3강

고급 스크래이핑

환경설정

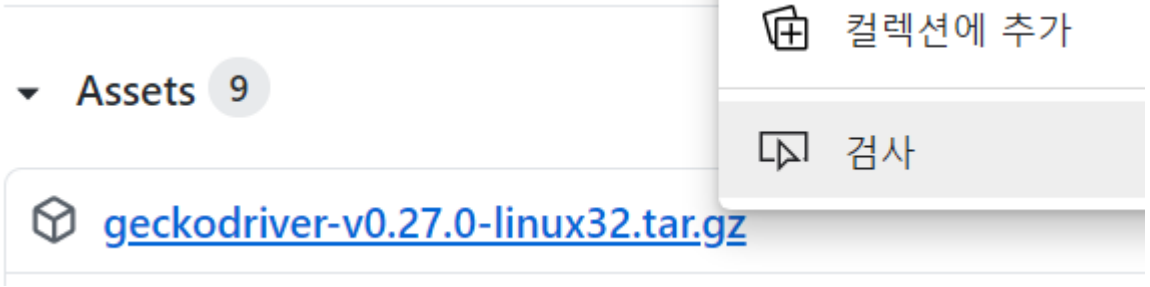
Docker 설치

```
docker pull ubuntu:16.04
docker run -it ubuntu:16.04

apt-get update
apt-get install -y python python3-pip
pip3 install selenium
pip3 install beautifulsoup4

apt-get install -y wget libfontconfig
```

firefox 설치

apt-get install -y firefox

https://github.com/mozilla/geckodriver/releases



wget https://github.com/mozilla/geckodriver/releases/download/v0.27.0/geckodriver-v0.27.0-linux32.tar.gz

tar -zxvf geckodriver-v0.27.0-linux32.tar.gz

mv geckodriver /usr/local/bin

도커 저장

apt-get install fonts-nanum* -y

exit

docker ps -a
docker commit b0f8758dbf12 ubuntu-phantomjs

Docker 실행

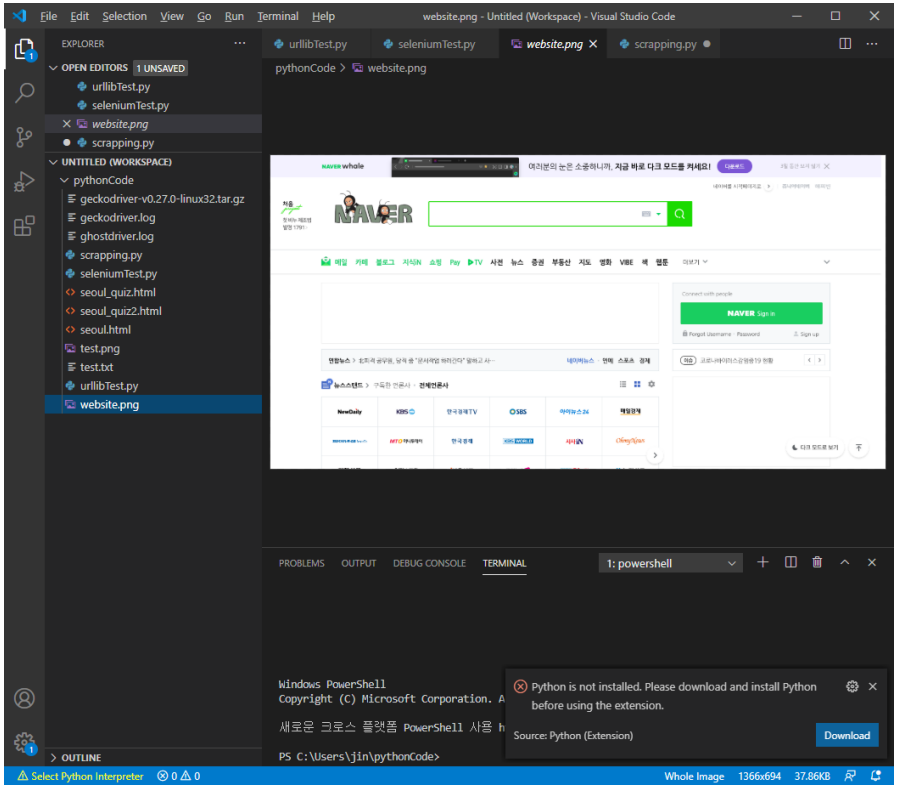
docker run -itv /c/users/jin/pythonCode:/pythonCode -e ko_kr.utf-8 -e PYTHONIOENCODING=utf-8 ubuntu-firefox /bin/bash

```
root@c316c6c89780:/# cd pythonCode/
root@c316c6c89780:/pythonCode# ls
scrapping.py  seoul.html  seoul_quiz.html  seoul_quiz2.html
```

Selenium 활용

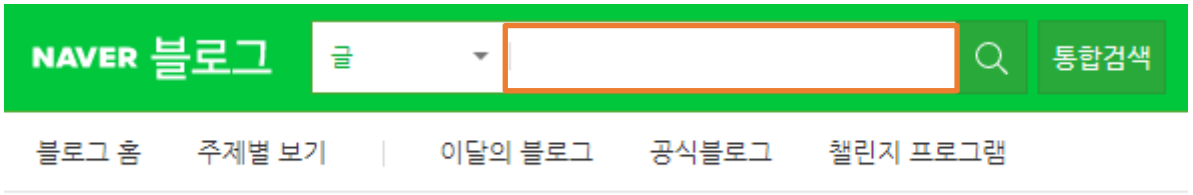
기본 코드

```
from selenium.webdriver import Firefox, FirefoxOptions

url="http://www.naver.com"

#ui제거를 위해 headless로 open
options=FirefoxOptions()
options.add_argument('-headless')
browser = Firefox(options=options)
#웹 접속
browser.get(url)
#현재 화면 스크린샷
browser.save_screenshot("website.png")
#종료
browser.quit()
```



행동패턴 분석

❖ https://section.blog.naver.com/




```
<input ng-click="navigationCtrl.checkIsEmptyAndSendNclick($event)"
ng-model="navigationCtrl.searchWord" bg-immediate-input bg-enter=
"navigationCtrl.search()" type="text" name="sectionBlogQuery"
class="textbox ng-pristine ng-valid ng-not-empty ng-valid-
maxlength ng-touched" accesskey title="검색어를 입력하고 버튼을 누르
세요" maxlength="255" autocomplete="off" value placeholder> == $0
```

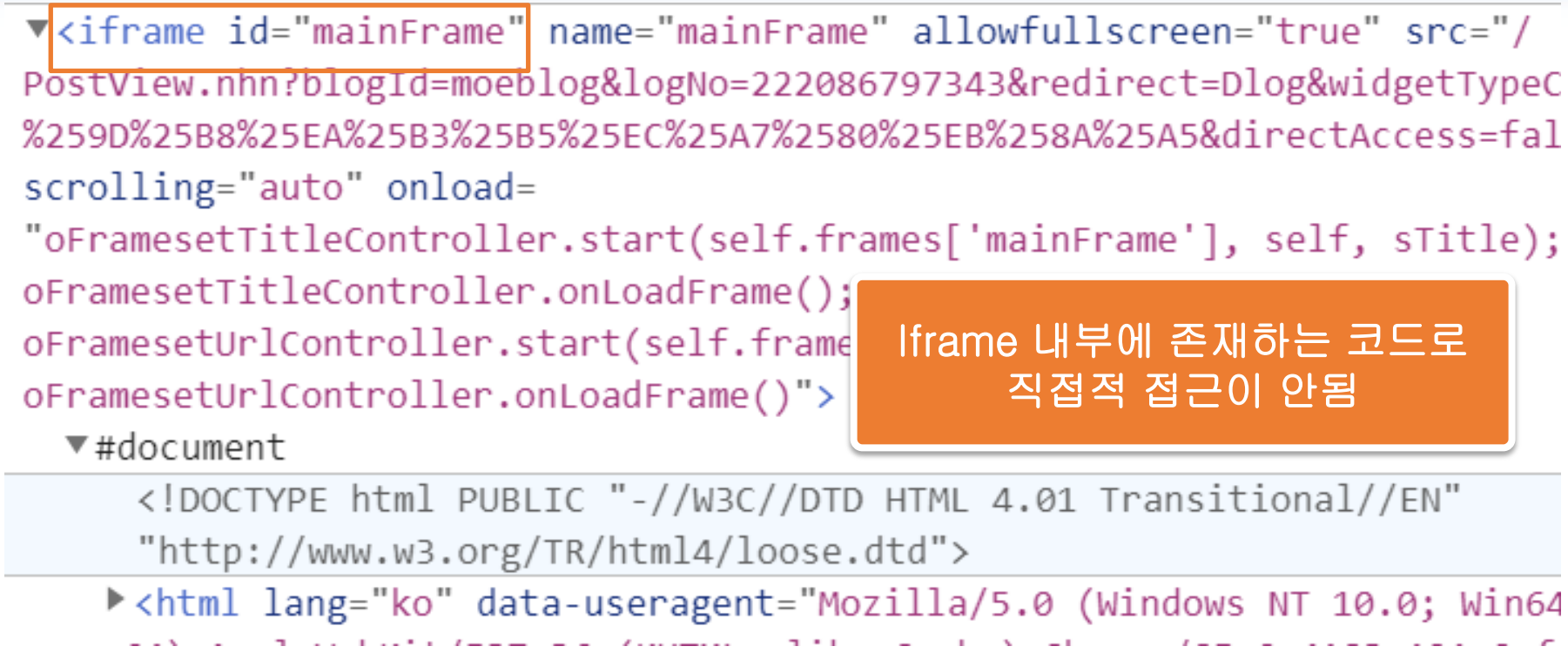
검색어 입력

```
from selenium.webdriver import Firefox, FirefoxOptions
import time

url="https://section.blog.naver.com/"
#ui제거를 위해 headless로 open
options=FirefoxOptions()
options.add_argument('-headless')
browser = Firefox(options=options)
#웹 접속
browser.get(url)
search = browser.find_element_by_css_selector("input.textbox[type=text]")
search.clear()
search.send_keys("인공지능")
#현재 화면 스크린샷
browser.save_screenshot("website2.png")
#종료
browser.quit()
```

검색 및 결과 출력

```
#anchor 태그를 이용한 이동
anchor = browser.find_element_by_css_selector("a.button.button_blog")
anchor.click()

titles = browser.find_elements_by_css_selector("div.desc")

for t in titles:
    el = t.find_element_by_css_selector("span.title")
    # print(el.get_property("innerText"))
    # print(el.get_property("innerHTML"))
    print(el.text)
    print("==============")
```

분석


<iframe id="mainFrame" name="mainFrame" allowfullscreen="true" src="/
PostView.nhn?blogId=moeblog&logNo=222086797343&redirect=Dlog&widgetTypeC
%259D%25B8%25EA%25B3%25B5%25EC%25A7%2580%25EB%258A%25A5&directAccess=fal
scrolling="auto" onload=
"oFramesetTitleController.start(self.frames['mainFrame'], self, sTitle);
oFramesetTitleController.onLoadFrame();
oFramesetUrlController.start(self.frame
oFramesetUrlController.onLoadFrame()">
#document
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html lang="ko" data-useragent="Mozilla/5.0 (Windows NT 10.0; Win64
Iframe 내부에 존재하는 코드로
직접적 접근이 안됨

분석

```
<span style="color:#000000;background-color:#ffffff;" class="se-
fs- se-ff- se-weight-unset  " id="SE-720bba40-700d-4fcf-b76a-
cf30c7b4bdab">맞는 학습 콘텐츠를 추천하고 학습 조언을 제공하는 시스
템이다. </span> == $0
```

Span의 classname으로
내부 내용 얻기

코드

```
options = Options()
options.headless = True
browser = webdriver.Firefox(options=options)
#웹 접속
browser.get("https://blog.naver.com/moeblog/222086797343")
time.sleep(3)
browser.save_screenshot("website2.png")
browser.switch_to_frame("mainFrame")
contents = browser.find_elements_by_class_name("se-fs-")
for content in contents:
    print(content.text)
```

Quiz

❖ 블로그에 있는 모든 정보를 얻어 올 수 있도록 코딩하시오