

데이터의 수집

방법

- API 를 활용하는 방법
 - API: Application Programming Interface 의 약자
 - 데이터를 가진 주체가 데이터를 조회 또는 조작할 수 있도록 제공하는 방법(Interface)
 - 예) 유튜브가 가진 데이터를 조회할 수 있도록 유튜브는 Youtube Data API를 제공
<https://developers.google.com/youtube/v3/getting-started?hl=ko>
 - 장점: 데이터를 가진 주체가 직접 제공하기 때문에 가이드가 있어 쉬움
 - 단점: 주체가 제공을 원하지 않는 경우 API가 없거나 제한 적일 수 있음

데이터의 수집

방법

- 웹 페이지 크롤링
 - 웹 페이지에 게재된 데이터를 수집하는 방법
 - 수많은 페이지로부터 수집을 하기 때문에 프로그래밍을 통해 수집
 - 예) 네이버뉴스에서 ‘코로나’를 검색하여 나온 기사 내용을 수집
 - 장점: 웹 페이지에 있는 모든 내용을 수집할 수 있음
 - 단점: 프로그래밍이 필요하여 어려운 편이고, 과도한 크롤링을 할 경우 사이트 접속에 블록(차단)당할 수 있음

데이터의 수집

방법

- 그 외 수집 방법
 - 내가 만든 프로그램, 웹사이트인 경우 직접 수집하는 코드 작성
 - 이동정보를 수집하기 위해서 별도의 모바일 어플리케이션 개발
 - 인터뷰를 통한 수집
 - 구글 Form과 같은 설문조사 용 플랫폼을 활용한 수집
 - 등 ...

데이터의 수집

작동 원리 및 실습 - API

- Youtube Data API를 활용한 데이터 수집
 - 채널에 올라온 영상 목록을 수집
 - API를 활용한 방법 => 가이드를 제공

Search: list

API 요청에 지정된 쿼리 매개변수와 일치하는 검색결과의 모음을 반환합니다. 기본적으로 검색결과의 집합은 쿼리 매개변수와 일치하는 `video`, `channel`, `playlist` 리소스를 식별하지만, 특정 유형의 리소스만 검색하도록 쿼리를 구성할 수도 있습니다. [지금 사용해 보거나 예를 참조하세요.](#)

요청

HTTP 요청

```
GET https://www.googleapis.com/youtube/v3/search
```

데이터의 수집

작동 원리 및 실습 - API

- Youtube Data API를 활용한 데이터 수집

```
...
채널의 영상 목록을 반환하는 함수

arguments
- channel: 채널의 고유 아이디 (str)
response
- items: 영상 리스트 (list)
...

def list_videos(channel):
    endpoint = 'https://www.googleapis.com/youtube/v3/search' → API 호출 주소
    args = {
        'part': 'id,snippet', → 수집할 데이터( snippet: 결과의 제목, 설명 등을 포함)
        'channelId': channel, → 채널 고유키
        'maxResults': '50', → 결과 최대 50개
        'type': 'video' → 검색 결과에 영상만 조회
    }
    items = call_api(endpoint, args) → API 사용
    return items
```

데이터의 수집

작동 원리 및 실습 - API

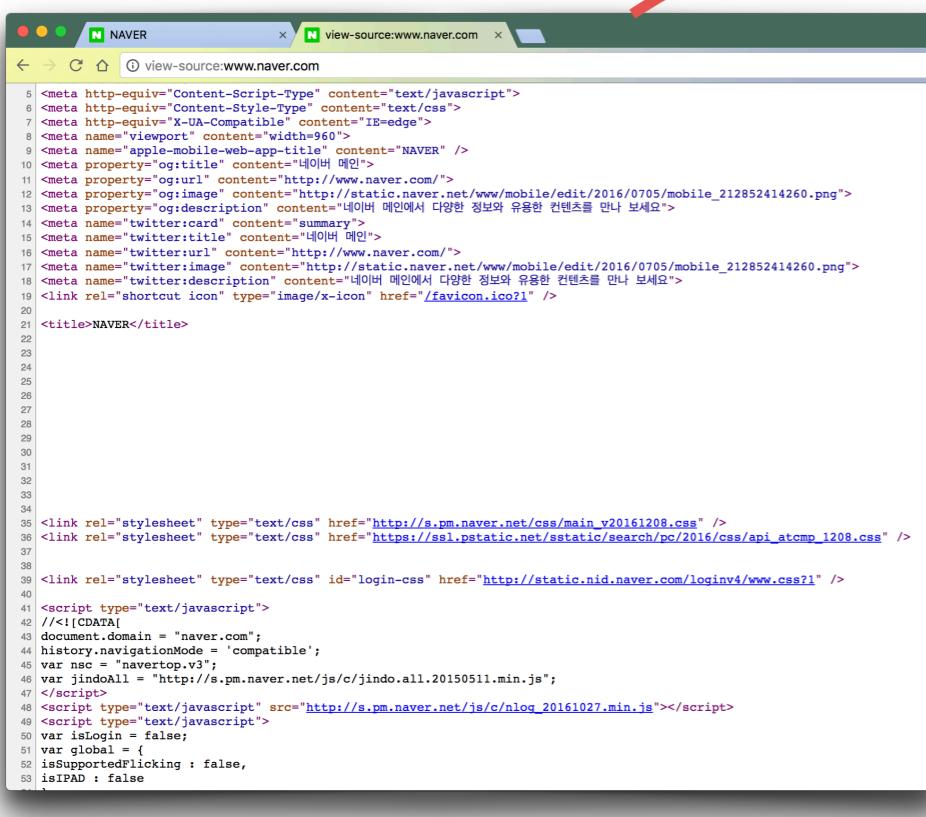
- Youtube Data API를 활용한 데이터 수집

```
{'etag': 'aHZQSLz4ETx690_TUYr3_wSnqWU',
'id': {'kind': 'youtube#video', 'videoId': 'UlfFjIDxL3o'},
'kind': 'youtube#searchResult',
'snippet': {'channelId': 'UC9EgNOu8Y9tY3zkrXFk0T_w',
    'channelTitle': '경기신용보증재단',
    'description': '안녕하세요 경기신용보증재단입니다. 경기신보에서 상담을 받고 서류제출을 어려워하시는 '
                   '소상공인 분들을 위해 준비했습니다! 경기신보 유튜브 구독! 좋아요!',
    'liveBroadcastContent': 'none',
    'publishTime': '2019-04-10T02:07:34Z',
    'publishedAt': '2019-04-10T02:07:34Z',
    'thumbnails': {'default': {'height': 90,
                                'url': 'https://i.ytimg.com/vi/UlfFjIDxL3o/default.jpg',
                                'width': 120},
                  'high': {'height': 360,
                           'url': 'https://i.ytimg.com/vi/UlfFjIDxL3o/hqdefault.jpg',
                           'width': 480},
                  'medium': {'height': 180,
                             'url': 'https://i.ytimg.com/vi/UlfFjIDxL3o/mqdefault.jpg',
                             'width': 320}},
    'title': '경기신보 제출서류 알아보기 소상공인편!'}}},
```

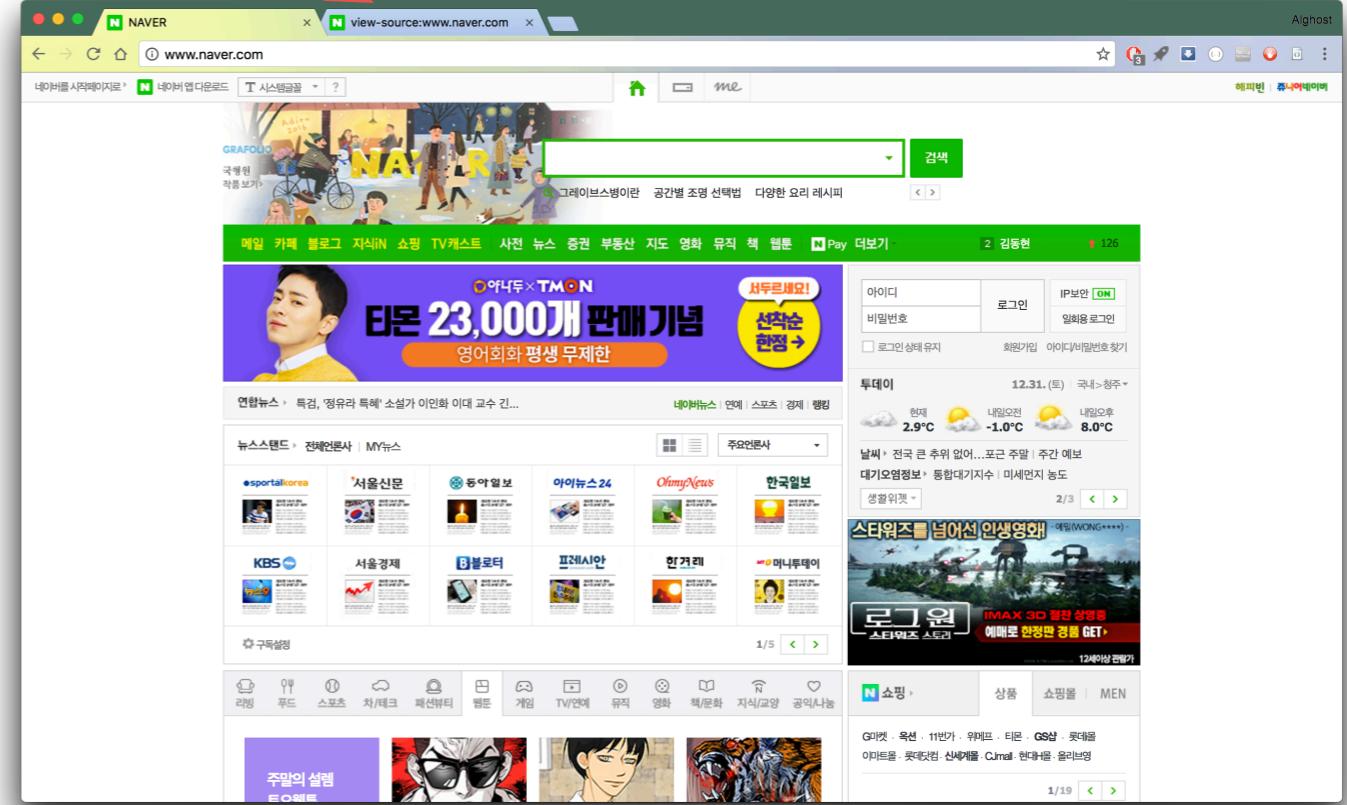
데이터의 수집

작동 원리 및 실습 - 크롤링

- 웹 페이지는 결국 ‘문서’
- 웹 페이지도 결국 문서이기 때문에 웹문서 라고도 함
웹브라우저: 문서 해석 후 보여줌



```
5 <meta http-equiv="Content-Type" content="text/javascript">
6 <meta http-equiv="Content-Style-Type" content="text/css">
7 <meta http-equiv="X-UA-Compatible" content="IE=edge">
8 <meta name="viewport" content="width=960">
9 <meta name="apple-mobile-web-app-title" content="NAVER" />
10 <meta property="og:title" content="네이버 메인">
11 <meta property="og:url" content="http://www.naver.com">
12 <meta property="og:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png">
13 <meta property="og:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요">
14 <meta name="twitter:card" content="summary">
15 <meta name="twitter:title" content="네이버 메인">
16 <meta name="twitter:url" content="http://www.naver.com"/>
17 <meta name="twitter:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png">
18 <meta name="twitter:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요">
19 <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico?1" />
20
21 <title>NAVER</title>
22
23
24
25
26
27
28
29
30
31
32
33
34
35 <link rel="stylesheet" type="text/css" href="http://s.pm.naver.net/css/main_v20161208.css" />
36 <link rel="stylesheet" type="text/css" href="https://ssl.pstatic.net/sstatic/search/pc/2016/css/api_atcmp_1208.css" />
37
38 <link rel="stylesheet" type="text/css" id="login-css" href="http://static.nid.naver.com/loginv4/www.css?1" />
39 <script type="text/javascript">
40 //<![CDATA[
41 document.domain = "naver.com";
42 history.navigationMode = 'compatible';
43 var nsc = "navertop.v3";
44 var jindoAll = "http://s.pm.naver.net/js/c/jindo.all.20150511.min.js";
45 &lt;/script&gt;
46 &lt;script type="text/javascript" src="http://s.pm.naver.net/js/c/nlog_20161027.min.js"&gt;&lt;/script&gt;
47 &lt;script type="text/javascript"&gt;
48 var isLogin = false;
49 var global = {
50     isSupportedFlicking : false,
51     isIPAD : false
52 }</pre>
```

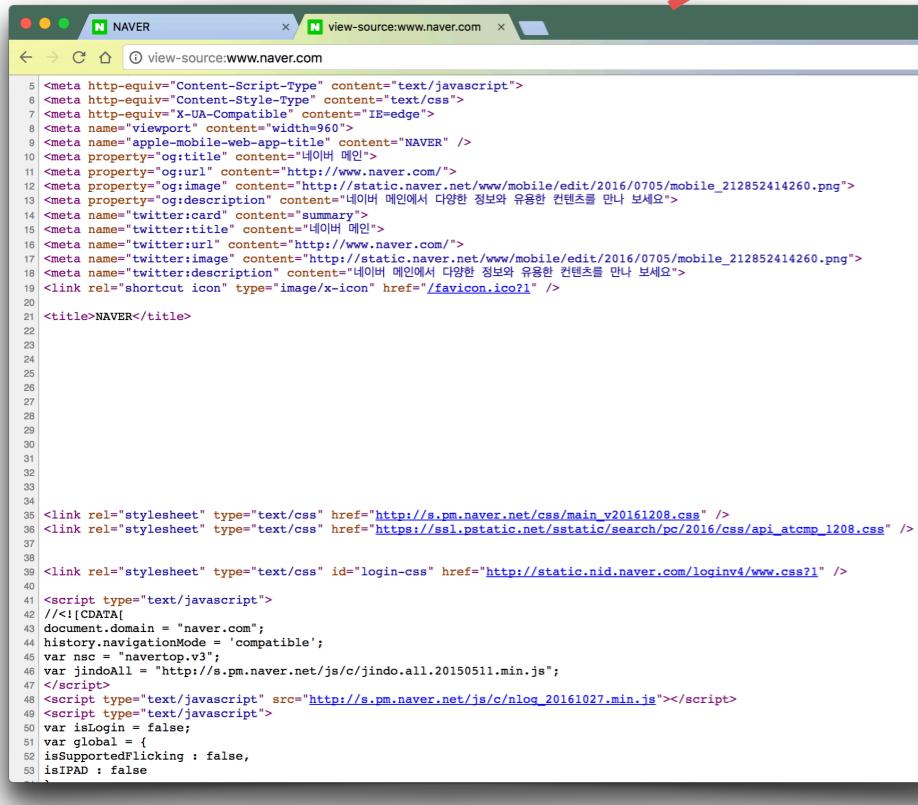


데이터의 수집

작동 원리 및 실습 - 크롤링

- 웹 페이지는 결국 ‘문서’
- 웹 페이지도 결국 문서이기 때문에 웹문서 라고도 함

웹 드라이버



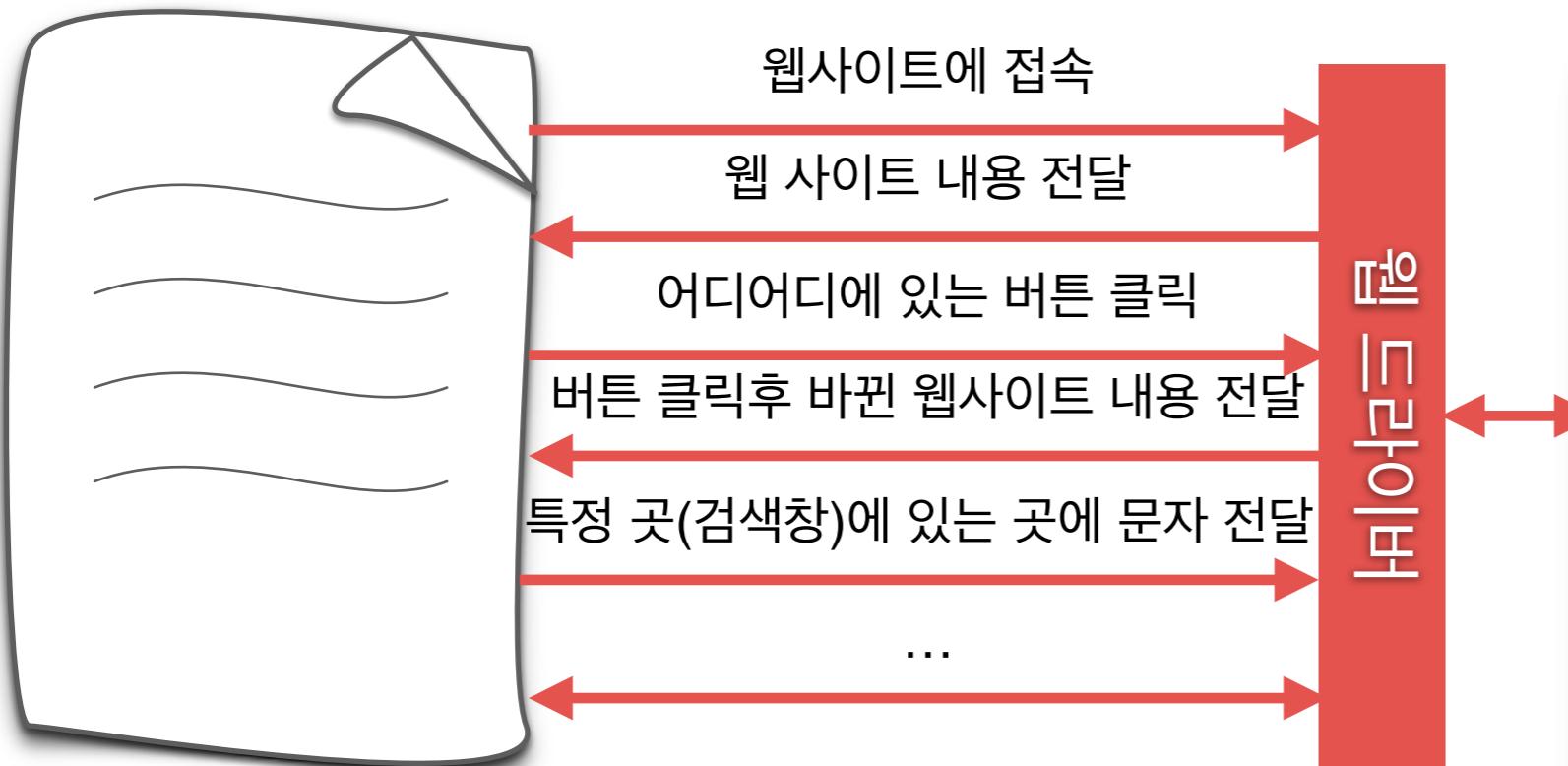
```
5 <meta http-equiv="Content-Type" content="text/javascript">
6 <meta http-equiv="Content-Type" content="text/css">
7 <meta http-equiv="X-UA-Compatible" content="IE=edge">
8 <meta name="viewport" content="width=960">
9 <meta name="apple-mobile-web-app-title" content="NAVER" />
10 <meta property="og:title" content="네이버 메인">
11 <meta property="og:url" content="http://www.naver.com/">
12 <meta property="og:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png">
13 <meta property="og:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요">
14 <meta name="twitter:card" content="summary">
15 <meta name="twitter:title" content="네이버 메인">
16 <meta name="twitter:url" content="http://www.naver.com/">
17 <meta name="twitter:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png">
18 <meta name="twitter:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요">
19 <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico1" />
20
21 <title>NAVER</title>
22
23
24
25
26
27
28
29
30
31
32
33
34
35 <link rel="stylesheet" type="text/css" href="http://s.pm.naver.net/css/main_v20161208.css" />
36 <link rel="stylesheet" type="text/css" href="https://ssl.pstatic.net/sstatic/search/pc/2016/css/api_atcmp_1208.css" />
37
38 <link rel="stylesheet" type="text/css" id="login-css" href="http://static.nid.naver.com/loginv4/www.css?1" />
39
40 <script type="text/javascript">
41 //<![CDATA[
42 document.domain = "naver.com";
43 history.navigationMode = 'compatible';
44 var ns = "navertop.v3";
45 var jindoAll = "http://s.pm.naver.net/js/c/jindo.all.20150511.min.js";
46 &lt;/script&gt;
47 &lt;script type="text/javascript" src="http://s.pm.naver.net/js/c/nlog_20161027.min.js"&gt;&lt;/script&gt;
48 &lt;script type="text/javascript"&gt;
49 var isLogin = false;
50 var global = {
51 isSupportedClicking : false,
52 isIPAD : false
53 }</pre>
```



데이터의 수집

작동 원리 및 실습 - 크롤링

- 웹 문서를 분석하여 이를 활용하여 화면을 구성
- 웹 문서에 이벤트를 전달하고 결과값을 받음
- 웹 드라이버가 제공하는 (어려운) 방법으로 서로 주고 받아야 함



python 프로그램

A screenshot of a web browser window titled "NAVER" showing the source code for "view-source:www.naver.com". The code includes meta tags for viewport, title, and Open Graph properties, as well as various script and link tags. The browser's status bar at the bottom shows the URL "view-source:www.naver.com".

```
5 <meta http-equiv="Content-Type" content="text/javascript">
6 <meta http-equiv="Content-Style-Type" content="text/css">
7 <meta http-equiv="X-UA-Compatible" content="IE=edge">
8 <meta name="viewport" content="width=960">
9 <meta name="apple-mobile-web-app-title" content="NAVER" />
10 <meta property="og:title" content="네이버 메인">
11 <meta property="og:url" content="http://www.naver.com"/>
12 <meta property="og:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png">
13 <meta property="og:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요">
14 <meta name="twitter:card" content="summary">
15 <meta name="twitter:title" content="네이버 메인">
16 <meta name="twitter:url" content="http://www.naver.com"/>
17 <meta name="twitter:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png">
18 <meta name="twitter:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요">
19 <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico?1" />
20
21 <title>NAVER</title>
22
23
24
25
26
27
28
29
30
31
32
33
34
35 <link rel="stylesheet" type="text/css" href="http://s.pm.naver.net/css/main_v20161208.css" />
36 <link rel="stylesheet" type="text/css" href="https://ssl.pstatic.net/sstatic/search/pc/2016/css/api_atcmp_1208.css" />
37
38
39 <link rel="stylesheet" type="text/css" id="login-css" href="http://static.nid.naver.com/loginv4/www.css?1" />
40
41 <script type="text/javascript">
42 //<![CDATA[
43 document.domain = "naver.com";
44 history.navigationMode = 'compatible';
45 var nsc = "navertop.v3";
46 var jindoall = "http://s.pm.naver.net/js/c/jindo.all.20150511.min.js";
47 &lt;/script&gt;
48 &lt;script type="text/javascript" src="http://s.pm.naver.net/js/c/nlog_20161027.min.js"&gt;&lt;/script&gt;
49 &lt;script type="text/javascript"&gt;
50 var isLogin = false;
51 var global = {
52 isSupportedFlicking : false,
53 isIPAD : false
54 }</pre>
```

데이터의 수집

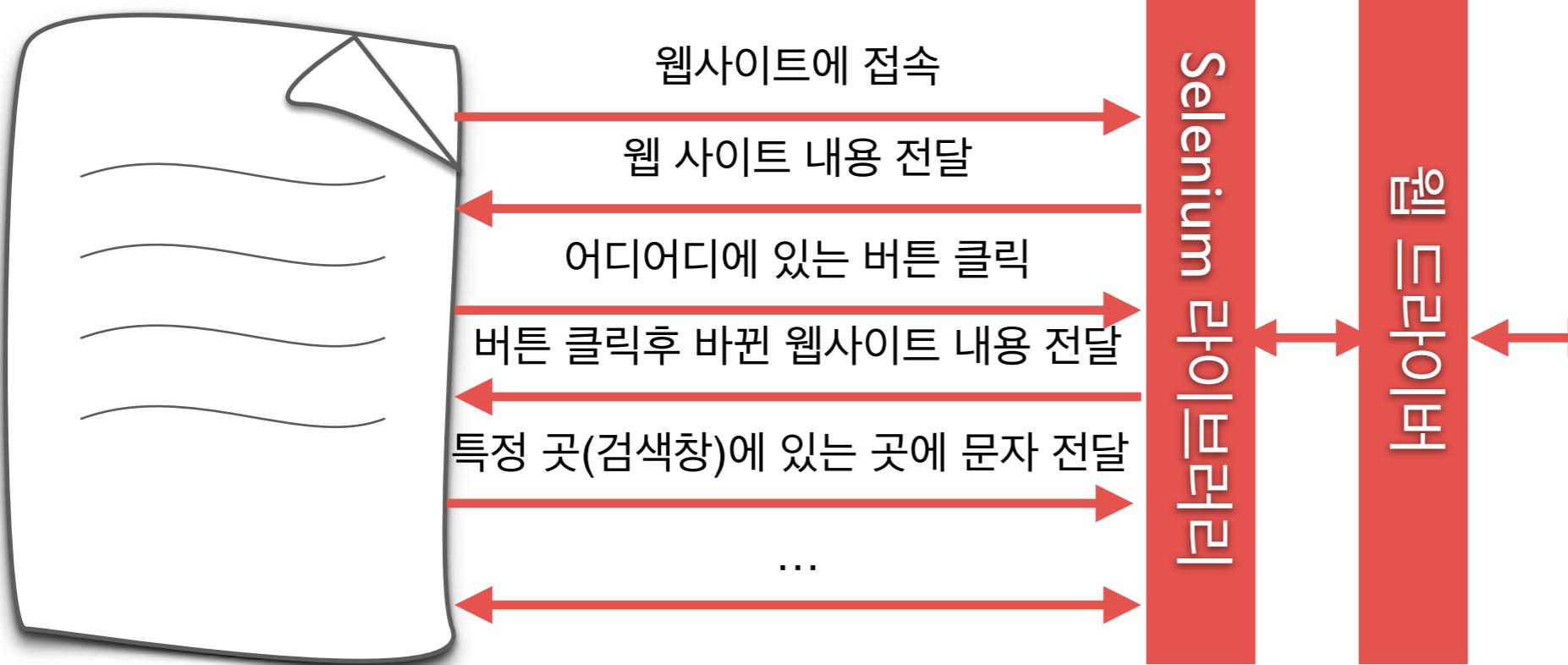
작동 원리 및 실습 - 크롤링

- 웹 드라이버를 ‘직접’ 다룬다?
 - 웹 드라이버를 직접 다룬다: 브라우저를 만들겠다!
 - 어렵기 때문에 라이브러리를 사용 => Selenium
- Selenium
 - 웹 브라우저 테스트 자동화 라이브러리
 - 다양한 브라우저의 웹 드라이버를 컨트롤 할 수 있는 라이브러리 제공

데이터의 수집

작동 원리 및 실습 - 크롤링

- Selenium을 사용하면
 - Selenium이 제공하는 편리한 라이브러리로 웹 드라이버를 제어
 - 동작 원리는 다음과 같음



python 프로그램

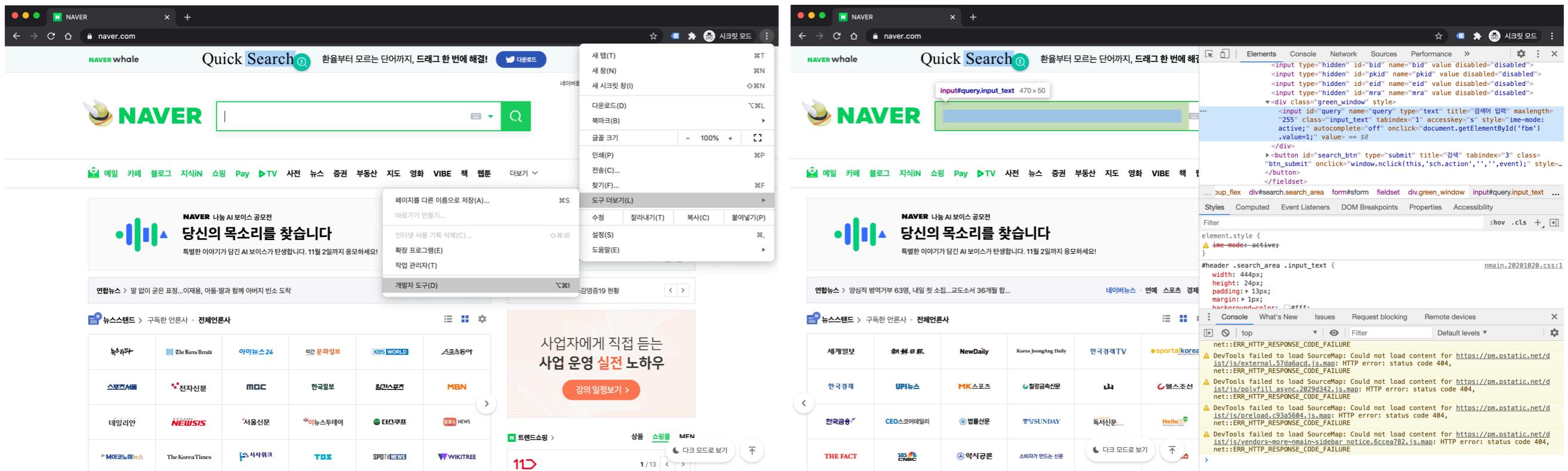
A screenshot of a web browser window titled 'NAVER'. The URL 'view-source:www.naver.com' is visible in the address bar. The page content shows the HTML source code of a search result page, including meta tags, script blocks, and various HTML elements. The code is mostly in Korean, with some English and URLs interspersed.

```
5 <meta http-equiv="Content-Script-Type" content="text/javascript">
6 <meta http-equiv="Content-Style-Type" content="text/css">
7 <meta http-equiv="X-UA-Compatible" content="IE=edge">
8 <meta name="viewport" content="width=device-width, initial-scale=1.0, user-scalable=0" />
9 <meta name="apple-mobile-web-app-title" content="NAVER" />
10 <meta property="og:title" content="네이버 메인" />
11 <meta property="og:url" content="http://www.naver.com" />
12 <meta property="og:image" content="http://static.naver.net/www/mobile_212852414260.png" />
13 <meta property="og:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요" />
14 <meta name="twitter:card" content="summary">
15 <meta name="twitter:title" content="네이버 메인" />
16 <meta name="twitter:site" content="http://www.naver.com" />
17 <meta name="twitter:image" content="http://static.naver.net/www/mobile/edit/2016/0705/mobile_212852414260.png" />
18 <meta name="twitter:description" content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요" />
19 <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico?1" />
20
21 <title>NAVER</title>
22
23
24
25
26
27
28
29
30
31
32
33
34
35 <link rel="stylesheet" type="text/css" href="https://s.pm.naver.net/css/main_v20161208.css" />
36 <link rel="stylesheet" type="text/css" href="https://ssl.static.naver.net/seach/pc/2016/css/api_atcmp_1208.css" />
37
38
39 <link rel="stylesheet" type="text/css" id="login-css" href="http://static.nid.naver.com/loginv4/www.css?1" />
40
41 <script type="text/javascript">
42 <!--[CDATA[
43 document.domain = "naver.com";
44 history.navigatMode = "compatible";
45 var isMobile = screen.width < 768;
46 var jindoAll = "http://s.pm.naver.net/js/c/jindo.all.20150511.min.js";
47 -->
48 <script type="text/javascript" src="http://s.pm.naver.net/is/c/nlog_20161027.min.js"></script>
49 <script type="text/javascript" src="http://s.pm.naver.net/is/c/nlog_20161027.min.js"></script>
50 var isLogin = false;
51 var global = {
52 isSupportDjlicking : false,
53 isPAU : false
54 }
```

데이터의 수집

작동 원리 및 실습 - 크롤링

- 웹 페이지를 분석해보자
- 크롬 브라우저의 개발자 도구 활용



LAH

데이터의 수집

작동 원리 및 실습 - 크롤링

- 크롤링을 활용한 네이버 뉴스 데이터 수집

The screenshot shows the Naver News homepage with a search bar containing the query "디지털화폐". The results page displays several news articles related to digital currencies, such as the introduction of CBDC and its impact on traditional currencies. On the right side, there is a sidebar titled "언론사별 가장 많이 본 뉴스" which lists the most viewed news stories from various media outlets like SBS and KBS.

NAVER 뉴스 | TV연예 | 스포츠 | 뉴스스탠드 | 날씨

로그인

뉴스 검색

10.25 (일) 헤드라인 뉴스 이재용, 팔리세이드 직접 몰고 아들·딸과 빈소로[이건희...]

팩트체크 | 언론사 구독 | 언론사 뉴스 | 라이브러리

IT/과학

모바일

인터넷/SNS

통신/뉴미디어

IT 일반

보안/해킹

컴퓨터

게임/리뷰

과학 일반

속보

모바일 메인에서
보고싶은 뉴스
구독하세요!
바로가기 >

① 헤드라인 뉴스 Beta

7 "디지털화폐" 급부상 • '디지털위안화는 암호화폐는 안되고' 中 법제화

中 디지털위안화 법제화, 암호화폐는 전면 금지

[아시아경제 나한아 기자] 법정 디지털화폐(CBDC)인 '디지털 위안화' 도입에 속도를 내고 있는 디지털 코드도 법정 화폐로 인정받을 수 있도록 법적 근 ...

아시아경제 | 10+

'디지털위안화는 되고 암호화폐는 안되고' 中 법제화 서울경제

비트코인 품은 페이팔에 주목해야 하는 이유 ZDNet Korea | 50+

세계 중앙은행은 '디지털화폐' 경쟁 중 [임주형의 테크토크] 아시아경제

④ 스마트폰 세계 1위 갤럭시...이건희 불호령 "애니콜 화형식" 있었다

스마트폰 세계 1위 '갤럭시'...이건희 불호령 '애니콜 화형식' 있었다

(서울=뉴스1) 김정현 기자 = "휴대폰 품질에 신경을 쓰십시오. 고객이 두렵지 않습니까? 비싼 휴대폰, 고장나면 누가 사겠습니까? 반드시 1명당 1대의 ..."

언론사별 가장 많이 본 뉴스 ⓘ

오늘 6시~오늘 7시까지 집계한 결과입니다.

더보기 ⓘ

어? 류현진·김광현이 다가 아니었네. 최지만도 있었어!

시사저널

[이건희 별세] 이재용 부회장, 자녀 2명 태우고 직접 운 ...

아이뉴스24

이건희 회장 빈소 마련...이재용 부회장 조문객 맞아

KBS

외톨이 소년에서 글로벌 선도자로... 지난 78년의 삶

SBS

文대통령 직접 조문은 4번뿐...김상조 대신 노영민 보내는 이 ...

머니투데이