

크롤링(Crawling)

1

기본설정

Crawling이란?

- Web상을 돌아다니면서 정보를 수집하는 행위.
- 필요로 되는 다양한 데이터를 수집 가능
ex) 뉴스, 유튜브 댓글, 유튜브 썸네일 등

Selenium이란?

- 브라우저 자동화, 크롤링과 관련된 라이브러리
- 웹 접속, 스크롤, 로그인 등 다양한 작업이 가능

1. 아나콘다 설치

다음의 사이트에서 anaconda설치 <https://www.anaconda.com/distribution/>



Windows



macOS



Linux

Anaconda 2020.02 for Windows Installer

Python 3.7 version

Download

64-Bit Graphical Installer (466 MB)

32-Bit Graphical Installer (423 MB)

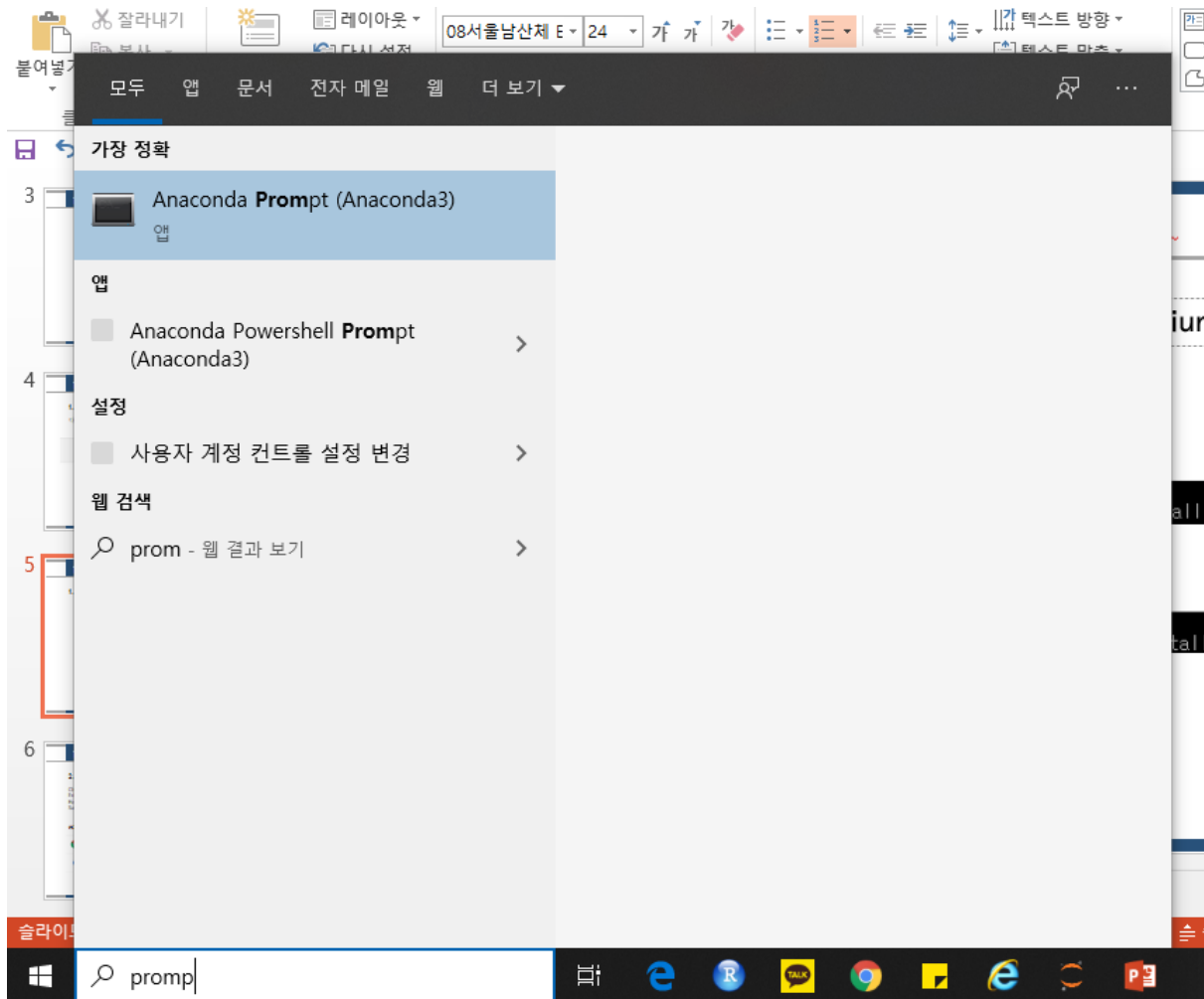
Python 2.7 version

Download

64-Bit Graphical Installer (413 MB)

32-Bit Graphical Installer (356 MB)

1. Prompt에서 'pip install selenium' & 'pip install bs4' 설치



1. Prompt에서 'pip install selenium' & 'pip install bs4'

관리자: Anaconda Prompt (Anaconda3)

```
(base) C:\WINDOWS\system32>pip install selenium
```

관리자: Anaconda Prompt (Anaconda3)

```
(base) C:\WINDOWS\system32>pip install bs4
```

2. 브라우저별 driever 설치

Chrome: <https://sites.google.com/a/chromium.org/chromedriver/downloads>

Edge: <https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/>

Firefox: <https://github.com/mozilla/geckodriver/releases>

Safari: <https://webkit.org/blog/6900/webdriver-support-in-safari-10/>

※본인 운영체제에 맞는 버전으로 드라이버를 다운로드 해주셔야 합니다. ※



Chrome



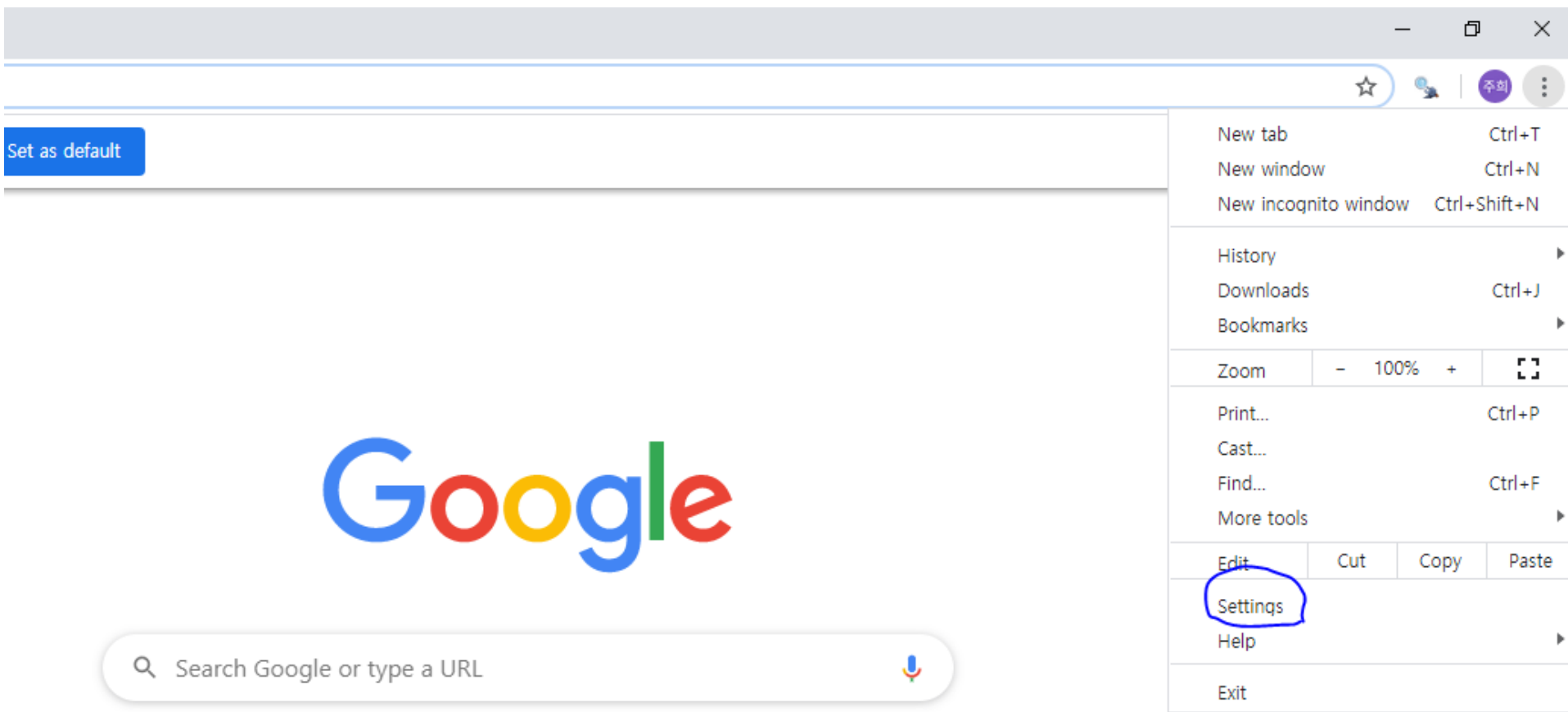
Chrome이 최신 버전입니다.

버전 80.0.3987.149(공식 빌드) (64비트)

출처: <https://sacko.tistory.com/13>

[데이터 분석하는 문과생, 싸코]

2-1. 크롬 버전 확인 방법



2-1. 크롬 버전 확인 방법


← → ↻ Chrome | chrome://settings/help

Settings

Search settings

- You and Google
- Autofill
- Privacy and security
- Appearance
- Search engine
- Default browser
- On startup
- Advanced
- Extensions
- About Chrome

About Chrome

 Google Chrome

Google Chrome is up to date
Version 80.0.3987.149 (Official Build) (64-bit)

Get help with Chrome

Report an issue

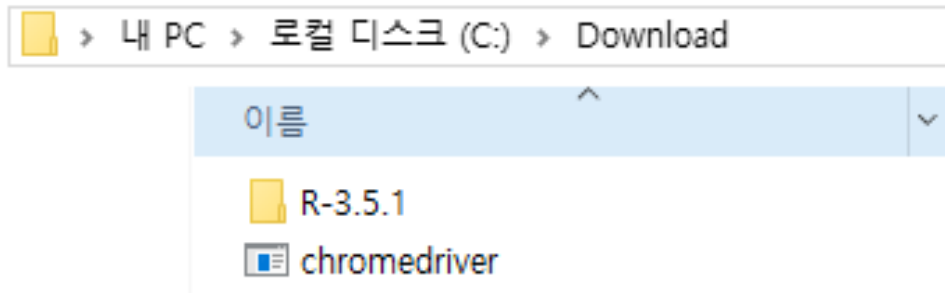
Google Chrome
Copyright 2020 Google LLC. All rights reserved.

Google Chrome is made possible by the [Chromium](#) open source project and other [open source software](#).

Google Chrome [Terms of Service](#)

2. 브라우저별 driver 설치

압축 해제 후 작업 공간에 위치시켜주세요.



본인의 경우 최종 파일 경로는 "C://Download//chromedriver.exe"

3. driver 시작

```
import os
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By

path = "C://Download//chromedriver.exe"
driver = webdriver.Chrome(path)
```

에러가 날 경우, 크롬버전 혹은 경로를 다시 한번 확인해주세요

2

크롤링 함수 설명

1. driver.get(url)

```
In [ ]: url = "http://~"  
        driver.get(url)
```

위의 함수는 지정해준 url로 이동할 수 있는 함수 입니다.

ex) driver.get(["https://www.naver.com/"](https://www.naver.com/)) -> 네이버로 이동

크롤링에서는 반복작업이 요구되기 때문에 url이 중요한 역할을 할 수 있습니다.
그렇기 때문에 만약 크롤링 하고 싶은 사이트가 있다면 url을 먼저 확인해주세요!

2. time.sleep(time)

```
In [ ]: import time  
        time.sleep(time)
```

위의 함수는 크롤러를 잠시 동안 재워주는 함수입니다.

브라우저가 크롤링을 거부하여 크롤링이 중단될 수 있습니다. 이러한 경우 다음과 같이 크롤러를 잠시 재워주세요.

ex) `time.sleep(5)` -> 5초간 크롤러를 재워라.

3. driver.find_element_by_xpath("xpath")

```
In [ ]: driver.find_element_by_xpath("xpath")
```

위의 함수는 홈페이지의 요소들을 'xpath'를 이용해 찾아주는 함수입니다.

우리는 크롤링을 할 때 홈페이지 전부가 아닌 '네이버의 실시간 검색어' 혹은 유튜브 영상의 댓글과 같이 홈페이지의 일부 요소를 긁어오고자 합니다.

*xpath란 XML 문서의 특정 요소나 속성에 접근하기 위한 경로를 지정하는 언어입니다.

4. 클릭하여 이동

```
driver.find_element_by_xpath('xpath').click()
```

다음과 같이 xpath를 찾아 클릭을 하면 해당 xpath로 이동할 수 있습니다.


```
In [ ]: source = driver.page_source  
        soup = BeautifulSoup(source, 'html.parser')  
        article = soup.select('#articleText > div:nth-child(3)')
```

`driver.page_source` – 웹페이지의 소스를 가져오는 함수입니다.

`BeautifulSoup(source, 'html.parser')` – 가져온 소스를 html로 변형시켜주는 함수입니다.

`soup.select('xpath')` – 우리가 필요한 부분만 가져올 수 있게 지정해주는 함수입니다.

3

연습을 해봅시다!

The image shows a screenshot of the Naver homepage in a web browser. The browser's address bar shows 'naver.com'. The Naver logo and search bar are visible at the top. Below the search bar, there are various links and a large red banner for AliExpress. The bottom section displays a grid of news and media logos.

The Chrome DevTools developer tools are open on the right side of the browser window. The 'Elements' panel shows the HTML structure of the page, including the head and body tags. The 'Styles' panel shows the CSS rules for the body element, including background color, font size, and font family. The 'Properties' panel shows the 'background-attachment' property set to 'scroll'.

HTML structure (Elements panel):

```
<!doctype html>
<html lang="ko">
<head>...</head>
<body class="">
  <!-- 스킵 내비게이션 -->
  <div class="u_skip">...</div>
  <!-- //스킵 내비게이션 -->
  <!-- 크롤/웨일/엣지 -->
  <div id="whale_promotion_banner" class="banner_area type_chrome">...</div>
  <a id="whale_promotion_download_file" style="display: none" href="http://update.whale.naver.net/downloads/installers/WhaleSetup.exe" download>...</a>
  <div id="PM_ID_ct" class="wrap">...</div>
  <script src="https://pm.ostatic.net/js/c/jindo v190909.js"></script>
</body>
</html>
```

CSS rules (Styles panel):

```
body {
  background-color: #fff;
  webkit-backface-visibility: hidden;
  backface-visibility: hidden;
}
body, html {
  height: 100%;
}
body, button, input,
select, table, textarea {
  font-size: 12px;
  font-family: Dotum, '돋움', Helvetica, 'Apple SD Gothic Neo', sans-serif;
}
```

Properties panel:

```
background-attachment: scroll
```

크롬에서 F12를 누르면 다음과 같은 창이 뜹니다.

NAVER whale 인터넷의 새로운 시작! 네이버 웨일로 차원이 다른 웹서핑을 경험해보세요

NAVER

매일 카페 블로그 지식iN 쇼핑 Pay TV 사전 뉴스 증권 부동산 지도 영화 뮤직 책 웹툰 더보기

AliExpress 알리익스프레스 10주년 감사세일
넘치는 즐거움 넘치는 50% 이상 할인

연합뉴스 > 서울지하철, 4월1일부터 밤 12시까지만 열차운행

네이버뉴스 연예 스포츠 총선

뉴스스탠드 > 전체 언론사 MY 뉴스

이데일리	뉴스라가	노컷뉴스	서울경제	헤럴드경제	KBS
KBS WORLD	NEWSIS	한국일보	한국경제	The Korea Herald	시사IN
TOPDAILY	MONEY	smartPC	Insight	DAILY NK	환경경제신문 그린포스트코리아

Elements

```

<!doctype html>
<html lang="ko">
<head>...</head>
<body class="">
  <!-- 스킵 내비게이션 -->
  <div class="u_skip">...</div>
  <!-- //스킵 내비게이션 -->
  <!-- 크롤/웨일/엣지 -->
  <div id="whale_promotion_banner" class="banner_area type_chrome">...</div>
  <a id="whale_promotion_download_file" style="display: none" href="http://update.whale.naver.net/downloads/installers/WhaleSetup.exe" download></a>
  <div id="PM_ID_ct" class="wrap">...</div>
  <script src="https://pm.ostatic.net/js/c/jindo v190909.js"></script>
</body>
</html>

```

html body

Styles

```

element.style {
}
body {
  background-color: #fff;
  backface-visibility: hidden;
  backface-visibility: hidden;
}
body, html {
  height: 100%;
}
body, button, input,
select, table, textarea {
  font-size: 12px;
  font-family: Dotum, '돋움', Helvetica, 'Apple
SD Gothic Neo', sans-serif;
}

```

DOM Breakpoints

Properties

Accessibility

margin

border

padding

794 x 640

Filter

Show all

backface-visibility hidden

background-attachment scroll

background-clip

빨간 동그라미로 표시된 곳을 눌러보세요.

The image shows a screenshot of the Naver homepage. A design tool overlay is visible on the left side of the page, showing the dimensions (29.41 x 20) and color (#000000) of the 'a.nav' element. The Chrome DevTools Elements panel is open on the right, showing the HTML structure of the page. The 'a.nav' element is highlighted in the 'li.nav_item' list, with its href attribute set to 'https://news.naver.com/'.

NAVER

보기 편하고 찾기 쉬운 모습으로 달라진 네이버를 만나세요!

NAVER

메일 카페 블로그 지식iN 쇼핑 Pay TV 사진 뉴스 증권 부동산 지도 영화 음악 책

지금! pm11:50~am1:00
실로 놀라운 침대 앤셀, 4월 마지막 방송!

연합뉴스 > 예결위 추경안 심사... "기부로 연대 발취" vs "취종고 ...

네이버뉴스 · 연예 스포츠 경제

뉴스스탠드 > 구독한 언론사 · 전체언론사

뉴스타파

2020.04.28. 18:11 편집

Chrome DevTools Elements panel:

```

<div id="gnb">
  <div id="NM_FAVORITE" class="gnb_inner">
    <div class="group_nav">
      <ul class="list_nav type_fix">...</ul>
      <ul class="list_nav NM_FAVORITE_LIST">
        <li class="nav_item">...</li>
        <li class="nav_item">...</li>
        <a href="https://news.naver.com/" class="nav" data-clk=
          "svc.news">뉴스</a> == $0
        </li>
        <li class="nav_item">...</li>
        <li class="nav_item">...</li>
        <li class="nav_item">...</li>
        <li class="nav_item">...</li>
        <li class="nav_item">...</li>
      </ul>
    </div>
  </div>

```

그리고 원하는 요소를 클릭하면 다음과 같이 xpath
를 확인 할 수 있습니다.

The screenshot shows the Naver homepage in a web browser. The Chrome DevTools 'Elements' panel is open on the right, displaying the DOM tree. A context menu is open over a navigation item, with the 'Copy' option selected, which has opened a sub-menu where 'Copy XPath' is highlighted. The DOM tree shows a list of navigation items with the class 'list_nav NM_FAVORITE_LIST'. The context menu options include: Add attribute, Edit as HTML, Delete element, Copy, Hide element, Force state, Break on, Expand recursively, Collapse children, Scroll into view, Focus, Store as global variable, Cut element, Copy element, Paste element, Copy outerHTML, Copy selector, Copy JS path, Copy styles, Copy XPath, and Copy full XPath.

Xpath는 다음과 같이 우클릭을 하면 복사할 수 있습니다.



THANK YOU

