25.Python Selenium

1. Web 개념

1] HTTP Protocol

1) WWW(웹)상에서 문서 전송을 위한 프로토콜

2) request(요청) / response(응답) 으로 구성

3) browser(클라이언트)가 요청하면 web server(서버)가 HTML 파일이나 다른 자원(이미지, 텍스트, 동영상 등)을 응답으로 전송

4) request의 유형에는 대표적으로 GET / POST 가 있음

2] request의 유형

1) GET 방식 : ① 데이터 전달을 URL 내에서 함(네이버 검색, 구글 검색 등)

② 예시 : **www.example.com?id=mommoo&pass=1234**

③ data Size 제한(. 최대 2048 문자의 실제 경로 있는 문자 수 뺀 제한 ,Browser마다 다를수 있음)

2) POST 방식 : ① 데이터 전송을 태그를 통해서 함(사용자에게 직접적으로 노출되지 않음)

② Data Size 제한 없음

③ BODY에 key 와 value 쌍으로 데이터를 넣는다. 똑같이 구분자 &를 사용

2. Python crawling

1] Crawling 개념

1) Web상에 존재하는 Contents를 수집하는 작업 (프로그래밍으로 자동화 가능

2) HTML 페이지를 **가져와서**, HTML/CSS등을 **파싱**하고, 필요한 데이터만 추출하는 기법

3) **Open API(Rest API)**를 제공하는 서비스에 Open API를 호출해서, 받은 데이터 중 필요한 데이터만 추출하는 기법

4) **Selenium**등 브라우저를 프로그래밍으로 조작해서, 필요한 데이터만 추출하는 기법

2] Rest API 개념

1) Rest 개념

- "Representational State Transfer" 의 약자
- 자원을 이름(자원의 표현)으로 구분하여 해당 자원의 상태(정보)를 주고 받는 모든 것을 의미
- 상태(정보) 전달데이터가 요청되어지는 시점에서 자원의 상태(정보)를 전달.
- JSON 혹은 XML를 통해 데이터를 주고 받는 것이 일반적
- HTTP URI(Uniform Resource Identifier)를 통해 자원(Resource)을 명시하고, HTTP Method(POST, GET, PUT, DELETE)를 통해 해당 자원에 대한 CRUD Operation을 적용하는 것을 의미

2) API(Application Programming Interface)

- 데이터와 기능의 집합을 제공하여 컴퓨터 프로그램간 상호작용을 촉진하며, 서로 정보를 교환가능

3) REST API의 정의

- REST 기반으로 서비스 API를 구현
- 최근 OpenAPI(누구나 사용할 수 있도록 공개된 API: 구글 맵, 공공 데이터 등), 마이크로 서비스(하나의 큰 애플리케이션을 여러 개의 작은 애플리케이션으로 쪼개어 변경과 조합이 가능하도록 만든 아키텍처) 등을 제공하는 업체 대부분은 REST API를 제공

3] JSON(JavaScript Object Notation)

- 웹 환경에서 서버와 클라이언트 사이에 데이터를 주고 받을 때 많이 사용

3. Python Selenium1

1] Selenium 개념

1) Web 을 Test 하기위한 Framework

- 테스트 코드를 사용하여 브라우져에서의 액션을 테스트할 수 있게 해주는 툴

2) 공식 home Page : www.seleniumhq.org

3) Web Driver Install

- https://sites.google.com/a/chromium.org/chromedriver (Chrome Browser용)

4) 설치후 다음 Sites에서 가장 최신 Version을 DownLoad받아서 덮어씌움

- https:/chromedriver.storage.googleapis.com/index.html
- Window :

c:/ C:/PyCharmProject/Sources/selenium/

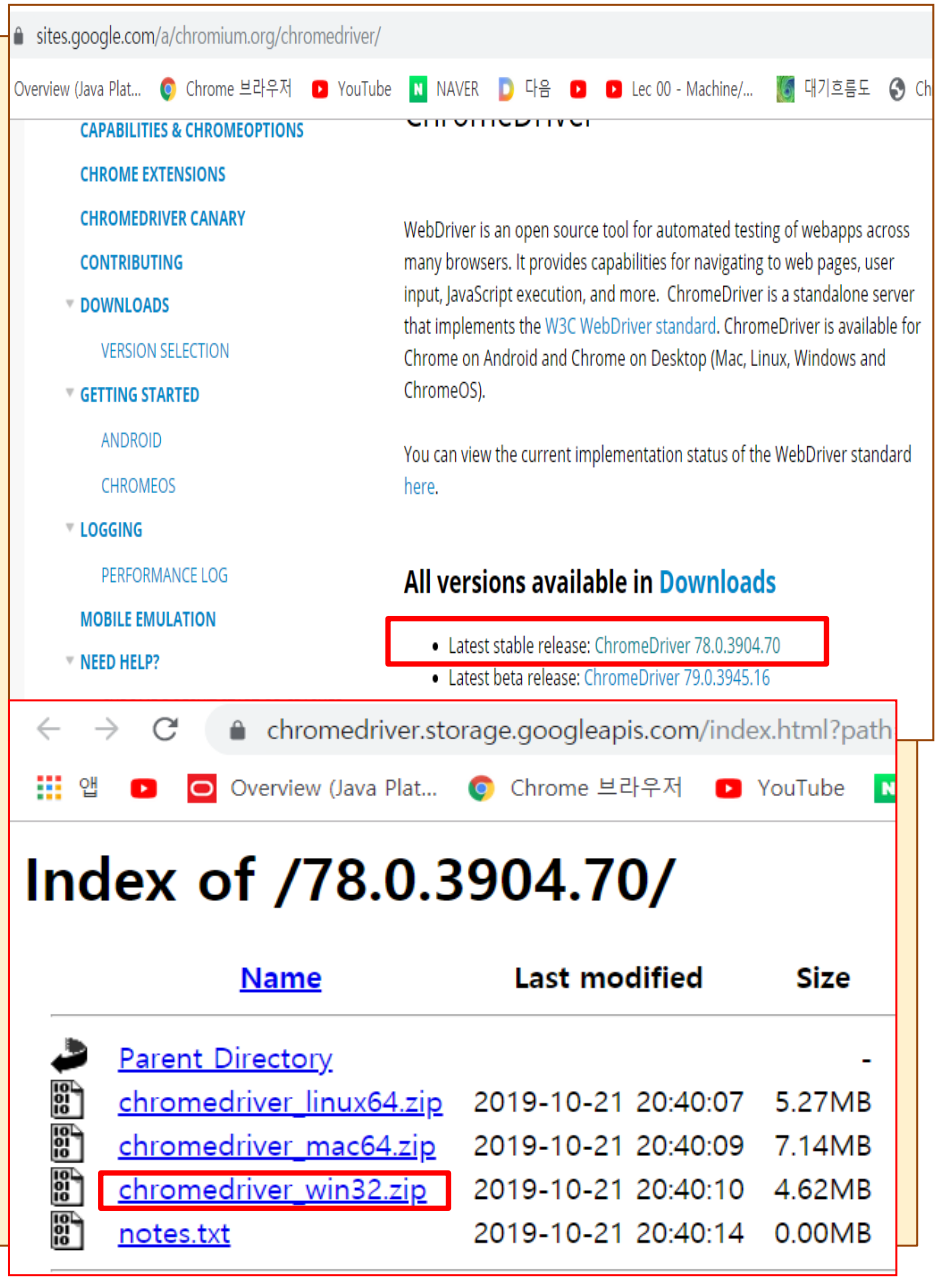
chromedriver_win32 /chromedriver.exe

2] **사전준비 (Selenium 설치)**

1) Selenium 인스톨: pip install selenium

2) 웹드라이버 인스톨: 웹 테스트 자동화를 위해 제공되는 툴 (각 browser및 os 별로 존재)

- Firefox, chromedriver 등 각 브라우저 마다 웹드라이버 다운로드 가능



3. Python Selenium2(Driver)

1. PhantomJS 개념

1) Web Testing을 위해 나온 화면이 존재하지 않는 브라우저

2) 터미널환경에서 동작하는 크롤러의 경우 PhantomJS 브라우저 사용 권장

3) 공식 home Page : www.seleniumhq.org

4) 사전준비 (PhantomJS 설치)

① 윈도우: PhantomJS 다운로드 후 절적한 디렉토리에 압축을 품

(http://phantomjs.org/download.html)

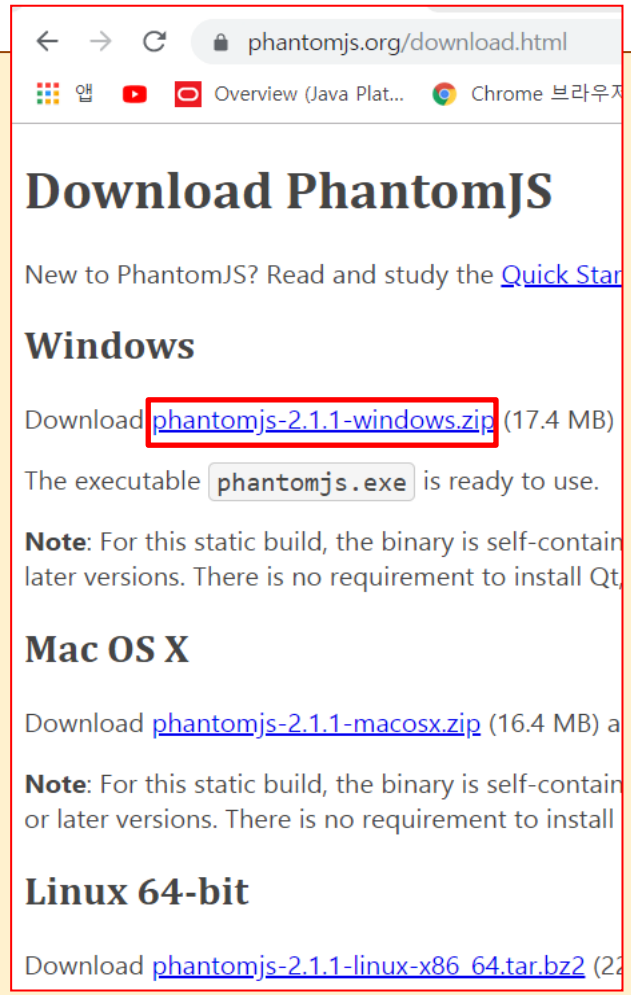
② 맥: brew install phantomjs 또는 윈도우에서 사용한 웹사이트를 활용

③ 리눅스: suto apt-get install phantomjs

(한글폰트가 없다면, 추가로 sudo apt-get install -y fonts-nanum*)

* 확인: 설치 디렉토리를 알아둬야 함

* Selenium을 사용할지, PhantomJS를 사용할지를 정해서 드라이버를 생성하는 코드를 자신의 로컬 환경에 맞게 넣어주신 후에 실행



2. Headless Chrome

1) 최신 Crawling 기술

2) PhantomJS유사 기술 , chrome Browser 기능으로 개발

3) 설치는 따로 필요없음 , 아래 4줄 선언

```
chromedriver = 'C:/PyCharmProject/Sources/selenium/chromedriver_win32/chromedriver.exe' # 윈도우
headlessOptions = webdriver.ChromeOptions()
headlessOptions.add_argument('headless')
driver = webdriver.Chrome(chromedriver , options=headlessOptions)
```

4. Python Selenium 검색 함수

3] **Selenium 태그 검색 주요 함수**

① find_element_by_name(): HTML name, 최초 발견된 name

② find_elements_by_tag_name(): 모든 태그 리스트로 가져오기

③ find_element_by_id: HTML id

④ find_element_by_class_name: some class_name

⑤ find_element_by_css_selector : CSS 규약 → 예시] #css > div.selector

⑥ find_element_by_xpath: /html/body/some/xpath

4] **XPATH 이용, Crawling**

1) 마크업에서 요소를 정의하기 위해 path 경로를 사용하는 방법

2) find_element_by_xpath(), find_elements_by_xpath() 메서드로 검색 가능

3) XPATH 문법 상세

항목	문법상세
/	절대경로
//	문서내 검색
//@href	href 속성이 있는 모든 태그 선택
//a[@href='http://google.com']	a 태그의 href 속성에 http://google.com 속성값을 가진 모든 태그 선택
(//a)[3]	문서의 세 번째 링크 선택
(//table)[last()]	문서의 마지막 테이블 선택
(//a)[position() < 3]	문서의 처음 두 링크 선택
//table/tr/*	모든 테이블에서 모든 자식 tr 태그 선택
//div[@*]	속성이 하나라도 있는 div 태그 선택

5. Python 페이지 로딩 시간을 기다린 후, 크롤링하기

1] 페이지 로딩 시간을 기다린 후, 크롤링하기

1) 몇몇 페이지의 경우, 페이지 로딩 지연이 발생하여(여러 요청이 병합하여 페이지 결과를 생성) tag를 못읽어오는 경우가 발생할 수 있음(아래의 코드를 이용하여 해결 가능)

e.g) 10초내에 해당 tag를 찾으면 반환, 그렇지 않으면 timeout 발생!

2] from selenium.webdriver.common.by import By 검색 지원 방법

항목	문법상세
By.ID	태그에 있는 ID 로 검색
By.CSS_SELECTOR	CSS Selector 로 검색
By.NAME	태그에 있는 name 으로 검색
By.TAG_NAME	태그 이름으로 검색

3] 주요 함수 요소 내용 가져오기

① head Tag 관련 요소 : get_attribute('text')

② body Tag 관련 요소 : text

6. Python **Open API 사용**

1. **정부3.0 공공 데이터 포털 API 사용**

   1) 공공 데이터 포털 가입하기 https://www.data.go.kr

   2) 회원가입 -> 로그인 ->

   https://www.data.go.kr/dataset/3050988/openapi.do -> 활용신청

   활용신청 메뉴 옆에 상세정보 메뉴의 활용사례 확인

2. 다양한 사이트 크롤링 – 트위터(트위터 개발자 등록)

   1) https://apps.twitter.com/app/new

   2) 트위터 개발자 등록 및 코드 개발 참고:
   https://m.blog.naver.com/PostView.nhn?blogId=acwboy&logNo=220541273950&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

   3) tweepy - 트위터 크롤링/트윗 라이브러리

   4) pip install tweepy

   사용법

   ① 검색 키워드를 활용한 트위터 게시글 검색

   ② 검색 조건을 활용한 트위터 게시글 검색

   ③ 트위터에 간단히 글쓰기