

Data Crawling

# 데이터 크롤링

류영표 강사

ryp1662@gmail.com

Copyright © "Youngpyo Ryu" All Rights Reserved.

This document was created for the exclusive use of "Youngpyo Ryu".

It must not be passed on to third parties except with the explicit prior consent of "Youngpyo Ryu".



# 류영표

Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~5기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 (주)셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

## 강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- (주)모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 이젠 종로 아카데미(파이썬, ADSP 강사) / 강남 : ADSP
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융, 한전 KDN 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원
- 2023년 인공지능 학습용 데이터 구축사업 품질검증 전문가 자문위원
- 보건·바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

## 주요 프로젝트

### 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 페플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

# 빅데이터 처리 과정



# 데이터 유형

유형	내용	예시
정형 데이터	<ul style="list-style-type: none"><li>정형화된 스키마 구조. 주로 관계형 데이터 (RDBMS)에 저장됨</li><li>데이터 수집 난이도가 낮고, 형식이 정해져 있어 처리가 쉬운 편</li></ul>	관계형 데이터 베이스, 스프레드 시트, CSV 등
반정형 데이터	<ul style="list-style-type: none"><li>데이터 내부의 데이터 구조에 대한 메타 정보가 포함된 구조</li><li>고정된 필드에 저장되어 있지는 않지만, 메타데이터나 데이터 스키마 정보를 포함하는 데이터</li></ul>	XML, HTML, JSON, 로그 형태 (웹 로그, 센서 데이터) 등
비정형 데이터	<ul style="list-style-type: none"><li>고정 필드 및 메타데이터(스키마 포함)가 정의되지 않음.</li><li>데이터 수집 난이도가 높으며, 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움</li></ul>	소셜 데이터(트위터, 페이스북), 영상, 이미지, 음성, 텍스트(word, pdf ...) 등

\*스키마(Schema) : 데이터베이스에서 자료의 구조, 자료의 표현 방법, 자료 간의 관계를 형식 언어로 정의한 구조.

\*메타데이터(Metadata) : 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터를 말함.

# 데이터 유형

유형	종류	예시
정형 데이터	<ul style="list-style-type: none"><li>RDB, 스프레드 시트</li></ul>	ETL, FTP, Open API
반정형 데이터	<ul style="list-style-type: none"><li>HTML, XML, JSON, 웹문서, 웹로그, 센서 데이터</li></ul>	Crawling, RSS, Open API, FTP
비정형 데이터	<ul style="list-style-type: none"><li>소셜 데이터, 문서(워드, 한글), 이미지, 오디오, 비디오, IoT</li></ul>	Crawling, RSS, Open API, Streaming, FTP

- Open API 수집기술 : 웹을 운영하는 운영주체가 정보를 제공하는 수집 기술
- RSS(Really Simple Syndication) : 새 기사들의 제목만, 또는 새 기사들 전체를 뽑아서 하나의 파일로 만들어 놓은 것
- FTP(File Transfer Protocol) : 인터넷을 통해 컴퓨터 간에 파일을 전송하는 방법

# 데이터 수집 개요

“데이터를 어디에서 수집할 것 인가?”

본인이 속한 단체 내 데이터가 없는 경우, 웹 상에 공유되어 있는 데이터를 수집한다.

웹 상에 공유되어 있는 데이터를 얻는 방법은 크게 두가지이다.

1. 분석용 데이터를 제공하는 플랫폼에서 공유 데이터를 얻는 방법
2. 크롤러 프로그램으로 웹 상에 공유되어 있는 데이터를 얻는 방법



# 데이터 수집 개요

“데이터를 공유하는 플랫폼을 정리해두자“

1. AIHub : <https://www.aihub.or.kr/>
2. Kaggle : <https://www.kaggle.com/>
3. Dacon : <https://dacon.io/>
4. 공공데이터포털 : <https://www.data.go.kr/>

이 외에도 여러 플랫폼이 많이 있으나 충분한 양질의 데이터가 많지는 않다.

# World Wide Web

- 인터넷에 연결된 컴퓨터들을 통해 사람들이 정보를 공유할 수 있는 전세계적인 정보 공간을 말합니다.
- 하이퍼텍스트 형식으로 표현된 인터넷상의 다양한 정보를 효과적으로 검색하는 시스템으로 전세계적으로 가장 널리 보급되어 있다.

\* 하이퍼텍스트(Hypertext) : 참고(하이퍼링크)를 통해 독자가 한 문서에서 다른 문서로 즉시 접근할 수 있는 텍스트.





# 다양한 웹 사이트



N



DB손해보험 다이렉트

7월, 내 차 보험료 확인하면  
**스타벅스 프라푸치노 1잔 즉시 지급!**

손해보험협회 심의필 제100450호 (2023.06.10)

뉴스스탠드 · 언론사편집 / 엔터 / 스포츠 **이이 LIVE** / 경제

전체언론사 ▾ | 연합뉴스 · 학생인권조례 공방...與 "교권붕괴 원인" 野 "객관적 근거 ...

뉴스스탠드 | 뉴스들

데일리안	한국경제	아이뉴스24	ZDNET Korea	NewDaily	朝鮮日報
ChosunBiz	아시아경제	SBS	마이뉴스	YTN	The JoongAng
일간스포츠	JIK	JIJI.COM	인원일보	경기일보	충북일보
주간조선	경매리더스	CEO스코어데일리	Digital Today	MK스포츠	여성경제신문

네이버를 더 안전하고 편리하게 이용하세요

NAVER 로그인

아이디 찾기 | 비밀번호 찾기 | 회원가입

SENSE MOM

**반값에 가져가요!**

[~58%] 시원한 여름나기 >

기상특보 남양주 폭염경보



Google Search

I'm Feeling Lucky

Google offered in: [한국어](#)

# 웹 크롤링(Web Crawling)

- 웹 상에 존재하는 데이터를 자동적으로 탐색하는 행위를 의미합니다.
- 컴퓨터 소프트웨어 기술로 웹 사이트 등에서 원하는 정보를 추출하는 것.
- 인터넷에 있는 웹페이지를 방문해서 자료를 수집하는 일을 하는 프로그램을 말합니다.

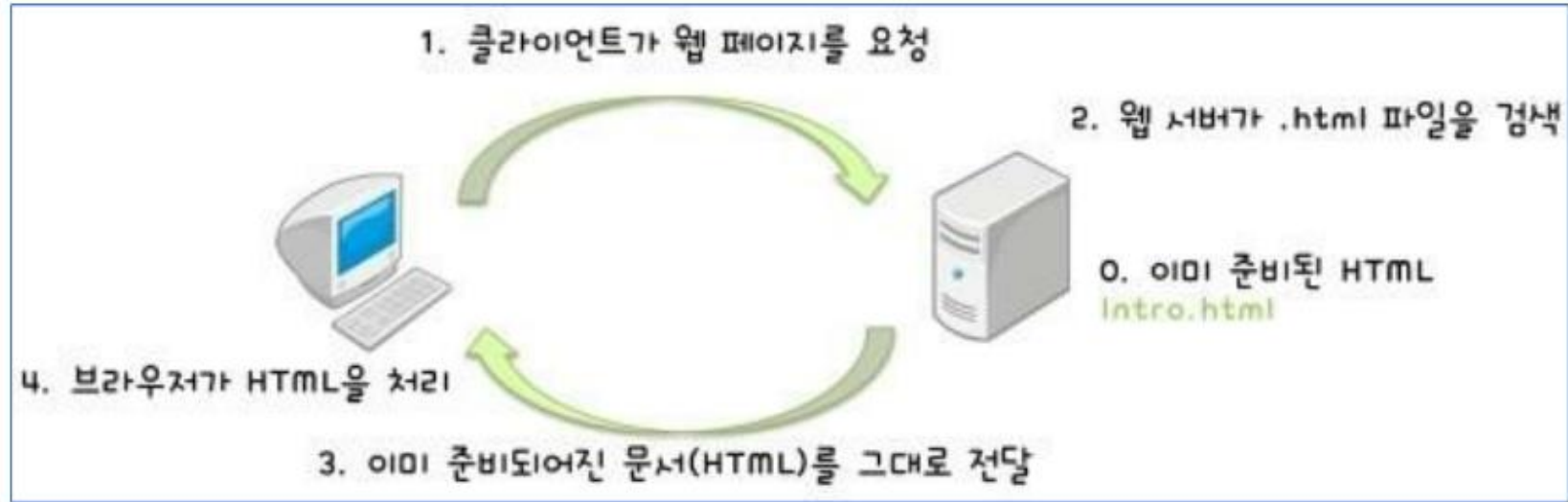


# 웹 크롤링 VS 웹 스크래핑

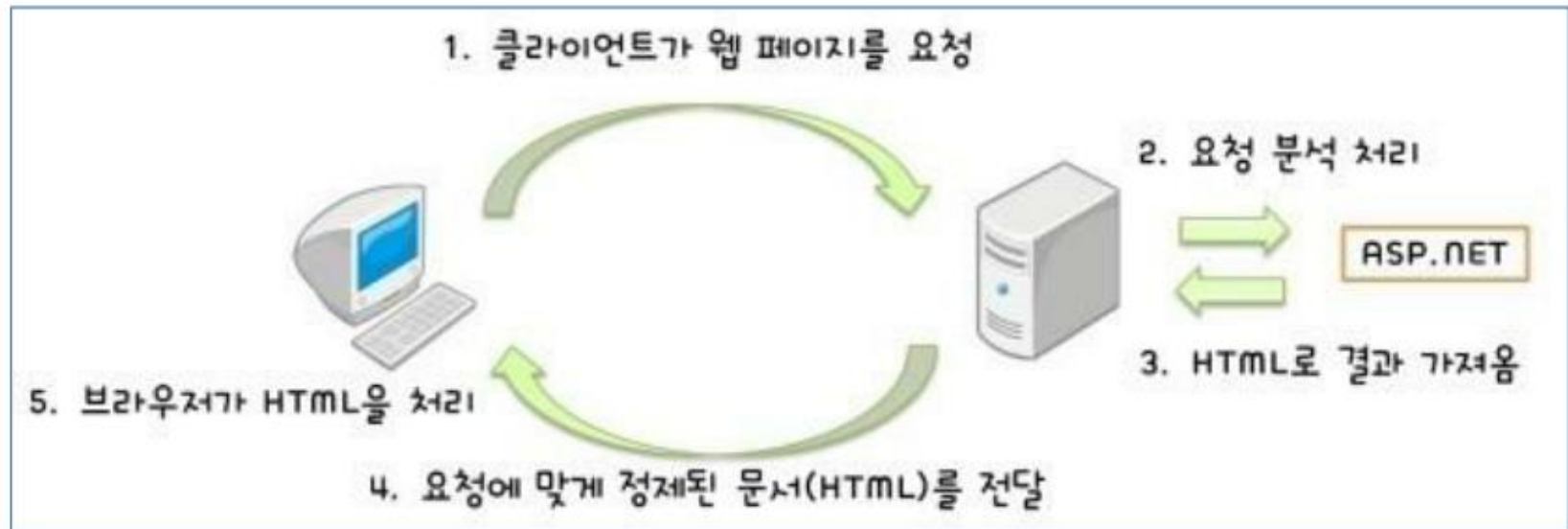
웹 크롤링 (Web Crawling)	웹 스크래핑 (Web Scraping)
웹에서 페이지 및 링크 다운로드 (웹을 기반으로 작동)	웹을 포함한 다양한 소스에서 데이터 추출 (반드시 웹과 관련된 것은 아님)
동일한 콘텐츠가 여러 페이지에 업로드 된 것을 인식하지 못하므로 중복 제거는 필수적	특정 데이터를 추출하는 것이므로 중복 제거가 반드시 필요한 것은 아님.

# 정적 페이지 VS 동적 페이지

정적:



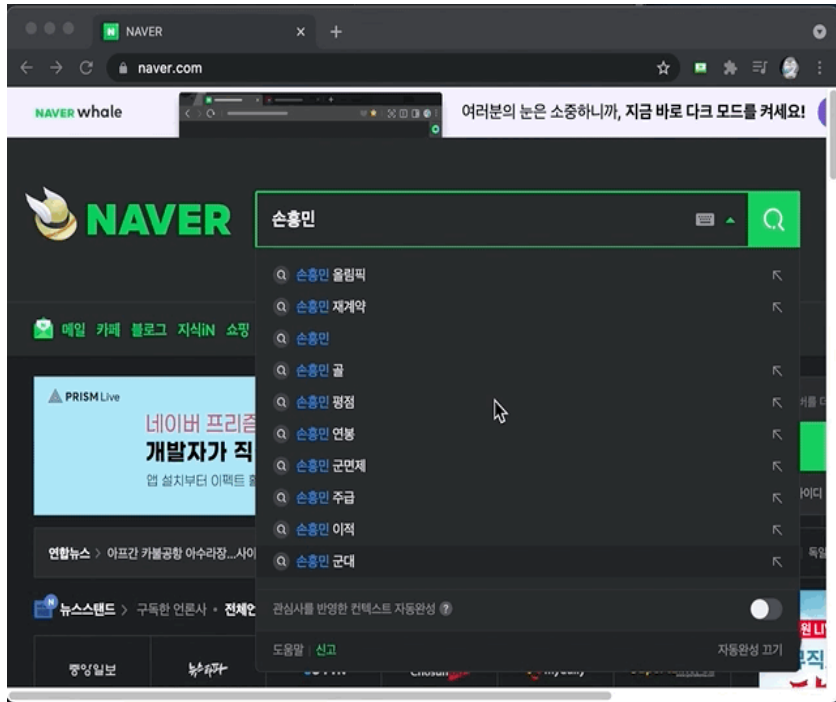
동적:



# 정적 페이지 VS 동적 페이지

- 정적 웹 페이지

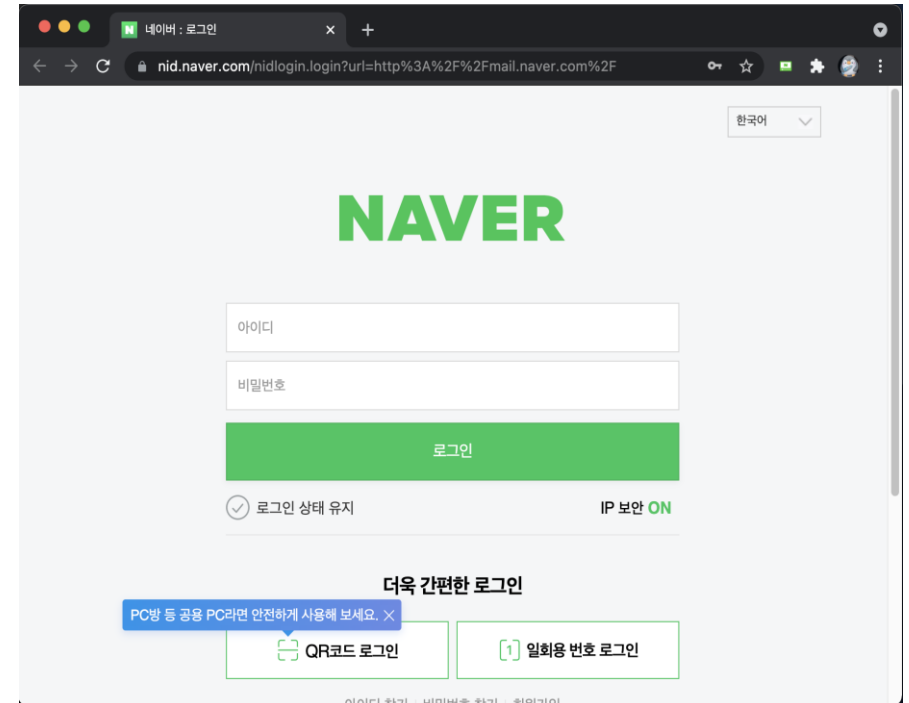
: 서버(Web Server)에 미리 저장된 파일이  
그래도 전달되는 페이지



- 1) 네이버 열어서 검색창에 '손흥민' 을 검색합니다.
- 2) 검색 결과의 url을 복사하시고, 다시 주소창에 해당 url을 입력합니다.
- 3) 처음 검색결과와 동일한 페이지를 볼 수 있습니다.

- 동적 웹 페이지(경로의 이동이 있음)

: URL만으로는 들어갈 수 없는 웹페이지를  
말합니다.



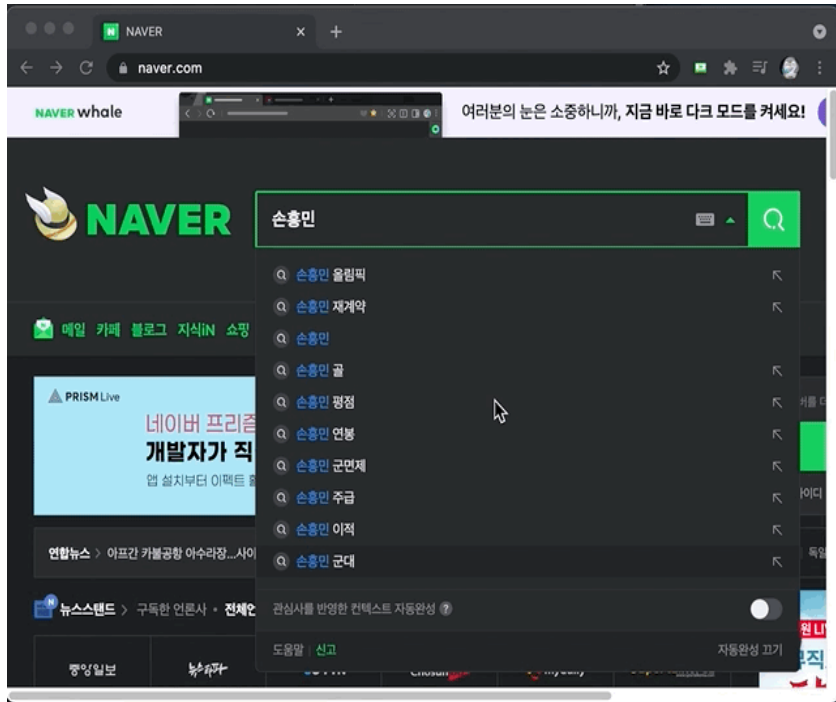
Case 1) 로그인을 해야만 접속 가능한 네이버 메일



# 정적 페이지 VS 동적 페이지

- 정적 웹 페이지

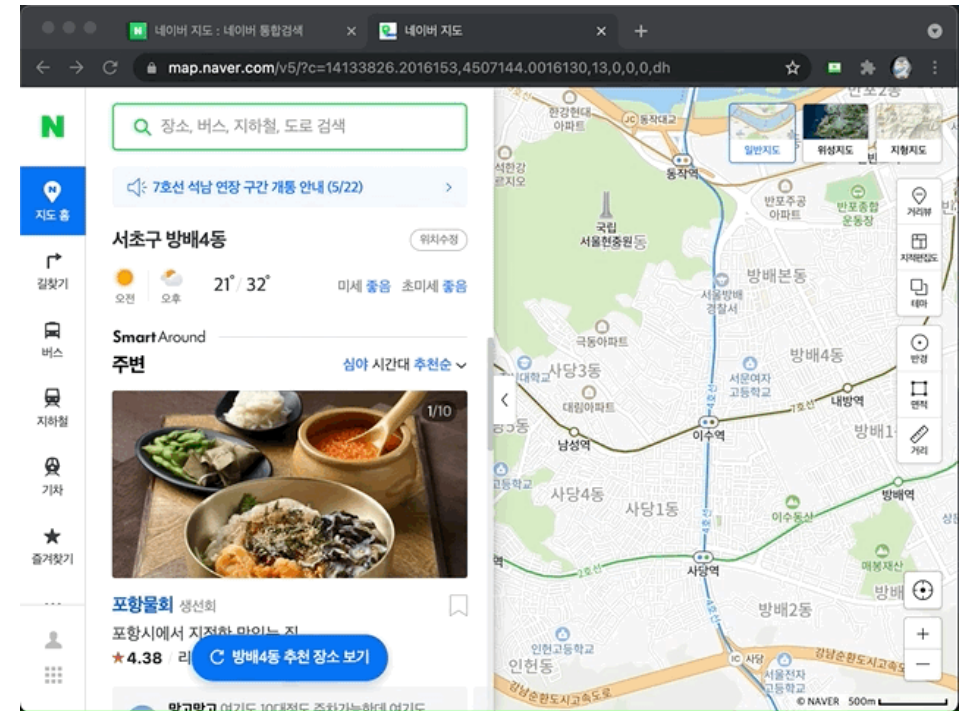
: 서버(Web Server)에 미리 저장된 파일이  
그래도 전달되는 페이지



- 1) 네이버 열어서 검색창에 '손흥민' 을 검색합니다.
- 2) 검색 결과의 url을 복사하시고, 다시 주소창에 해당 url을 입력합니다.
- 3) 처음 검색결과와 동일한 페이지를 볼 수 있습니다.

- 동적 웹 페이지(경로의 이동이 있음)

: URL만으로는 들어갈 수 없는 웹페이지를  
말합니다.

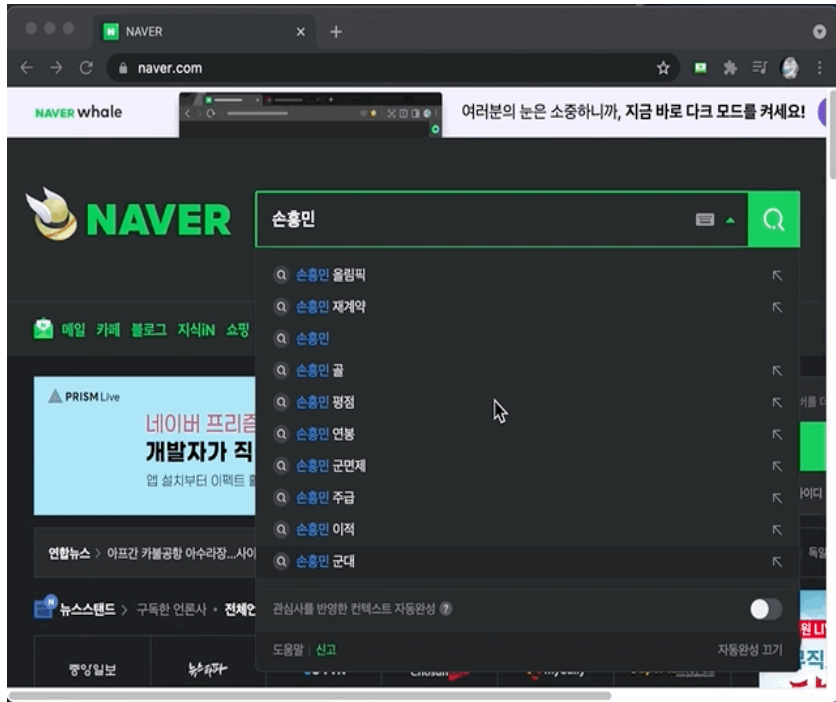


Case 2) 보고 있는 위치에 출력 결과와 url이 계속 변하는 네이버 지도

# 정적 페이지 VS 동적 페이지

- 정적 웹 페이지

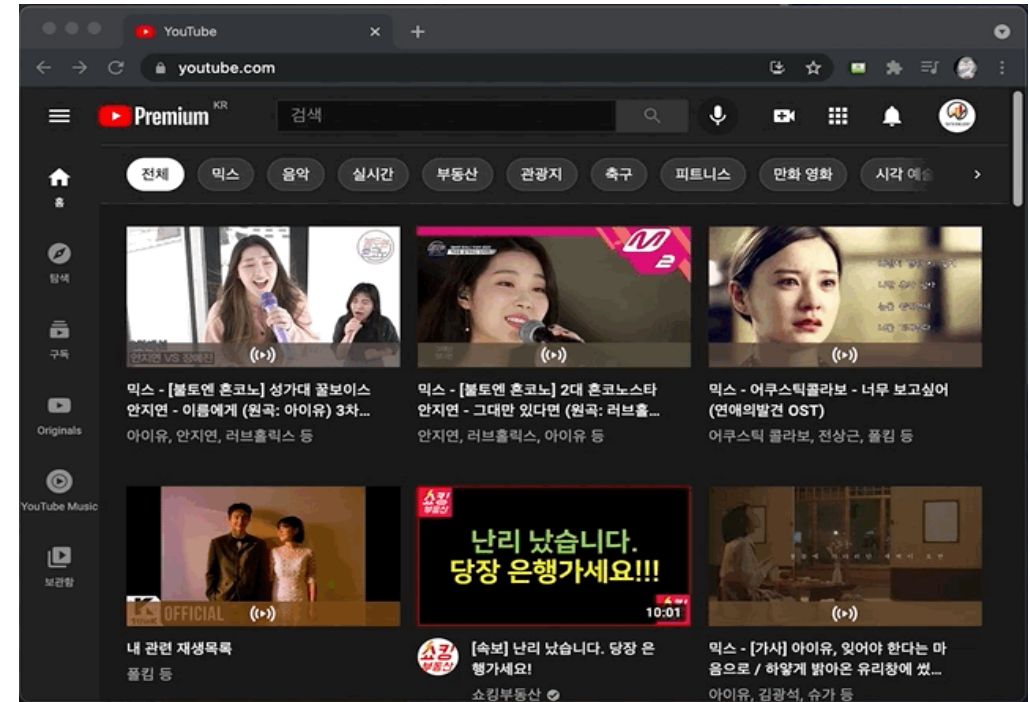
: 서버(Web Server)에 미리 저장된 파일이  
그래도 전달되는 페이지



- 1) 네이버 열어서 검색창에 '손흥민' 을 검색합니다.
- 2) 검색 결과의 url을 복사하시고, 다시 주소창에 해당 url을 입력합니다.
- 3) 처음 검색결과와 동일한 페이지를 볼 수 있습니다.

- 동적 웹 페이지(경로의 이동이 있음)

: URL만으로는 들어갈 수 없는 웹페이지를  
말합니다.



Case 3) 드래그를 아래로 내리면 계속 새로운 사진과 영상이  
나타나는 인스타그램과 유튜브

# 정적 페이지 VS 동적 페이지

	정적 크롤링	동적 크롤링
연속성	주소를 통해 단발적으로 접근	브라우저를 사용하여 연속적으로 접근
수집 능력	수집 데이터의 한계가 존재	수집 데이터의 한계가 없음
속도	빠름(별도 페이지 조작 필요 X)	상대적으로 느림
특징	모두 사용할 수 있는 범용성은 떨어짐.	수집 대상에 한계가 거의 존재 하지 않는다.
라이브러리	requests, BeautifulSoup	Selenium, Chromedriver

## 형사는 무죄, 민사는 “10억 배상”...데이터 크롤링 어디까지 되나



최민영 기자 +구독

등록 2022-10-09 16:00

수정 2022-10-09 22:12

f t talk link star printer plus

### [뉴스AS]

야놀자 vs 여기어때, 4년8개월 법정 다툼 막내려  
형사 재판에선 “서버 열어놔으니 무단 침입 아냐”  
선두 기업 투자 노력 인정 취지로 민사에선 인정  
크롤링 분쟁 핵심은 ‘부정경쟁방지법’ 위반 여부

### 사회 많이 보는 기사

1. 1300만원 들고 농막 라이프...아파트 돌아오면 생각나는 ‘집’
2. “경찰 보복인사 배후는 이상민 장관보다 앞선” 류삼영의 직격



여행·숙박 정보 플랫폼 야놀자와 여기어때의 ‘크롤링 갈등’은 6년 전으로 거슬러 올라간다. 이 사업의 후발주자인 여기어때는 2016년 1월부터 10개월 동안 크롤링 프로그램을 이용해서 야놀자의 플랫폼에 게시된 숙박업소 정보를 대량 수집해 플랫폼 영업에 이용했다. 여기어때의 데이터 복제 행위에 대해 야놀자는 2018년 초 민사소송을 제기하고 형사고소도 했다. 여기어때는 “모든 사람에게 공개된 정보로, 이 데이터에 저작권이 있다고 보기 어렵다”는 입장이었다. 반면 야놀자는 “해당 정보는 자신들이 시간과 비용 등 노력을 들여서 정보로서의 부가가치를 만든 것”이라고 맞섰다.

## 형사는 무죄, 민사는 “10억 배상”...데이터 크롤링 어디까지 되나



최민영 기자 +구독

등록 2022-10-09 16:00

수정 2022-10-09 22:12

f t talk link star printer plus

### [뉴스AS]

야놀자 vs 여기어때, 4년8개월 법정 다툼 막내려  
형사 재판에선 “서버 열어놔으니 무단 침입 아냐”  
선두 기업 투자 노력 인정 취지로 민사에선 인정  
크롤링 분쟁 핵심은 ‘부정경쟁방지법’ 위반 여부

### 사회 많이 보는 기사

1. 1300만원 들고 농막 라이프...아파트 돌아오면 생각나는 ‘집’
2. “경찰 보복인사 배후는 이상민 장관보다 뒤편” 류삼영의 직격



여행·숙박 정보 플랫폼 야놀자와 여기어때의 ‘크롤링 갈등’은 6년 전으로 거슬러 올라간다. 이 사업의 후발주자인 여기어때는 2016년 1월부터 10개월 동안 크롤링 프로그램을 이용해서 야놀자의 플랫폼에 게시된 숙박업소 정보를 대량 수집해 플랫폼 영업에 이용했다. 여기어때의 데이터 복제 행위에 대해 야놀자는 2018년 초 민사소송을 제기하고 형사고소도 했다. 여기어때는 “모든 사람에게 공개된 정보로, 이 데이터에 저작권이 있다고 보기 어렵다”는 입장이었다. 반면 야놀자는 “해당 정보는 자신들이 시간과 비용 등 노력을 들여서 정보로서의 부가가치를 만든 것”이라고 맞섰다.



# 원 데이터 저작권 보호

## 이용허락조건 (4 종류)



저작자와 출처를 표시해야 합니다.



비영리 목적으로만 사용할 수 있습니다.



변경하거나 다른 창작물에 이용하지 말아주세요.



내 저작물을 이용해 새로운 저작물을 창작한 경우,  
동일한 라이선스를 붙여야 합니다.

## CC 라이선스 (6 종류)



저작자 표시 (CC BY)



저작자표시-비영리(CC BY-NC)



저작자표시-변경금지 (CC BY-ND)



저작자표시-동일조건변경허락 (CC BY-SA)



저작자표시-비영리-동일조건 변경 허락 (BY-NC-SA)



저작자 표시-비영리-변경금지 (BY-NC-ND)



# 프론트 엔드 VS 백엔드

## 프론트엔드 VS 백엔드

### 프론트엔드

#### 언어



HTML



CSS



Javascript

#### 프레임워크



React



Angular



Bootstrap



Vue.js

#### 사용자 관점

사용자가 볼 수 있고  
상호작용할 수 있는 파트  
[클라이언트]

### 백엔드

#### 언어



Python



PHP



Ruby



Java

#### 프레임워크



Node.js



django



Flask



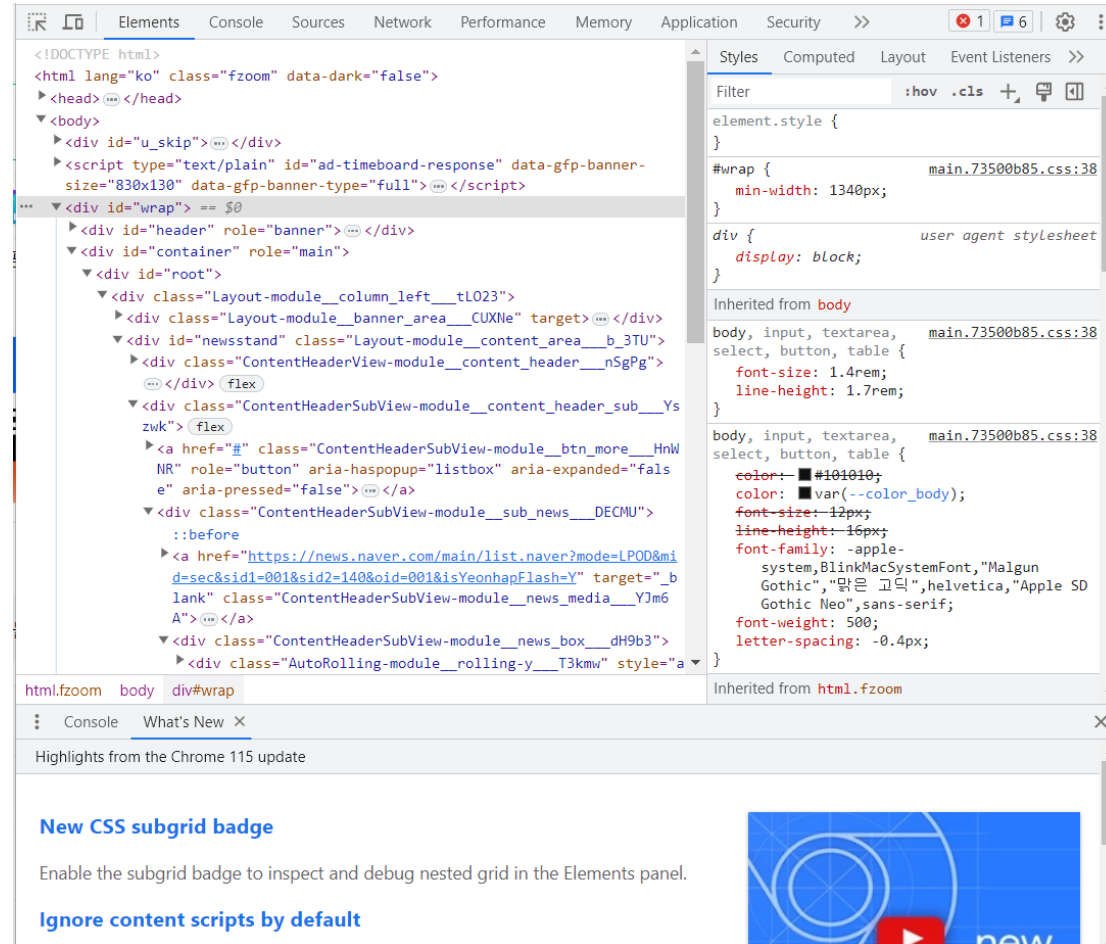
Spring

#### 사용자 관점

사용자에게 보이지 않는 파트.  
데이터베이스, 정보 처리  
[서버]

# 웹 크롤링 대상

- 웹 상의 데이터 -> HTML 혹은 JSON



- 크롬 개발자 도구 HTML 요소 찾는 방법 참고 : <https://developer-talk.tistory.com/826>

# HTML

- Hyper Text Markup Language
- 하이퍼링크를 통해 문서 사이를 옮겨 다닐 수 있는 페이지를 표시하는 언어

```
<!DOCTYPE html>
<html>
<!-- created 2010-01-01 -->
<head>
  <title>sample</title>
</head>
<body>
  <p>Voluptatem accusantium
  totam rem aperiam.</p>
</body>
</html>
```

HTML

# JSON

- JavaScript Object Notation
- 데이터를 쉽게 '교환'하고 '저장'하기 위한 텍스트 기반의 데이터 교환 표준

```
{
  "employees": [
    {
      "name": "Surim",
      "lastName": "Son"
    },
    {
      "name": "Someone",
      "lastName": "Huh"
    },
    {
      "name": "Someone else",
      "lastName": "Kim"
    }
  ]
}
```



# 하이퍼링크(Hyperlink)

- 다른 HTML 페이지로의 연결 고리입니다.
- 사이트의 html 페이지는 물론이고, 다른 웹 사이트의 html 페이지도 연결이 가능합니다.
- 즉, World Wide Web 상에서 어떤 대상에 대한 연결을 말함.
- 마크업(Markup)으로 표시하고자 하는 어떤 정보나 기능들을 태그(Tag)라는 형식으로 감싸서 사용.
- <a>, </a> 태그와 href 속성으로 만들 수 있다. A는 Anchor, href는 hyperlink refernece의 줄임말.

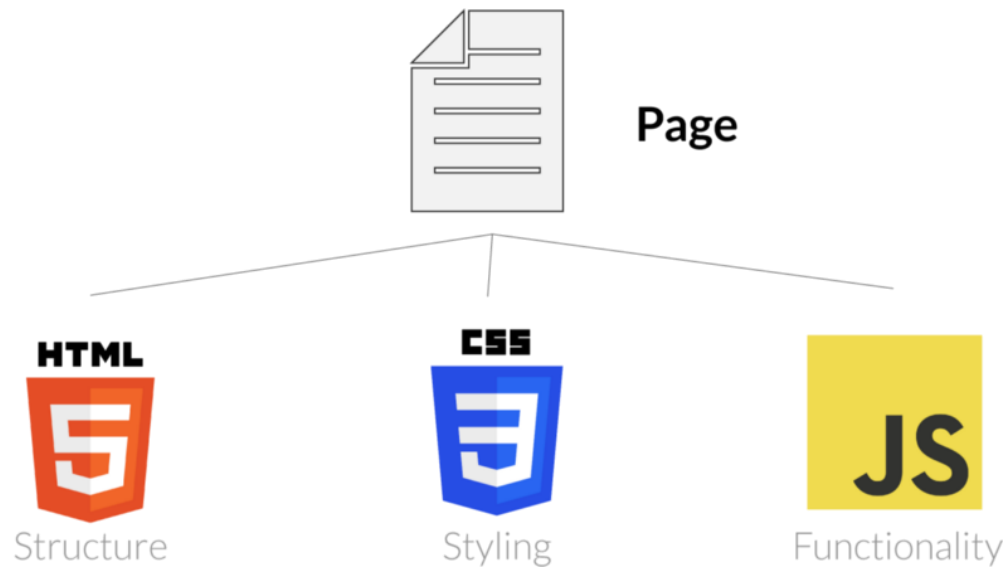
# 웹 크롤러

- 방대한 웹 페이지를 두루두루 방문하여, 각종 정보를 자동적으로 수집하는 일을 하는 프로그램으로서 검색엔진의 근간이 됨.
- 크롤러(Crawler)란 기어가는 사람 혹은 포복동물 이라는 의미로, 조직적, 자동적인 방법으로 각종 웹 페이지들을 돌아다니며 웹 문서의 URL, 링크정보, 문서내용 등 다량의 정보들을 수집해 오는 기능들로 인해 이름이 붙었습니다.



# HTML, CSS, JS

- HTML : 문장의 구조를 만드는 것
- CSS : HTML 요소의 스타일을 선택적으로 지정하는데 사용
- JS(JavaScript) : 웹 브라우저 내 동적인 요소를 구현하는 객체 기반의 스크립트 언어



# HTML

- 문서 간의 이동이 가능한 문서의 문서 형식을 정의하는데 사용하는 언어

- HTML 태그

: 문서의 모양과 행동양식을 정해주는 명령어 이름.

: < > 속에 HTML 태그 명령어의 이름을 작성하는 형태로 사용.

: 자신이 사용하고자 하는 기능을 가진 HTML 태그 명령어를 < >속에 작성하며 HTML 문서를 작성함.

```
▼<div class="AutoRolling-module__rolling-y__T3kmw" style="animation-duration: 0.5s;">
  ▼<div class="ContentHeaderSubView-module__news_box__dH9b3">
    <a href="https://media.naver.com/press/016" target="_blank" class="ContentHeaderSubView-module__news_media__YJm6A">헤럴드경제</a>
    ▼<div class="ContentHeaderSubView-module__news_desc__ztwqZ">
      ::before
      ▶<span class="ContentHeaderSubView-module__news_breaking__DZkm4">...</span>
      ▼<div class="ContentHeaderSubView-module__news_title__wuetX">
        <a href="https://n.news.naver.com/article/016/0002179488?type=breakingnews" target="_blank">한덕수 "각국 대표단 회의서 잼버리 계속 진행 결정"</a> == $0
      </div>
    </div>
  </div>
</div>
```

# HTML 태그

1. HTML은 태그들로 이루어져 있다.
2. 태그에 정보가 들어 있다.
3. 태그는 계층적인 구조로 구성되어 있다.

```
▼ <div class="AutoRolling-module__rolling-y__T3kmw" style="animation-duration: 0.5s;">
  ▼ <div class="ContentHeaderSubView-module__news_box__dH9b3">
    <a href="https://media.naver.com/press/016" target="_blank" class="ContentHeaderSubView-module__news_media__YJm6A">헤럴드경제 </a>
    ▼ <div class="ContentHeaderSubView-module__news_desc__ztwqZ">
      ::before
      ▶ <span class="ContentHeaderSubView-module__news_breaking__DZkm4">...</span>
      ▼ <div class="ContentHeaderSubView-module__news_title__wuetX">
        <a href="https://n.news.naver.com/article/016/0002179488?type=breakingnews" target="_blank">한덕수 "각국 대표단 회의서 잼버리 계속 진행 결정"</a> == $0
      </div>
    </div>
  </div>
</div>
```



# HTML 태그

1. 태그 = <시작태그> + 하위태그(or Text) + </끝태그>
2. 시작태그 = 이름 + 속성
3. 태그의 속성 = 태그의 세부 정보

```
▼ <div class="AutoRolling-module__rolling-y__T3kmw" style="animation-duration: 0.5s;">
  ▼ <div class="ContentHeaderSubView-module__news_box__dH9b3">
    <a href="https://media.naver.com/press/016" target="_blank" class="ContentHeaderSubView-module__news_media__YJm6A">헤럴드경제 </a>
    ▼ <div class="ContentHeaderSubView-module__news_desc__ztwqZ">
      ::before
      ▶ <span class="ContentHeaderSubView-module__news_breaking__DZkm4">... </span>
      ▼ <div class="ContentHeaderSubView-module__news_title__wuetX">
        <a href="https://n.news.naver.com/article/016/0002179488?type=breakingnews" target="_blank">한덕수 "각국 대표단 회의서 잼버리 계속 진행 결정"</a> == $0
      </div>
    </div>
  </div>
</div>
```

# HTML 기본

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8">
    <title>하이퍼 링크</title>
  </head>
  <body>
    <a href="https://www.naver.com/">네이버
    </a>
    <br> <!-- 'break'의 약자로 단순히 HTML에서 줄바꿈을 할 때 사용된다-->
    <a href="https://www.google.com/">구글
    </a>
  </body>
</html>
```

## 1. <html> </html>

: html 문서임을 알리는 태그로, 웹 문서의 시작과 끝에 위치할 수 있습니다.

: html 태그는 마치 사람처럼 머리 <head> 태그와 몸통에 해당되는 <body> 태그를 가짐.

# HTML 기본

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8">
    <title>하이퍼 링크</title>
  </head>
  <body>
    <a href="https://www.naver.com/">네이버
    </a>
    <br> <!-- 'break'의 약자로 단순히 HTML에서 줄바꿈을 할 때 사용된다-->
    <a href="https://www.google.com/">구글
    </a>
  </body>
</html>
```

## 2. <head></head> 태그

: 문서의 머리말, head 태그 영역에 작성된 내용은 웹브라우저 창에 표시되지 않음.

: head 태그 영역 안에 들어갈 수 있는 태그로는 <title>, <meta>, <script>, <style>태그 등이 있음.

# HTML 기본

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8">
    <title>하이퍼 링크</title>
  </head>
  <body>
    <a href="https://www.naver.com/">네이버
    </a>
    <br> <!-- 'break'의 약자로 단순히 HTML에서 줄바꿈을 할 때 사용된다-->
    <a href="https://www.google.com/">구글
    </a>
  </body>
</html>
```

### 3. <body></ body > 태그

: html 문서에서 문서의 작성자가 실제로 원하는 내용이 담기는 곳으로, 브라우저의 창 부분에 보여짐.

# HTML 기본

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8">
    <title>하이퍼 링크</title>
  </head>
  <body>
    <a href="https://www.naver.com/">네이버
    </a>
    <br> <!-- 'break'의 약자로 단순히 HTML에서 줄바꿈을 할 때 사용된다-->
    <a href="https://www.google.com/">구글
    </a>
  </body>
</html>
```

### 3. <body></ body > 태그

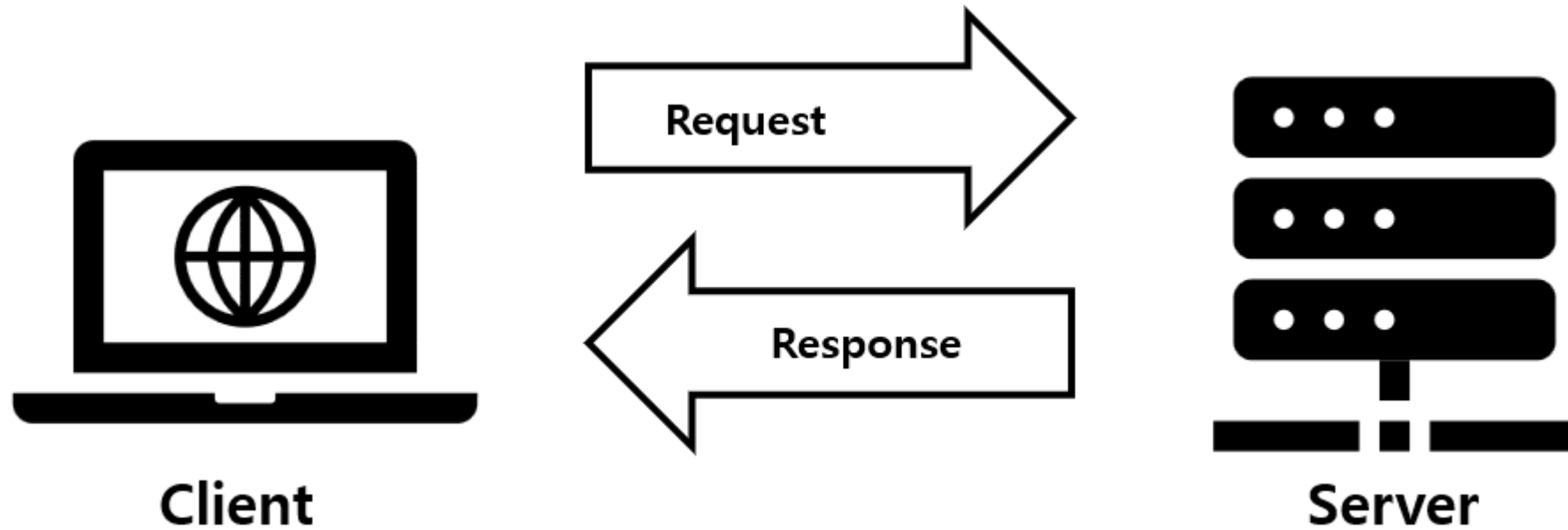
: html 문서에서 문서의 작성자가 실제로 원하는 내용이 담기는 곳으로, 브라우저의 창 부분에 보여짐.

# 웹 크롤링 과정

1. 어떤 정보를 얻고자 하는 웹 사이트에 접속하여 수집할 대상을 정한다.(ex. 연예인 사진)
2. 키보드의 F12(개발자 도구) 키를 눌러 내가 원하는 정보의 위치를 확인하고 분석합니다.
3. 해당 웹 페이지 내 원하는 정보가 있는지, 어떠한 구조로 되어 있는지 살펴본다.(정보 및 패턴 인식)
4. 불러온 데이터(html)에서 원하는 정보를 가공한 후 추출합니다.
5. 원하는 정보가 수집되면 전처리 과정을 통해 데이터(CSV, 데이터 베이스)로 활용하거나 시각화를 진행합니다.

# 서버와 클라이언트

- 서버 : 정보를 제공하는 사업자가 사용하는 컴퓨터 또는 컴퓨터 위에 설치되어 있는 소프트웨어
- 클라이언트 : 무언가를 요청하는 사람을 클라이언트

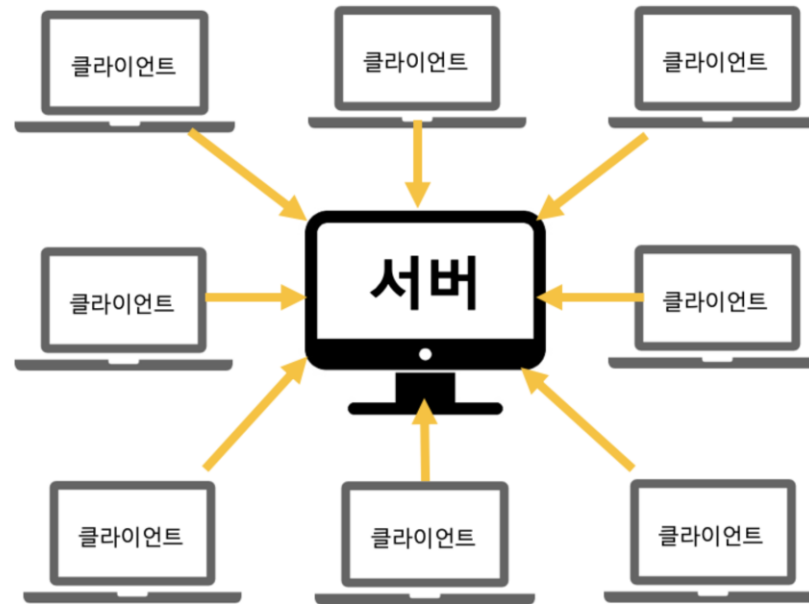




# 트래픽

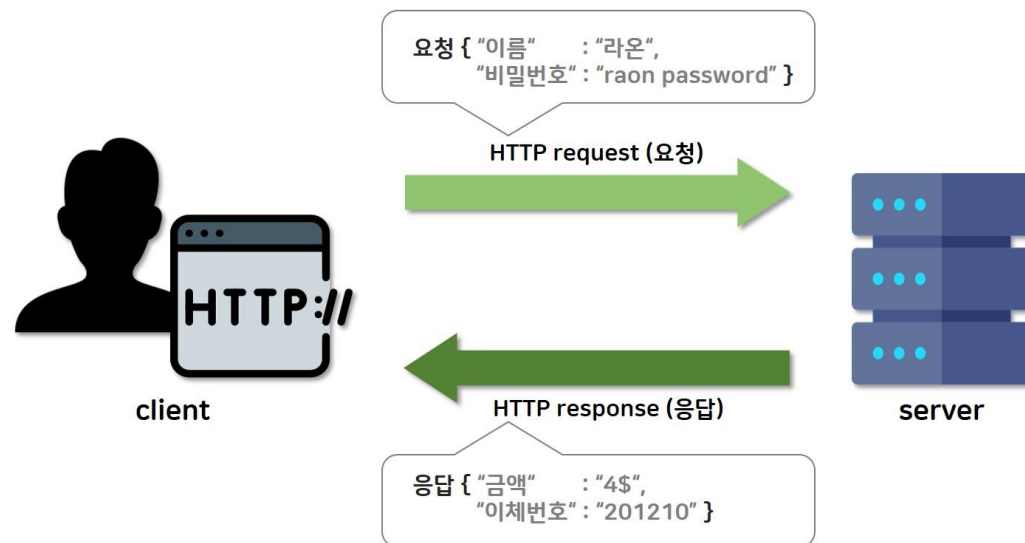
- 많은 사용자가 서비스를 이용하면 그만큼 서버에게 많은 요청이 가게 됨.  
-> 요청이 많아졌다는 말을 “트래픽이 높아졌다”라고 이야기 함.

## 트래픽



# HTTP

- HyperText Transfer Protocol
- 웹에서 데이터를 전달할 때, 사용하는 프로토콜(Protocol)
  - \* 프로토콜 : 웹에서 데이터를 주고받을 때, 지켜야 할 규칙.
- HTTP는 어떤 종류의 데이터든지 전송할 수 있도록 설계되어 있지만 주로 HTML문서를 주고 받는데 쓰임.
- 클라이언트가 서버에게 요청할 때는 어떻게 요청해야 되고 또 그 요청에 대해서 서버가 응답할 때에는 어떻게 응답해야하는가?라고 하는 것인 약속 규칙으로 미리 정해져 있음.



# 웹 크롤링하는 방법

“크롤러를 만든다는 것은 브라우저를 대신하는 프로그램을 만든다는 것”

1. 브라우저에 URL을 입력하여 원하는 웹페이지에 접근한다.
2. 전체 페이지 중에서 원하는 정보를 찾는다.

# 웹 크롤링하는 방법 : Get, Select

“2가지 파이썬 라이브러리가 필요하다.”

1. 브라우저에 URL을 입력하여 원하는 웹페이지에 접근한다. -> requests
2. 전체 페이지 중에서 원하는 정보를 찾는다. -> BeautifulSoup

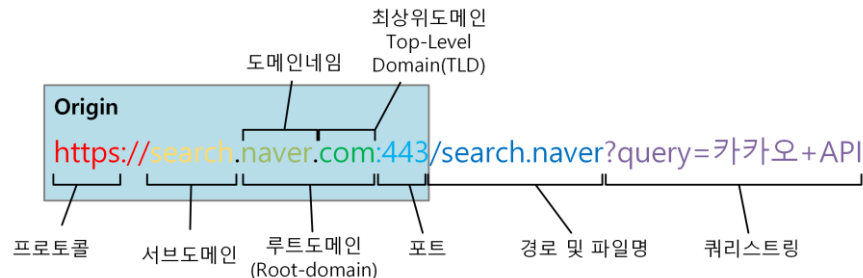
# Requests

- Python용 HTTP 라이브러리이다.
- Python에서 특정 사이트에 HTTP 요청을 보내는 모듈  
-> 특정 웹 사이트에 HTTP 요청을 보내 HTML 문서를 받아 볼 수 있는 라이브러리
- 웹페이지 = requests.get(url) -> url에 할당된 웹페이지 정보 가져오기

```
import requests  
res = requests.get('http://naver.com')  
print('응답코드 : ', res.status_code) #200이면 정상
```

응답 코드 : 200

- URL 참고)



# Requests

- Python용 HTTP 라이브러리이다.
- Python에서 특정 사이트에 HTTP 요청을 보내는 모듈
  - > 특정 웹 사이트에 HTTP 요청을 보내 HTML 문서를 받아 볼 수 있는 라이브러리
- 웹페이지 = requests.get(url) -> url에 할당된 웹페이지 정보 가져오기

```
import requests
res = requests.get('http://naver.com')
print('응답코드 : ', res.status_code) #200이면 정상

if res.status_code == requests.codes.ok:
    print('정상입니다.')
else :
    print("문제가 생겼습니다. [에러코드 ", res.status_code, "]")
```

```
응답코드 : 200
정상입니다.
```

# BeautifulSoup

- HTML, XML, JSON 등 파일의 구문을 분석하는 모듈.
  - HTML 정보로부터, 원하는 데이터를 가져오기 쉽게 비슷한 분류의 데이터별로 나누어주는(Parsing)
- ## 파이썬 라이브러리
- HTML에서 필요한 정보들을 뽑아낼 때 사용합니다.

```
[3] import requests
from bs4 import BeautifulSoup

url = 'https://comic.naver.com/webtoon/weekday.nhn'
res = requests.get(url)
res.raise_for_status() # 문제있는지 확인

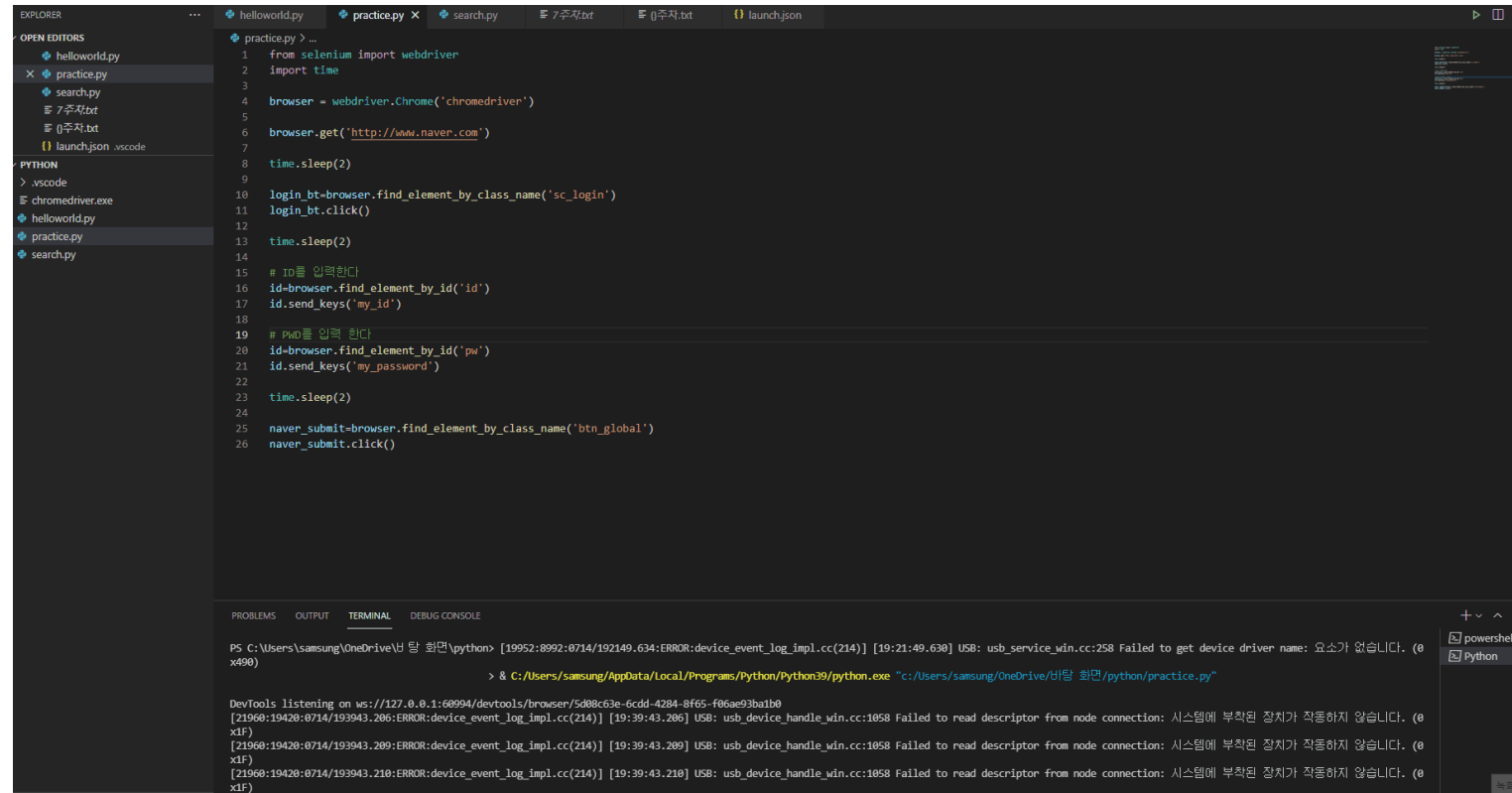
soup = BeautifulSoup(res.text, "lxml")
soup

<!DOCTYPE html>
<html lang="ko">
<head>
<title>네이버 웹툰</title>
<link href="https://ssl.gstatic.net/static/n/comic/im/favicon/1804/webtoon_favicon_32x32.ico" rel="shortcut icon" type="image/x-icon"/>
<meta charset="utf-8"/>
<meta content="ie=edge" http-equiv="x-ua-compatible"/>
<meta content="article" property="og:type"/>
<meta content="네이버 웹툰" property="og:article:author"/>
<meta content="https://comic.naver.com" property="og:article:author:url"/>
<meta content="네이버 웹툰" property="og:title"/>
<meta content="https://ssl.gstatic.net/static/comic/images/gg_tag_v2.png" property="og:image"/>
<meta content="매일매일 새로운 재미, 네이버 웹툰." property="og:description"/>
<script>
  if (/MSIE |Trident.*rv:/.test(navigator.userAgent)) {
    window.location = 'microsoft-edge:' + window.location;
    setTimeout(function () {
      window.location = 'https://go.microsoft.com/fwlink/?linkid=2135547';
    }, 1);
  }
</script>
<script async="" src="https://ssl.gstatic.net/tveta/iibs/glad/qrod/gfq-core.js"></script>
<script>
  var ccscr = 'cc.naver.com';
  window.gladsdk = window.gladsdk || { cmd: [] };
</script>
</head>
<body>
<div id="root"></div>
</body>
<script src="/runtime-5ea87e9aac5b0a7d1e87.js" type="text/javascript"></script>
<script src="/vendor-react-d37d8c657a271200d9cf.js" type="text/javascript"></script>
<script src="/vendor-react-common-39f644b98f3af612d766.js" type="text/javascript"></script>
<script src="/vendor-common-4c04532899aef03d14c.js" type="text/javascript"></script>
<script src="/vendor-log-feb99cf7b041c7e3b64d.js" type="text/javascript"></script>
<script src="/router-99a400e9d215c0b15950.js" type="text/javascript"></script>
</html>
```



# Selenium

- 웹 상에서 정적인 페이지를 탐색하는데 사용하던 BeautifulSoup같은 패키지가 하지 못하는 동적인 크롤링을 지원한다.
- 동적인 크롤링이란 url상에는 아무런 변화없이 동작하는 웹 페이지에 대한 크롤링을 의미합니다.



The screenshot shows a VS Code editor with a Python script named 'practice.py' open. The script uses Selenium WebDriver to interact with a web browser (Chrome). The code includes imports for Selenium and time, initializes a Chrome browser, navigates to 'http://www.naver.com', waits for 2 seconds, finds the login button by class name, clicks it, waits for 2 seconds, finds the ID input field by ID, sends the text 'my\_id', finds the password input field by ID, sends the text 'my\_password', waits for 2 seconds, finds the submit button by class name, and clicks it.

```
1 from selenium import webdriver
2 import time
3
4 browser = webdriver.Chrome('chromedriver')
5
6 browser.get('http://www.naver.com')
7
8 time.sleep(2)
9
10 login_bt=browser.find_element_by_class_name('sc_login')
11 login_bt.click()
12
13 time.sleep(2)
14
15 # ID를 입력한다
16 id=browser.find_element_by_id('id')
17 id.send_keys('my_id')
18
19 # PW를 입력 한다
20 id=browser.find_element_by_id('pw')
21 id.send_keys('my_password')
22
23 time.sleep(2)
24
25 naver_submit=browser.find_element_by_class_name('btn_global')
26 naver_submit.click()
```

The terminal at the bottom shows the execution of the script. It displays the path to the Python interpreter and the Selenium WebDriver binary. The output shows that the script is running successfully, with some warnings about USB device connections.

```
PS C:\Users\samsung\OneDrive\바탕 화면\python> [19952:8992:0714/192149.634:ERROR:device_event_log_impl.cc(214)] [19:21:49.630] USB: usb_service_win.cc:258 Failed to get device driver name: 요소가 없습니다. (0x490)
> & C:/Users/samsung/AppData/Local/Programs/Python/Python39/python.exe "c:/Users/samsung/OneDrive/바탕 화면/python/practice.py"
DevTools listening on ws://127.0.0.1:60994/devtools/browser/5d88c63e-6cdd-4284-8f65-f06ae93ba1b0
[21960:19420:0714/193943.206:ERROR:device_event_log_impl.cc(214)] [19:39:43.206] USB: usb_device_handle_win.cc:1058 Failed to read descriptor from node connection: 시스템에 부착된 장치가 작동하지 않습니다. (0x1f)
[21960:19420:0714/193943.209:ERROR:device_event_log_impl.cc(214)] [19:39:43.209] USB: usb_device_handle_win.cc:1058 Failed to read descriptor from node connection: 시스템에 부착된 장치가 작동하지 않습니다. (0x1f)
[21960:19420:0714/193943.210:ERROR:device_event_log_impl.cc(214)] [19:39:43.210] USB: usb_device_handle_win.cc:1058 Failed to read descriptor from node connection: 시스템에 부착된 장치가 작동하지 않습니다. (0x1f)
```



# Thank you.

데이터 크롤링 / 류영표 강사  
ryp1662@gmail.com

Copyright © “Youngpyo Ryu” All Rights Reserved.  
This document was created for the exclusive use of “Youngpyo Ryu”.  
It must not be passed on to third parties except with the explicit prior consent of “Youngpyo Ryu”.