

제 14 강 Web Scraping (1)

학습 목차

- 웹 스크래핑 절차
- 구글 검색

웹 스크래핑(Web Scraping)

- 웹 사이트로부터 필요로 하는 데이터를 추출하는 것.

Web scraping 절차

- 웹 사이트 성격에 따른 scraping 방식 결정
- 웹 구조 분석, 데이터 검색 및 인터랙션 절차 파악
- 실제 데이터 추출 및 가공
- 데이터 저장

Scraping 방식 결정

- 아예 원하는 데이터를 액세스할 수 있는 API를 제공하는가?
 - requests 를 이용하여 API 호출
- 간단한 URL 만으로 원하는 데이터에 접근할 수 있는가?
 - BeautifulSoup, Scrapy 등의 검색 추출 도구 활용
- 사용자의 인터랙션이 필요한가?
 - 인터랙션 자동화 도구 활용(MechanicalSoup, Selenium, Playwright 등)

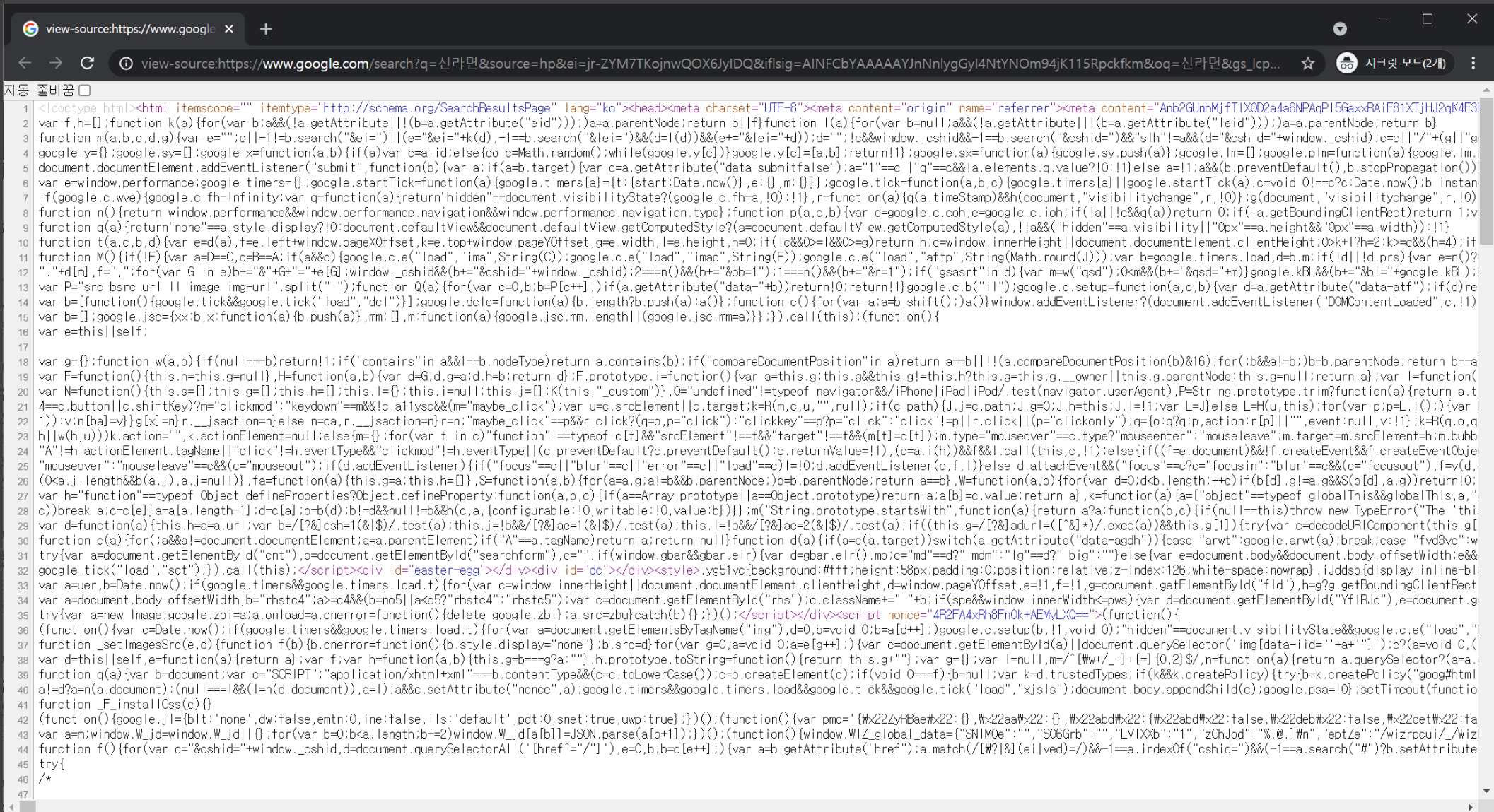
Note: 실제로는 복합적인 접근이 필요한 경우가 대다수

과업 #1

- 구글 검색 결과를 “제목”, “URL 주소” 형태의 CSV 파일로 만들기.

The screenshot shows a Google search for '신라면' (Shin Ramen). The search bar at the top contains '신라면'. Below the search bar, there are tabs for '전체' (All), '이미지' (Images), '동영상' (Videos), '뉴스' (News), '쇼핑' (Shopping), and '더보기' (More). The search results are displayed in a list format. The first result is from 'namu.wiki' with the title '신라면 - 나무위키'. The second result is from 'shinramyun.com' with the title '농심 신라면'. The third result is from 'brand.nongshim.com' with the title '신라면 | 브랜드관'. The fourth result is from 'ko.wikipedia.org' with the title '신라면 - 위키백과, 우리 모두의 백과사전'. The fifth result is from 'emart.ssg.com' with the title '신라면'. On the right side of the search results, there are two product listings for '신라면' (Shin Ramen) from '농심' (Nongshim). The first listing is for '신라면 120gx30개입' (Shin Ramen 120gx30 packs) with a price of ₩21,800 and a shipping fee of ₩2,500. The second listing is for '신라면 120g x40' (Shin Ramen 120g x40) with a price of ₩32,960 and free shipping. Below the product listings, there is a section titled 'Google은 상품 판매의 당사자가 아닙니다.' (Google is not the merchant of the goods) with a link to 'Google에서 더보기' (See more on Google). At the bottom of the search results, there is a section titled '라면은 역시 신라면' (Ramen is still Shin Ramen) with a link to '사나이 올리는 신라면' (Shin Ramen for men). This section includes a collage of images showing various Shin Ramen products and packaging.

페이지 소스 보기 - Ctrl + U (크롬)



검사(Inspect) - F12 또는 Ctrl+Shift+I

The screenshot shows a Google search for '신라면' (Shin Ramyun). The search results are displayed on the left, and the developer tools (F12) are open on the right, showing the HTML structure of the search results.

Search Results:

- 검색결과 약 2,320,000개 (0.52초)
- <https://namu.wiki> > 신라면 ▾
신라면 - 나무위키
신라면이 더 이상 농심 라인업에 프리미엄 라면이 아닌 것이 되면서 나타나는 지속 하락은 막을 수 없게 되었다. 원가 절감을 아무리 ...
2021. 4. 21. · 업로더: 농심기획(NongShim Communications)
신라면 블랙 · 농심라면 · 농심 메밀소바 · 신라면블랙컵
- <http://www.shinramyun.com> ▾
농심 신라면
오랜시간 변함없이 사랑받아 온 한국인의 매운맛, 사나이울리는 신라면, 세계인의 한 한국의 매운맛 농심 신라면.
- <http://brand.nongshim.com> > shinramyun > main ▾
신라면 | 브랜드관
신라면은 소고기를 주원료로 완전진공 농축 시킨 소고기 엑기스에 각종 천연 양념 만든, 전통 가정요리로부터 유래된 얼큰한 소고기국 맛의 제품 ...
- <https://ko.wikipedia.org> > wiki > 신라면 ▾
신라면 - 위키백과, 우리 모두의 백과사전
신라면은 1986년 10월 대한민국에서 출시된 대한민국의 식품회사인 농심의 라면(고; 2 이름; 3 원재료; 4 종류. 4.1 포장 구분; 4.2 맛 구분; 4.3 할랄 ...

Developer Tools (F12) HTML Structure:

```
<div id="center_col">
  <style>...</style>
  <div id="taw">...</div>
  <div class="eqAnXb" id="res" role="main">
    <div id="topstuff">...</div>
    <div id="search">
      <div data-hveid="CAEQNA" data-ved="2ahUKEwj1r4qboMDwAhWQZt4KHT2IDw0QGnoECAEQNA">
        <h1 class="Uo8X3b">검색결과</h1>
        <div eid="kr-ZYPWeBJDN-Qa9kL5o" data-async-context="query:%EC%8B%A0%EB%9D%BC%EB%A9%B4" id="rso">
          <div class="g">
            <h2 class="Uo8X3b">웹 검색결과</h2>
            <div data-hveid="CAUQAA" data-ved="2ahUKEwj1r4qboMDwAhWQZt4KHT2IDw0QtgIoADAAegQIBRA">
              <div class="tf2Cxc">
                <div class="yuRuf">
                  <a href="https://namu.wiki/w/%EC%8B%A0%EB%9D%BC%EB%A9%B4" data-ved="2ahUKEwj1r4qboMDwAhWQZt4KHT2IDw0QtgIoADAAegQIBRA" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%25EB%259D%25BC%25EB%25A9%25B4&ved=2ahUKEwj1r4qboMDwAhWQZt4KHT2IDw0QtgIoADAAegQIBRA">
                    <br>
                    <h3 class="LC201b DKV0Md">신라면 - 나무위키</h3> == $0
                    <div class="TbwUpd NJjxre">...</div>
                    </a>
                    <div class="B6fmyf">...</div>
                    </div>
                    <div class="IsZvec">...</div>
                    <div jscontroller="m6a0l" id="eob_15" jsdata="fxg5tf;_A4jth4" jsaction="rcuq6b:nPT2md" data-ved="2ahUKEwj1r4qboMDwAhWQZt4KHT2IDw0Q2Z08MAB6BAGFEAo">...</div>
                    </div>
                    </div>
                    <div class="g">...</div>
                    <div class="g">...</div>
                    <div class="g">...</div>
                    <script nonce="4R2FA4xRh8Fn0k+AEMyLXQ==">...</script>
                    <style>...</style>
                    <span id="fld"></span>
                    <div class="g">...</div>
                    <div class="g">...</div>
                    <div class="hlcw0c">...</div>
                    <div class="ULSxyf">...</div>
                  </div>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
```

Styles Panel:

- search?..hov .cls**
element.
style {
}
- search?..#res h3, #botstuff h3 {**
font-size: 20px;
line-height: 1.3;
- search?..DKV0Md {**
padding-top: 4px;
- padding-top: 5px;**
- search?..LC201b {**
display: inline-block;
line-height: 1.3;
margin-bottom: 3px;

HTML 문서 구조 - 엘리먼트 트리 구조

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
    "http://www.w3.org/TR/html4/loose.dtd">
<html>

  <head>
    <title>The Dormouse's story</title>
  </head>

  <body>
    <p class="title"><b>The Dormouse's story</b></p>

    <p class="story">Once upon a time there were three little sisters; and their names were
      <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
      <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
      <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
      and they lived at the bottom of a well.</p>
  </body>

</html>
```

Scraping 방식의 결정 - URL 만으로 원하는 정보 추출 가능

- <https://www.google.com/search?q=신라면>

The screenshot shows a Google search for '신라면' (Shin Ramyun) on a desktop browser. The search bar at the top contains the text '신라면'. Below the search bar, there are tabs for '전체' (All), '이미지' (Images), '동영상' (Videos), '뉴스' (News), '쇼핑' (Shopping), and '더보기' (More). The search results show approximately 2,320,000 results in 0.48 seconds.

The first search result is from <https://namu.wiki> titled '신라면 - 나무위키'. The snippet mentions that Shin Ramyun is a premium ramen brand and has been a popular product for a long time.

The second search result is from <http://www.shinramyun.com> titled '농심 신라면'. The snippet describes it as a Korean instant noodle brand with a long history.

The third search result is from <http://brand.nongshim.com> titled '신라면 | 브랜드관'. The snippet mentions that Shin Ramyun is a popular instant noodle brand.

The fourth search result is from <https://ko.wikipedia.org> titled '신라면 - 위키백과, 우리 모두의 백과사전'. The snippet provides a brief history of the brand, mentioning its founding in 1986.

On the right side of the search results, there is a knowledge panel for '신라면'. It includes a title '신라면' with a share icon, a description '신라면은 1986년 10월 대한민국에서 출시된 대한민국의 식품회사인 농심의 라면이다. 위키백과', and several facts: '원산지: 대한민국', '제조사: 농심', '만들어진 연도: 1986년 10월 1일', '음식 에너지 (120 g 음식 당): 500 kcal (2093 kJ)', and '비슷한 음식: 삼양 삼양라면'. At the bottom of the panel, there is a section '관련 검색어' (Related search terms) with a link to '5개 이상 항목 더보기' (View more than 5 items).

BeautifulSoup

- Html 문서를 파이썬 객체들의 트리 형태로 구조화하여 처리하는 모듈.
- 트리 구조 상의 오브젝트들을 손쉽게 검색할 수 있음.
- 구성 요소의 검색, 변경, 출력 함수들을 제공함.
- BeautifulSoup4 패키지 설치, lxml 파서 라이브러리도 함께 설치



검색 기본

```
from bs4 import BeautifulSoup

f = open('alice.html')

soup = BeautifulSoup(f, 'lxml')

f.close()

soup.title
soup.title.string
soup.title.parent.name

soup.p
soup.p['class']

soup.a
soup.find_all('a')
soup.find(id='link3')
soup.find('a', attrs={'id': 'link3'})
for link in soup.find_all('a'):
    print(link['href'])

# soup.select('a')
# soup.select('#link3')
# soup.select('a[id="link3"]')
```

Google 신라면

전체 이미지 쇼핑 동영상 뉴스 더보기

검색결과 약 3,230,000개 (0.25초)

h3.LC201b.MBeuO.DKV0Md 157.78 × 31

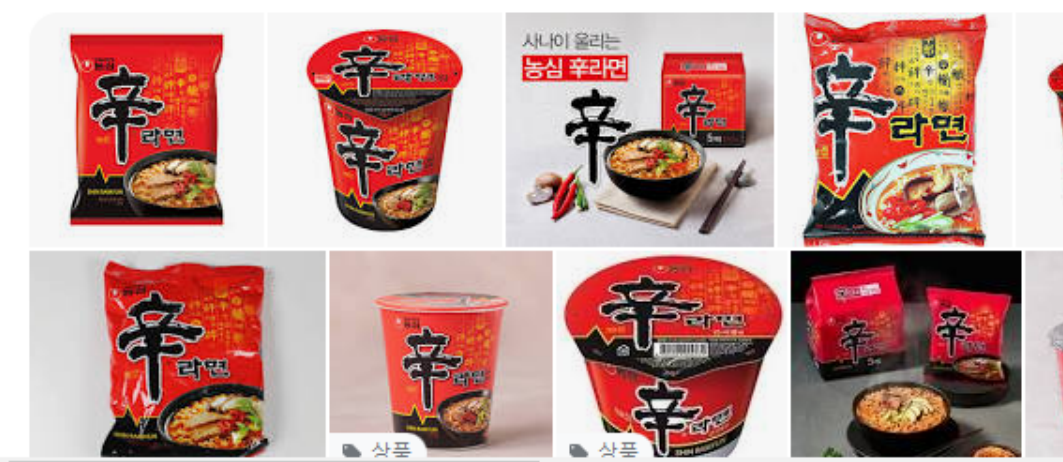
https://namu.wiki/신라면

신라면 - 나무위키

2023. 4. 25. — 신라면은 '매울 신(辛)'자의 '맵다'는 뜻과 농심 신준호 회장의 성을 동시에 의미하는 중의적인 글자이다.[1] 농심 메밀소바의 전신인 짭어먹는 춘면의 '...

신라면 관련 이미지

컵라면 수채화 농심 옛날 멀티 일본



DevTools is now available in Korean! Always match Chrome's language Switch DevTools to Korean Don't show again

Elements Console Sources Network Performance Memory Application

```
lang="ko" style="width:652px" jsaction="QyLbLe:OMITjf;ewaord:qsYrDe;xd28Mb:A6j43c" data-hveid="CBQQAA" data-ved="2ahUKEwiir4jOs07-AhUDQ94KHdD0AtEQFSgAegQIFBAA">
<div class="GLI88c UK95Uc" data-snc="ih6Jnb_LGyvfv">
  <div class="Z26q7c UK95Uc jGGQ5e VGXe8" data-snf="x5WNvb" data-snhf="0" style="grid-area:x5WNvb">
    <div class="yuRUBf">
      <a href="https://namu.wiki/w/%EC%8B%A0%EB%9D%BC%EB%A9%B4" data-ved="2ahUKEwiir4jOs07-AhUDQ94KHdD0AtEQFnoECAsQAQ" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%25EB%259D%25BC%25EB%25A9%25B4&ved=2ahUKEwiir4jOs07-AhUDQ94KHdD0AtEQFnoECAsQAQ">
        <br>
        <h3 class="LC201b MBeuO DKV0Md">신라면 - 나무위키
        </h3>
        <div class="TbwUpd NJjxre iUh30 apx8Vc oJE3Fb">
          <div class="B6fmyf byrV5b Mg1HEd">
            <div class="Z26q7c UK95Uc Sth6v" data-sncf="0,1,2" data-snf="Vjbam" style="padding-left:20px;grid-area:Vjbam;width:92px">
              <div class="Z26q7c UK95Uc VGXe8" data-sncf="2" data-snf="nke7rc" style="grid-area:nke7rc">
                <div class="Z26q7c UK95Uc VGXe8" data-snf="bvRF1f" style="grid-area:bvRF1f">
                  <span id="z9PoV"></span>
                  <script nonce=></script>
                </div>
              <div class="ULSxyf">
                <div class="MidYud">
```

request 결과를 저장

```
import re
import requests
from bs4 import BeautifulSoup

keyword = '신라면'
url = f'https://www.google.com/search?q={keyword}'

r = requests.get(url)
r.raise_for_status()

soup = BeautifulSoup(r.text, features='lxml')
elms = soup.find_all('h3')

with open('downloaded.html', 'w', encoding='utf-8') as wf:
    wf.write(r.text)
```

User Agent

- 웹에 접근하여 액세스하는 사용자의 정보
- 웹 서버는 이를 근거로 user agent 마다 다른 내용의 응답을 함.

HTTP에서의 사용 [\[편집 \]](#)

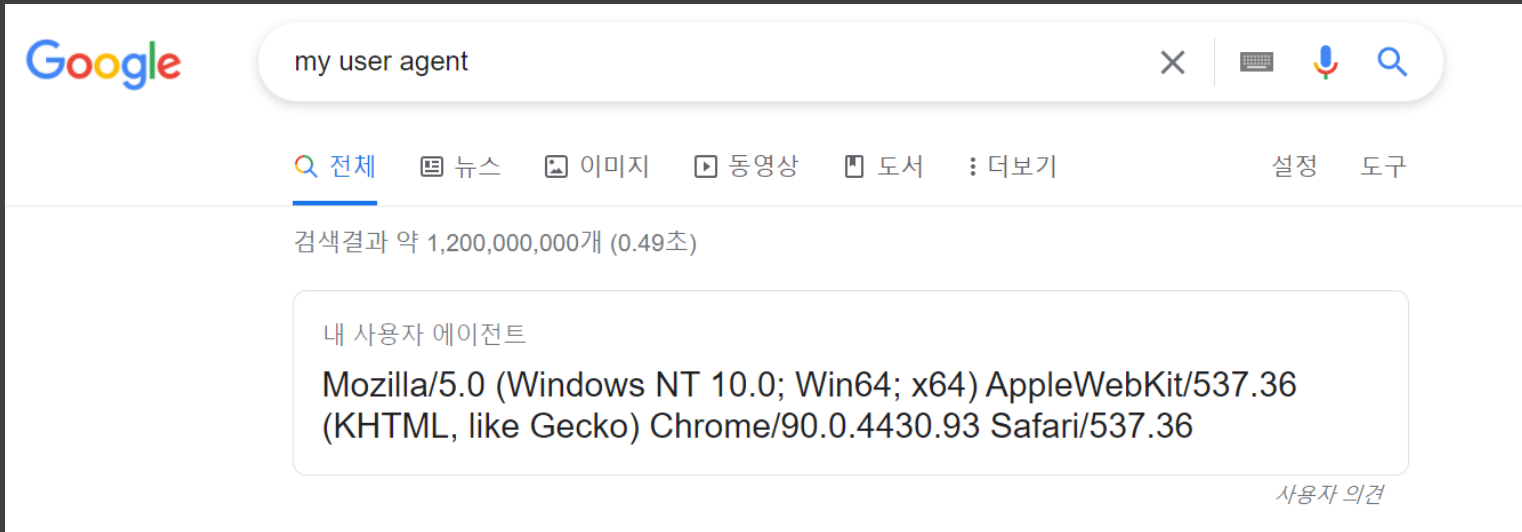
인간이 조작하는 웹 브라우저 형식 [\[편집 \]](#)

```
Mozilla/5.0 (iPad; U; CPU OS 3_2_1 like Mac OS X; en-us) AppleWebKit/531.21.10 (KHTML, like Gecko) Mobile/7B405
```

자동화된 에이전트(봇)의 형식 [\[편집 \]](#)

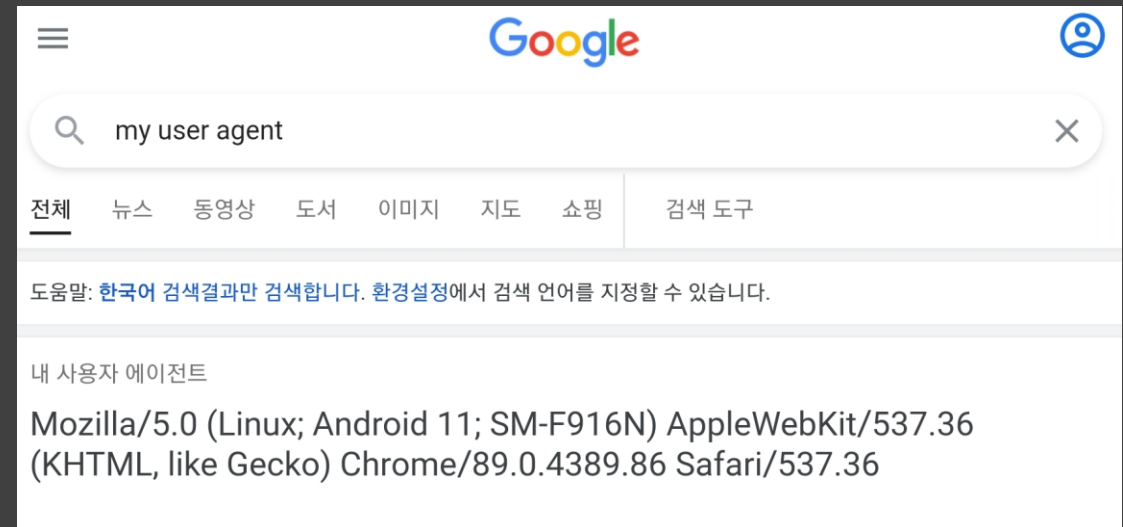
```
Googlebot/2.1 (+http://www.google.com/bot.html)
```


My User Agent



PC

스마트폰



User agent header 정보 제공

```
headers = {  
    'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)  
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.54 Safari/537.36'  
}  
  
r = requests.get(url, headers=headers)  
r.raise_for_status()  
  
soup = BeautifulSoup(r.text, features='lxml')
```

타이틀 추출

```

▼<div class="g tF2Cxc" style="width:652px" data-hveid="CB4QAA" data-ved="2ahUKEwi3t60VwN73AA">
  ▼<div class="kwxLod" data-sokoban-container="ih6Jnb_LGyvvf"> flex
    ▼<div class="jtfYYd" style="flex-grow:1"> flex
      ▼<div class="NJo7tc Z26q7c jGGQ5e" data-header-feature="0">
        ▼<div class="yuRUBf">
          ▼<a href="https://namu.wiki/w/%EC%8B%A0%EB%9D%8C%EB%A9%B4" data-ved="2ahUKEwi3t60VwN73AA" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%25B4&ved=2ahUKEwi3t60VwN73AhVDA4KHTtsBtwQFnoECAoAQ">
            <br>
            <h3 class="LC201b MBeuO DKV0Md">신라면 - 나무위키:대문</h3> == $0
          ▶<div class="TbwUpd NJjxre">...</div>
        </a>
        ▶<div class="B6fmyf">...</div>
      </div>
    </div>
  </div>

```

```
for e in elms:
    print(e.find('h3').string)
```

하지만, 찾지 못하는 것이 있다.

```
> <div class="g tF2Cxc" style="width:600px" data-hveid="CCcQAA" data-ved="AA">...</div>  
> <div data-hveid="CEMQAA">...</div> == $0  
> <div class="hlcw0c">...</div>  
> <div class="ULSxyf">...</div>  
> <div class="hlcw0c">...</div>  
</div>
```

<http://prod.danawa.com> > info ▼

농심 신라면 120g (20개) : 다나와 가격비교

라면중에 최고 내가 태어나 제일 많이 먹어봤던것중에 맛도 찐이고 언제 어디서나 끓여 먹어도 맛 있고 부서 먹어도 맛 좋고 아주 자주 잘 먹는 농심 신라면!!!

2021. 8. 23. · ★★★★★ 평점: 4.7 · 리뷰 29,200개 · 최저가: ₩11,330

<http://prod.danawa.com> > info ▼

농심 신라면 120g (1개) : 다나와 가격비교

식품/유아/완구>라면/밥/찌개>라면, 요약정보 : 봉지라면 / 일반라면 / 실온보관 / [영양 정보] / 표시기준량: 120g / 열량: 500kcal.

★★★★★ 평점: 4.6 · 리뷰 266개 · 최저가: ₩510



CSS Selector

- 선택을 좀 더 간결하고 직관적으로 할 수 있음.

Selector passed to the <code>select()</code> method	Will match . . .
<code>soup.select('div')</code>	All elements named <code><div></code>
<code>soup.select('#author')</code>	The element with an <code>id</code> attribute of <code>author</code>
<code>soup.select('.notice')</code>	All elements that use a CSS class attribute named <code>notice</code>
<code>soup.select('div span')</code>	All elements named <code></code> that are within an element named <code><div></code>
<code>soup.select('div > span')</code>	All elements named <code></code> that are <i>directly</i> within an element named <code><div></code> , with no other element in between
<code>soup.select('input[name]')</code>	All elements named <code><input></code> that have a <code>name</code> attribute with any value
<code>soup.select('input[type="button"]')</code>	All elements named <code><input></code> that have an attribute named <code>type</code> with value <code>button</code>

```

<style>...</style>
<div id="taw">...</div>
<div class="eqAnXb" id="res" role="main">
  <div id="topstuff"></div>
  <div id="search">
    <div data-hveid="CAIQNg" data-ved="2ahUKEwjN5tztz_N_3AhXKCaYKHw5rBpMQGnoECAIQNg">
      <h1 class="Uo8X3b OhScic zsYMMe">검색결과</h1>
      <div class="v7W49e" eid="bySAYqfzIcqTmAXu1pmYCQ" data-async-context="query:%EC%8B%A0%EB%9D%
      <div class="g tF2Cxc" style="width:652px" data-hveid="CBcQAA" data-ved="2ahUKEwjN5tztz_N_3.
      AA">
        <div class="kwxLod" data-sokoban-container="ih6Jnb_LGyvfvf"> flex
        <div class="jtfYYd" style="flex-grow:1"> flex
        <div class="NJo7tc Z26q7c jGGQ5e" data-header-feature="0">
          <div class="yuRUBf">
            <a href="https://namu.wiki/w/%EC%8B%A0%EB%9D%BC%EB%A9%B4" data-ved="2ahUKEwjN5tztz_N_3AhXKCaYKHw5rBpMQGnoECAIQNg">
              AsQAQ" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%
              5B4&ved=2ahUKEwjN5tztz_N_3AhXKCaYKHw5rBpMQGnoECAIQNg">
                <h3 class="LC201b MBeu0 DKV0Md">신라면 - 나무위키:대문</h3> == $0
            <div class="TbwUpd NJjxre">...</div>
          </a>
        <div class="B6fmyf">

```

```

elems = soup.select('#search a h3[class="LC201b MBeu0 DKV0Md"]')
for e in elems:
    print(e.get_text(), " : ", e.parent.get('href'))

```

CSV 출력

```
import csv

elms = soup.select('#search a h3[class="LC201b MBeu0 DKV0Md"]')
with open('result.csv', 'w') as wf:
    for e in elms:
        csv.writer(wf).writerow([e.get_text(), e.parent['href']])
```

webbrowser

```
import webbrowser  
webbrowser.open('https://www.google.com/search?q=신라면')
```