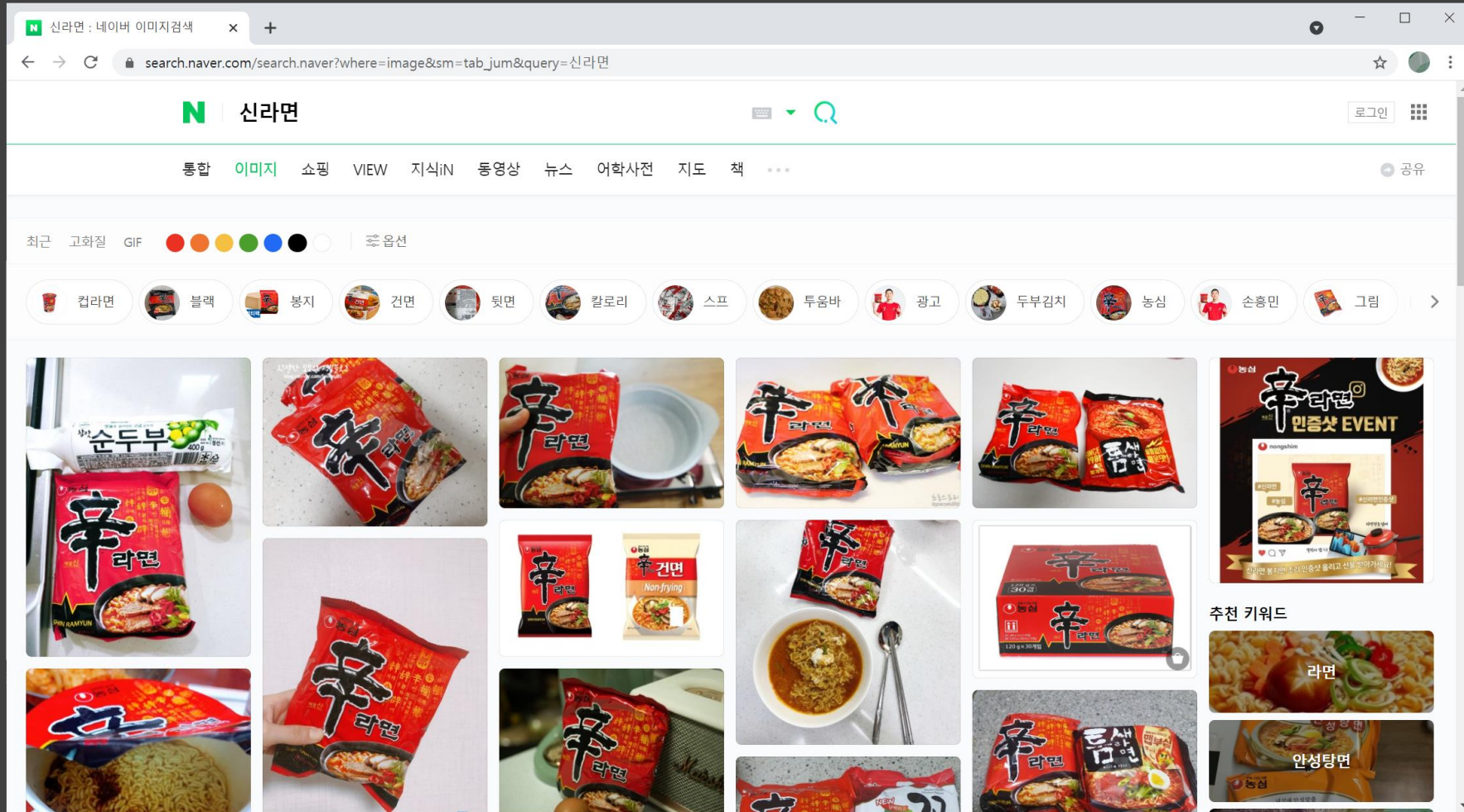


제 14 강 Web Scraping (3)

과업 #3.

신라면 이미지 자동 다운로드



```
01a4ecce5c96b13d3230451e&g=663960999185414" class="link_thumb _imageBo
x _infoBox" title="[라면추천] 모두가 알고있고 모두가 맛있어하는 신기한 얼
큰라면 신라면 !!!" role="button" aria-pressed="false">
 == $0
<i class="spimg ico_selected"></i>
::after
```

```
keyword = '신라면'
url = 'https://search.naver.com/search.naver?where=image&sm=tab_jum&query=' + keyword

p = sync_playwright().start()
browser = p.chromium.launch(headless=False).new_context(
    user_agent='Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/109.0.0.0 Safari/537.36'
)
page = browser.new_page()
page.goto(url)

soup = BeautifulSoup(page.content(), 'lxml')

elms = soup.select('img._image._listImage')
for n, e in enumerate(elms):
    print(n, e['src'], e)
```

로딩 늦어지는 리소스에 대한 처리

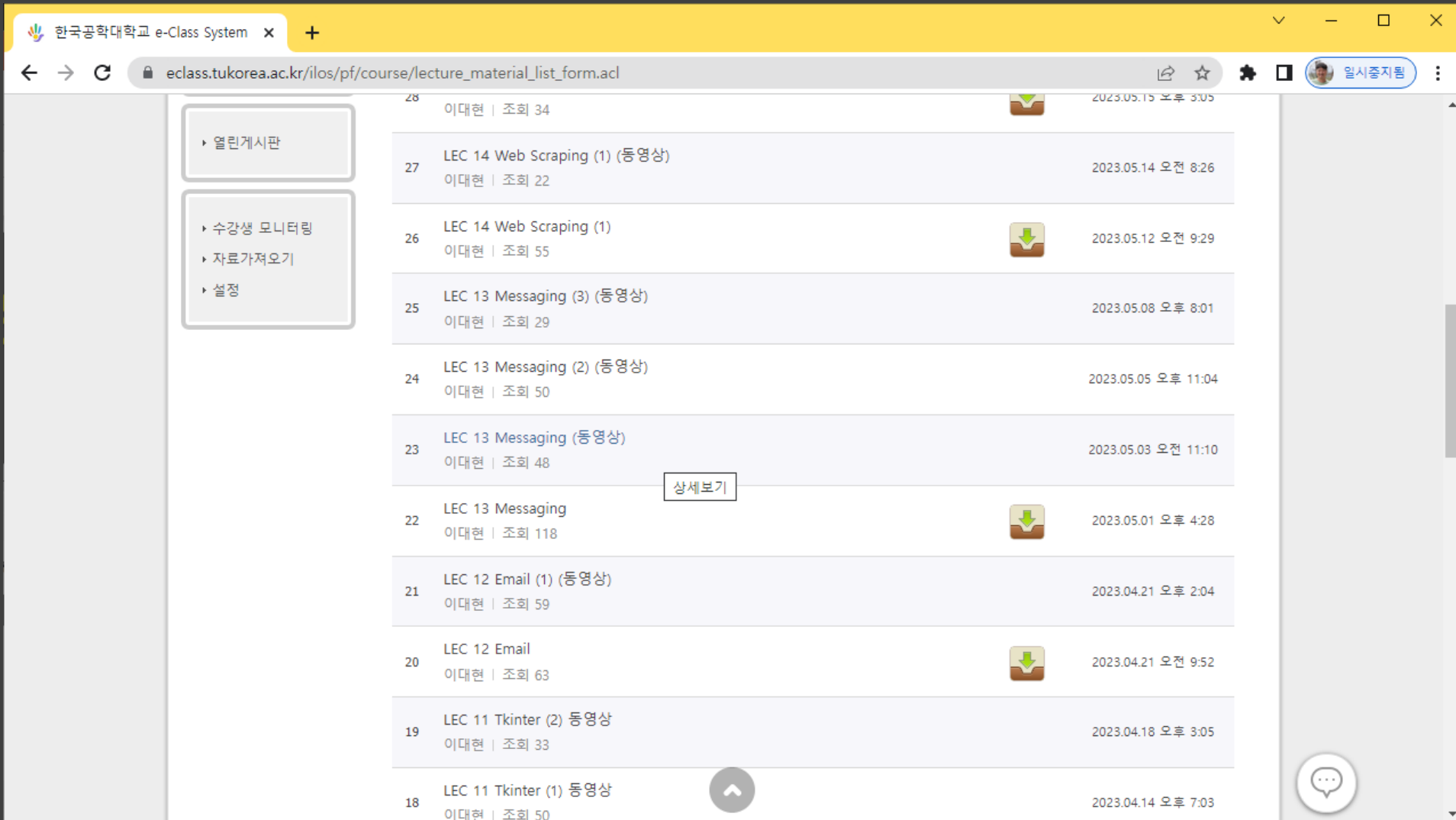
```
for n, e in enumerate(elms):  
    caption = e['alt']  
    image_url = e.get('data-lazy-src', e['src'])  
    print(f'{n} {caption} {image_url}')
```

requests.get 를 이용한 다운로드

```
def download_image(image_url):
    r = requests.get(image_url)
    r.raise_for_status()
    if r.headers['Content-Type'] == 'image/jpeg':
        fname = f'img_{uuid.uuid4().int}.jpg'
        with open(fname, 'wb') as wf:
            print(f'Downloading {fname} .....')
            wf.write(r.content)

for e in elms:
    image_url = e.get('data-lazy-src', e['src'])
    download_image(image_url)
```

과업 #4. eclass 에서 강의 자료 일괄 다운로드



The screenshot shows the 'e-Class System' interface of TUKorea University. The browser address bar displays 'eclass.tukorea.ac.kr/ilos/pf/course/lecture_material_list_form.acl'. On the left, a sidebar contains navigation links: '열린게시판', '수강생 모니터링', '자료가져오기', and '설정'. The main content area is a table listing lecture materials, with columns for item number, title, author, and date. A '상세보기' button is visible next to item 23. Download icons are present for items 26, 22, 20, and 19.

번호	제목	작성자	작성일
28	이대현 조회 34		2023.05.15 오후 3:05
27	LEC 14 Web Scrapping (1) (동영상)	이대현 조회 22	2023.05.14 오전 8:26
26	LEC 14 Web Scrapping (1)	이대현 조회 55	2023.05.12 오전 9:29
25	LEC 13 Messaging (3) (동영상)	이대현 조회 29	2023.05.08 오후 8:01
24	LEC 13 Messaging (2) (동영상)	이대현 조회 50	2023.05.05 오후 11:04
23	LEC 13 Messaging (동영상)	이대현 조회 48	2023.05.03 오전 11:10
22	LEC 13 Messaging	이대현 조회 118	2023.05.01 오후 4:28
21	LEC 12 Email (1) (동영상)	이대현 조회 59	2023.04.21 오후 2:04
20	LEC 12 Email	이대현 조회 63	2023.04.21 오전 9:52
19	LEC 11 Tkinter (2) 동영상	이대현 조회 33	2023.04.18 오후 3:05
18	LEC 11 Tkinter (1) 동영상	이대현 조회 50	2023.04.14 오후 7:03

로그인

```
from playwright.sync_api import sync_playwright
import time

url = 'https://eclass.tukorea.ac.kr/ilos/main/member/login_form.ac1'
p = sync_playwright().start()

browser = p.chromium.launch(headless=False).new_context(
    viewport={'width': 1920, 'height': 1024}
)

page = browser.new_page()
page.goto(url)

page.locator('input[name="usr_id"]').fill('*****')
page.locator('input[name="usr_pwd"]').fill('*****')
page.locator('#login_btn').click()
```


강의 자료 페이지 이동

```
page.locator('em[title="스크립트언어"]').click()  
page.locator('#menu_lecture_material').click()  
page.reload()
```

개별 게시물 이동

```
elms = page.locator('a.site-link').all()
text_list = [e.text_content() for e in elms]
for t in text_list[:3]: # 상위 3개만
    page.get_by_text(t).click()
    time.sleep(1)
    page.goto('https://eclass.tukorea.ac.kr/ilos/pf/course/lecture_material_list_form.ac1')
```

자동화 추출 : playwright codegen

The screenshot shows a web browser window displaying the e-Class System interface. The URL is `eclass.tukorea.ac.kr/ilos/pf/course/lecture_material_view_form.acl?ARTL_NUM=4463200&SCH_KEY=&SCH_VALUE=&display=1&start=1`. The interface includes a sidebar with a course list and a main content area showing lecture materials for "LEC 13 Messaging".

Overlaid on the browser window is the Playwright Inspector window, which displays the recorded code for the current session. The code is as follows:

```
37 with page.expect_download() as download2_info:
38     page.get_by_role("link", name="LEC12 - Email.pdf (1.3MB)").click()
39     download2 = download2_info.value
40     page.locator("#menu_lecture_material").click()
41     page.get_by_text("LEC 13 Messaging (2) (동영상) 이대현 조회 50").click()
42     page.locator("#menu_lecture_material").click()
43     page.get_by_text("LEC 7 8 동영상 이대현 조회 67").click()
44     page.locator("#menu_lecture_material").click()
45     page.get_by_text("LEC 12 Email 이대현 조회 63").click()
46     with page.expect_download() as download3_info:
47         page.get_by_role("link", name="LEC12 - Email.pdf (1.3MB)").click()
48         download3 = download3_info.value
49         page.locator("#menu_lecture_material").click()
50         page.get_by_text("LEC 14 교수 참고 자료 이대현 조회 0").click()
51         with page.expect_download() as download4_info:
52             page.get_by_role("link", name="2023_Lab_WebScraping.py (7.8KB)").cl
53             download4 = download4_info.value
54             page.locator("html").click()
55             page.locator("#menu_lecture_material").click()
56             page.get_by_text("LEC 13 Messaging 이대현 조회 118").click()
57             with page.expect_download() as download5_info:
58                 page.get_by_role("link", name="LEC13 - Messaging.pdf (4.6MB)").clid
59                 download5 = download5_info.value
60
61 # -----
62 context.close()
63 browser.close()
64
65
66 with sync_playwright() as playwright:
67     run(playwright)
68
```

At the bottom of the browser window, a console log entry is visible:

```
get_by_text("본문내용 바로가기 상단메뉴 바로가기 왼쪽메뉴 바로가기 한국어 English 로그인 이대현 66 20 23 로그아웃 Back to the top ")
```

파일 다운로드

```
elms = page.locator('a.site-link').all()
text_list = [e.text_content() for e in elms]
for t in text_list[:3]: # 상위 3개만
    page.get_by_text(t).click()

    for l in page.locator('a.site-link').all():
        with page.expect_download() as download_info:
            l.click()
        download = download_info.value
        download.save_as(download.suggested_filename)

time.sleep(1)

page.goto('https://eclass.tukorea.ac.kr/ilos/pf/course/lecture_material_list_form.ac1')
```

네이버 로그인

```
from playwright.sync_api import sync_playwright
import time

url = 'https://nid.naver.com/nidlogin.login?mode=form&url=https://www.naver.com/'
p = sync_playwright().start()

browser = p.chromium.launch(headless=False).new_context(
    viewport={'width': 1920, 'height': 1024}
)

page = browser.new_page()
page.goto(url)
print(page.title())

page.get_by_role('textbox', name='아이디').fill('*****')
page.get_by_role('textbox', name='비밀번호').fill('*****')
page.get_by_role('button', name='로그인').click()
page.get_by_text('등록', exact=True).click()
```