Selenium과 Scrapy 정리

- 파이썬입문과 크롤링기초 부트캠프 강좌를 통해 익힌 내용 정리

크롤링 핵심 코드 패턴으로 이해하기

```
import requests
from bs4 import BeautifulSoup
```
①

```
res = requests.get('http://v.media.daum.net/v/20170615203441266')
```
②

```
soup = BeautifulSoup(res.content, 'html.parser')
```
③

```
mydata = soup.find('title')
```
④

```
print(mydata.get_text())
```
⑤

필요한 데이터 추출하기

> 이 부분이 크롤링 핵심!

- soup.find() 함수로 원하는 부분을 지정하면 됨
- 변수.get_text() 함수로 추출한 부분을 가져올 수 있음

```
mydata = soup.find('title')
print(mydata.get_text())
```

> 이를 위해 HTML/CSS 언어로 어떻게 웹페이지를 만드는지, 기본 내용을 이해할 필요가 있음!

필요 기술

1. 태그와 속성으로 선택하기
2. CSS Selector로 선택하기
3. XPATH로 선택하기
4. 데이터 후처리
    - 파이썬 문자열 함수/정규 표현식

다양한 크롤링 기술

- 유용한 데이터를 크롤링할 수 있는 Open API
- 로그인이 필요한 웹페이지 크롤링 기법
    - 가장 많은 환경에서 가능한 크롤링 기법: Selenium/Headless Chrome

업무 자동화 기술도 덤으로 얻게 됨

- 크롤링 데이터 기반, 엑셀 보고서 만들기
- 크롤링 데이터 기반, 구글 쉬트 보고서 만들기

풀스택과 빅데이터/데이터 과학의 기본