구글 크롤러 사용

크롤링

참고 전용

크롤링

인터넷에 존재하는 다량의 정보를 분석,활용을 위해 수집하는 행위

- 크롤러 : 크롤링 하는 프로그램
- 스크래핑 : 웹 사이트 상에서 내가 원하는 정보 추출

- 크롤링시 알아야 할 것
  - 크롤링이 가능한 웹사이트인지 확인 : robots.txt
  - 웹 구조 : js , html , css
  - 전처리 : numpy, pandas , 정규표현식 (import re) 등

라이브러리

SELENIUM

웹 페이지를 자동화하여 크롤링하는 도구

chromdriver 사용
동적 크롤링

BEAUTIFUL SOUP

HTML, XML 등의 마크업 언어를 파싱하기 위한 라이브러리

정적 크롤링

REQUESTS

HTTP 요청을 보내고 받는 라이브러리

PYQUERY

jQuery 문법을 사용하여 HTML을 파싱할 수 있는 라이브러리

SCRAPY

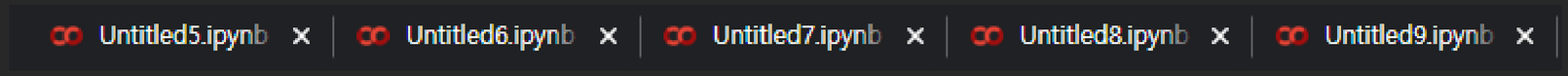
크롤링과 스크래핑에 사용되는 웹 프레임워크

고마워요 GPT

Anaconda , Jupyter 설치

코랩에서는 구글 드라이버 버전 문제로 귀찮음

-> 하다가 화남



셀레니움이 버전 4로 업그레이드 되면서 구글드라이버 자동 설치

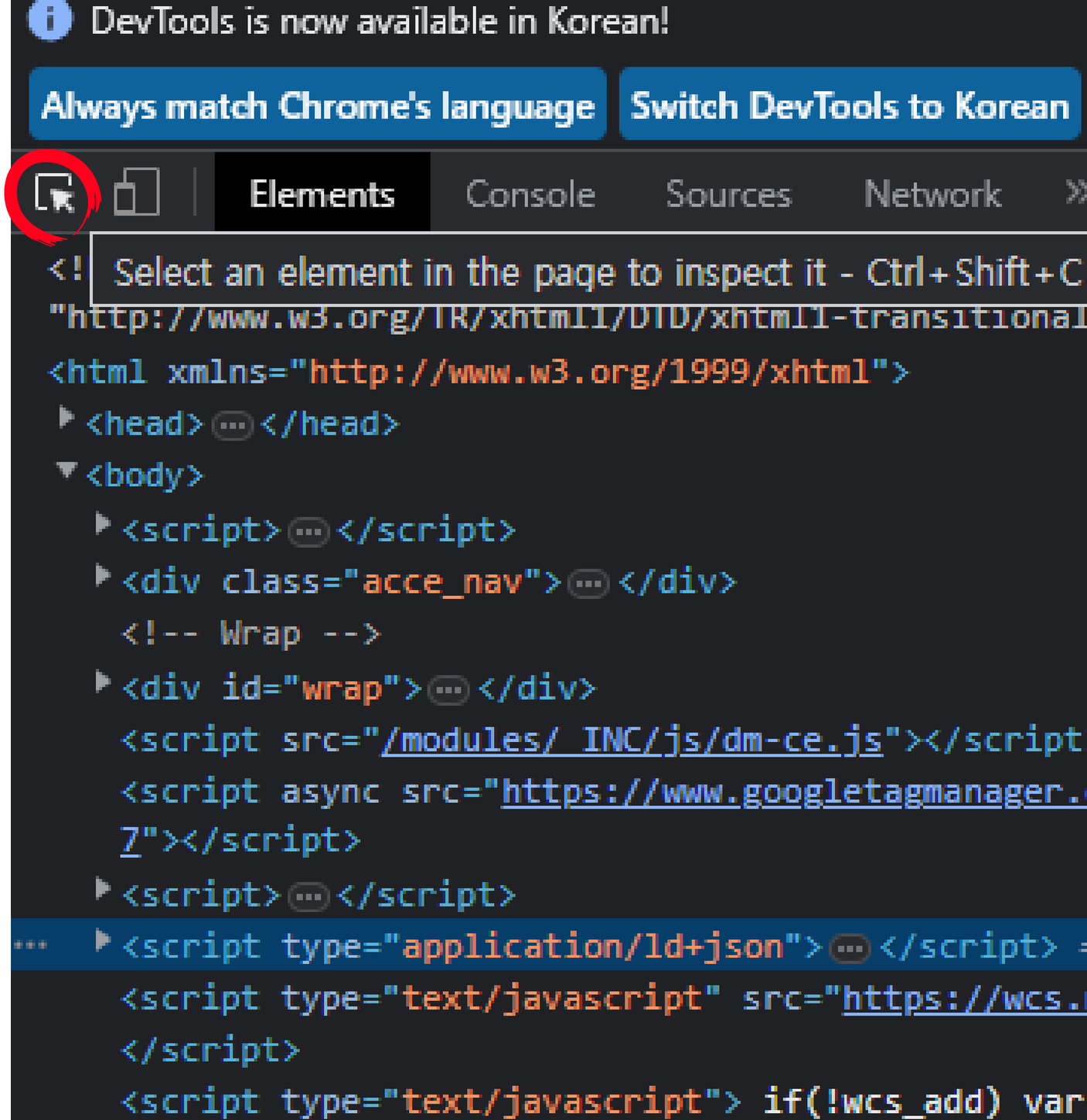
설치 방법 : Anaconda 검색 후 설치 -> Anaconda Navigator 실행 -> Jupyter 설치

다음, 코드로 진행

- 크롤링에 너무 많은 시간 쓰지 말 것 : 별로 중요하지 않다
- 지금은 앞에서 명시한 '크롤링시 알아야 할 것'은 알고 진행하지 않음
  -> 야매에 가깝다
- 인터넷에 크롤링 관련 글이 정말 많음

DevTools

개발자 툴
웹페이지의 html, css, js를 확인할 수 있다



스크래핑은 주소다

배달시 상대방의 주소를 알아야한다.
경기도 > 용인시 > 수지구 > 죽전로 152 > IT관 203
해당 주소로 배달원이 찾아온다.

스크래핑시 해당 데이터의 주소를 알아야 한다.
id > div > div > class > div (예시)
해당 주소에 저장된 값을 가져온다.

find_element

하나만 가져온다.

주요 뉴스

SK그룹 5개사, 문화 후원으로 '글로벌 스토리' 만든다

find_elements

리스트 형식으로 여러개 혹은 전부를 가져온다.

find_elements는 text로 출력시
find_elements[1]처럼 인자 지정을 해주자

주요 뉴스

SK그룹 5개사, 문화 후원으로 '글로벌 스토리' 만든다

'코인 논란' 김남국, 민주당 탈당… "무소속으로 공세 맞...

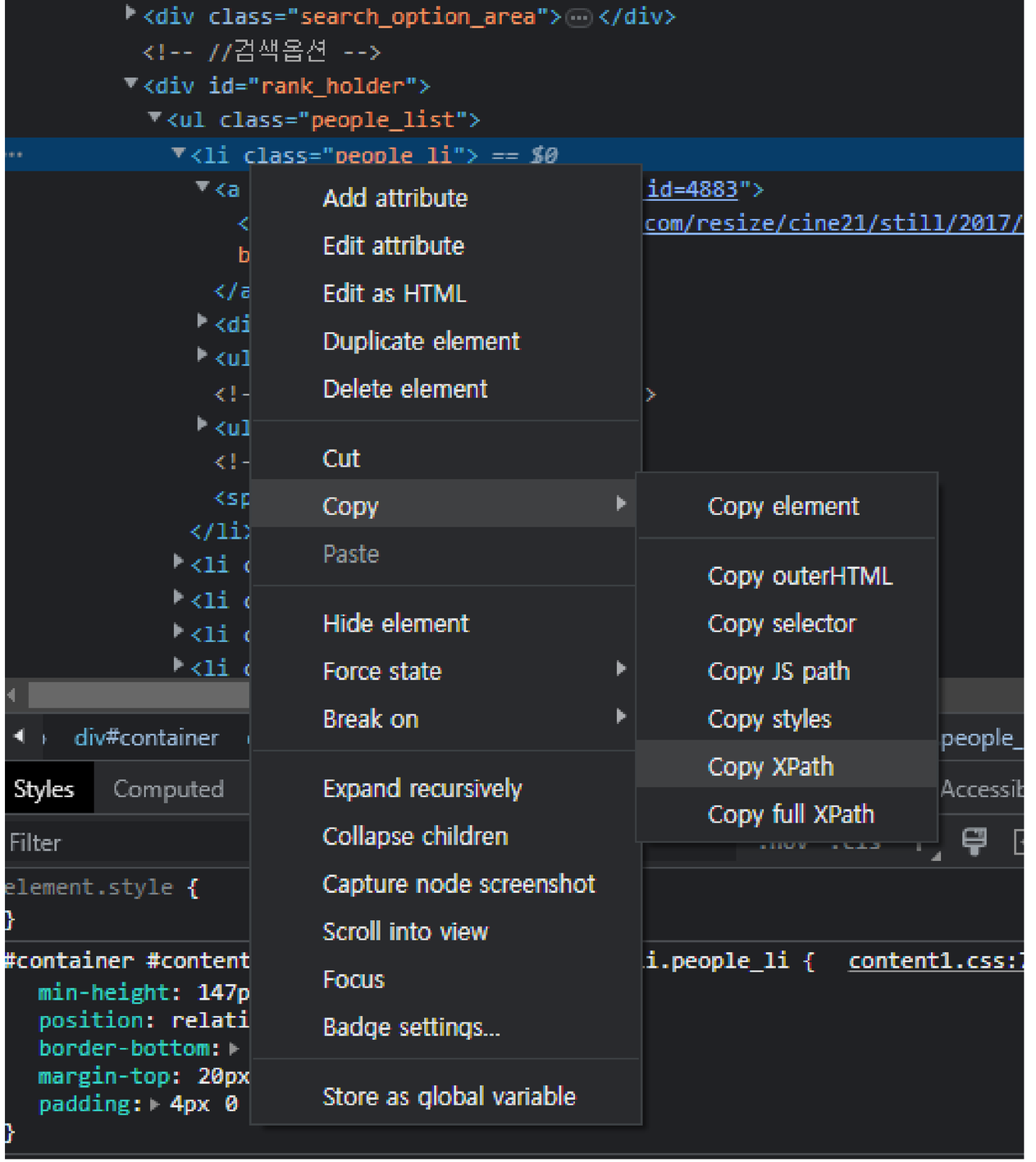
KT, 종합 헬스케어 사업 본격화…베트남서 비대면 케어 ...

김동연, "오월을 그리고 광주를 기억하고 또 기억하겠습...

김남국 "에어드롭으로 코인 무상 지급"…與 "코인계의 ...

XPATH

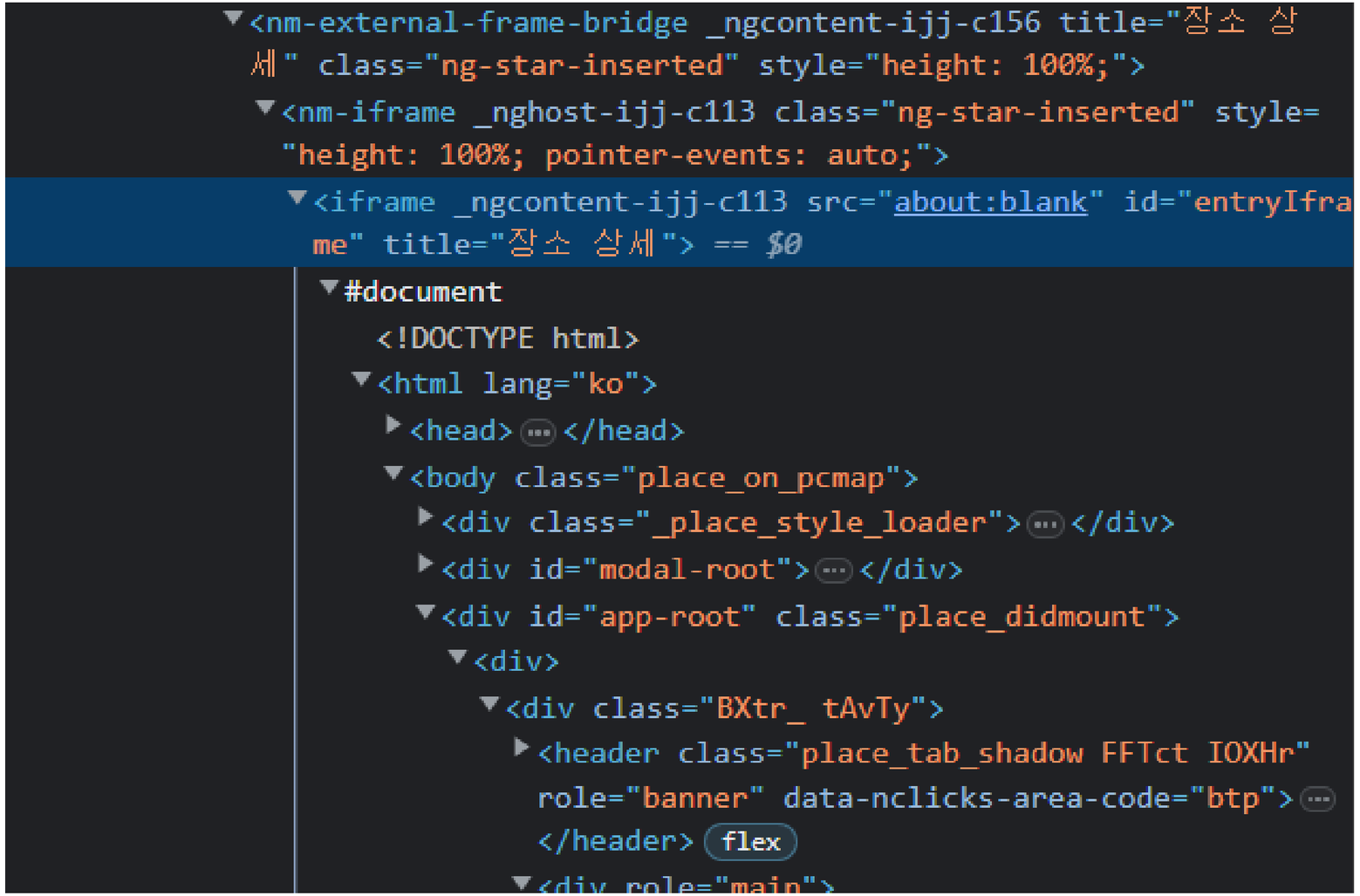
정말 편한 녀석이다



다중 html 구조에서는 iframe을 찾자

driver.switch_to.frame('entryIframe')

Iframe의 id가 entryIframe인 html에서 서칭 시작



끝