Fastcampus Data Science Extension SCHOOL

Web Scraping - Dynamic Pages

Index

- review
- install selenium
- dynamic web pages
- scrape with selenium

Requirements

- `$ pip install selenium`
- ChromeDriver: https://sites.google.com/a/chromium.org/chromedriver/downloads

Review

requests && BeautifulSoup

- 정적인 페이지를 수집할 때
- requests: HTTP 요청 -> HTML 응답
- BeautifulSoup: HTML 응답 -> 분석 후 요소 접근

But..

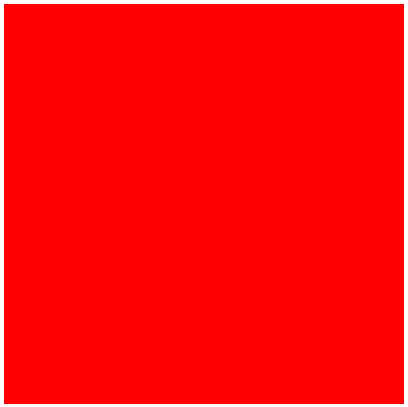
- BeautifulSoup은 AJAX나 javaScript로 그려지는(렌더링) 요소나 행동은 접근할 수 없음

Selenium!

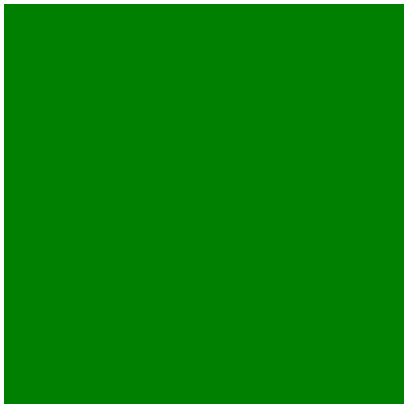
- Web Application User test tool
- `$ pip install selenium`

Dynamic Contents

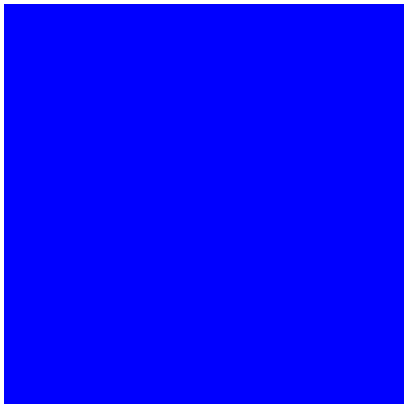
Static Web site - 1



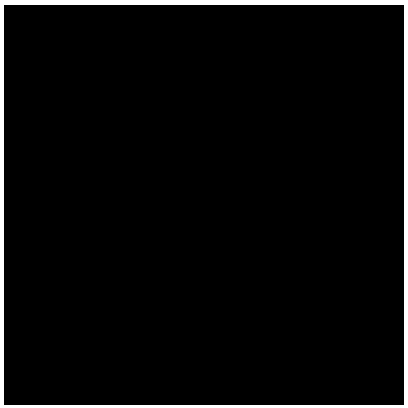
Static Web site - 2



Static Web site - 3



Dynamic Web site



Red Green Blue

Pros & Cons

Pros

- 동적 페이지 제어 가능
- 사용자처럼 행동 가능
- iframe 제어 가능

Cons

- 느림
- BS4에 비해 신경써야 할 것이 많음

Route is important while using Selenium

- BeautifulSoup : 수집할 요소 선택 -> url 정보 수집 -> 스크래핑 수행
- Selenium: 수집할 요소 선택 -> 요소까지의 경로 선정 -> 스크래핑 수행

Web Scraping with Selenium

naver cafe

- login
- execute script
- iframe - switch frame

Quora

- scroll