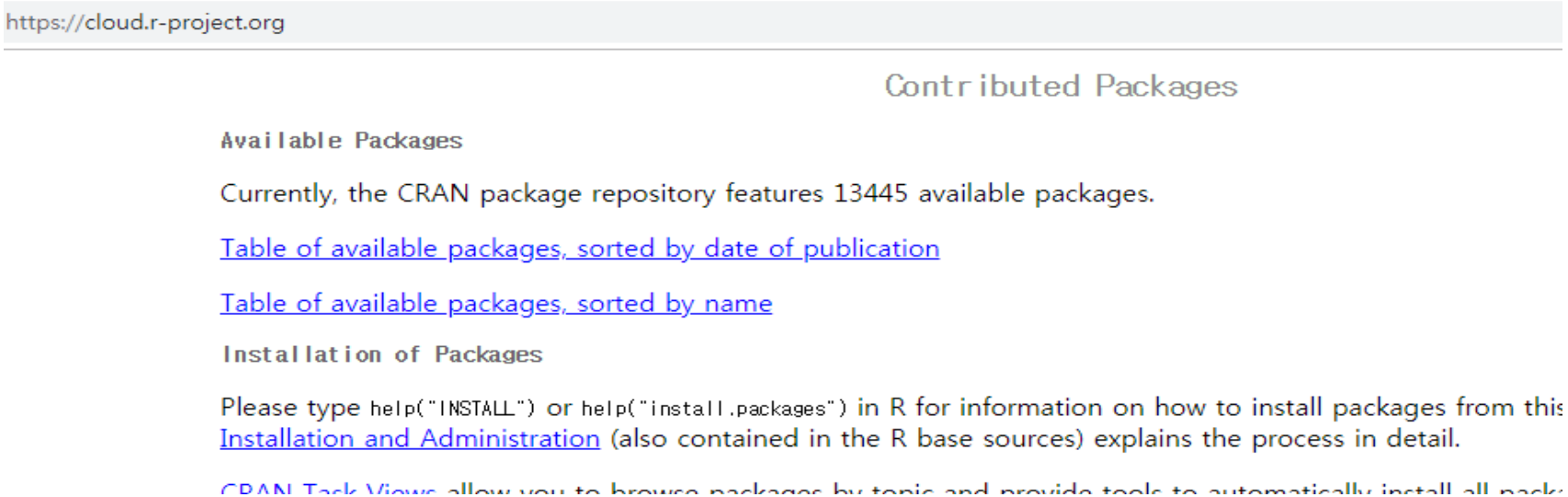
R 패키지

- 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.



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()

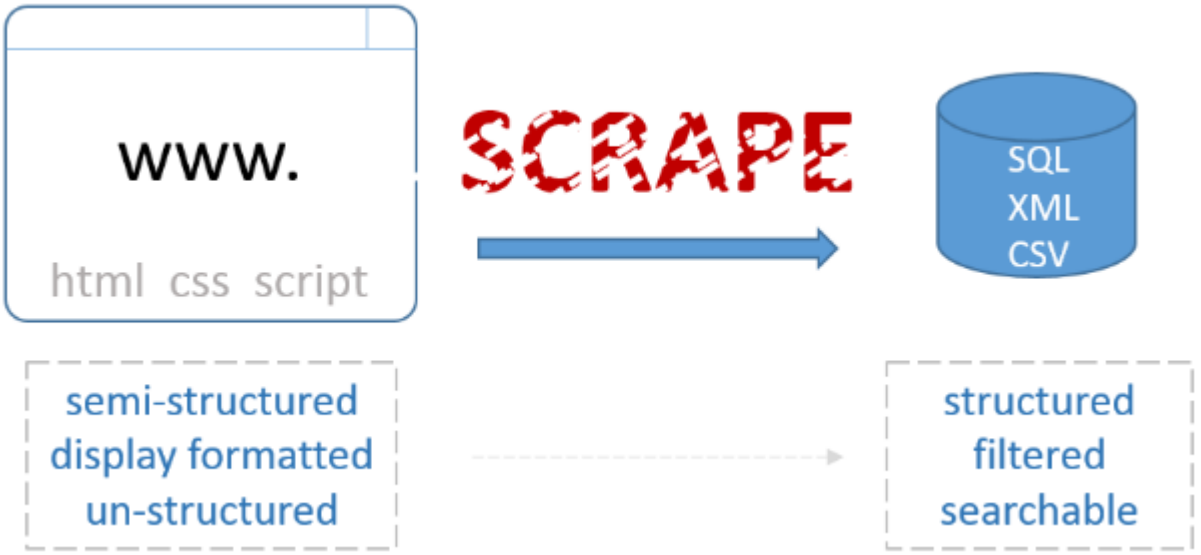
데이터 수집

[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



데이터 수집

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 *English*인 모든 `edition` 요소의 문자열)를 선택하고, 아래의 XPath 식은

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

모든 위키백과의 주소(*Wikipedia*의 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 컨텐트 추출하기

  html_attrs(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**

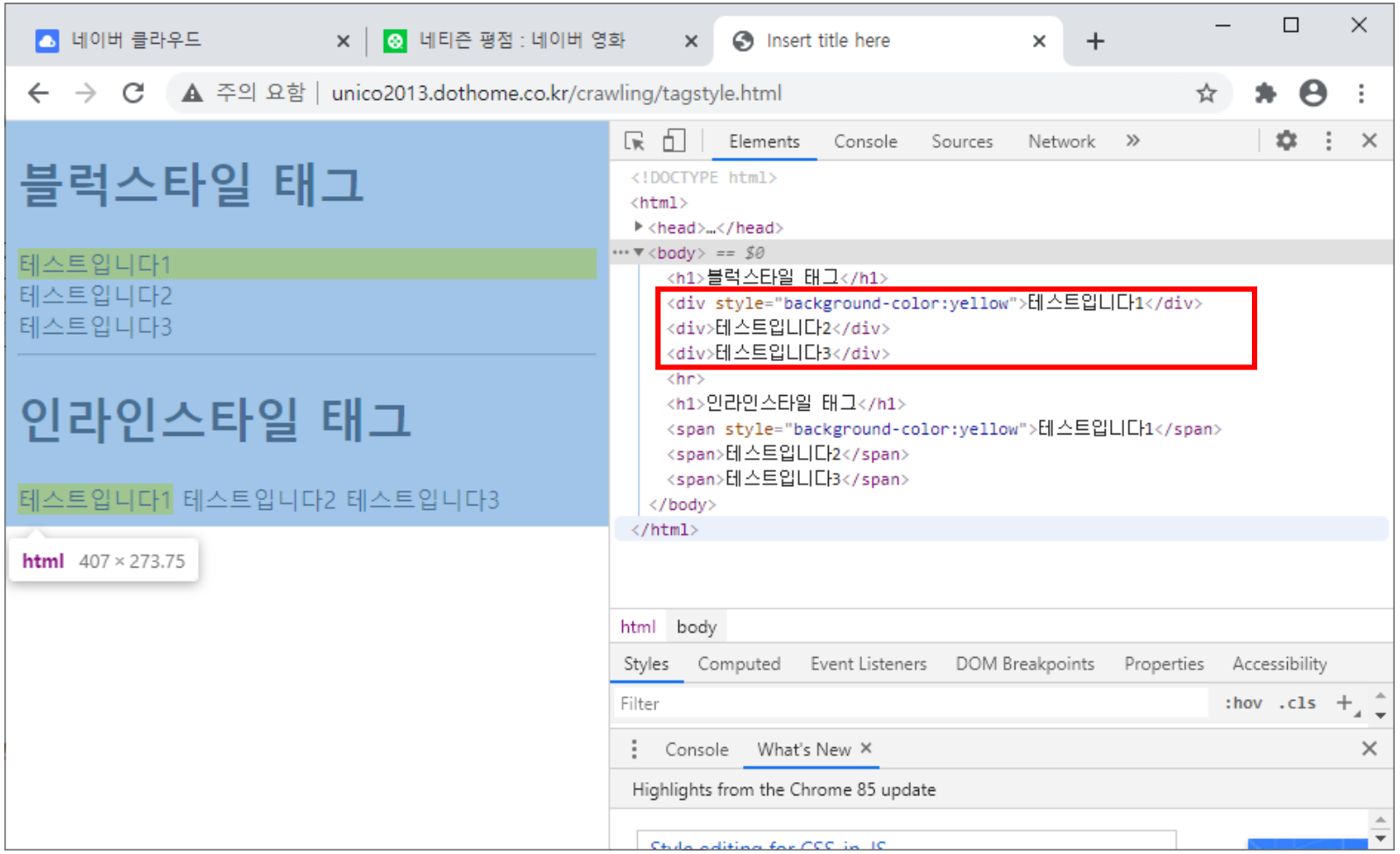
  GET(url) : HTML 웹 페이지를 요청해서 받아오기

  요청헤더에 계정 또는 패스워드 등의 정보 전달 가능

  응답 내용이 바이너리인 경우에도 사용 가능

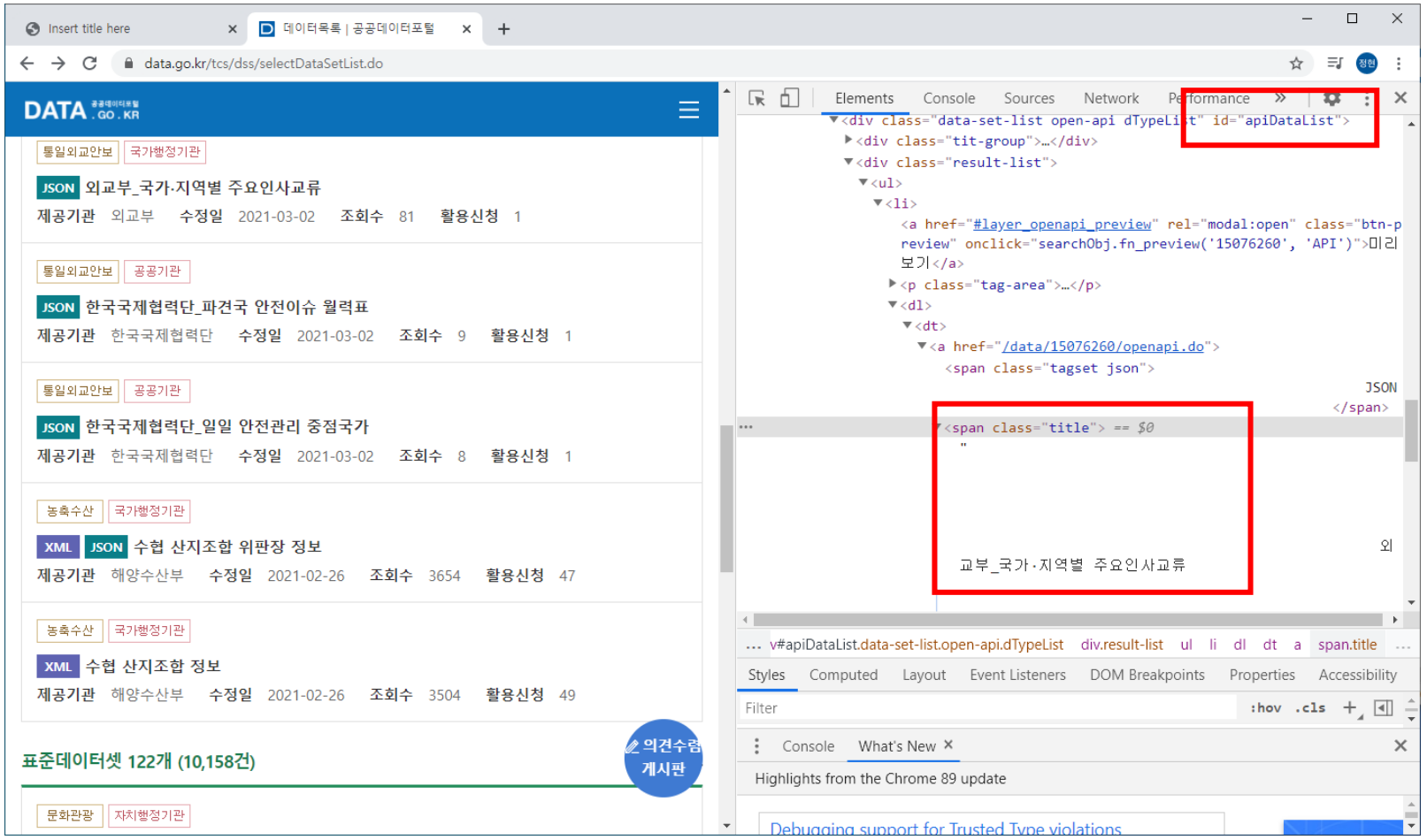
데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)



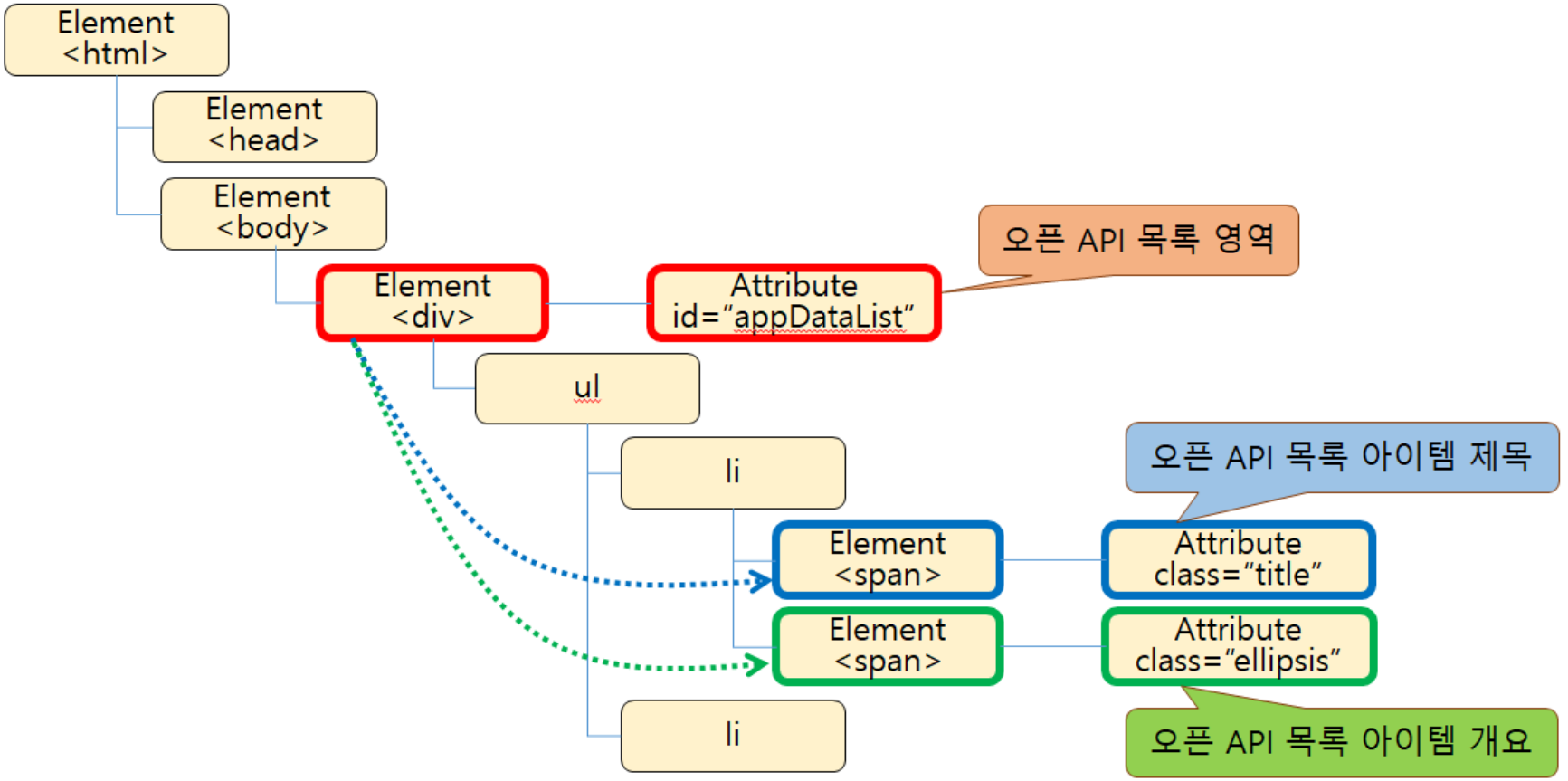
데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)



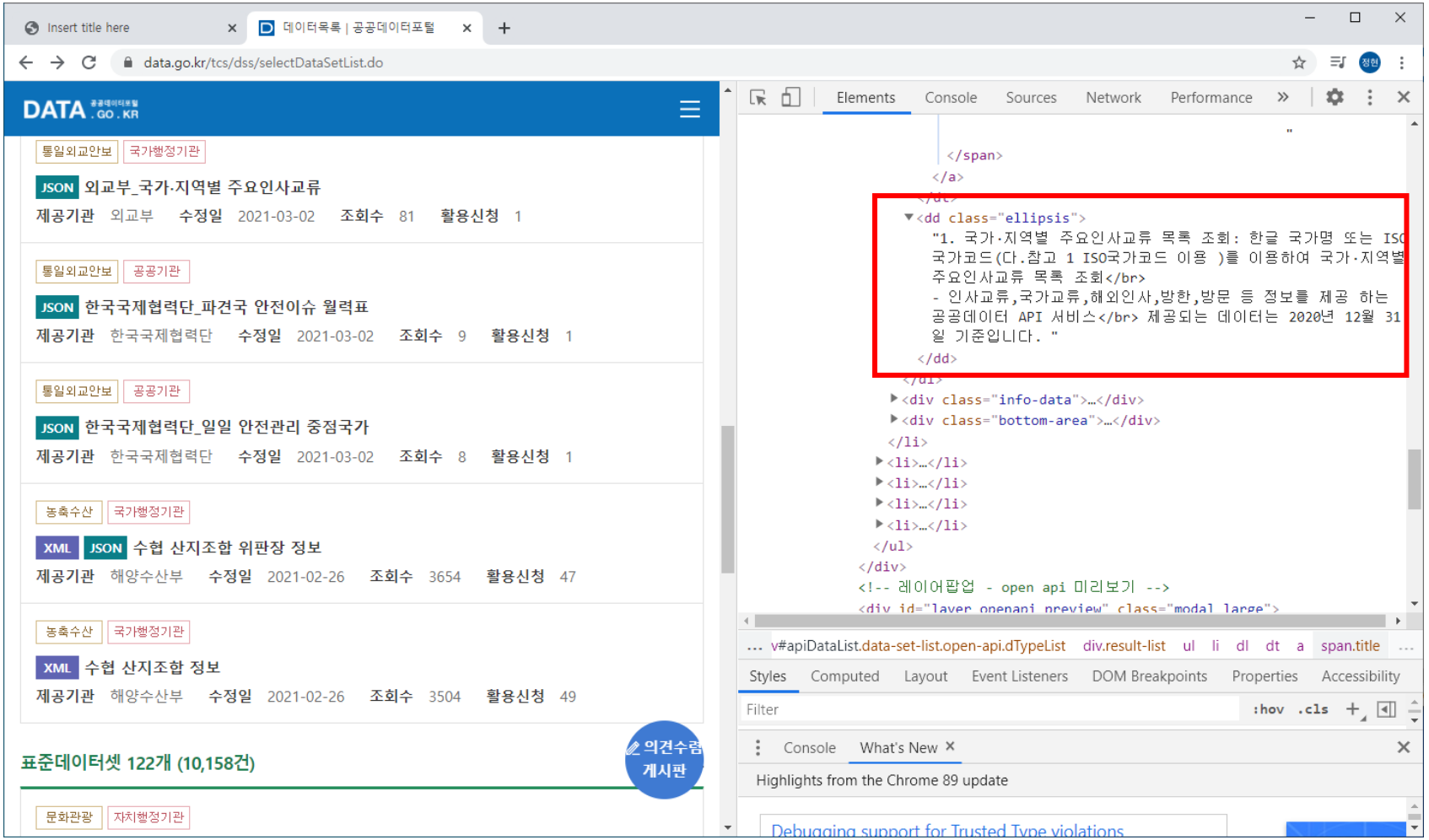
데이터 수집

스크래핑할 태그 정보 얻기(DOM 객체의 구조)



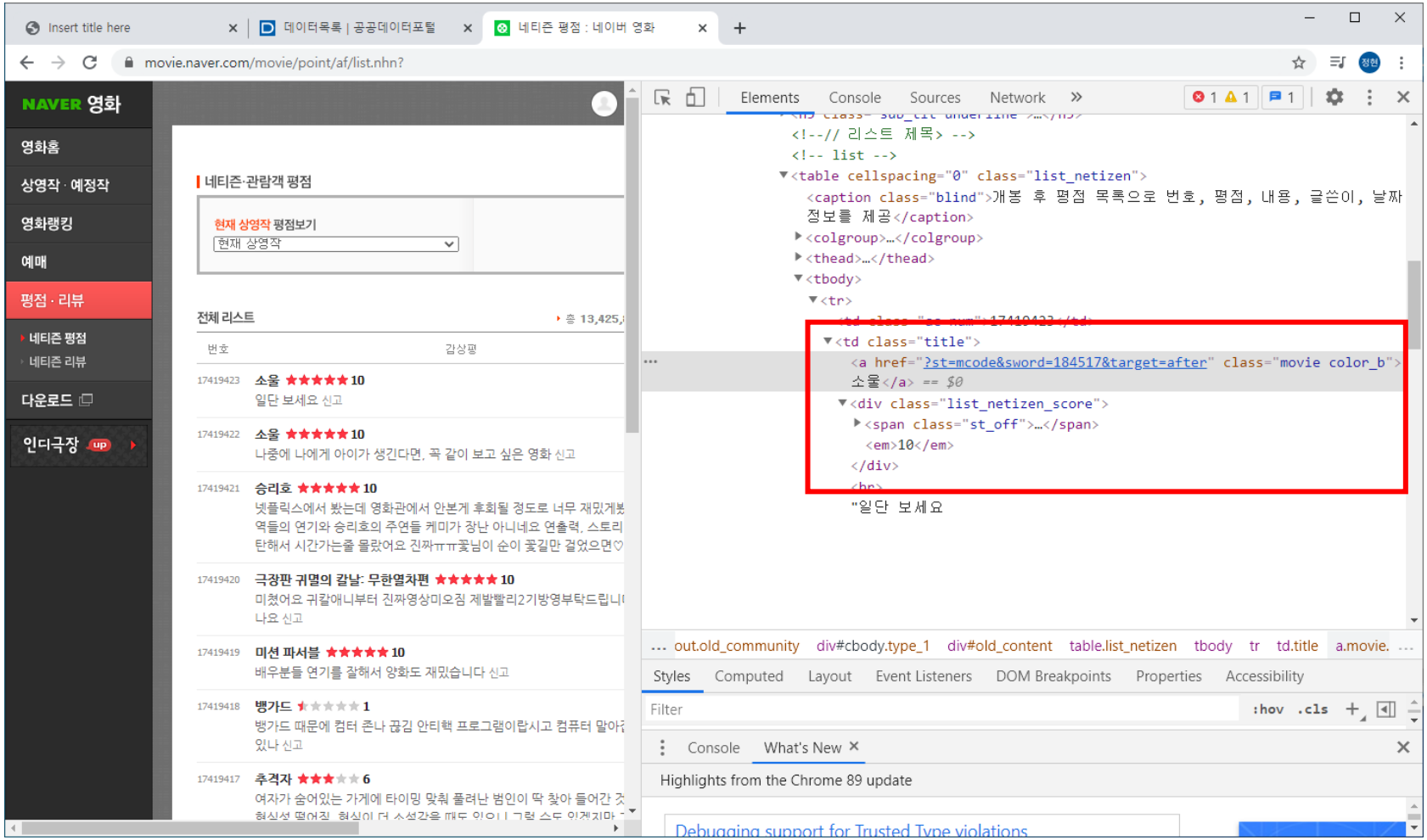
데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)



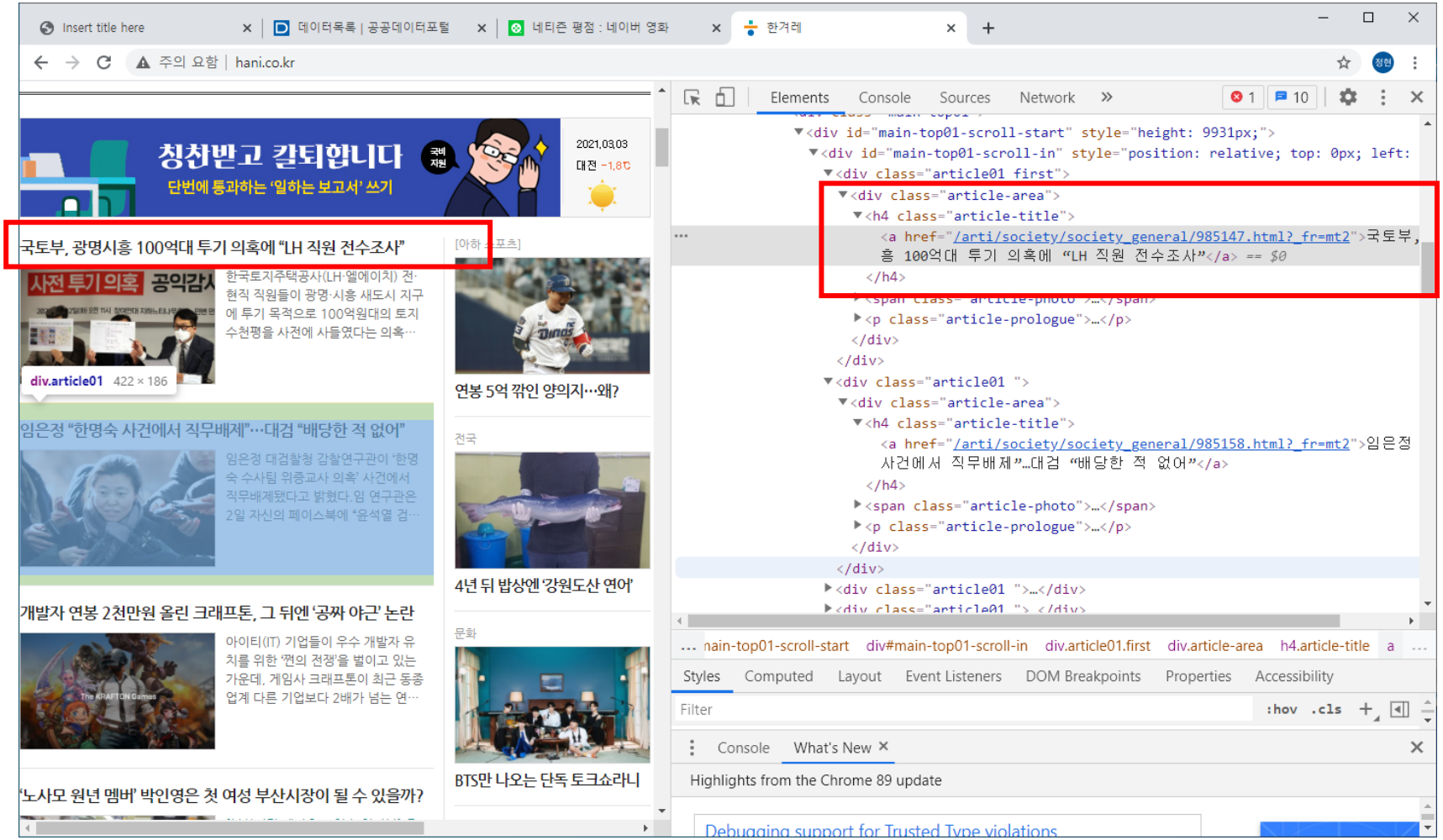
데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)



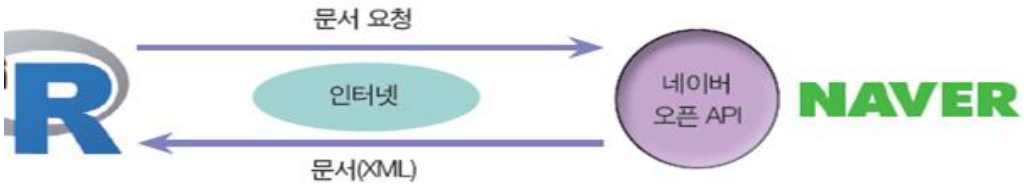
데이터 수집

스크래핑할 태그 정보 얻기(chrome의 개발자도구 활용)



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

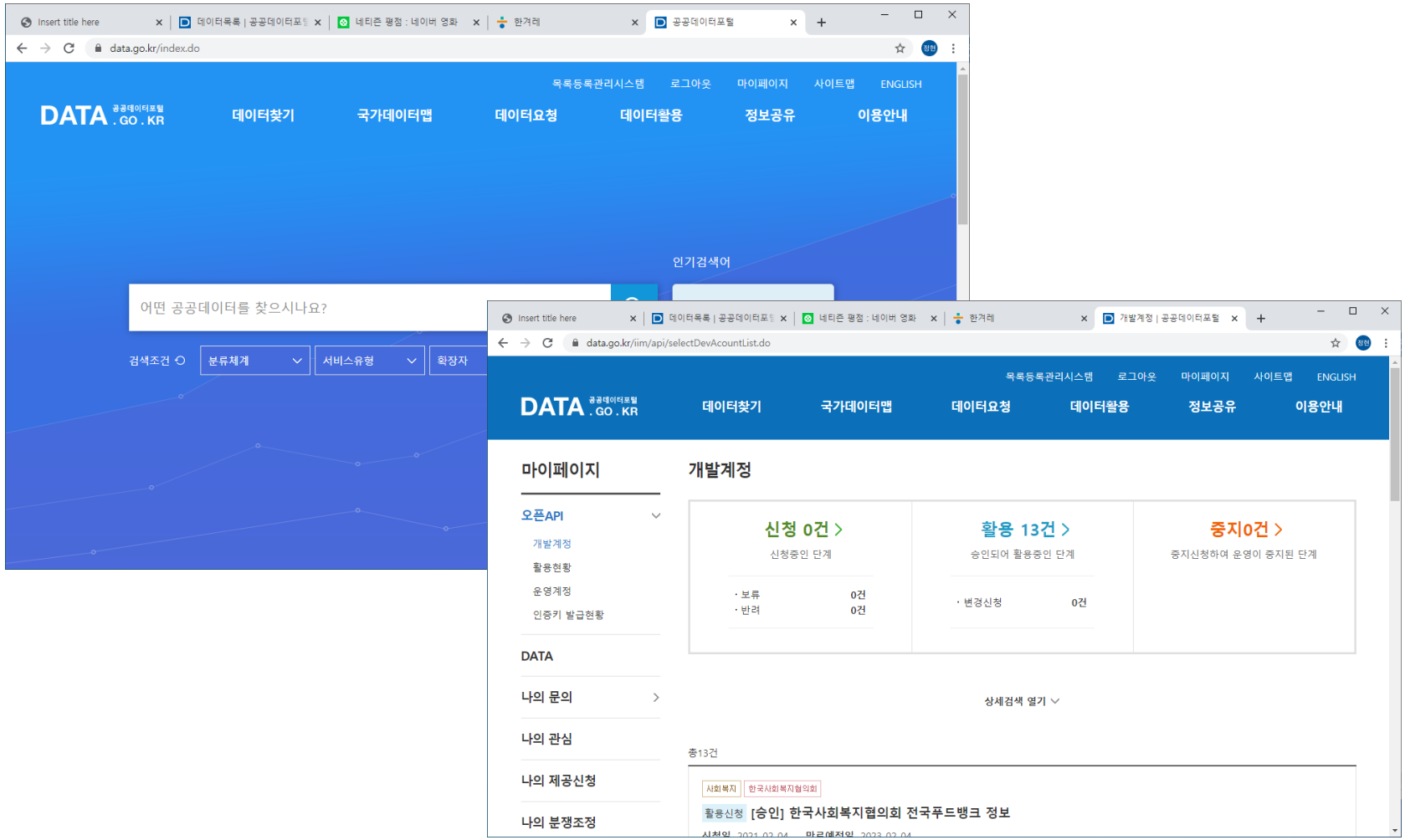
트위터에서는 rtweet 이라는 패키지를 제공하여 트위터에 올려진 글을 수집하는데 도움을 준다.

```
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 인증토큰를 내려받는 기능
result<- search_tweets(key, n=100, token)	key 에 해당되는 트위터 글 읽어 오기

데이터 수집

공공DB 읽어오기



데이터 수집

공공DB 읽어오기