# (가제) 나만의 웹 크롤러 만들기

패스트 캠퍼스 4주 커리큘럼 (주2회 총 8회차)

기본 기획:
웹 크롤링에 필요한 파이썬의 기본을 알아보고 웹 크롤러를 만들기 위한 전반적인 개념과 실제 사례에 적용을 통해 웹 크롤링 기법을 익혀봅니다.

W1 D1 (2.5H기준)
Summery: 8주간 진행할 전반적인 내용에 대해 알아봅니다. 웹 크롤러에 대해 알아보고 파이썬 기본문법을 알아봅니다.

- 우리가 할 것(결과물 예시)
- 우리가 사용하는 것: 크롬브라우저, 파이썬, 그리고 크롤링 도구들
- 크롤링 프로그래밍 환경 잡기: 크롬, 파이썬, 도구(라이브러리)들 설치하기
- 파이썬의 기본: 변수, 자료형, 객체
- 파이썬 반복문 이용해보기

W1 D2
Summery: 파이썬의 문법을 조금 더 알아보고 크롤러에 필요한 라이브러리들의 사용법을 알아봅니다.

- 파이썬의 기본: 함수, 라이브러리 import해보기
- 웹은 어떻게 구성되어있나?: 서버와 클라이언트, 네트워크가 이루어지는 방법

W2 D1
Summery: 본격적으로 웹 크롤링을 시작해 봅니다. HTML의 구조를 이해하고 Requests를 설치해 기본적인 크롤링을 진행해 봅니다.

- 웹 사이트는 어떻게 만들어져있나?: HTML + CSS + JS 간단한 소개
- HTML의 구조
- CSS Selector란?
- 크롤링 실습:
    - 네이버 홈페이지의 구조 뜯어보기
    - 네이버 실시간 검색어 크롤링하기

W2 D2
Summery: 크롤링을 도와주는 크롬 사용법, 로그인해서 크롤링하기

- CSS 속성 조금 더 알아보기
- CSS Selector 연습해보기
- 크롤링 실습1:
    - 온오프믹스(IT행사 신청 사이트) 구조 뜯어보기
    - 온오프믹스 검색 결과 크롤링하기
- 웹에서 로그인이 어떻게 이루어지는지 알아보기
- 로그인이 유지되는 방법들 알아보기
- 크롤링 실습2:
    - 뽐뿌 로그인 구조 뜯어보기
    - 뽐뿌 로그인 하고 장터 크롤링하기

W3 D1
Summery: 진짜 크롬으로 크롤링하는 무적 크롤러를 만들어봅니다.

- Selenium이란?
- Selenium 설치 (pip로 설치 / selenium chrome driver 받기)
- 크롤링 실습:
    - Selenium 사용법 익히기 (브라우저 이동, 클릭, 입력 / 웹 페이지 Element가져오기)
    - 페이스북 로그인하기
    - 페이스북 타임라인 크롤링하기

W3 D2
Summery: PhantomJS와 Headless Chrome으로 CLI에서 크롤링 코드를 실행합니다.

- PhantomJS 설치하기
- Chrome을 Headless하게 이용하기
- 실습: 기존 Selenium이용한 코드들을 Headless 환경에서 실행하기
    - 트위터 로그인하기
    - 트위터 해쉬태그 검색결과 크롤링하기

W4 D1
Summery: 크롤링 프로그램을 원격 서버에 올려서 주기적으로 크롤링 하는 방법을 알아봅니다.

- VPS란?
- 원격 서버에 코드를 올려 동작시키는 방법
- Crontab이란?
- 시간 주기에 맞춰 크롤링 프로그램을 실행시키는 방법 알아보기
- 실습:
    - VPS에 크롤링 프로그램 올려서 실행해보기
    - 주기적으로 크롤링하도록 설정해보기

W4 D2(완결)

Summery: re(정규표현식)알아보기, 크롤링할 때 주의사항

- 정규표현식이란?
- 원하는 데이터 추출하기(re.pattern)
- 실습1:
    - regexr.com에서 정규표현식 연습하기
    - 크롤링한 HTML에서 정규표현식으로 단어들 추출하기
- 크롤링의 법적 문제와 한계
- 크롤러를 좀더 사람처럼 보이게 만들기
- 실습2:
    - 진짜 브라우저처럼 만들기
    - 랜덤하게 쉬면서 크롤링하기