2019.11.19.

PYTHON으로

2019.11.19.

<PYTHON으로> 개요

- **Pandas (데이터 분석 도구)**

  실거래가(부동산 가격) 변화 분석

  가계금융복지조사 마이크로데이터 분석

- **Selenium (Web Crawling 도구)**

  마이홈포털 임대료 데이터

마이홈포털 데이터 크롤링

- https://www.myhome.go.kr/hws/portal/sch/selectRentalHouseInfoListView.do

총 **4,460** 건의 검색결과

임대종류	주소	단지명	세대수	주택유형
50년임대	서울특별시 노원구 노원로19길 31	중계4단지	1,979	아파트
50년임대	서울특별시 강남구 양재대로55길 10	수서1단지	1,230	아파트
50년임대	서울특별시 노원구 월계로55길 16	월계사슴2단지	775	아파트
국민임대	서울특별시 마포구 월드컵로42길 12	상암3-8단지(임대)	840	아파트
국민임대	서울특별시 강동구 아리수로93가길 25	강일지구1단지(임대)	11	아파트
장기전세	서울특별시 강동구 아리수로97길 19	강일지구4단지(임대)	71	아파트
장기전세	서울특별시 은평구 진관4로 17	상림마을8-2단지(은평1-1,임대)	25	아파트

마이홈포털 데이터 크롤링

- from ~ import ~

```
In [ ]: from selenium import webdriver
```

webdriver에서 selenium만 불러오기

자동 제어를 위한 크롬 열기

```
In [ ]: driver = webdriver.Chrome('chromedriver')
        driver.get("https://www.myhome.go.kr/hws/portal/sch/selectRentalHouseInfoListView.do")
```

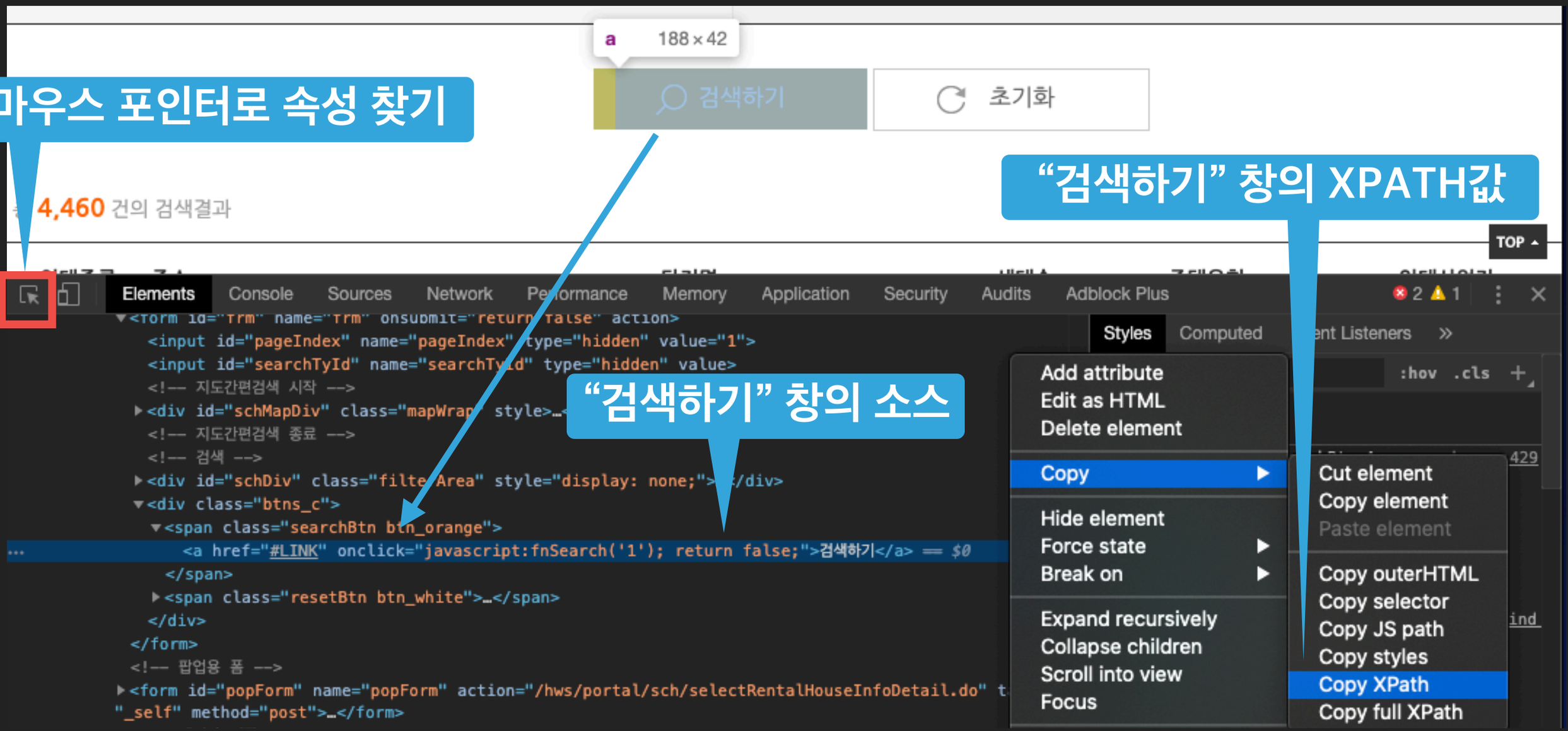
접근할 URL을 “ ” 안에 입력

마이홈포털 데이터 크롤링

- 크롬의 "검사" 기능으로 제어할 소스 찾아내기

마이홈포털 데이터 크롤링

- 웹페이지의 구조 분석 후 selenium으로 제어하기



마이홈포털 데이터 크롤링

- 웹페이지의 구조 분석하는 방법

```
▼<div class="bbs_min">
  <!-- 게시판 목록 -->
  ▼<table class="bbs_type1" cellspacing="0" summary="유형별 찾기 게시판">
    <caption>유형별 찾기 게시판</caption>
    ▶<colgroup>…</colgroup>
    ▶<thead>…</thead>
    ▼<tbody id="schTbody">
      ▼<tr>
        ▶<td>…</td>
        ▶<td class="al">…</td>
        ▼<td class="al">
          <a href="javascript:fnRentalHouseInfoDetail('C','30582262','XXX','03');">중계4단지</a>
        </td>
        <td>1,979</td>
        <td>아파트</td>
        ▶<td>…</td>
        ▶<td>…</td>
      </tr>
```

<태그명> … </태그명>

들여쓰기 차이

<TD>는 <A>의 PARENT

웹페이지의 구조:
서로 위상이 다른 여러 태그의 결합

마이홈포털 데이터 크롤링

- .find_element_by_ 로 원하는 요소에 접근하기

```
driver.find_element_by_link_text(임대유형).click()
```

링크가 연결된 텍스트

```
driver.find_element_by_xpath("""//*[@id="frm"]/div[3]/span[1]/a""").click()
```

XPATH값

```
driver.find_element_by_tag_name('select')
```

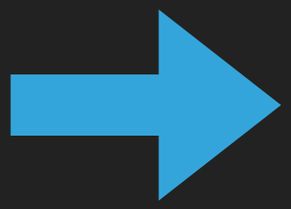
태그명

원하는 요소를 특정하기 위해서
웹페이지 구조에서 희소성 있는 값을 찾아야 함

마이홈포털 데이터 크롤링

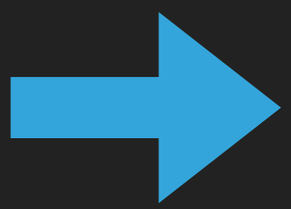
- soup.find() 로 원하는 요소에 접근하기
  soup.find_all() 로 원하는 '모든' 요소에 접근하기

```
soup.find('select')
```



태그명이 select인 (첫번째) 요소 반환

```
soup2.find_all('th')
```



태그명이 th인 모든 요소를 [목록]으로 반환

마이홈포털 데이터 크롤링

- .join(DF, how = ‘outer’)로 길이가 다른 목록 결합하기
  .fillna(method=‘ffill’)로 NaN값 채우기

임대종류	주소	단지명	세대수	주택유형	임대사업자
행복주택	서울특별시 구로구 천왕로 21	천왕이펜하우스 7	374	아파트	SH공사

형(Type)	공급면적(전용)	공급면적(공용)	임대보증금	임대료	전환보증금
29	29.96㎡	16.75㎡	46,400,000원	213,000원	0원
29S	29.96㎡	16.75㎡	42,400,000원	212,000원	0원

단일한 단지의 복수의 임대료 정보를 저장

마이홈포털 데이터 크롤링

- **def** 함수명(입력변수):　로 복잡한 작업을 함수로 정리하기
  return(출력변수)

입력변수

```
def 페이지넘김(driver):
    소스 = driver.page_source
    soup = BeautifulSoup(소스, 'html.parser')
    페이지수 = soup.find('div', {'id': 'pageDiv'}).find_all('li')
    통합 = pd.DataFrame()
    for 페이지 in 페이지수:
        driver.find_element_by_xpath("""//*[@id="pageDiv"]/ul/li
        time.sleep(4)
        임시저장 = 단지정보(driver)
        통합 = pd.concat([통합, 임시저장], sort=False)
    return(통합)
```

출력변수

‘driver’를 입력해 ‘통합’을 출력함

웹 크롤링에서 유의할 점

- 컴퓨터 사양
- 인터넷 환경
- 운영체제와 브라우저 간의 호환
- 웹사이트의 반응속도
- 화면의 크기
- **tqdm_notebook()**으로 진행 과정 모니터링하기

웹 크롤링 사용 예시

- 한국감정원 주간가격동향조사
- 렌트홈(등록임대주택 정보)
- 어린이집 유치원 통합정보공시
- 네이버 뉴스검색
- 내한공연 정보
- 지역별 블루리본 맛집

https://www.aladin.co.kr/shop/wproduct.aspx?ItemId=126093708

감사합니다

hong@pspd.org