빅데이터 수집 개요
(웹 크롤링, 공공 DB, SNS)

한국기술교육대학교
온라인평생교육원

학습목표

- 웹 크롤링, 공공 DB, SNS를 활용한 **빅데이터 수집 방법**을 설명할 수 있다.
- **웹 페이지의 구조**를 파악하고 원하는 요소에 접근할 수 있다.

학습내용

- 빅데이터 수집 방법
- 웹 페이지를 구성하는 기술

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

웹 크롤링	▪ 웹 페이지의 하이퍼링크를 순회하면서 웹 페이지를 다운로드하는 작업
웹 스크래핑	▪ 다운로드한 웹 페이지에서 필요한 콘텐츠를 추출하는 작업 ▪ 웹 페이지를 구성하고 있는 HTML 태그의 콘텐츠나 속성의 값을 읽는 작업

<td>빨강 머리 앤</td>

태그의 콘텐츠

파이썬

태그의 속성값

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

1 URL(Uniform Resource Locator)

- 네트워크 상에서 자원이 어디 있는지를 알려주기 위한 규약
- 컴퓨터 네트워크와 검색 메커니즘에서의 자원의 위치를 지정하는 문자열

http://e-koreatech.step.or.kr/page/lms

프로토콜명 | 도메인명 | 요청 대상(URI)

URI(Uniform Resource Identity)

- 웹 사이트에 요청하고자 하는 대상의 패스정보와 파일명으로 구성
- 파일명이 생략되면 디폴트로 index.html 사용

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

2 HTTP(HyperText Transfer Protocol)

- 웹상에서 클라이언트와 서버 간에 정보를 주고받을 수 있는 통신 규약(프로토콜)

- URL 문자열을 직접 입력하거나, 하이퍼링크 텍스트 또는 이미지를 클릭하여 HTML 문서를 주고받는 데 사용
- 디폴트로 80번 포트 사용
- 다른 포트 번호를 사용하는 웹 서버에 요청 시 도메인명 뒤에 : 기호와 함께 포트 번호 지정
- 웹 클라이언트에서 웹 서버에 HTTP 요청을 전달할 때 요청 방식 명시
- 일반적으로 2가지 방식 사용

GET 방식	POST 방식

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

3 GET 방식과 POST 방식

GET 방식	▪ 브라우저에서 직접 요청하려는 페이지의 URL 문자열을 입력하여 요청 ▪ 하이퍼링크가 설정된 텍스트나 이미지를 클릭하여 요청
〈form〉 태그를 통한 요청	▪ method 속성값에 따라서 GET 방식 요청과 POST 방식 요청 모두 가능

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

3 GET 방식과 POST 방식

GET 방식

- Query 문자열 없는 요청과 Query 문자열을 추가한 요청 모두 가능
- Query 문자열이 URL 문자열 뒤에 추가되어 전달

https://movie.daum.net/moviedb/main?movieId=126260

웹 브라우저가 웹 서버에게 요청을 보내면서
함께 전달되는 name과 value로 구성되는 문자열

POST 방식

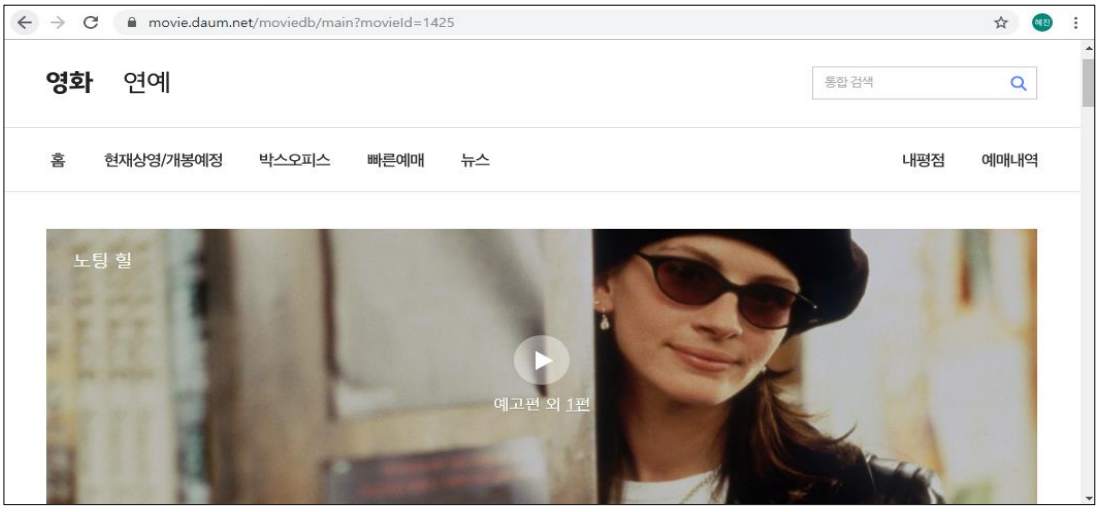
- Query 문자열을 추가한 요청만 가능
- Query 문자열이 요청 바디에 따로 담겨서 전달되므로 요청 URL 문자열에서는 볼 수 없음

빅데이터 수집 방법

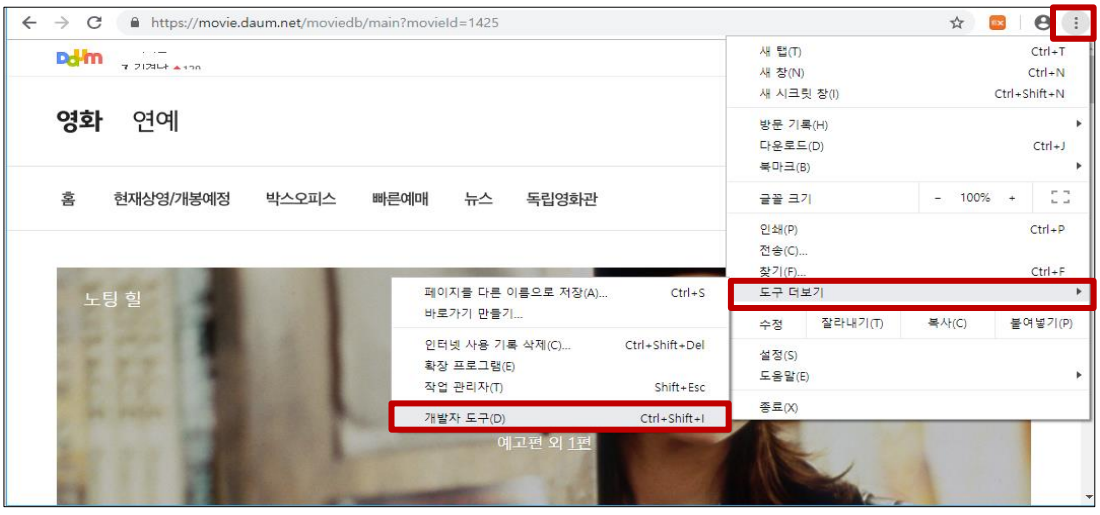
1 웹 크롤링과 웹 스크래핑

4 크롬(Chrome) 브라우저의 개발자 도구

1 크롬 브라우저에서 URL을 입력하고 요청



2 오른쪽 상단의 'Chrome 맞춤설정 및 제어' 메뉴를 클릭한 다음 '도구 더보기' 메뉴의 '개발자 도구' 메뉴 클릭



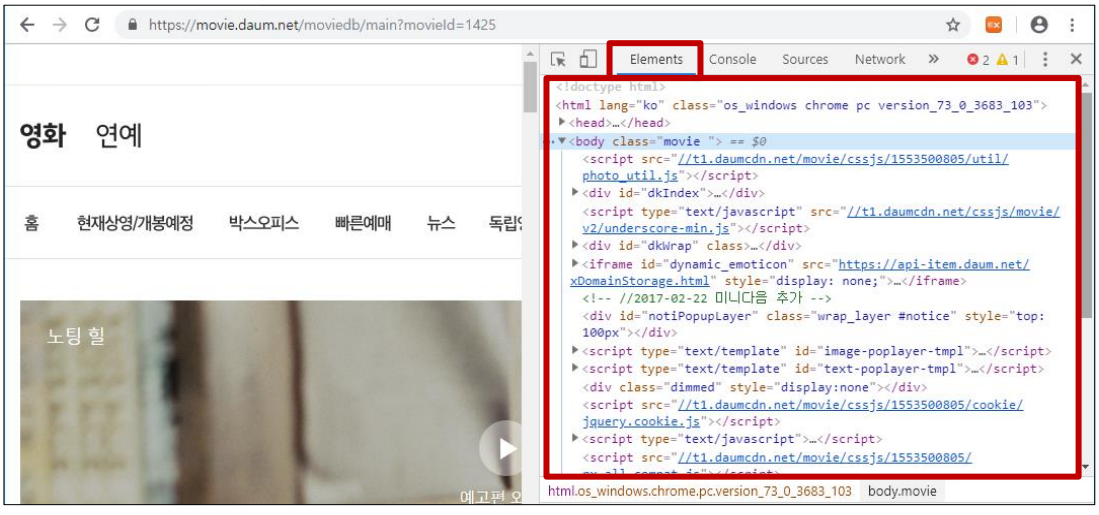
[출처 : https://movie.daum.net/]

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

4 크롬(Chrome) 브라우저의 개발자 도구

3 오른편으로 개발자 도구가 출력되고, 'Elements' 탭을 클릭하면 브라우저에서 렌더링되고 있는 웹 페이지의 HTML 소스가 출력됨



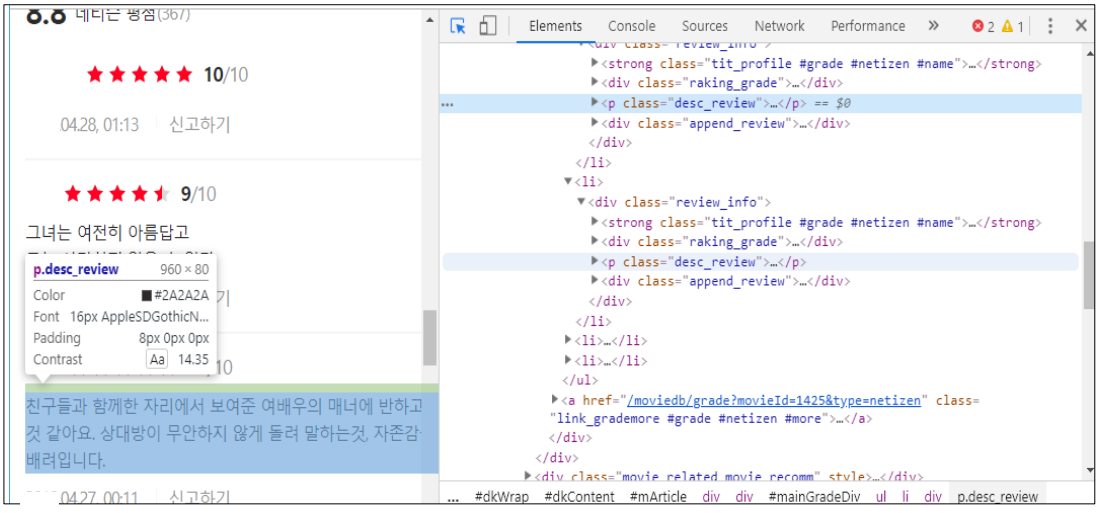
[출처 : https://movie.daum.net/]

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

4 크롬(Chrome) 브라우저의 개발자 도구

4 개발자 도구의 왼쪽 상단 버튼 클릭



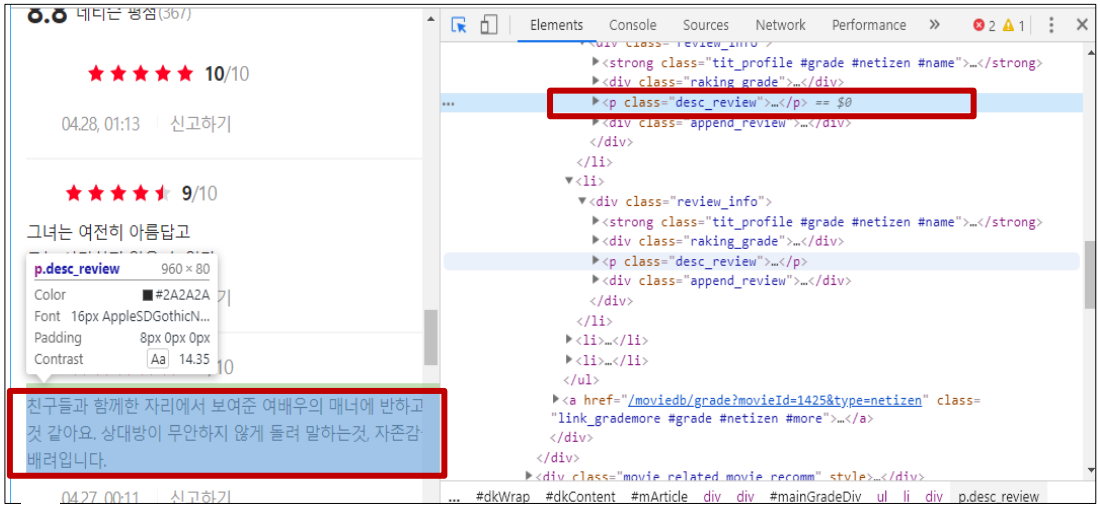
[출처 : https://movie.daum.net/]

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

4 크롬(Chrome) 브라우저의 개발자 도구

5 찾으려는 콘텐츠에 마우스를 올리면, 콘텐츠를 담고 있는 HTML 태그 부분이 개발자 도구의 태그 영역을 표시해주어 찾고자 하는 콘텐츠의 태그를 쉽게 찾을 수 있음



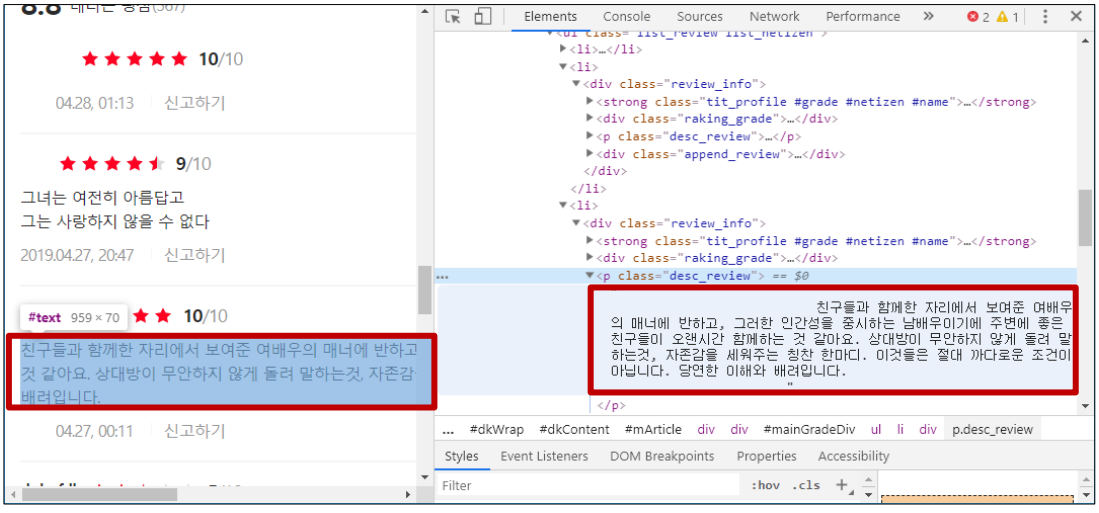
[출처 : https://movie.daum.net/]

빅데이터 수집 방법

1 웹 크롤링과 웹 스크래핑

4 크롬(Chrome) 브라우저의 개발자 도구

6 해당 태그 영역을 클릭하면, 태그의 콘텐츠 영역의 출력 내용이 웹 페이지에 렌더링된 내용과 동일함을 확인할 수 있음



[출처 : https://movie.daum.net/]

빅데이터 수집 방법

2 공공데이터

공공데이터	▪ 공공기관이 전자적으로 생성 또는 취득하여 관리하고 있는 모든 데이터베이스(DB), 전자화된 파일
공공데이터 개방	▪ 공공기관이 이용자에게 정보를 재활용할 수 있도록 제공하고, 제공받은 정보를 상업적·비영리적으로 이용할 권한 부여

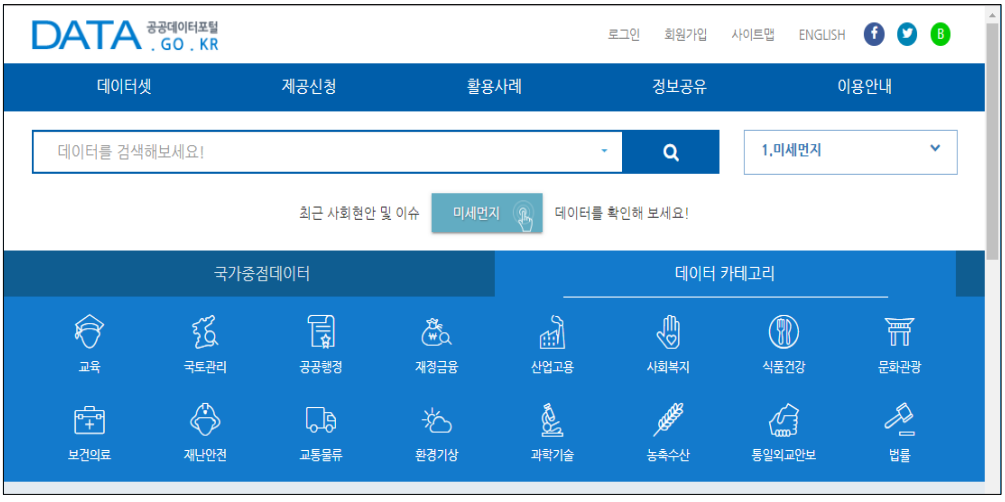
- 보유하고 있는 공공데이터를 적극적으로 개방하여 국민과 공유함으로써 소통과 협력을 확대하기 위해 공공데이터 정책 추진
- 2013년 7월 공공데이터법을 제정하고 공공데이터 개방을 10월부터 시행

빅데이터 수집 방법

2 공공데이터

1 공공데이터포털

- 공공기관이 생성 또는 취득하여 관리하고 있는 공공데이터를 한 곳에서 제공하는 통합 창구



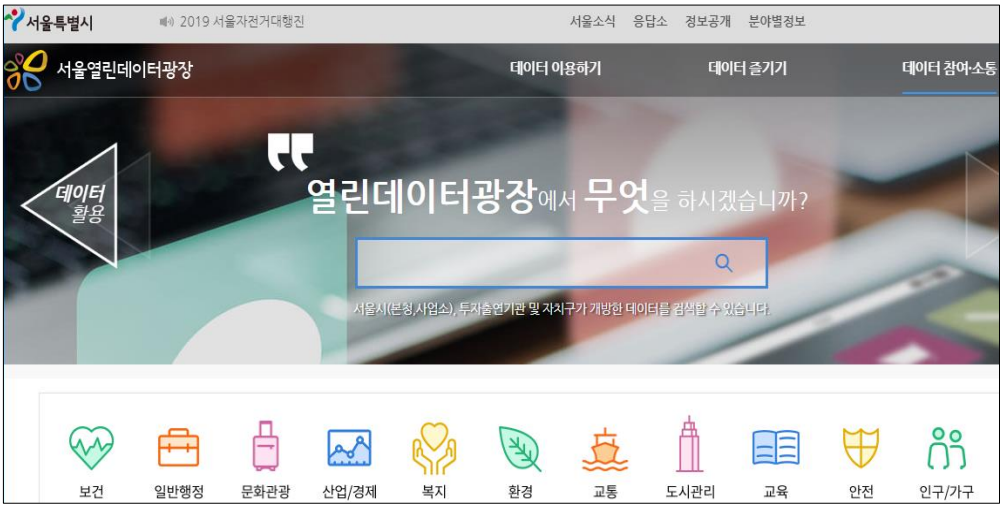
[출처 : https://www.data.go.kr/]

빅데이터 수집 방법

2 공공데이터

2 서울열린데이터광장

- 열린 시정 3.0에 의해 공공데이터를 민간에 개방하고 소통함으로써 공익성, 업무 효율성, 투명성을 높이고 시민의 자발적 참여로 새로운 서비스와 공공의 가치를 창출할 수 있도록 하는 서비스



[출처 : https://data.seoul.go.kr/]

빅데이터 수집 방법

2 공공데이터

3 국가통계포털(KOSIS, Korean Statistical Information Service)

- 국내, 국제, 북한의 주요 통계를 한 곳에 모아 이용자가 원하는 통계를 한 번에 찾을 수 있도록 통계청이 제공하는 One-Stop 통계 서비스

[출처 : http://kosis.kr/index/index.do]

빅데이터 수집 방법

3 SNS

소셜 네트워킹 서비스(Social Networking Service)

- 사용자 간의 자유로운 의사소통과 정보 공유, 인맥 확대 등을 통해
  사회적 관계를 생성하고 강화해주는 온라인 플랫폼
- 최근 스마트폰 이용자의 증가와 무선인터넷 서비스의 확장과 더불어 SNS의 이용자 또한 급증하고 있음

빅데이터 수집 방법

3 SNS

1 OPEN API

- 인터넷 이용자가 웹 검색 결과 및 사용자 화면 등을 제공받는데 그치지 않고 직접 응용 프로그램과 서비스를 개발할 수 있도록 공개된 개발자를 위한 인터페이스

- 대부분의 SNS 사이트들은 개발자로 등록하고 인증키를 받아 제공되는 API 사용

- 트위터 : https://developer.twitter.com/
- 네이버 블로그 검색 : https://developers.naver.com/docs/search/blog/
- 네이버 뉴스 검색 : https://developers.naver.com/docs/search/news/

빅데이터 수집 방법

3 SNS

2 RSS(Really Simple Syndication/Rich Site Summary)

- 뉴스나 블로그와 같이 콘텐츠 업데이트가 자주 일어나는 웹 사이트에서 업데이트된 정보를 정해진 규격의 XML 형식으로 자동화하여 사용자에게 제공하기 위한 서비스

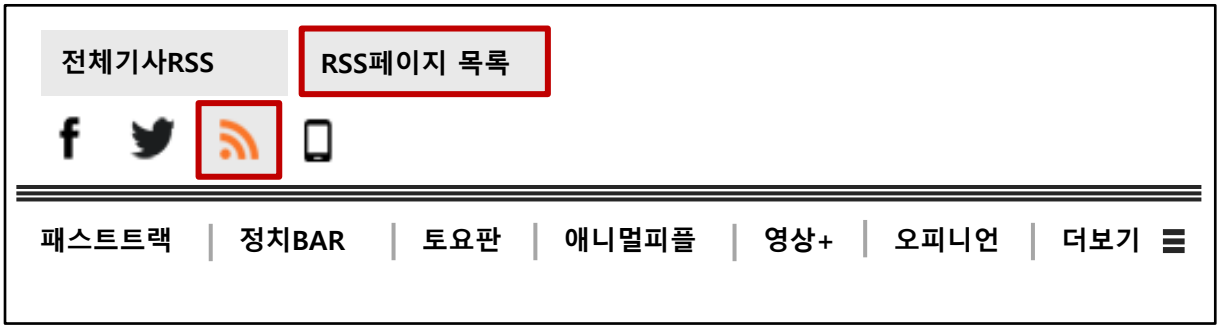
- RSS가 등장하기 전에는 원하는 정보를 얻기 위해 해당 사이트를 직접 방문해야 했음
- RSS 관련 프로그램(혹은 서비스)을 이용하여 자동 수집이 가능해졌기 때문에 사용자는 각각의 사이트 방문 없이 최신 정보들만 골라 한 자리에서 볼 수 있음

빅데이터 수집 방법

3 SNS

2 RSS(Really Simple Syndication/Rich Site Summary)

1 홈페이지에서 RSS 이미지를 클릭한 후 RSS페이지 목록 선택



빅데이터 수집 방법

3 SNS

2 RSS(Really Simple Syndication/Rich Site Summary)

2 과학 관련 RSS 콘텐츠 URL을 파악하고 이 URL로 요청한 결과 화면

This XML file does not appear to have any style information associated with it.
The document tree is shown below.

```
▼ <rss xmlns:dc="http://pu**.org/dc/elements/1.1/" version="2.0">
  ▼ <channel>
      <title> 전체기사 : 과학 : 뉴스 : OOO 뉴스 – 인터넷 OOO </title>
      <link> http://www.ha**.co.kr/arti/science/</link>
    ▼ <description>
        <![CDATA[ 인터넷 OOO – 토론이 살아 있는 생생한 인터넷 뉴스 ]]>
      </description>
      <dc:language>ko</dc:language>
      <copyright>Copyright The OOO. </copyright>
      <lastBuildDate>Sun, 28, Oct, 20**, 18:45:57 +0900</lastBuildDate>
```

```
  <item>
      <title>직원 해고를 결정하는 인공지능</title>
      <link>http://www.ha**.co.kr/arti/science//891764.html</link>
      <description>
        <![CDATA[ <table border='0px' cellpadding='0px' cellspacing='0px'
width='107px'><tr>
          <td bgcolor='#DDDDDD'
```

웹 페이지를 구성하는 기술

1 HTML과 CSS

1 HTML(HyperText Markup Language)

- 웹 페이지를 만들 때 사용하는 마크업 언어
  ➡ 태그를 사용하여 내용 작성

- 전체적으로 <html> 태그로 감싸짐
- 문서의 정보를 제공하는 <head> 태그와 브라우저에 렌더링되는 내용을 작성하는 <body> 태그로 구성

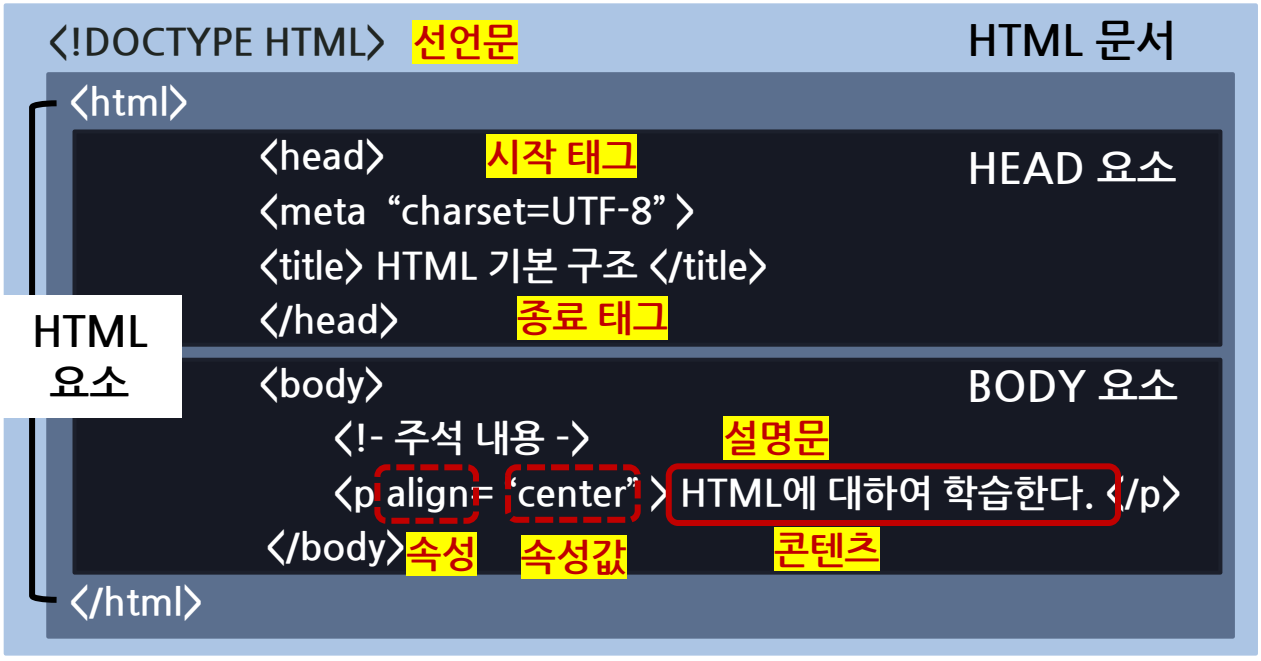
웹 페이지 데이터 추출

- 추출하려는 콘텐츠의 태그를 찾아서 속성의 값이나 콘텐츠 부분을 추출하는 것

웹 페이지를 구성하는 기술

1 HTML과 CSS

1 HTML(HyperText Markup Language)



웹 페이지를 구성하는 기술

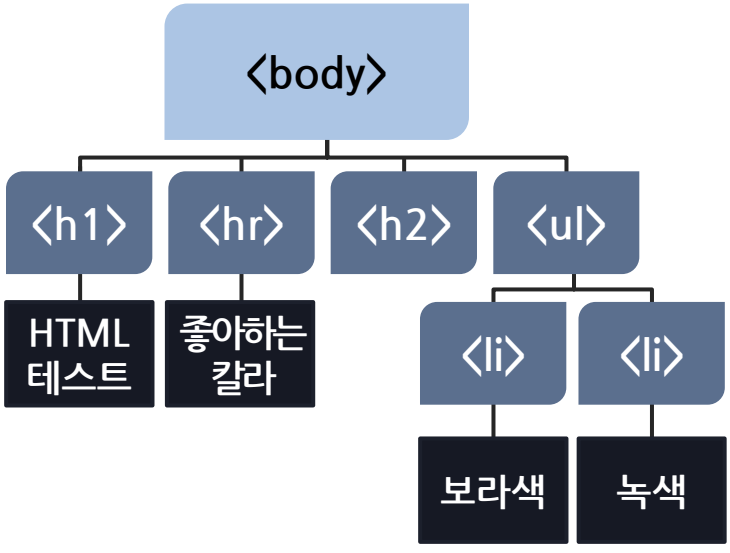
1 HTML과 CSS

1 HTML(HyperText Markup Language)

- 브라우저가 HTML 문서를 파싱하여 브라우저의 도큐먼트 영역에 렌더링할 때 HTML 문서를 구성하는 모든 태그와 속성, 콘텐츠들을 DOM(Document Object Model)이라는 규격을 적용하여 JavaScript 객체 생성

HTML 문서의 내용으로 구성되는 DOM 객체들은 HTML 문서 그대로 계층 구조를 이룸

```
<head>
  <meta charset="utf-8">
</head>
<body>
  <h1>HTML 테스트</h1>
  <hr>
  <h2>좋아하는 칼라</h2>
  <ul>
   <li>보라색</il>
   <li>보라색</il>
  </ul>
 <body>
</html>
```



웹 페이지를 구성하는 기술

1 HTML과 CSS

2 CSS(Cascade Style Sheet)

CSS	▪ HTML과 같은 마크업 언어가 브라우저에 표시되는 방법을 기술하는 언어 ▪ HTML과 XHTML에 주로 사용되는 W3C의 표준

1 CSS 사용의 이점

- 웹 표준에 기반한 웹 사이트 개발 가능(페이지의 내용과 디자인 분리)
- 클라이언트 기기에 알맞는 반응형 웹 페이지 개발 가능
- 이미지의 사용을 최소화시켜 가벼운 웹 페이지 개발 가능

웹 페이지를 구성하는 기술

1 HTML과 CSS

2 CSS(Cascade Style Sheet)

2 CSS의 작성 규칙

```
선택자 {
        CSS속성명 : CSS속성값;
        CSS속성명 : CSS속성값;
              :
}
```

```
<style>
  h1{
    color : red:
    background-color : yellow;
    width : 200px;
    border : 3px solid magenta;
    border-radius : 10px;
    padding : 3px
    text-align : center;
}
h2{
    color : blue;
    text-shadow : 2px 2px 2px skyblue;
}
</style>
```

웹 페이지를 구성하는 기술

1 HTML과 CSS

2 CSS(Cascade Style Sheet)

2 CSS의 작성 규칙

HTML 문서를 CSS 없이 렌더링한 경우

HTML 테스트

좋아하는 칼라

- 보라색
- 녹색

HTML 문서를 CSS를 적용하여 렌더링한 경우

HTML 테스트

좋아하는 칼라

- 보라색
- 녹색

웹 페이지를 구성하는 기술

1 HTML과 CSS

3 CSS 선택자

CSS 선택자 (Selector)	▪ 스타일을 적용하기 위해 대상 태그를 선택하는 방법

1 태그 선택자

- 태그명으로 태그를 선택하려는 경우로 태그명을 그대로 사용

```
h2 { color : blue; }
```

```
<h2> CSS(Cascade Style Sheet) </h2>
```

웹 페이지를 구성하는 기술

1 HTML과 CSS

3 CSS 선택자

2 클래스 선택자

- 태그에 정의된 class 속성의 값으로 태그를 선택하려는 경우로 . 과 함께 작성

```
.redtext { color : red; }
```

```
<h2  class="redtext"> CSS(Cascade Style Sheet) </h2>
```

3 id 선택자

- 태그에 정의된 id 속성의 값으로 태그를 선택하려는 경우로 #과 함께 작성

```
#t1 { color : green; }
```

```
<h2 id="t1"> CSS(Cascade Style Sheet)  </h2>
```

웹 페이지를 구성하는 기술

1 HTML과 CSS

3 CSS 선택자

4 자식 선택자

- 지정된 부모 태그의 자식 태그에만 스타일이 적용

```
section > p { color : blue; }
```

```
<section>
  <p>
   선택됨
  </p>
</section>
```

```
<nav>
  <p>
   선택되지 않음
  </p>
</nav>
```

웹 페이지를 구성하는 기술

1 HTML과 CSS

3 CSS 선택자

5 자손 선택자

- 지정된 부모 태그의 자식 태그에만 스타일이 적용

```
div  p { color : yellow; }
```

```
<div>                    </section>
  <p>                   </div>
   선택됨
  </p>
  <section>
    <p>
     선택됨
    <p>
```

6 속성 선택자

- 태그에 정의된 속성과 값으로 태그를 선택하려는 경우로 [ ]와 함께 작성

```
img[src=duke.png] { radius : 0.5; }
```

```
<img src="duke.png">
```

웹 페이지를 구성하는 기술

2 JavaScript

JavaScript	▪ 스크립트 방식으로 구현되는 OOP 프로그래밍 언어 ▪ 최근에는 다양한 기능의 프로그래밍에 사용 가능해짐 ▪ 주로 웹 페이지 개발 시 동적인 처리를 구현하기 위해 사용
JavaScript 코드	▪ <script> 태그와 함께 HTML 문서 내에 작성 ▪ xxx.js 라는 독립된 파일로 만들어 <script> 태그를 이용하여 HTML 문서에서 호출하는 방식으로 작성

웹 페이지를 구성하는 기술

2 JavaScript

정적 콘텐츠로 구성된 웹 페이지

- HTML 태그와 CSS만으로 구성되는 웹 페이지는 간단하게 콘텐츠 추출 가능

동적 컨텐츠로 구성된 웹 페이지

- JavaScript를 이용하여 웹 페이지의 콘텐츠가 동적으로 구성되는 경우, Selenium 같은 기술을 추가로 사용해야 함

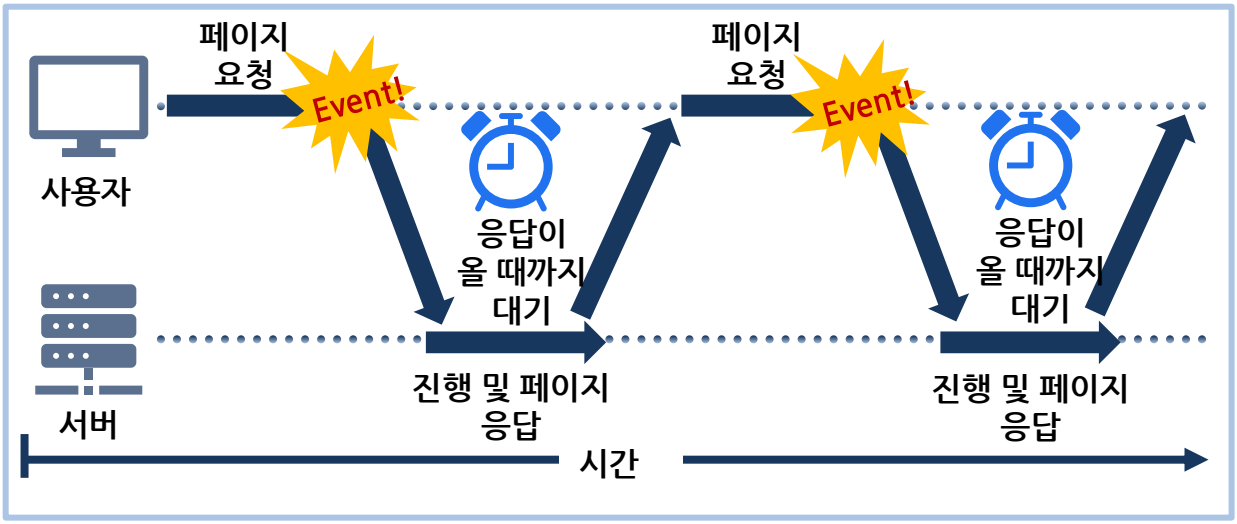
➡ 웹 크롤링을 할 때는 크롤링하려는 콘텐츠 부분이 정적으로 만들어진 것인지 JavaScript에 의해서 동적으로 만들어지는 것인지부터 파악

웹 페이지를 구성하는 기술

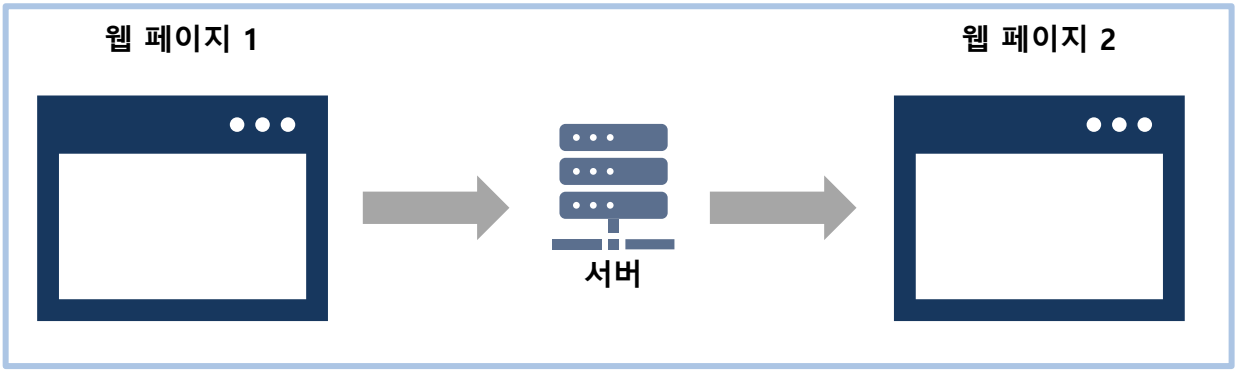
3 Ajax

1 전통적 웹 통신 방법

- 동기통신으로 서버에 요청할 때마다 응답이 올 때까지 대기



- 서버에 요청을 할 때마다 브라우저에 보여지고 있는 끝난 현재 페이지는 지워짐
- 페이지의 일부분만 변경하려는 경우에도 전체 페이지 내용을 요청하여 받아와야 함



웹 페이지를 구성하는 기술

3 Ajax

2 Ajax의 특징

Ajax	
	▪ Asynchronous JavaScript and XML의 약어 ▪ JavaScript 코드로 서버와 통신하는 기술 ▪ 통신 방식을 비동기적으로 처리하여 요청하고 나서 대기하지 않고 다른 작업을 처리할 수 있음 ▪ 전체 페이지가 아닌 필요한 일부분만 요청하여 받아올 수 있는 통신

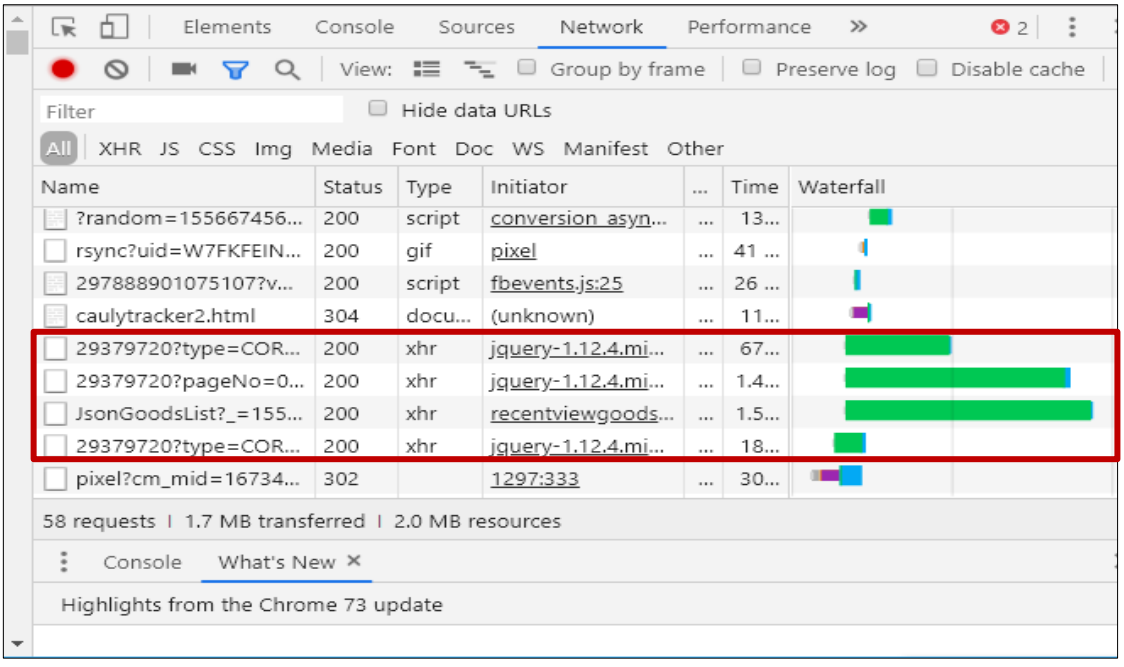
1. 전체 페이지를 리로드(Refresh)하지 않고 보여지고 있는 현재 페이지 내에 서버로부터 받아온 내용을 자연스럽게 추가할 수 있음
2. 필요한 만큼의 일부 데이터만 요청하여 받아 빠르게 동적 웹 페이지 생성
3. 데이터 전송량이 중요한 모바일 웹과 최근 많이 활용되는 SPA(Single Page Application)에서 더욱 중요해진 통신 기술

웹 페이지를 구성하는 기술

3 Ajax

3 Ajax 기술을 사용한 페이지 확인 방법

1. 크롬 브라우저의 개발자 도구를 열고 네트워크 탭 선택
2. 네트워크 탭에서 웹 브라우저와 웹 서버 간 통신 상태 정보 출력
3. 확인하고자 하는 웹 페이지 요청
4. 웹 페이지의 렌더링이 끝난 후 개발자 도구의 통신상태 정보 리스트에서 통신 Type열을 체크했을 때 xhr로 출력되는 통신 Ajax를 이용한 통신



[출처 : https://www.yes24.com/]

학습정리

1. 빅데이터 수집 방법

- 웹 크롤링 : 웹 페이지의 하이퍼링크를 순회하면서 웹 페이지를 다운로드하는 작업
- 웹 스크래핑 : 다운로드한 웹 페이지에서 필요한 콘텐츠를 추출하는 작업
- 공공데이터 : 공공기관이 전자적으로 생성 또는 취득하여 관리하고 있는 모든 데이터베이스(DB), 전자화된 파일
- SNS : 사용자 간의 자유로운 의사소통과 정보 공유, 인맥 확대 등을 통해 사회적 관계를 생성하고 강화해주는 온라인 플랫폼
- RSS : 뉴스나 블로그와 같이 콘텐츠 업데이트가 자주 일어나는 웹 사이트에서 업데이트된 정보를 정해진 규격의 XML 형식으로 자동화하여 사용자에게 제공하기 위한 서비스

학습정리

2. 웹 페이지를 구성하는 기술

- 웹 페이지를 만들 때 사용하는 HTML은 태그를 사용하여 내용 작성
- CSS : HTML과 같은 마크업 언어가 실제 표시되는 방법을 기술하는 언어로 HTML과 XHTML에 주로 사용
- CSS 선택자(Selector) : 스타일을 적용하기 위해 대상 태그를 선택하는 방법
- JavaScript를 이용하여 웹 페이지의 콘텐츠가 동적으로 구성되는 경우 Selenium과 같은 기술을 추가로 사용해야 함
- 크롬 브라우저의 개발자 도구를 사용하여 웹 페이지에서 Ajax 통신으로 콘텐츠를 구성하고 있는지 확인할 수 있음