Chapter 11

웹 데이터 수집(크롤링)

셀레니움

- 셀레니움: 웹 자동화 및 웹의 소스코드를 수집하는 모듈

- cmd -> pip install selenium (셀레늄 라이브러리 다운로드)

기능	설명
driver = webdriver.Chrome(경로)	크롬 드라이버 준비
driver.get(웹페이지주소)	크롬 드라이버 페이지 이동명령
driver.find_element_by_xpath(xpath값) -> click() -> send_keys()	Xpath로 태그찾기
driver.back()	뒤로가기
…이하생략	

뷰티풀숩

- 뷰티풀솝(Beautiful Soup)은 스크린 스크래핑(screen-scraping) 프로젝트를 위해 설계된 파이썬 라이브러리

- 구문 분석, 트리 탐색, 검색 및 수정을 위한 몇 가지 간단한 방법과 파이썬 관용구를 제공하며 문서를 분석하고 필요한 것을 추출하는 도구

- 들어오는 문서를 유니코드로 보내고 문서를 UTF-8로 자동 변환

공식 사이트
https://www.crummy.com/software/BeautifulSoup/
Documentation
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

뷰티풀숩

주요 함수	설명
BeautifulSoup(src, 파싱타입);	셀레니움이 가지고 온 html소스코드를 변환
find('태그명', {'속성명': '값' ...})	조건에 맞는 하나 태그를 가져온다
find_all('태그명', {'속성명': '값' ...})	조건에 맞는 모든 태그를 리스트로 가져온다
Select(선택자)	Css선택자로 태그를 가져온다
...이하생략	

예시

```
1  from bs4 import BeautifulSoup
```

```
1  soup = BeautifulSoup(res.content, "html.parser")
```

```
1  el = soup.select_one("h1")
2  print(el)
```

```
<h1>Hello CSS</h1>
```

뷰티풀숩의 선택함수들로 태그를 선택한다

뷰티풀숍

Css 선택자란?

CSS 선택자

CSS를 이용해서 HTML 문서에 디자인 할 때 사용하는 것
CSS 선택자(Selector)는 HTML 문서의 태그 이름, class 속성, id 속성 등을 이용해서 작성할 수 있음

```
1  <html>
2  <head><title>HTML Sample</title>
3  </head>
4  <body>
5      <h1>Hello CSS</h1>
6      <div id="subject">선택자</div>
7      <div class="contents">선택자를 어떻게 작성하느냐에 따라
   <span>다른 <b>요소가 반환</b></span> 됩니다.</div>
8      <div>CSS 선택자는 다양한 곳에서 <b>활용</b>됩니다.</div>
9  </body>
10 </html>
```

Id는 #
Class는 .
을 사용하여 태그를 지목할 수 있음

선택자 예시

태그 선택자 ("element")

태그 선택자는 일반적으로 스타일 정의하고 싶은 html 태그 이름을 사용
요소 안의 텍스트는 text, 태그이름은 name 그리고 태그의 속성들은 attrs를 이용해 조회

```
soup.select("h1")
```

다중(그룹) 선택자 ("selector1, selector2")

선택자를 ","(comma)로 분리하여 선언하면 여러 개 선택자 적용.
해당하는 모든 선택자의 요소를 찾기 때문에 select_one()아닌 select() 메서드를 이용

```
soup.select("h1, p")
```

내포 선택자 ("ancestor descendant")

요소가 내포 관계가 있을 때 적용시키기 위한 선택자. 선택자와 선택자 사이를 공백으로 띄우고 나열

```
soup.select("div b")
```

자식 선택자 ("parent > child")

선택자와 선택자 사이에 >를 입력하며 반드시 부모자식간의 관계에만 스타일이 적용되도록 함.
두 단계 이상 건너뛴 관계에서는 자식 선택자가 작동하지 않음

```
soup.select("div > b")
```

클래스(class) 선택자 (".class")

HTML 문서에서 class 속성의 값과 일치하는 요소를 선택
선택자 이름 앞에 "."을 이용하여 선언.

```
soup.select("div.contents")
```

아이디(id) 선택자 (" #id ")

HTML 문서에서 id 속성의 값과 일치하는 요소를 선택
id 선택자는 #으로 정의합니다.

```
soup.select_one("#subject")
```

속성 선택자 [name="value"]

특정한 속성을 갖는 요소만 선택. 속성 선택자는 [와]사이에 속성의 이름과 값을 지정

```
soup.select_one("[id=subject]")
```

뷰티풀숩

예제 and 실습
1. 도서 데이터 수집
2. 환율 데이터 수집

Chapter 11

수고하셨습니다