

# 웹 크롤링(Web Crawling)

2021년 8월

대진대학교 컴퓨터공학과 서혜선교수

[jako403@daejin.ac.kr](mailto:jako403@daejin.ac.kr)

---

# 웹 크롤링(Web Crawling) 준비

---

- 크롤링이란?
- 셀레니움 소개

## 크롤링이란?

---

인터넷에서 존재하는 데이터를 컴퓨터 프로그램을 통하여 자동화된 방법으로 웹에서 데이터를 수집하는 모든 작업

**‘연금’** 2019.01.01 ~ 2020.12.31 ( 2년치 )  
이라는 키워드

네이버 , 다음 – 웹문서, 블로그, 카페 . . .

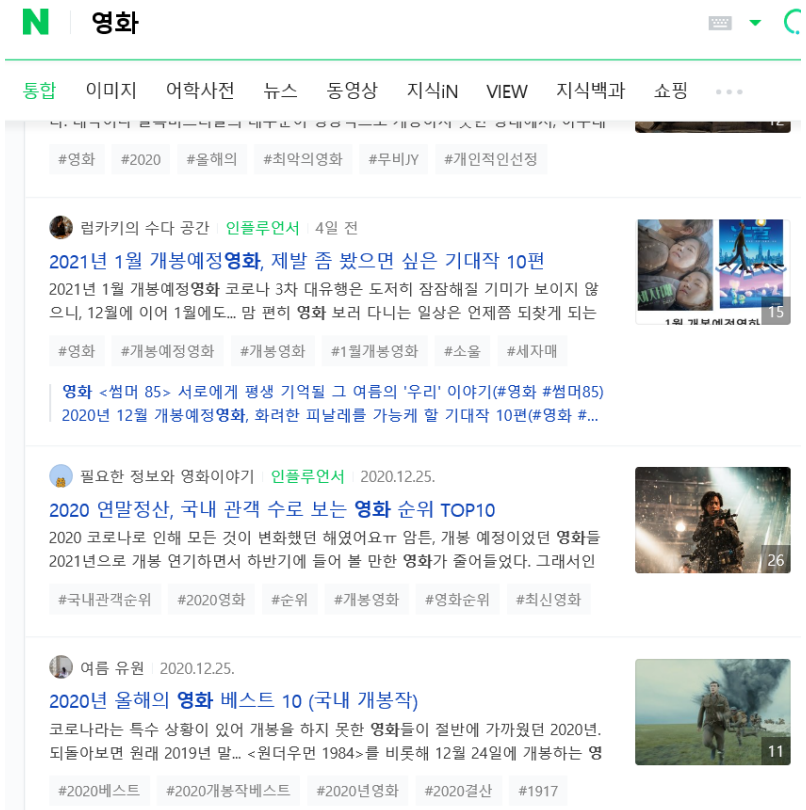


데이터 수집

# 크롤링이란?

Ex) 영화 후기와 평점,  
제품 사용에 관한 리뷰,  
맛집 정보 등의 방대한 데이터를 수집하는 행위

Ex) 오늘과 내일 날씨 정보  
특정 키워드의 뉴스 수집  
영화 리뷰정보 등 등



# 크롤링이란?

## 크롤링(Crawling)과 스크래핑(Scraping)

Crawling : 웹 크롤러가 일정한 규칙으로 웹페이지를 브라우징 하는것 (돌아다니는것)

Scraping : 웹사이트 상에서 원하는 정보를 추출하는 기술

→ 통틀어 크롤링이라 칭함



데이터를 크롤링 해오는데 유용한 도구

크롤링시  
주의점

대부분의 크롤링은 위법의 범주에 들지 않지만

- 1) 웹페이지 운영자가 굶어가지 못하도록 조치한 데이터
- 2) 가져온 데이터를 사용해 부당이득을 얻는 경우
- 3) 민감한 개인정보를 가져오는 경우

→ 저작권법이나 부정경쟁방지법 등의 제재를 받을 수 있음 (주의!!)

# 크롤링이란?

## 크롤링 준비\_라이브러리 설치

### ■ 라이브러리 설치

- `pip install requests` (웹사이트에 HTTP를 요청할 때 사용하는 모듈)  
BS과 같은 기능으로 설치 안해도 무방
- `pip install BS4` (html을 쉽게 사용할 수 있도록 파싱해 주는 모듈)
- `pip install selenium` (크롤링을 위한 모듈, 어느 사이트든 쉽게 접근, 동적 크롤링 – page in page)
- `pip install pandas` (오류시 `pip install numpy==1.19.3`, 현재 1.19.4 버전에서 오류가 발생함)

```
(semina) C:\Windows\system32>pip install BS4
Collecting BS4
  Using cached bs4-0.0.1-py3-none-any.whl
Collecting beautifulsoup4
  Using cached beautifulsoup4-4.9.3-py3-none-any.whl (115 kB)
Collecting soupsieve>1.2
  Downloading soupsieve-2.1-py3-none-any.whl (32 kB)
Installing collected packages: soupsieve, beautifulsoup4, BS4
Successfully installed BS4-0.0.1 beautifulsoup4-4.9.3 soupsieve-2.1
```

# 크롤링이란?

---

## 크롤링 준비\_라이브러리 설치

### ■ Request

- Python 에서 HTTP 요청을 보내는 모듈인 requests

### ■ BeautifulSoup4 (BS4)

- HTML 및 XML 파일에서 원하는 데이터를 손쉽게 Parsing 할 수 있는 Python 라이브러리 (특정 패턴이나 순서대로 추출)

### ■ Pandas (pd)

- Pandas는 파이썬 데이터 처리를 위한 라이브러리

# 크롤링이란?

---

## 크롤링 준비\_필요라이브러리 import

### 쥬피터에서의 준비작업

```
from selenium import webdriver
import requests
from bs4 import BeautifulSoup as BS
import pandas as pd
import time
import re
import pickle
```

re 모듈 : 문자열 처리를 위한 정규표현식 사용 라이브러리



# 크롤링이란?

---

## 크롤링 준비\_크롬 드라이버 설치

1. 크롬 드라이버를 설치한다.
2. 크롬드라이버의 파일 경로 설정
  - 크롬 드라이버 파일 경로 지정 . 셀레니움의 필수조건
  - 크롤링을 하고 데이터를 저장할 폴더와, 셀레니움을 사용하기 위해 크롬드라이버 경로 지정

## 셀레니움 소개

---

보통은 BeautifulSoup (BS) 라이브러리를 이용하기도 하지만 크롤링이 잘 안되는 경우가 많아 Selenium을 주로 사용

- 셀레니움의 유용성

- 시각화를 거친 셀레니움은 사람에게 최적화
- 디버깅(코드 실행)시 브라우저에서 눈으로 직접 크롤링 진행과정 확인가능
- 사용자가 보는 웹페이지의 모든 정보를 가져올 수 있음
- 오픈소스로 라이선스 비용 무료

- 셀레니움의 단점

- 메모리를 많이 차지하고 속도가 느림
- 설치, 구성 등이 다소 복잡 (크롬드라이브 설치 필수 )

# 셀레니움 소개

---

- Selenium 실행을 위해 웹 드라이버가 필요한 브라우저들

- **Apple Safari**

- 사파리 10은 WebDriver API를 지원.

- **Google Chrome**

- ChromeDriver를 별도로 다운로드 하고 경로를 지정해 주어야 함

- **Internet Explorer**

- Selenium project에서 IE용 Driver를 제공.

- **Microsoft Edge**

- Microsoft Edge WebDriver도 Selenium을 위한 WebDriver 사용 가능.

- **Mozilla Firefox**

- Firefox 47.0.1이상을 테스트하기 위해서는 Mozilla GeckoDriver가 필요.

# 셀레니움 소개

## 크롬드라이버 설치\_내 PC의 크롬 버전 확인

- Chrome 이용시 웹 드라이버 설치하기

- Chrome의 맨 우측 상단의 세 개의 점을 클릭하여 크롬의 설정페이지로 들어감



(크롬 버전 확인은 크롬의 설정 > 크롬 정보 에서 확인 )

# 셀레니움 소개

## 크롬드라이버 설치\_내 PC의 크롬 버전 확인

- 왼쪽 메뉴에서 Chrome 정보를 클릭하여 버전 확인 (최신 버전이 아니면 업데이트)

The screenshot displays the Chrome Settings application. On the left, a sidebar lists various settings categories. The 'Chrome 정보' (Chrome Info) option is highlighted with a red circle. The main panel on the right is titled 'Chrome 정보' and contains the following information:

- Chrome** (with the Chrome logo)
- A blue checkmark icon followed by the text: 'Chrome이 최신 버전입니다. 버전 87.0.4280.88(공식 빌드) (64비트)' (Chrome is the latest version. Version 87.0.4280.88 (Official Build) (64-bit)).
- A link labeled 'Chrome 도움말 보기' (View Chrome Help) with an external link icon.
- A link labeled '문제 신고' (Report Problem) with an external link icon.
- A section titled 'Chrome' with the text 'Copyright 2020 Google LLC. All rights reserved.' and a paragraph stating 'Chrome은 Chromium 오픈소스 프로젝트를 비롯한 여러 오픈소스 소프트웨어에 기초해 만들어진 브라우저입니다.' (Chrome is a browser created based on the Chromium open-source project and other open-source software).
- A link labeled '서비스 약관' (Terms of Service).

## 셀레니움 소개

### 크롬드라이버 설치\_내 PC의 크롬 버전 확인

- Chrome의 버전을 기억한다. 현재 컴퓨터의 Chrome 버전은 87.0.4280.88

Chrome 정보



Chrome



Chrome이 최신 버전입니다.

버전 87.0.4280.88(공식 빌드) (64비트)

Chrome 도움말 보기



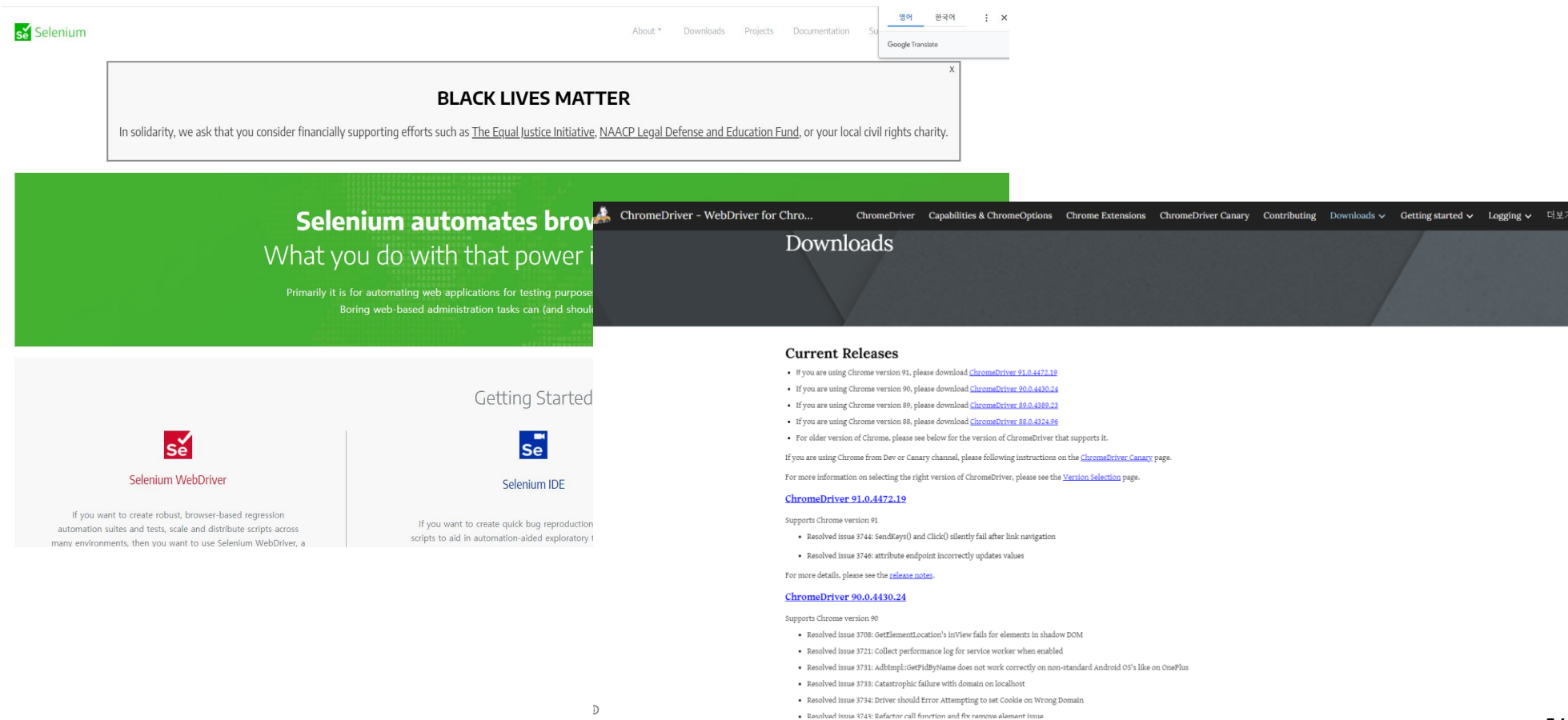
문제 신고



# 셀레니움 소개

## 크롬드라이버 설치

- 주소창에 'selenium.dev', 또는 chromedriver를 검색하여 버전에 맞는 chromedriver 설치



The image shows a composite of two web pages. The top part is the Selenium website (selenium.dev) with a green header and a black banner for 'BLACK LIVES MATTER'. Below the banner, it says 'Selenium automates browsers' and 'What you do with that power is'. The bottom part is the ChromeDriver website (chromedriver.chromium.org) with a dark header and a 'Downloads' section. The 'Current Releases' section lists versions 91.0.4472.19, 90.0.4430.24, 89.0.4389.23, and 88.0.4324.56. It also includes instructions for using Chrome from Dev or Canary channels and links to release notes.

**BLACK LIVES MATTER**

In solidarity, we ask that you consider financially supporting efforts such as [The Equal Justice Initiative](#), [NAACP Legal Defense and Education Fund](#), or your local civil rights charity.

**Selenium automates browsers**

What you do with that power is

Primarily it is for automating web applications for testing purposes  
Boring web-based administration tasks can (and should) be automated too

**Getting Started**

**Selenium WebDriver**

If you want to create robust, browser-based regression automation suites and tests, scale and distribute scripts across many environments, then you want to use Selenium WebDriver, a

**Selenium IDE**

If you want to create quick bug reproduction scripts to aid in automation-aided exploratory t

**ChromeDriver - WebDriver for Chromium**

**Downloads**

**Current Releases**

- If you are using Chrome version 91, please download [ChromeDriver 91.0.4472.19](#)
- If you are using Chrome version 90, please download [ChromeDriver 90.0.4430.24](#)
- If you are using Chrome version 89, please download [ChromeDriver 89.0.4389.23](#)
- If you are using Chrome version 88, please download [ChromeDriver 88.0.4324.56](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

If you are using Chrome from Dev or Canary channel, please following instructions on the [ChromeDriver Canary](#) page.

For more information on selecting the right version of ChromeDriver, please see the [Version Selection](#) page.

**ChromeDriver 91.0.4472.19**

Supports Chrome version 91

- Resolved issue 3744: SendKeys() and Click() silently fail after link navigation
- Resolved issue 3746: attribute endpoint incorrectly updates values

For more details, please see the [release notes](#).

**ChromeDriver 90.0.4430.24**

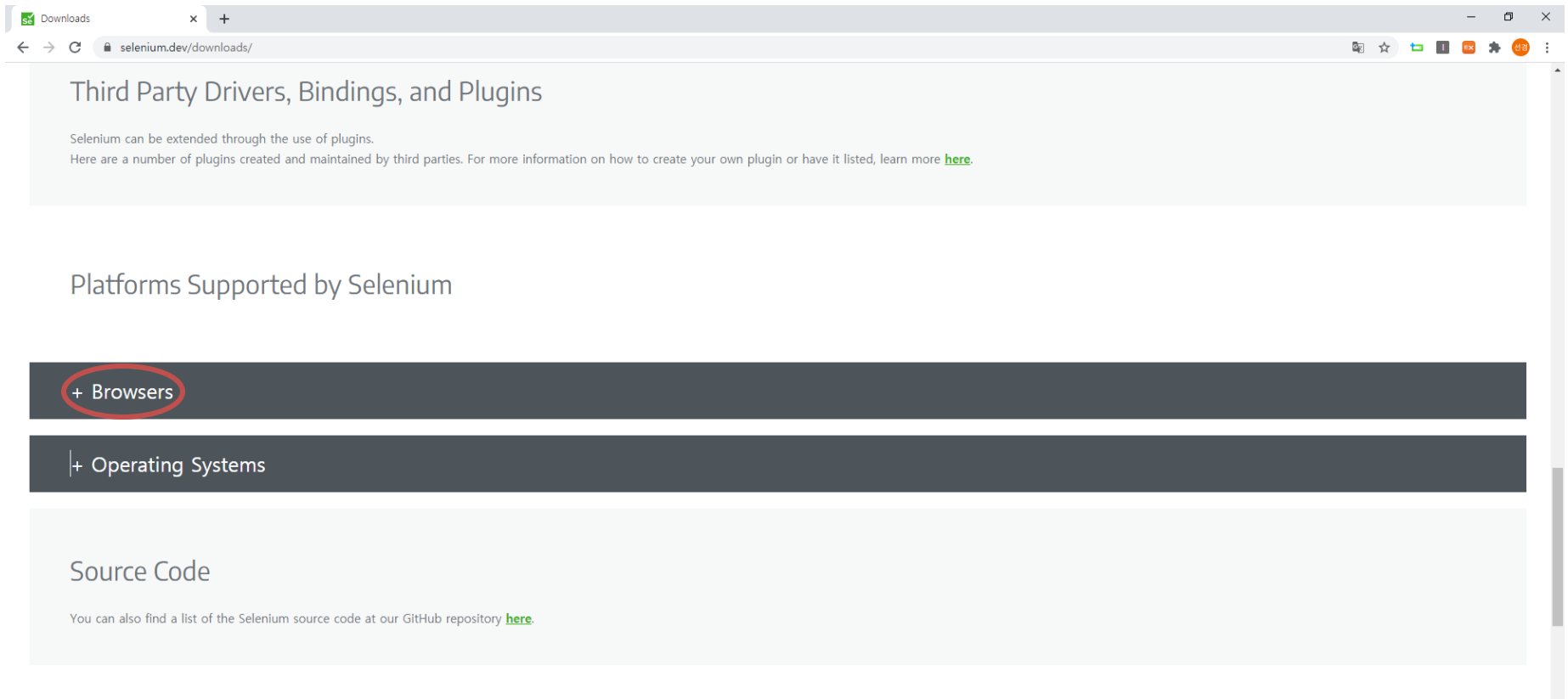
Supports Chrome version 90

- Resolved issue 3708: GetElementLocation's inView fails for elements in shadow DOM
- Resolved issue 3721: Collect performance log for service worker when enabled
- Resolved issue 3731: AddImpl::GetPidByName does not work correctly on non-standard Android OS's like on OnePlus
- Resolved issue 3733: Catastrophic failure with domain on localhost
- Resolved issue 3734: Driver should Error Attempting to set Cookie on Wrong Domain
- Resolved issue 3745: Refactor call function and fix remove element issue

# 셀레니움 소개

## 크롬드라이버 설치

- Download 페이지로 이동하였으면 아래로 스크롤 하여 내려가 +Browsers 버튼 클릭

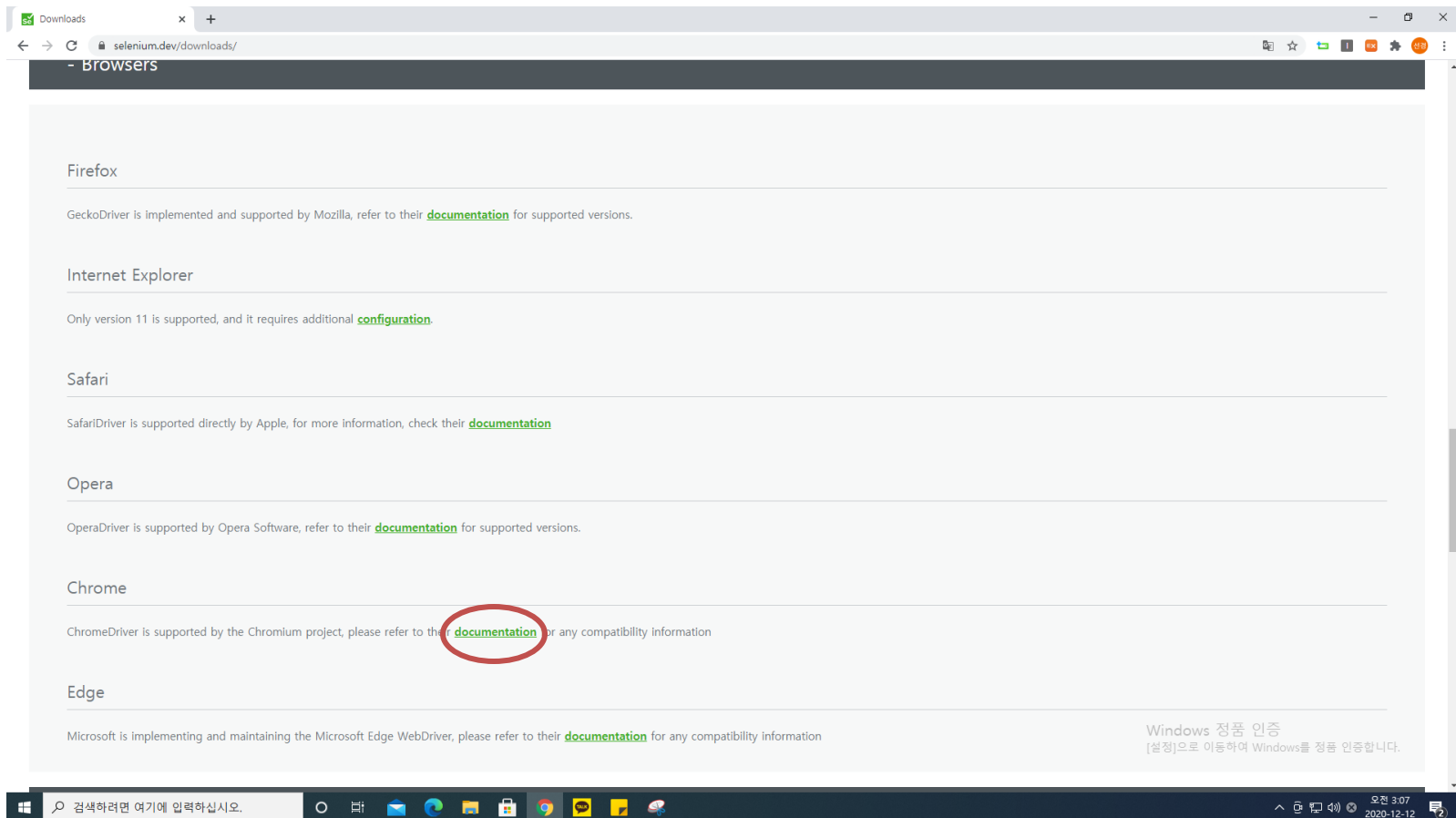




# 셀레니움 소개

## 크롬드라이버 설치

- 여기서 Chrome 항목에 있는 [documentation](#) 버튼 클릭



# 셀레니움 소개

## 크롬드라이버 설치

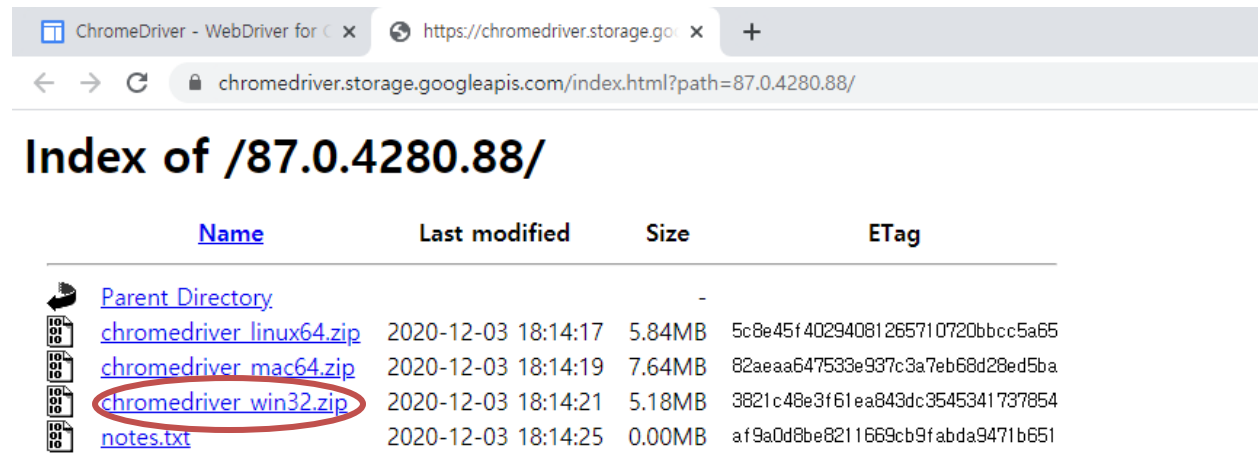
- 미리 확인했던 본인의 Chrome 버전에 맞는 드라이버를 다운(여기서는 87.0.4280.88)

The screenshot shows the ChromeDriver website (sites.google.com/a/chromium.org/chromedriver/). The page title is "ChromeDriver - WebDriver for Chrome". On the left is a navigation menu with links like "CHROMEDRIVER", "CAPABILITIES & CHROMEOptions", "CHROME EXTENSIONS", "CHROMEDRIVER CANARY", "CONTRIBUTING", "DOWNLOADS", "VERSION SELECTION", "GETTING STARTED", "LOGGING", "MOBILE EMULATION", and "NEED HELP?". The "DOWNLOADS" section is expanded, showing "All versions available in Downloads". Under this section, there are two bullet points: "Latest stable release: ChromeDriver 87.0.4280.88" and "Latest beta release: ChromeDriver 88.0.4324.27". The version "87.0.4280.88" is circled in red. Below this is the "ChromeDriver Documentation" section with links for getting started, capabilities, mobile emulation, security considerations, and verbose logging. At the bottom is the "Troubleshooting" section with links for common issues like crashes, clicking issues, and remote debugging.

# 셀레니움 소개

## 크롬드라이버 설치






- chromedriver\_win32.zip을 눌러 다운로드 (win64인경우도 win32로 다운로드)



ChromeDriver - WebDriver for C x    https://chromedriver.storage.go x    +

← → ↻    chromedriver.storage.googleapis.com/index.html?path=87.0.4280.88/

### Index of /87.0.4280.88/

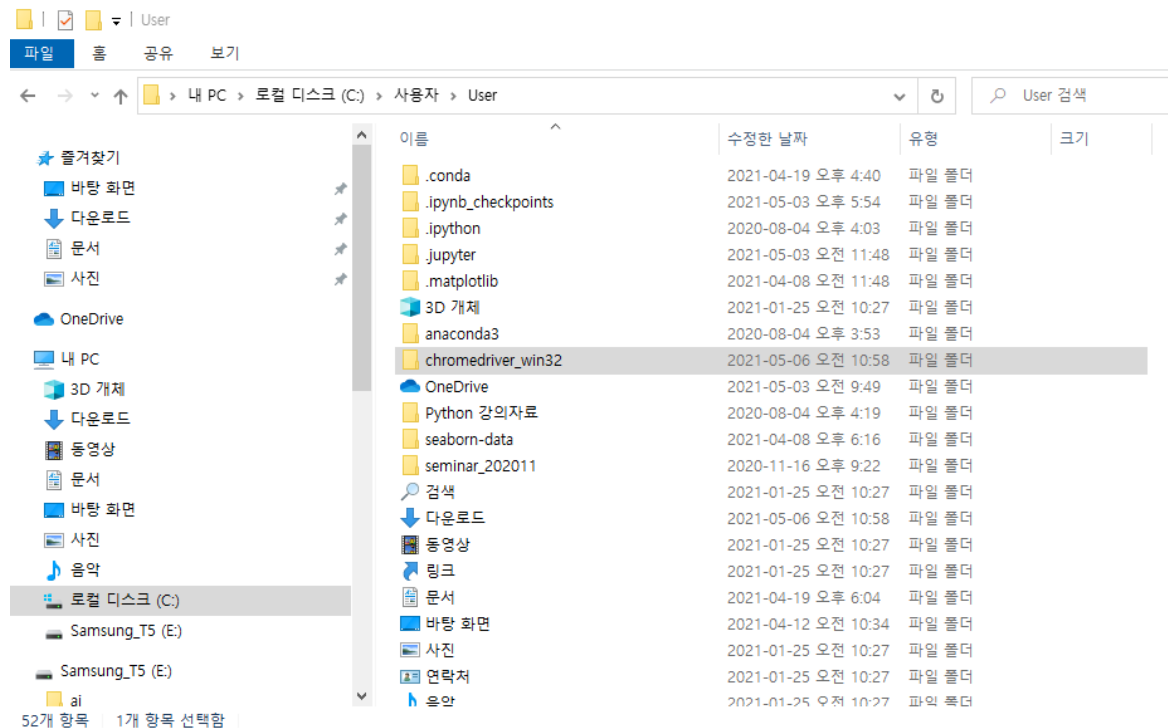
	<a href="#">Name</a>	Last modified	Size	ETag
	<a href="#">Parent Directory</a>		-	
	<a href="#">chromedriver_linux64.zip</a>	2020-12-03 18:14:17	5.84MB	5c8e45f40294081265710720bbcc5a65
	<a href="#">chromedriver_mac64.zip</a>	2020-12-03 18:14:19	7.64MB	82aeaa647533e937c3a7eb68d28ed5ba
	<a href="#">chromedriver_win32.zip</a>	2020-12-03 18:14:21	5.18MB	3821c48e3f61ea843dc3545341737854
	<a href="#">notes.txt</a>	2020-12-03 18:14:25	0.00MB	af9a0d8be8211669cb9fabda9471b651

# 셀레니움 소개

## 크롬드라이버 설치

- 다운받은 zip파일을 열어준 후 Selenium를 사용하는 코드를 저장할 폴더에 압축 풀기

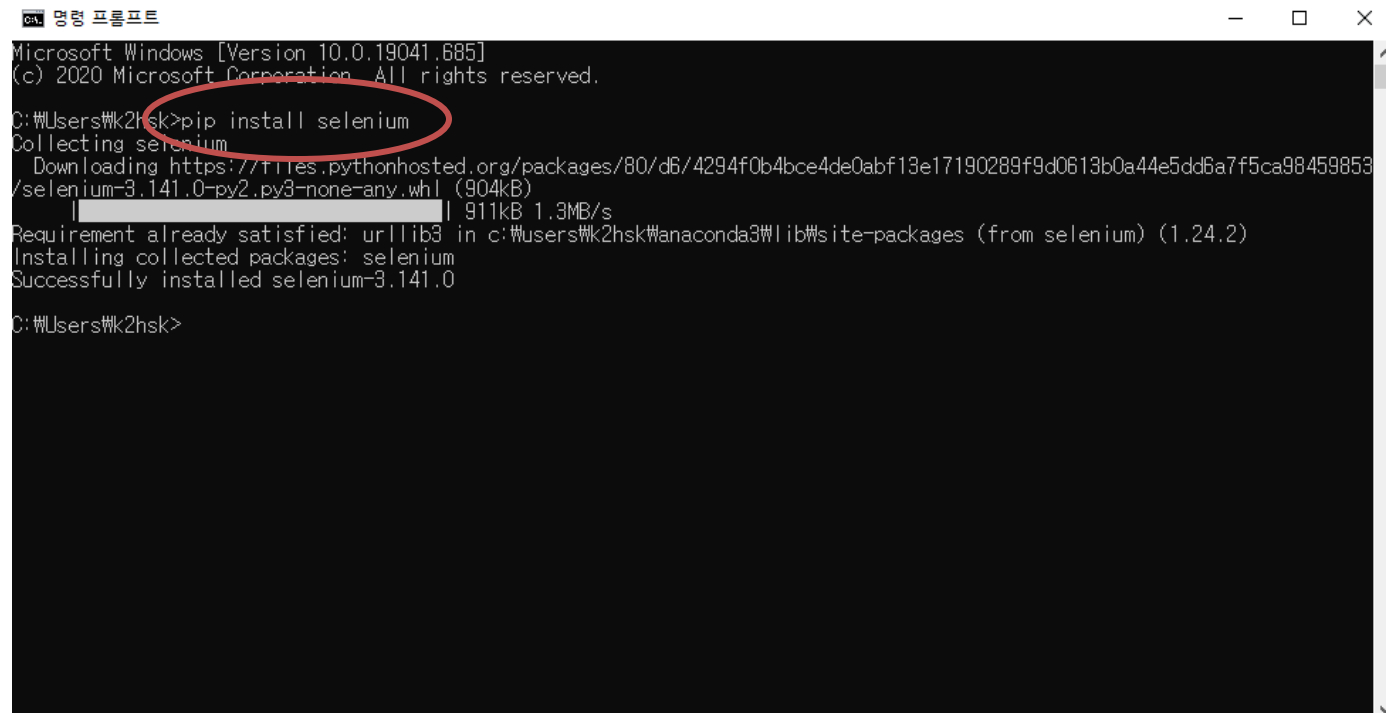
(설치 드라이브 경로 기억 : 이동식 드라이브에 설치하면 안됨)



# 셀레니움 소개

## 셀레니움 설치

- 명령프롬프트(anaconda prompt)를 열고 'pip install selenium' 을 입력하여 Selenium 설치  
(앞에서 진행했으므로 재설치하지 않아도 됨)



```
명령 프롬프트
Microsoft Windows [Version 10.0.19041.685]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\k2hsk>pip install selenium
Collecting selenium
  Downloading https://files.pythonhosted.org/packages/80/db/4294f0b4bce4de0abf13e17190289f9d0613b0a44e5dd6a7f5ca98459853/selenium-3.141.0-py2.py3-none-any.whl (904kB)
    | 911kB 1.3MB/s
Requirement already satisfied: urllib3 in c:\users\k2hsk\anaconda3\lib\site-packages (from selenium) (1.24.2)
Installing collected packages: selenium
Successfully installed selenium-3.141.0

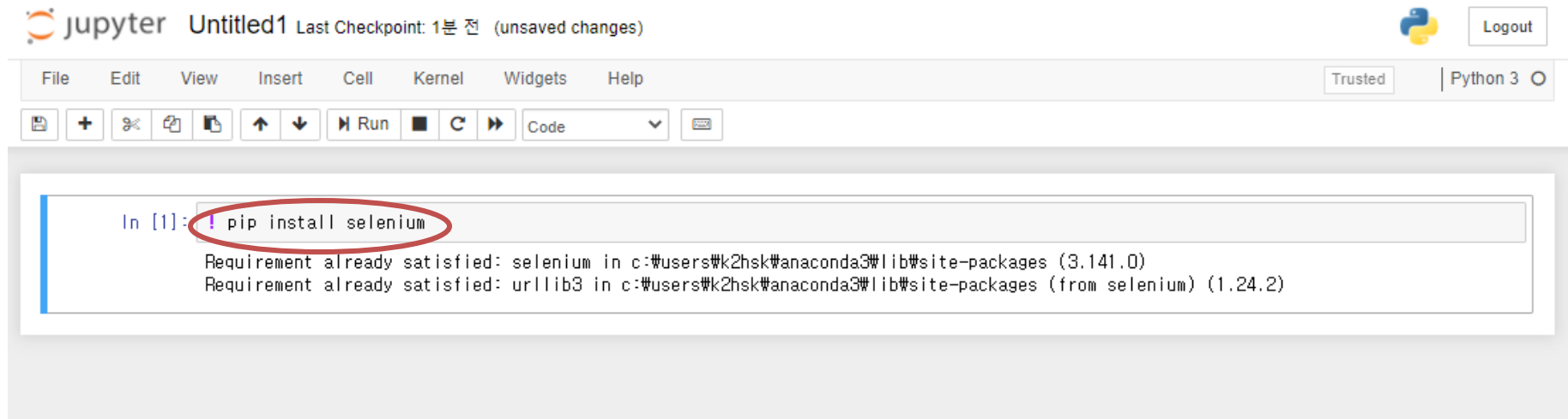
C:\Users\k2hsk>
```

# 셀레니움 소개

## 셀레니움 설치

- jupyter notebook에서는 ' ! pip install selenium ' 을 입력하여 설치("!" : 쥬피터 노트북에서 설치시 필수)

간혹 쥬피터 노트북에서 설치시 에러 발생 → cmd 상에서 설치 권장



The screenshot shows the Jupyter Notebook interface with the following components:

- Header: jupyter Untitled1 Last Checkpoint: 1분 전 (unsaved changes)
- Menu: File, Edit, View, Insert, Cell, Kernel, Widgets, Help
- Buttons: Trusted, Python 3
- Toolbar: +, -, Run, Stop, Refresh, Code
- Code Cell: In [1]: ! pip install selenium
- Output: Requirement already satisfied: selenium in c:\users\k2hsk\anaconda3\lib\site-packages (3.141.0)  
Requirement already satisfied: urllib3 in c:\users\k2hsk\anaconda3\lib\site-packages (from selenium) (1.24.2)

# 셀레니움 소개

---

## 셀레니움 준비

- 크롬 드라이브가 잘 작동하는지 확인 (드라이브의 버전에 따라 작동 오류 발생)

```
▶ from selenium import webdriver
```

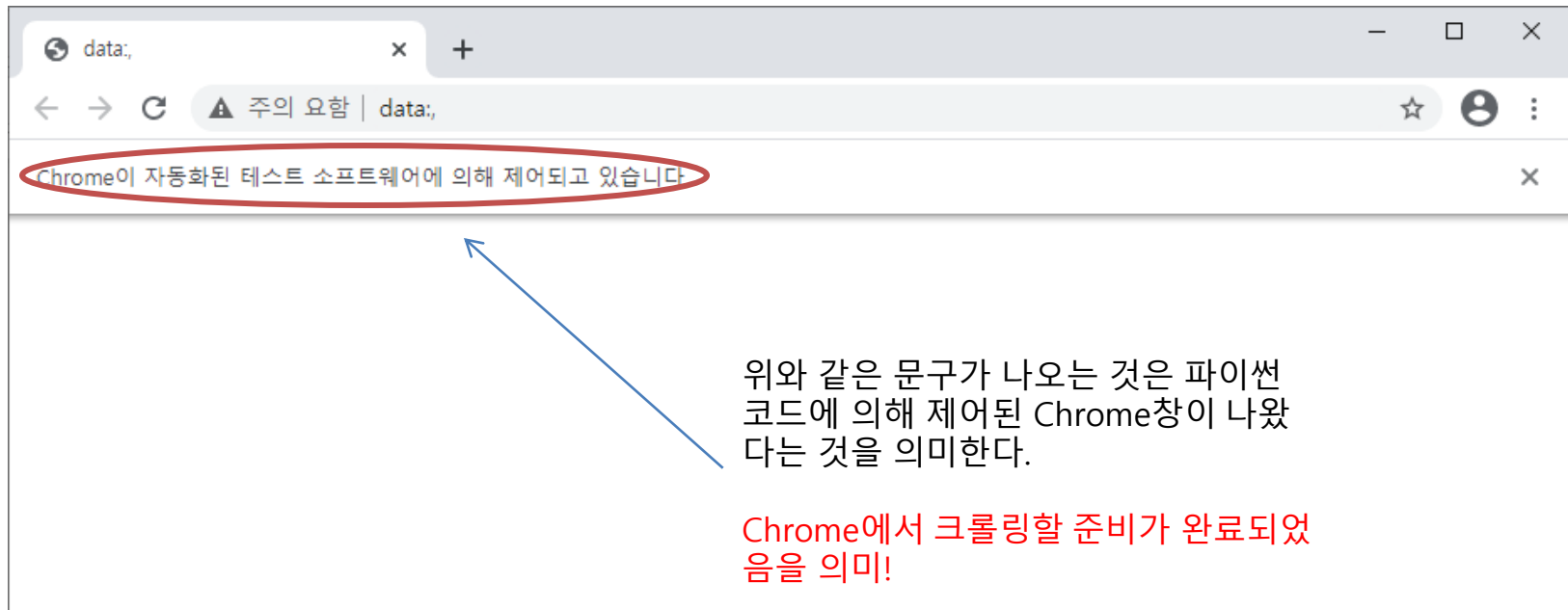
```
▶ driver = webdriver.Chrome('C:\\\\chrome\\\\chromedriver_win32\\\\chromedriver.exe')
```

크롬 드라이브가 설치되어 있는 폴더 및 파일명 지정

# 셀레니움 소개

## 웹 크롤링 준비완료

- 아래와 같이 실행된다면 성공





# 셀레니움 소개

## 크롤링에서 사용되는 Selenium 함수

함수명	설명
find_element_by_id	요소의 속성 <u>id</u> 로 찾는 오브젝트를 찾습니다.
find_element_by_class_name	요소의 속성 <u>class</u> 가 포함된 오브젝트를 찾습니다.
find_element_by_name	요소의 속성 <u>name</u> 로 찾는 오브젝트를 찾습니다.
find_element_by_xpath	<u>xpath</u> 를 이용해서 오브젝트를 찾습니다.
find_element_by_link_text	<u>하이퍼 링크의 텍스트</u> 로 오브젝트를 찾습니다.(완전 일치) - 탐색이 잘 안됩니다.
find_element_by_partial_link_text	<u>하이퍼 링크의 텍스트</u> 로 오브젝트를 찾습니다.(포함) - 탐색이 잘 안됩니다.
find_element_by_tag_name	요소의 <u>태그 이름</u> 으로 찾습니다.
find_element_by_css_selector	css selector(sizzle)로 오브젝트를 찾습니다.

## 셀레니움 소개

- URL(Uniform Resource Locator) : URL 은 웹 사이트 주소만이 아니라 컴퓨터 네트워크상의 자원을 모두 나타낼 수 있음



- **Directory** : 연결할 파일이 들어있는 폴더 디렉토리(경로명)
- **Filename** : 연결되어 보여줄 파일의 실제 이름
- **Query parameters (Query String)**: 정보에 따라서 페이지의 콘텐츠가 가변적일 수 있을 때 많이 사용, 이름과 값으로 구성되어 있으며, &로 구분한다.