

Crawling

17기 Engineering team 윤수진

크롤링?

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

크롤링(crawling)

파싱(parsing)

스크래핑(scraping)

크롤링?

1. HTML 페이지를 가져와서, HTML/CSS등을 파싱하고, 필요한 데이터만 추출하는 기법

HTML (Hyper Text Markup Language)

:웹 사이트를 생성하기 위한 언어로 문서와 문서가 링크로 연결되어있고, 태그를 사용하는 언어

HTML문서의 기본 구조

<태그명 속성1 = '속성값1' 속성2 = '속성값2' > Value </태그명>

<a href="<https://news.naver.com/>" class="nav" data-clk="svc.news">뉴스

크롤링?

```
<html lang="ko" data-dark="false" data-useragent="Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
e/87.0.4280.141 Safari/537.36">
  <head>...</head>
  <body>
    <div id="u_skip">...</div>
    <div id="wrap">
      <style type="text/css">...</style>
      <div id="NM_TOP_BANNER" data-clk-prefix="top" class="_1hiMWemA" style="background-color: #f3f1ff;">...</div>
      <div id="header" role="banner">
        <div class="special_bg">...</div>
        <!--EMPTY-->
        <div id="gnb">
          <div id="NM_FAVORITE" class="gnb_inner">
            <div class="group_nav">
              <ul class="list_nav type_fix">
                <li class="nav_item">
                  <a href="https://mail.naver.com/" class="nav" data-clk="svc.mail">...</a>
```

〈html〉 : html 문서의 시작과 끝

〈head〉 : 문서의 머리 (브라우저에 직접적으로 보이지 않음)

〈body〉 : 문서의 콘텐츠 (브라우저에 직접적으로 보이는 부분)

〈div〉 : 문서 구역(공간)을 나누는 기준

〈h〉 : 머리글

〈p〉 : 단락을 나누는 기준

크롤링?

2. Open API를 제공하는 서비스에 open API를 호출해서, 받은 데이터 중 필요한 데이터만 추출하는 기법

NAVER DevelopersProductsDocumentsApplicationNAVER D2SupportForumAPI 상태Search Here

Products | 네이버에서 제공하는 다양한 서비스와 콘텐츠를 소개합니다.

Products > API 이용 안내 > API 소개

API 이용 안내

API 소개
운영 정책
FAQ
BI 가이드
이용약관
상표사용 가이드

CLOVA
네이버 아이디로 로그인
파파고

네이버 오픈 API 목록

네이버 오픈API 목록 및 안내입니다.

API명	설명	호출제한
검색	네이버 블로그, 이미지, 웹, 뉴스, 백과사전, 책, 카페, 지식IN 등 검색	25,000회/일
네이버 아이디로 로그인	외부 사이트에서 네이버 아이디로 로그인 기능 구현	없음
네이버 회원 프로필 조회	네이버 회원 이름, 이메일 주소, 휴대전화번호, 별명, 성별, 생일, 연령대, 출생연도, 프로필 조회	없음

크롤링?

1.



BeautifulSoup

Request

: HTTP Request를 파이썬에서 가능하게 해주는 모듈

ex. `response = requests.get('https://www.naver.com/')`

➤ 'https://www.naver.com/' 주소에 get 요청을 하고 받은 응답(HTML등)을 저장

Beautiful Soup

: HTML과 같은 문서를 가져와 파싱해주는 모듈

크롤링?

2. Selenium

: 웹사이트 테스트 자동화용 모듈 (실제 사용자가 사용하는 것 처럼 동작)

크롤링?



장점 :
거의 모든 상황에서 크롤링 가능

단점 :
속도가 느림



BeautifulSoup

장점 :
심플함, 속도가 빠름

단점 :
크롤링 불가능한 경우가 있을 수 있음



Selenium

실습!

Selenium

1. 크롬 웹 드라이버 설치

chrome://version/

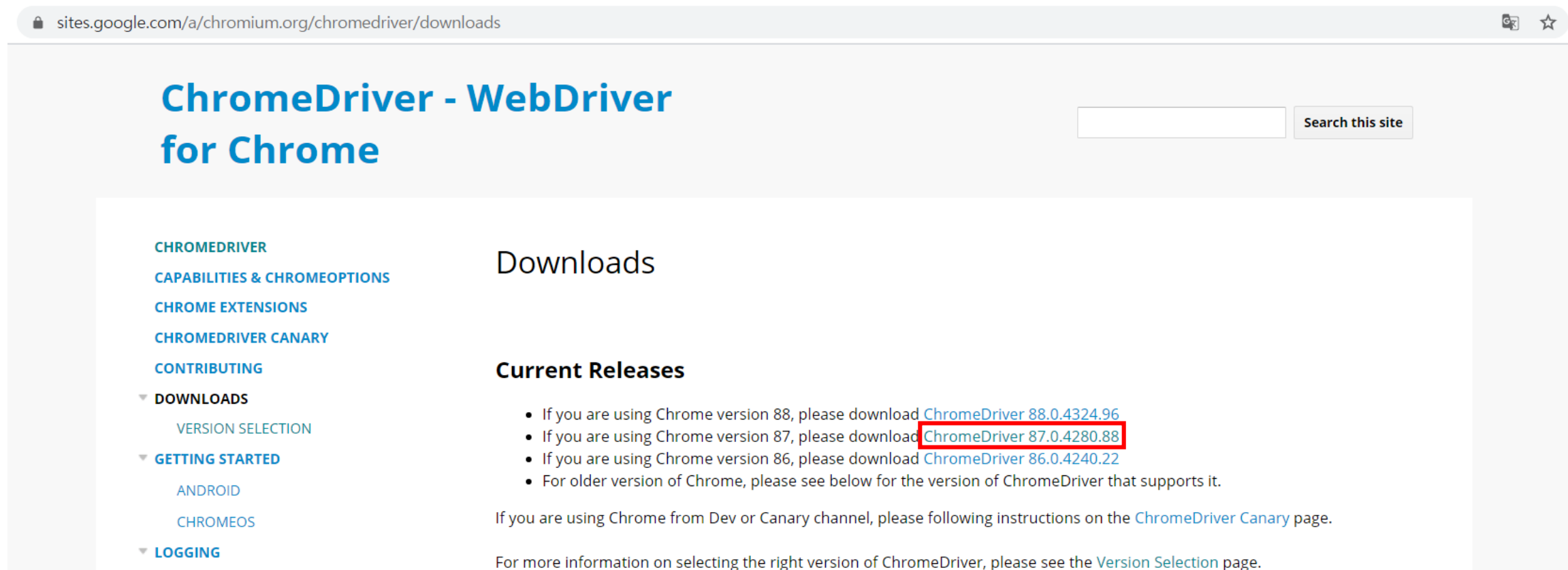
Chrome | chrome://version

Chrome: 87.0.4280.141 (공식 빌드) (64비트) (cohort: Stable)

Selenium

1. 크롬 웹 드라이버 설치

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



The screenshot shows the official ChromeDriver website. The browser's address bar displays the URL `sites.google.com/a/chromium.org/chromedriver/downloads`. The page features a navigation menu on the left with links to CHROMEDRIVER, CAPABILITIES & CHROMEOPTIONS, CHROME EXTENSIONS, CHROMEDRIVER CANARY, CONTRIBUTING, DOWNLOADS (which is expanded to show VERSION SELECTION), GETTING STARTED (with links to ANDROID and CHROMEOS), and LOGGING. The main content area is titled 'Downloads' and includes a 'Current Releases' section with a bulleted list of download links for different Chrome versions. The link for Chrome version 87, 'ChromeDriver 87.0.4280.88', is highlighted with a red box. Below the list, there is a note about Dev or Canary channels and a link to the 'ChromeDriver Canary' page, followed by a link to the 'Version Selection' page for more information.

ChromeDriver - WebDriver for Chrome

Search this site

CHROMEDRIVER
CAPABILITIES & CHROMEOPTIONS
CHROME EXTENSIONS
CHROMEDRIVER CANARY
CONTRIBUTING
▼ DOWNLOADS
 VERSION SELECTION
▼ GETTING STARTED
 ANDROID
 CHROMEOS
▼ LOGGING

Downloads

Current Releases

- If you are using Chrome version 88, please download [ChromeDriver 88.0.4324.96](#)
- If you are using Chrome version 87, please download [ChromeDriver 87.0.4280.88](#)
- If you are using Chrome version 86, please download [ChromeDriver 86.0.4240.22](#)
- 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.

Selenium

1. 크롬 웹 드라이버 설치


Index of /87.0.4280.88/

	<u>Name</u>	Last modified	Size	ETag
	Parent Directory		-	
	chromedriver_linux64.zip	2021-01-06 20:30:29	5.84MB	1dd81cad235eb14478543d67c27d351d
	chromedriver_mac64.zip	2021-01-06 20:30:31	7.64MB	82aeaa647533e937c3a7eb68d28ed5ba
	chromedriver_mac64_m1.zip	2021-01-06 20:30:33	6.89MB	69d0807ba09a5bf61c9c14c1f217f95f
	chromedriver_win32.zip	2021-01-06 20:30:35	5.18MB	3821c48e3f61ea843dc3545341737854
	notes.txt	2020-12-03 18:14:25	0.00MB	af9a0d8be8211669cb9fabda9471b651

 chromedriver.exe

2020-12-02 오전 2:33

응용 프로그램

 naverblogcrawler.ipynb

2020-12-22 오후 2:46

IPYNB 파일

ETC

데이터 수집부터 겁나는 기업들... "불법인지 합법인지 모르겠다"

조선비즈 | 박현익 기자



입력 2020.09.23 06:00 | 수정 2020.09.23 06:53

데이터3법 시행은 됐는데...

공개된 정보, 무턱대고 가져왔다가 DB권 침해 소지

데이터 수집 방법인 '크롤링'... "안 쓰는 기업 없다"

"무조건 위법은 부적절... 어디까지 괜찮은지 합의 필요"

"데이터3법 활성화 발목"... 명확한 기준 정립 시급



픽사베이

#속박업소 플랫폼 '여기어때'를 운영하는 위드노베이션의 심명섭 전 대표는 경쟁사 '야놀자'의 데이터를 무단 복제했다는 혐의로 지난 2월 법원에서 징역 1년2개월에 집행유예 2년을 선고받았다. 심 전 대표는 법정에서 "닐리 행해지는 정보수집 방법을 통해 공개된 정보를 수집했을 뿐"이라고 주장했지만 1심 법원은 "피해 회사인 야놀자 측 의사에 반해 데이터를 가져온 것"이라며 "타인의 정보통신망에 대한 무단침입"이라고 판단했다.

사이트명+robots.txt 확인하기!

*robots.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

User-agent: *

Disallow: /

Allow : /\$

과제

네이버 검색창에 '제주도' 검색해서 나오는 기사 제목
한페이지 크롤링한 화면 캡처해서 제출 (Selenium/bs4중 선택)