웹크롤링(Web Crawaling)1

requests

웹사이트의 정보를 가져오기 위한 파이썬 라이브러리

설치

```
pip3 install requests
```

형식	import requests response = requests.get('https://www.naver.com/') print(response.status_code) # 응답코드를 출력 print(response.text) # HTML 코드를 출력

파라미터를 전달해야 할 경우 JSON형식으로 작성후 추가

형식	param = {
	'pageNo' : 1,
	'rangeType' : 'ALL',
	'orderBy' : 'sim',
	'keyword' : '파이썬'
	}
	
	response = requests.get('https://section.blog.naver.com/Search/Post.nhn',
	params=params)

BeautifulSoup

HTML정보로 부터 원하는 데이터를 가져오기 쉽게, 비슷한 분류의 데이터별로 나누어주는(parsing) 파이썬 라이브러리

설치

```
pip3 install beautifulsoup4
```

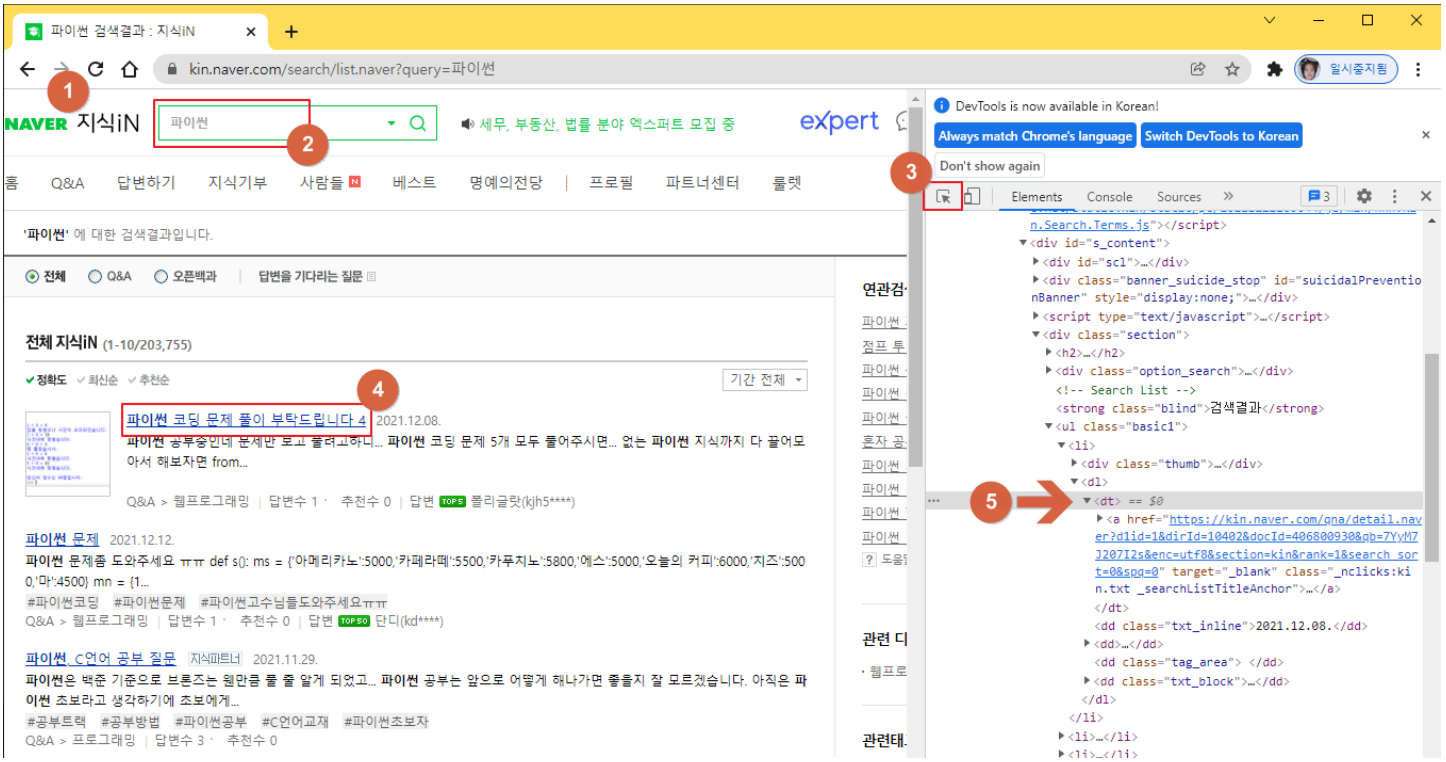
형식

```
import requests
from bs4 import BeautifulSoup

url = 'http://daum.net/'
response = requests.get(url)

if response.status_code == 200:
    html = response.text
    soup = BeautifulSoup(html, 'html.parser')
    print(soup)
else :
    print(response.status_code)
```

from bs4 import BeautifulSoup 에러날때 참조 => `python -m pip install bs4`

크롤링 실습하기

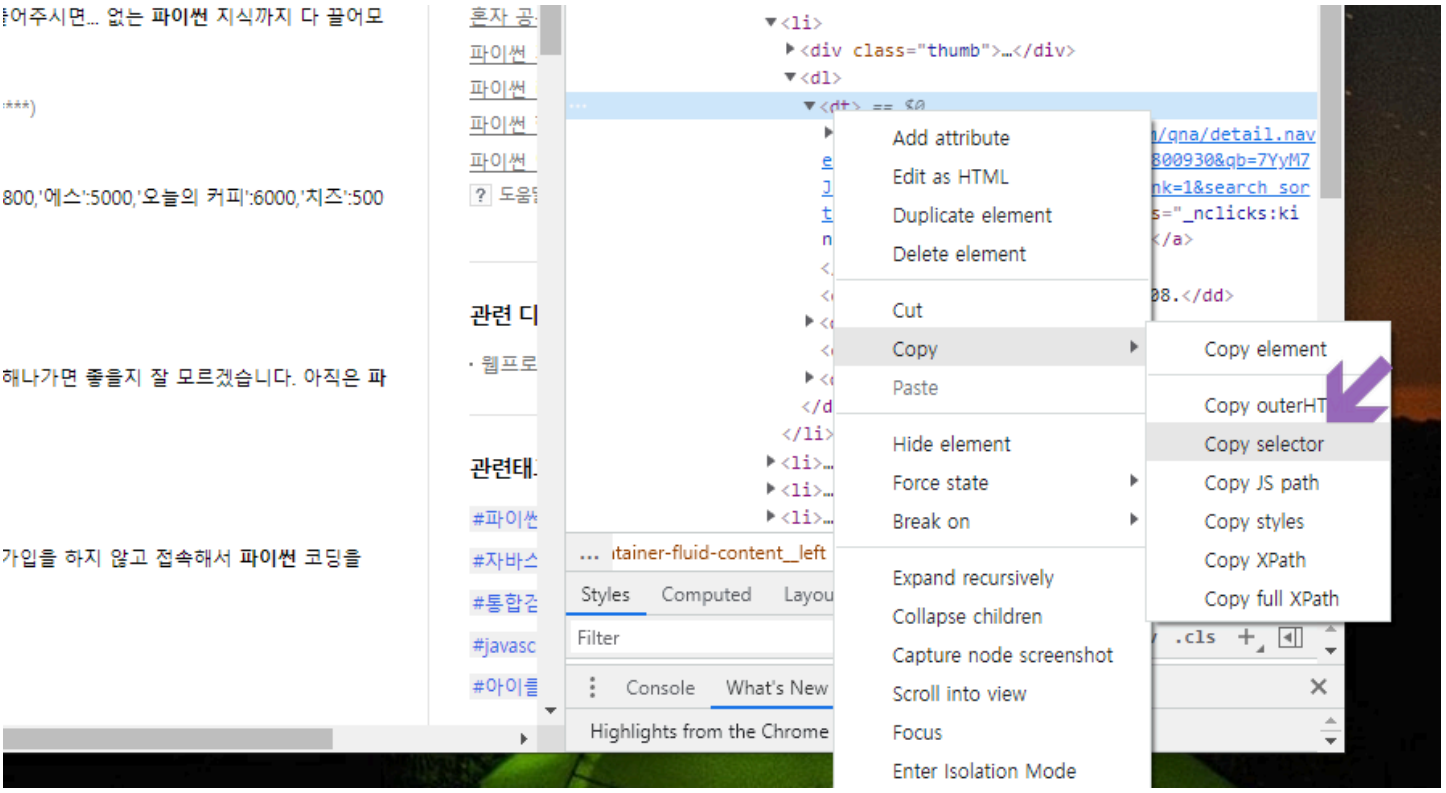
먼저 네이버 메인에서 지식in으로 이동한 후 파이썬을 검색한다.



검색 결과에서 그림의 순서대로 개발자도구의 인스펙트(Inspect)를 이용해서 첫번째 항목을 선택한다.

그러면 Elements 항목에서 <dt> 태그가 보일것이다.

이 부분을 우클릭한다.



Copy > Copy selector를 찾아 클릭한다.

그러면 CSS선택자가 복사된다.

크롬에서의 복사 결과

#s_content > div.section > ul > li:nth-child(1) > dl > dt

파이어폭스에서의 복사 결과

.basic1 > li:nth-child(1) > dl:nth-child(2) > dt:nth-child(1)

웹브라우저에 따라 복사된 결과는 조금 다르지만 크롤링한 결과는 동일하다.

예제] 19webCrawling01.py

```
1   import requests
2   from bs4 import BeautifulSoup
3
4   url = 'https://kin.naver.com/search/list.nhn?query=%ED%8C%8C%EC%9D%B4%EC%8D%AC'
5   response = requests.get(url)
6
7   if response.status_code==200:
8       html = response.text
9       soup = BeautifulSoup(html, 'html.parser')
10
11      # 셀렉터로 추출(크롬)
12      title1_1 = soup.select_one('#s_content > div.section > ul > li:nth-child(1) > dl > dt')
13      #print("추출1_1:", title1_1)
14      # 셀렉터로 추출(파이어폭스)
15      #title1_2 = soup.select_one('.basic1 > li:nth-child(1) > dl:nth-child(2) > dt:nth-child
        (1)')
16      #print("추출1_2:", title1_2)
17
```

네모 부분은 복사해서 작성하세요.

```
18      text = title1_1.get_text()
19      print("추출2:", text)
20
21      ul = soup.select_one('ul.basic1')
22      print("추출3:", ul)
23
24      #print("추출4")
25      titles2 = ul.select('li > dl > dt > a')
26      for tit in titles2:
27          print(tit.get_text())
28  else:
29      print(response.status_code)
```

예제] 19webCrawling02.py

https://www.koreabaseball.com/Record/Player/HitterBasic/BasicOld.aspx?sort=HRA_RT

```python
import requests
from bs4 import BeautifulSoup

response = requests.get("https://www.koreabaseball.com/Record/Player/HitterBasic/BasicOld.aspx?sort=HRA_RT")
html = response.text
soup = BeautifulSoup(html, 'html.parser')
#print(soup)

title = soup.select_one('#contents > h4')
#print("title 요소 :", title)

title_txt = title.get_text()
#print("title 텍스트 :", title_txt)

record_table = soup.select_one('#cphContents_cphContents_cphContents_udpContent > div.record_result > table')
#print("타자기록 요소 :", record_table)

record_tr = soup.select_one('#cphContents_cphContents_cphContents_udpContent > div.record_result > table > tbody')
repeat_tr = record_tr.select('tr')
for rec in repeat_tr:
    #print("dddd", rec)

    d1 = rec.select_one('td:nth-child(1)').get_text() # 순위
    d2 = rec.select_one('td:nth-child(2)').get_text() # 선수명
    d3 = rec.select_one('td:nth-child(3)').get_text() # 팀명
    d4 = rec.select_one('td:nth-child(4)').get_text() # 타율
    d5 = rec.select_one('td:nth-child(5)').get_text() # 경기
    d6 = rec.select_one('td:nth-child(6)').get_text() # 타석
    d7 = rec.select_one('td:nth-child(7)').get_text() # 타수
    d8 = rec.select_one('td:nth-child(8)').get_text() # 안타
    d9 = rec.select_one('td:nth-child(9)').get_text() # 2루타
    d10 = rec.select_one('td:nth-child(10)').get_text() # 3루타

    d11 = rec.select_one('td:nth-child(11)').get_text() # 홈런
    d12 = rec.select_one('td:nth-child(12)').get_text() # 타점
    d13 = rec.select_one('td:nth-child(13)').get_text() # 도루
    d14 = rec.select_one('td:nth-child(14)').get_text() # 도루실패
    d15 = rec.select_one('td:nth-child(15)').get_text() # 볼넷
    d16 = rec.select_one('td:nth-child(16)').get_text() # 사구
    d17 = rec.select_one('td:nth-child(17)').get_text() # 삼진
    d18 = rec.select_one('td:nth-child(18)').get_text() # 병살타
    d19 = rec.select_one('td:nth-child(19)').get_text() # 실책
```

```
42      #출력
43      print(d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12, d13, d14, d15, d16, d17,
        d18, d19)
44      #DB입력
45
```