

Web crawling with selenium

셀레니움, 뷰티플수프 등을 활용한 텍스트 자료 크롤링



목차

- selenium 소개 및 컨트롤 방법
- beautifulsoup 소개 및 기본적 명령어
- 환경설정, 라이브러리 설치 및 크롤링 루틴
- 텍스트 크롤링에 대한 고급 루틴

Selenium이란?

- Selenium은 웹 테스트를 목적으로 사용되는 라이브러리
- 물리 드라이버를 활용하기 때문에 실제로 사용자가 사이트를 방문하는 것처럼 구현할 수 있음.
- Chrome, firefox 등의 물리 드라이버가 존재하며 본 교안에서는 크롬 드라이버를 사용함.

Selenium 기본 명령어

```
driver = webdriver.Chrome('크롬드라이버경로/chromedriver.exe')
```

가장 기본이 되는 명령어로, 물리 드라이버 가동 명령.

```
driver.get("사이트명")
```

물리드라이버를 사이트로 이동시킴

```
driver.find_element_by_xpath("xpath")
```

xpath를 이용해서 요소를 특정짓기.

```
username = driver.find_element_by_xpath("xpath")  
username.send_keys("입력할 문자열")
```

로그인 창 등 입력창에 문자열 입력하기

Selenium 기본 명령어

`driver.find_element_by_xpath("xpath명").click()`
특정 요소를 찾아서 그 요소 클릭하기.

`driver.page_source`
현재 페이지의 소스코드를 전부 긁어옴

Xpath 추출하는 방법

크롬드라이버에서 F12를 누르면 요소분석 창이 나타남.

요소분석창 왼쪽 위 버튼을 눌러 사용할 구역을 찾아낸 다음

클릭하면 요소분석창에서 그 부분의 소스가 선택됨

선택된 부분 우클릭->copy->copy xpath를 클릭해 복사 가능.

Beautifulsoup란?

뷰티풀수프는 데이터를 수집해오는 기능과는 전혀 상관 없는 라이브러리이다.

하지만 수집해온 텍스트 데이터를 처리하는데 있어서 강력한 기능을 제공하기 때문에 데이터 가공에 많이 활용한다.

특히 본 교안에서는 웹에서 자료를 가져올때는 html.parser 기능을 이용해 Html 구조로 작성된 문서를 처리하는 기능을 활용하게 된다.

Beautifulsoup 기본 명령어

Beautifulsoup(소스코드, "html.parser")
가져온 소스코드를 html로 파싱하는것.

파싱이란 비유하자면 두꺼운 책을 가지고만 있느냐 읽어서 이해하느냐의 차이이다.

파싱을 해야만 컴퓨터가 코드의 구조를 파악할 수 있다.

Beautifulsoup 기본 명령어

파싱한자료.find_all("태그명", id="아이디명", class_="클래스명")

파싱한 자료 내부에서 원하는 태그명 및 아이디, 클래스에 해당하는 자료만 선별하는 코드.

이 코드를 활용해 원하는 텍스트자료가 있는 부분을 선택적으로 수집할 수 있게 된다.

만약 두 개 이상의 요소를 동일하게 수집하고 싶다면 리스트 처리해 find_all(["태그명1", "태그명2"...]) 과 같이 처리한다.

Beautifulsoup 기본 명령어

.text 명령어

가져온 코드에서 태그(<태그명> </태그명>) 를 제외하고 외부에 있는 자료만 보여주는 명령어.

이 기능을 이용하면 손쉽게 텍스트부분만 떼어낼 수 있다.

다만 a태그 내부의 링크()이 같이 소거되므로 링크를 가져올때는 문자열 슬라이싱을 활용해야 한다.

환경설정, 라이브러리 설치

환경설정은 크게 두 가지로 나뉩니다.

첫번째는 크롬드라이버 세팅으로 제가 제공하는 링크인

<https://drive.google.com/open?id=18mtOiYaPj5kBXzfPxGMeQnXnN3FLAfad>

에서 크롬드라이버를 다운받은 다음 원하는 폴더에 넣습니다.

이후 `driver = webdriver.Chrome('크롬드라이버경로/chromedriver.exe')`
와 같이 물리드라이버 코드를 세팅해주면 완료됩니다.

환경설정, 라이브러리 설치

두 번째 환경설정은 라이브러리 설치입니다.

Cmd창 실행 ->

pip install beautifulsoup4 입력 ->

설치 완료 후 pip install selenium 입력

이후 파이썬 코드 첫번째줄, 두번째줄에

```
from selenium import webdriver
```

```
from bs4 import BeautifulSoup
```

를 입력하면 각 라이브러리 사용이 가능해집니다.

설치가 안 될 경우는 제가 드린 파이썬 설치교안을 보고 다시 절차대로 설치해주세요.

크롤링의 기본적인 원리

기본적인 루틴은 다음과 같습니다.

셀레니움을 이용해 크롤링할 페이지 접근

전체 소스코드 수집

파싱 후 원하는 부분에 가깝게 소스코드 수집

수집된 소스코드 다시 가공

가공된 코드 .text를 이용해 태그 제거

정리한 텍스트자료 txt파일로 export 시키기