Python 프로그래밍
#6. 브라우저 자동화 및 데이터 분석

목표

- selenium으로 브라우저를 코드로 다룰 수 있다
- IFrame으로 이루어진 사이트들을 파싱하며, javascript를 강제로 실행시킬 수 있다

python 로컬에 설치하기

- https://www.python.org/downloads/
- 파이썬 3 버전으로 설치
- 윈도우는 설치시 **Add Python to the PATH Environmental Variable** 체크

모듈 설치하기: Windows

시작 -> 검색 -> cmd 검색 (명령프롬프트) -> 오른쪽 클릭 -> 관리자 권한으로 실행 -> 아래 명령어를 하나하나 넣고 엔터

- pip install requests
- pip install beautifulsoup4
- pip install selenium
- pip install jupyter
- pip install pandas

모듈 설치하기: MAC

Launchpad 혹은 Spotlight (ctrl + space) 에서 terminal을 검색 후 실행. Windows 모듈설치법처럼 넣어서 실행하는데 앞에다 sudo를 붙이고 pip3이라고 적는다

- sudo pip3 install request

비밀번호를 물어보면 맥에 로그인했던 비밀번호를 넣으면 된다

Jupyter의 실행

- Windows: 명령프롬프트 창에서 jupyter notebook 명령어로 실행
- Mac: 터미널 창에서 jupyter notebook 명령어로 실행

만약 토큰 관련 문제가 있을 경우

- 터미널에서 jupyter notebook을 실행중이였다면 ctrl + c를 눌러서 중단한다
- jupyter notebook --generate-config를 실행시켜서 설정 파일을 만든다
- jupyter notebook password를 실행시키고 암호를 설정한다
- 다시 jupyter notebook을 실행한다

브라우저 드라이버 설치

- 크롬 드라이버:
  https://sites.google.com/a/chromium.org/chromedriver/downloads
- 파이어폭스 드라이버
  https://github.com/mozilla/geckodriver/releases
- 다운로드한 드라이버 파일을 파이썬이 설치된 장소로 이동시키기

selenium

- 웹 어플리케이션을 위한 테스트 프레임워크
- 테스트 자동화를 위해 여러 기능을 지원해준다
- 우리는 크롤링을 수월하게 하기 위해서 사용한다

데이터 수집의 유형

1. 정적인 사이트 (네이버)
2. 동적인 사이트 (직방, 왓챠, ...)
3. **한국형 사이트: IFrame**
4. **한국형 사이트: javascript**

selenium

```
from selenium import webdriver

# 크롬 브라우저 실행
driver = webdriver.Chrome()
# 파이어폭스 브라우저 실행
driver = webdriver.Firefox()

# 네이버 페이지 접속
driver.get("http://www.naver.com")
```

selenium

- css selector로 찾기

  ```
  driver.find_element_by_css_selector("")
  driver.find_elements_by_css_selector("")
  ```

- 선택된 element 클릭

  ```
  element.click()
  ```

- 선택된 element에 문자열 보내기

  ```
  element.send_keys("")
  ```

CSS (Cascading Style Sheets)

- css selector:
  - class: .class
  - id: #id
  - 요소 안의 하위 요소: element element
  - 하위 요소: element > element
  - 요소안의 속성: element[id='id']

C. 한국형 사이트: IFrame

- 하나의 HTML 문서 안에서 다른 HTML 문서를 보여주고자 할때 사용
- 다른 사이트가 안에 들어가는 것이므로 일반적인 방식으론 파싱 불가
- 개발자 도구를 통해서 IFrame으로 이루어진 사이트 임을 확인한다
- selenium을 사용해서 IFrame의 컨텐츠를 탐색할 수 있다

C. 한국형 사이트: IFrame

- 중고나라 크롤링중 iframe으로 전환하기

```
# iframe이 담긴 element를 선택 후
iframe = driver.find_element_by_css_selector("#cafe_main")

# frame으로 focus변경
driver.switch_to_frame(iframe)

# 원래 접속했던 사이트로 focus변경
driver.switch_to_default_content()
```

D. 한국형 사이트: javascript

- 자바스크립트로 내용을 변경하는 사이트들
- 크롤링을 위해서 페이지에 있는 자바스크립트를 강제로 실행시켜야 할 필요가 있다
- 화장품 성분 분석 사이트

  https://www.kcia.or.kr/cid/Document/020.Ingredient_shis/INGREDIENT_SHIS_10L.asp

D. 한국형 사이트: javascript

- 페이지를 옮기기 위해서 어떻게 해야할까?

```
# 태그로 앨리먼트 찾기
driver.find_element_by_tag_name("")

# fGoPage() 자바스크립트 함수를 실행시켜서 페이지 이동
driver.execute_script("fGoPage(2)")
```