

웹크롤링 기본

- 라이브러리 설치

`pip install requests`

`pip install beautifulsoup4(bs4)`

requests

사용법 실습

- 라이브러리 사용법

```
import requests
```

```
response = requests.get("https://www.naver.com")
```

```
print(response.text)
```

* `response.text` = 문자열

내가 원하는 부분만 쉽게 추출하기 위해, BeautifulSoup 객체로 만들어 준다

Beautifulsoup

사용법 실습

- 라이브러리 사용법

```
import requests
```

```
from bs4 import BeautifulSoup
```

```
response = request.get("https://www.naver.com")
```

```
html = response.text
```

```
soup = BeautifulSoup(html, 'html.parser')
```

* soup : 객체

특정 태그를 select, select_one 메서드로 추출할 수 있다.

- 태그 한개 선택하기

```
tag = soup.select_one("선택자")
```

* 찾았을 때 : Tag 객체 반환 * 못 찾았을 때 : None 객체 반환

- 태그 여러개 선택하기

```
tags = soup.select ("선택자")
```

* 찾았을 때 : Tag 객체 리스트 반환 * 못 찾았을 때 : 빈 리스트 반환

- Tag 객체

HTML 태그의 텍스트, 속성 등의 다양한 정보를 가지고 있다.

```
['DEFAULT_INTERESTING_STRING_TYPES', '__bool__', '__call__', '__class__', '__contains__', '__copy__', '__delattr__', '__delitem__', '__dict__', '__dir__', '__doc__', '__eq__', '__format__', '__ge__', '__getattr__', '__getattribute__', '__getitem__', '__gt__', '__hash__', '__init__', '__init_subclass__', '__iter__', '__le__', '__len__', '__lt__', '__module__', '__ne__', '__new__', '__reduce__', '__reduce_ex__', '__repr__', '__setattr__', '__setitem__', '__sizeof__', '__str__', '__subclasshook__', '__unicode__', '__weakref__', '__all_strings__', 'find_all', 'find_one', 'is_xml', 'lastRecursiveChild', 'last_descendant', 'namespaces', '_should_pretty_print', 'append', 'attrs', 'can_be_empty_element', 'cdata_list_attributes', 'childGenerator', 'children', 'clear', 'contents', 'decode', 'decode_contents', 'decompose', 'decomposed', 'default', 'descendants', 'encode', 'encode_contents', 'extend', 'extract', 'fetchNextSiblings', 'fetchParents', 'fetchPrevious', 'fetchPreviousSiblings', 'find', 'findAll', 'findAllNext', 'findAllPrevious', 'findChild', 'findChildren', 'findNext', 'findNextSibling', 'findNextSiblings', 'findParent', 'findParents', 'findPrevious', 'findPreviousSibling', 'findPreviousSiblings', 'find_all', 'find_all_next', 'find_all_previous', 'find_next', 'find_next_sibling', 'find_next_siblings', 'find_parent', 'find_parents', 'find_previous', 'find_previous_sibling', 'find_previous_siblings', 'format_string', 'formatter_for_name', 'get', 'getText', 'get_attribute_list', 'get_text', 'has_attr', 'has_key', 'hidden', 'index', 'insert', 'insert_after', 'insert_before', 'interesting_string_types', 'isSelfClosing', 'is_empty_element', 'known_xml', 'name', 'namespace', 'next', 'nextGenerator', 'nextSibling', 'nextSiblingGenerator', 'next_element', 'next_elements', 'next_sibling', 'next_siblings', 'parent', 'parentGenerator', 'parents', 'parserClass', 'parser_class', 'prefix', 'preserve_whitespace_tags', 'prettify', 'previous', 'previousGenerator', 'previousSibling', 'previousSiblingGenerator', 'previous_element', 'previous_elements', 'previous_sibling', 'previous_siblings', 'recursiveChildGenerator', 'renderContents', 'replaceWith', 'replaceWithChildren', 'replace_with', 'replace_with_children', 'select', 'select_one', 'setup', 'smooth', 'sourceline', 'sourcepos', 'string', 'strings', 'stripped_strings', 'text', 'unwrap', 'wrap']
```


- 자주 사용하는 Tag 객체의 속성과 메서드

명칭	기능
tag.text	태그안의 모든 텍스트 요소
tag.attrs["속성명"]	태그의 속성값
tag.select_one("선택자")	해당 태그 안에서 한개 선택
tag.select("선택자")	해당 태그 안에서 여러개 선택

정적인 페이지 크롤링 실습

크롤링 연습 하자

startcoding's crawling practice website

<https://startcoding-crawling.herokuapp.com/>

크롤링 초급

가장 기본적인 형태의 웹페이지를 크롤링 할 수 있습니다. request, beautifulsoup4 라이브러리를 이용해 줍니다.

▶ [연습하러가기](#)

Step1. 하나의 게시물 크롤링하기

숲 이론

실전! 네이버 뉴스 크롤링 - 파이썬으로 데이터 수집 쉽게 하자 (1단계)

****주의 사항**** 1. 상업적 용도로 크롤링한 정보를 이용하지 말 것 2. 서버에 부담을 줄 정도로 많은 요청을 하지 말 것 네이버 뉴스 크롤링을 초보자 분들도 쉽게 할 수 있도록 영상을 제작하였습니다. (1단계) 네이버 뉴스 1페이지 제목과 링크 크롤링하기 (2단계) 네이버 뉴스 검색어 변경하면서 크롤링하기 (3단계) 네이버 뉴스 여러 페이지 가져오기

[유튜브보러가기](#)

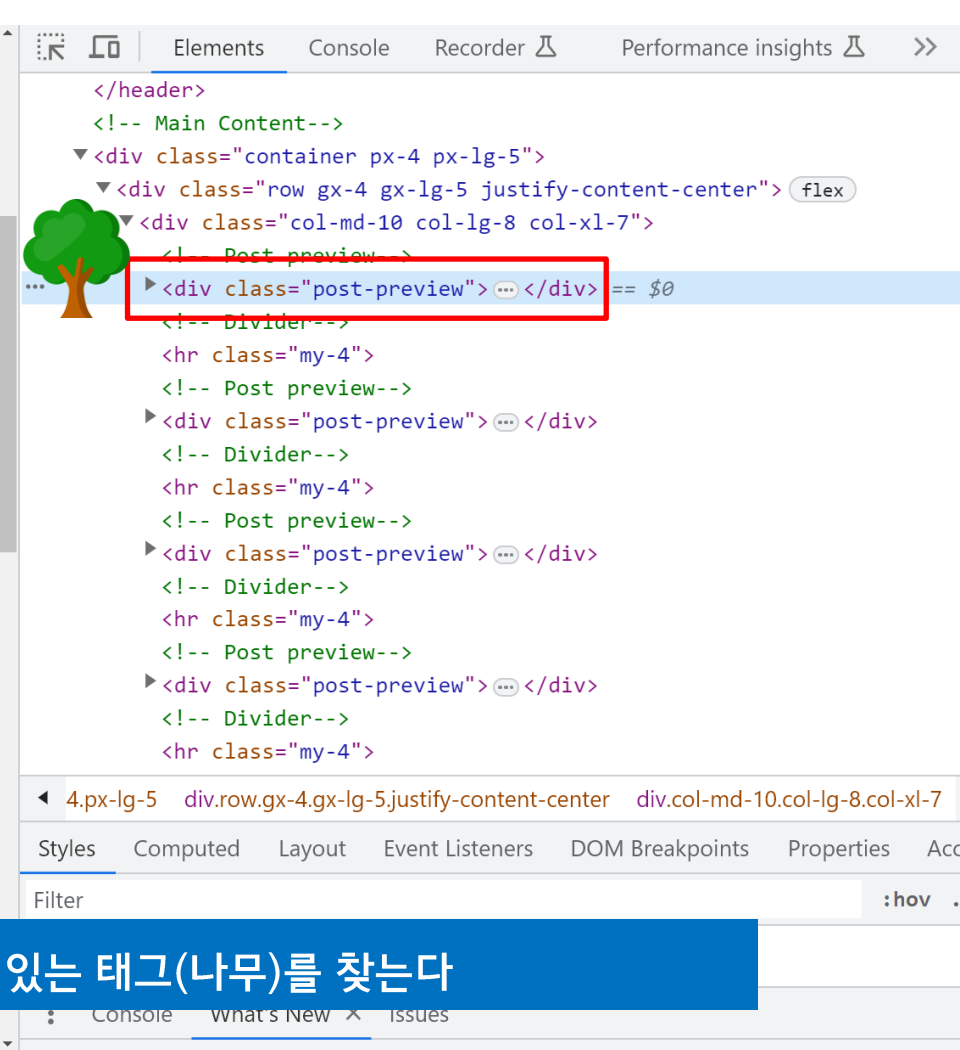
30만원 짜리 외주 프로그램 만들기 - 파이썬으로 네이버 쇼핑 크롤링하기 1편 (ft. 셀레니움 selenium)

30만원을 받고 만들어 준 외주 프로그램 만드는 과정을 공개합니다. 이런 거 다 공개해도 괜찮겠죠.? 네이버 쇼핑 상품 정보를 검색어 별로 크롤링 후 엑셀로 저장하는 프로그램입니다. 여러분도 강의를 듣고 나면 프로그램을 제작해서 돈을 벌 수 있습니다.

[유튜브보러가기](#)

(지금 당장) 우리가 코딩을 배워야 하는 3가지 이유

정리를 하면 코딩을 배워야 하는 이유는 첫 번째, 컴퓨터에게 일을 시킬 수 있고 두 번째, 세상에 없는 나만의 서비스를 개발할 수 있고 세 번째, 실력이 생기면 개발자로 취업이 가능하다. 스타트코딩이 여러분의 코딩 입문을 도와드



1. 원하는 정보를 모두 담고 있는 태그(나무)를 찾는다

실전! 네이버 뉴스 크롤링 - 파이썜으로 데이터 수집 쉽게 하자 (1단계)

****주의 사항**** 1. 상업적 용도로 크롤링한 정보를 이용하지 말 것 2. 서버에 부담을 줄 정도로 많은 요청을 하지 말 것 네이버 뉴스 크롤링을 초보자 분들도 쉽게 할 수 있도록 영상을 제작하였습니다. (1단계) 네이버 뉴스 1페이지 제목과 링크 크롤링하기 (2단계) 네이버 뉴스 검색어 변경하면서 크롤링하기 (3단계) 네이버 뉴스 여러 페이지 가져오기

[유튜브보러가기](#)

30만원 짜리 외주 프로그램 만들기 - 파이썜으로 네이버 쇼핑 크롤링하기 1편 (ft. 셀레니움 selenium)

30만원을 받고 만들어 준 외주 프로그램 만드는 과정을 공개합니다. 이런 거 다 공개해도 괜찮겠죠..? 네이버 쇼핑 상품 정보를 검색어 별로 크롤링 후 엑셀로 저장하는 프로그램입니다. 여러분도 강의를 듣고 나면 프로그램을 제작해서 돈을 벌 수 있습니다.

[유튜브보러가기](#)

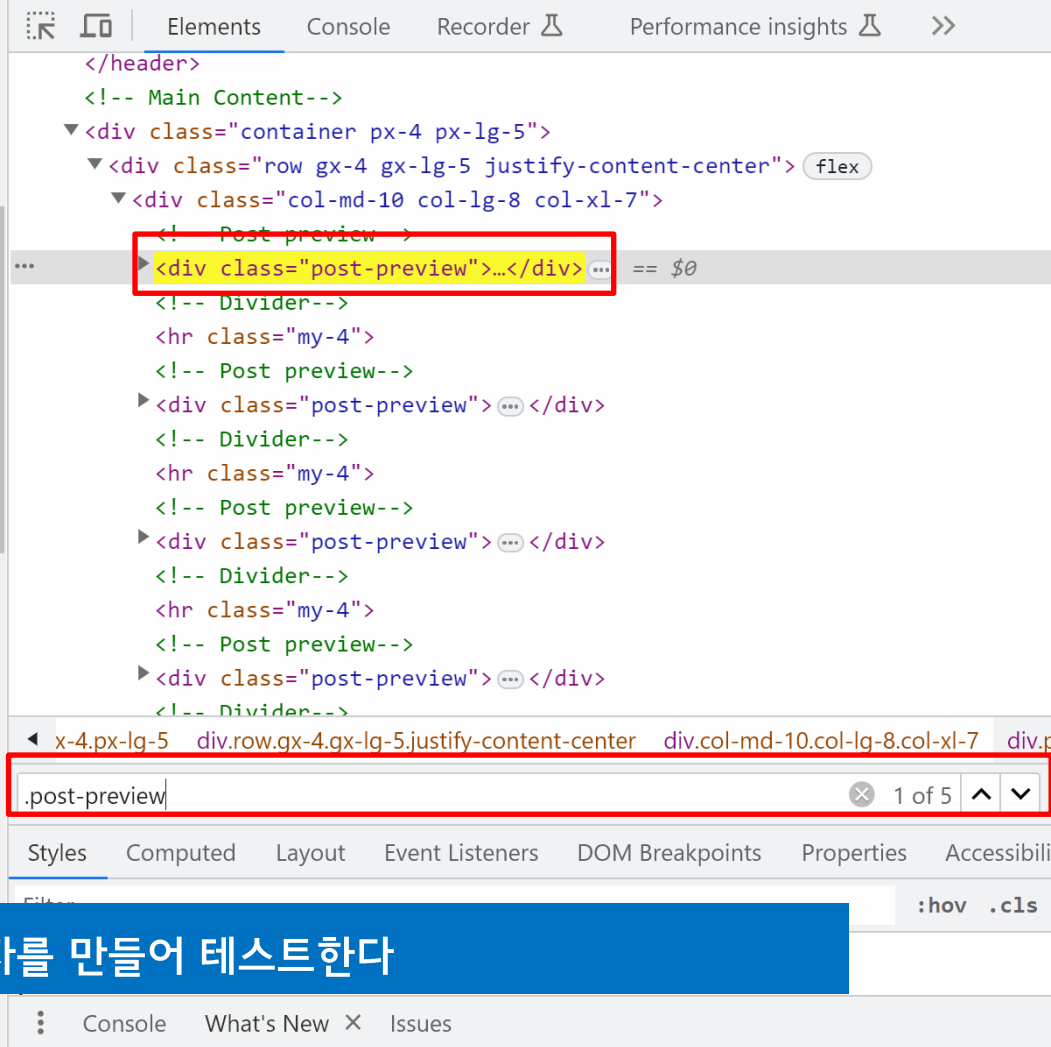
(지금 당장) 우리가 코딩을 배워야 하는 3가지 이유

정리를 하면 코딩을 배워야 하는 이유는 첫 번째, 컴퓨터에게 일을 시킬 수 있고 두 번째, 세상에 없는 나만의 서비스를 개발할 수 있고 세 번째, 실력이 생기면 개발자로 취업이 가능하다. 스타트코딩이 여러분의 코딩 입문을 도와드리겠습니다.

[유튜브보러가기](#)




2. CSS 선택자를 만들어 테스트한다





```
1 import requests
2 from bs4 import BeautifulSoup
3
4 response = requests.get("https://startcoding-crawling.herokuapp.com/lv1")
5 html = response.text
6 soup = BeautifulSoup(html, 'html.parser')
7 posts = soup.select(".post-preview")
```

3. soup.select("CSS선택자")로 숲에서 나무들을 뽑는다



```
1 import requests
2 from bs4 import BeautifulSoup
3
4 response = requests.get("https://startcoding-crawling.herokuapp.com/lv1")
5 html = response.text
6 soup = BeautifulSoup(html, 'html.parser')
7 posts = soup.select(".post-preview")
8 for post in posts:
9     title = post.select_one(".post-title").text
10    link = post.select_one("a").attrs['href']
11    content = post.select_one(".post-subtitle").text
12    print(title, link, content)
```

4. 반복문을 돌면서 나무에서 하나씩 열매를 추출한다

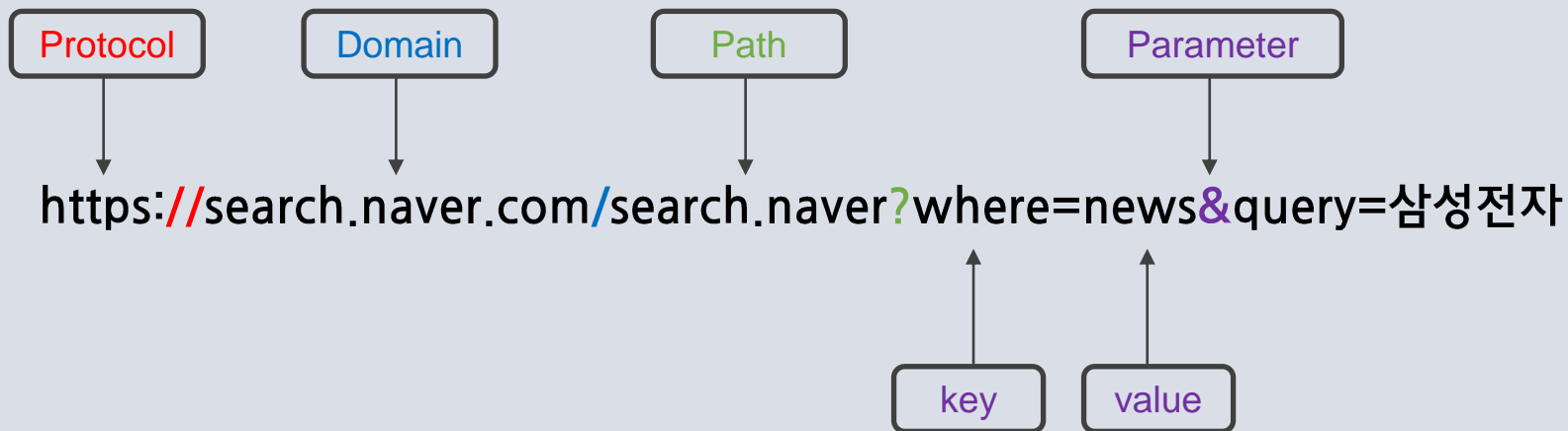
Step2. 여러 개의 게시물 크롤링하기

URL 조작자

URL

- 인터넷 주소 형식
- Protocol - Domain - Path - Parameter

URL



URL

A screenshot of a web browser window illustrating the components of a URL. The browser's address bar shows the URL: `https://comic.naver.com/webtoon/list?titleId=736277&weekday=sun`. Below the address bar, four labels in colored boxes point to specific parts of the URL: 'Protocol' (red) points to 'https', 'Domain' (blue) points to 'comic.naver.com', 'Path' (green) points to '/webtoon/list', and 'Parameter' (purple) points to '?titleId=736277&weekday=sun'. Above the address bar, a tab is labeled '싸움독학 :: 네이버 만화'. To the right of the tab, four boxes labeled 'key', 'value', 'key', and 'value' are shown, corresponding to the query parameters in the URL. Below the address bar, a row of bookmarks is visible: '앱', '패스트캠퍼스', '자기계발강의', '강의제작', 'Tools', and '마케팅'. At the bottom of the browser window, the 'NAVER 만화' logo is visible, along with navigation links: '홈', '웹툰' (highlighted in green), and '베스트 도전'.

싸움독학 :: 네이버 만화

key value key value

← → ↻ <https://comic.naver.com/webtoon/list?titleId=736277&weekday=sun>

앱 패스트캠퍼스 자기계발강의 강의제작 Tools 마케팅

Protocol Domain Path Parameter

NAVER 만화 | 웹소설

홈 웹툰 베스트 도전

페이징 알고리즘

페이징 알고리즘

<https://startcoding-crawling.herokuapp.com/lv1?page=1>
<https://startcoding-crawling.herokuapp.com/lv1?page=2>
<https://startcoding-crawling.herokuapp.com/lv1?page=3>
<https://startcoding-crawling.herokuapp.com/lv1?page=4>

1. 페이지를 바꾸면서 URL이 변경되는 부분을 찾는다

페이징 알고리즘

```
1 import requests
2 from bs4 import BeautifulSoup
3
4 for i in range(1, 5):
5     response = requests.get(f"https://startcoding-crawling.herokuapp.com/lv1?page={i}")
6     html = response.text
7     soup = BeautifulSoup(html, 'html.parser')
8     posts = soup.select(".post-preview")
9     for post in posts:
10         title = post.select_one(".post-title").text
11         content = post.select_one(".post-subtitle").text.replace('\r\n', '')
12         link = post.select_one("a").attrs['href']
13         print(title, content, link, sep='\n')
```

2. 페이지를 증가시키면서 요청을 보낸다

Step3. 여러 페이지 크롤링하기

엑셀 저장

- 라이브러리 설치

`pip install pandas`

`pip install openpyxl`



```
1 import requests
2 from bs4 import BeautifulSoup
3 import pandas as pd
4
5 data = []
6 for i in range(1, 5):
7     response = requests.get(f"https://startcoding-crawling.herokuapp.com/lv1?page={i}")
8     html = response.text
9     soup = BeautifulSoup(html, 'html.parser')
10    posts = soup.select(".post-preview")
11    for post in posts:
12        title = post.select_one(".post-title").text
13        content = post.select_one(".post-subtitle").text.replace('\r\n', '')
14        link = post.select_one("a").attrs['href']
15        print(title, content, link, sep='\n')
16        data.append([title, content, link])
```

1. 비어있는 리스트를 만들고 데이터를 한행씩 추가한다



```
1 # 데이터 프레임 만들기
2 df = pd.DataFrame(data, columns=['제목', '내용', '링크'])
3
4 # 엑셀로 저장하기
5 df.to_excel('result.xlsx')
```

2. 데이터 프레임을 만들고 엑셀로 저장한다

엑셀 저장 결과

	A	B	C	D
1		제목	내용	링크
2	0	실전! 네이버 뉴스 크롤링 - 파이썬으로 데이터 수집 쉽게	****주의 사항****1. 상업적 용도로 크롤링	https://www.youtube.com/watch?v=U1amkBqKF5g
3	1	30만원 짜리 외주 프로그램 만들기 - 파이썬으로 네이버 쇼	30만원을 받고 만들어 준 외주 프로그램 만	https://www.youtube.com/watch?v=ZHx6oATaI28
4	2	(지금 당장) 우리가 코딩을 배워야 하는 3가지 이유	정리를 하면코딩을 배워야 하는 이유는 첫	https://www.youtube.com/watch?v=WqeEIOYzhkM
5	3	[초보도 가능] 30만원 외주 프로그램 만들기 - 파이썬 웹	[강의 설명]초보자도 가능한 30만원 짜리 프	https://www.youtube.com/watch?v=qRU94vtUb7c
6	4	잠 안잘고 하루에 1시간 버는 법 #반복업무탈출 #업무지	영상을 보고 나면, 여러분이 직접 로봇을 가	https://www.youtube.com/watch?v=rMRQo9XbPgk
7	5	컴퓨터에게 일 시키고 노는 방법 #키보드자동화 #업무자동	여러분이 지금 하고 있는 반복 업무를 자동	https://www.youtube.com/watch?v=NI5dEsKsaSE
8	6	파이썬 셀레니움 네이버 로그인을 만들면서 배우는 웹사이	원래 방법을 공개하지 않으려 했는데요...	https://www.youtube.com/watch?v=fy107mUHapQ&t=22s
9	7	누구나 할 수 있는 파이썬 코딩을 활용한 업무 자동화 기초	파이썬 코딩을 활용해서 네이버 메일을 자	https://www.youtube.com/watch?v=3VRw7UVJPQk
10	8	사무직 칼퇴 필수 기술 파이썬 엑셀 업무 자동화 배우기	반복되는 엑셀 업무로 고통 받고 있나요?	https://www.youtube.com/watch?v=EjTGSYCWmEo
11	9	(NEW) 파이썬 기초 강의, "당신의 커리어에 파이썬을 더하	코딩을 가장 쉽게 알려주는 크리에이터, 스	https://www.youtube.com/watch?v=REUu0T1xsiU&t=33s
12	10	오늘은 엑셀 매크로 대신, 파이썬을 배우고 싶다	엑셀 매크로 보다, 쓸모가 많은 파이썬 진짜	https://www.youtube.com/watch?v=Z34SVo_jGr4
13	11	엑셀 VBA 대신 파이썬 - 실무 중심 예제 1탄	코딩을 가장 쉽게 알려주는 크리에이터, 스	https://www.youtube.com/watch?v=UT68mutiJJI&t=971s
14	12	파이썬 실행파일 (EXE) 만들기, 배포, 변환 이 영상만 보세	유튜브 구독자 분들이 가장 많이 질문 주셨	https://www.youtube.com/watch?v=nuZAmSyCMgY
15	13	직장인 엑셀 자동화, 똑똑하게 일할 사람만 보세요 (Feat. 피	직장인으로 엑셀 작업이 많으신가요?반복!	https://www.youtube.com/watch?v=29Vh_RgQNhM&t=511s
16	14	팀장님께 인정받는 일잘러는 '이걸' 활용한다며? (크롤링	(일잘러 = 일을 잘하는 사람)요즘 대부분의	https://www.youtube.com/watch?v=dBQ0tbYjgMc
17	15	파이썬 초보에서 벗어나려면 꼭 봐야하는 영상 TOP1 (full)	코딩을 가장 쉽게 알려주는 크리에이터 스	https://www.youtube.com/watch?v=rWiM-QjBRIs
18	16	파이썬 클래스, 객체, 인스턴스, 생성자, 메서드, self 개념	중간명하세요.코딩을 가장 쉽게 알려주는 크	https://www.youtube.com/watch?v=FRHGtAvU03Q