데이터사이언스를위한 지식관리시스템

4주차: 웹스크래핑

4주차 목표

- 파일형태로도, OpenAPI로도 배포되지 않는 데이터의 수집 방법입니다
- 단. 허가되지 않은 웹스크래핑을 하지 않도록 주의해야 합니다

우리가 웹사이트를 열면 일어나는 일

- 웹페이지를 요청합니다
- 웹페이지의 응답이 옵니다
- 이를 렌더링 합니다
- 그 과정에서 js(javascript)와 같은 스크립트가 실행됩니다
- 웹페이지의 응답이 오는것 자체는 정적이지만 이후의 동적인 변화가 있을 수 있습니다

정적 스크래핑

- 웹페이지의 응답을 받아서 데이터를 즉시 얻을 수 있을 때 수행합니다
- 응답 내에 우리가 원하는 데이터가 있으며 이를 파싱하여 데이터를 얻을 수 있습니다

동적 스크래핑

- 응답 내에 우리가 원하는 데이터 대신에 스크립트가 있으며 이를 실행하면서 데이터가 보이게 됩니다
- 웹페이지의 응답을 받은 이후 추가적인 스크립트의 실행에 의한 변화를 모니터링 하면서 데이터를 수집합니다
- 필요시 추가적인 웹페이지 핸들링도 필요합니다

정적 스크래핑과 동적 스크래핑 비교

구분	정적 스크래핑	동적 스크래핑
데이터 위치	초기 **HTML 응답** 안에 이미 존재	초기 HTML에는 없음 **Javascript 실행 후** 추가 로딩
도구	`BeautifulSoup`	`Selenium`
속도/성능	빠르고 가벼움	상대적으로 느리고 무거움
난이도	비교적 쉬움	더 복잡 (JS 실행, DOM 변화 감시, 페이지 조작 필요)
예시 상황	블로그 글	무한 스크롤 쇼핑몰
핵심 포인트	응답만 보면 된다	응답 이후 변화를 봐야 한다