X

툴 스터디 2주차 – 웹 스크래핑

정은진

목차

웹 스크래핑이란?

웹 스크래핑 (Web Scraping): 웹 사이트 상에서 '원하는 정보'를 추출하여 수집하는 것

웹 크롤링 (Web Crawling): 자동화된 봇 (bot)을 이용해 웹 페이지 상에
허용된 범위 내에 존재하는 '모든 데이터'를 마구잡이로 수집하는 것

웹 스크래핑의 목적

- **웹 스크래핑을 통해서 무엇을 할 수 있나요?**

  뉴스 기사, SNS (instagram, facebook, blog) 콘텐츠, 주식 시장 데이터, 비즈니스 트렌드 등… 추출 가능

- **우리가 웹 스크래핑을 공부하는 이유는 무엇인가요?**

  우리가 얻고자 하는 데이터가 있다면, 수많은 데이터가 존재하는 웹 페이지 상에서
  우리에게 필요한, 우리의 목적에 부합하는 데이터들을 골라 수집해야 합니다.
  이렇게 골라 모은 데이터들을 토대로 우리는 인사이트(insight)를 도출할 수 있습니다.

웹 스크래핑 환경 구축

1) Chrome 브라우저
2) Python
3) VS Code
4) VS Code 확장 프로그램에 Python 설치
5) 파이썬 라이브러리 (requests, BeautifulSoup4, Selenium)
6) Chrome Web Driver

라이브러리 소개

1) requests: 웹 페이지에 접근하여 웹 페이지 정보를 받아오는 라이브러리입니다.

2) Beautiful Soup: 웹 사이트의 html 파일을 읽어오거나, requests 라이브러리를 통해 웹 소스를 가져옵니다.

3) Selenium: 웹 드라이버를 사용해 웹 페이지를 동적으로 스크래핑할 수 있게 하는 라이브러리입니다.

html이란

- HTML (Hyper Text Markup Language)
  웹 페이지를 만들기 위한 언어입니다.
  태그와 태그로 연결되어 웹 페이지를 구성하고, 조직화합니다.

- 브라우저에 들어가서 아무 웹 사이트나 접속한 후, f12 (개발자 도구)를 누르면 해당 웹 사이트의 html 파일을 확인할 수 있습니다.

html이란

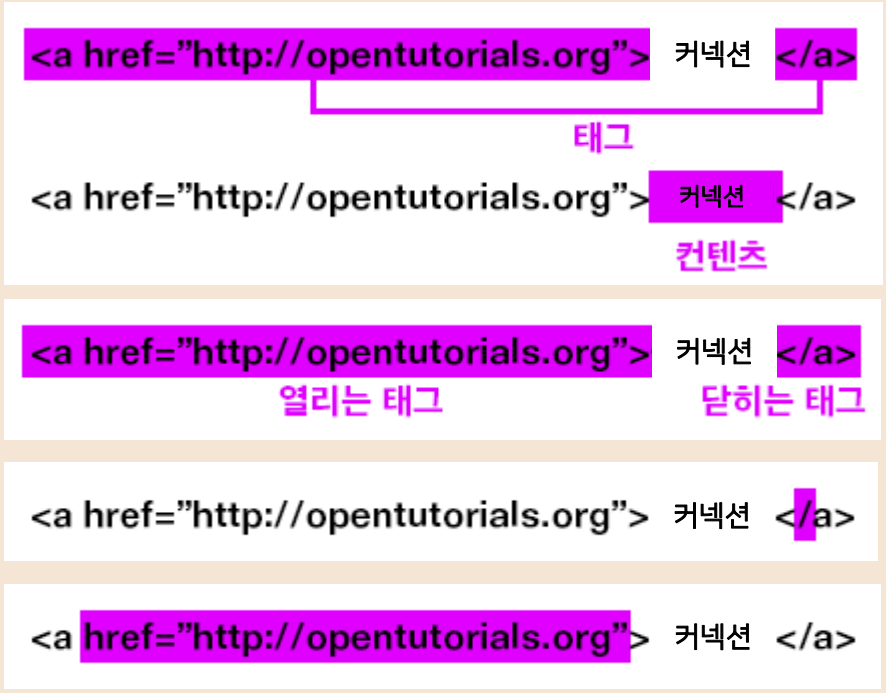
- HTML 문서의 파일 확장자는 '.html' 혹은 '.htm' 으로 끝납니다.
- 최상위 태그는 <html>이며, 하위에 <head>와 <body> 태그를 가지고 있습니다.
- <head> 태그: 문서를 설명하는 태그, 제목, 혹은 키워드와 같은 정보를 담고 있음
- <boay> 태그: 문서의 내용이 담겨 있는 태그

```
1  <html>
2      <head>
3          문서를 정의하는 데이터가 위치함
4      </head>
5      <body>
6          문서에 표시되는 컨텐츠가 위치함
7      </body>
8  </html>
```

html이란

<'태그명' '속성명1'="속성값1" '속성명2'="속성값2"> 컨텐츠 </태그명>

'태그'는 컨텐츠를 감싸 그 컨텐츠의 성격과 의미를 정의합니다.



열리는 태그가 있으면 닫히는 태그가 있어야 하며,

닫히는 태그의 태그 명 앞에는 '/'가 붙습니다.

속성은 태그의 부가적인 정보가 담겨있습니다.

예제에서 'href'는 속성명, 'http://opentutorials.org'은 속성값입니다.

href 속성은 컨텐츠인 '커넥션'이 opentutorials.org와 연결되어 있다는 것을 의미합니다.

실습 start!

- Beautiful Soup

  실습 1. 오늘의 1위 인기 프로그램 정보 추출하며 html 구조 이해하기

  실습 2. 오늘의 국내 넷플릭스 인기 top 10 프로그램 리스트를 추출해 csv 파일에 저장하기

- Selenium

  실습 3. 동적 페이지 구글 무비 사이트에서 인기 영화 리스트 추출하기

웹 스크래핑을 통해 완성도 높은 프로젝트를 진행해봅시당

프로젝트 1팀 '딥 페이크': 딥 페이크 관련 통계, 뉴스 등

프로젝트 3팀 '플랫폼 노동': 노동 플랫폼 사이트 내 콘텐츠, 플랫폼 노동 관련 통계 등

프로젝트 4팀 '탐닉과 쾌락': 금융 트렌드, 네이버 주식 지수 등

프로젝트 5팀 '환경': 환경 관련 뉴스, 다큐 콘텐츠, 통계 등

※웹 스크래핑 시 주의해야 해요※

- 무분별한 웹 크롤링 / 웹 스크래핑은 대상 서버에 부하를 주게 되고, 계정이나 IP가 차단될 수 있습니다.

- 스크래핑을 통해 얻은 이미지나 텍스트 등의 데이터를 무단으로 활용할 시 저작권이 침해될 우려가 있고, 법적인 제재를 받을 수 있습니다.

웹 스크래핑을 더 자세히 공부해보고 싶다면?

나도코딩 유튜브 웹 스크래핑 강의 참고

https://www.youtube.com/watch?v=yQ20jZwDjTE

Thank You

Q & A

궁금한 점이 있으시다면 말씀해 주세요 😊