

웹크롤링 & 스크래핑

김혜경

웹크롤링 & 스크래핑 필요 기술셋

- html
- css
- java script
- 정규표현식
- 데이터 수집 기술 – python, non-python
- python
 - 데이터 수집 library 다양
 - 동적 자바 스크립트 기반의 데이터 수집
 - 브라우저 드라이버등도 필요
 - 별도의 설치

-
- I 크롤링 & 스크래핑 개요
 - II 데이터 저장 형식
 - III Python & DataBase
 - IV 크롤링 고려 사항 및 분류
 - V 크롤링 주요 library

| 크롤링 & 스크래핑 개요

크롤링과 스크래핑 비교

크롤링

웹 페이지의 하이퍼링크를 순회화면서
웹 페이지를 다운로드 하는 작업

웹 사이트를 정기적으로 돌며 정보를 추출하는 기술 의미

(정해진 시간대에 또는 정해진 이벤트 발생시 자동 실행...기술이 적용)

크론 & 스케줄러 : 정해진 시간에 업무 자동화 처럼 작업을 수행

스크래핑

다운로드한 웹 페이지에서 필요한
특정 정보를 추출하는 기술

7월 2일

- 정규표현식 리뷰
- oracle db에 정제시킨 데이터 CRUD

스크래핑 필요성

- 데이터 가져오기
 - 예시 : 소셜 데이터 -> 누구와 연관이 있는지
- 외부로 내보내는 기능이 없는 시스템에서 데이터 가져오기
- 사이트를 모니터링하며 새로운 정보 감지
- 검색 엔진의 데이터베이스를 구축하기 위한 스크래핑
- 고려사항
 - 웹 페이지의 내용을 마음대로 발췌한다는 부분에서 논란의 여지가 있음
 - copyright된 정보로 출판하는 것은 허용되지 않음
 - 이용 약관을 이용하지 않도록 유의

데이터 수집과 머신러닝

스크래핑(Scraping)

크롤링(Crawling)

머신러닝에서 사용 가능한 데이터 구조

- 데이터의 구조를 분석, 정제 후 추출등의 과정으로 CSV, JSON, XML 등의 형식으로 가공 후 사용

파이썬으로 크롤링 & 스크래핑

- 파이썬 기반의 작업시 장점

- 언어 자체의 특성
- 강력한 라이브러리

- Python Package Index(PyPI)에 수많은 개발자들이 개발한 라이브러리 공개
- 유명한 스크래핑 라이브러리 제공
 - BeautifulSoup
 - lxml
 - Selenium

- 스크래핑 후처리의 편리성

- 크롤링/ 스크래핑으로 데이터를 추출한 후 데이터 분석 등의 처리시 파이썬은 강력한 무기로 작용 함

```
>pip list
```

아나콘다 기반에선 존재 단,
Selenium 차후에 설치 예정

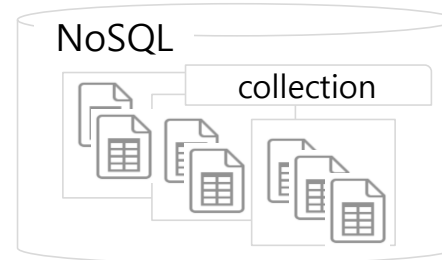
쉬어가기 - 상식

표기	설명	예시
r'...' or r"..."	정규 표현식으로 스크래핑 할 경우, 정규 표현식 패턴에는 백슬래시가 자주 나옴, raw 문자열 표기 형식 사용시 백슬래시가 이스케이프 문자로 사용되지 않음	re.search(r'charset=["\W']*?([\Ww-]+)', scanned_text)
rss	블로그 또는 뉴스 사이트 등의 웹 사이트는 변경 정보 등을 RSS라는 이름의 XML 형식으로 제공 XML기반으로 구성되었기 때문에 HTML보다 간결하게 파싱 가능	
CP949	EUC-KR에 확장 문자 추가한 문자 코드	

I 데이터 저장 형식

크롤링한 데이터 저장 형식

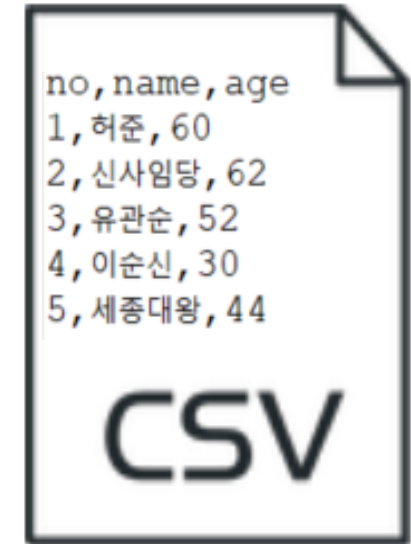
- 크롤링한 데이터를 파일 등에 저장해 두면 데이터를 쉽게 활용할 수 있음



CSV 형식으로 저장하기

- 데이터를 수집해서 가공하는 필수 작업

- Comma -Seperated Values
 - 하나의 recode를 한 줄로 표현
 - 각 줄의 데이터는 , 를 기준으로 구분하는 형식
- csv 형식을 만드는 쉬운 방법
 - str.join() 함수 사용
 - csv 모듈 사용



CSV 형식으로 저장하기

- str.join()

```
2
3  # 첫 번째 줄에 헤더를 작성
4  print('no,name,age')
5
6  # join() 메서드
7  print(', '.join(['1', '허준', '60']))
8  print(', '.join(['2', '신사임당', '62']))
9
```

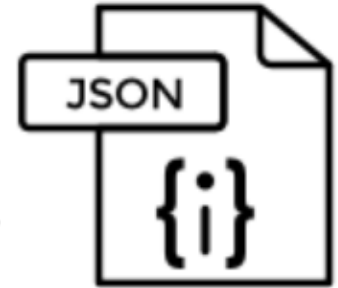
CSV 형식으로 저장하기

- csv 모듈 사용

```
15 import csv
16
17 # 파일을 열기
18 # newline=''으로 줄바꿈 코드의 자동 변환 제어
19 # 한글 문제 발생시 : encoding="utf-8-sig"
20 with open('history.csv', 'w', newline='', encoding="utf-8-sig") as f:
21
22     # csv.writer는 파일 객체를 매개변수로 지정
23     writer = csv.writer(f)
24
25     # 첫 번째 줄에는 헤더를 작성
26     writer.writerow(['no', 'name', 'age'])
27
28     # writerows()에 리스트를 전달하면 여러 개의 값 출력
29     writer.writerows([
30         [1, '허준', 60],
31         [2, '신사임당', 62],
32         [3, '유관순', 52],
33         [4, '이순신', 30],
34         [5, '세종대왕', 44],
35     ])
```

JSON 형식으로 저장하기

- JavaScript Object Notation
 - 자바스크립트에서 객체를 표현하는 방법을 사용하는 텍스트 형식
- 장점
 - list or dict를 조합한 복잡한 데이터 구조를 쉽게 다룰 수 있음
- 파이썬에서 JSON 형식으로 저장하기
 - json 모듈 제공
 - json.dump() 함수 사용시 list, dict 등의 객체를 JSON 형식 문자열로 변환 가능
 - 한글 데이터 및 들여쓰기 고려 하기



```
json.dumps(dataset, ensure_ascii=False, indent=2)
```

JSON 형식으로 저장하기

```
1 import json
2
3 friends = [
4     {'f1':1, 'name' : '허준'},
5     {'f2':2, 'name' : '이이'},
6     {'f3':3, 'name' : '신사임당'}
7 ]
8
9 print("1. dict로 구성된 list 출력 ---")
10 print(friends)
11
12 print("2. list의 첫번째 요소 출력 ---")
13 print(friends[0])
14
15 print("3. list를 JSON 객체로 변환해 보기 ---")
16 jsonData = json.dumps(friends)
17 print(jsonData)
18
19 print("4. 유니코드 표기로 출력된 데이터를 한글로 변환하기 ---")
20 jsonData = json.dumps(friends, ensure_ascii=False)
21 print(jsonData)
22
23 print("5. 들여쓰기 반영하기 ---")
24 jsonData = json.dumps(friends, ensure_ascii=False, indent=2)
25 print(jsonData)
```

```
[{"f1": 1, "name": "\ud5c8\uc900"}, {"f2": 2, "name": "\uc774\uc774"}, {"f3": 3, "name": "\uc2e0\uc0ac\uc784\uc2f9"}]
4. 유니코드 표기로 출력된 데이터를 한글로 변환하기 ---

(base) C:\0.ITStudy\9.ctest\crawlingLab\01.basic>python step01JSONsave.py
1. dict로 구성된 list 출력 ---
[{'f1': 1, 'name': '허준'}, {'f2': 2, 'name': '이이'}, {'f3': 3, 'name': '신사임당'}]
2. list의 첫번째 요소 출력 ---
{'f1': 1, 'name': '허준'}
3. list를 JSON 객체로 변환해 보기 ---
[{"f1": 1, "name": "\ud5c8\uc900"}, {"f2": 2, "name": "\uc774\uc774"}, {"f3": 3, "name": "\uc2e0\uc0ac\uc784\uc2f9"}]
4. 유니코드 표기로 출력된 데이터를 한글로 변환하기 ---
[{"f1": 1, "name": "허준"}, {"f2": 2, "name": "이이"}, {"f3": 3, "name": "신사임당"}]

(base) C:\0.ITStudy\9.ctest\crawlingLab\01.basic>python step01JSONsave.py
1. dict로 구성된 list 출력 ---
[{'f1': 1, 'name': '허준'}, {'f2': 2, 'name': '이이'}, {'f3': 3, 'name': '신사임당'}]
2. list의 첫번째 요소 출력 ---
{'f1': 1, 'name': '허준'}
3. list를 JSON 객체로 변환해 보기 ---
[{"f1": 1, "name": "\ud5c8\uc900"}, {"f2": 2, "name": "\uc774\uc774"}, {"f3": 3, "name": "\uc2e0\uc0ac\uc784\uc2f9"}]
4. 유니코드 표기로 출력된 데이터를 한글로 변환하기 ---
[{"f1": 1, "name": "허준"}, {"f2": 2, "name": "이이"}, {"f3": 3, "name": "신사임당"}]
5. 들여쓰기 반영하기 ---
[
  {
    "f1": 1,
    "name": "허준"
  },
  {
    "f2": 2,
    "name": "이이"
  },
  {
    "f3": 3,
    "name": "신사임당"
  }
]
```


JSON 구조로 파일로 생성하기

```
28  """
29  1. 파일로 출력시 한글 인코딩 문제 고려해야 함
30  2. 추가 속성
31      encoding="UTF-8-sig
32      fp=f
33      ensure_ascii=False
34  """
35  print("6. JSON 형식의 문자열을 출력하지 않고 파일에 저장할 경우 json.dump() 함수 사용 ---")
36  with open("friends.json", "w", encoding="UTF-8-sig") as f:
37      json.dump(friends, fp=f, ensure_ascii=False)
38      #json.dump(friends, f) 파일에 유니코드 형식으로 출력됨
```

Python & DataBase

Oracle DB 연동을 위한 작업

- 1단계 : oracle db 접속 모듈인 "cx_Oracle" 존재 여부 확인하기

```
pip list
```

- 2단계 : 설치하기

```
conda install -c http://conda.anaconda.org/anaconda cx_oracle
```

- 3단계 : oracle db 접속 모듈 설치 확인하기

```
pip list
```

```
conda-build 3.17.8
conda-package-handling 1.3.10
conda-verify 3.1.1
contextlib2 0.5.5
cryptography 2.6.1
cx-Oracle 7.0.0
cycler 0.10.0
Cython 0.29.6
```

Oracle DB 연동을 위한 개발 단계

접속 : connect()

커서 추출 : cursor()

SQL 구문 실행 : execute() or executemany()

변경 사항 저장 : commit()

데이터 활용 : fetchone() & fetchall()

연결 해제 : close()

Oracle DB 연동을 위한 기초

```
6 import cx_Oracle
7 import pandas as pd
8
9 # 한글 문제 해결을 위한 접속전 설정 정보
10 import os
11 os.putenv('NLS_LANG', 'KOREAN_KOREA.KO16MSWIN949')
12
13 # 데이터베이스 연결
14 con = cx_Oracle.connect("SCOTT/TIGER@localhost:1521/xe")
15
16 # 커서 추출
17 cur = con.cursor()
18
19 # sql 문장 실행
20 cur.execute("select * from emp")
21
22 # 검색된 데이터 row 단위로 출력
23 for row in cur:
24     print(row)
25
26 # 커서 및 접속 해제
27 cur.close()
28 con.close()
```

Kyung

```
(7369, 'SMITH', 'CLERK', 7902, datetime.datetime(1980, 12, 17, 0, 0), 800.0, None, 20)
(7499, 'ALLEN', 'SALESMAN', 7698, datetime.datetime(1981, 2, 20, 0, 0), 1600.0, 300.0, 30)
(7521, 'WARD', 'SALESMAN', 7698, datetime.datetime(1981, 2, 22, 0, 0), 1250.0, 500.0, 30)
(7566, 'JONES', 'MANAGER', 7839, datetime.datetime(1981, 4, 2, 0, 0), 2975.0, None, 20)
(7654, 'MARTIN', 'SALESMAN', 7698, datetime.datetime(1981, 9, 28, 0, 0), 1250.0, 1400.0, 30)
(7698, 'BLAKE', 'MANAGER', 7839, datetime.datetime(1981, 5, 1, 0, 0), 2850.0, None, 30)
(7782, 'CLARK', 'MANAGER', 7839, datetime.datetime(1981, 6, 9, 0, 0), 2450.0, None, 10)
(7839, 'KING', 'PRESIDENT', None, datetime.datetime(1981, 11, 17, 0, 0), 5000.0, None, 10)
(7844, 'TURNER', 'SALESMAN', 7698, datetime.datetime(1981, 9, 8, 0, 0), 1500.0, 0.0, 30)
(7900, 'JAMES', 'CLERK', 7698, datetime.datetime(1981, 12, 3, 0, 0), 950.0, None, 30)
(7902, 'FORD', 'ANALYST', 7566, datetime.datetime(1981, 12, 3, 0, 0), 3000.0, None, 20)
(7934, 'MILLER', 'CLERK', 7782, datetime.datetime(1982, 1, 23, 0, 0), 1300.0, None, 10)
```

Oracle DB 연동을 위한 기초

- RDBMS에 저장된 데이터들을 DataFrame 으로 생성 가능
 - 데이터 전처리 쉽게 수행 가능
 - ML/DL등으로 서비스 로직을 client에게 최적화 해서 개발해서 제공

```
# 참고 : sql 실행 및 데이터 검색해서 DataFrame으로 생성
emp = pd.read_sql("select * from emp", con=con)
print(type(emp))
print(emp.head())
```

--- DataFrame 단위로 검색 후 바로 변환 ---

<class 'pandas.core.frame.DataFrame'>

	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
0	7369	SMITH	CLERK	7902.0	1980-12-17	800.0	NaN	20
1	7499	ALLEN	SALESMAN	7698.0	1981-02-20	1600.0	300.0	30
2	7521	WARD	SALESMAN	7698.0	1981-02-22	1250.0	500.0	30
3	7566	JONES	MANAGER	7839.0	1981-04-02	2975.0	NaN	20
4	7654	MARTIN	SALESMAN	7698.0	1981-09-28	1250.0	1400.0	30

SQLite 개요

- 가볍게 파일 하나로 사용할 수 있는 데이터베이스
 - 파일 하나가 하나의 데이터베이스
 - 굉장히 가볍다는 것이 특징
- **다양한 곳에서 사용되는 데이터베이스**
 - 웹 브라우저 내부, 안드로이드/iOS에서 표준으로 제공되는 데이터베이스
- 별도의 데이터베이스 전용 애플리케이션을 사용하지 않아도 됨
- sqlite3 모듈은 파이썬 표준 라이브러리로 SQLite에 대한 인터페이스를 제공
 - 별도의 설치없이 데이터베이스를 쉽게 이용할 수 있음



- <https://sqlite.org/index.html>
- <https://sqlite.org/cli.html> - Command Line Shell For SQLite

SQLite 사용

데이터베이스 연결 : `connect()`

테이블 생성 : **`cursor()`** / `execute("sql")`

CRUD : **`cursor()`** / `execute("sql")` / `commit()` : insert/update/delete

검색 : **`cursor()`** / `execute("query")`

검색된 데이터 활용하기 : **`cursor()`** / `fetchall()` : for or `fetchone()`

<https://docs.python.org/ko/3.6/library/sqlite3.html>

SQLite 실습

```
5  import sqlite3
6
7  # 데이터베이스 연결
8  filepath = "encore.sqlite"
9  conn = sqlite3.connect(filepath)
10
11 # 테이블 생성
12 cur = conn.cursor()
13 cur.execute("DROP TABLE IF EXISTS books")
14 cur.executescript("""CREATE TABLE books (
15     key text primary key,
16     title text,
17     content text)""")
18
19 # 모든 데이터 저장
20 cur.execute('insert into books values("002", "data3", "data3 이해하기")')
21 conn.commit()
22
23 # 모든 데이터 검색 - fetchall()
24 cur.execute("select * from books")
25 for data in cur.fetchall():
26     print(data)
27
28 # 특정 데이터 하나만 검색 - fetchone()
29 cur.execute("select * from books where key='002'")
30 data = cur.fetchone()
31 print(data)
```

동적 데이터

value = ("003", "data3", "data3 이해하기")

cur.execute('insert into books values(?, ?, ?)', value)

cur.execute('insert into books values(?, ?, ?)' % value)

mysql인 경우

value = (" 003 " , " data3 " , " data3 이해하기 ")

cur.execute(' insert into books values(%s, %s, %s)', value)

크롤링 고려 사항

크롤러 구성시 주의 사항

- 저작권 고려
- 너무 많은 부하(엄청나게 지속적으로 계속 수집하는 작업)는 업무 방해등으로 고소 당할 수 있음
 - 수집하는 user들의 접속 정보를 로그기록화
 - 1초단위? 3초단위?...접속인 경우 등 비정상적으로 간주해서 차단
- 종량제 사용하는 서버인 경우 부하가 많이 걸리면 과금 발생
- 웹 사이트들의 **robots.txt** 확인 하기
 - <https://www.google.com/robots.txt>

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/static
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&gws_rd=ssl$
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
Disallow: /u/
Disallow: /preferences
Disallow: /setprefs
Disallow: /default
Disallow: /m?
Disallow: /m/
```

크롤러 구성시 주의 사항

- robots.txt 파일 parsing으로 크롤링 가능 여부 확인하기
 - 연관 예제 : step01robotsParser.py

```
24 # 파이썬 표준 library
25 import urllib.robotparser
26
27 # robots.txt 파일을 위한 RobotFileParser 클래스
28 rp = urllib.robotparser.RobotFileParser()
29
30 # 해당 사이트 크롤링 해도 되는지 확인
31
32 # http://wikibook.co.kr/ 사이트 크롤링 가능 여부 확인
33 rp.set_url("http://wikibook.co.kr/robots.txt")
34
35 # robots.txt 파일 read
36 rp.read()
37
38 # wikibook.co.kr 사이트 크롤링 여부 확인
39 data = rp.can_fetch("mybot", "http://wikibook.co.kr/")
40 print(data) # True
```

크롤러와 URL

- 크롤러
 - 웹 페이지에 존재하는 하이퍼링크를 따라 돌아야 함
- 크롤러와 URL
 - 브라우저에서는 링크를 클릭하면 되지만 크롤러로 링크를 돌아다니면 URL과 관련된 기초적인 지식을 이해하고 활용 해야 함
- URL 구조

<http://www.google.com:80/main/index?q=python#lead>

schema

authority

path

web query
string

flagment

URL을 구성하는 각 부분의 의미

URL 구성 요소	설명
schema	http or https와 같은 프로토콜 의미
authority	//뒤에 나오는 일반적인 호스트 이름 사용자 이름, 비밀번호, 포트 번호 등을 포함
path	/로 시작하는 해당 호스트 내부에서의 리소스 경로
query	? 뒤에 나오는 경로와는 다른 방법으로 리소스를 표현하는 방법 Key=value 구조의 데이터 전송 미 존재하는 경우도 있음
fragment	# 뒤에 나오는 리소스 내부의 특정 부분 등을 표현 미 존재하는 경우도 있음

크롤링시 고려사항

- url 구분

절대 url

http:// 로 시작

상대 url

// 로 시작
/ 로 시작
이 밖의 상대 경로 형식을 사용하는
url

상대 URL을 절대 URL로 변환하기

- 변환 API
 - urllib.parse 모듈에 포함되어 있는 urljoin() 함수 사용
 - urljoin(절대 url, 상대 url)

Kim Hye Kyung



제품 관리 분류

크롤러 분류하기

- 크롤러는 대상 웹사이트에 따라 다양한 성질 보유 가능
- 크롤러 분류를 알면 웹사이트에 맞게 크롤러를 쉽게 설계 할 수 있음
- 크롤러 분류

상태를 가지고 있는지

자바스크립트를 실행 할 수 있는지

크롤러 분류하기

- 상태를 가지고 있는지

Statefull 크롤러

HTTP는 stateless로 설계된 프로토콜

따라서 모든 요청이 연계된 요청이
아닌 독립적인 요청으로 간주

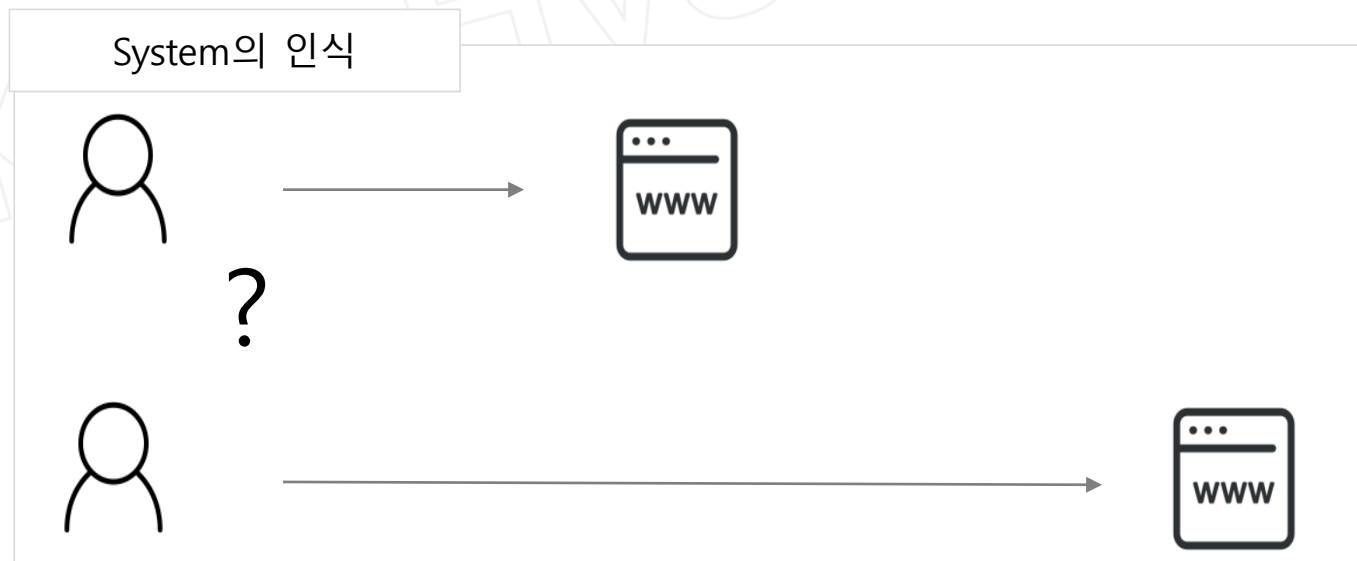
Stateless 크롤러

Client의 상태 유지 기술인
Session/Cookie 기술을 활용하여
개발자들이 stateless 기술 적용

로그인이 필요한 웹사이트 크롤링을
위해 필요한 크롤러
Cookie 지원해야 함

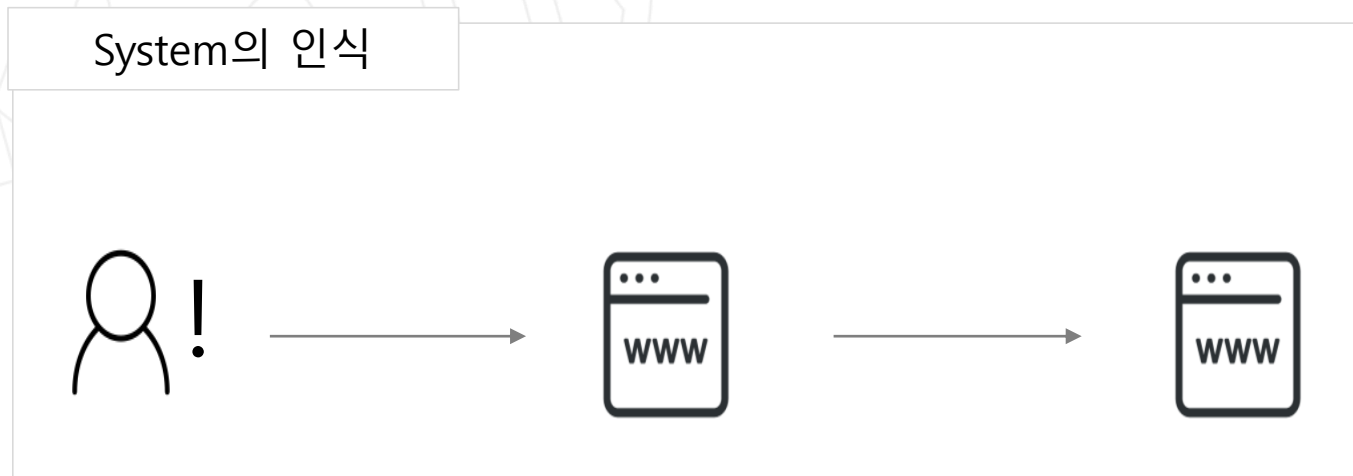
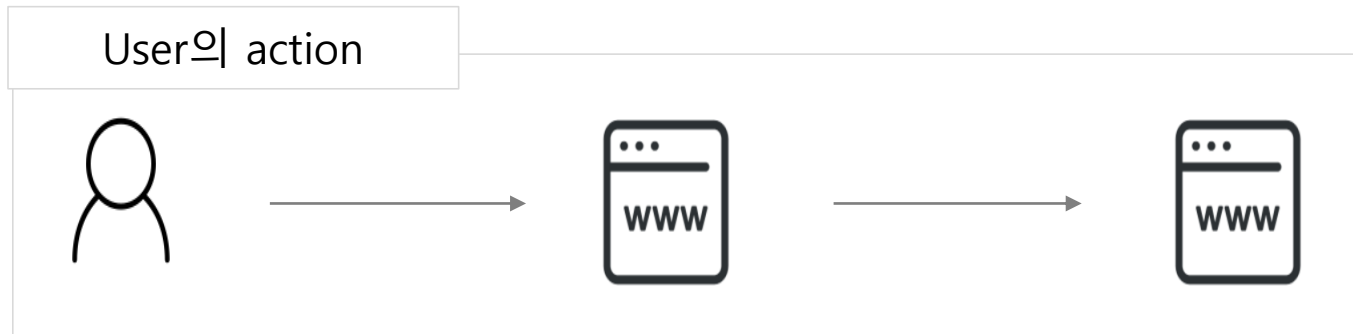
크롤러 분류하기

- 상태를 가지고 있는지
 - Stateless 동작 원리



크롤러 분류하기

- 상태를 가지고 있는지
 - Statefull 동작 원리



크롤러 분류하기

- 자바스크립트를 실행 할 수 있는지



자바스크립트를 실행 할 수 없는 크롤러

자바스크립트 실행 할 수 있는 크롤러



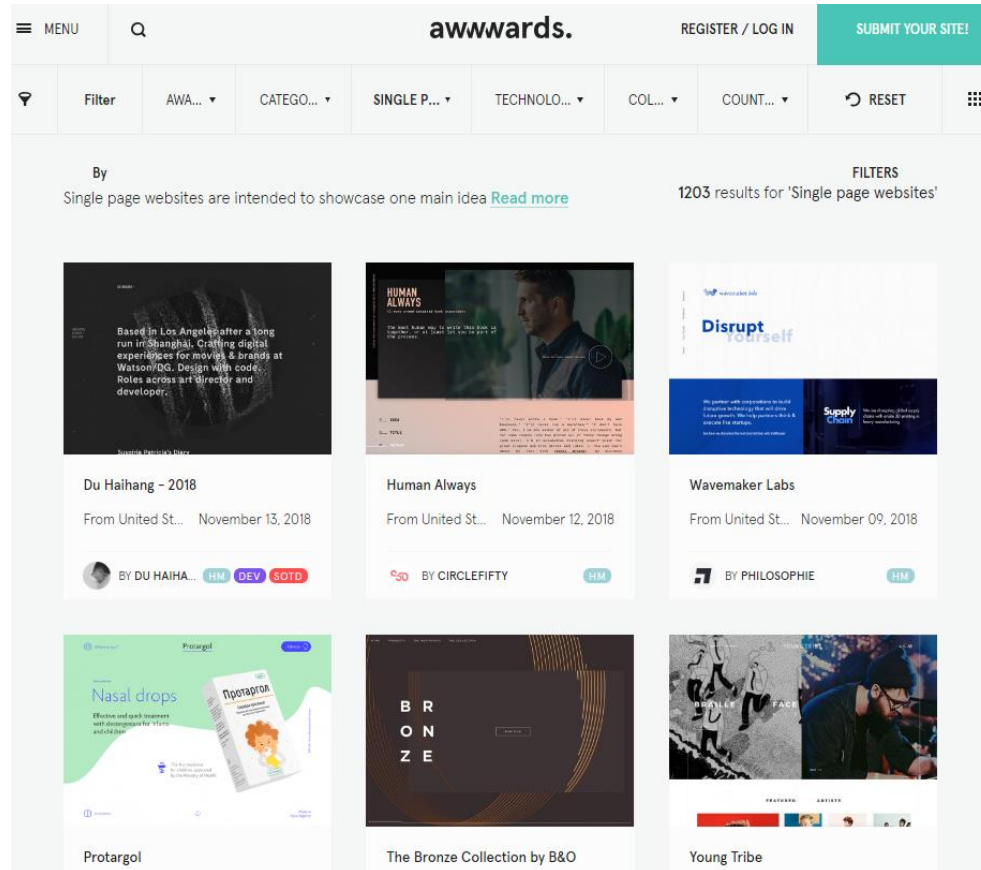
크롤러 분류하기

자바스크립트 실행 할 수 있는 크롤러

- Single page Application란?
 - 사용자가 처음 웹사이트 접근시 HTML과 Java Script등의 필요한 리소스 로드
 - 이후 모두 자바스크립트로 운용해서 페이지 전환이 발생되지 않게 만드는 애플리케이션 의미

크롤러 분류하기

- Single page Application 예시



스크롤을 내릴 때마다
자바스크립트를 활용해서
서버로 부터 데이터
받아와서 화면에 출력

<https://www.awwwards.com/websites/single-page/>

| 크롤링 주요 library

크롤링을 위한 주요 library

- requests
- lxml
- BeautifulSoup
- Selenium
 - 다양한 브라우저를 자동 조작하는 도구
 - 자바스크립트를 해석할 수 있는 library
 - 참고
 - 초기 개발 목적
 - 웹 애플리케이션 자동 테스트 도구로 개발
 - 최근 자바스크립트를 활용한 웹 페이지를 스크래핑할 때 사용

주요 라이브러리 사전 준비

- Python Package Index(PyPI)에서 제공
 - python package 저장소
- library 다운로드 방법

```
pip install library이름
```

```
pip install library이름=버전
```

- 설치된 library 확인

```
pip list
```

주요 library - requests

- requests
 - pip install requests 명령어로 설치
 - 아나콘다에는 이미 설치되어 있음
 - 장점
 - 표준 library인 urllib를 사용 하는 것에 비해 쉽게 웹페이지 내용 추출이 가능
 - 사용하기 굉장히 쉬운 라이브러리
 - HTTP 헤더 추가 또는 Basic 인증, 문자 코드 변환, 압축등 간결한 코드만으로 사용 가능
 - 주요 API
 - Session
 - 여러 개의 페이지를 연속으로 크롤링할 때 효과적으로 사용 가능
 - HTTP 헤더 또는 Basic 인증 등의 설정을 한 번만 하고 여러 번 재사용 가능
 - HTTP Keep-Alive라는 접속 방식 사용(한번 확립한 TCP 요청을 계속 활용하므로 오버헤드가 되는 TCP 커넥션 확립 처리를 줄일 수 있어서 성능 향상 효과)

주요 library – BeautifulSoup

- 단순한 API가 특징인 스크래핑 라이브러리
- HTML or XML을 분석해주는 라이브러리
 - 용어
 - parser – parsing 작업을 지원하는 주체
 - parsing - 문서 분석, 검증, 변환..등의 모든 작업을 parsing
- URL
 - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/#>

```
pip install beautifulsoup4
```
- 설치 명령어
 - anaconda에는 이미 설치되어 있음

```
pip show beautifulsoup4
```



주요 library – BeautifulSoup

- Beautiful Soup에서 사용 가능한 parser

파서	매개변수에 지정하는 문자열	특징
표준 라이브러리의 html.parse	'html.parser'	기본적으로 보유 되어 있음
lxml의 HTML parser	'lxml'	빠른 처리가 가능
lxml의 XML parser	'lxml-xml' 또는 'xml'	Xml에 대해 빠른 처리가 가능
Html5lib	html5lib	Html5의 사양에 맞게 파싱 가능

주요 library – Selenium

- 자바스크립트를 활용한 웹 페이지를 스크래핑 할 수 있음
- 프로그램에서 브라우저를 자동으로 조작할 수 있게 해주는 도구
- 데이터 출력을 자바스크립트로 처리하는 웹사이트인 경우에 사용
- 단점
 - 단순 html만 해석하는 크롤러에 비해 실행 속도가 느림
 - 외부 자바스크립트, css, 이미지등 모두 read한 후에 자바스크립트 코드를 실행
 - 메모리도 많이 소비

- <https://selenium.dev/>

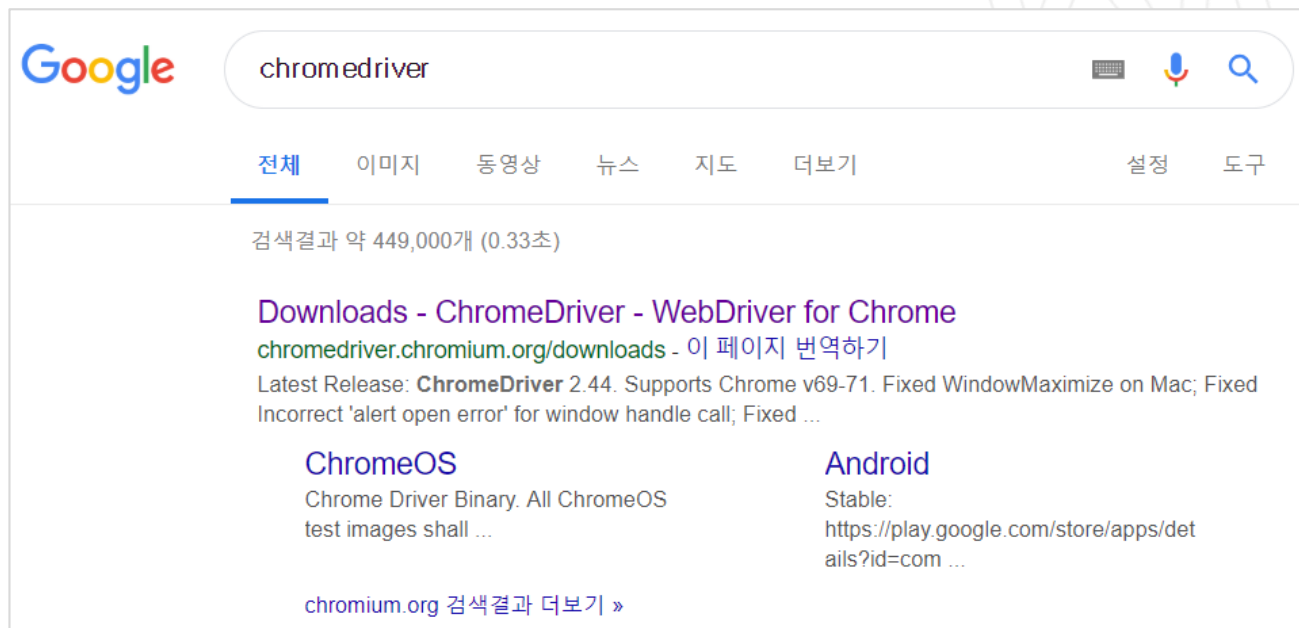
```
pip install selenium
```

Selenium WebDriver



주요 library – Selenium

- Chrome driver 다운로드
- 주의사항
 - 시스템에 설치되어 있는 chrome 브라우저 보다 하위 버전 사용 불가



주요 library – Selenium

- Selenium으로 DOM 요소를 선택하는 방법
 - DOM 내부에 있는 여러 개의 요소 중 처음 찾아지는 요소를 추출

메서드 이름	설명
find_element_by_id(id)	id 속성으로 요소를 하나 추출합니다.
find_element_by_name(name)	name 속성으로 요소를 하나 추출합니다.
find_element_by_css_selector(query)	CSS 선택자로 요소를 하나 추출합니다.
find_element_by_xpath(query)	XPath를 지정해 요소를 하나 추출합니다.
find_element_by_tag_name(name)	태그 이름이 name에 해당하는 요소를 하나 추출합니다.
find_element_by_link_text(text)	링크 텍스트로 요소를 추출합니다.
find_element_by_partial_link_text(text)	링크의 자식 요소에 포함돼 있는 텍스트로 요소를 하나 추출합니다.
find_element_by_class_name(name)	클래스 이름이 name에 해당하는 요소를 하나 추출합니다.

주요 library – Selenium

- Selenium으로 DOM 요소를 선택하는 방법
 - DOM 내부에 있는 모든 요소 추출

메서드 이름	설명
<code>find_elements_by_css_selector(query)</code>	CSS 선택자로 요소를 여러 개 추출합니다.
<code>find_elements_by_xpath(query)</code>	XPath를 지정해 요소를 여러 개 추출합니다.
<code>find_elements_by_tag_name(name)</code>	태그 이름이 <code>name</code> 에 해당하는 요소를 여러 개 추출합니다.
<code>find_elements_by_class_name(name)</code>	클래스 이름이 <code>name</code> 에 해당하는 요소를 여러 개 추출합니다.
<code>find_elements_by_partial_link_text(text)</code>	링크의 자식 요소에 포함돼 있는 텍스트로 요소를 여러 개 추출합니다.

주요 library – Selenium

- Selenium으로 요소 조작하기
 - DOM 요소에 적용할 수 있는 메서드와 속성

메서드 또는 속성	설명
clear()	글자를 입력할 수 있는 요소의 글자를 지웁니다.
click()	요소를 클릭합니다.
get_attribute(name)	요소의 속성 중 name에 해당하는 속성의 값을 추출합니다.
is_displayed()	요소가 화면에 출력되는지 확인합니다.
is_enabled()	요소가 활성화돼 있는지 확인합니다.
is_selected()	체크박스 등의 요소가 선택된 상태인지 확인합니다.
screenshot(filename)	스크린샷을 찍습니다.
send_keys(value)	키를 입력합니다.
submit()	입력 양식을 전송합니다.

주요 library – Selenium

- Selenium으로 요소 조작하기
 - DOM 요소에 적용할 수 있는 메서드와 속성

메서드 또는 속성	설명
value_of_css_property(name)	name에 해당하는 CSS 속성의 값을 추출합니다.
id	요소의 id 속성입니다.
location	요소의 위치입니다.
parent	부모 요소입니다.

주요 library – Selenium

- DOM 요소에 적용할 수 있는 메서드와 속성

메서드 또는 속성	설명
rect	크기와 위치 정보를 가진 딕셔너리 자료형을 리턴합니다.
screenshot_as_base64	스크린샷을 Base64로 추출합니다.
screenshot_as_png	스크린샷을 PNG 형식의 바이너리로 추출합니다.
size	요소의 크기입니다.
tag_name	태그 이름입니다.
text	요소 내부의 글자입니다.

크롤링 & 스크래핑 실전

파이썬으로 스크래핑 흐름 이해하기

1. 한빛출판네트워크

주소: hanbit.co.kr/store/books/full_book_list.html

한빛출판네트워크

카테고리: 새로나온 책, 베스트셀러, 전체도서목록, 교육, 기타

전체도서목록

전체 목록 다운로드

브랜드	도서명	저자
한빛아카데미	NO BULLSHIT 수학&물리 가이드	Ivan Savov
한빛아카데미	IT CookBook, UI/UX 디자인 이론과 실습 with Adobe XD	이영주
한빛아카데미	IT CookBook, 파이썬 for Beginner(2판)	우재남
한빛미디어	쿠버네티스를 활용한 클라우드 네이티브 데브옵스	존 어런들 외 1
한빛라이프	리얼 상하이 황저우-쑤저우 [2020~2021년 개정판]	도선미
한빛라이프	리얼 광 [2020~2021년 최신판]	민정아
한빛미디어	약속달수 C 언어 180제	시바타 부유 오

3. NO BULLSHIT 수학&물리 가이드

저자: Ivan Savov

IT CookBook, UI/UX 디자인 이론과 실습 with Adobe XD

이영주

IT CookBook, 파이썬 for Beginner(2판)

우재남

주소: hanbit.co.kr/store/books/look.php?p_code=B5203031354

2. NO BULLSHIT 수학&물리 가이드

저자: Ivan Savov

IT CookBook, UI/UX 디자인 이론과 실습 with Adobe XD

이영주

IT CookBook, 파이썬 for Beginner(2판)

우재남

쿠버네티스를 활용한 클라우드 네이티브 데브옵스

존 어런들 외 1

리얼 상하이 황저우-쑤저우 [2020~2021년 개정판]

도선미

리얼 광 [2020~2021년 최신판]

민정아

주소: hanbit.co.kr/store/books/look.php?p_code=B5203031354

4. 최종 결론 : 상세 page의 url 발체

주소: hanbit.co.kr/store/books/look.php?p_code=B5203031354

NO BULLSHIT 수학&물리 가이드

한빛아카데미

저자: Ivan Savov

번역: 권기영

출간: 2020-01-06

페이지: 568 쪽

ISBN: 9791156644644

파이썬으로 스크래핑 흐름 이해하기

- RDBMS에 저장된 데이터 검색

```
SQL> select * from books;
TITLE
-----
URL
-----
NO BULLSHIT 수학&물리 가이드
http://www.hanbit.co.kr/store/books/look.php?p_code=B5203031354
IT CookBook, UI/UX 디자인 이론과 실습 with Adobe XD
http://www.hanbit.co.kr/store/books/look.php?p_code=B5034837432
IT CookBook, 파이썬 for Beginner&#40;2판&#41;
http://www.hanbit.co.kr/store/books/look.php?p_code=B3780991491
쿠버네티스를 활용한 클라우드 네이티브 데브옵스
http://www.hanbit.co.kr/store/books/look.php?p_code=B4886455651
리얼 상하이 항저우 · 쑤저우 [2020~2021년 개정판]
http://www.hanbit.co.kr/store/books/look.php?p_code=B4507382865
리얼 광 [2020~2021년 최신판]
http://www.hanbit.co.kr/store/books/look.php?p_code=B5732120765
알쏭달쏭 C 언어 180제
http://www.hanbit.co.kr/store/books/look.php?p_code=B1265473016
```


퍼머링크와 링크 구조 패턴

- 퍼머링크란?
 - 웹사이트는 하나의 콘텐츠가 하나의 URL에 대응
 - 이처럼 하나의 콘텐츠에 대응되는 URL이 시간이 흘러도 대응되는 콘텐츠가 변하지 않는 URL 의미
 - 불변(permanent) + 링크(link) 조합
- 퍼머링크를 사용하는 웹사이트인 경우
 - 퍼머링크를 가진 페이지로 연결되는 링크가 목록으로 존재하는 페이지가 있음
 - 예시 - http://www.hanbit.co.kr/store/store_submain.html

퍼머링크를 사용하는 웹사이트인 경우

한빛출판네트워크

목록 페이지

[HOME](#)
[한빛미디어](#)
[한빛아카데미](#)
[한빛비즈](#)
[한빛라이프](#)
[한빛에듀](#)
[리얼타임](#)
[한빛정보교과서](#)
[한빛대강서비스](#)

로그인 | 회원가입 | 마이한빛 | 장바구니

한빛출판네트워크

[BRAND](#)
[Channel.H](#)
[STORE](#)
[SUPPORT](#)
[EVENT](#)

카테고리

[새로운 책](#)
[베스트셀러](#)
[전체도서목록](#)
[교육](#)
[Item & Maker Shed](#)

새로운 책

재미있고 빠른 한글 1권 : 기본 모음과 자음

재미있고 빠른 한글 2권 : 기본 자음과 쌍자음

재미있고 빠른 한글 3권 : 받침

재미있고 빠른 한글 4권 : 복잡한 모음

이도한글학습연구회 김두섭(대표 저자)

이도한글학습연구회 김두섭(대표 저자)

이도한글학습연구회 김두섭(대표 저자)

이도한글학습연구회 김두섭(대표 저자)

베스트셀러

click시 상세 page로 이동

퇴근길 인문학 수업

퇴근길 인문학 수업

지적 대화를 위한 넓고 얇은 지식

지적 대화를 위한 넓고 얇은 지식

이도한글학습연구회 김두섭(대표 저자)

이도한글학습연구회 김두섭(대표 저자)

이도한글학습연구회 김두섭(대표 저자)

이도한글학습연구회 김두섭(대표 저자)

상세 페이지

한빛출판네트워크

[BRAND](#)
[Channel.H](#)
[STORE](#)
[SUPPORT](#)
[EVENT](#)

카테고리

[새로운 책](#)
[베스트셀러](#)
[전체도서목록](#)
[교육](#)
[Item & Maker Shed](#)

재미있고 빠른 한글 1권 : 기본 모음과 자음

이도한글학습연구회 김두섭(대표 저자)

출간: 2018-12-01

페이지: 100 쪽

ISBN: 9791162241110

출판코드: 10111

가격: 7,000원

판매가: 6,300원 (10% off)

마일리지: 359원 (5%)

장바구니

위시리스트

구매하기

파이썬으로 크롤러 만들기 실습[step03]

- 크롤링 -> 제목과 상세 url 스크래핑 -> DB에 저장

1. 크롤링 사이트에 접속

2. 정규화를 통한 도서명과 상세 url 스크래핑

브랜드	도서명	이도한글 김두섭(대표 저자)
한빛에듀	재미있고 빠른 한글 1권 : 기본 모음과 자음	이도한글 김두섭(대표 저자)
한빛에듀	재미있고 빠른 한글 2권 : 기본 자음과 쌍자음	이도한글 김두섭(대표 저자)
한빛에듀	재미있고 빠른 한글 3권 : 받침	이도한글 김두섭(대표 저자)

재미있고 빠른 한글 1권 : 기본 모음과 자음

재미있고 빠른 한글 1권 : 기본 모음과 자음

한빛에듀 김필서 판매중

저자 : 이도한글학습연구회 김두섭(대표 저자)
출간 : 2018-12-01
페이지 : 100 쪽
ISBN : 9791162241110

파이썬으로 크롤러 만들기

- requests – 웹 페이지 추출
- lxml – 웹 페이지 스크래핑
- process

목록 페이지에서 퍼머링크 목록 추출하기

상세 페이지에서 스크래핑 하기

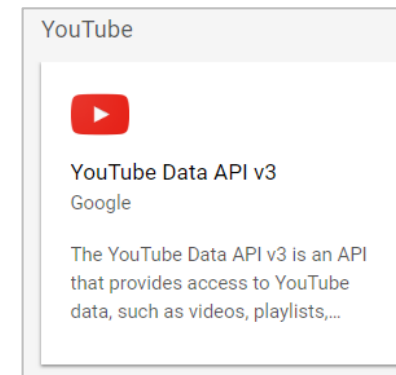
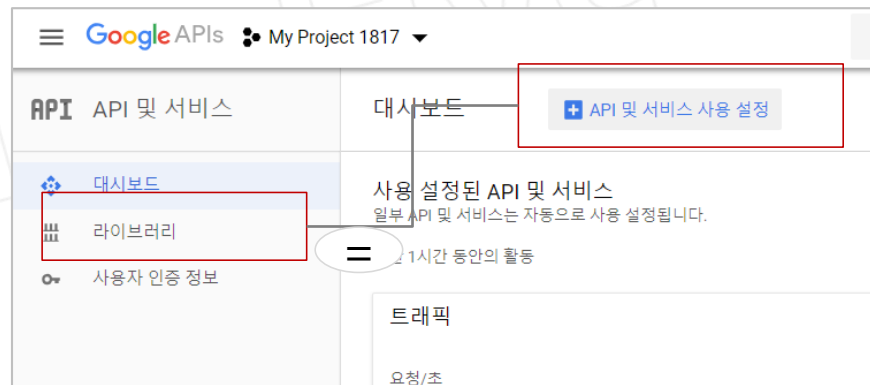
상세 페이지 크롤링하기

스크래핑한 데이터 저장하기

유튜브에서 동영상 정보 수집하기

- Youtube Data API
 - 동영상 검색 및 채널 내용 확인등의 API는 API키만으로 사용 가능
 - API로 추출할 수 있는 것은 동영상과 관련된 메타데이터
 - 동영상 자체는 추출할 수 없음

구글 계정으로 API 신청
<https://console.developers.google.com>



구글 API 사용을 위한 library
`pip install google-api-python-client`

유튜브에서 동영상 정보 수집하기

YouTube KR

요리

홈 인기 구독 라이브러리 최근 본 동영상 나중에 볼 동영상 좋아요 표시한 동... Pandas 팬더스 강... 더보기

구독 Sung Kim Big Data Koo 1 todaycodes오늘... 바람 3 윤인성

관련 필터

승기오빠가 강추하는 맛있는 커피 노블
커피의 건강성분을 등백 담다 향산화 성분과 플레페놀 뽐냄!

매우 쉽다는 놀라운 27가지 요리 법
요리와 같은 요리법 우리의 삶을 열심히 만들고 그들을 피할 수 있는 완벽한 해결책을 발견하는 인기있는 주방 실수를 살펴 봅시다!

라면으로 호텔요리 만든 얼굴천재 고등학생 [맨vs차일드 코리아]
라면으로 코스요리 만드는 16살.avi #구승민 얼굴천재 요리까지 잘해 9:50 라면요리 대결 승자는?!

[백종원 레시피] 김치볶음밥 황금 따라잡기
집밥 백선생2 가 시작했죠~~ 요리다나와에서 먹는 김치볶음밥 황금 ...

code

```
search_response = youtube.search().list(
    part='snippet',
    q='요리',
    type='video',
).execute()
```

실행결과

매우 쉽다는 놀라운 27가지 요리 법
청국장 맛있게 끓이는 법 | 함께요리해요 | 영자씨의 부엌
이렇게 귀여운 향아리가 요리하겠다는데 좀 냅두라 쫌!!!!
라따뚜이 속 바로 그 요리?? 라따뚜이 만들기 : Ratatouille inspired by the Pixar film Ratatouille | Honeykki 꿀키
만화책에 나온 요리를 먹을 수 있는 만화방에 다녀왔습니다.슈슁~장지동 만화 동아리 '장만동'

검색 요청 매개변수의 의미

매개변수 이름	설명
key	API 키
part	응답에 포함할 속성을 쉼표로 구분해서 지정 id와 snippet을 지정 할 수 있음 (id는 default로 자동 설정됨)
q	검색 쿼리
type	검색 대상 리소스의 종류를 쉼표로 구분해서 지정

유튜브에서 동영상 정보 수집하기

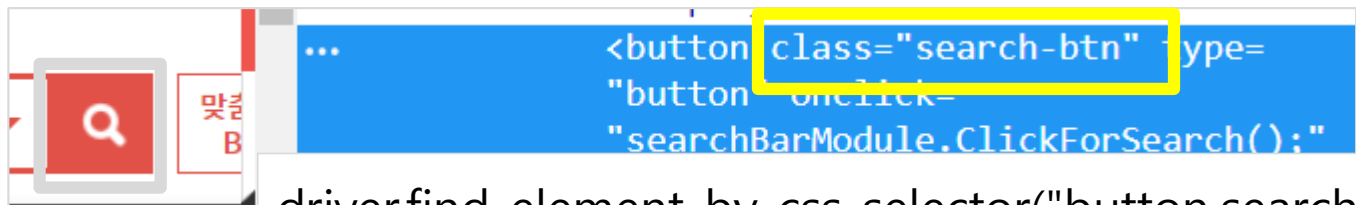
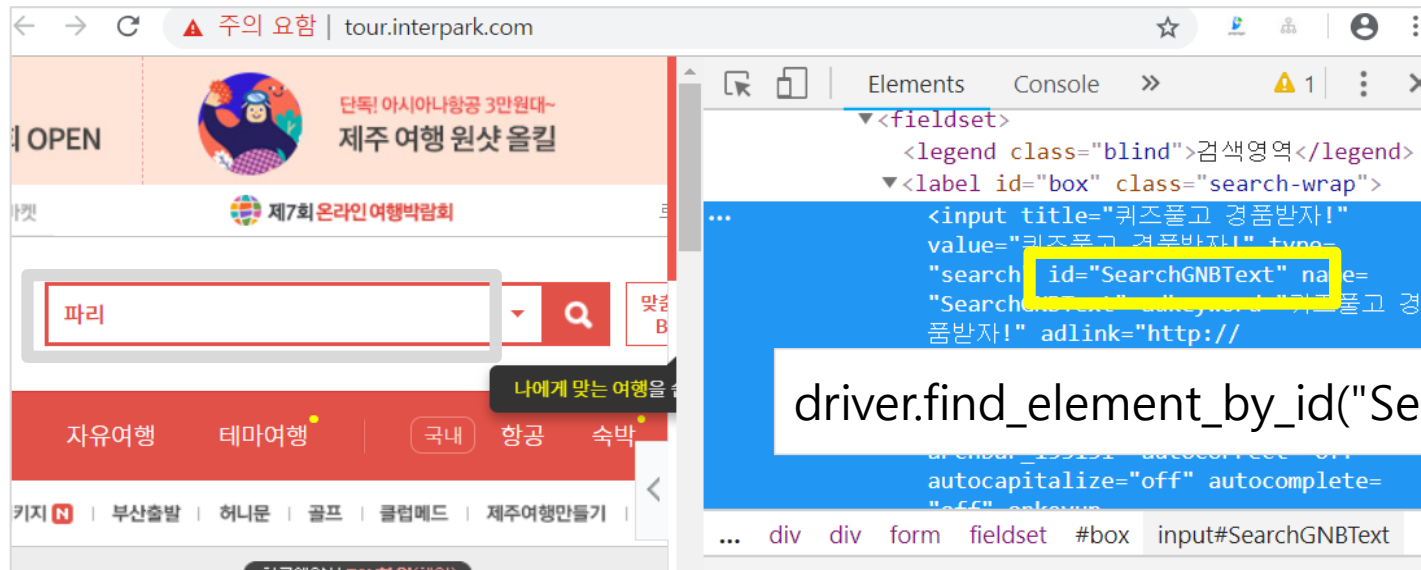
- 선행 조건
 - Google API key

```
1 import os
2
3 from apiclient.discovery import build
4 YOUTUBE_API_KEY = API Key
5
6 youtube = build('youtube', 'v3', developerKey=YOUTUBE_API_KEY)
7
8 search_response = youtube.search().list(
9     part='snippet',
10     q='BTS', ←
11     type='video',
12 ).execute()
13
14 for item in search_response['items']:
15     print(item['snippet']['title'])
```

검색 데이터

여행 정보 수집하기

- <http://tour.interpark.com/>



여행 정보 수집하기

해외여행 [141]

자유여행

[온라인 박람회] OZ사전발권특가-런던/파리 자유 8일

직항으로 편안하게 떠나는

여행 기간 : 6박8일 인천-런던-파라-인천
 출발 가능 기간 : 2018.12.02~2019.10.31

상품득점②
 평점 0
[0개](#)의 상품평

하나로

♥직항으로 떠나는♥ 파리 자유히문 6일

여행 기간 : 4박6일 인천-경유지-파라-경유지-인천
출발 가능 기간 : 2018.12.02~2019.09.30

상품득점⑦
 평점 9
[2개](#)의 상품평

하나로

L.o.v.e. In 파리♥프라하 자유히문 8일

여행 기간 : 6박8일 인천-파라-프라하-경유지-인천
출발 가능 기간 : 2018.12.02~2019.09.30

상품득점⑧
 평점 10
[1개](#)의 상품평

자유여행

[대한항공 직항] 파리/베니스/로마 8일

[대한항공 직항]

여행 기간 : 6박8일 파리-베니스-로마
 출발 가능 기간 : 2018.12.02~2019.06.14

<

<<

1

2

3

4

5

6

7

8

9

10


```
</div>
<div id="tempdiv" style="text-align:
center; display: none;"...</div>
<div class="oTravelBox">...</div>
<div class="pageNumBox">
  <button type="button" class=
  "prevAllBtn" onclick=
  "searchModule.SetCategoryList(1, '')">
  처음</button>
  <ul>
```

여행 정보 수집하기

싸니까 믿으니까 인터파크 투어 x +

← → ↺ 주의 요함 | search-tour.interpark.com/PC/Result?search=파리&code1=R&code2=

인기있는 파리 해외여행



해커지

[두바이★핵심투어]서유럽 3개국9일_알프스뮤젠+피사

■ 관광주거시설 ■ 프랑스의 도시 두바이에서 물의 도시 베니스...


여행 기간 : 9박9일 | 마르세유-파리-로망-스위스-이탈리아
출발 가능 기간 : 2018.11.07~2019.04.30

1,590,000 원~

상품특징
여행코스
11개의 상품할

해외여행 (141)

연관도순 ▼



자유여행

1,000,000 원

Elements Console Sources Network Performance Memory Application Security Audits

```
<!--해외여행-->
<div class="oTravelBox">
  <ul class="boxlist">
    <li class="boxItem">
      <a href="javascript:;" onclick="searchModule.OnClickDetail('http://tour.interpark.com/goods/detail/?BaseGoodsCd=A3013666','')" class="detail8tn" data-click="target">
        
      </a>
      <div class="boxTables">
    </li>
  </ul>
</div>
```

Styles Computed Event Listeners

Filter :hov .<

element.style {

html, body, div, span, object, common

iframe, h1, h2, h3, h4, h5,

h6, p, blockquote, pre, abbr, address

code, del, dfn, em, img, ins, kbd, q,

small, strong, sub, sup, var, b, i, d

dd, ol, ul, li, fieldset, form, label

legend, table, caption, tbody, tfoot,

tr, th, td, article, aside, canvas, d

figcaption, figure, footer, header, h

열린 데이터

- 열린 데이터란?
 - 정부, 자치단체, 기업 등이 보유하고 있는 데이터를 공개적으로 자유롭게 활용 할 수 있게 하는 것
 - 기대효과
 - 데이터 활용에 의한 투명성 제고, 민관 협동, 행정 효율화, 경제 활성화 등의 효과 기대
 - 공공데이터 포털
 - <https://www.data.go.kr>
 - 서울 데이터 열린 광장
 - <http://data.seoul.co.kr>

참고 문헌 및 사이트

- <http://www.w3schools.com>
- 파이썬을 이용한 머신러닝, 딥러닝 실전 개발 입문[위키북스]

Kim Hye Kyung