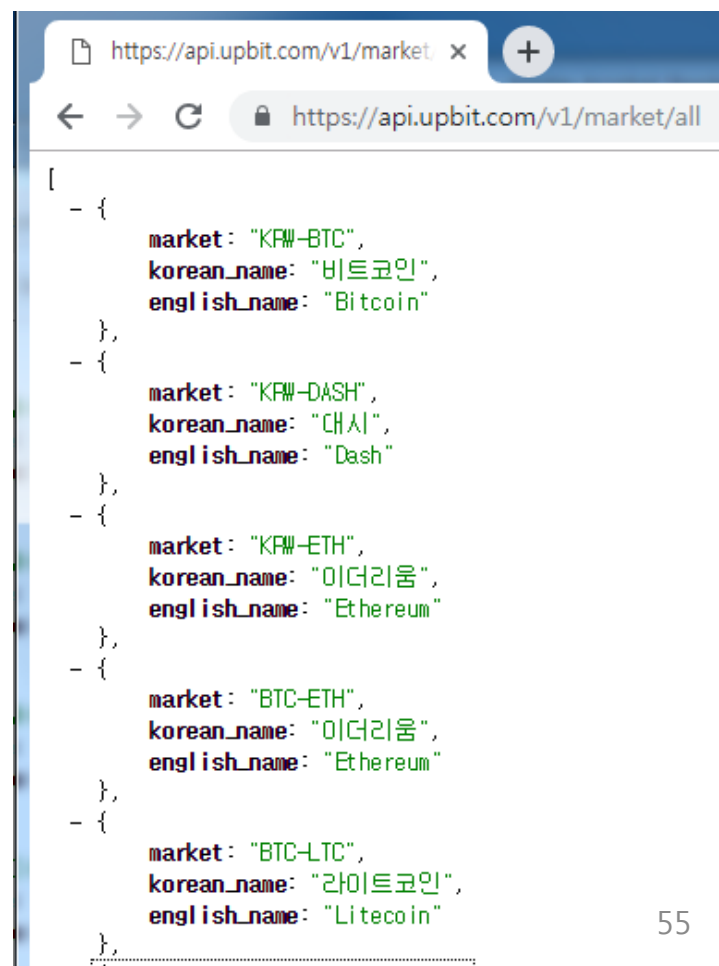
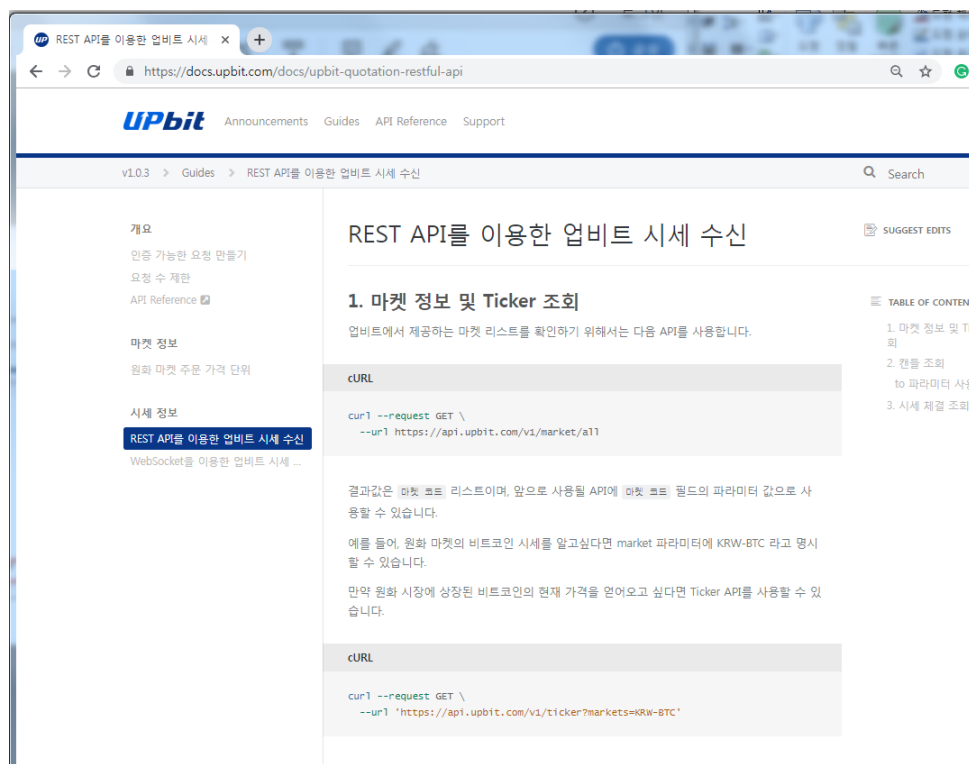
다음은 한 사람에 관한 정보를 갖는 JSON 객체이다.

키-값 쌍(이름:값)의 패턴으로 표현된다.

```
1 {  
2   "이름": "홍길동",  
3   "나이": 25,  
4   "성별": "여",  
5   "주소": "서울특별시 양천구 목동",  
6   "특기": ["농구", "도술"],  
7   "가족관계": {"#": 2, "아버지": "홍판서", "어머니": "춘섬"},  
8   "회사": "경기 수원시 팔달구 우만동"  
9 }
```

OPENAPI 활용 크롤링 예시

- 암호화폐 가격 수집



OPENAPI 활용 크롤링 예시

```
1 # 필요한 패키지를 불러옵니다.
2 library(httr)
3 library(rvest)
4 library(jsonlite)
5
6 # 업비트 거래 암호화폐 출력
7 res <- GET(url = 'https://api.upbit.com/v1/market/all')
8
9 # 응답 결과
10 coinList <- fromJSON(content(res, as = 'text'))
11
12 # 코인 선택
13 coinName <- '비트코인'
14
15 # 코인 코드 얻기
16 coinList[coinList$korean_name == coinName, 'market']
17
18 # 관심 코인 조회
19 res <- GET(url = 'https://api.upbit.com/v1/ticker', query = list(markets = 'KRW-BTC'))
20 coinInfo <- fromJSON(content(res, as = 'text'))
21 coinInfo$trade_price
```


참고자료

- R로 웹 데이터를 가져오는 4가지 방법(은 크롤링), https://tidyverse-korea.github.io/r-meetup-x-presser/kaggle/Meetup_3/crawling/getWebR.pdf
- 최대한 친절하게 쓴 R 크롤러 만들기, <https://kuduz.tistory.com/1041>
- R 웹 크롤링 기초, https://statklee.github.io/yonsei/data/R_Web_Crawling.pdf
- 패스트 캠퍼스 R 웹크롤링, <https://github.com/j2hoon85/FastCampus>