

데이터 수집

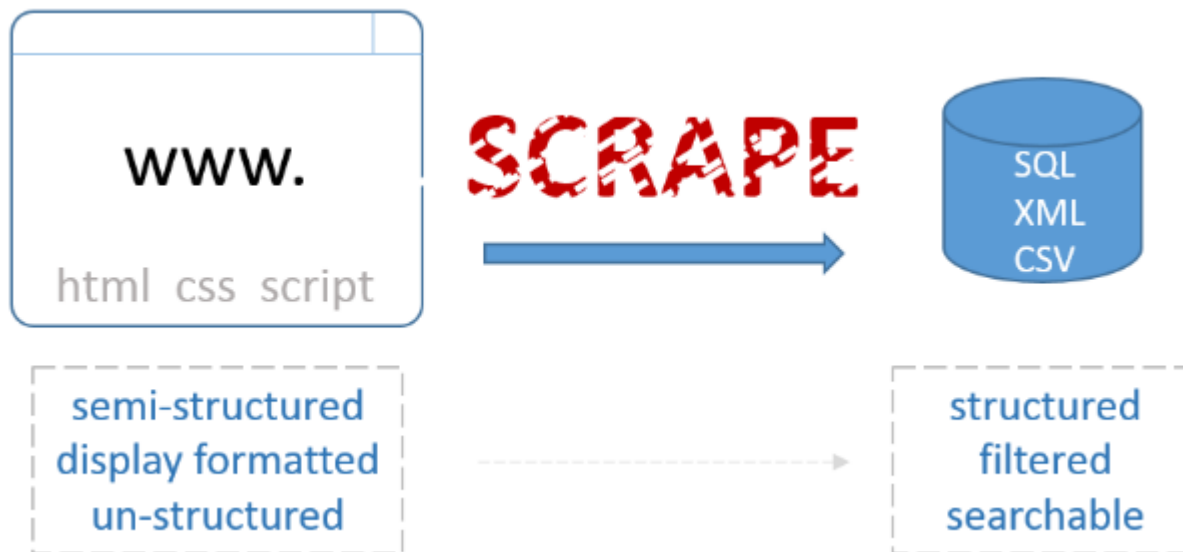
1. 정적 스크래핑(크롤링)

[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

1. 정적 스크래핑(크롤링)

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

1. 정적 스크래핑(크롤링)

(1) 네이버 영화 사이트 댓글정보 스크래핑

네이버 영화 사이트의 데이터 중 영화제목, 평점, 리뷰만을 추출하여 CSV 파일의 정형화된 형식으로 저장한다.

- 스크래핑하려는 웹페이지의 URL 구조와 문서 구조를 파악해야 한다.

- URL 구조 : <http://movie.naver.com/movie/point/af/list.nhn?page=1>

The screenshot shows the Naver movie review page for the movie '고지전' (Gojigen). The URL in the browser's address bar is <http://movie.naver.com/movie/point/af/list.nhn?page=1>, with the page number '1' highlighted in a red box. The page displays a list of reviews with columns for '번호' (Number), '평점' (Rating), and '140자평' (140-character review). The first review is for '고지전' with a rating of 10. The HTML structure is shown on the right, with the 'title' element of the first review highlighted in a red box. The HTML code for the first review is as follows:

```
<tr>
  <td class="ac num">
    14911148</td>
  <td>...</td>
  <td class="point">10</td>
  <td class="title"> == $0
    <a href="?
      st=mcode&sword=74315&ta
      rget=after" class=
        "movie">고지전</a>
    <br>
    "박평식 평론가의 성을 차
      는 영화는 어디 있는가? 내
      가 평론가였으면, 평론가가
      된다면, 박평식 평론가부터
      지적할 것이다
```

1. 정적 스크래핑(크롤링)

- 문서 구조

영화 제목 class="movie"

영화 평점 class="point"

영화 리뷰 class="title"

[rvest 패키지의 주요 함수]

html_nodes(x, css, xpath), html_node(x, css, xpath)

html_text(x, trim=FALSE)

html_attrs(x)

html_attr(x, name, default = "")

[1페이지 스크래핑]

```
install.packages("rvest");  
library(rvest)  
url <- "http://movie.naver.com/movie/point/af/list.nhn?page=1"  
text <- read_html(url, encoding="CP949")  
# 영화제목  
nodes <- html_nodes(text, ".movie")  
title <- html_text(nodes)  
# 영화평점  
nodes <- html_nodes(text, ".point")  
point <- html_text(nodes)  
# 영화리뷰  
nodes <- html_nodes(text, ".title")  
review <- html_text(nodes, trim=TRUE); review  
review <- gsub("\t", "", review)  
review <- gsub("\r\n", "", review)  
review <- gsub("신고", "", review); review  
page <- cbind(title, point)  
page <- cbind(page, review)  
write.csv(page, "movie_reviews.csv")
```

1. 정적 스크래핑(크롤링)

[여러 페이지 스크래핑]

```
site <- "http://movie.naver.com/movie/point/af/list.nhn?page="
movie.review <- NULL
for(i in 1: 100) {
  url <- paste(site, i, sep="")
  text <- read_html(url, encoding="CP949")
  nodes <- html_nodes(text, ".movie")
  title <- html_text(nodes)
  nodes <- html_nodes(text, ".point")
  point <- html_text(nodes)
  nodes <- html_nodes(text, ".title")
  review <- html_text(nodes, trim=TRUE)
  review <- gsub("http", "", review); review <- gsub("www", "", review)
  review <- gsub("신고", "", review)
  page <- cbind(title, point)
  page <- cbind(page, review)
  movie.review <- rbind(movie.review, page)
}
write.csv(movie.review, "movie_reviews2.csv")
```

데이터 수집

1. 정적 스크래핑(크롤링)

한국일보 헤드라인
기사 스크래핑

→ ↻ ⓘ 주의 요함 | www.hankookilbo.com

고 보는 동영상 PRAM

정치 경제 사회 국제 문화 연예 라이프 스포츠 피플 지역 | 오피니언 기획·특집 디지털스페셜 멀티미디어

"문 대통령, 링컨이나 물태우냐" 정동영, 선거제 개편 촉구

바른미래당 · 민주평화당 · 정의당 등 3당이 연동형 비례대표제 관철을 위해 문재인 대통령을 거론하며 더불어민주당에 대한 압박 수위를 높였다. 더보기

간판 없는 비밀의 공간 "취향을 팝니다"
요즘 뜨는 명소들은 손맛이나 목 좋은 곳이 아니다. 찾아오는 방법부터 공간이 가진 매력을 느낄 수 있어야 비로소 명소가 된다.

(겨울) "신지도 않는 에어컨조단을 왜 수집하느냐고?"
대중문화부터 상품까지, 레트로에 열광하는 이유는
미국서도 멀레니얼 세대 겨냥 레트로 마케팅 활발

"그때 그 아이 맞습니다" 아역출신 배우들이 떴다
"여기 착하는 70대 배우" 기구에 아역 출신 배우들이 확연히 반증과

단독 이영렬, 돈봉투 만찬 '무혐의' 받았지만 명예는...
형사상 혐의 벗었지만 정권초 과도한 망신주기 조치 비판도

(뒤끝뉴스) '4조원'에 막힌 470조 슈퍼 예산심의
문희상 의장, 중부세 인상 등 예산부수법안 28건 지정

조명래 "미세먼지 중국 탓 하기 전 우리가 줄여야"
30일 또 스모그 밀려온다... "中 환경규제 늦춘 탓"
흡입하면 몸에 축적되는 미세먼지, 이렇게 대처해야

속보 부산 폐수처리업체 황화수소 누출...4명 의식불명
대처, 아이스크림 개발자로 었지페 모델 후보에

이번엔 '분빠이'... 이은재 의원의 일본어 사랑
맞춤법은 틀리면서... 이은재 의원, 또 일어 사용

(줌인뉴스) 배달음식 하나 당 딸려오는 일회용품 7개
유치원생이 아파트 2채... '금수저' 미성년자들 조사

"대기업 임원들, 2년 차에 옷 가장 많이 벗는다"
친일했는데 독립운동가? "가짜 100명 색출한다"

음주 뺑소니로 50대 사망 이르게 한 20대
"광개이불 화재 처음 본다" KT화재 원인 미스터리

합계출산율 2.3분기 연속 '0명대'... 역대 최저
수면유도제 졸피뎀 처방, 4주 넘길 수 없다

김관영 "여당 스스로 대통령 레임덕 부추겨"

데이터 수집

1. 정적 스크래핑(크롤링)

[XML 패키지의 주요 함수]

htmlParse (file, encoding="...")

xpathSApply(doc, path, fun)

fun : **xmlValue**, **xmlGetAttr**, **xmlAttrs**

library(XML)

t <- htmlParse("http://hankookilbo.com")

content <- xpathSApply(t, "//p[@class='title']", xmlValue);

Content

content <- gsub("[:punct:][:cntrl:]", "", content)

content

content <- trimws(content)

content

```
<div class="splash" /.../div>
▼<ul class="headline-related">
  ▼<li>
    ▶<div class="frame"> </div>
    ▼<p class="title">
      <a href="/News/Read/201811262370046811?
      did=PA&dtype=3&dtypecode=571" target>
        간판 없는 비밀의 공간 “취향을 팝니다”
      </a>
    </p>
    ▶<p class="preview">...</p>
  </li>
  ▶<li>...</li>
  ▶<li>...</li>
```

1. 정적 스크래핑(크롤링)

[Xpath]

XPath란?

XPath란 XML Path Language를 의미한다.

XPath는 XML 문서의 특정 요소나 속성에 접근하기 위한 경로를 지정하는 언어이다.

/bookstore 최상위 엘리먼트로서의 bookstore 태그를 찾는다.

/bookstore/book 최상위 엘리먼트 bookstore 태그를 찾은 후에 book이라는 자식 태그를 찾는다.

/bookstore//book 최상위 엘리먼트 bookstore 태그를 찾은 후에 book이라는 자손 태그를 찾는다.

//book 조상이 누구든 book이라는 태그를 찾는다.

//@lang lang이라는 속성을 갖는 태그를 찾는다.

//title[@lang='en'] lang이라는 속성의 값이 'en' 인 title 태그를 찾는다.

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
<book>
  <title lang="en">Harry Potter</title>
  <price>29.99</price>
</book>
<book>
  <title lang="en">Learning XML</title>
  <price>39.95</price>
</book>
</bookstore>
```


1. 정적 스크래핑(크롤링)

그 외의 웹 스크래핑시 알고 있으면 도움되는 내용들

[R에서 GET으로 사이트 내용 가져오기 : httr 패키지 사용]

```
library(httr)
http.standard <- GET('http://www.worg/Protocols/rfc2616/rfc261html')
title2 = html_nodes(read_html(http.standard), 'div.toc h2')
title2 = html_text(title2)
```

[R에서 POST로 사이트 내용 가져오기 : httr 패키지 사용]

```
library(httr)
# POST 함수를 이용해 모바일 게임 랭킹 10월 29일 주 모바일 게임 랭킹을 찾는다
#(http://www.gevolution.co.kr/score/gamescore.asp?t=3&m=0&d=week)
game = POST('http://www.gevolution.co.kr/score/gamescore.asp?t=3&m=0&d=week',
            encode = 'form', body=list(txtPeriodW = '2018-10-29'))
title2 = html_nodes(read_html(game), 'a.tracktitle')
title2 = html_text(title2)
title2[1:10]
```

1. 정적 스크래핑(크롤링)

[뉴스, 게시판 등 글 목록에서 글의 URL만 뽑아내기]

```
res = GET('https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&sid1=001')
htxt = read_html(res)
link = html_nodes(htxt, 'div.list_body a')
article.href = unique(html_attr(link, 'href'))
```

[이미지, 첨부파일 다운 받기]

```
# pdf
res = GET('http://cran.r-project.org/web/packages/htr/htr.pdf')
writeBin(content(res, 'raw'), 'c:/Temp/htr.pdf')

# jpg
h = read_html('http://unico201dothome.co.kr/productlog.html')
imgs = html_nodes(h, 'img')
img.src = html_attr(imgs, 'src')
for(i in 1:length(img.src)){
  res = GET(paste('http://unico201dothome.co.kr/',img.src[i], sep=""))
  writeBin(content(res, 'raw'), paste('c:/Temp/', img.src[i], sep=""))
}
```