[크롤링(Crawling)]

- 웹 페이지의 내용을 가지고 오는 것을 크롤링 또는 스크래핑(Scraping) 이라고 함
- 구글이나 네이버, 다음과 같은 검색 엔진 사이트들은 검색이 속도를 높이기 위해 로봇(robot)이라는 프로그램을 만들어서 자동으로 웹 페이지를 크롤링
- 무분별한 크롤링을 막고 제어하기 위해 1994년 6월 로봇 배제 규약 - robot.txt(로봇 접근 관련 내용) : 크롤링 허가/불허가 여부를 이 파일에 적어 놓자고 약속한 규약
- 크롤링하는 로봇프로그램은 크롤링하고 자하는 사이트의 http://www.aaa.com/robot.txt 파일을 찾아 분석 후 수집해도 되는 콘텐츠만 수집해야 함. 단, 강제는 아닌 권고.

[Robot.txt]

예)홈페이지 전체가 모든 검색엔진에 노출되기를 원치 않음

User-agent: *

Disallow: /

예) 홈페이지 전체가 모든 검색엔진에 노출되기를 원함

User-agent: *

Disallow:

예)홈페이지 디렉토리중 일부만 검색엔진에 노출하고 싶음

User-agent: *

Disallow: /my_photo/

Disallow: /my_diary/

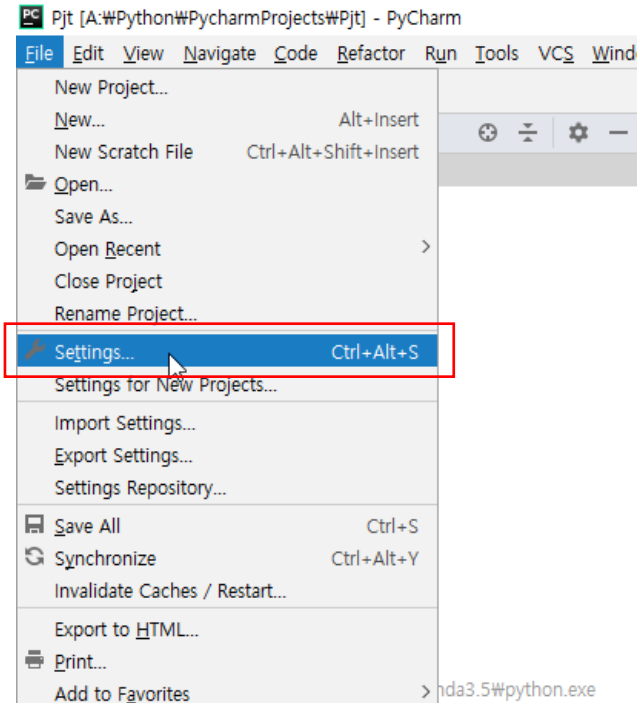
예)홈페이지 전체를 노출시키지만 특정 검색엔진 (EvilRobot)만 거부

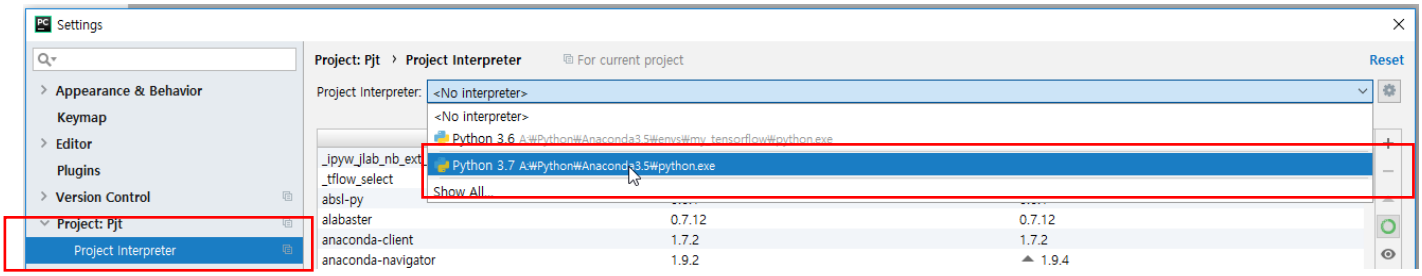
User-agent: EvilRobot

Disallow: /

예) https://www.google.com/robots.txt

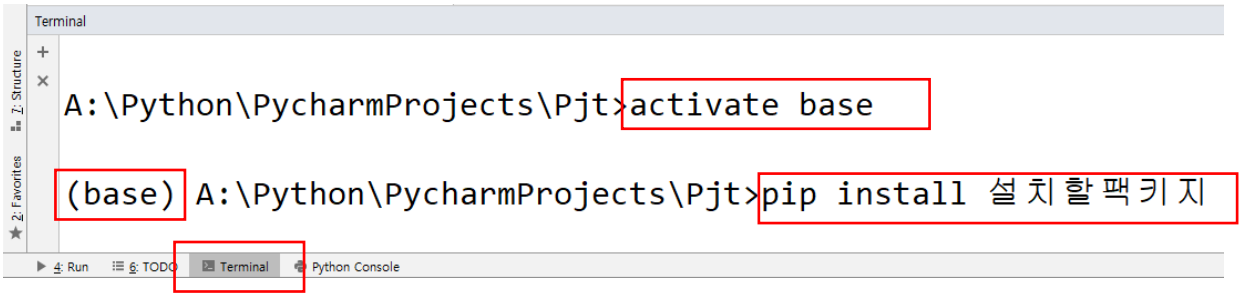
파이참에서 python interpreter 변경 방법





Ananconda3/python.exe 파일 선택합니다. (Ananconda base env)

추가 팩키지 설치 방법)

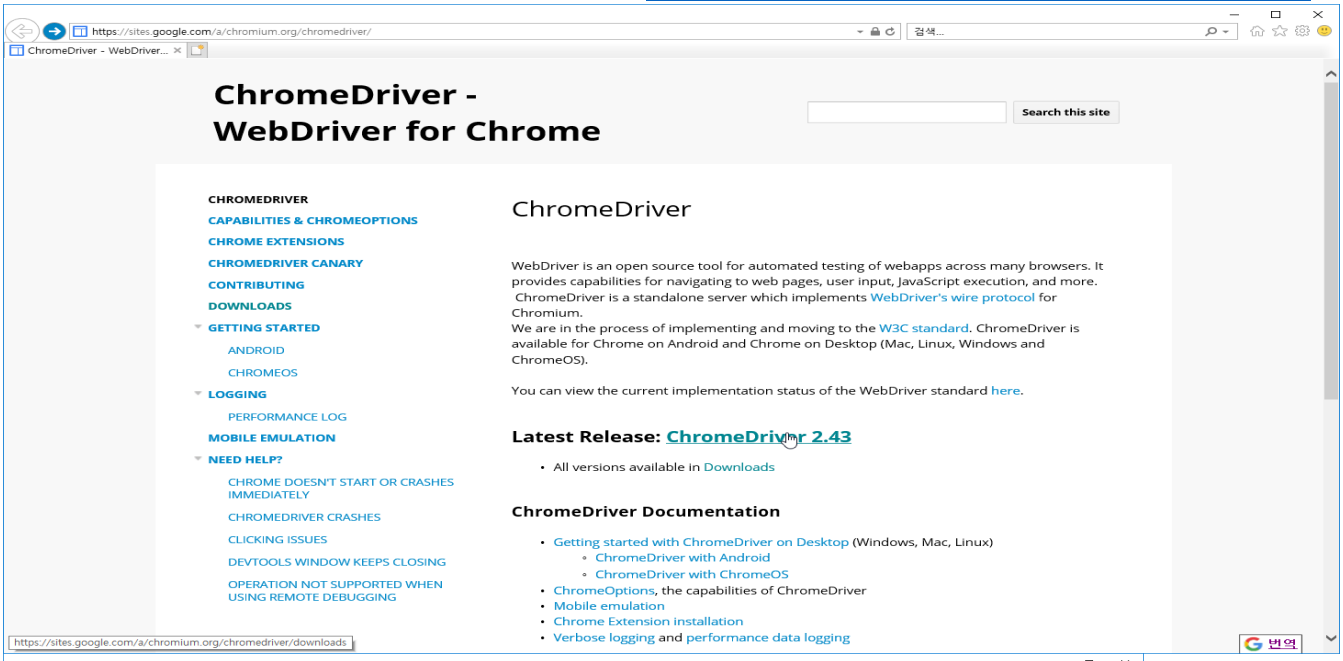


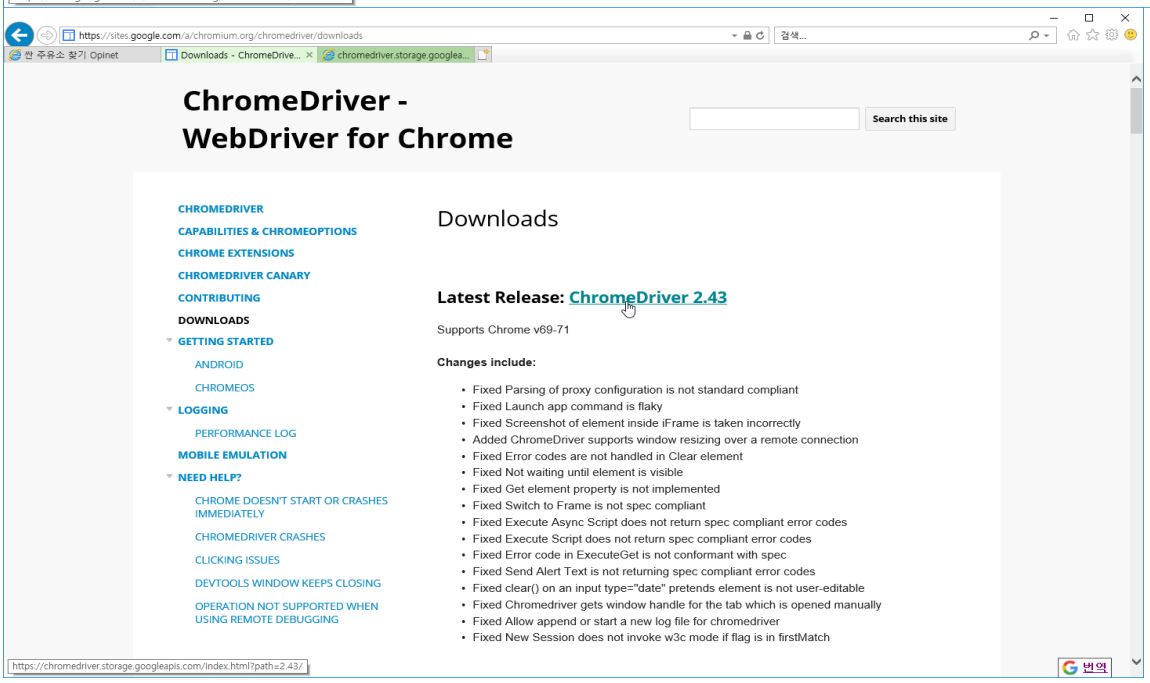
[Selenium] 웹 페이지에 자동 접근, 원하는 정보 얻기

1) selenium 설치

pip install selenium

2) '크롬 웹 드라이버' Chrome Driver 다운로드 https://sites.google.com/a/chromium.org/chromedriver

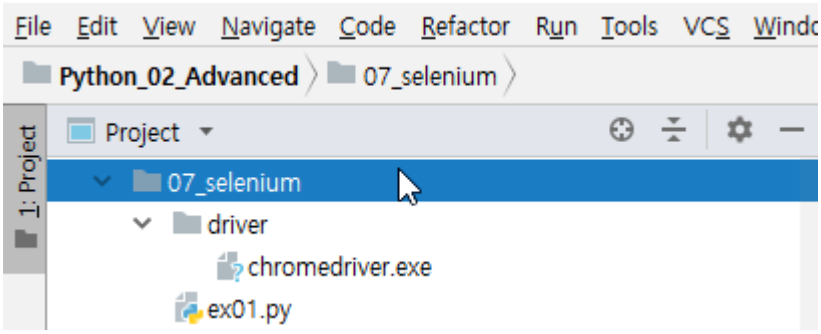




Index of /2.43/

Name	Last modified	Size	ETag
Parent Directory		-	
chromedriver_linux64.zip	2018-10-17 02:46:13	3.89MB	1a67148288f4320e5125649f66e02962
chromedriver_mac64.zip	2018-10-17 04:09:49	5.71MB	249108ab937a3bf8ae8fd22366b1c208
chromedriver_win32.zip	2018-10-17 03:01:50	3.45MB	d238c157263ec7f668e0ea045f29f1b7
notes.txt	2018-10-17 05:00:45	0.02MB	a84902c9429641916b085a72ad5de724

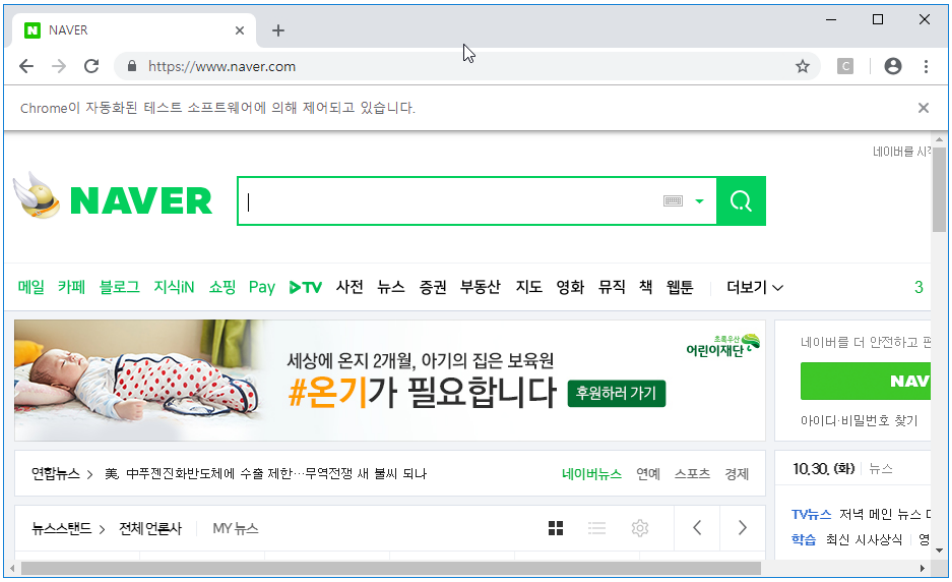
다음과 같이 driver 폴더 생성하고 크롬 드라이버 복사



ex01.py

```
from selenium import webdriver

driver = webdriver.Chrome('driver/chromedriver')
driver.get("http://naver.com")
```

실행 결과, 자동으로 크롬 웹 브라우저 뜸.

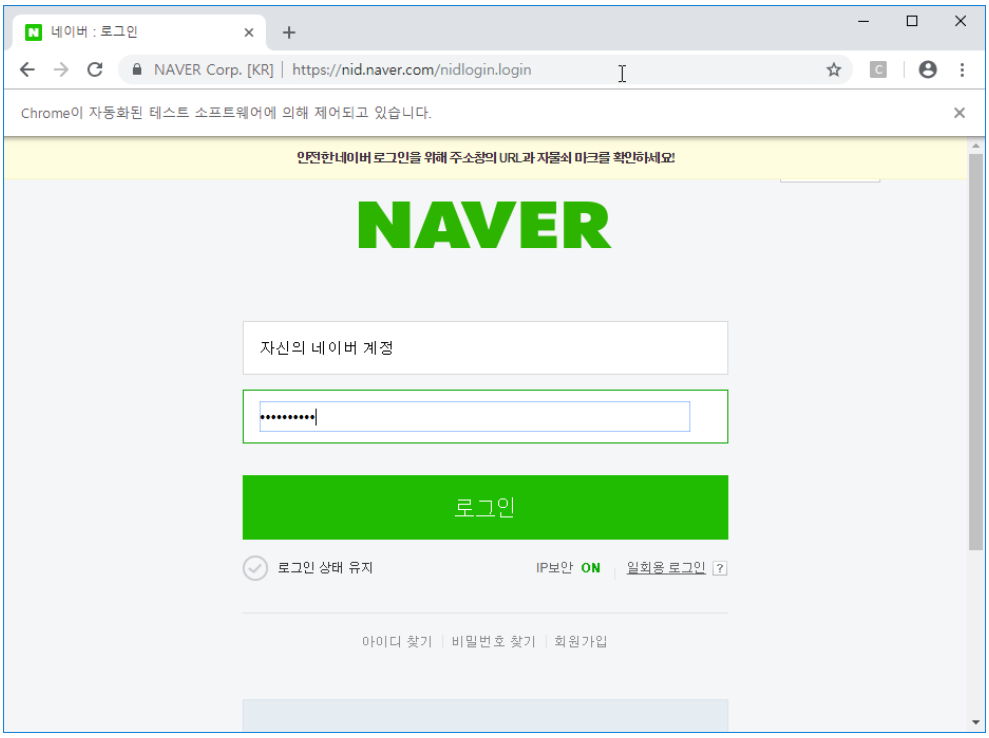


ex02_login.py 특정 위치에서 타이핑

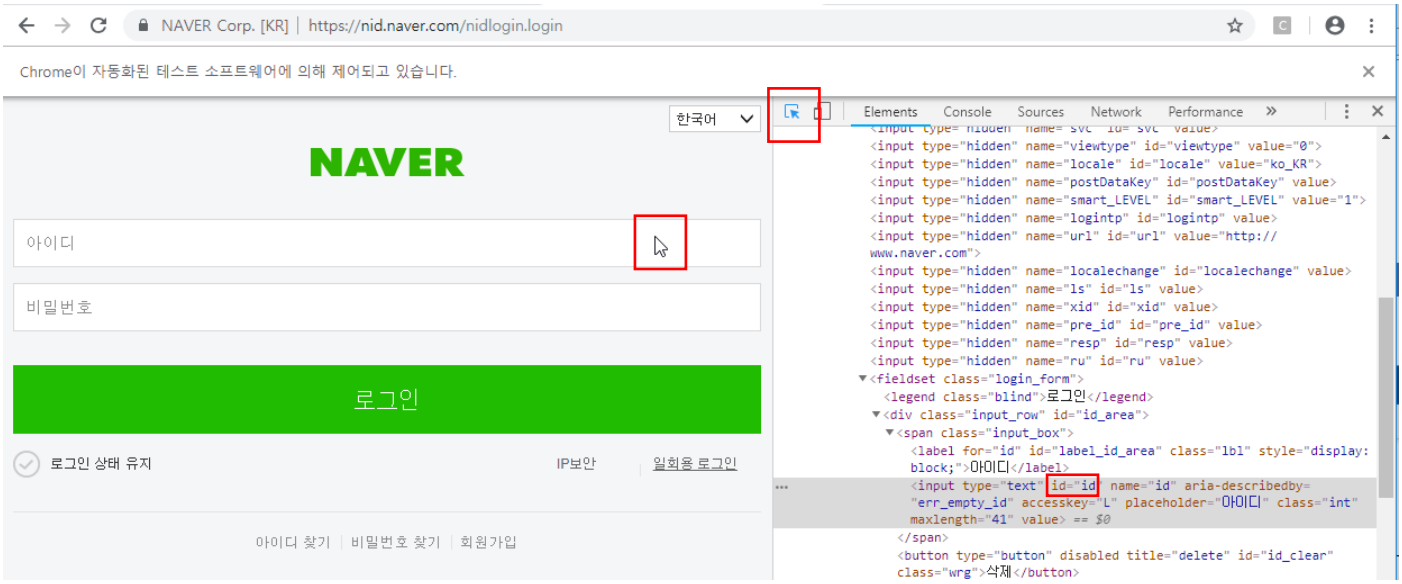
```
from selenium import webdriver

driver = webdriver.Chrome('driver/chromedriver')
driver.get("https://nid.naver.com/nidlogin.login")

elem_login = driver.find_element_by_id("id")
elem_login.clear()
elem_login.send_keys("자신의 네이버 계정")

elem_login = driver.find_element_by_id("pw")
elem_login.clear()
elem_login.send_keys("자신의 네이버 비번")
```

elem_login = driver.find_element_by_id("id")
--> 현재 크롬에 떠 있는 웹페이지에서 id 속성 값이 id인 element 찾음
elem_login.clear()
elem_login.send_keys("자신의 네이버 계정") -->그곳에 타이핑

실행 결과

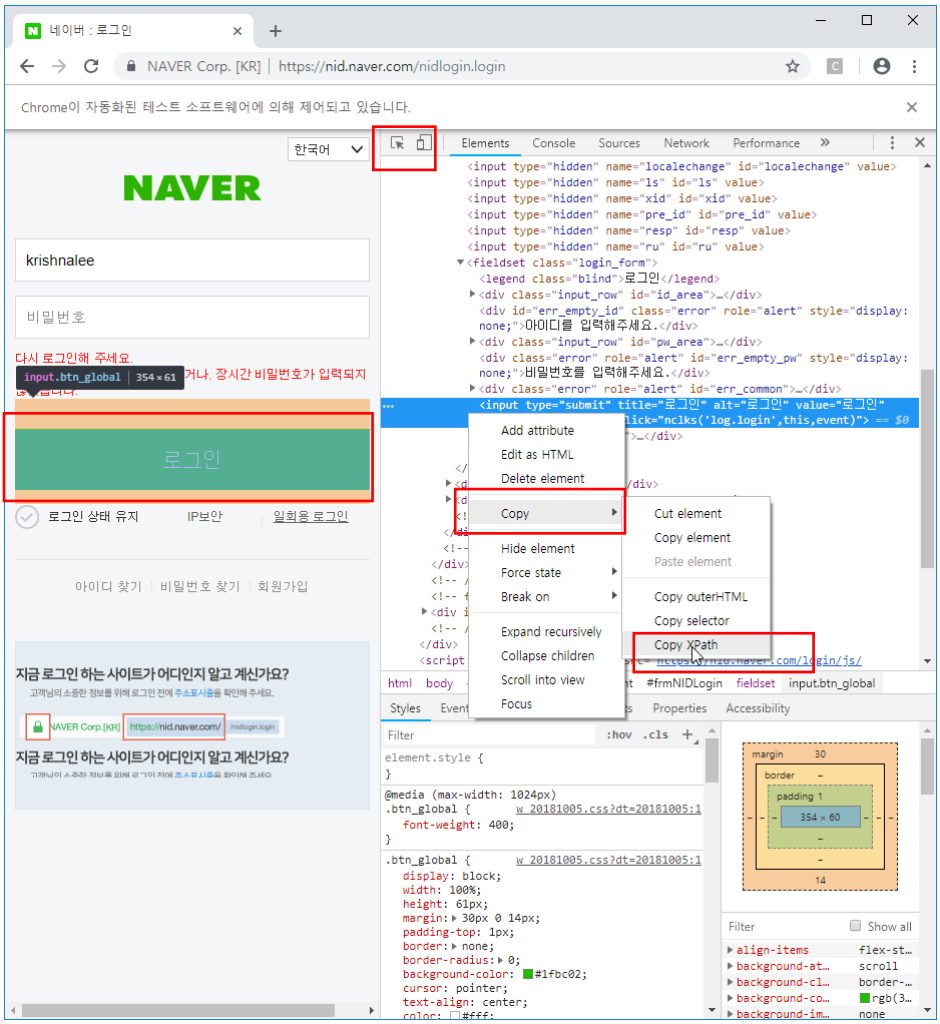


<참고> 찾고자하는 element 의 id 확인 방법. 크롬에서 F12



ex03_login.py 특정 위치에서 클릭 예

1) 클릭할 곳의 XPath 찾기



복사한 xpath를 다음 코드에 붙여 넣음

```
from selenium import webdriver

driver = webdriver.Chrome('driver/chromedriver')
driver.get("https://nid.naver.com/nidlogin.login")

elem_login = driver.find_element_by_id("id")
elem_login.clear()
elem_login.send_keys("자신의 네이버 계정")

elem_login = driver.find_element_by_id("pw")
elem_login.clear()
elem_login.send_keys("자신의 네이버 비번")

xpath = """//*[@id="frmNIDLogin"]/fieldset/input"""
driver.find_element_by_xpath(xpath).click()
```

실행 결과:

로그인 버튼이 클릭됨. 네이버는 보안문자를 입력하게 되어 있지만,
보안문자 입력이 없는 사이트는 자동로그인가능.

ex04.py Beautiful soup 과 함께 사용 가능

```
from selenium import webdriver

driver = webdriver.Chrome('driver/chromedriver')
driver.get("https://movie.naver.com/movie/bi/mi/basic.nhn?code=160487")

from bs4 import BeautifulSoup

html = driver.page_source
soup = BeautifulSoup(html, 'html.parser')

raw_list = soup.find_all('div', class_="story_area")
print( raw_list )
```

실행 결과 :

[

<h4 class="h_story">줄거리</h4>

<h5 class="h_tx_story">야귀떼가 온 세상을 집어삼켰다!</h5>
밤에만 활동하는 산 자도 죽은 자도 아닌 '야귀(夜鬼)'가 창궐한 세상,

 위기의 조선으로 돌아온 왕자 '이청'(현빈)은

 도처에 창궐한 야귀떼에 맞서 싸우는 최고의 무관 '박종사관'(조우진)

 일행을 만나게 되고,

 야귀떼를 소탕하는 그들과 의도치 않게 함께하게 된다.

 한편, 조선을 집어삼키려는 절대악 '김자준'(장동건)은 이 세상을 뒤엎기 위한

 마지막 계획을 감행하는데...

 조선필생 VS 조선필망

 세상을 구하려는 자와 멸망시키려는 자!

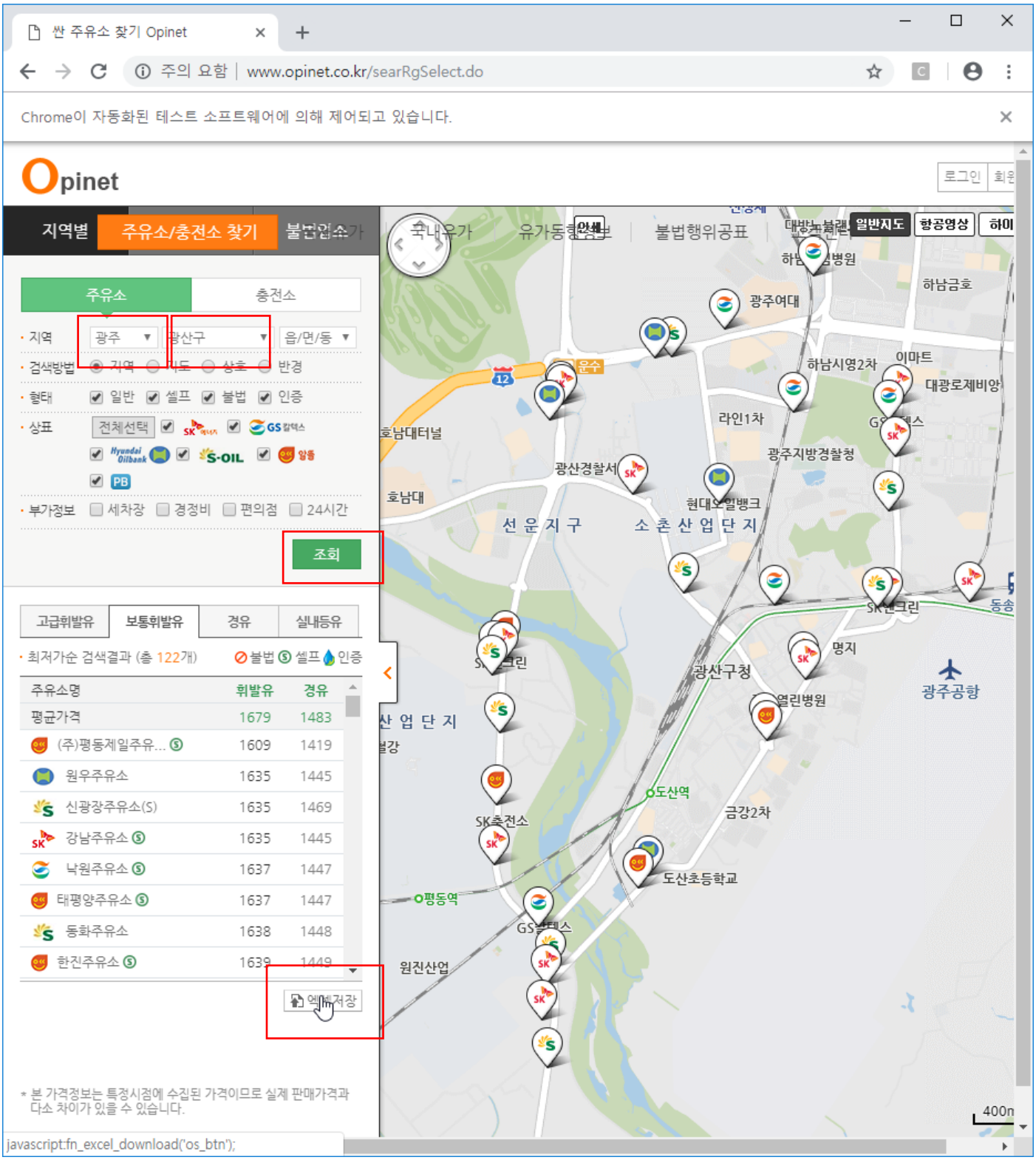
 오늘 밤, 세상에 없던 혈투가 시작된다!
<button class="story_more" id="toggleMakingnoteButton" onclick="storyAndNote.toggleMakingnote();" type="button">제작노트 보기
</button><!-- N=a:mai.story -->
]

mission

http://www.opinet.co.kr/ 사이트는 주유소 가격비교 사이트입니다.

사이트 html을 분석하고, Seleium을 이용하여 다음 과정을 프로그램으로 자동으로 처리하세요.

1) 다음 사이트(http://www.opinet.co.kr/searRgSelect.do)를 Seleium을 이용하여 자동 접속하고,
2) 모든 지역 이름을 크롤링하여 출력하고,
   (출력결과 -->서울, 부산, 대구, ...)
3) 광주, 광산구 지역을 타이핑 후, 조회 버튼을 클릭.
4) 엑셀 저장 버튼을 클릭하여, 엑셀파일을 다운 받습니다.


싼 주유소 찾기 Opinet
주의 요함 | www.opinet.co.kr/searRgSelect.do
Chrome이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.
Opinet
로그인
지역별
주유소/충전소 찾기
불법업소가
국내유가
유가동향분석
불법행위공표
일반지도
항공영상
주유소
충전소
지역
광주
광산구
읍/면/동
검색방법
거리
지도
상호
반경
형태
일반
셀프
불법
인증
상표
전체선택
부가정보
세차장
경정비
편의점
24시간
조회
고급휘발유
보통휘발유
경유
실내등유
최저가순 검색결과 (총 122개)
불법 셀프 인증
주유소명
휘발유
경유
평균가격 1679 1483
(주)평동제일주유... 1609 1419
원우주유소 1635 1445
신광장주유소(S) 1635 1469
강남주유소 1635 1445
낙원주유소 1637 1447
태평양주유소 1637 1447
동화주유소 1638 1448
한진주유소 1639 1449
엑셀저장
* 본 가격정보는 특정시점에 수집된 가격이므로 실제 판매가격과 다소 차이가 있을 수 있습니다.
javascript:fn_excel_download('os_btn');
광주여대
하남금호
하남시영2차
이마트
대광로제비앙
라인1차
광주지방경찰청
광산경찰서
현대오일뱅크
호남대터널
호남대
선 운 지 구
소 촌 산 업 단 지
광산구청
광주공항
명지
열린병원
도산역
금강2차
도산초등학교
SK충전소
평동역
원진산업
400m