제 14 강 Web Scraping (1)

학습 목차

- 웹 스크래핑 절차
- 구글 검색

웹 스크래핑(Web Scraping)

- 웹 싸이트로부터 필요로 하는 데이터를 추출하는 것.

Web scraping 절차

- 웹 싸이트 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

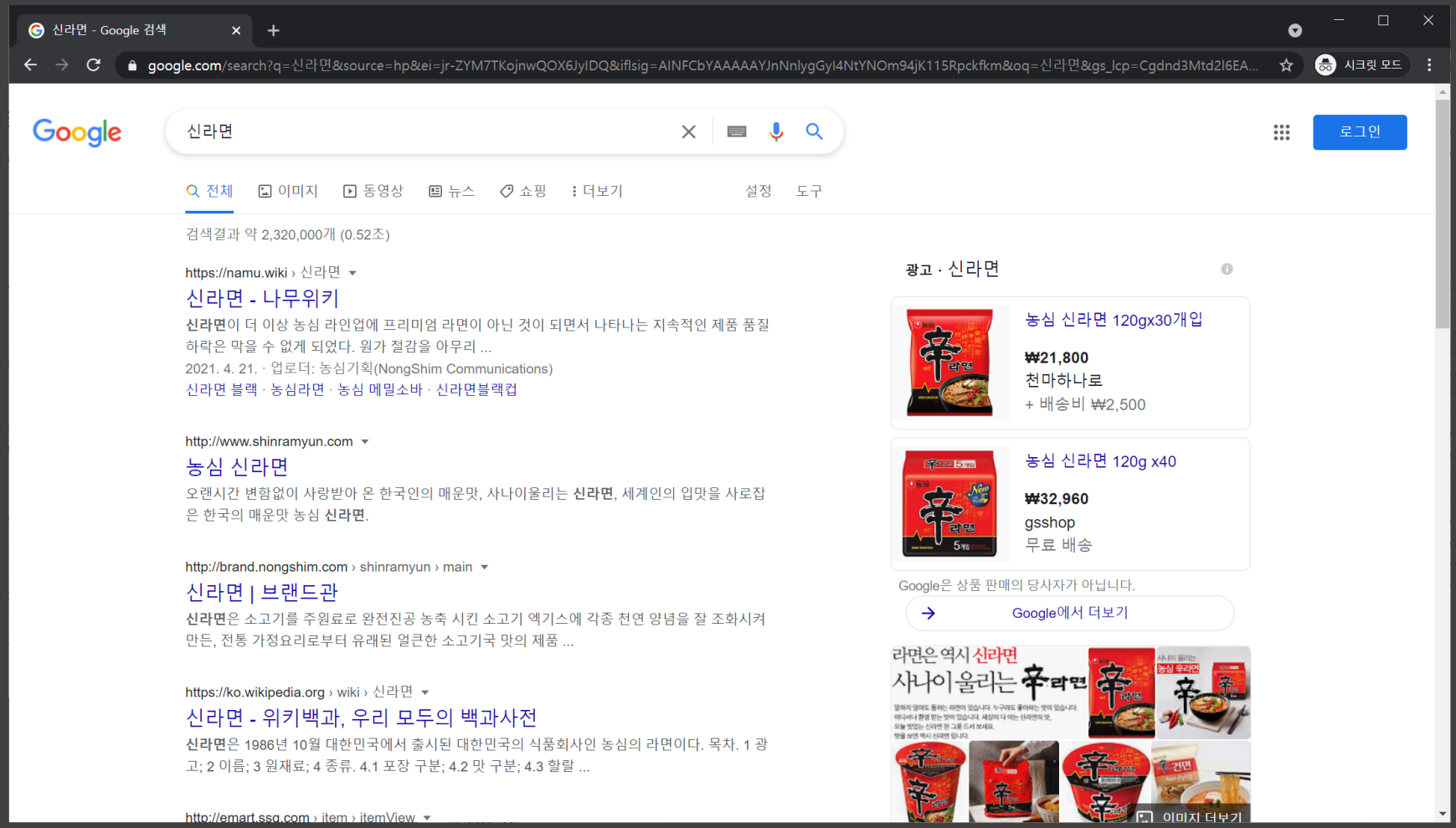
- 간단한 URL 만으로 원하는 데이터에 접근할 수 있는가?
  - BeautifulSoup, Scrapy 등의 검색 추출 도구 활용

- 사용자의 인터랙션이 필요한가?
  - 인터랙션 자동화 도구 활용(MechanicalSoup, Selenium, Playwright 등)

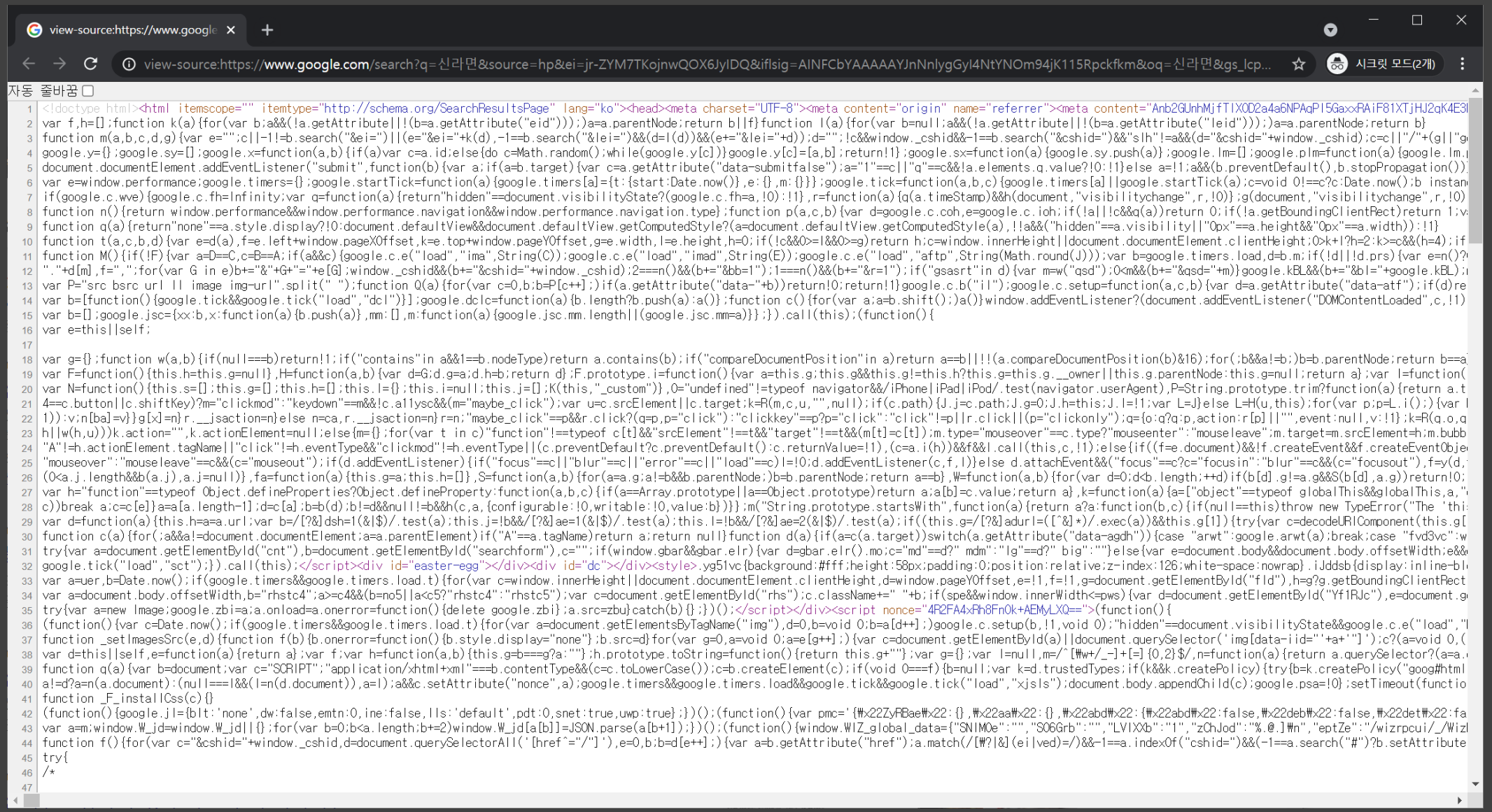
Note: 실제로는 복합적인 접근이 필요한 경우가 대다수

과업 #1

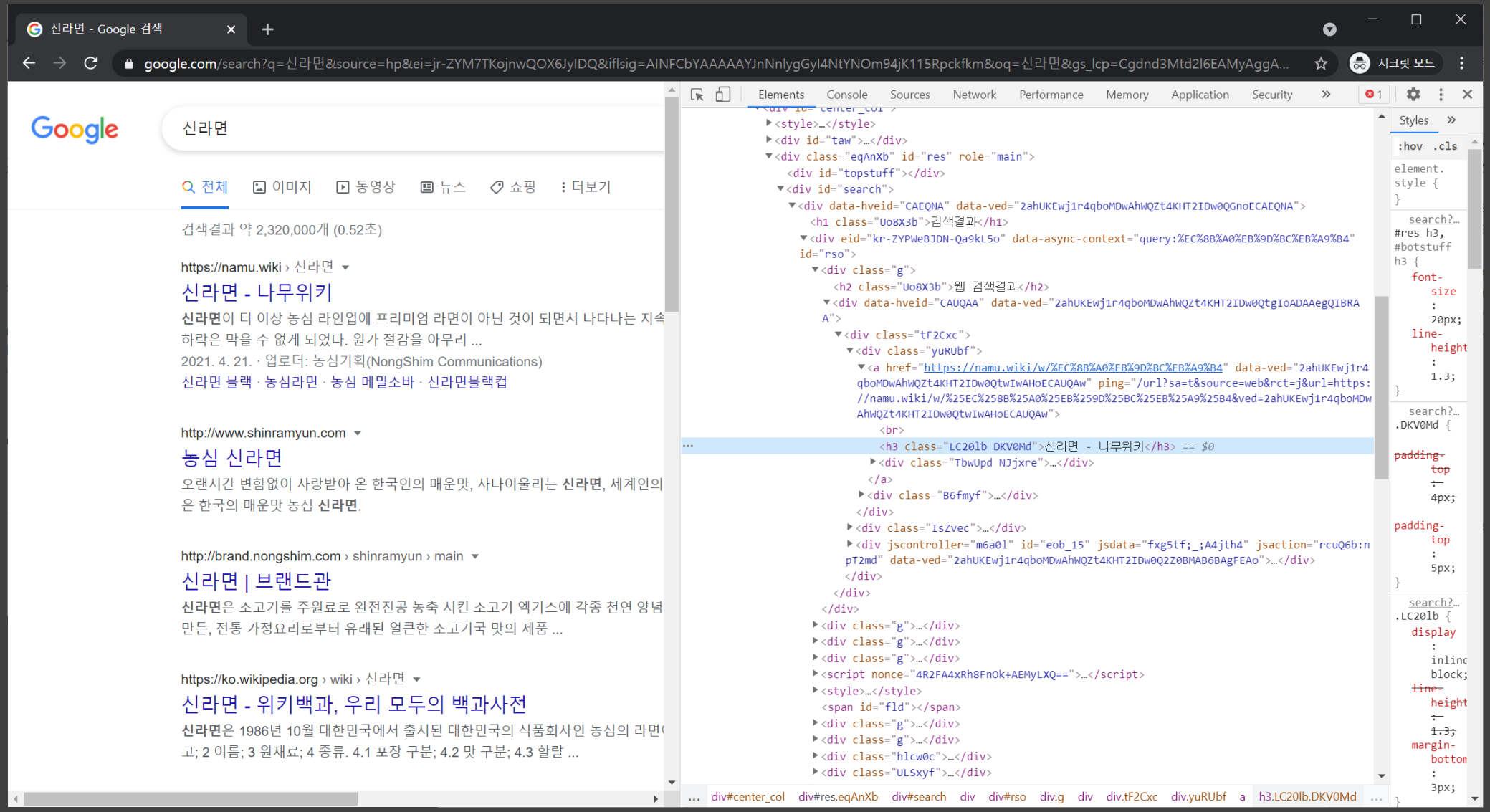
- 구글 검색 결과를 “제목”, “URL 주소” 형태의 CSV 파일로 만들기.



페이지 소스 보기 – Ctrl + U (크롬)



검사(Inspect) – F12 또는 Ctrl+Shift+I

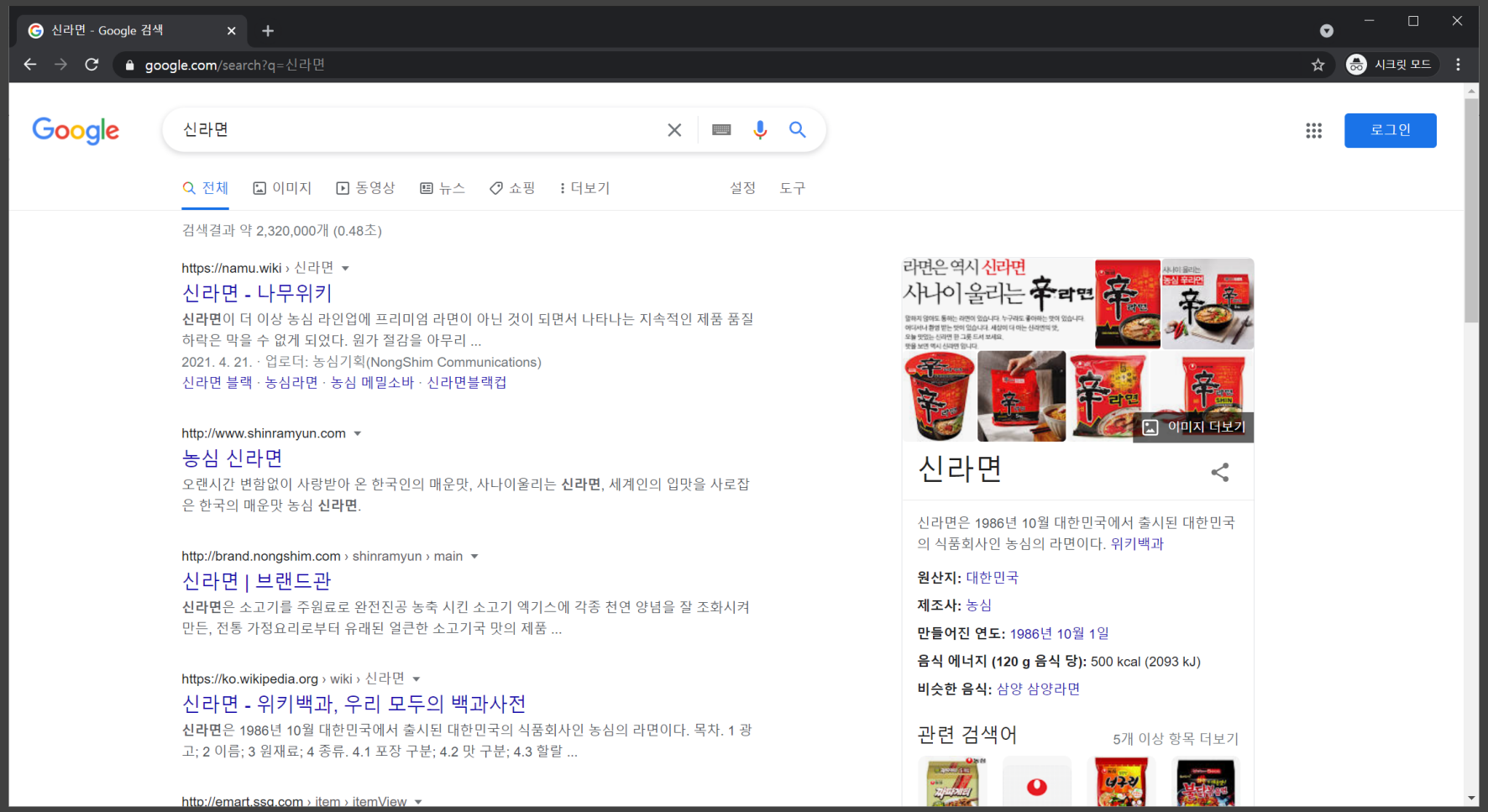


HTML 문서 구조 – 엘리먼트 트리 구조

```
<!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



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

신라면 - Google 검색 | python - Google 검색 | The Dormouse's story

google.com/search?q=신라면&source=hp&ei=1XBdZJ_YKdHAhwOAvIGoAw&iflsig=AOEireoAAAAAZF1-5UrMM_tX-G2WrrYqY-X565sLmXEg&ved=0ahUKEwjfl...

Google 신라면

전체 이미지 쇼핑 동영상 뉴스 더보기

검색결과 약 3,230,000개 (0.25초)

h3.LC20lb.MBeuO.DKV0Md 157.78 × 31

https://namu.wiki › 신라면

신라면 - 나무위키

2023. 4. 25. — **신라면**은 '매울 신(辛)'자의 '맵다'는 뜻과 농심 신춘호 회장의 성을 동시에 의미하는 중의적인 글자이다.[1] 농심 메밀소바의 전신인 찍어먹는 춘면의 ' ...

신라면 관련 이미지

컵 라면 | 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="2ahUKEwiir4jOsO7-AhUDQ94KHdD0AtEQFSgAegQIFBAA">
▼<div class="GLI8Bc UK95Uc" data-snc="ih6Jnb_LGyvvf"> grid
 ▼<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="2ahUKEwiir4jOsO7-AhUDQ94KHdD0AtEQFnoECAsQAQ" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%25EB%259D%25BC%25EB%25A9%25B4&ved=2ahUKEwiir4jOsO7-AhUDQ94KHdD0AtEQFnoECAsQAQ">
     <br>
     <h3 class="LC20lb MBeuO DKV0Md">신라면 - 나무위키</h3> == $0
    ▶<div class="TbwUpd NJjxre iUh30 apx8Vc ojE3Fb">…</div> flex
   </a>
  ▶<div class="B6fmyf byrV5b Mg1HEd">…</div> flex
  </div>
 </div>
 ▶<div class="Z26q7c UK95Uc Sth6v" data-sncf="0,1,2" data-snf="Vjbam" style="padding-left:20px;grid-area:Vjbam;width:92px">…</div>
 ▶<div class="Z26q7c UK95Uc VGXe8" data-sncf="2" data-snf="nke7rc" style="grid-area:nke7rc">…</div>
 ▶<div class="Z26q7c UK95Uc VGXe8" data-snf="bvRF1f" style="grid-area:bvRF1f">…</div>
 </div>
</div>
<span id="z9PoV"></span>
▶<script nonce>…</script>
</div>
</div>
▶<div class="ULSxyf">…</div>
```

Styles

```
:hov .cls +
element.style {
}
search?q=%E…=gw… .DKV0Md {
  margin-top: 18px;
}
search?q=%E…=gw… .DKV0Md {
  padding-top: 4px;
  padding-top: 5px;
}
search?q=%E…=gw… .LC20lb {
  display: inline-block;
  line-height: 1.3;
  margin-bottom: 3px;
}
search?q=%E…=gw… .MBeuO {
  line-height: 24px;
}
search?q=%E…=gw… .MBeuO {
  font-family: Apple SD Gothic Neo,arial,sans serif;
  font-size:
```

div.GLI8Bc.UK95Uc div.Z26q7c.UK95Uc.jGGQ5e.VGXe8 div.yuRUbf a h3.LC20lb.MBeuO.DKV0Md

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 엑세스하는 사용자의 정보
- 웹 서버는 이를 근거로 user agent 마다 다른 내용의 응답을 함.

HTTP에서의 사용 [편집]

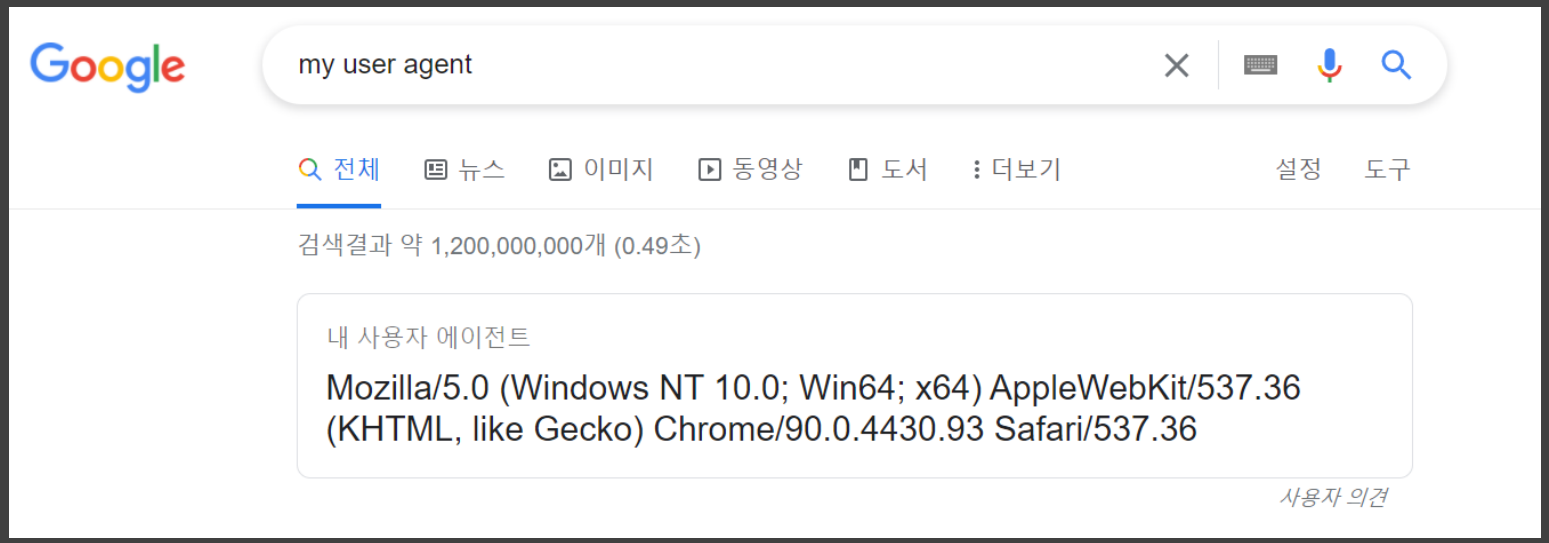
인간이 조작하는 웹 브라우저 형식 [편집]

```
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]

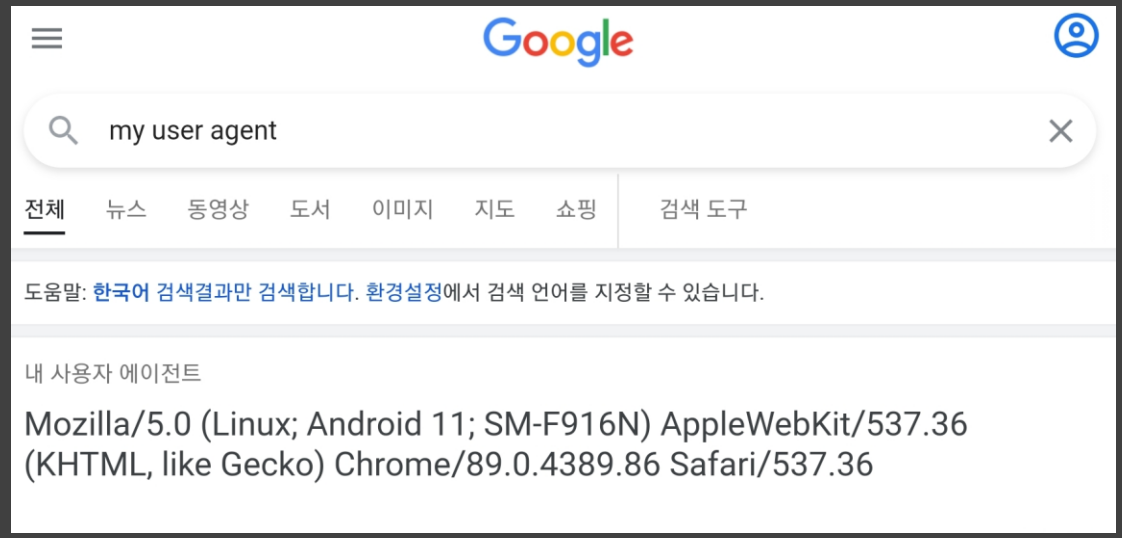
```
Googlebot/2.1 (+http://www.google.com/bot.html)
```

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


```
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 맛있고 뿌셔 먹어도 맛 좋고 아주 자주 잘 먹는 농심 **신라면!!!**

2021. 8. 23. · ★★★★½ 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 / [영양정보] / 표시기준량: 120g / 열량: 500kcal.

★★★★½ 평점: 4.6 · 리뷰 266개 · 최저가: ₩510

CSS Selector

- 선택을 좀 더 간결하고 직관적으로 할 수 있음.

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


```
elms = soup.select('#search a h3[class="LC20lb MBeuO DKV0Md"]')
for e in elms:
    print(e.get_text(), " : ", e.parent.get('href'))
```

CSV 출력

```
import csv

elms = soup.select('#search a h3[class="LC20lb MBeuO 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')
```