

# 파이썬으로 웹 크롤링 하기

이기은

# 들어가기 전에

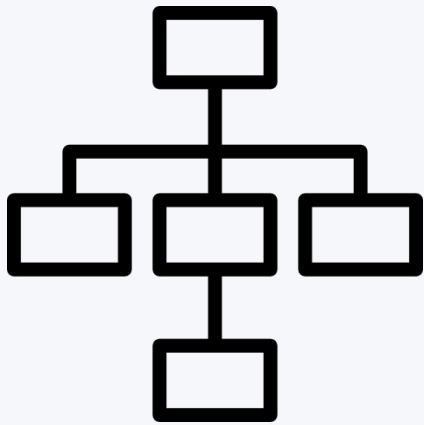
파이썬, 웹 관련 코딩을 모두 모른다는 가정 하에 진행  
-> **어려운 것이 당연하다!**

1차 고비: 파이썬 환경, 라이브러리 설치  
2차 고비: 웹 구조 이해  
3차 고비: for문 사용

**코드는 원리이기 때문에, 흐름을 따라온다고 생각해주세요**

**파이썬은 0부터 세요**

## 웹 크롤링을 위해서 알아야하는 것 : 웹은 어떻게 표출되는가?



HTML

뼈대



CSS

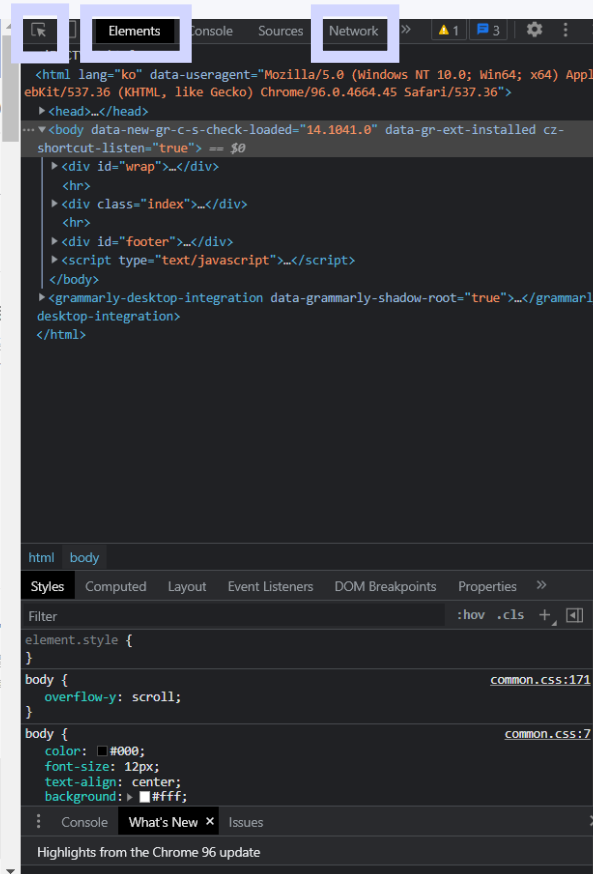
꾸미기



JS

애니메이션

# 웹 크롤링을 위해서 알아야하는 것 : 냅다 개발자도구 켜기



중요한 건

Elements / Network

맥 크롬 기준, Alt(Option) + Command + i)  
윈도우 크롬 기준, F12 키 누르기

스크래핑 하고싶은 사이트를 찾으면  
먼저 개발자 도구를 켜보자

## 웹 크롤링을 위해서 알아야하는 것 : GET vs POST

### GET

페이지를 바꾸면 url이 바뀌는 방식

[https://gall.dcinside.com/board/lists/?id=exam\\_new2&page=2](https://gall.dcinside.com/board/lists/?id=exam_new2&page=2)

[https://gall.dcinside.com/board/lists/?id=exam\\_new2&page=3](https://gall.dcinside.com/board/lists/?id=exam_new2&page=3)

...

[https://gall.dcinside.com/board/lists/?id=exam\\_new2&page=50](https://gall.dcinside.com/board/lists/?id=exam_new2&page=50)

### POST

페이지를 바꿔도 url이 고정되어 있는 방식

웹 크롤링을 위해서 알아야하는 것  
: GET vs POST

GET

페이지를 바꾸면 url이 바뀌는 방식

Ajax 와 깊은 연관!  
(Network를 뜯어보자)

POST

페이지를 바꿔도 url이 고정되어 있는 방식

## 파이썬으로 웹 크롤링은 어떻게?



### Requests + BeautifulSoup

url을 요청해 html로 파싱한 후, 원하는 태그 속 내용만을 스크래핑해오는 것

### Selenium

자바스크립트 기반의 동적으로 작동하는 웹  
(ex. 팝업창을 꺼줘야하는 경우, 버튼을 클릭해야하는 경우, 로그인 필요한 경우)

## 파이썬으로 웹 크롤링은 어떻게?



### Requests + BeautifulSoup

url을 요청해 html로 파싱한 후, 원하는 태그 속 내용만을 스크래핑해오는 것

### Selenium

자바스크립트 기반의 동적으로 작동하는 웹  
(ex. 팝업창을 꺼줘야하는 경우, 버튼을 클릭해야하는 경우, 로그인 필요한 경우)



# 오늘 크롤링할 사이트는?



<https://www.assembly.go.kr/assm/memact/congressman/memCond/memCond.do>