

R 패키지

- R 패키지라는 것은 R를 가지고 할 수 있는 통계, 분석 그리고 시각화와 관련하여 기능을 정의한 함수들의 묶음이라 할 수 있다.
- R 패키지는 R을 설치할 때 함께 설치되는 기본 패키지가 있고 만약 찾는 기능이 없다면 원하는 기능을 처리해주는 패키지를 찾아서 추가로 설치 한 후 사용하면 된다.
- R 패키지는 CRAN(<https://cran.r-project.org/>) 사이트에서 모두 검색 가능하고 다운로드 받을 수 있다.
- R은 무료라는 장점 외에 일정 규칙에 맞춰 누구나 제작하고 배포할 수 있는 Package를 통해 기능 확장을 유연하게 할 수 있는 큰 장점을 갖고 있다.

<https://cloud.r-project.org>

Contributed Packages

Available Packages

Currently, the CRAN package repository features 13445 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this [Installation and Administration](#) (also contained in the R base sources) explains the process in detail.

[CRAN Task Views](#) allow you to browse packages by topic and provide tools to automatically install all pack-

R 패키지

- 새로운 R 패키지의 설치
`install.packages("패키지명")`
- 이미 설치된 R 패키지 확인
`installed.packages()`
- 설치된 패키지 삭제
`remove.packages("패키지명")`
- 설치된 패키지의 버전 확인
`packageVersion("패키지명")`
- 설치된 패키지 업데이트
`update.packages("패키지명")`
- 설치된 패키지 로드
`library(패키지명)`
`require(패키지명)`
- 로드된 패키지 언로드(로드상태 해제)
`detach("package:패키지명")`
- 로드된 패키지 점검
`search()`
- 설치된 패키지의 버전 점검
`packageVersion(패키지명)`

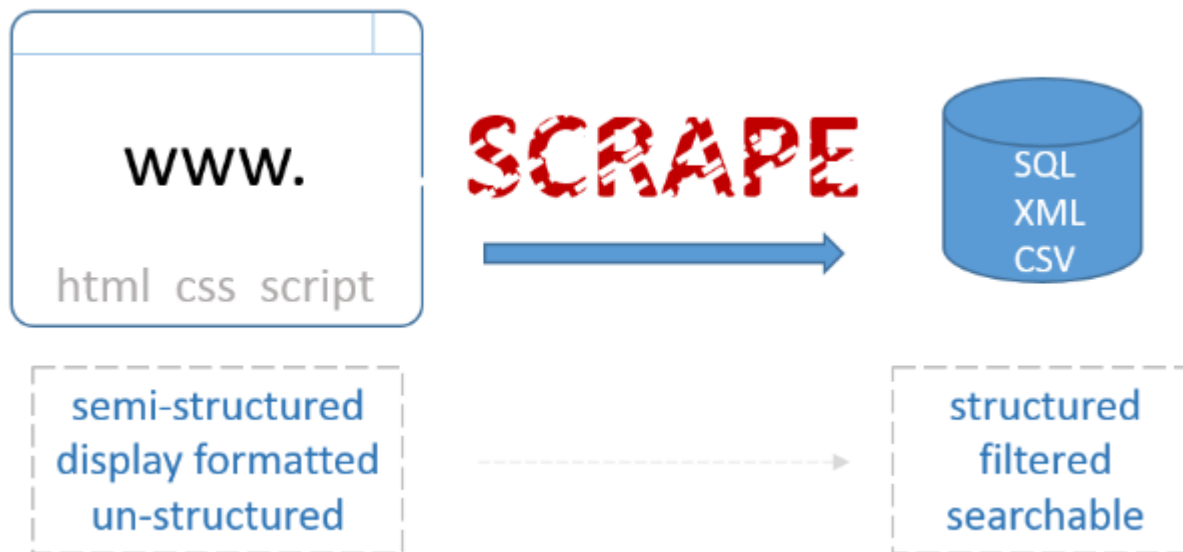
데이터 수집

[웹 스크래핑(web scraping)]

웹 사이트 상에서 원하는 부분에 위치한 정보를 컴퓨터로 하여금 자동으로 추출하여 수집하는 기술

[웹 크롤링(web crawling)]

자동화 봇(bot)인 웹 크롤러가 정해진 규칙에 따라 복수 개의 웹 페이지를 브라우징 하는 행위



스크래핑하려는 페이지에서 원하는 태그 찾기 – CSS Selector

Selectors

Basics

#id
element
.class,
.class.class
*
selector1,
selector2

Hierarchy

ancestor
descendant
parent > child
prev + next
prev ~ siblings

Basic Filters

:first
:last
:not(selector)
:even
:odd
:eq(index)
:gt(index)
:lt(index)

Content Filters

:contains(text)
:empty
:has(selector)
:parent

Visibility Filters

:hidden
:visible

Child Filters

:nth-child(expr)
:first-child
:last-child
:only-child

Attribute Filters

[attribute]
[attribute=value]
[attribute!=value]
[attribute^=value]
[attribute\$=value]
[attribute*=value]
[attribute|=value]
[attribute~=value]
[attribute]
[attribute2]

Forms

:input
:text
:password
:radio
:checkbox
:submit
:image
:reset
:button
:file

Form Filters

:enabled
:disabled
:checked
:selected

데이터 수집

스크래핑하려는 페이지에서 원하는 태그 찾기 - XPath

XPath(XML Path Language)는 W3C의 표준으로 확장 생성 언어 문서의 구조를 통해 경로 위에 지정한 구문을 사용하여 항목을 배치하고 처리하는 방법을 기술하는 언어이다.

```
<wikimedia>
  <projects>
    <project name="Wikipedia" launch="2001-01-05">
      <editions>
        <edition language="English">en.wikipedia.org</edition>
        <edition language="German">de.wikipedia.org</edition>
        <edition language="French">fr.wikipedia.org</edition>
        <edition language="Polish">pl.wikipedia.org</edition>
      </editions>
    </project>
    <project name="Wiktionary" launch="2002-12-12">
      <editions>
        <edition language="English">en.wiktionary.org</edition>
        <edition language="French">fr.wiktionary.org</edition>
        <edition language="Vietnamese">vi.wiktionary.org</edition>
        <edition language="Turkish">tr.wiktionary.org</edition>
      </editions>
    </project>
  </projects>
</wikimedia>
```

아래의 XPath 식은

```
/wikimedia/projects/project/@name
```

모든 project 요소의 name 속성을 선택하고, 아래의 XPath 식은

```
/wikimedia/projects/project/editions/edition[@language="English"]/text()
```

모든 영문 Wikimedia 프로젝트의 주소(language 속성이 English인 모든 edition 요소의 문자열)를 선택하고, 아래의 XPath 식은

```
/wikimedia/projects/project[@name="Wikipedia"]/editions/edition/text()
```

모든 위키백과의 주소(Wikipedia의 이름 특성을 가진 project 요소 아래에 존재하는 모든 edition 요소의 문자열)를 선택한다.

데이터 수집

[정적 웹 페이지에 스크래핑에 사용되는 주요 API]

- xml2 패키지

`read_html(url)` : HTML 웹 페이지를 요청해서 받아오기

- rvest 패키지

`html_nodes(x, css, xpath)`, `html_node(x, css, xpath)` : 원하는 노드(태그) 추출하기

`html_text(x, trim=FALSE)` : 노드에서 콘텐츠 추출하기

`html_attr(x)` : 노드에서 속성들 추출하기

`html_attr(x, name, default = "")` : 노드에서 주어진 명칭의 속성값 추출하기

- XML 패키지

`htmlParse (file, encoding="...")` : `xpathSApply()` 사용 가능한 객체로 변환

`xpathSApply(doc, path, fun)` : 원하는 노드(태그) 추출하고 전달된 함수 수행하기

fun : `xmlValue`, `xmlGetAttr`, `xmlAttrs`

- httr 패키지

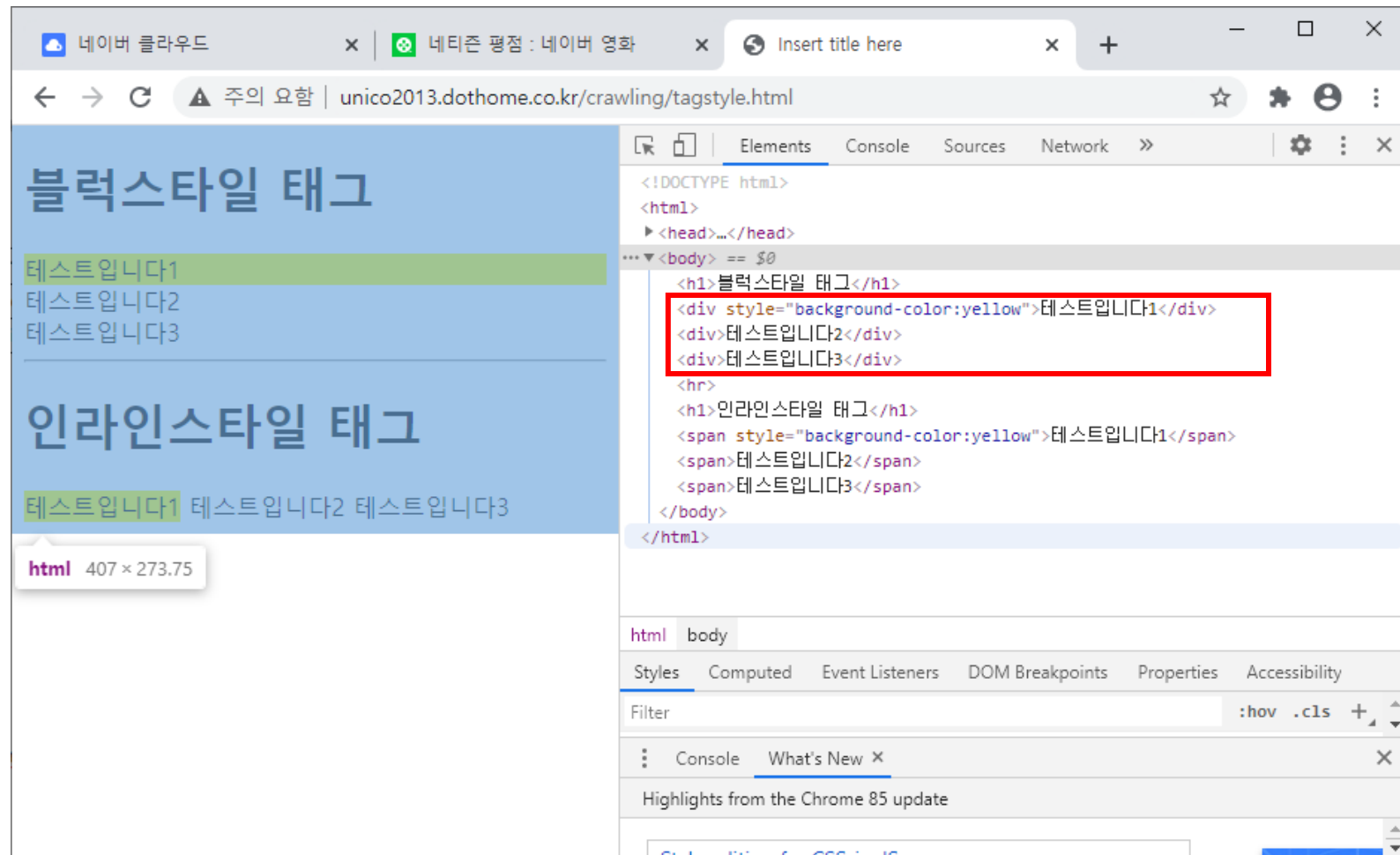
`GET(url)` : HTML 웹 페이지를 요청해서 받아오기

요청헤더에 계정 또는 패스워드 등의 정보 전달 가능

응답 내용이 바이너리인 경우에도 사용 가능

데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)



데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)

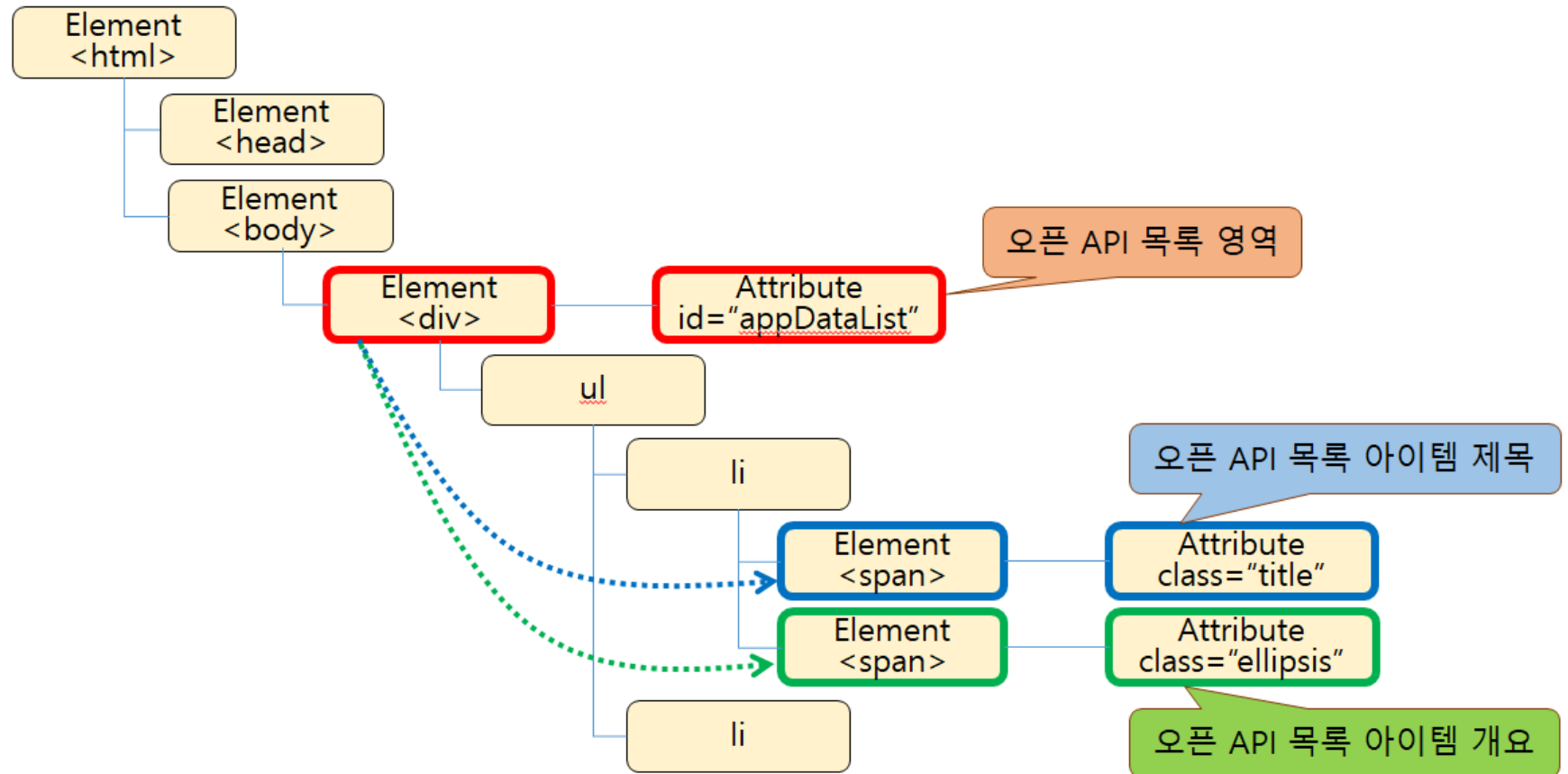
The screenshot shows a web browser window with the URL `data.go.kr/tcs/dss/selectDataSetList.do`. The page displays a list of datasets under the heading "표준데이터셋 122개 (10,158건)". The datasets listed include:

- 외교부_국가·지역별 주요인사교류** (JSON format, provided by 외교부, updated 2021-03-02, 81 downloads, 1 application)
- 한국국제협력단_파견국 안전이슈 월력표** (JSON format, provided by 한국국제협력단, updated 2021-03-02, 9 downloads, 1 application)
- 한국국제협력단_일일 안전관리 중점국가** (JSON format, provided by 한국국제협력단, updated 2021-03-02, 8 downloads, 1 application)
- 수협 산지조합 위판장 정보** (XML/JSON format, provided by 해양수산부, updated 2021-02-26, 3654 downloads, 47 applications)
- 수협 산지조합 정보** (XML format, provided by 해양수산부, updated 2021-02-26, 3504 downloads, 49 applications)

The developer tools on the right show the DOM structure. A red box highlights the `<div class="data-set-list open-api dTypeList" id="apiDataList">` element. Another red box highlights a specific data item's title tag: ` == $0`. The text "교부_국가·지역별 주요인사교류" is visible below the title tag.

데이터 수집

스크래핑할 태그 정보 얻기(DOM 객체의 구조)



데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)

The screenshot shows the data.go.kr website with the Chrome DevTools developer tools open. The website displays a list of datasets. The second dataset, '외교부_국가·지역별 주요인사교류' (Ministry of Foreign Affairs - Main Personnel Exchange by Country/Region), is selected. The DevTools 'Elements' panel shows the HTML structure of the selected dataset's description. A red box highlights the following HTML snippet:

```
<dd class="ellipsis">
  "1. 국가·지역별 주요인사교류 목록 조회: 한글 국가명 또는 ISO
  국가코드 (다.참고 1 ISO국가코드 이용 )를 이용하여 국가·지역별
  주요인사교류 목록 조회</br>
  - 인사교류,국가교류,해외인사,방한,방문 등 정보를 제공 하는
  공공데이터 API 서비스</br> 제공되는 데이터는 2020년 12월 31
  일 기준입니다. "
</dd>
```

The website interface includes filters for '통일외교안보' and '국가행정기관', and a 'JSON' icon. The dataset details show it was provided by the Ministry of Foreign Affairs, last updated on 2021-03-02, with 81 records and 1 active application. A blue circular badge at the bottom right of the website says '의견수렴 게시판' (Opinion Collection Board).

데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)

The screenshot shows a web browser with the Naver movie review page open. The page displays a list of movie reviews with columns for rank, number, title, and score. The right side of the browser shows the Chrome DevTools 'Elements' panel, which is expanded to show the HTML structure of the page. A red box highlights the HTML code for a specific review item, showing the title, a link to the movie page, and the score.

NAVER 영화

영화홈
상영작 · 예정작
영화랭킹
예매
평점 · 리뷰
네티즌 평점
네티즌 리뷰
다운로드
인디극장

네티즌·관람객 평점

현재 상영작 평점보기
현재 상영작

전체 리스트 총 13,425건

번호	감상평
17419423	소울 ★★★★★ 10 일단 보세요 신고
17419422	소울 ★★★★★ 10 나중에 나에게 아이가 생긴다면, 꼭 같이 보고 싶은 영화 신고
17419421	승리호 ★★★★★ 10 넷플릭스에서 봤는데 영화관에서 안본게 후회될 정도로 너무 재밌게 봤 역들의 연기와 승리호의 주연을 케미가 장난 아니네요 연출력, 스토리 탄해서 시간가는줄 몰랐어요 진짜ㅠㅠ꽃님이 순이 꽃길만 걸었으면♡
17419420	극장판 귀멸의 칼날: 무한열차편 ★★★★★ 10 미쳤어요 귀칼애니부터 진짜영상미오짐 제발빨리2기방영부탁드립니다 나요 신고
17419419	미션 파서블 ★★★★★ 10 배우분들 연기를 잘해서 양화도 재밌습시다 신고
17419418	벵가드 ★★★★★ 1 벵가드 때문에 컴퓨터 존나 끊길 안티핵 프로그램이랍시고 컴퓨터 말아 있나 신고
17419417	추격자 ★★★★★ 6 여자가 숨어있는 가계에 타이밍 맞춰 풀려나 범인이 딱 찾아 들어간 것 현실성 떨어진 형식이 더 수성감을 때드 인이! 그럴 수도 있겠지만...

Elements

```
<!-- list -->
<table cellpadding="0" class="list_netizen">
  <caption class="blind">개봉 후 평점 목록으로 번호, 평점, 내용, 글쓴이, 날짜
  정보를 제공</caption>
  <colgroup>...</colgroup>
  <thead>...</thead>
  <tbody>
    <tr>
      <td class="sc_num">17419423</td>
      <td class="title">
        <a href="?st=mcode&sword=184517&target=after" class="movie_color_b">
          소울</a> == $0
        <div class="list_netizen_score">
          <span class="st_off">...</span>
          <em>10</em>
        </div>
      </td>
    </tr>
  </tbody>
</table>
```

... out.old_community div#cbody.type_1 div#old_content table.list_netizen tbody tr td.title a.movie. ...

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

Console What's New

Highlights from the Chrome 89 update

Debugging support for Trusted Type violations

데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)

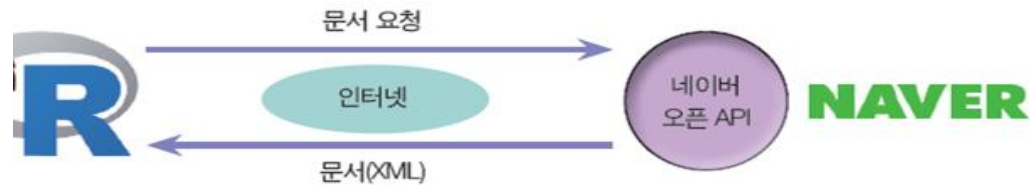
The screenshot shows a web browser with a news article on the left and the Chrome DevTools developer tools on the right. The article is titled "국토부, 광명시흥 100억대 투기 의혹에 'LH 직원 전수조사'" (Ministry of Land, Infrastructure and Transport, 100 billion won speculation in Gwangmyeong-Siheung, 'LH employee exhaustive investigation'). The article text mentions that LH employees involved in the speculation of 100 billion won in Gwangmyeong-Siheung are being investigated. The developer tools on the right show the HTML structure of the article, with the following code highlighted in a red box:

```
<div id="main-top01-scroll-start" style="height: 9931px;">
  <div id="main-top01-scroll-in" style="position: relative; top: 0px; left: 0px;">
    <div class="article01 first">
      <div class="article-area">
        <h4 class="article-title">
          <a href="/arti/society/society_general/985147.html?_fr=mt2">국토부,
            광명시흥 100억대 투기 의혹에 "LH 직원 전수조사" </a> == $0
        </h4>
        <span class="article-photo">...</span>
        <p class="article-prologue">...</p>
      </div>
    </div>
  </div>
</div>
```

The browser's address bar shows the URL: hani.co.kr. The page title is "국토부, 광명시흥 100억대 투기 의혹에 'LH 직원 전수조사'". The page content includes a large image of a person in a suit, a smaller image of a person in a suit, and a section titled "연봉 5억 짜인 양의자...왜?". The developer tools show the HTML structure of the article, with the following code highlighted in a red box:

데이터 수집

네이버의 뉴스와 블로그 글 읽어오기



<https://developers.naver.com/docs/search/blog/> 에서 내용 검토

API	요청	출력 포맷
뉴스	https://openapi.naver.com/v1/search/news.xml	XML
블로그	https://openapi.naver.com/v1/search/blog.xml	XML

요청 변수	값	설명
query	문자(필수)	검색을 원하는 질의, UTF-8 인코딩
display	정수: 기본값 10, 최대 100	검색 결과의 출력 건수(최대 100까지 가능)
start	정수: 기본값 1, 최대 1000	검색의 시작 위치(최대 1000까지 가능)
sort	문자: date(기본값), sim	정렬 옵션 • date : 날짜순(기본값) • sim : 유사도순

데이터 수집

트위터 글 읽어오기

트위터에서는 rtweet 이라는 패키지를 제공하여 트위터에 올려진 글을 수집하는데 도움을 준다.

```
install.packages("rtweet")
```

```
library(rtweet)
```

```
appname <- "edu_data_collection"
```

```
api_key <- "..."; api_secret <- "..."; access_token <- "..."; access_token_secret <- "..."
```

```
twitter_token <- create_token(
```

```
  app = appname,
```

```
  consumer_key = api_key,
```

```
  consumer_secret = api_secret,
```

```
  access_token = access_token,
```

```
  access_secret = access_token_secret)
```

```
key <- "취업"; key <- enc2utf8(key)
```

```
result <- search_tweets(key, n=100, token = twitter_token)
```

```
str(result)
```

```
result$retweet_text
```

```
content <- result$retweet_text
```

```
content <- gsub("[[:lower:]][[:upper:]][[:digit:]][[:punct:]][[:cntrl:]]", "", content)
```

```
content
```

create_token(appname, api_key,api_secret, access_token,access_token_secret)	현재의 R세션에 인증토큰을 내려받는 기능
result<- search_tweets(key, n=100, token)	key 에 해당되는 트위터 글 읽어 오기

데이터 수집

공공DB 읽어오기

The image displays two screenshots of the data.go.kr website, which is the official portal for public data in South Korea.

Left Screenshot (Homepage):

- URL: `data.go.kr/index.do`
- Navigation Menu: 데이터찾기, 국가데이터맵, 데이터요청, 데이터활용, 정보공유, 이용안내
- Search Bar: 인기검색어 (Popular Searches)
- Form: 어떤 공공데이터를 찾으시나요? (Which public data are you looking for?)
- Filters: 검색조건 (Search Conditions) with dropdowns for 분류체계 (Classification System), 서비스유형 (Service Type), and 확장자 (Extension).

Right Screenshot (Developer Account Page):

- URL: `data.go.kr/lim/api/selectDevAccountList.do`
- Section: 마이페이지 (My Page) - 개발계정 (Developer Account)
- Statistics:

신청 0건 >	활용 13건 >	중지 0건 >
신청중인 단계 (Application in progress)	승인되어 활용중인 단계 (Approved and in use)	중지신청하여 운영이 중지된 단계 (Suspended and operation stopped)
· 보류: 0건 · 반려: 0건	· 변경신청: 0건	

Below the statistics, there is a list of developer accounts. The first entry is:

사회복지	한국사회복지협의회
활용신청 [승인] 한국사회복지협의회 전국푸드뱅크 정보	
신청일: 2021.02.04 관리종료일: 2023.02.04	

데이터 수집

공공DB 읽어오기

오픈API 상세

XML 노선정보조회 서비스
노선에 대한 정보 제공
👍 0 🗨️ 0 📄 관심

OpenAPI 정보

메타데이터 다운로드

분류체계	교통및물류 - 도로
관리부서명	교통정보과
API 유형	REST
활용신청	10483
등록	2011-12-03
심의유형	개발단계 : 허용 / 운영단계 : 허용
비용부과유무	무료
이용허락범위	출처표시
참고문서	서울특별시 노선정보조회 서비스

[상세기능](#)[활용사례](#)

오픈API 상세

XML 버스위치정보조회 서비스
실시간 버스위치 정보 제공
👍 5 🗨️ 0 📄 관심

OpenAPI 정보

메타데이터 다운로드

분류체계	교통및물류 - 도로	제공기관	서울특별시
관리부서명	교통정보과	관리부서 전화번호	02-2133-4969
API 유형	REST	데이터포맷	XML
활용신청	10096	키워드	
등록	2011-12-03	수정	2020-06-25
심의유형	개발단계 : 허용 / 운영단계 : 허용		
비용부과유무	무료		
이용허락범위	출처표시		
참고문서	서울특별시 버스위치정보조회 서비스 활용가이드 20190110.docx		