Fastcampus Sprint - Programming

Day 4. Advanced Web Scraping

Index

- Recap
- Advanced Web Scraping with Selenium
- Web Scraping with cloud service

Recap(1)

```
num_list = [0, -1, 10, 3.14, 2.71828, 10000, 2736, 2847, 25, 287, 1, 50]

for i in num_list:
    if i%5==0:
        print("{}는 5의 배수입니다.".format(i))
```

Recap(2)

```
def get_query():
    # TODO
    # requests, bs4를 이용하여 N사 실시간 검색어
    # 가져오는 거 까지만 하세요.
    # result = [(1,"검색어"), (2,"검색어")]
    return result
```

Requirements

- `$ pip install selenium`
- ChromeDriver: https://chromedriver.storage.googleapis.com/index.html?path=78.0.3904.11/

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`

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

N사 포털 카페 서비스

> 특정 카페의 검색 결과물을 가져와봅시다.

Quora

> 더 많은 데이터 로딩을 위해 스크롤 후 데이터를 가져와봅시다.

Google Cloud

Cloud?

- 인터넷에 연결된 다른 컴퓨터로 연산을 하는 기술
- 접근성, 주문형 서비스 제공으로 경제적이고 효율적인 컴퓨팅 서비스 제공
- Amazon Web Service(Amazon), Google Cloud Platform(Google), Microsoft Azure(Microsoft), ..
- Virtual Machine, Cloud Storage, Database, Docker Engine 등 다양한 서비스 제공

Google Cloud Flatform

- 2011년 Google이 출시한 클라우드 컴퓨팅 솔루션서비스
- 20개의 region과 61개의 zone 서비스 중
- 2020년 서울 region 오픈예정
- https://cloud.google.com/

Google Cloud Functions

- Pricing: https://cloud.google.com/functions/pricing

Store data with mlab

- https://mlab.com/