

빅데이터 수집시스템 개발



동적 페이지의 웹 크롤링 실습

학습목표

- JavaScript에 의해서 생성되는 웹 콘텐츠의 실전 예제를 통해 스크래핑 할 수 있다.
- 브라우저를 기동시킨 상태와 브라우저를 기동시키지 않은 상태에서 웹 페이지의 렌더링 결과를 캡처할 수 있다.

학습내용

- 카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제
- 웹 페이지의 화면 캡처

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제



1 Selenium을 활용한 웹 크롤링과 스크래핑을 고려해야 하는 경우

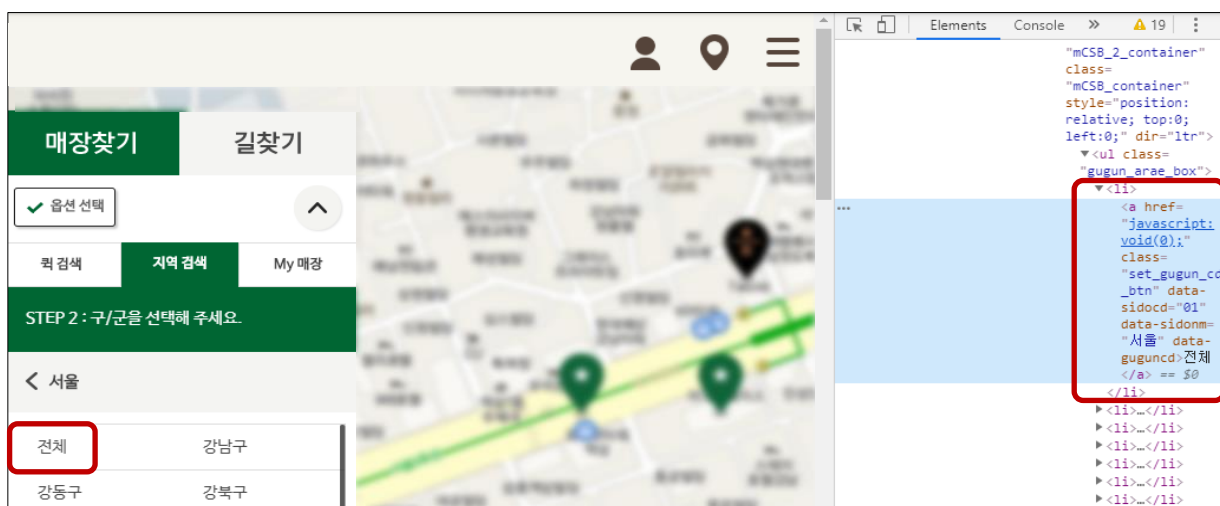
- 1 화면에 렌더링된 웹 페이지의 내용을 서버로부터 전송된 소스 코드에서 찾을 수 없는 경우
- 2 페이지 내의 링크를 클릭할 때 이동되는 페이지의 URL이 주소 필드에 출력되지 않는 경우
- 3 웹 페이지를 크롤링하기 전에 로그인 과정을 거쳐서 인증 처리를 해야 하는 경우
- 4 추출하려는 웹 페이지의 내용이 스크롤과 같은 이벤트가 발생해야 화면에 렌더링되는 경우
- 5 버튼을 클릭해야 웹 페이지의 콘텐츠가 출력되는 경우

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

2 관심 지역의 카페 매장 정보 읽어 오기

1 스크래핑 내용

1 전체의 WebElement 객체를 추출하여 클릭 이벤트 발생



```

s_link =
driver.find_element_by_css_selector("#mCSB_2_
container > ul > li:nth-child(1) > a")
s_link.click()

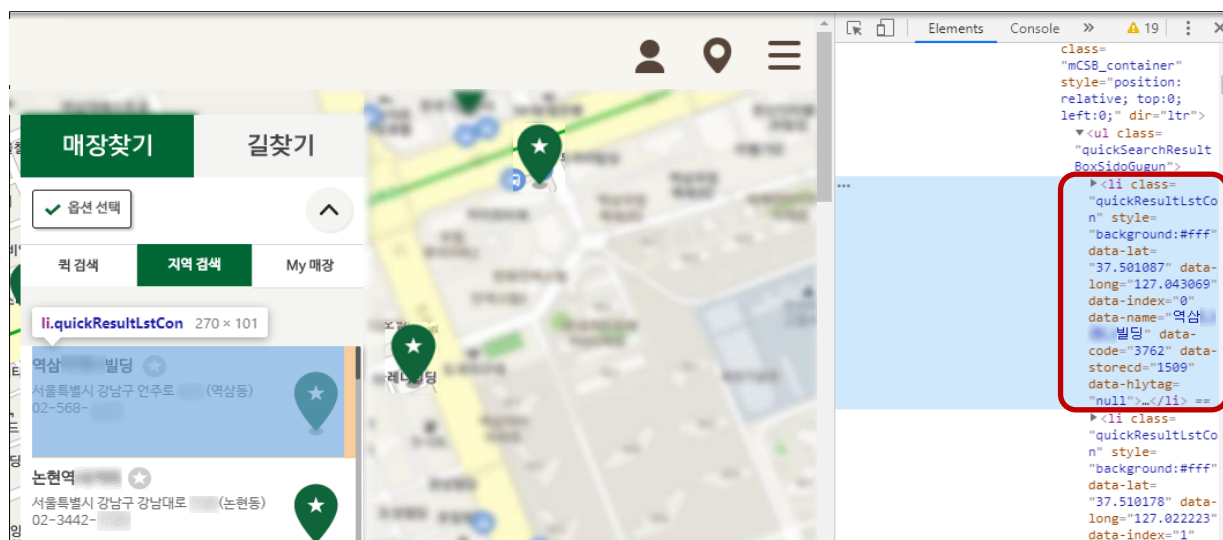
```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

2 관심 지역의 카페 매장 정보 읽어 오기

1 스크래핑 내용

- 매장 리스트가 출력되면 다음과 같은 CSS 선택자로 매장 리스트의 WebElement 객체를 리스트 객체로 추출



```
shopList =
driver.find_elements_by_css_selector("#mCSB_3_
container > ul > li")
```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

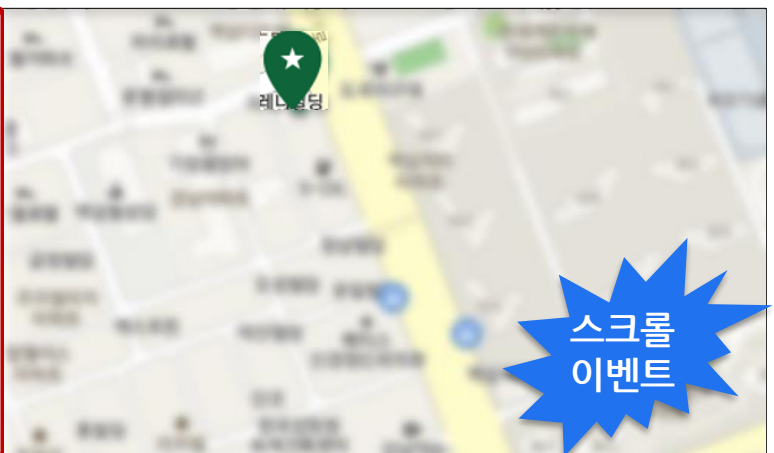
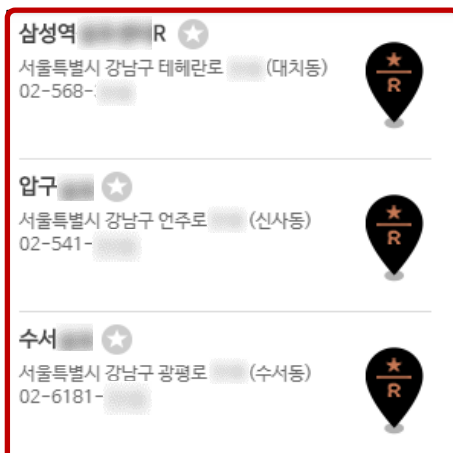
2 관심 지역의 카페 매장 정보 읽어 오기

1 스크래핑 내용

3 해당 뷰 내에서 스크롤 이벤트 강제 발생

- 기본적으로 3개의 매장 정보 출력
 ➔ 해당 뷰의 스크롤을 아래로 내리면 그 다음 3개의 매장 정보 출력
- 화면에 렌더링된 매장 리스트의 세 번째 항목에서 JavaScript 코드 실행

```
driver.execute_script("var su = arguments[0];
var dom=document.querySelectorAll
( '#mCSB_3_container > ul > li' ) [su];
dom.scrollToView();", count)
```



카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

2 관심 지역의 카페 매장 정보 읽어 오기

2 소스

```
#파일명 : exam7_1.py
from selenium import webdriver
driver = webdriver.Chrome('C:/Temp/chromedriver')
driver.implicitly_wait(3)
driver.get("https://www.is*****ks.co.kr/store/store_map.do")
import time
time.sleep(2)
loca = driver.find_element_by_class_name('loca_search')
loca.click()
time.sleep(2)
f_link = driver.find_element_by_css_selector("div.loca_step1_cont > ul > li:nth-child(1) > a")
f_link.click()
time.sleep(2)
s_link = driver.find_element_by_css_selector("#mCSB_2_container > ul > li:nth-child(1) > a")
s_link.click()
time.sleep(2)
shopList = driver.find_elements_by_css_selector("#mCSB_3_container > ul > li")
temp_list = []
time.sleep(3)
count = 0
total = len(shopList)
for shop in shopList :
    count += 1
    shoplatt = shop.get_attribute("data-lat")
    shoplont = shop.get_attribute("data-long")
```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

2 관심 지역의 카페 매장 정보 읽어 오기

2 소스

```
shopname = shop.find_element_by_tag_name("strong")
shopinfo = shop.find_element_by_tag_name("p")
splitinfo = shopinfo.text.split('\n')
if(len(splitinfo) == 2):
    addr = splitinfo[0]
    phonenum = splitinfo[1]
temp_list.append([shopname.text, shoplat, shoplong, addr, phonenum])
if count != total and count % 3 == 0:
    driver.execute_script("var su = arguments[0]; var dom=
    document.querySelectorAll(
    '#mCSB_3_container > ul > li')[su]; dom.scrollToView();", count)

for item in temp_list:
    print(item)
```


카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

2 관심 지역의 카페 매장 정보 읽어 오기

3 실행 결과

```

starbucks_shop.txt - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말
['역삼 빌딩', '37.5010', '127.0430', '서울특별시 강남구 언주로 (역삼동)', '02-568-']
['논현역', '37.5101', '127.0222', '서울특별시 강남구 강남대로 (논현동)', '02-3442-']
['국기원', '37.4995', '127.0314', '서울특별시 강남구 테헤란로 (역삼동)', '02-568-']
['대치 빌딩R', '37.4946', '127.0625', '서울특별시 강남구 남부순환로 (대치동)', '02-568-']
['삼성역 센터R', '37.5077', '127.0606', '서울특별시 강남구 테헤란로 (대치동)', '02-568-']
['압구', '37.52736', '127.0330', '서울특별시 강남구 언주로 (신사동)', '02-541-']
['수서', '37.4880', '127.1026', '서울특별시 강남구 광평로 (수서동)', '02-6181-']
['양재 빌딩R', '37.4851', '127.0366', '서울특별시 강남구 남부순환로 (도곡동)', '02-571-']
['선릉 빌딩R', '37.5053', '127.0504', '서울특별시 강남구 테헤란로 (삼성동)', '02-2051-']
['봉선정릉', '37.5112', '127.0484', '서울특별시 강남구 봉은사로 (삼성동)', '02-539-']
['강남', '37.5021', '127.0266', '서울특별시 강남구 봉은사로2길 (역삼동)', '02-557-']
['스타 R', '37.509', '127.0614', '서울특별시 강남구 영동대로 (삼성동)', '02-6002-']
['강남구', '37.5181', '127.0459', '서울특별시 강남구 학동로 (청담동)', '02-514-']

```

```

starbucks_shop.txt - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말
['순 구', '37.5333', '127.005', '서울특별시 용산구 대사관로 (한남동)', '02-758-']
['숙', '37.544604', '126.96722', '서울특별시 용산구 청파로47길 (청파동3가)', '02-758-']
['동빙', '37.52898', '126.9917', '서울특별시 용산구 장문로 (동빙 동)', '02-798-']
['숙 역', '37.5441', '126.9718', '서울특별시 용산구 한강대로 (갈월동)', '02-758-']
['용산 물', '37.52885', '126.964044699999', '서울특별시 용산구 한강대로23길 55', '02-2012-']
['동부 동', '37.521', '126.969', '서울특별시 용산구 이촌로 (이촌동)', '02-758-']
['상', '37.5978', '127.0925', '서울특별시 중랑구 상봉로 (상 동, 상 주상복합)', '02-438-5']
['중 역', '37.59303', '127.074735799999', '서울특별시 중랑구 망우로30길 (상봉동)', '02-758-']
['중 청', '37.605389', '127.09575', '서울특별시 중랑구 신내로', '02-758-']
['사 역', '37.5795', '127.0879', '서울특별시 중랑구 면목로', '02-758-']
['상', '37.596', '127.086', '서울특별시 중랑구 망우로, 3,4번지 (상봉동)', '02-758-']
['묵동', '37.6134', '127.0774', '서울특별시 중랑구 농밀로, 묵동 B1층 (묵동)', '02-758-']

```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

1 도서 댓글 스크래핑

1 요약 댓글과 전체 댓글의 두 가지 타입 댓글로 구성

| 전체 리뷰(97) | 포토 리뷰(9) | 스타블로거 리뷰(45) |
|--|----------|--------------|
| << 1 2 3 4 5 6 7 8 9 10 >> 구매리뷰 최근순 추천순 별점순 | | |
| <div>구매</div> <div>지금을 보는 눈</div> <div> 내용 ★★★★★ 편집/디자인 ★★★★★ 2050-06-17 </div> <p>내가 살고 있는 지금 사회 문제는 무엇인지, 무엇을 고민하고 생각해봐야 많다. 바로 나와 내 가족, 내가 먹고 사는 일에 직결이 되는 일임에도 불구하고 다가오는 자동차와 같은 일들. 이것을 잡아 타면 남들보다 빠르고 편안하게 먼 길을 이를 미처 알지 못하고 있으면 놓치는 것은 물론...</p> <div> 이 리뷰가 도움이 되었나요? ♡0 댓글 0 > </div> <div> ‘펼쳐보기’ 링크를 클릭해야 전체 댓글 내용을 볼 수 있음 </div> <div> 펼쳐보기 </div> | | |

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

1 도서 댓글 스크래핑

2 '펼쳐보기'의 WebElement 객체들을 추출하여 각각 클릭 이벤트 발생

리뷰

지금들 보는 눈

내용 ★★★★★ 편집/디자인 ★★★★★ | 2050-06-17

내가 살고 있는 지금 사회 문제는 무엇인지, 무엇을 고민하고 생각해봐야 하는지 놓치는 경우가 많다. 바로 나와 내 가족, 내가 먹고 사는 일에 직결이 되는 일임에도 불구하고 전조등과 소리 없이 다가오는 자동차와 같은 일들. 이것을 잡아 타면 남들보다 빠르고 편안하게 먼 길을 갈 수 있지만, 이를 미처 알지 못하고 있으면 놓치는 것은 물론...

이 리뷰가 도움이 되었나요? ♡ 0 댓글 0 >

펼쳐보기 ▾

```

<div class="reviewContentList">...</div>
<div class="btn_halfMore">
  <a href="javascript:void(0);" onclick="toggleInfoSubSet(this)"
    >== $0
    <em class="txt txt_open">펼쳐보기</em>
    <em class="txt txt_close">접어보기</em>
    <em class="bgVUI ico_arr"></em>
  </a>
</div>
<!-- ##### 리뷰 하나 반복 끝 ##### -->
<div class="review_sort sortBot">...</div>
<script type="text/javascript">...</script>
</div>
<input type="hidden" id="hdnStartScore">
<input type="hidden" id="hdnEndScore">
<input type="hidden" id="hdnScoreYn">

```

- 댓글 한 페이지당 5개의 댓글이 있음

```

readLinks =
driver.find_elements_by_css_selector('#infoSet_
reviewContentList div.btn_halfMore > a') # 펼쳐보기
for readlink in readLinks :
    readlink.click()
    time.sleep(1)

```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

1 도서 댓글 스크래핑

3 전체 댓글이 보여지면 댓글 추출

리뷰

지금을 보는 눈

내용 ★★★★★ 편집/디자인 ★★★★★ | 2050-06-17

내가 살고 있는 지금 사회 문제는 무엇인지, 무엇을 고민하고 생각해봐야 하는지 놓치는 경우가 많다. 바로 나와 내 가족, 내가 먹고 사는 일에 직결이 되는 일임에도 불구하고 전조등과 소리 없이 다가오는 자동차와 같은 일들. 이것을 잡아 타면 남들보다 빠르고 편안하게 먼 길을 갈 수 있지만, 이를 미처 알지 못하고 있으면 놓치는 것은 물론, 어물쩍 하다가는 차에 치여 치명적인 부상과 뒤쳐짐에 직면할지도 모른다. 새로운 사회 편은 내게 빛 없이 소리 없이 조용히 다가오는 그런 자동차처럼 느껴지는 편이었다. 이 책이 사랑받는 이유는 우리가 현재와 미래를 살아 가는데 있어서 꼭 한번 이상은 생각해보고 준비해야 하는 일들을 쉽고 재미있게 풀어쓴 까닭이 아닐까 생각한다.

이 리뷰가 도움이 되었나요?

♡ 0

댓글 0 >

<< 1 2 3 4 5 6 7 8 9 10 >>

접어보기 ▾

‘펼쳐보기’
링크
문자열



‘접어보기’
링크 문자열로
변경



댓글 추출

```

reviewList =
driver.find_elements_by_css_selector('#infoSet_
reviewContentList div.reviewInfoBot.origin
div.review_cont')
    
```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

1 도서 댓글 스크래핑

4 여러 페이지의 댓글 내용을 읽어오기 위해 다음 페이지 링크 클릭

| 전체 리뷰(97) | 포토 리뷰(9) | 스타블로거 리뷰(45) |
|---|----------|--------------|
| <div> <div> << 1 2 3 4 5 6 7 8 9 10 >> </div> <div> 구매투표 </div> <div> 최근순 </div> <div> 추천순 </div> <div> 별점순 </div> </div> | | |

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

1 도서 댓글 스크래핑

5 페이지 이동을 위해 각 페이지 링크 숫자에 적용되는 URL 규칙 파악

| | |
|--|---|
| <div>리뷰</div> <p>지금들 보는 눈</p> <p>내용 ★★★★★ 편집/디자인 ★★★★★</p> <p>내가 살고 있는 지금 사회 문제는 무엇인지, 무엇 바로 나와 내 가족, 내가 먹고 사는 일에 직결이 자동차와 같은 일들. 이것을 잡아 타면 남들보다 알지 못하고 있으면 놓치는 것은 물론, 어물쩍하 직면할지도 모른다. 새로운 사회 편은 내게 빛 느껴지는 편이었다. 이 책이 사랑받는 이유는 우 이상은 생각해보고 준비해야 하는 일들을 쉽고</p> <p>이 리뷰가 도움이 되었나요? <input type="button" value="0"/> 댓글 0</p> <p>« < 1 2 3 4 5 6 7 8 9 10 > »</p> | <pre> 이전 <strong class="num">1 "" 2 == \$0 3 <a href="/Product/communityModules/GoodsReviewList/ </pre> |
|--|---|

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

1 도서 댓글 스크래핑

- 5 페이지 이동을 위해 각 페이지 링크 숫자에 적용되는 URL 규칙 파악

```
#infofet_reviewContentList >
div.review_sort.sortBot > div.review_sortLft > div >
a:nth-child(4)
```

```
#infofet_reviewContentList >
div.review_sort.sortBot > div.review_sortLft > div >
a:nth-child(5)
```

```
#infofet_reviewContentList >
div.review_sort.sortBot > div.review_sortLft > div >
a:nth-child(11)
```

```
#infofet_reviewContentList >
div.review_sort.sortBot > div.review_sortLft > div >
a:nth-child(12)
```

```
#infofet_reviewContentList >
div.review_sort.sortBot > div.review_sortLft > div >
a.bgYUI.next
```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

2 소스

```
#파일명 : exam7_2.py
from selenium import webdriver

driver = webdriver.Chrome('C:/Temp/chromedriver')
driver.implicitly_wait(3)

driver.get("http://www.y***4.com/Product/goods/40936880")
import time
time.sleep(2)
readLinks = driver.find_elements_by_css_selector
('#infoSet_reviewContentList div.btn_halfMore > a') # 펼쳐보기

for readlink in readLinks :
    readlink.click()
    time.sleep(1)
reviewList = driver.find_elements_by_css_selector
('#infoSet_reviewContentList div.reviewInfoBot.origin div.review_cont')

temp_list = []
time.sleep(3)
for review in reviewList :
    temp_list.append(review.text)
stopFlag = False
while True :
    for n in range(4, 13) :
        linkurl = '#infoSet_reviewContentList > div.review_sort.sortBot >
div.review_sortLft > div > a:nth-child('+str(n)+' )'
        linkNum = driver.find_element_by_css_selector(linkurl)
        linkNum.click()
        time.sleep(3)
    readLinks = driver.find_elements_by_css_selector
        ('#infoSet_reviewContentList div.btn_halfMore > a')
```


카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

2 소스



```
for readlink in readLinks :
    #readlink.click()
    driver.execute_script("arguments[0].click();", readlink)
    time.sleep(3)
```

```
reviewList = driver.find_elements_by_css_selector
('#infoSet_reviewContentList div.reviewInfoBot.origin
div.review_cont')
time.sleep(2)
```

```
for review in reviewList :
    temp_list.append(review.text)
```

```
if len(reviewList) < 5 :
    stopFlag = True
    break
```

```
if stopFlag == True :
    break
```

```
nextPage = '#infoSet_reviewContentList > div.review_sort.sortBot >
div.review_sortLft > div >
```

```
a.bgYUI.next'
```

```
linkNum = driver.find_element_by_css_selector(nextPage)
linkNum.click()
time.sleep(1)
```

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제



3 서점 사이트의 도서 댓글 읽어 오기

2 소스

```
for item in temp_list :  
    print(item)
```

```
wfile = open("c:/Temp/서점file.txt","w")  
wfile.writelines(temp_list)  
wfile.close()
```



카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제



3 서점 사이트의 도서 댓글 읽어 오기

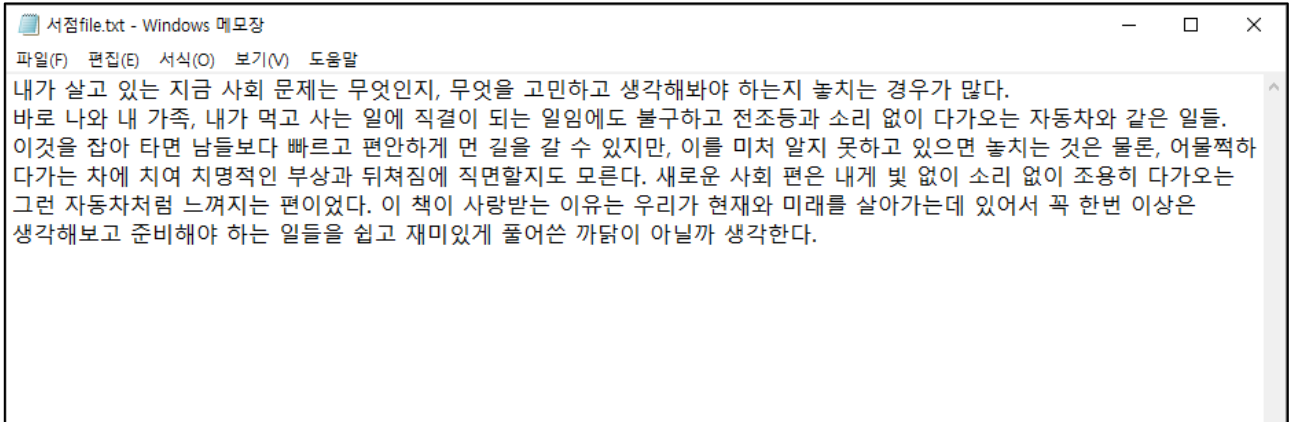
3 실행 결과 콘솔 화면

내가 살고 있는 지금 사회 문제는 무엇인지, 무엇을 고민하고 생각해봐야 하는지 놓치는 경우가 많다. 바로 나와 내 가족, 내가 먹고 사는 일에 직결이 되는 일임에도 불구하고 전조등과 소리 없이 다가오는 자동차와 같은 일들. 이것을 잡아 타면 남들보다 빠르고 편안하게 먼 길을 갈 수 있지만, 이를 미처 알지 못하고 있으면 놓치는 것은 물론, 어물쩍하다가 차에 치여 치명적인 부상과 뒤통수에 직면할지도 모른다. 새로운 사회 편은 내게 빛 없이 소리 없이 조용히 다가오는 그런 자동차처럼 느껴지는 편이었다. 이 책이 사랑받는 이유는 우리가 현재와 미래를 살아가는데 있어서 꼭 한번 이상은 생각해보고 준비해야 하는 일들을 쉽고 재미있게 풀어쓴 까닭이 아닐까 생각한다.

카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제

3 서점 사이트의 도서 댓글 읽어 오기

4 파일 저장 화면



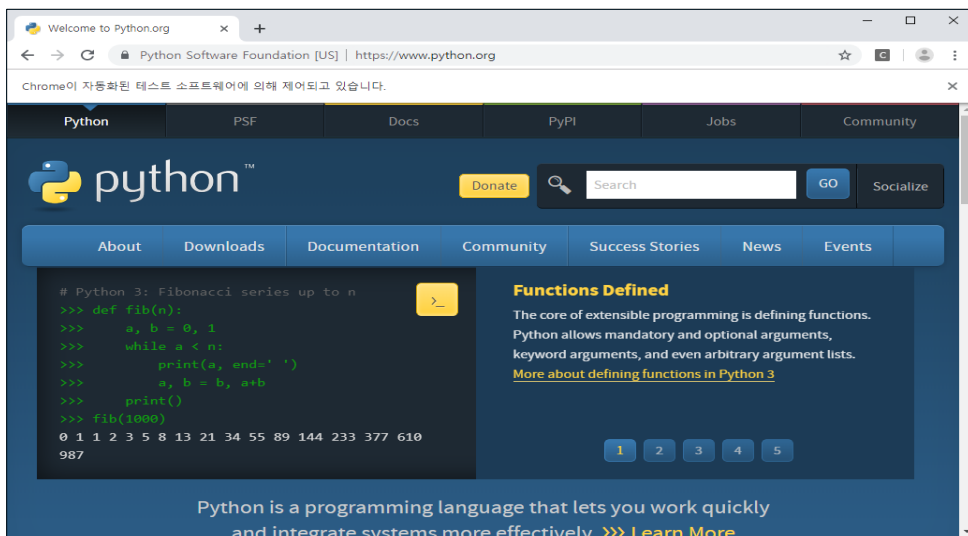
웹 페이지의 화면 캡처

1 웹 페이지 화면을 캡처하여 이미지 저장하기

1 Selenium에서 캡처 파일 이미지로 저장

- Selenium에 의해 기동된 크롬 브라우저의 페이지 렌더링 화면을 캡처하여 이미지 파일로 저장

```
driver.get_screenshot_as_file('c:/Temp/python_main.png')
```



[출처 : www.python.org]

웹 페이지의 화면 캡처

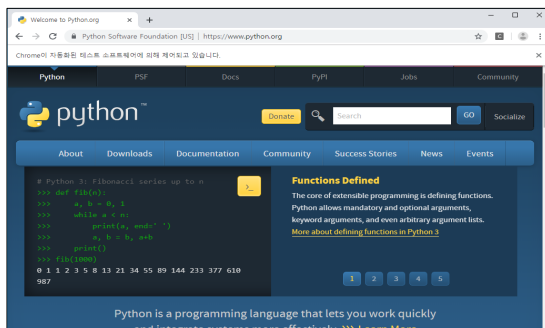
1 웹 페이지 화면을 캡처하여 이미지 저장하기

2 소스

```
#파일명 : exam7_3.py
from selenium import webdriver

driver = webdriver.Chrome('C:/Temp/chromedriver')

driver.get('http://www.python.org/')
driver.implicitly_wait(3)
driver.get_screenshot_as_file('c:/Temp/python_main.png')
print('캡처 저장 완료')
import time
time.sleep(2)
driver.quit()
```



캡처



[출처 : www.python.org]

웹 페이지의 화면 캡처



1 웹 페이지 화면을 캡처하여 이미지 저장하기

3 실행 결과 콘솔 화면

```
PS C:\example\myvscode> & C:/Users/UNICO/Anaconda3/python.exe c:/example/myvscode/unit7/exam7_3.py
```

```
DevTools listening on ws://127.0.0.1:56345/devtools/browser/e754b8ab-3ff5-45fd-9f00-9e6323d5e8bc
```

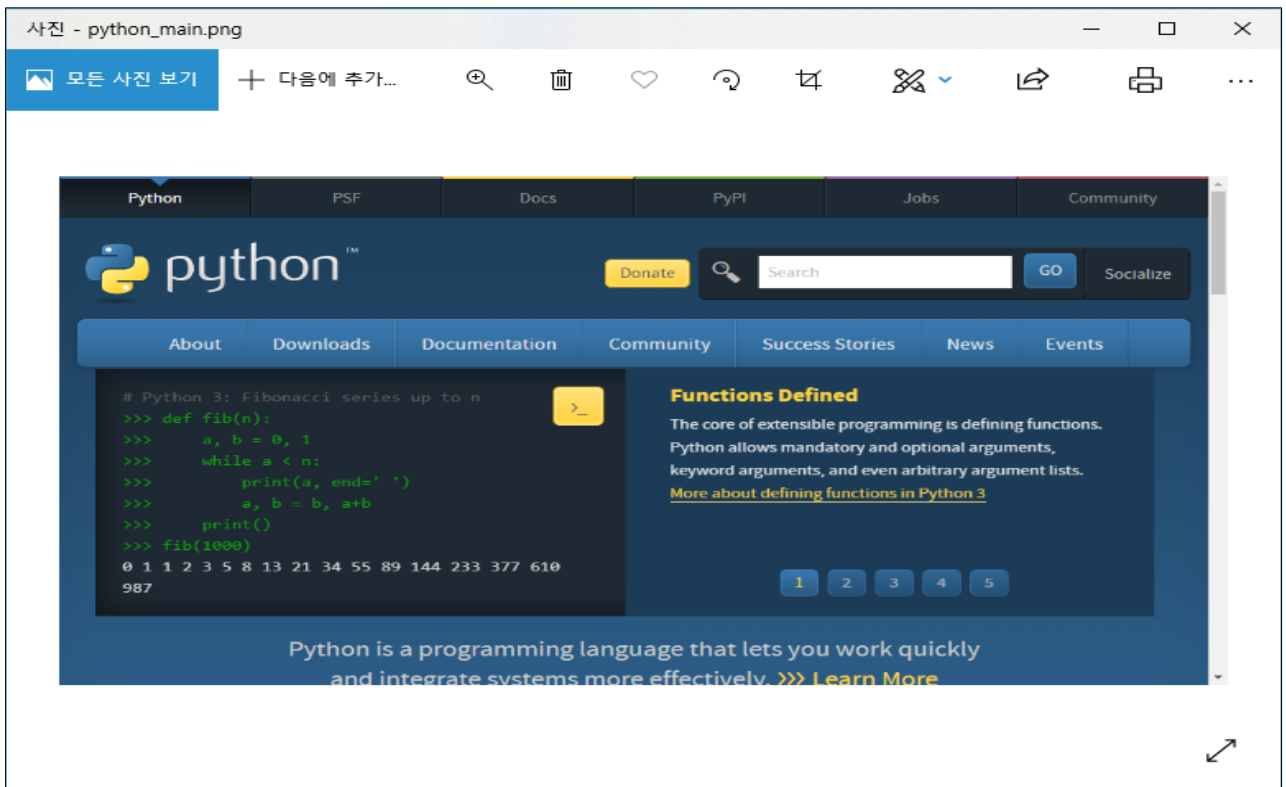
```
KLIB_SelfTest return : KLR_OK
```

```
캡처 저장 완료
```

웹 페이지의 화면 캡처

1 웹 페이지 화면을 캡처하여 이미지 저장하기

4 실행 결과 이미지 파일 화면



[출처 : www.python.org]

웹 페이지의 화면 캡처



2 웹 브라우저를 기동시키지 않고 웹 페이지 화면을 캡처하여 이미지 저장하기

1 Headless 모드

Headless 모드

- 브라우저 창을 실제 운영체제의 '창'으로 띄우지 않고 웹 페이지의 화면을 그려주는 작업(렌더링)을 가상으로 진행해주는 방법
- 실제 브라우저와 동일하게 동작하지만 창은 띄지 않는 방식으로 동작 가능

윈도우 기준
크롬 59 버전

맥/리눅스 기준
크롬 60 버전

Headless Mode 정식 추가

- 브라우저가 최신이라면 크롬의 Headless모드를 쉽게 이용 가능

웹 페이지의 화면 캡처



2 웹 브라우저를 기동시키지 않고 웹 페이지 화면을 캡처하여 이미지 저장하기

1 Headless 모드

- 다음 코드 사용

```
options = webdriver.ChromeOptions()
options.add_argument('headless')
options.add_argument('window-size=1920x1080')
driver = webdriver.Chrome('C:/Temp/chromedriver',
options=options)
```

웹 페이지의 화면 캡처

2 웹 브라우저를 기동시키지 않고 웹 페이지 화면을 캡처하여 이미지 저장하기

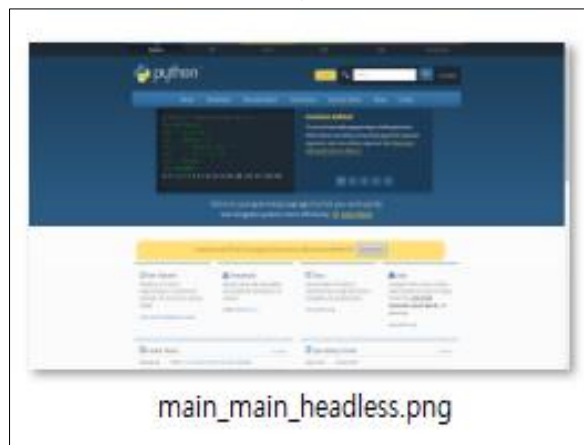
2 소스

```
#파일명 : exam7_4.py
from selenium import webdriver
options = webdriver.ChromeOptions()
options.add_argument('headless')
options.add_argument('window-size=1920x1080')
driver = webdriver.Chrome('C:/Temp/chromedriver', options=options)

driver.get('http://www.python.org/')
driver.implicitly_wait(3)
driver.get_screenshot_as_file('c:/Temp/main_main_headless.png')
print('캡처 저장 완료')
import time
time.sleep(2)
driver.quit()
```

브라우저 기동 없이

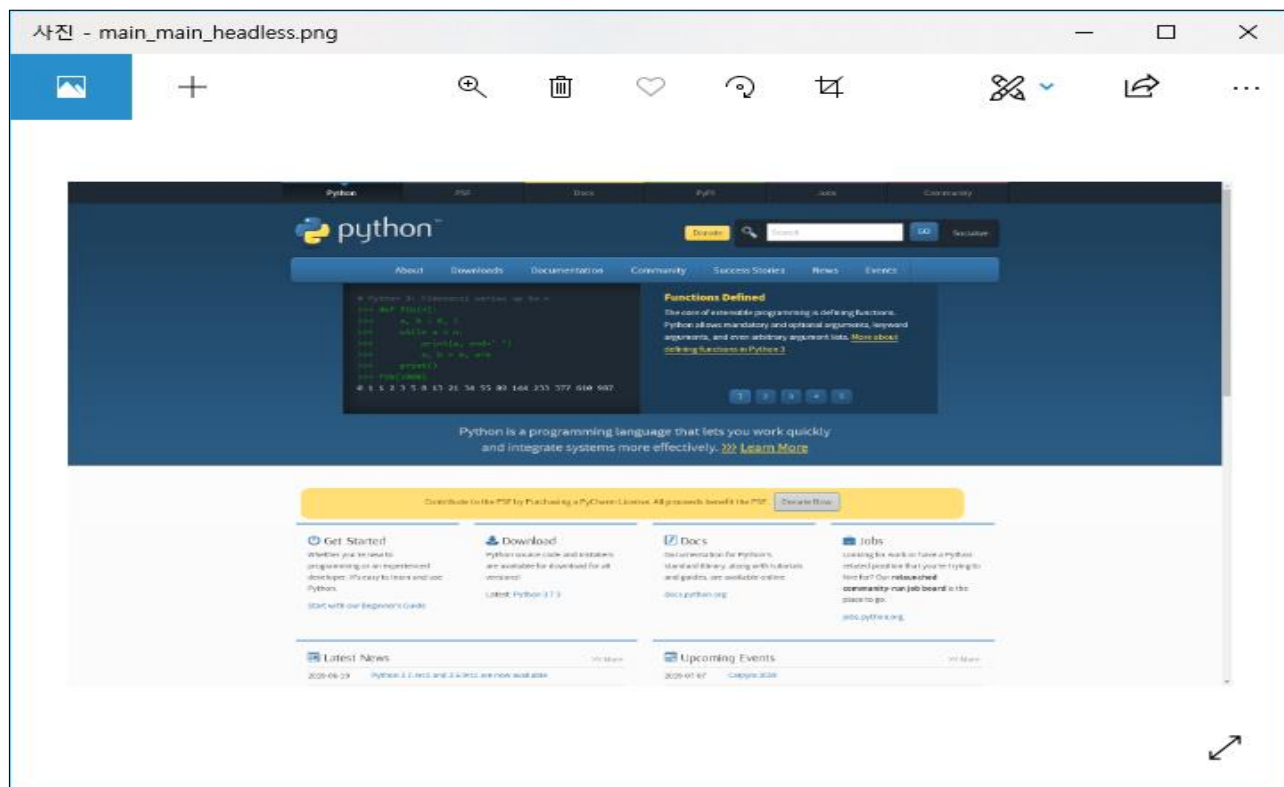
캡처



웹 페이지의 화면 캡처

2 웹 브라우저를 기동시키지 않고 웹 페이지 화면을 캡처하여 이미지 저장하기

3 실행 결과 이미지 파일 화면



학습정리

1. 카페 및 서점 동적 웹 페이지 크롤링 및 스크래핑 예제



- 동적 웹 페이지 : 태그의 콘텐츠 즉, 웹 페이지의 내용을 브라우저에 함께 전송된 JavaScript 코드의 수행 결과로 생성시키는 페이지
- 태그에서 강제로 클릭 이벤트를 발생시킬 때 WebElement 객체의 click() 메서드를 호출

학습정리

2. 웹 페이지의 화면 캡처



- Selenium에 의해 제어되는 크롬 브라우저에 렌더링된 웹 페이지의 화면을 캡처하여 파일로 저장할 때 WebDriver 객체의 `get_screenshot_as_file()` 메서드 사용
- Headless 모드
 - 브라우저 창을 실제 운영체제의 '창'으로 띄우지 않고 웹 페이지의 화면을 그려주는 작업(렌더링)을 가상으로 진행해주는 방법
 - 실제 브라우저와 동일하게 동작하지만 창은 띄지 않는 방식으로 동작