

---

# 데이터 구축을 위한 웹 크롤링

2024. 11. 11

2024 내일배움캠프

설무아 AI튜터

# 정적 웹 크롤링

# 크롤링

---

- 웹 사이트에서 자동화된 방법으로 데이터를 수집하는 과정
- 주요 절차
  1. 크롤링할 Web URL 분석
  2. HTTP GET 요청 송신 및 응답 수신
  3. HTML 파싱
    - BeautifulSoup 등의 파서를 이용해 HTML에서 원하는 정보 추출
  4. 데이터 정제 후 저장
    - CSV, JSON 등의 형태로 파일로 저장
    - 데이터 베이스에 저장

# 크롤링 코드 예시

---

```
import requests
from bs4 import BeautifulSoup

# 웹 페이지 요청
url = "https://example.com"
response = requests.get(url)

# HTML 파싱
soup = BeautifulSoup(response.text, 'html.parser')

# 데이터 추출 (예: 모든 제목 태그)
titles = soup.find_all('h1')

# 결과 출력
for title in titles:
    print(title.text)
```

# 웹의 URL

- 웹의 자원은 URL을 통해 요청

http://www.domain.com:1234/path/to/resource?a=b&x=y

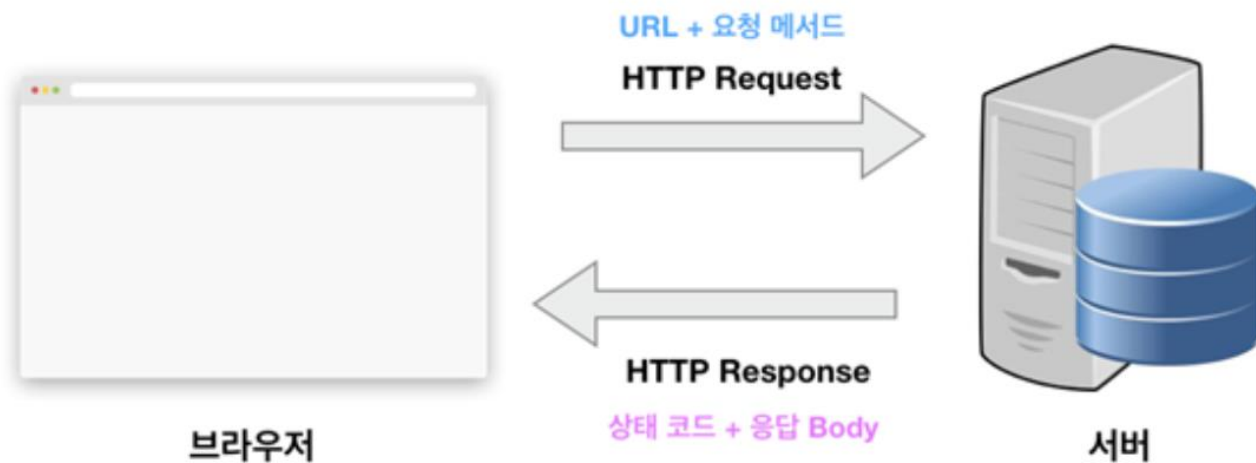
↑  
protocol

↑  
host

↑  
port

↑  
resource path

↑  
query



# HTTP 요청 메소드

---

메소드(Method)

데이터 조회

GET



GET Request

데이터 추가

POST



POST Request

데이터 수정

PUT



PUT Request

데이터 삭제

DELETE



DELETE Request

# 크롤링 허용 여부 확인하기


- ‘크롤링할 주소/ robots.txt’를 입력
- robots.txt 파일이 없다면 수집에 대한 정책이 없으니 크롤링을 해도 된다는 의미

표시	허용 여부
User-agent: * Allow: / 또는 User-agent: * Disallow:	모든 접근 허용
User-agent: * Disallow: /	모든 접근 금지
User-agent: * Disallow: /user/	특정 디렉토리만 접근 금지

# 크롤링 허용 여부 확인하기

---

- 네이버 robots.txt

 robots.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

User-agent: \*

Disallow: /

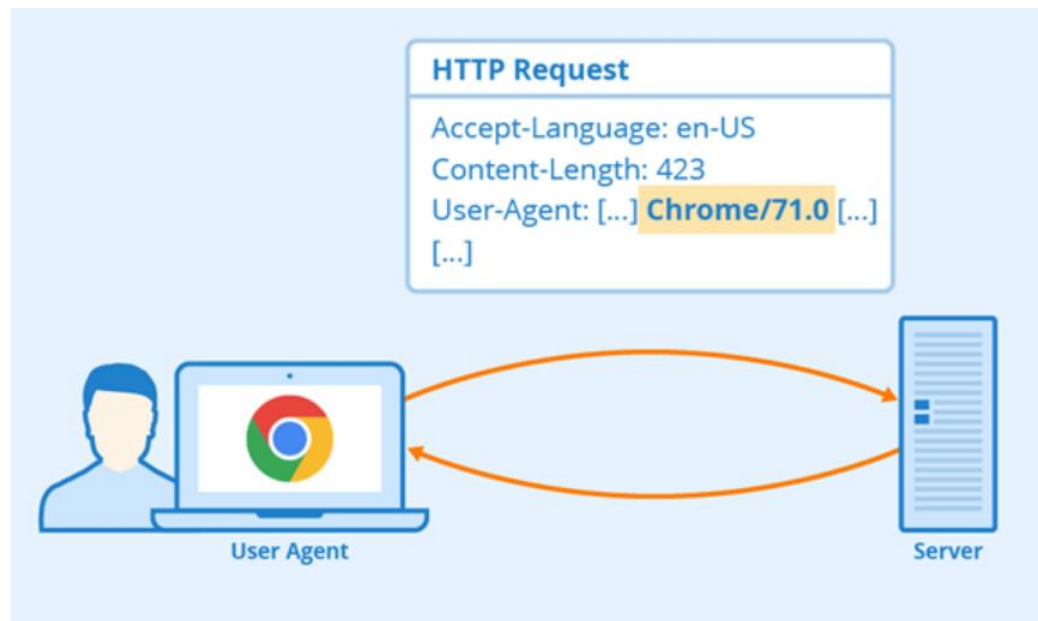
Allow : /\$

Allow : /.well-known/privacy-sandbox-attestations.json



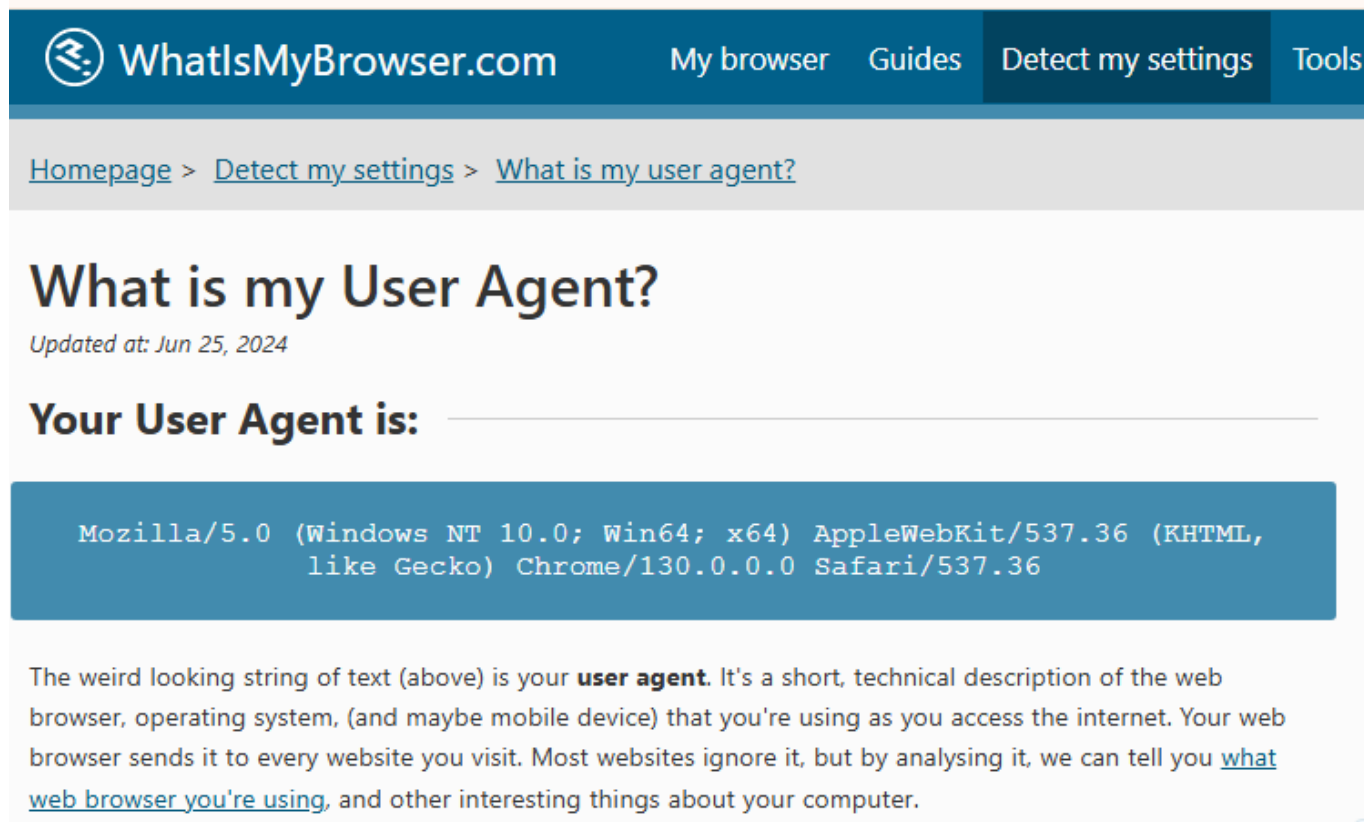
# HTTP 요청 시 : User-Agent

- 사용자 에이전트 (User Agent) : 웹 브라우저나 애플리케이션이 웹 서버에 자신을 식별하기 위해 보내는 문자열로, 브라우저 종류, 버전, 운영체제 등의 정보를 포함.



# HTTP 요청 시 : User-Agent

- 나의 User-Agent 확인



The screenshot shows the homepage of WhatIsMyBrowser.com. The navigation bar includes links for 'My browser', 'Guides', 'Detect my settings' (which is highlighted), and 'Tools'. A breadcrumb trail shows the path: 'Homepage > Detect my settings > What is my user agent?'. The main heading is 'What is my User Agent?' with a subtext 'Updated at: Jun 25, 2024'. Below this, it says 'Your User Agent is:' followed by a text box containing the user agent string: 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Safari/537.36'. A paragraph explains that this string is the user agent, a technical description of the browser and system used to access the internet, which is sent to every website visited.

WhatIsMyBrowser.com My browser Guides Detect my settings Tools

[Homepage](#) > [Detect my settings](#) > [What is my user agent?](#)

## What is my User Agent?

Updated at: Jun 25, 2024

Your User Agent is: \_\_\_\_\_

```
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Safari/537.36
```

The weird looking string of text (above) is your **user agent**. It's a short, technical description of the web browser, operating system, (and maybe mobile device) that you're using as you access the internet. Your web browser sends it to every website you visit. Most websites ignore it, but by analysing it, we can tell you [what web browser you're using](#), and other interesting things about your computer.

## 웹과 HTML

- HTML은 웹상의 정보를 구조적으로 표현하기 위한 언어



<h1>손하트 날리는 BTS 지민</h1>

<image src=http://img.yonhapnews.co.kr/photo/yna/YH/2018/10/07//PYH2018100706920007200\_P4.jpg>

<p>(뉴욕=연합뉴스) 이준서 특파원 = 세계적 케이팝 그룹 방탄소년단(BTS)이 6일 밤(현지시간) 미국 뉴욕의 시티필드에서 '러브 유어셀프'(Love Yourself) 북미투어의 대미를 장식하는 피날레 공연을 하고 있다. 2018.10.7 [빅히트 엔터테인먼트 제공]</p>

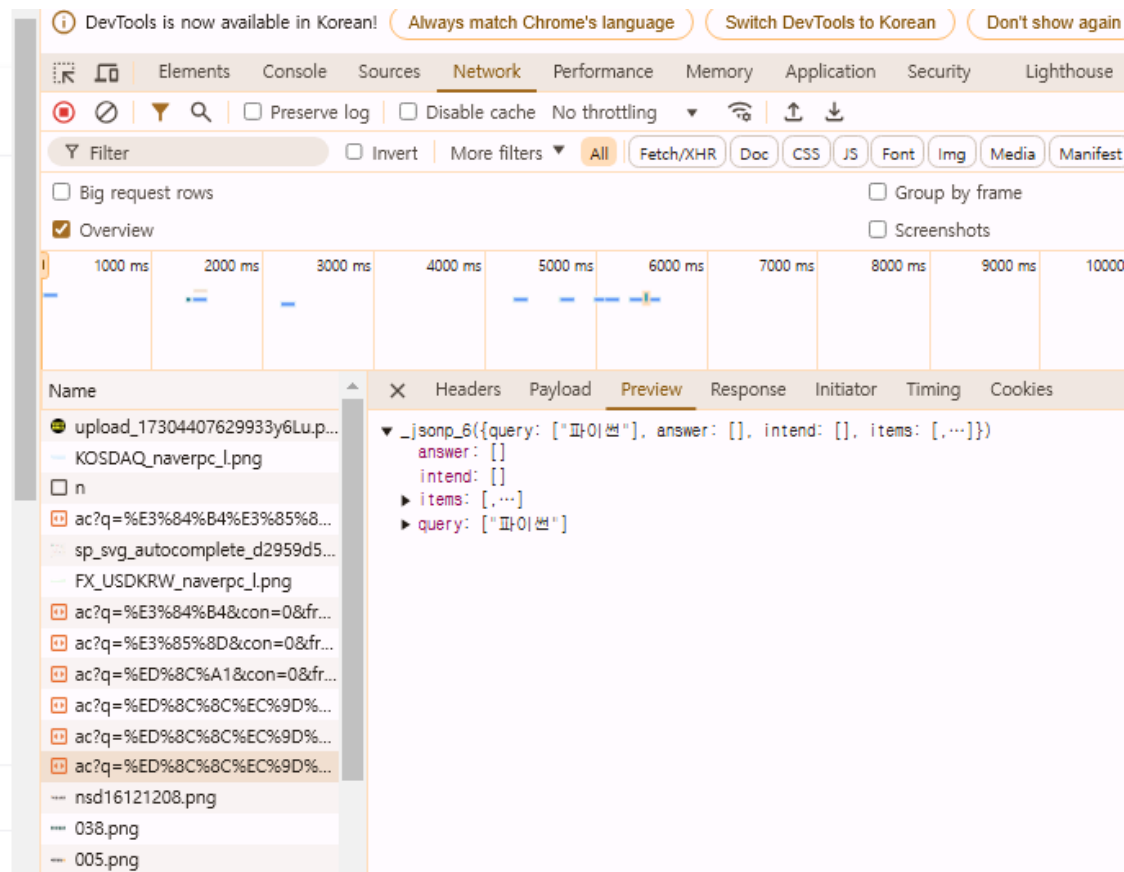
# HTML과 태그

---

- HTML은 태그들로 이루어져 있음
  - 주요 태그 역할:
    - `<html>`: HTML 문서의 시작과 끝
    - `<head>`: 문서의 메타데이터, 제목 등 정보 포함
    - `<body>`: 실제 웹 페이지에 표시되는 내용
    - `<div>`: 구역을 나누는 컨테이너
    - `<p>`: 문단
    - `<a>`: 하이퍼링크
    - `<img>`: 이미지 삽입
    - `<ul>`, `<ol>`, `<li>`: 목록 만들기
    - `<table>`: 표 만들기
    - `<form>`: 입력 양식

## 1-1. 크롤링 할 Web URL 분석

- 네이버 연관 검색어 가져오기



## 1-2. 크롤링 할 Web URL 분석

### ■ 할리스 커피 매장 정보 가져오기

\* 할리스 전 매장에서 무선인터넷 서비스 이용이 가능합니다 (단, 휴게소 및 특수매장 제외)

지역	매장명	현황	주소	매장 서비스	전화번호
인천 남동구	간석오거리점2	영업중	인천광역시 남동구 남동대로 931 (간석동) 씨앤티이 워딩홀	☎	032-425-0915
대구 수성구	대구범어천로점2	영업중	대구광역시 수성구 범어천로 200 (범어동, 범어월드메르디앙웨스턴 카운티) 102호, 201호	☎	053-759-5779
서울 강동구	길동푸유르센티점	영업중	서울특별시 강동구 진항대로 104 (길동) 1층 (101~102호)		02-487-9997
서울 강동구	고덕비즈밸리점	영업중	서울특별시 강동구 고덕비즈밸리로 38 (고덕동) KS타워 101호 ~102호	↑ ☎	.
경남 창원시 의창구	창원팔용점	영업중	경상남도 창원시 의창구 창원대로 18번길 6-8 (팔용동) 할리스 창원 팔용점	↑ ↻ ☎	.
서울 강남구	파르나스점	영업중	서울특별시 강남구 테헤란로 521 (삼성동, 파르나스타워) 파르나스 풀 내 B1층 F-3a호	☎	02-3453-1024
강원 동해시	동해록호점	영업중	강원특별자치도 동해시 해안로 368 (평릉동, 할리스커피) 1~3층	↑ ☎	033-534-8902
충남 홍성군	홍남도청점	영업중	충청남도 홍성군 홍북읍 청사로 174번길 25 (성원타워)	☎	041-631-4725
서울 동작구	신대방삼거리역점	영업중	서울특별시 동작구 상도로 60 (대방동) 1층~2층	☎	02-823-2377
세종	세종국책연구단지점	영업중	세종특별자치시 시청대로 370 (반곡동, 나라키움세종국책연구단지)	☎	044-865-4684

☎ hollys.co.kr/store/korea/korStore2.do?pageNo=1&sid=&gugun=&store=



## 2. HTTP 요청 (GET)

---

```
import requests

def get_hollys_store_info(page: int) -> str:
    """
    할리스 매장 정보 페이지의 HTML을 가져오는 함수
    """
    url = "https://www.hollys.co.kr/store/korea/korStore2.do"
    params = {"pageNo": page}
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) Chrome/120.0.0.0'
    }
    try:
        response = requests.get(url, params=params, headers=headers)
        response.raise_for_status()
        return response.text
    except requests.exceptions.RequestException as e:
        print(f"페이지 {page} 요청 중 에러 발생: {e}")
        return ""
```

## 3. HTML 파싱

- BeautifulSoup 라이브러리를 사용하기
  - 라이브러리 설치 `pip install beautifulsoup4`
  - 공식 문서 : <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>





# 4. 데이터 정제 후 저장

- JSON : 계층적 구조 표현이 가능하고 다양한 데이터 타입을 지원하는 데이터 포맷
- CSV : 표로 구분된 단순한 표 형태의 데이터 포맷

```
{
  "region": "충남 홍성군",
  "name": "충남도청점",
  "status": "영업중",
  "address": "충청남도 홍성군 홍북읍 청사로174번길 25 (성원타워) .",
  "service": "주차",
  "phone": "041-631-4725"
},
{
  "region": "서울 동작구",
  "name": "신대방삼거리역점",
  "status": "영업중",
  "address": "서울특별시 동작구 상도로 60 (대방동) 1층~2층",
  "service": "주차",
  "phone": "02-823-2377"
},
```

```
region,name,status,address,service,phone
인천 남동구,간석오거리점2,영업중,인천광역시 남동구 남동대로 931 (간석동
대구 수성구,대구범어천로점2,영업중,"대구광역시 수성구 범어천로 200 (범
서울 강동구,길동포유르센티점,영업중,서울특별시 강동구 진황도로 104 (길
서울 강동구,고덕비즈밸리점,영업중,서울특별시 강동구 고덕비즈밸리로 38 (
경남 창원시 의창구,창원팔용점,영업중,경상남도 창원시 의창구 창원대로185
서울 강남구,파르나스몰점,영업중,"서울특별시 강남구 테헤란로 521 (삼성동
강원 동해시,동해목호점,영업중,"강원특별자치도 동해시 해안로 368 (평릉동
충남 홍성군,충남도청점,영업중,충청남도 홍성군 홍북읍 청사로174번길 25
서울 동작구,신대방삼거리역점,영업중,서울특별시 동작구 상도로 60 (대방동
세종,세종국책연구단지점,영업중,"세종특별자치시 시청대로 370 (반곡동, N
부산 사하구,동아대 인문대점,영업중,"부산광역시 사하구 낙동대로550번길 3
```

동일업종 PER ▶ 17.13배  
동일업종 등락률 ▶ -3.68%

# 삼성전자 주가 데이터

- 날짜, 종가, 전일비, 시가, 고가, 저가, 거래량

```
날짜,종가,전일비,시가,고가,저가,거래량
2024.11.11,55000,"하락2,000",56700,56800,55000,29587277
2024.11.08,57000,하락500,58000,58300,57000,13877396
2024.11.07,57500,상승200,56900,58100,56800,17043102
2024.11.06,57300,하락300,57600,58000,56300,22092218
2024.11.05,57600,"하락1,100",57800,58100,57200,17484474
2024.11.04,58700,상승400,58600,59400,58400,15586947
2024.11.01,58300,하락900,59000,59600,58100,19083180
2024.10.31,59200,상승100,58500,61200,58300,35809196
2024.10.30,59100,하락500,59100,59800,58600,19838511
2024.10.29,59600,"상승1,500",58000,59600,57300,28369314
2024.10.28,58100,"상승2,200",55700,58500,55700,27775009
2024.10.25,55900,하락700,56000,56900,55800,25829315
2024.10.24,56600,"하락2,500",58200,58500,56600,31499922
2024.10.23,59100,"상승1,400",57500,60000,57100,27300780
2024.10.22,57700,"하락1,300",58800,58900,57700,27582527
2024.10.21,59000,하락200,59000,59600,58500,18514905
2024.10.18,59200,하락500,59900,60100,59100,14420260
2024.10.17,59700,상승200,59400,60100,59100,23372873
2024.10.16,59500,"하락1,500",59400,60000,59200,23303268
2024.10.15,61000,상승200,61100,61400,60100,22715239
2024.10.14,60800,"상승1,500",59500,61200,59400,20886249
```

Thank you