04.웹크롤링

1 웹크롤링

1.1 BeautifulSoup

1) HTML 파싱 라이브러리로 bs4버젼부터 Python 3.x 지원

2) 특징

- 단순한 몇 개의 메소드를 가지고 웹 페이지의 내용 추출이 가능(DOM 탐색이 가능하다)
- HTML 뿐만 아니라 XML도 지원한다.
- UTF-8 형식이 기본이지만 CP949 엔코딩도 지원한다.

3) 설치

```
> > pip install beautifulsoup4
```

4) 테스트 하기

1. tag 조회
2. 속성값
3. Attributes 조회

1.2 Selenium

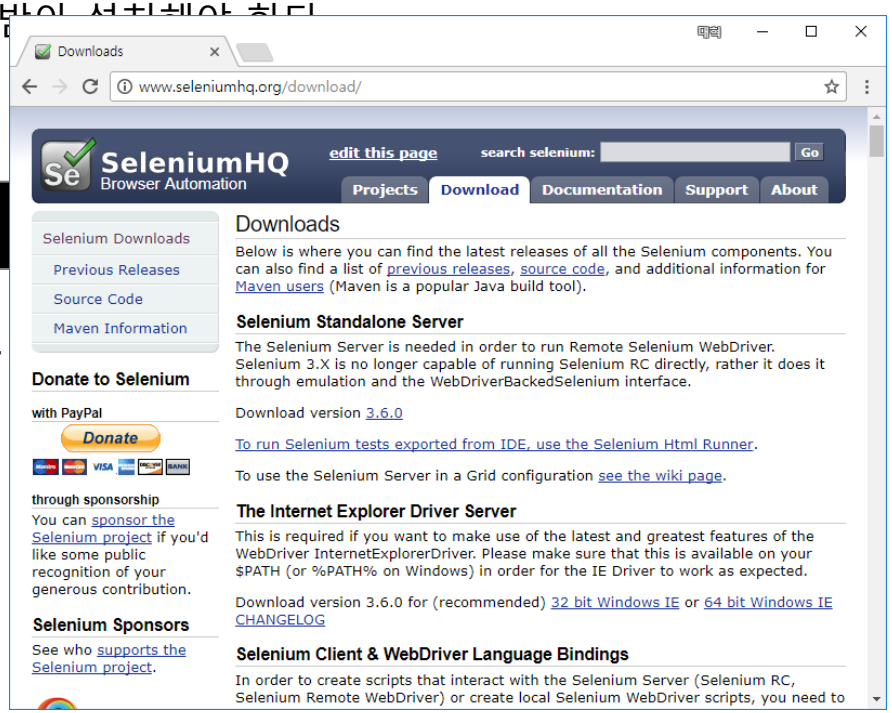
- 원래 웹 사이트 UI(화면) 테스트 목적으로 제작된 라이브러리
- 테스트 절차에 따라 개발한 웹 화면들을 브라우저에서 직접 동작시키고 화면을 캡쳐하여 확인하는 테스트 용도로 활용된다.
- 브라우저에서 동작 시킬 때 자바스크립트 코드 동작도 가능하기 때문에 동적 웹페이지 테스트에 필수 도구 이다.
- Selenium 파이썬 라이브러리 자체는 브라우저를 포함하고 있지 않다.
- WebDriver 라는 인터페이스와 함께 구동된다.
- WebDriver는 운영체제 및 브라우저에 맞게 다운로드 받아 설치해야 한다.

1) 설치

```
> > pip install selenium
```

설치가 완료되면 사용할 브라우저에 맞는 WebDriver 다운로드 한다.

http://www.seleniumhq.org/download 로 이동



2) 테스트

```
from selenium import webdriver


wd =
webdriver.Chrome('D:₩javaStudy₩webDriver₩chromedriver.exe')
wd.get('http://www.google.com')
```