검색어를 입력하세요.

파이썬 웹크롤링 과 자동화에 대한 A to Z
(/book/4706)

1. 준비단계 - 스트레칭

1.1 마음가짐 - 무엇이든 할 수 있게 하는 힘

1.2 목 돌리기 - 필요한 파이썬 문법만 알고가자

2. 기초단계 - 유산소운동

2.1 마우스 자동화 - pyautogui 사용법 (1)

2.2 키보드 자동화 - pyautogui 사용법 (2)

2.3 메세지 박스 - pyautogui 사용법(3)

2.4 이미지로 좌표찾기 - pyautogui 사용법 (4)

2.5 사이트 정보 가져오기 - requests 사용법

2.6 사이트 정보 추출하기 - beautifulsoup 사용법 (1)

2.7 사이트 정보 추출하기 - beautifulsoup 사용법 (2)

2.8 사이트 자동화하기 - selenium 사용법(1)

3. 응용단계 - 웨이트 트레이닝

3.0 파이썬 엑셀 다루기 - openpyxl 사용법

3.1 웹스크래핑 예제(1) - 네이버금융 실시간 검색 순위 스크래핑 후 엑셀에 저장하기

3.2 웹스크래핑 예제(2) - 네이버금융 실시간 주가 크롤링하기

2.6 사이트 정보 추출하기 - beautifulsoup 사용법 (1)

BeautifulSoup가 필요한 이유

request.text를 이용해 가져온 데이터는 텍스트형태의 html 입니다.

텍스트형태의 데이터에서 어떻게 원하는 html 요소에 접근할 수 있을까요?

이를 쉽게 할 수 있게 도와주는 녀석이 바로 "뷰티풀수프"입니다!! (이름이 특이하죠)

즉, 날 것의 html을 의미있는 객체로 만들어서 사용자가 요리하기 쉽게 만드는 겁니다.

BeautifulSoup 설치

```
pip install beautifulsoup4
```

BeautifulSoup 사용법

```
import requests
from bs4 import BeautifulSoup

url = 'https://kin.naver.com/search/list.nhn?query=%ED%8C%8C%EC%9D%B4%EC%8D%AC'

response = requests.get(url)

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

else :
    print(response.status_code)
```

네이버 지식인에 파이썬을 검색한 url 입니다. 응답 코드가 200 일때, html 을 받아와 soup 객체로 변환 합니다.

네이버 지식인 크롤링 예제

네이버 지식인에 "파이썬"을 검색하면 첫 번째로 나오는 제목을 가져오려면 어떻게 할까요?

먼저, 구글 개발자 도구를 사용할줄 알아야해요!!! 제가 친절하게 알려 드리겠습니다.

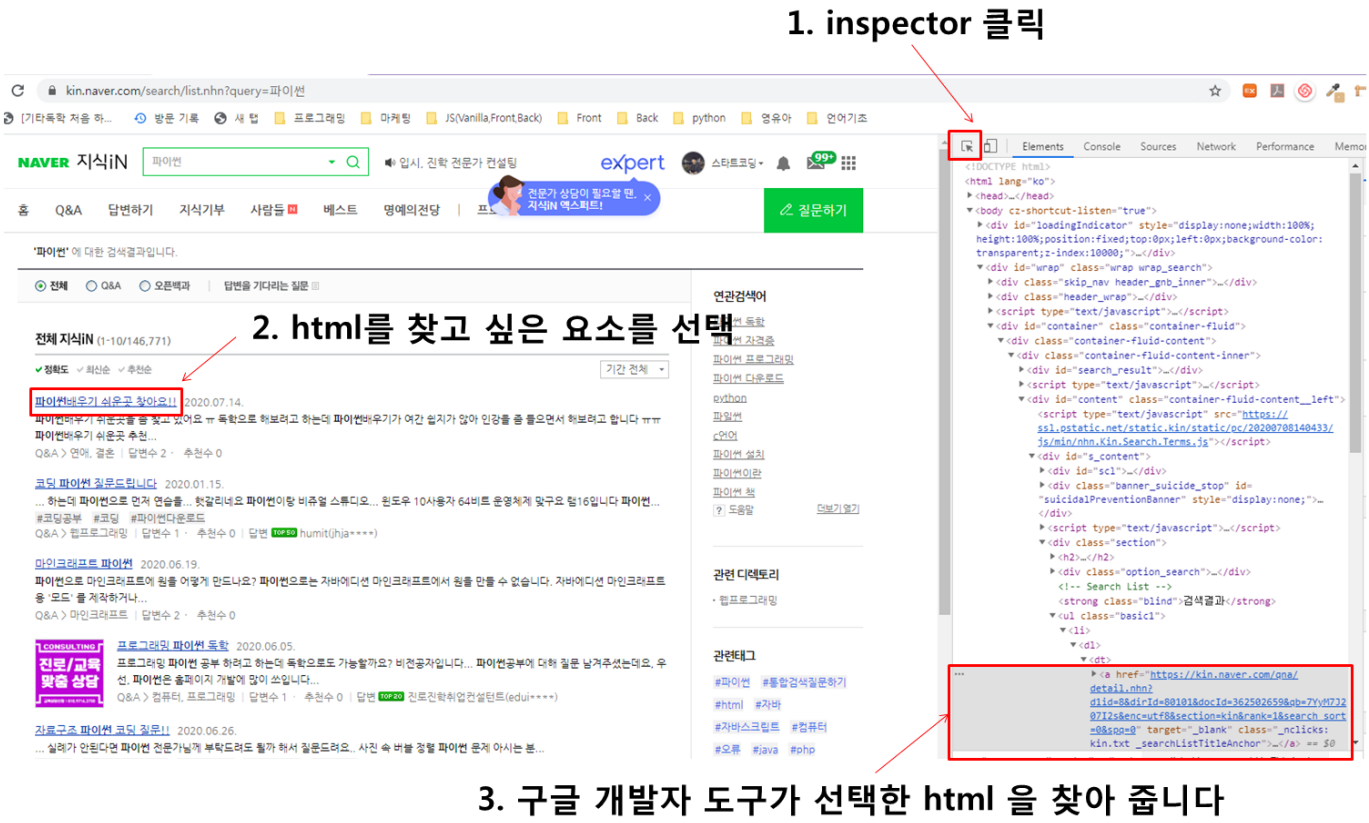
크롬 브라우저를 열고 "https://kin.naver.com/search/list.nhn?query=%ED%8C%8C%EC%9D%B4%EC%8D%AC" 주소로 가봅시다!

그리고 f12 버튼을 눌러보세요.

오른쪽에서 뽕하고 나오는게 바로 구글 개발자 도구입니다. (크롤링을 도와줄 아주 멋진 녀석이죠)

여러분에게 필요한 건 원하는 html 요소가 어디있는지 찾아주는 inspector 기능입니다

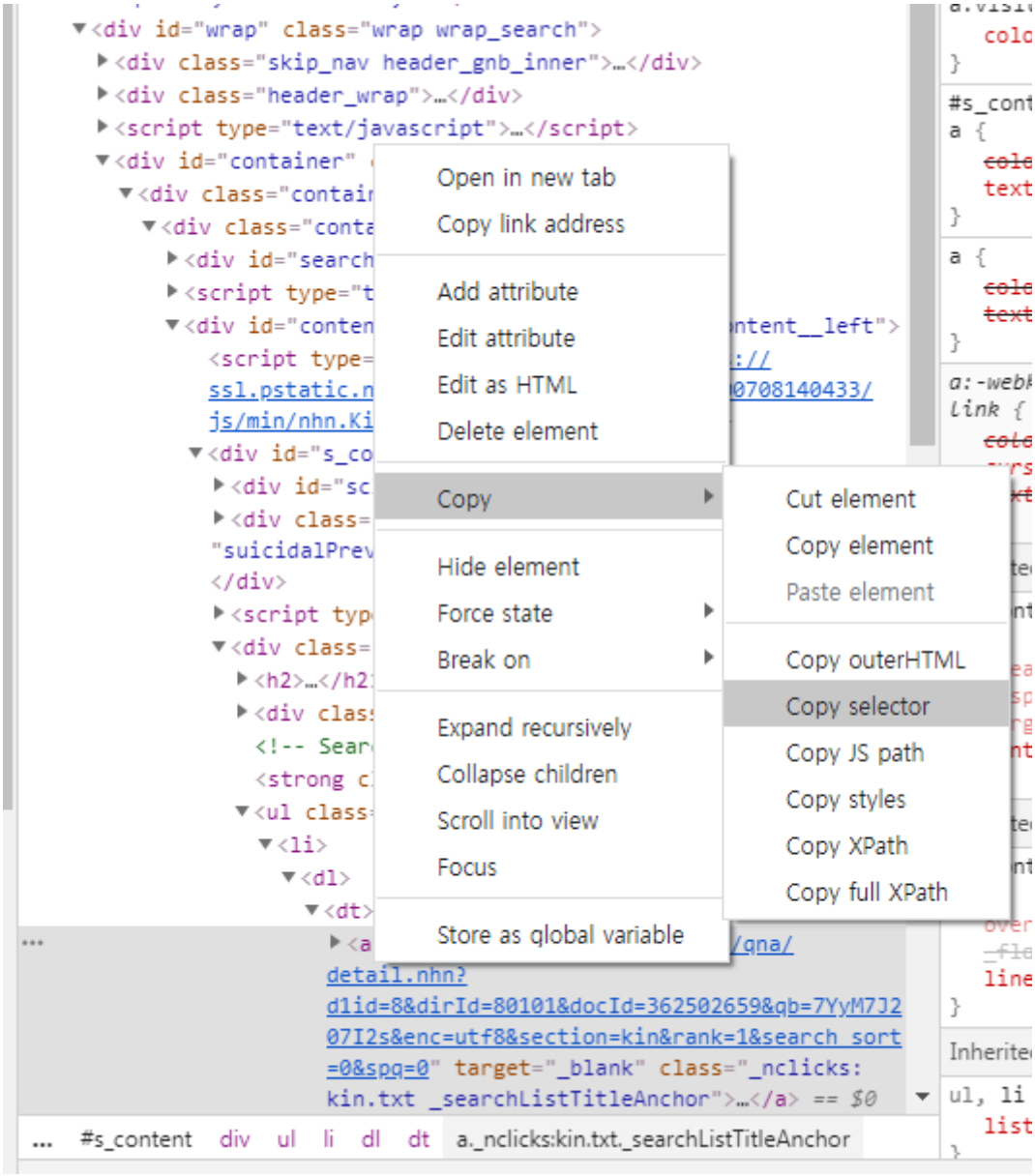
사용방법은 이미지에 첨부할께요



찾은 html에 copy selector 기능을 이용해 봅시다.

Css 선택자를 자동으로 찾아주는 엄청난 기능입니다.

html에 오른쪽 클릭을 한 후 Copy -> Copy Selector 를 선택해 줍니다.



이제 클립보드에 css 선택자가 복사되었는데요.

이를 코드에 적용해 보겠습니다.

```
import requests
from bs4 import BeautifulSoup

url = 'https://kin.naver.com/search/list.nhn?query=%ED%8C%8C%EC%9D%B4%EC%8D%AC'

response = requests.get(url)

if response.status_code == 200:
    html = response.text
    soup = BeautifulSoup(html, 'html.parser')
    title = soup.select_one('#s_content > div.section > ul > li:nth-child(1) > dl > dt > a')
    print(title)
else :
    print(response.status_code)
```

select_one 은 하나의 html 요소를 찾는 함수인데, css 선택자를 사용해서 찾을 수 있습니다. 복사한 css 선택자를 select_one 함수의 인자로 넣어주세요.

```
결과물
<a class="_nclicks:kin.txt _searchListTitleAnchor" href="https://kin.naver.com/qna/detail.nhn?d1id=8&amp;dirId=80101&amp;docId=362502659&amp;qb=7YyM7J207I2s&amp;enc=utf8§ion=kin&amp;rank=1&amp;search_sort=0&amp;spq=0" target="_blank"><b>파이썬</b>배우기 쉬운곳 찾아요!!</a>
```

텍스트만 뽑아오고 싶다면 get_text() 함수를 이용하면 됩니다.

```
print(title.get_text())
```

어때요! 사이트에서 원하는 정보를 파이썬으로 가져오다니 정말 놀랍지 않나요??!

다음 장에서는 네이버 지식인의 제목 10개를 모두 가져오는 방법에 대해서 알아보도록 하겠습니다.

마지막 편집일시 : 2020년 7월 17일 12:55 오전

댓글 0 피드백

- **이전글** : 2.5 사이트 정보 가져오기 - requests 사용법
- **다음글** : 2.7 사이트 정보 추출하기 - beautifulsoup 사용법 (2)

TOP