

파이썬을 이용한 데이터수집 및 스마트공장 견학

Crawling 활용
Selenium

2021년 1월 14일
안재관

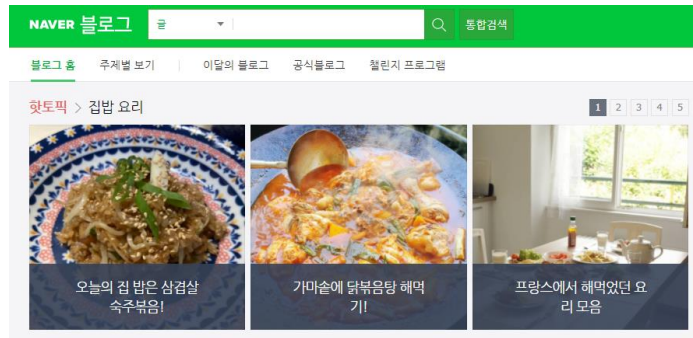
금일 목표

- BeautifulSoup로 crawling이 안되는 경우에 대해 알아본다
- Javascript에 대해 간단히 이해한다
- Selenium 사용법을 알아본다

기존 방식으로 안되는 페이지?

● 찾고자 하는 내용을 HTML 소스에서 찾을 수가 없음?

- <https://section.blog.naver.com/BlogHome.nhn?directoryNo=0¤tPage=1&groupld=0>
- <https://blog.naver.com/> 를 들어가면 redirect 되는 주소
- 80줄 소스코드 안에 블로그 포스트 내용은 하나도 없음



아직 이곳이 없습니다.

아래 주제별 블로거 추천을 통해 관심 주제의 블로거 이웃을 만들어보세요.



```
<!DOCTYPE html>
<html lang="ko">
<head>
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <base href="/" />
  <meta name="robots" content="noindex,nofollow">
  <meta name="referrer" content="always">
  <meta name="format-detection" content="telephone=no">
  <link rel="shortcut icon" type="image/x-icon" href="https://section.blog.naver.com/favicon.ico?3">

  <meta property="og:title" content="네이버 블로그">
  <meta property="og:image" content="https://bloglog.nstatic.net/nblog/eylog/post/og/default_image_160610.png">
  <meta property="og:description" content="당신의 모든 기록을 담은 공간" />

  <meta property="se:feed:serviceId" content="blog" />
  <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
  <title>네이버 블로그</title>

  <link rel="stylesheet" type="text/css" href="https://ssl.static.net/t-static.blog/section/versionina/section-page-202197315_https.css" charset="UTF-8">

  <script>
    var angularConfig = {"isDev":false,"isDebugEnabled":false};
    var readOnlyStatus = {"isReadOnly":false,"startTime":null,"endTime":null,"reason":null,"userAgent":null,"maintenanceType":null,"maintenancePeriod":null};
    var urlMap = {"tvec2":"https://tvec2.naver.com","ad":"https://veta.naver.com","id_secret":"https://nid.naver.com","likeit":"https://blog.like.naver.com","nadbile":"https://admin.blog.naver.com","section":"https://section.blog.naver.com","blog":"https://blog.naver.com","sai":"https://ssl.static.net/sai","tvec2":"https://tvec2.naver.com","kin":"https://kin.naver.com","help":"https://help.naver.com","search":"https://search.naver.com","guestbook":"https://guestbook.blog.naver.com","static":"https://static.nstatic.net/templates/gnb_ut18.nhn","thumbn12":"https://blogthumb.nstatic.net","lcs":"lcs.naver.com"};
    var nsc = "blog.mainv2";

    // 팝업을 통해서 부모함을 제거하기 위해 필요
    document.domain = "naver.com";
    //후로라일 css 가 제대로 동작안해서 필요
    var doc = document.documentElement;
    doc.setAttribute("data-useragent",navigator.userAgent);
  </script>
  <script type="text/javascript" src="https://ssl.static.net/t-static.blog/section/versionina/koecBundle-162545304_https.js" charset="UTF-8"></script>
  <script type="text/javascript" src="https://ssl.static.net/t-static.blog/section/versionina/koecBundle-7343430_https.js" charset="UTF-8"></script>
  <script type="text/javascript" src="https://ssl.static.net/t-static.blog/section/versionina/koecBundle-34182188_https.js" charset="UTF-8"></script>
</head>
<body bg-body-click>
  <ui-view autoscroll="true"></ui-view>
</body>
<script>
  if(window.isNotSupportBrowser()){
    alert('IE8 이하 브라우저에서는 "신규버전 보기"를 제공하지 않습니다. 최신버전 업데이트를 부탁드립니다.');
```

기존 방식으로 안되는 페이지?

● 찾고자 하는 내용을 HTML 소스에서 찾을 수가 없음?

- <http://147.46.178.16:33333/javascript.html>
- 처음에 페이지에 들어갔을 때와 "로또번호 생성하기" 를 눌렀을 때와 화면에 표시되는게 다름

[로또번호 생성하기](#)

[로또번호 생성하기](#)

40

4

26

24

28

42

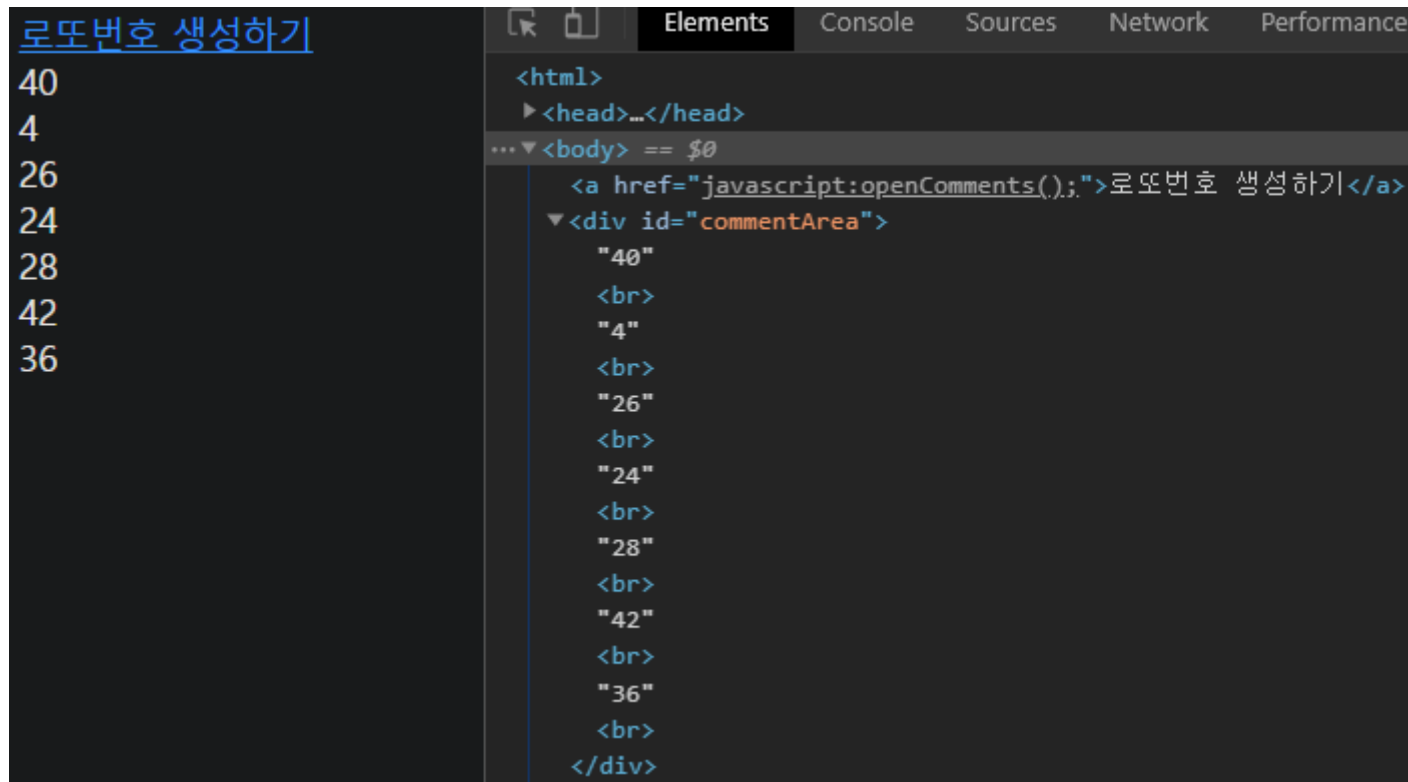
36

- 소스코드에는? 7개 숫자들이 전혀 없음

```
1 <a href="javascript:openComments();">로또번호 생성하기</a>
2 <div id="commentArea"></div>
3
4 <script>
5   function openComments() {
6     a = "";
7     for (i = 0; i < 7; i++) {
8       a += Math.ceil(Math.random() * 45) + "<br>";
9     }
10    document.getElementById("commentArea").innerHTML = a;
11  }
12
13  function displayTime() {
14    document.getElementById('demo').innerHTML = Date()
15  }
16
17 </script>
```

기존 방식으로 안되는 페이지?

- 찾고자 하는 내용을 HTML 소스에서 찾을 수가 없음?
 - <http://147.46.178.16:33333/javascript.html>
 - 브라우저 검사에는 분명히 나옴



- How? Javascript

Javascript 소개

- 웹페이지의 추가적인 기능이 동작하는 것을 담당하는 언어
- 확장자는 *****.js**
- HTML이 웹페이지의 내용/뼈대/구조라면, CSS는 옷입히기/디자인, Javascript는 부가기능
- 웹을 구성하는 마지막 단추 (HTML, CSS, Javascript)
- 동적 웹 페이지를 가능케 함, Javascript 관련 라이브러리와 프레임워크들이 많음
 - jQuery
 - AJAX
 - AngularJS
 - Node.js
- **Requests, BeautifulSoup**는 브라우저와 달리 Javascript 실행을 할 수 없음
 - Javascript 실행이 가능한 패키지가 필요 : Selenium

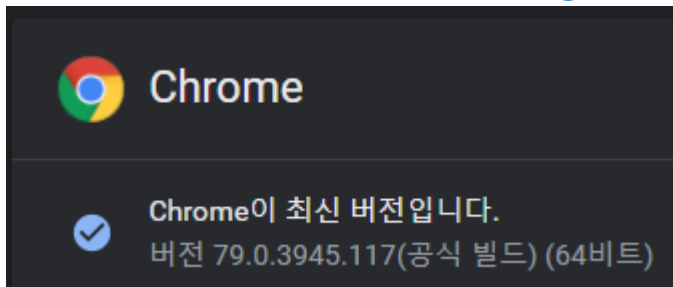
Selenium 설치

- **Requests, BeautifulSoup는 브라우저와 달리 Javascript 실행을 할 수 없음**
 - Javascript 실행이 가능한 패키지가 필요
 - Selenium : 가장 널리 쓰이는 브라우저 자동화 도구, 실제 브라우저를 띄움
- **Selenium 설치**
 - `Pip install selenium`

Selenium 설치

● Selenium 설치

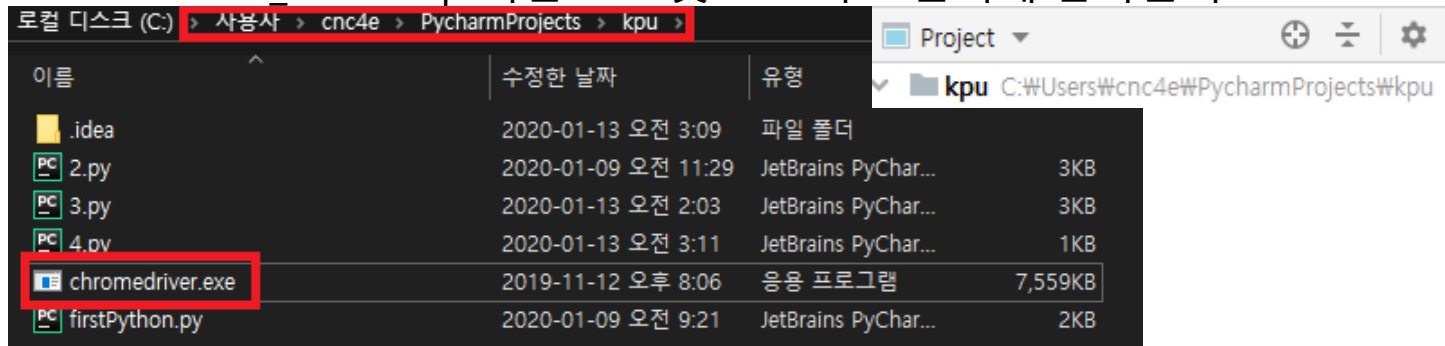
- Chrome에서 <chrome://settings/help> 에 접속



- <https://chromedriver.chromium.org/downloads> 에 들어가서 위 버전에 맞는 링크를 선택

Chrome version 80, please download [ChromeDriver 80.0.3987.16](#)
Chrome version 79, please download [ChromeDriver 79.0.3945.36](#)
Chrome version 78, please download [ChromeDriver 78.0.3904.105](#)

- chromedriver_win32.zip 다운로드 및 프로젝트 폴더에 압축풀기



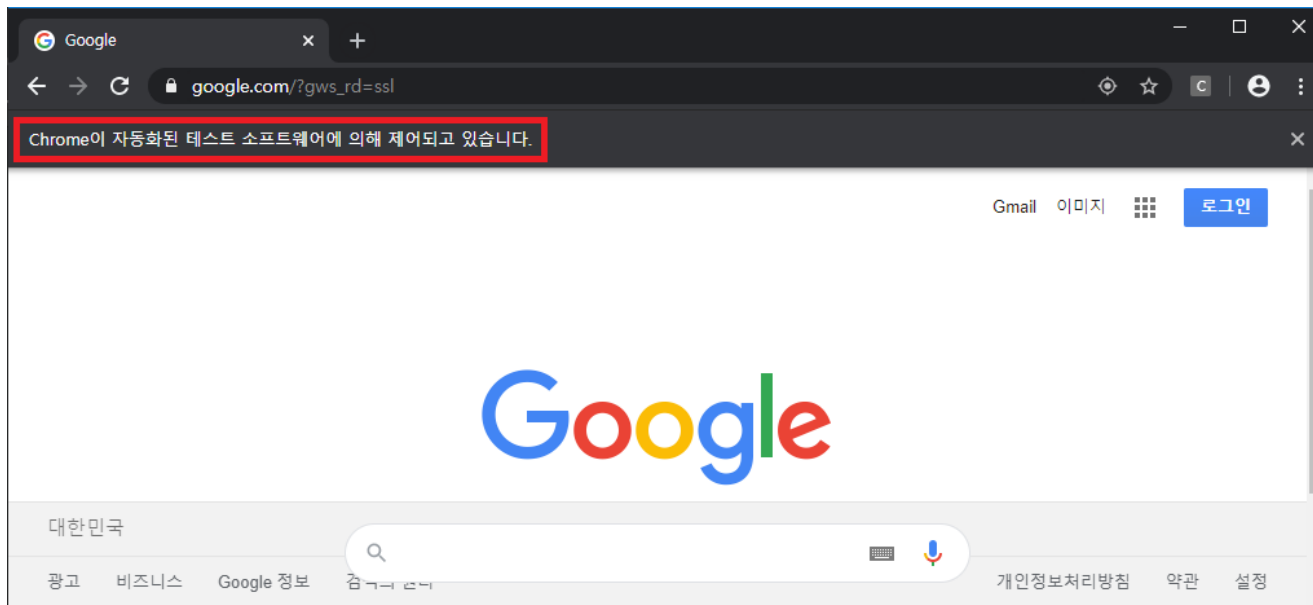
Selenium을 이용한 Crawling

● Selenium으로 웹페이지의 HTML 소스 받기

- Selenium 패키지 사용

```
from selenium import webdriver  
browser = webdriver.Chrome("./chromedriver")  
browser.get("http://www.google.com")
```

- 브라우저가 새로 뜨며 지정한 주소로 이동함



Selenium을 이용한 Crawling

● 기존에 실패했던 페이지들 재시도

- <https://section.blog.naver.com/BlogHome.nhn?directoryNo=0¤tPage=1&groupId=0> 재방문하여 HTML 소스 출력

```
from selenium import webdriver
browser = webdriver.Chrome("./chromedriver")
browser.get("https://section.blog.naver.com/BlogHome.nhn?directoryNo=0&currentPage=1&groupId=0")
print(browser.page_source)
```



```
1 from selenium import webdriver
2 browser = webdriver.Chrome("./chromedriver")
3 url="https://section.blog.naver.com/BlogHome.nhn?directoryNo=0&currentPage=1&groupId=0"
4 browser.get(url)
5 print(browser.page_source)

<div class="area_guide">
  <a bg-nclink="htm.news" ng-href="https://blog.naver.com/blogpeople?Redirect=Category&categoryNo=27&parentCategoryNo=27"
target="_blank" class="title" href="https://blog.naver.com/blogpeople?Redirect=Category&categoryNo=27&parentCategoryNo=27">
  누구보다 발빠르게<br><span class="point">블로그 새소식</span>
  <i class="sp_common icon_news"></i>
```

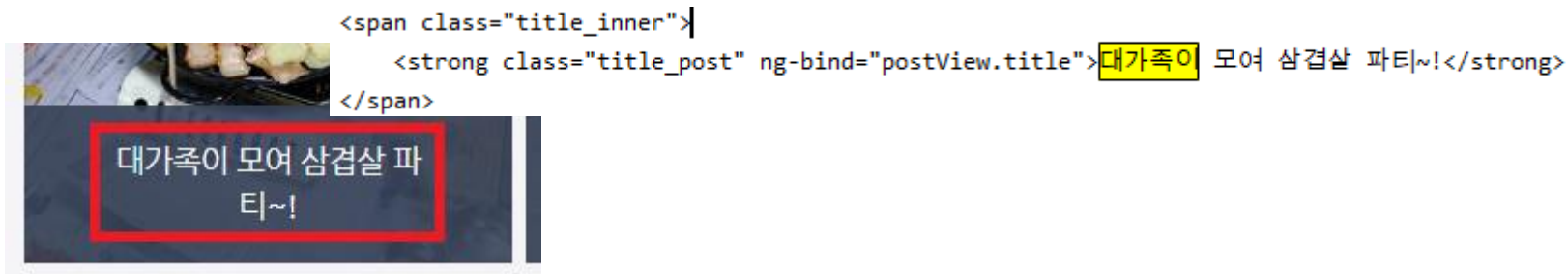
Selenium을 이용한 Crawling

● 기존에 실패했던 페이지들 재시도

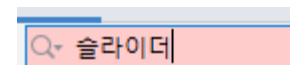
- <https://section.blog.naver.com/BlogHome.nhn?directoryNo=0¤tPage=1&groupId=0> 재방문하여 HTML 소스 출력

```
from selenium import webdriver
browser = webdriver.Chrome("./chromedriver")
browser.get("https://section.blog.naver.com/BlogHome.nhn?directoryNo=0&currentPage=1&groupId=0")
print(browser.page_source)
```

- 그래도 안나오거나, 나오는 내용이 있고 안나오는 내용이 있음?!?!



카메라 슬라이더 멋진 영상을 위한 HORUSBENNU V2-260 슬라이드캠 구입후기



- Why?

Selenium을 이용한 Crawling: time 모듈과 sleep() 함수

● 기존에 실패했던 페이지들 재시도

- 페이지가 로딩될 시간을 줘야함, 페이지의 일부만 로딩된 상태에서 소스코드를 출력하라고 했음
- 1초 sleep 후 출력시

```
from selenium import webdriver
browser = webdriver.Chrome("./chromedriver")
browser.get("https://section.blog.naver.com/BlogHome.nhn?directoryNo=0&currentPage=1&groupId=0")
import time
time.sleep(1)
print(browser.page_source)
```

카메라 슬라이더 멋진 영상을 위한 HORUSBENNU V2-260 슬라이드캠 구입후
기

```
<a ng-href="https://blog.naver.com/dogslife78/221769374691" class="desc_inner" target="_blank" bg-nclickf="{
  <strong class="title_post" ng-bind-html="post.noTagTitle || post.title">카메라 슬라이더 멋진 영상을 위한 +
</a>
<a ng-href="https://blog.naver.com/dogslife78/221769374691" class="text" ng-bind-html="post.briefContents ||
.v>
class="comments">
<!-- ngIf: post.sympathyEnable --><span class="like" ng-if="post.sympathyEnable">공감 <em>15</em></span><!--
<!-- ngIf: post.commentEnable --><span class="reply" ng-if="post.commentEnable">댓글 <em>2</em></span><!-- e
```

Selenium을 이용한 Crawling: click() 함수

● 기존에 실패했던 페이지들 재시도

- <http://147.46.178.16:33333/javascript.html> Javascript가 있던 페이지 방문
- Selector로 "로또번호 생성하기" 링크 골라내기
 - find_elements_by_css_selector 함수 = Soup변수.select("selector")
 - find_element_by_css_selector 함수 = Soup변수.select_one("selector")
 - body a

```
from selenium import webdriver
browser = webdriver.Chrome("./chromedriver")
browser.get("http://147.46.178.16:33333/javascript.html")
import time
time.sleep(1)
browser.find_element_by_css_selector("body a").click()
print(browser.page_source)
```

```
<html><head></head><body><a href="javascript:openComments();">로또번호 생성하기</a>
<div id="commentArea">34<br>1<br>41<br>14<br>45<br>44<br>36<br></div>
```

로또번호 생성하기

- Javascript로 생성된 내용도 성공적으로 가져옴

34
1
41
14
45
44
36

Selenium을 이용한 Crawling

- 실습 : BeautifulSoup로 추출한 정보들 Selenium으로도 추출해보기
 - <http://147.46.178.16:33333/table.html>
 - <http://www.kpu.ac.kr/contents/main/cor/kcollege.html>
 - <http://www.kpu.ac.kr/contents/main/cor/sanhak.html>
 - <https://sports.news.naver.com/index.nhn>

Selenium을 이용한 Crawling

● 실습 : 멜론 top100: 순위, 타이틀, 가수

MelOn

주간인기상에서 이번주 인기곡을 확인하세요!

김광중 3. 실어제인2

LF mall 첫구매 전용 스타벅스 기프티콘 지급

멜론차트 최신음악 장르음악 멜론DJ 멜론TV 스타포스트 매거진 뮤직어워드 어학 마이뮤직


























TOP100 일간 주간 월간 시대

TOP100 ?

2022.01.14 07:00

서랍장 전체듣기 듣기 + 담기 다운로드 FLAC 선택

새로고침

순위	곡정보	앨범	좋아요	듣기	담기	다운	유비
1	 취중고백 김민석 (멜로망스)	취중고백	38,840				
2	 회전목마 (Feat. Zion.T, 원슈타... sokodomo	쇼미더머니 10 Episode 2	178,493				
3	 Counting Stars (Feat. Beenzino) BE'O (비오)	Counting Stars	139,194				
4	 리무진 (Feat. MINO) (Prod. G... BE'O (비오)	쇼미더머니 10 Episode 3	154,515				
5	 ELEVEN IVE (아이브)	ELEVEN	89,885				

Selenium을 이용한 Crawling

● 실습 : Selenium 사용 정보 추출

- 대상 : <https://news.naver.com/main/ranking/read.nhn?oid=011&aid=0003814563>
- 댓글 수 출력

서울경제

독감백신 사망 전국 14명 넘어서...당국 '인과권 변만

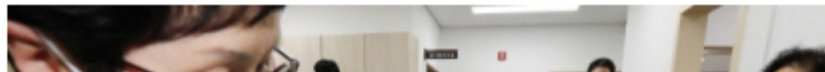
기사입력 2020.10.22. 오전 10:54 최종수정 2020.10.22. 오전 11:10 기사원문 스크랩 본문듣기 · 설

👍👎 1,168

💬 521

전국 곳곳 사망자 발생...현재 14명

명확한 사망 원인 몰라... 추측만 난무



Selenium을 이용한 Crawling: send_keys() 함수

● Selenium만의 기능들

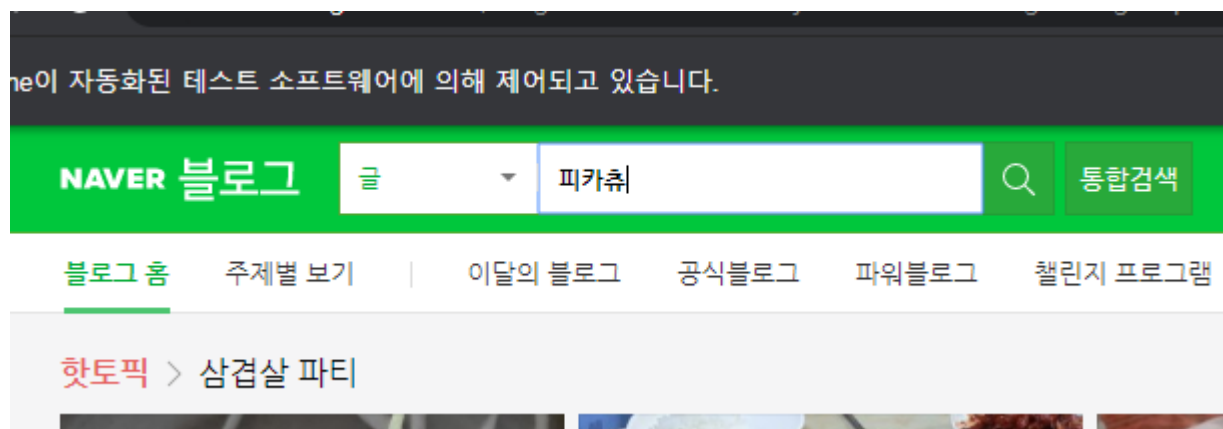
- `https://section.blog.naver.com/BlogHome.nhn?directoryNo=0¤tPage=1&groupId=0`
- `.send_keys("타이핑할 문구")` : 선택된 HTML 요소에 타이핑하기
 - 검색창을 선택한 후 피카츄 라고 타이핑하기

textbox =

```
browser.find_element_by_css_selector("div.area_dropdown + input.textbox")
```

```
textbox.send_keys("피카츄")
```

- 이를 사용하여 아이디/비밀번호 입력 가능



Selenium을 이용한 Crawling: click() 함수

● Selenium만의 기능들

- <https://section.blog.naver.com/BlogHome.nhn?directoryNo=0¤tPage=1&groupId=0>
- .click() : 선택된 HTML 요소를 클릭하기
 - 검색 버튼을 찾아서 클릭하기

```
searchButton = browser.find_element_by_css_selector("fieldset a.button.button_blog").click()
```



여러 페이지에서 정보 추출하기

● 주소를 변경하는 방법

- 페이지 숫자를 변경하기

- <https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=피카츄>
- for문을 통해 숫자를 바꿔가며 웹페이지를 방문

```
for i in range(1, 5):  
    print("https://section.blog.naver.com/Search/Post.nhn?pageNo=" +  
          str(i) + "&rangeType=ALL&orderBy=sim&keyword=피카츄");
```

```
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=피카츄  
https://section.blog.naver.com/Search/Post.nhn?pageNo=2&rangeType=ALL&orderBy=sim&keyword=피카츄  
https://section.blog.naver.com/Search/Post.nhn?pageNo=3&rangeType=ALL&orderBy=sim&keyword=피카츄  
https://section.blog.naver.com/Search/Post.nhn?pageNo=4&rangeType=ALL&orderBy=sim&keyword=피카츄
```

여러 페이지에서 정보 추출하기

● 주소를 변경하는 방법

- 키워드를 변경하기

- <https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=피카츄>
- 검색하고자 하는 키워드가 정해져 있을 때 유용함

```
keywords = ["피카츄", "라이츄", "파이리", "꼬부기"];  
for i in range(len(keywords)):
```

```
    print("https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=" +  
          keywords[i]);
```

```
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=피카츄  
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=라이츄  
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=파이리  
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=꼬부기
```

여러 페이지에서 정보 추출하기

● 주소를 변경하는 방법

- 페이지 숫자 & 키워드를 변경하기

- <https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=피카츄>
- for문을 중첩해서 사용

```
keywords = ["피카츄", "라이츄", "파이리", "꼬부기"];
for i in range(len(keywords)):
    for j in range(1, 4):
```

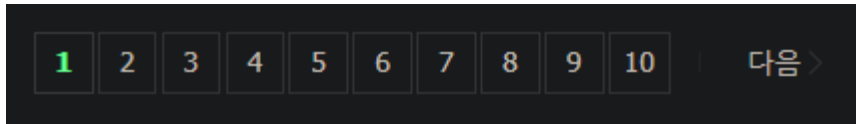
```
print("https://section.blog.naver.com/Search/Post.nhn?pageNo="
      + str(j) + "&rangeType=ALL&orderBy=sim&keyword=" +
      keywords[i]);
```

```
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=피카츄
https://section.blog.naver.com/Search/Post.nhn?pageNo=2&rangeType=ALL&orderBy=sim&keyword=피카츄
https://section.blog.naver.com/Search/Post.nhn?pageNo=3&rangeType=ALL&orderBy=sim&keyword=피카츄
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=라이츄
https://section.blog.naver.com/Search/Post.nhn?pageNo=2&rangeType=ALL&orderBy=sim&keyword=라이츄
https://section.blog.naver.com/Search/Post.nhn?pageNo=3&rangeType=ALL&orderBy=sim&keyword=라이츄
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=파이리
https://section.blog.naver.com/Search/Post.nhn?pageNo=2&rangeType=ALL&orderBy=sim&keyword=파이리
https://section.blog.naver.com/Search/Post.nhn?pageNo=3&rangeType=ALL&orderBy=sim&keyword=파이리
https://section.blog.naver.com/Search/Post.nhn?pageNo=1&rangeType=ALL&orderBy=sim&keyword=꼬부기
https://section.blog.naver.com/Search/Post.nhn?pageNo=2&rangeType=ALL&orderBy=sim&keyword=꼬부기
https://section.blog.naver.com/Search/Post.nhn?pageNo=3&rangeType=ALL&orderBy=sim&keyword=꼬부기
```

여러 페이지에서 정보 추출하기

● 주소를 변경하는 방법

- 다음 페이지 숫자 버튼을 (pagination) 클릭하기



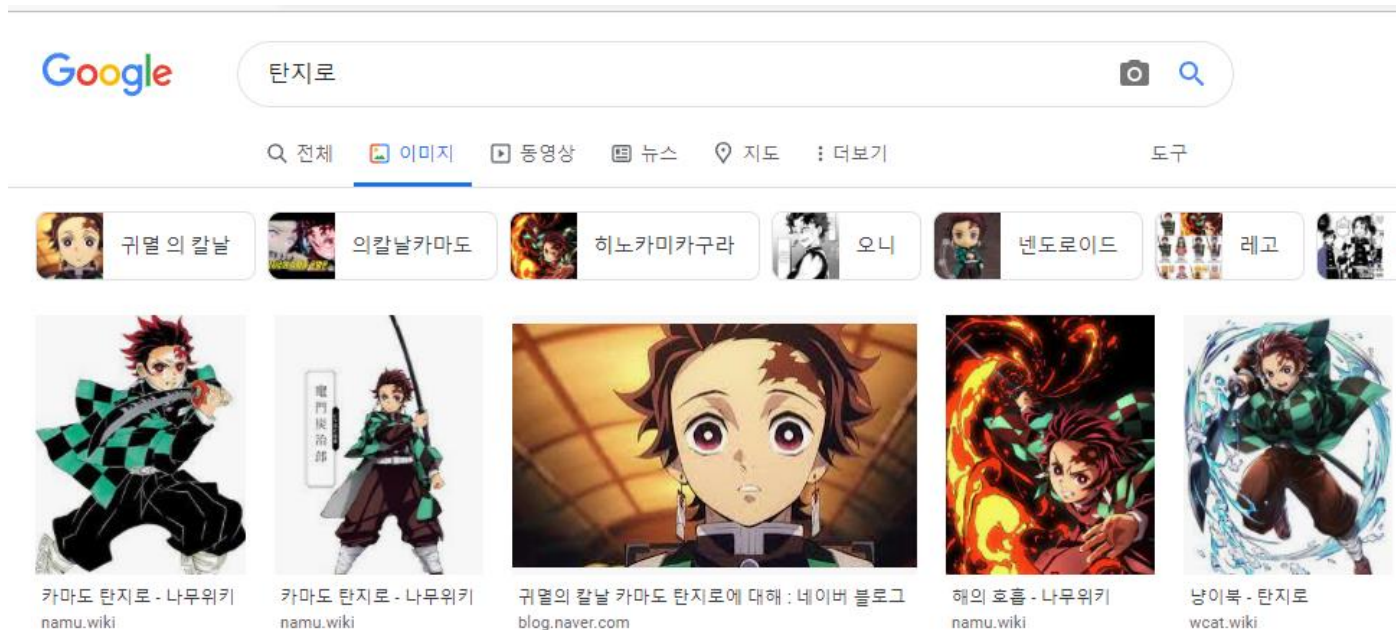
- 주소에 페이지 정보 등이 없을 경우 사용하게 됨
- if문을 통해 10페이지까지 갔을 경우 '다음' 버튼을 눌러주도록 해야함
- 마지막 페이지까지 갔을 때 멈춰주는 코드 필요

Selenium을 이용한 Google Image crawling

● 실습 : "탄지로" 구글 검색

- 대상 : "http://www.google.co.kr/imghp?hl=ko"
- 검색창에 "탄지로" 검색

```
driver = webdriver.Chrome()
driver.get('http://www.google.co.kr/imghp?hl=ko')
elem = driver.find_element_by_name("q")
elem.send_keys("탄지로")
elem.send_keys(Keys.RETURN)    #enter 키 치는거 자동화임
```



Selenium을 이용한 Google Image crawling

● 실습 : “탄지로” 구글 검색

- 대상 : “<http://www.google.co.kr/imghp?hl=ko>”
- 구글 이미지 검색의 문제는 scroll 후, “결과 더보기” 버튼을 클릭해야 한다!

```
SCROLL_PAUSE_TIME = 1
# Get scroll height
last_height = driver.execute_script("return document.body.scrollHeight")
while True:
    # Scroll down to bottom
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    # Wait to load page
    time.sleep(SCROLL_PAUSE_TIME)
    # Calculate new scroll height and compare with last scroll height
    new_height = driver.execute_script("return document.body.scrollHeight")
    if new_height == last_height:
        try:
            driver.find_element_by_css_selector(".mye4qd").click()
        except:
            break
    last_height = new_height
```

참조: 창 사이즈 읽어오기 [scrollHeight / clientHeight / scrollTop \(브라우저별 차이점\)](#) | 개발자 남인식 Lab. (naminsik.com)

[파이썬 셀레니움 이미지 크롤링으로 배우는 업무 자동화의 기초 - YouTube](#)

Selenium 설치 – Google Colab

● Selenium 실행시

- 다음 코드를

```
browser = webdriver.Chrome("./chromedriver.exe")
```

- 다음 코드로 바꿔서 실행해야함
 - 브라우저 창을 띄우지 말고 실행하라는 옵션임
 - PyCharm에서도 사용가능하며, 저사양 컴퓨터에서 유용함

```
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
browser = webdriver.Chrome("chromedriver", options=options)
```