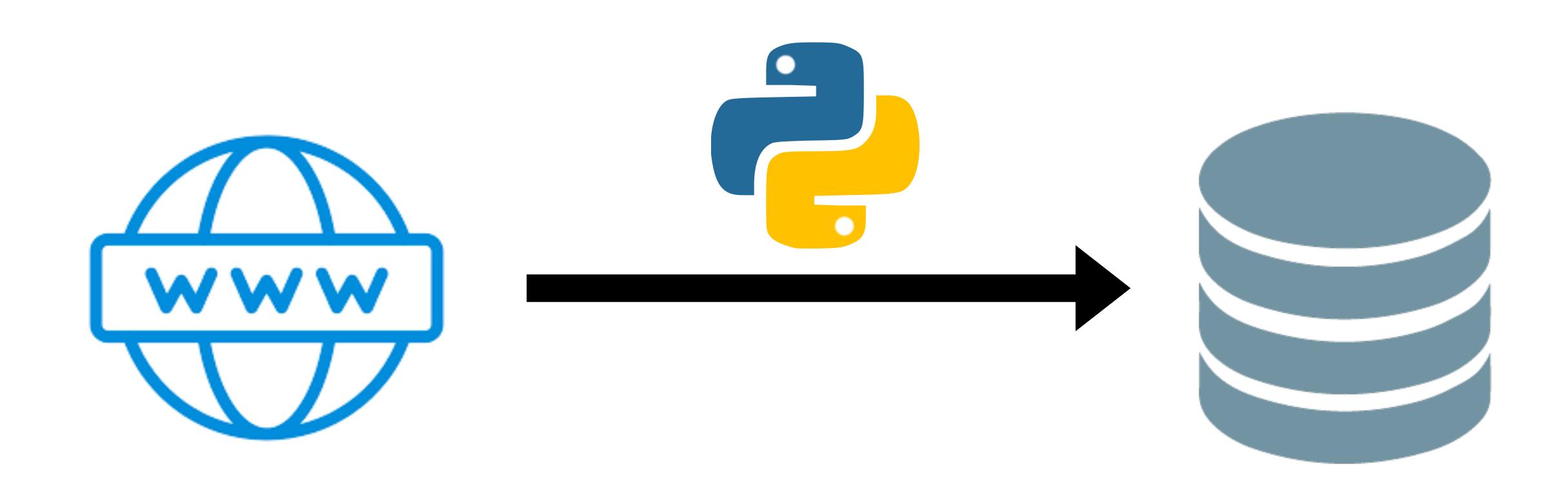
# Python을 활용한 웹 크롤링

2018.04.03. 시범강의 크롤링, 스크래핑이란??

컴퓨터 소프트웨어 기술을 사용하여 웹에서 원하는 정보를 자동적으로 추출하는 것

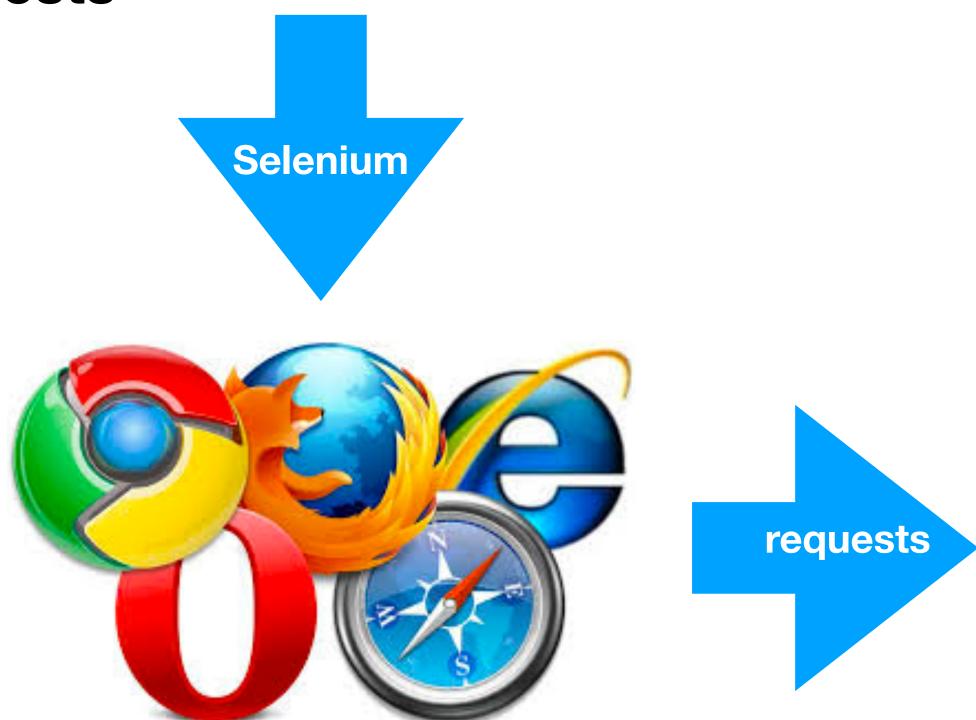
## 크롤링, 스크래핑이란??

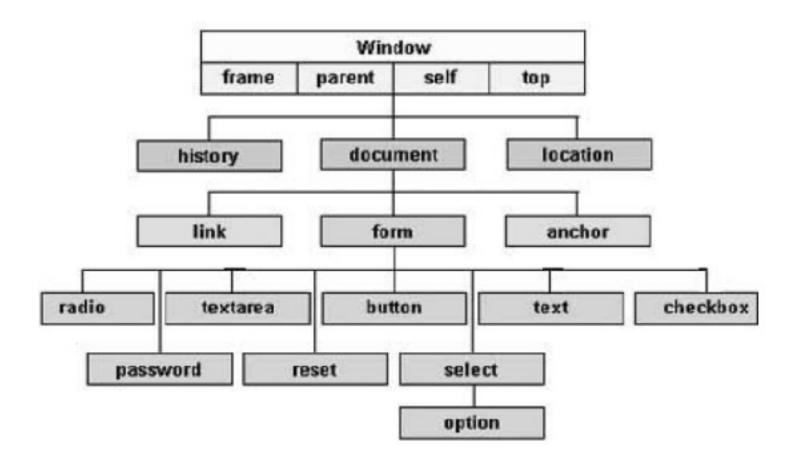
컴퓨터 소프트웨어 기술을 사용하여 웹에서 원하는 정보를 자동적으로 추출하는 것

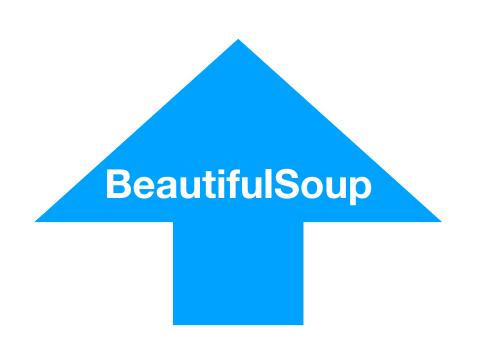


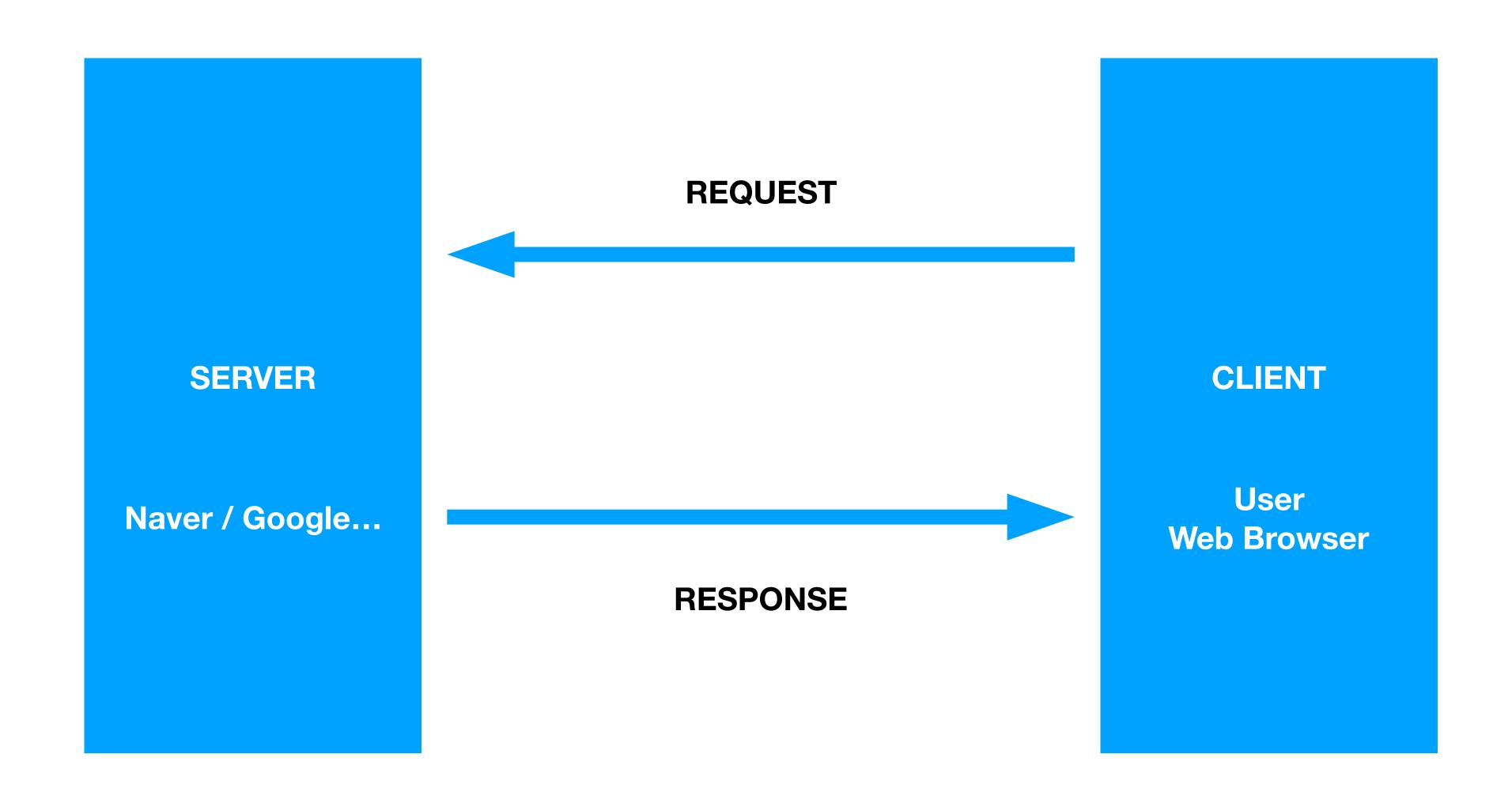
- 1. Python
  - 1. Selenium
  - 2. BeautifulSoup
  - 3. Requests
- **2. WEB** 
  - 1. HTML DOM
  - 2. GET / POST
- 3. DATABASE
  - 1. SQL
  - 2. Python 연동

- 1. Python
  - 1. Selenium
  - 2. BeautifulSoup
  - 3. Requests



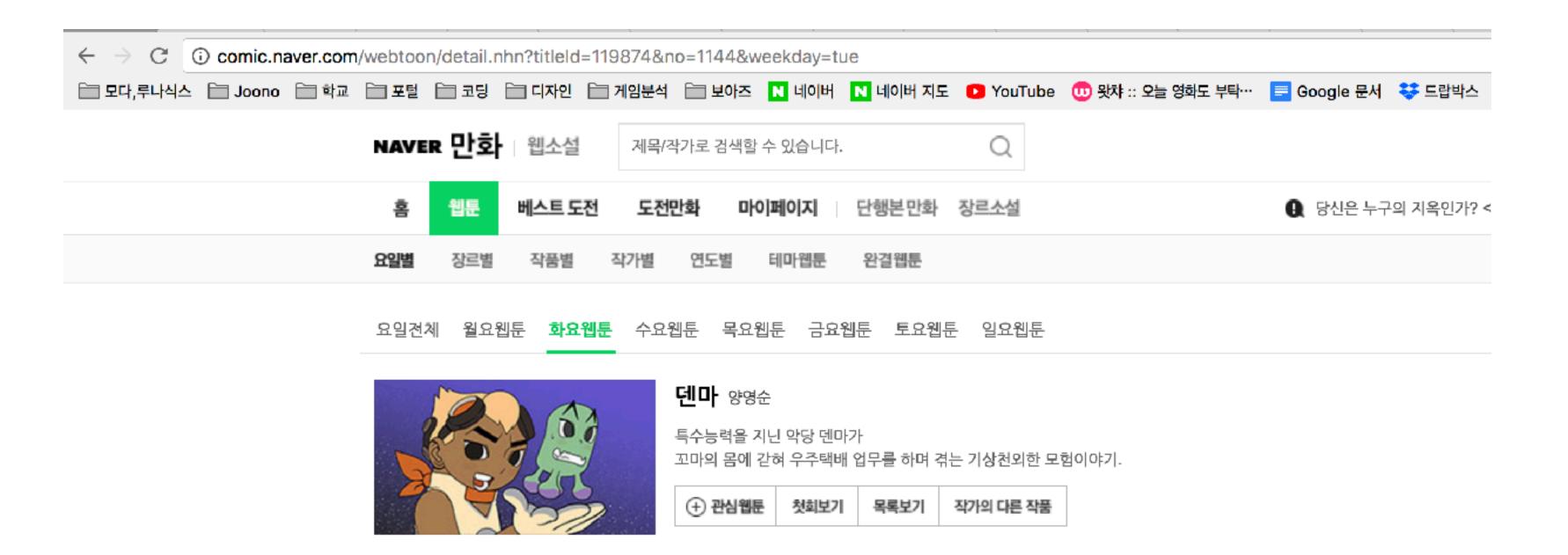






#### **GET**

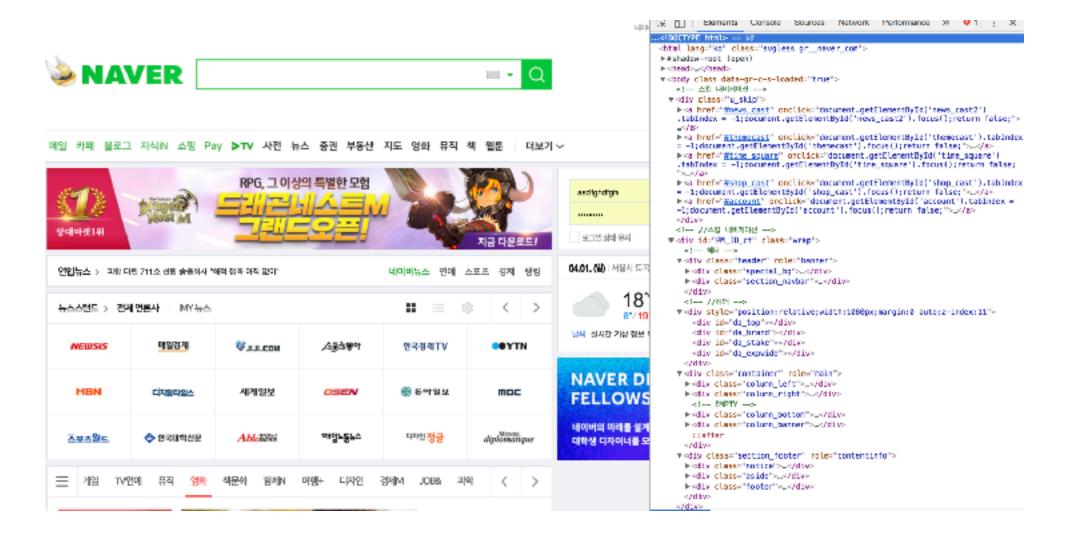
http://comic.naver.com/webtoon/detail.nhn?titleld=119874&no=1144&weekday=tue

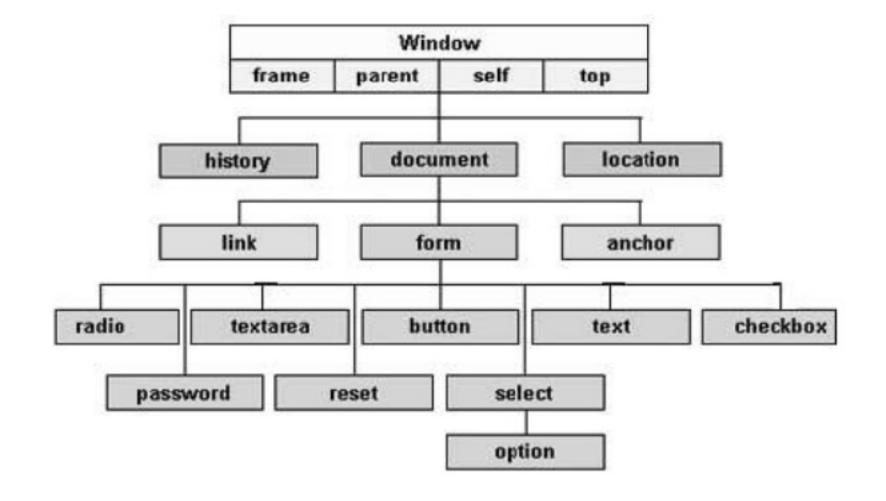


#### **POST**

#### http://cafe.naver.com/boazbigdata

```
Lr i | Elements Console Sources Network Performance Memory Application ≫
Smart Editor™ ?
                                                                                                                         ▼<<mark>form</mark> id="frmWrite" name="frmWrite" method="post" action="/
                                                                                                                        ArticlePost.nhn"> == $0
                                                                                                                            <input type="hidden" name="preAccessMenuid" value="427">
           모두를 위한 딥러닝_신준호 ♣│ 말머리선택
게시판
                                                                                                                            <input type="hidden" name="clubid" value="26506451">
                                                                                                                            <input type="hidden" name="menuidForList" value="427">
제목
                                                                                  임시 저장된 글 (1)
           게시글 제목을 입력하세요
                                                                                                                            <input type="hidden" name="clubTempId" value="pdKfzfs//m+b+FXiNfDsb6">
                                                                                                                            <input type="hidden" name="branchCode" value=</pre>
파일첨부
                  🕟 동영상 🚻 📗  지도 📗 필정 📗 🔗 링크 📗 🖉 파일 📗 투표 📗 🕞 음악
                                                                                                                            "291ky7zrYvASjN9Mjg@YcPHWK87gbdEvQPkomVfVN+TeLWziko8bCs8FlmipMeGyP/
                                                                                                                            cFxZreYeY7fbuDcMy/
정보첨부
                영화DB | 드라마DB | 음악DB | 상품DB | 인물DB | 날씨DB | 지식백과DB
                                                                                                                            USGozQHZ679mI+kX3tmJm039CqYQ+BvU2iS0mNW7remzBgA39uRW85RxBKuKijzIcErQ3EluK
                                                                                                                            T4Hi93piBR12RA=">
                                                                                                                            <input type="hidden" name="boardtype" value="L">
  기본쓰기
                                                                                                                            <input type="hidden" name="page" value="1">
                                                                                                                            <input type="hidden" name="m" value="write">
  글꼴 - 크기 - 가 가 가 가 를 통 물 물 물 물 표 표 66 至 URL @ 图 ※ 오 ♡ 😂 스E터 🖭 🗈
                                                                                                                            <input type="hidden" name="attachpollyn" id="attachpollyn">
                                                                                                                            <input type="hidden" name="attachPollids" id="attachPollids">
                                                                                                                            <input type="hidden" name="attachfileyn" id="attachfileyn" value>
                                                                                                                            <input type="hidden" name="attachimageyn" id="attachimageyn" value>
                                                                                                                            <input type="hidden" name="attachfiles" id="attachfiles">
                                                                                                                            <input type="hidden" name="attachsizerealsum" id="attachsizerealsum"</pre>
                                                                                                                            <input type="hidden" name="attachmodifylist" id="attachmodifylist">
                                                                                                                            <input type="hidden" name="attachsizes" id="attachsizes">
                                                                                                                            <input type="hidden" name="attachimagesizesum" id="attachimagesizesum"</pre>
                                                                                                                            value="0">
                                                                                                                            <input type="hidden" name="attachfilesizesum" id="attachfilesizesum"</pre>
                                                                                                                            value="0">
                                                                                                                            <input type="hidden" name="scrapedyn" id="scrapedyn" value>
                                                                                                                            <input type="hidden" name="attachinfolist" id="attachinfolist" value>
                                                                                                                            <input type="hidden" name="attachinfo" id="attachinfo" value="0">
                                                                                                                            <input type="hidden" name="attachmaplist" id="attachmaplist" value>
                                                                                                                            <input type="hidden" name="attachmovie" id="attachmovie" value>
                                                                                                                            <input type="hidden" name="attachmovielist" id="attachmovielist" value>
                                                                                                                            <input type="hidden" name="article.leveragecode" id=</pre>
                                                                                                                            "article.leveragecode" value="0">
                                                                                                                            <input type="hidden" name="article.attachCalendarList" id=</pre>
                                                                                                                            "calendarAttachmentStruct">
                                                                                                                            <input type="hidden" name="attachCalendar" id="calendarAttachment">
                                                                                                                            <input type="hidden" name="attachedCalendar" value>
                                                                                                                            <input name="tempsaveid" id="tempsaveid" type="hidden" value>
                                                                                                                            <input type="hidden" name="owfs" id="owfs" value="false">
                                                                                                                            <input type="hidden" name="ndriveid" id="ndriveid" value="0">
                                                                                                                            <input type="hidden" name="hadNaverPoll" id="hadNaverPoll" value>
                                                                                                                            <input type="hidden" name="article.attachMusic" id="musicAttachment">
                                                                                                                            <input type="hidden" name="alimCode" value="lCQeLUxdntRMWnCrsqhlU=">
                                                                                                                            <input type="hidden" name="parameterString" id="parameterString" value>
                                                                                               나의 태그!
          태그와 태그는 쉼표로 구분하며, 10개까지 입력하실 수 있습니다.
                                                                                                                            <input type="hidden" name="representImagePath" id="representImagePath"</pre>
```





#### SQL이란?

- RDBMS(관계형 데이터베이스 관리 시스템)의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어

왜 필요한가?

- 크롤링한 데이터가 방대한 경우, 혹은 관계형 데이터베이스에 관리해야 할 경우

Python에서는 어떻게 사용할까?

- Pymysql 패키지를 사용하여, Python script안에서 SQL을 사용

- 3. DATABASE
  - 1. SQL
  - 2. Python 연동

## **PRACTICE**

https://github.com/jjjjooonno/fc\_wc/blob/master/코드/test\_case/exam1.py