

데이터 수집과 관리

임정빈

CONTENTS

1. 데이터 수집

2. 데이터 관리(DBMS)

데이터 수집 방법

Web Crawling(웹 크롤링)

- Requests
- Beautifulsoup
- Selenium
- 이외에도 여러 방법이 존재 (ex : ajax 렌더링 크롤링)

OPEN API

Web Crawling(웹 크롤링)

웹이란 ?

- 월드 와이드 웹(World Wide Web,WWW)의 약자
인터넷에 연결된 클라이언트들이
정보를 공유할 수 있는 공간을 의미

크롤링의 과정

- 1. 웹에서 **HTML** 파일을 다운로드 (Requests)
- 2. 다운로드 한 **HTML**에서 원하는 데이터를 파싱 (BS4)

HTML(Hyper Text Markup Language)

- HTML은 하이퍼텍스트 마크업 언어(HyperText Markup Language)의 약자.
- 데이터의 구조나 형식을 지정하는 언어로 **요소,태그,속성** 등으로 이루어져 있다.
- 크롤링을 할 때, 먼저 원하는 데이터의 HTML(+CSS)의 요소, 태그 등을 확인한다.

HTML(Hyper Text Markup Language)

The screenshot shows the Korean Film Council (KOFIC) website with movie rankings for the US, UK, and Germany. The browser's developer tools are open, showing the HTML structure of the page.

KOFIC 영화진흥위원회 Korean Film Council

로그인 | ID/PW찾기 | 회원가입 | 영화정보

DB검색 | 박스오피스 | 테마통계 | 공식통계 | 온라인상영관 | 고객센터

미국	영국	독일
1 Captain Marvel \$67,988,130 - 0	1 Captain Marvel £ 6,643,217 - 0	1 Captain Marvel € 4,279,492 - 0
2 Wonder Park \$ 15,853,646 NEW	2 Fisherman's Friends £ 1,154,864 NEW	2 Green Book € 1,023,217 - 0
3 Five Feet Apart \$ 13,190,286 NEW	3 What Men Want £ 836,612 NEW	3 Asterix The Secret Of T... € 879,819 NEW
4 How To Train Your Dragon... \$ 9,277,310 ↓ 2	4 Lego Movie 2: The Second... £ 643,623 ↓ 1	4 Escape Room € 817,432 ↑ 1
5 Tyler Perry's A Madea Fa... \$ 7,836,167 ↓ 2	5 Fighting With My Family £ 632,598 ↓ 3	5 Ostwind 4Arts Arrival € 766,403 ↓ 1
6 No Manches Frida 2 \$ 3,831,401 NEW	6 How To Train Your Dragon... £ 530,499 ↓ 1	6 How To Train Your Dragon... € 736,452 ↓ 3
7 Captive State \$ 3,131,525 NEW	7 Instant Family £ 467,902 ↓ 3	7 Pocca Changes The World! € 424,509 NEW
8 Lego Movie 2: The Second... \$ 2,150,853 ↓ 4	8 Green Book £ 351,502 ↓ 2	8 Cold Pursuit € 416,927 ↓ 2
9 Alita: Battle Angel \$ 1,900,355 ↓ 4	9 Eggnog 3 £ 167,934 NEW	9 Bohemian Rhapsody € 385,607 - 0
10 Green Book \$ 1,258,795 ↓ 4	10 Kid Who Would Be King, T... £ 108,728 ↓ 2	10 Trautmann € 373,097 NEW

US : 03.15 ~ 03.17 UK : 03.15 ~ 03.17 DE : 03.15 ~ 03.17

출처: RENTRAK

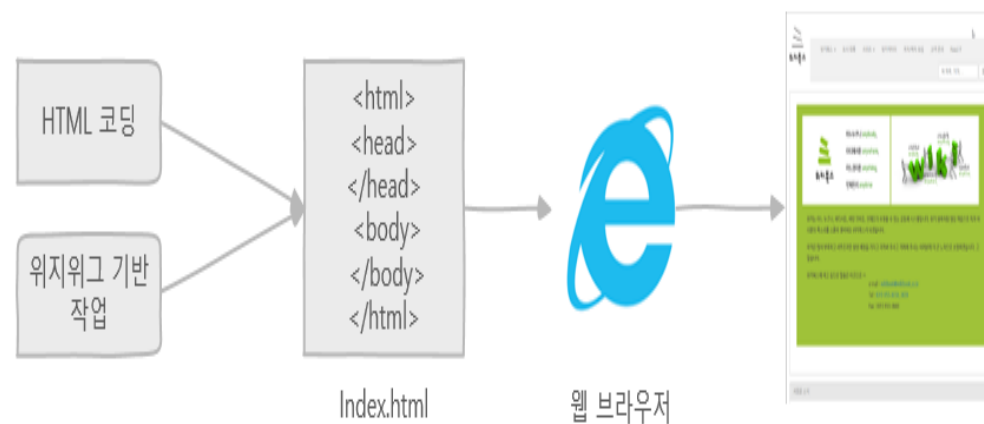
html.gr_kobis_or_kr body

```
<!doctype html>
<html lang="ko" class="gr_kobis_or_kr">
  <head>...</head>
  <body data-gr-c-s-loaded="true"> == $0
    <!-- layerPopup -->
    <section id="wrap_layer">
    </section>
    <div id="wrap_layer_1" style="display: none; z-index: auto;">...</div>
    <div id="wrap_layer_2" style="display: none; z-index: auto;">...</div>
    <div id="wrap_layer_3" style="display: none; z-index: auto;">...</div>
    <div id="wrap_layer_4" style="display: none; z-index: auto;">...</div>
    <!-- //layerPopup -->
    <div id="m_wrap" style="background-color:rgb(242,242,242);">...</div>
  </body>
</html>
```

- 크롬에서 **F12**를 누르면 위의 사진 처럼 해당 페이지의 HTML 구조를 확인해 볼 수 있다.

Requests

- 웹에서 **HTML** 파일을 **다운로드** 하는 단계에서 사용
- urllib 모듈 보다 편리하고 빠름.
 - conda install requests (or pip install requests)
 - GET method를 이용해 HTTP 요청을 보낼 수 있다 (HTML을 가져오는 것)
 - 그 외 자세한 내용은 아래의 공식 문서 참조.



<http://docs.python-requests.org/en/latest/>

BeautifulSoup

- 다운로드 한 **HTML**에서 원하는 데이터를 **파싱**, 추출할 때 사용
- `conda install bs4` (or `pip install bs4`)
- 사용할 `parser`를 지정 해줘야 함 (lxml 추천)
- bs4 객체 생성 후, `find` 등의 `method`를 활용해 데이터를 추출.
- 실습 코드 참조

Selenium

- 웹 브라우저를 컨트롤하는 라이브러리

JavaScript 요소가 있는 페이지를 다루어야 할 때,
requests를 통한 접속이 막혀 있는 사이트 등
Bs4로 크롤링이 **불가능한** 부분에서 사용한다.

크롬 및 크롬 드라이버 설치 필요 :

<https://sites.google.com/a/chromium.org/chromedriver/downloads>

*직관적이나 속도가 느리므로, bs4와 적절히 섞어서 사용하는 것이 좋다.

ChromeDriver - WebDriver for Chrome

CHROMEDRIVER

CAPABILITIES & CHROME OPTIONS

CHROME EXTENSIONS

CHROMEDRIVER CANARY

CONTRIBUTING

▼ DOWNLOADS

VERSION SELECTION

▼ GETTING STARTED

ANDROID

CHROME OS

▼ LOGGING

PERFORMANCE LOG

Downloads

Current Releases

- If you are using Chrome version 74, please download [ChromeDriver 74.0.3729.6](#)
- If you are using Chrome version 73, please download [ChromeDriver 73.0.3683.68](#)
- If you are using Chrome version 72, please download [ChromeDriver 2.46](#) or [ChromeDriver 72.0.3626.69](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

If you are using Chrome from Dev or Canary channel, please download [ChromeDriver 2.46](#). This is not officially :
in most cases it should work without major issues.

For more information on selecting the right version of ChromeDriver, please see the [Version Selection](#) page.

Selenium 명령어 (참고)

Selenium

find_element_by_id(id)	id 속성으로 요소 하나 추출	
find_element_by_name(name)	Name 속성으로 요소 하나 추출	
find_element_by_class_name(name)	클래스 이름이 name에 해당되는 요소 하나 추출	find_elements_by_class_name
find_element_by_partial_link(text)	링크의 자식요소에 포함되어 있는 텍스트로 요소 하나 추출	find_elements_by_partial_link
find_element_by_tag_name(name)	태그 이름이 name에 해당하는 요소 하나 추출	find_elements_by_tag_name
find_element_by_link_text(text)	링크 텍스트로 요소 하나 추출	
find_element_by_xpath(query)	xpath를 지정해 요소 하나 추출	find_elements_by_xpath
find_element_by_css_selector(query)	css 선택자 요소 하나 추출	find_elements_by_css_selector

Selenium 명령어 (참고)

Selenium으로 요소 조종하기

<code>clear()</code>	글자 입력란에 글자를 지움	<code>id</code>	요소의 id속성
<code>click()</code>	요소를 클릭	<code>location</code>	요소의 위치
<code>get_attribute(name)</code>	Name에 해당되는 값을 추출	<code>parent</code>	부모 요소
<code>is_displayed()</code>	요소가 화면에 출력되는지 확인	<code>rect</code>	크기와 위치정보를 가진 딕셔너리 자료형을 리턴
<code>is_selected()</code>	체크박스 등의 요소가 선택된 상태인지 확인	<code>screenshot_as_base64</code>	BASE64로 스크린샷을 추출
<code>is_enabled()</code>	요소가 활성화되어 있는지 확인	<code>screenshot_as_png</code>	PNG형식으로 스크린샷 추출
<code>screenshot(filename)</code>	스크린샷을 찍는다.	<code>size</code>	요소의 크기
<code>send_keys(value)</code>	키를 입력한다.	<code>tag_name</code>	태그 이름
<code>submit()</code>	입력 양식을 전송한다.	<code>text</code>	요소의 내부글자
<code>value_of_css_property(name)</code>	Name에 해당하는 CSS속성의 값을 추출		

Selenium 명령어 (참고)

Selenium 드라이버 조작

<code>add_cookie(cookie_dict)</code>	쿠키 값을 딕셔너리 형식으로 지정	<code>get_screenshot_as_base64()</code>	Base64형식으로 스크린샷을 추출
<code>back() / forward()</code>	이전 페이지 또는 다음 페이지로 이동	<code>get_screenshot_as_png()</code>	PNG형식으로 스크린샷을 추출
<code>close()</code>	브라우저를 닫는다	<code>get_window_position(windowHandel = "current")</code>	브라우저의 위치를 추출
<code>current_url</code>	현재 url을 추출	<code>get_window_size(windowHandel = "current")</code>	브라우저의 크기를 추출
<code>delete_all_cookies()</code>	모든 쿠키를 제거	<code>implicitly_wait(sec)</code>	대기시간을 토 단위로 지정해 대기
<code>delete_cookie(name)</code>	특정 쿠키를 제거	<code>quit()</code>	selenium자체를 종료
<code>execute(command, params)</code>	브라우저의 고유 명령어를 실행	<code>save_screenshot(filename)</code>	스크린 샷을 저장
<code>excute_async_script(script,*args)</code>	비동기 처리하는 자바스크립트를 실행	<code>set_page_load_timeout(time_to_wait)</code>	페이지를 읽는 타임아웃 시간을 지정
<code>execute_script(script,*args)</code>	동기 처리하는 자바스크립트를 실행	<code>set_script_timeout(time_to_wait)</code>	스크립트의 타임아웃 시간을 지정
<code>get(url)</code>	url로 브라우저 이동	<code>set_window_position(x,y>windowHandle='current')</code>	브라우저의 위치를 지정
<code>get_cookie(name)</code>	특정 쿠키 값을 추출	<code>set_window_size(가로,세로>windowHandle='current')</code>	브라우저의 크기를 지정
<code>get_cookies</code>	모든 쿠키 값을 딕셔너리 형식으로 추출	<code>title</code>	현재페이지의 타이틀을 추출
<code>get_log(type)</code>	로그를 추출(browser/drvrt/client/server)		

Open API

- 공개 API라고도 불리며, 누구나 사용할 수 있도록 공개된 API
(주로 Rest API 기술을 많이 사용함)

- 일반적으로 json 형태로 데이터를 보내줌.
필요한 데이터를 open API를 통해 구할 수 있다면
크롤링 대신 **API**를 사용하는 게 효율적임.

ex : 공공데이터포털(<https://www.data.go.kr/>), 네이버 개발자 센터
(<https://developers.naver.com>), 등등..

데이터 관리

- 수집한 데이터를 어떻게 관리할 것인가?
 1. CSV(or json) 파일로 저장
 2. PICKLE 형태로 저장(파이썬 객체 저장용)
 3. **Database**에 저장
- 실습 코드 참고

데이터 베이스와 SQL

데이터베이스 (DataBase)

데이터란 의미 있는 정보를 가진 모든 값, 사람이나 자동 기기가 생성 또는 처리하는 형태로 표시된 것.
여기에 특정한 의미가 부여될 때, '정보'가 된다.

데이터베이스란 여러 사람에 의해 공유되어 사용될 목적으로 통합되어 관리되는 데이터의 집합을 말한다.

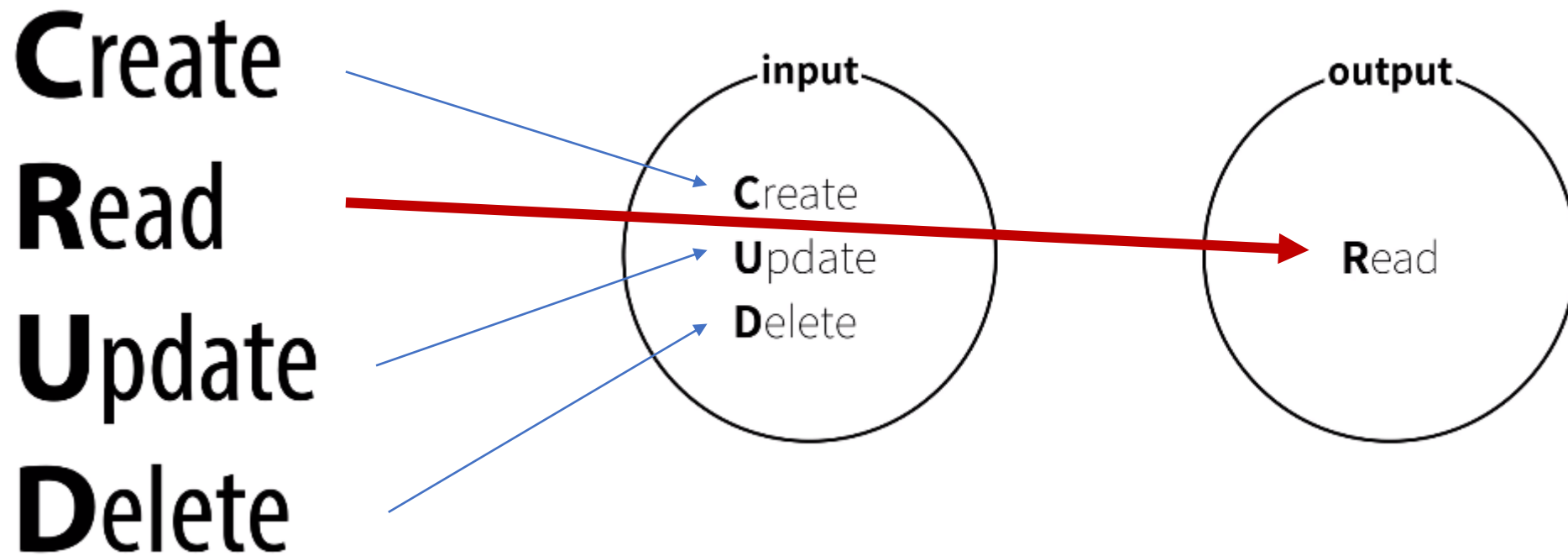
데이터베이스를 효율적으로 관리하는 소프트웨어

→ DBMS (데이터베이스 관리 시스템; Database Management System)

데이터베이스는 다양한 종류가 있다. 계층형 데이터베이스, 관계형 데이터베이스, 키-벨류 스토어 등...

데이터 베이스와 SQL

데이터베이스의 본질



데이터 베이스와 SQL

■ 관계형 데이터베이스 구조

스키마: 데이터베이스의 구조와 제약조건을 정의하는 가장 큰 틀

테이블: 스키마에 적합하게 만들어진, 데이터가 저장되는 작은 틀
(보통 특정 주제, 기능에 맞는 데이터들이 저장된다. user, order, production table...)

행, 열, 기본키, 외래키: 테이블속의 ‘실직적인 데이터’를 구성하는 요소들

데이터 베이스와 SQL

관계형 데이터베이스 구조

[표 II-1-2] K-리그 2차 자료 정리

선수	팀	포지션	백넘버	생년월일	키	몸무게	...
박지성	서울FC	MF	7	1981/02/25	178cm	73kg	⋮
이청용	블루윙즈	MF	17	1988/07/02	180cm	69kg	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



[그림 II-1-3] 데이터베이스의 테이블

SQLITE란 ?

- SQLite는 MySQL나 PostgreSQL와 같은 데이터베이스 관리 시스템이지만, 서버가 아니라 응용 프로그램에 넣어 사용하는 비교적 가벼운 데이터베이스이다.
- 일반적인 RDBMS에 비해 대규모 작업에는 적합하지 않지만, 중소 규모라면 속도에 손색이 없다. 또 API는 단순히 라이브러리를 호출하는 것만 있으며, 데이터를 저장하는 데 하나의 파일만을 사용하는 것이 특징이다. 컬럼을 삭제하거나 변경하는 것 등이 제한된다.
- 따로 복잡한 설치과정 없이 python에서 바로 쓸 수 있으므로, DBMS에 대해 알기 위해 간단히 실습해볼 예정.
- <https://www.sqlite.org/> => 제대로 배워보고 싶으면 공식문서 참조!