

2차시

데이터 수집하기



I . API 알아보기

- API 란?

Application Programming Interface: 하나의 소프트웨어가 다른 소프트웨어와 소통할 수 있도록 소통 방법을 정해둔 규칙.



→ 데이터를 제공하는 API를 사용할 수 있으면 처음부터 데이터를 구축할 필요가 없다!

API 종류	특징
REST API	HTTP 프로토콜 기반, JSON 형식 사용, GET/POST/PUT/DELETE 메서드
SOAP API	XML 기반, 보안성이 강하지만 상대적으로 무겁고 느림
GraphQL API	클라이언트가 원하는 데이터만 선택적으로 요청 가능, REST보다 유연하고 빠름
WebSocket API	실시간 양방향 통신 지원, 채팅/주식 거래/실시간 알림 등에 사용

- HTTP 란?

Hyper Text Transfer Protocol: 인터넷을 통해 HTML 또는 JSON 파일 형태의 데이터를 주고받는 데에 사용하는 통신 규칙.

- HTTP와 HTTPS의 차이

구분	HTTP	HTTPS
보안	암호화 없음	SSL/TLS 암호화 적용
사용목적	일반적인 웹페이지	로그인, 결제, 개인정보 전송 등
주소형식	http://example.com	https://example.com

http 요청 예시

```
GET /index.html HTTP/1.1
Host: www.example.com
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64)
Accept: text/html,application/xhtml+xml,application/xml;q=0.9
Accept-Language: en-US,en;q=0.5
Connection: keep-alive
```

http 요청시 실제 전달되는 메시지

```
GET /index.html HTTP/1.1\r\nHost: www.example.com\r\nUser-Agent: Mozilla/5.0 (Win\r\ndows NT 10.0; Win64; x64)\r\nAccept: text/html,application/xhtml+xml,application/\r\nxml;q=0.9\r\nAccept-Language: en-US,en;q=0.5\r\nConnection: keep-alive\r\n
```

http 응답 예시

```
HTTP/1.1 200 OK
Date: Sun, 03 Mar 2025 12:34:56 GMT
Server: Apache/2.4.41 (Ubuntu)
Content-Type: text/html; charset=UTF-8
Content-Length: 125
```

```
<html>
<body>
  <h1>Welcome to Example</h1>
</body>
</html>
```

http 응답시 실제 전달되는 메세지

```
HTTP/1.1 200 OK\r\nDate: Sun, 03 Mar 2025 12:34:56 GMT\r\nServer: Apache/2.4.41
(Ubuntu)\r\nContent-Type: text/html; charset=UTF-8\r\nContent-Length: 125\r\n
<html>\r\n<body>\r\n  <h1>Welcome to Example</h1>\r\n</body>\r\n</html>\r\n
```

https 암호화 예시

```
Ⓜ?İýÜäİ$F.ìóëë@Ü%öÔß.ÔçpbÆÝþÿöë
óø@öÈÄÄðäöÊ€İýİßÿİþÿİôİÿpßöûð
```

- JSON이란?

JavaScript Object Notation: 인간이 읽을 수 있는 텍스트 키와 값 쌍으로 이루어진 데이터 객체를 전달하기 위해 사용하는 개방형 표준 포맷.

JSON 객체 예시

```
{
  "name": "홍길동",
  "age": 30,
  "isStudent": false,
  "hobbies": ["독서", "축구", "코딩"],
  "address": {
    "city": "서울",
    "zipcode": "12345"
  }
}
```

API에 JSON 객체를 요청했을 때 응답 예시

```
HTTP/1.1 200 OK\r\nDate: Sun, 03 Mar 2025 12:34:56 GMT\r\nContent-Type: application/json; charset=UTF-8\r\nContent-Length: 156\r\n{\r\n  "name": "홍길동",\r\n  "age": 30,\r\n  "isStudent": false,\r\n  "hobbies": ["독서", "축구", "코딩"],\r\n  "address": {\r\n    "city": "서울",\r\n    "zipcode": "12345"\r\n  }\r\n}
```

Q. JSON 데이터를 어떻게 활용할 수 있을까?

II. Python으로 JSON 객체 다루기

1. Python 객체와 JSON String

- Python에서는 json모듈을 사용하여 JSON 데이터를 다룰 수 있다.
- json.dumps()를 사용하면 Python 객체를 JSON 문자열로 변환할 수 있으며, json.loads()를 사용하면 JSON 문자열을 Python 객체로 변환할 수 있다.

2. JSON String과 DataFrame

- pandas라이브러리의 json_normalize()함수 또는 pd.DataFrame.from_dict()메서드를 활용하면 JSON 데이터를 DataFrame으로 변환할 수 있다.
- API 응답을 DataFrame 형태로 변환하여 데이터 분석에 활용할 수 있다.

III. API 활용하기

1. API KEY 발급받기

- API를 사용하려면 대부분의 서비스에서 API 키(API Key)를 요구한다.
- API 키는 보안이 중요하며, 공개 저장소(GitHub 등)에 노출되지 않도록 환경 변수 또는 설정 파일을 활용해 관리하는 것이 좋다.

2. API 요청 구조 알아보기

- API 요청은 일반적으로 HTTP GET또는 HTTP POST방식으로 이루어진다.
- requests라이브러리를 사용하여 API 요청을 보낼 수 있으며, requests.get(url, params=params)형식으로 요청을 수행한다.
- 응답 데이터는 JSON 형식으로 반환되는 경우가 많으며, response.json()을 통해 파싱할 수 있다.

ex) <http://www.yes24.com/Product/Search?domain=B00K&query={isbn}>



IV. Web Scraping

1. Web Scraping이란?

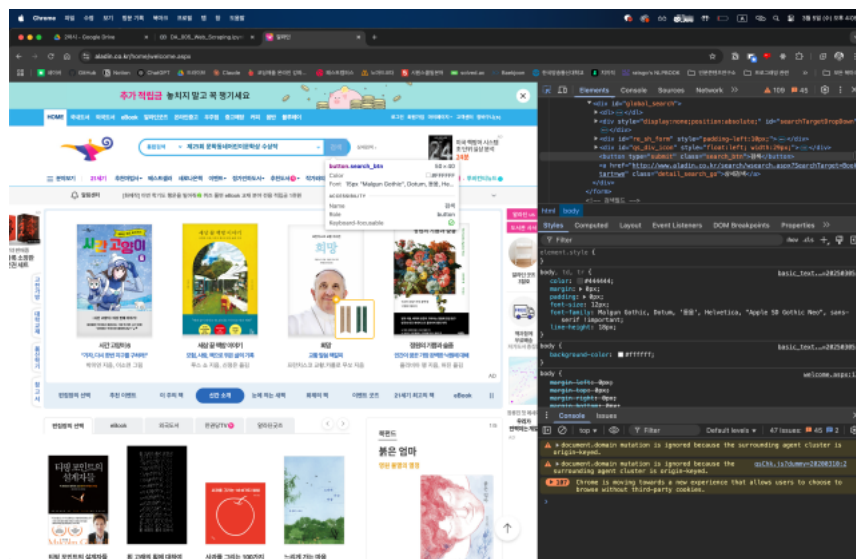
- 웹 스크래핑(Web Scraping)은 특정 웹사이트에서 데이터를 추출하는 기술을 의미한다.
- 주로 BeautifulSoup, Scrapy, Selenium같은 라이브러리를 사용하여 웹페이지의 HTML을 파싱하고 필요한 데이터를 추출한다.

2. Web Scraping의 기본 과정

- requests를 사용하여 웹페이지 HTML 가져오기
- BeautifulSoup을 활용하여 HTML 파싱 및 원하는 데이터 추출
- pandas를 사용하여 데이터 정리 및 저장

3. HTML 분석하기

- 개발자도구 사용하기
- > Chrome에서 F12 누르기



- 내가 원하는 정보가 어떤 태그로 둘러싸여있는지 찾아내보자!

V. 실습

실습 상황: 스마트팜 관련 서적 정보 크롤링 & 분석

a) 개요:

스마트팜을 운영하는 데 도움이 될 만한 책 목록을 만들고, 책의 페이지 수, 출판 연도, 출판사를 분석해 유용한 정보를 도출해보자.

b) 실습 내용:

1. 스마트팜 관련 도서 목록 수집

- YES24, 교보문고, 알라딘 등에서 "스마트팜", "농업 자동화", "IoT" 등의 키워드로 검색
- 책 제목, 저자, 출판사, 출판 연도, 페이지 수 크롤링

2. 도서 데이터 정리 및 시각화

- requests를 활용해 데이터를 수집
- pandas를 활용해 데이터를 정리
- csv 파일로 내보내기
- 출판 연도별 스마트팜 관련 도서 트렌드 분석
- 페이지 수 분포 분석 (어떤 책이 더 깊이 있는 내용을 다루는지 파악 가능)
- 출판사별 도서 수 비교

기대 결과:

최신 스마트팜 관련 도서를 한눈에 볼 수 있는 데이터셋 구축
출판 트렌드를 분석해 스마트팜 학습을 위한 최적의 도서 선택 가능