2019/09/26

크롤링과 데이터 관리

5기 백승렬

Growth Hackers

CONTENTS

Growth Hackers

데이터 수집 방법

Web Crawling(웹 크롤링)

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

OPEN API

Web Crawling(웹 크롤링)

웹 이란 ?

- 월드 와이드 웹 (World Wide Web,WWW) 의 약자

인터넷에 연결된 클라이언트들이
정보를 공유할 수 있는 공간을 의미

크롤링 의 과정

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

Growth Hackers

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.

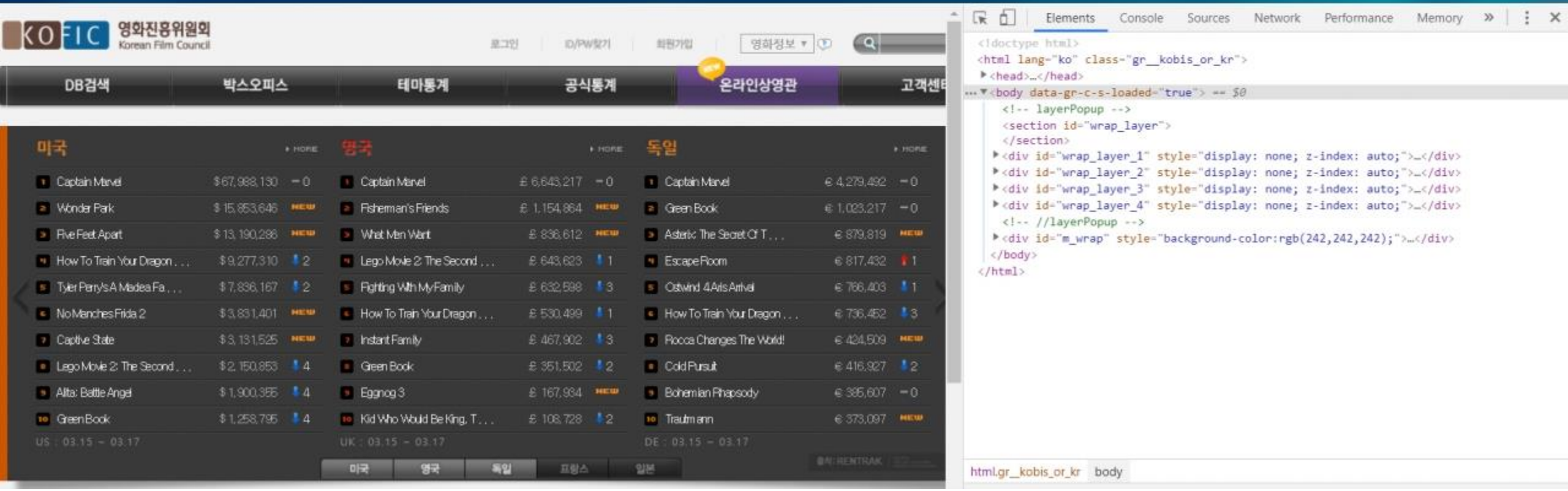
HTML(Hyper Text Markup Language)

- 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)



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

Growth Hackers

웹의 구성 요소 – HTML

HTML 태그

```
<!DOCTYPE html>
<html lang='en'>
<head>
<title></title>
</head>
<body>

    <p>Hello GH!</p>
    <h1><p>Hello GH!</p></h1>
    <h2><p>Hello GH!</p></h2>

    <ul>
        <li>이재경</li>
    </ul>

    <ol>
        <li>이재경</li>
    </ol>
</body>
</html>
```

Hello GH!

Hello GH!

Hello GH!

- 이재경

1. 이재경

HTML 에는 어떤 태그들이 있는지 몇가지 알아보겠다 .
이런 태그들은 body 태그 안에 포함된다 .

p 태그
문단을 나타내는 태그
h 태그
폰트 크기를 설정하는 태그 ; h 태그는 숫자가 작을수록 폰트가 커지며 1 에서 6 까지 지원합니다 .
ul, ol, li 태그
리스트를 만드는 태그

[그림 1-2]

Growth Hackers

웹의 구성 요소 - HTML

HTML 태그

```
<!DOCTYPE html>
<html lang='en'>
<head>
<title></title>
</head>
<body>

    <table>
        <tr>
            <td>1</td>
            <td>2</td>
            <td>3</td>
        </tr>
        <tr>
            <td>1</td>
            <td>2</td>
            <td>3</td>
        </tr>
    </table>
    <table>
        <thead>
            <tr>
                <th>11</th>
                <th>22</th>
                <th>33</th>
            </tr>
        </thead>
        <tbody>
            <tr>
                <td>111</td>
                <td>222</td>
                <td>333</td>
            </tr>
        </tbody>
    </table>
</body>
</html>
```

```
1 2 3
1 2 3
11 22 33
111 222 333
```

[그림 1-3]

table 태그

table 태그를 이용하여 표를 표현할 수 있다 . 과거에는 table 태그를 이용하여 테이블 뿐만 아니라 페이지의 레이아웃을 잡는 역할로 많이 사용한다 .

table 태그는 내부적으로 thead, tbody 를 가지고 있을 수 있으며 , tr 을 이용하여 행을 표현하고 td 와 th 를 이용하여 각 행의 열을 표현한다 .

th 태그 와 td 태그는 모두 한 행에서 열을 나타내지만 th 태그를 줄 경우 열 가운데 정렬과 굵은 글씨체가 된다 .

웹의 구성 요소 - HTML

HTML 태그

```
<!DOCTYPE html>
<html>
<head>
<title></title>
</head>
<body>
    <ul>
        <li>
            <a href="http://www.naver.com/">네이버로 이동하기</a>
        </li>
    </ul>
    <img src="GH.png" alt= "GH">

    <p>Growth</p>
    <span>Growth</span>
    <span>Hackers</span>
    <span><br>Growth</span>
</body>
</html>
```

• 네이버로 이동하기

GH

Growth

Growth Hackers
Growth

[그림 1-4]

a 태그
다른 페이지로 이동할 때의 태그
a 태그는 herf 를 속성으로 가지며 이는 이동하게 될 링크입니다 .

img 태그
이미지를 띄우게 하는 태그
img 는 src 와 alt 속성을 가집니다 . src 를 통하여 이미지를 불러오게 되며 , alt 를 통하여 이미지가 정상정으로 불려지지 않았을 때 , 텍스트로 대체하여 나타냅니다 .
img 태그는 닫는 태그를 요구하지 않습니다 .

span 태그
p 태그와 같이 텍스트를 추가적으로 넣을 수 있습니다 .
다만 문단이 바로 나누어 지는 p 태그와 달리 span 태그는 옆으로 나열되게 됩니다 .
,</br> 태그를 사용하여 p 태그와 같이 사용할 수 있으며 , 이들 또한 굳이 쌍을 이룰 필요는 없습니다 .

웹의 구성 요소 – HTML

HTML 태그

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>
    <div>
        <h1>첫 번째 div</h1>
        <span>이재경</span>
    </div>

    <div>
        <h1>두 번째 div</h1>
    </div>
</body>
</html>
```

첫 번째 div

이재경

두 번째 div

[그림 1-4]

div 태그

가장 자주 사용하는 태그로 , 태그는 특정 기능이 있는 것이 아니라 단순히 영역을 잡아서 레이아웃을 만드는 역할로 사용한다 .

div 태그를 사용하면 눈에는 보이지 않지만 그 하위 태그들의 영역을 잡아 주는 역할을 한다 .

예전에는 div 태그를 이용하지 않고 table 태그를 이용하여 웹 사이트 구조를 잡았었기에 간혹 table 을 중첩해서 만든 사이트들이 존재한다 .

웹의 구성 요소 – CSS 와 JavaScript

CSS 와 JavaScript

```
<!DOCTYPE html>
<html>
<head>
<title>Title</title>
<style>
    p{
    font-size: 17px;
    color: blue;
    }
    li.li-target{
    color: red;
    }
    #target{
    color: green;
    }
</style>

</head>
<body>

    <p>p 태그1</p>
    <p>p 태그2</p>
    <span id="target">스팬</span>
    <ul>
        <li>li1</li>
        <li class="li-target">li3</li>
    </ul>
</body>
</html>
```

p 태그1

p 태그2

스팬

- li1
- li3

[그림 1-4]

CSS (Cascading Style Sheets) 은 웹 사이트를 꾸며주는 역할을 한다 .
CSS 를 이용하여 꾸미기 위해 특정 요소에 접근하는 것을 셀렉터 (selector) 라고 부르며 , 특정한 태그를 이용하거나 id 와 class 라는 속성을 이용하는 방식이 크롤러에서 똑같이 사용 가능하다 .

class 는 마침표 (.) 를 붙여 class 임을 나타냅니다 . class 를 이용하면 태그와 상관없이 같은 class 를 공유하는 항목에 접근한다 .

id 는 class 와 달리 id 값이 고유해야 하기에 중복 사용하지 않기를 권장하고 있다 .

JavaScript 는 script 태그를 이용하여 웹 사이트에 기능을 넣어줄 수 있다 .
script 태그는 head 에 들어가도 되지만 , body 의 가장 하단 부분에 넣어주는 것을 권장하고 있다 .

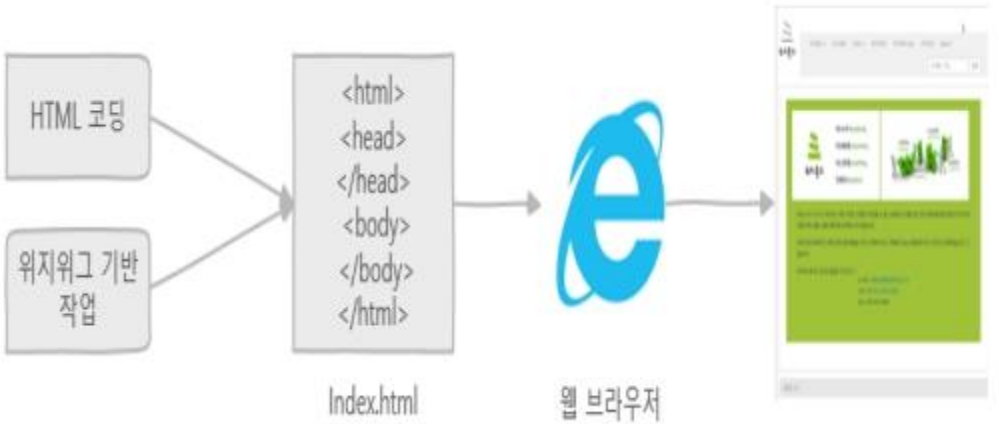
Requests

- 웹에서 HTML 파일을 다운로드 하는 단계에서 사용
- urllib 모듈 보다 편리하고 빠름 .

- conda install requests (or pip install requests)

- GET method 를 이용해 HTTP 요청을 보낼 수 있다 (HTML 을 가져오는 것)

- 그 외 자세한 내용은 아래의 공식 문서 참조 .



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

Growth Hackers

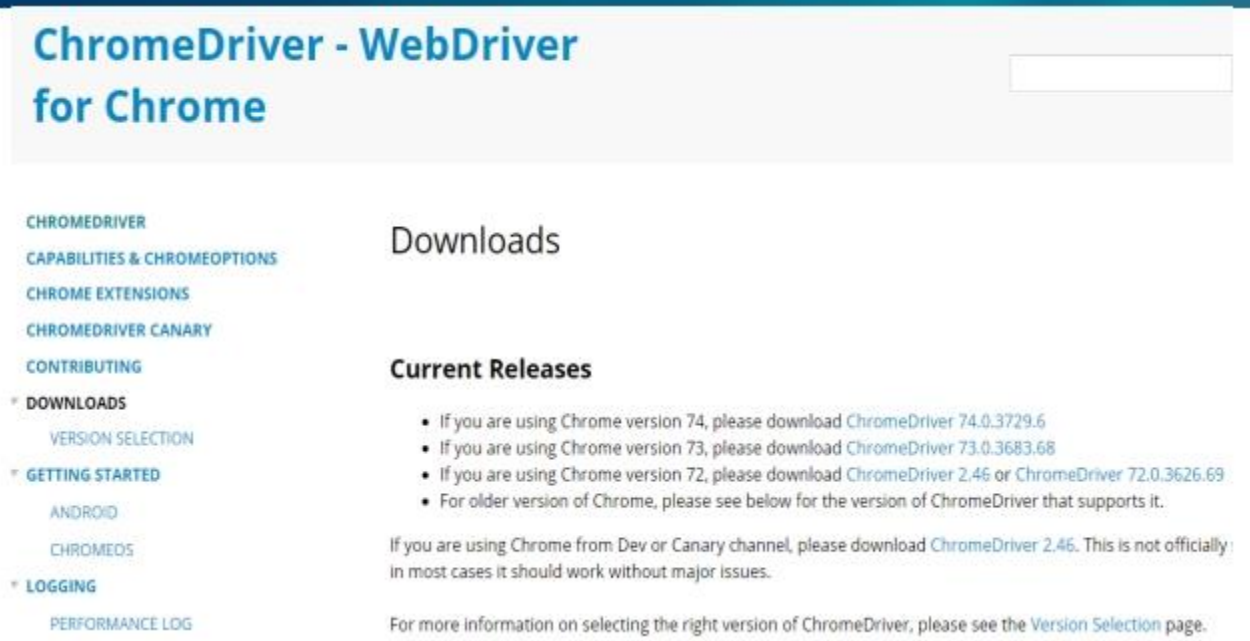
BeautifulSoup

- 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 ,
requests 를 통한 접속이 막혀 있는 사이트 등
Bs4 로 크롤링이 불가능한 부분에서 사용한다 .



크롬 및 크롬 드라이버 설치 필요 :
https://sites.google.com/a/chromium.org/chromedriver/downloads

* 직관적이나 속도가 느리므로 , bs4 와 적절히 섞어서 사용하는 것이 좋다 .

Growth Hackers

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

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

Growth Hackers

Selenium 명령어 (참고)

Selenium 드라이버 조작

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

Growth Hackers

Open API

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

- 일반적으로 json 형태로 데이터를 보내줌.
필요한 데이터를 open API 를 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 표시 된 것.
여기에 특정한 의미가 부여될 때, '정보'가 된다.
데이터베이스란 **여러 사람에 의해 공유되어 사용될 목적으로 통합되어 관리되는 데이터의 집합**을 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

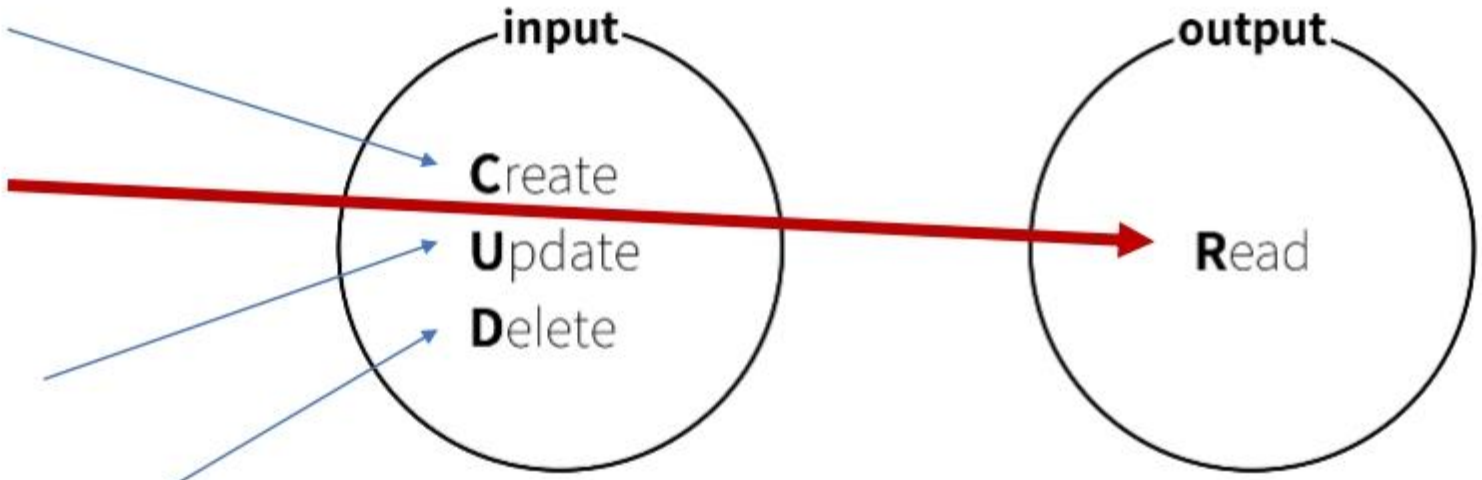
데이터베이스의 본질

Create

Read

Update

Delete



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 '실질적인 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

Growth Hackers

SQLITE란?

- SQLite 는 MySQL 나 PostgreSQL 와 같은 데이터베이스 관리 시스템이지만 , 서버가 아니라 응용 프로그램에 넣어 사용하는 비교적 가벼운 데이터베이스이다 .
- 일반적인 RDBMS 에 비해 대규모 작업에는 적합하지 않지만 , 중소 규모라면 속도에 손색이 없다 . 또 API 는 단순히 라이브러리를 호출하는 것만 있으며 , 데이터를 저장하는 데 하나의 파일만을 사용하는 것이 특징이다 . 컬럼을 삭제하거나 변경하는 것 등이 제한된다 .
- 따로 복잡한 설치과정 없이 python 에서 바로 쓸 수 있으므로 , DBMS 에 대해 알기 위해 간단히 실습해볼 예정 .

- https://www.sqlite.org/ => 제대로 배워보고 싶으면 공식문서 참조 !

QUEST

- 실습코드를 활용해서, 최근 일주일 간의 일일 박스오피스 50위까지의 데이터를 selenium 과 beautiful soup를 활용해서 가져오는 코드를 작성해주세요
- 가져온 데이터를 날짜별로 dataframe을 만든 후, sqlite3를 이용해 dailyboxoffice.db 형태로 저장해주세요.
- db 안에는 날짜별로 나눠진 총 7개의 테이블이 포함되어 있어야 합니다.
- 작성한 코드와 db 파일을 올려주시면 됩니다