

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.

신라면 - Google 검색

google.com/search?q=신라면&source=hp&ei=jr-ZYM7TKojnwQOX6JyIDQ&ifsig=AINFCbYAAAAAYJnNlygGyl4NtYNOM94jK115Rpckfkm&oq=신라면&gs_lcp=Cgdn3Mtd2I6EA...

신라면

전체 이미지 동영상 뉴스 쇼핑 더보기 설정 도구

검색결과 약 2,320,000개 (0.52초)

<https://namu.wiki> > 신라면 ▾
신라면 - 나무위키
신라면이 더 이상 농심 라인업에 프리미엄 라면이 아닌 것이 되면서 나타나는 지속적인 제품 품질 하락은 막을 수 없게 되었다. 원가 절감을 아무리 ...
2021. 4. 21. · 업로더: 농심기획(NongShim Communications)
신라면 블랙 · 농심라면 · 농심 메일소바 · 신라면블랙컵


<http://www.shinramyun.com> ▾
농심 신라면
오랜시간 변함없이 사랑받아 온 한국인의 매운맛, 사나이올리는 신라면, 세계인의 입맛을 사로잡은 한국의 매운맛 농심 신라면.


<http://brand.nongshim.com> > shinramyun > main ▾
신라면 | 브랜드관
신라면은 소고기를 주원료로 완전친공 농축 시킨 소고기 엑기스에 각종 천연 양념을 잘 조화시켜 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 라면이다. 목차. 1 광고; 2 이름; 3 원재료; 4 종류. 4.1 포장 구분; 4.2 맛 구분; 4.3 할랄 ...

<http://emart.ssg.com> > item > itemView ▾


광고 · 신라면

 **농심 신라면 120gx30개입**
₩21,800
천마하나로
+ 배송비 ₩2,500

 **농심 신라면 120g x40**
₩32,960
gsshop
무료 배송

Google은 상품 판매의 당사자가 아닙니다.
→ Google에서 더보기

라면은 역시 신라면
사나이올리는 신라면

 이미지 더보기

페이지 소스 보기 - Ctrl + U (크롬)

```
view-source:https://www.google.com/search?q=신라면&source=hp&ei=jr-ZYM7TKojnwQOX6JyIDQ&isfsl=AINFcBYAAAAAYJnNlyGyL4NtYNOM94jK115Rpkfkm&oq=신라면&gs_lcp...
자동 줄바꿈
1 <!doctype html><html itemscope="" itemtype="http://schema.org/SearchResultsPage" lang="ko"><head><meta charset="UTF-8"><meta content="origin" name="referrer"><meta content="Anb2GJnhMjFT1X0D2a4a6NPAqP15GaxRAiF81XTjHJ2qK4E3I
2 var f,h=[];function k(a){for(var b;a&&(!a.getAttribute()||(b=a.getAttribute("eid"))):a=a.parentNode;return b}function l(a){for(var b=null;a&&(!a.getAttribute()||(b=a.getAttribute("leid"))):a=a.parentNode;return b}
3 function m(a,b,c,d,g){var e="";c||-1!=b.search("&ei=")||e=b.search("&ei=")+k(d),-1==b.search("&ei=")&&(d=l(d))&&(e+=k(d));d="";!c&&window._cshid&&-1==b.search("&cshid=")&&"sh"!a&&(d="&cshid="+window._cshid);c=c||"/"+(g||"g
4 google.y={};google.sy={};google.x=function(a,b){if(a)var c=a;else do c=Math.random();while(google.y[c]&&google.y[c]=[a,b];return l}google.sx=function(a){google.sy.push(a)};google.lm=[];google.plm=function(a){google.lm.
5 document.documentElement.addEventListener("submit",function(b){var a;if(a=b.target){var c=a.getAttribute("data-submitfalse");a="1"==c||"a"==c&&!a.elements.q.value?!0:1}else a=1;a&&(b.preventDefault(),b.stopPropagation())}
6 var e=window.performance;google.timers={};google.startTick=function(a){google.timers[a]={start:Date.now(),e:{},m:{}}};google.tick=function(a,b,c){google.timers[a]||google.startTick(a);c=void 0!==c?c:Date.now();b instan
7 if(google.c.wve){google.c.fh=Infinity;var q=function(a){return"hidden"==document.visibilityState?(google.c.fh=a,!0):!1},r=function(a){q(a.timeStamp)&&(document,"visibilitychange",r,!0)};g(document,"visibilitychange",r,!0)
8 function n(){return window.performance&&window.performance.navigation&&window.performance.navigation.type};function p(a,c,b){var d=google.c.coh,e=google.c.ioh;if(!a||c&&q(a))return 0;if(!a.getBoundingClientRect)return 1:v
9 function q(a){return"none"==a.style.display?!0:document.defaultView&&document.defaultView.getComputedStyle(a,document.defaultView.getComputedStyle(a),!a&&"hidden"==a.visibility||"Opx"==a.height&&"Opx"==a.width):!1}
10 function t(a,c,b,d){var e=d(a),f=e.left+window.pageXOffset,k=e.top+window.pageYOffset,g=e.width,l=e.height,h=0;if(!c&&0==l&&0==g)return h;c=window.innerHeight||document.documentElement.clientHeight;0>k+1?h=2:k>=c&&(h=4);if
11 function M(){if(!F){var a=D=C,c=B=A;if(a&&c){google.c.e("load","ima",String(C));google.c.e("load","imad",String(E));google.c.e("load","aftp",String(Math.round(J)))};var b=google.timers.load,d=b.m;if(!d||d.prs){var e=n()?"
12 ".+d[m],f=",";for(var g in e)b+="&"+g+"="+e[g];window._cshid&&(b+="&cshid="+window._cshid);2==n()&&(b+="&bb=");1==n()&&(b+="&r=");if("gsasrt"in d){var m=w("qsd");0<m&&(b+="&qsd="+m)}google.kL&&(b+="&b="+google.kL);
13 var P="src bsrc url ll image img-url".split(" ");function Q(a){for(var c=0,b=P[c++];)if(a.getAttribute("data-"+b))return!0;return!1}google.c.b("il");google.c.setup=function(a,c,b){var d=a.getAttribute("data-atf");if(d)re
14 var b=[function(){google.tick&&google.tick("load","dcl")}]};google.dclc=function(a){b.length?b.push(a):a()};function c(){for(var a=a-b.shift():a())window.addEventListener?(document.addEventListener("DOMContentLoaded",c,!1)
15 var b=[];google.jsc={x:b,x:function(a){b.push(a)},mm:[],m:function(a){google.jsc.mm.length||(google.jsc.mm=a)}}}.call(this);(function(){
16 var e=this|self;
17
18 var g={};function w(a,b){if(null==b)return!1;if("contains"in a&&!b.nodeType)return a.contains(b);if("compareDocumentPosition"in a)return a==b||!(a.compareDocumentPosition(b)&16);for(b&&a!=b;)b=b.parentNode;return b==a
19 var F=function(){this.h=this.g=null},H=function(a,b){var d=G;d.g=a;d.h=b;return d};F.prototype.i=function(){var a=this.g;this.g&&this.g!=this.h?this.g=this.g.__owner||this.g.parentNode:this.g=null;return a};var l=function(
20 var N=function(){this.s=[];this.g=[];this.h=[];this.l=[];this.i=null;this.j=[];K(this,"_custom"),O="undefined"!=typeof navigator&&iPhone iPad iPod.test(navigator.userAgent),P=String.prototype.trim?function(a){return a.t
21 4==c.button||c.shiftKey)?m="clickmod":"keydown"==m&&!c.allisclick?(m="maybe_click":var u=c.srcElement||c.target;k=R(m,c,u,"",null);if(c.path;J.g=J.this;J.l=J;var L=J}else L=H(u,this);for(var p=L.i():)for(var l
22 1)):v:n[ba]=v}g[x]=n}r.__jsaction=n}else n=c,a,r.__jsaction=n}r.n="maybe_click"==p&&r.click?(q=p,p="click":"click"!p||r.click||(p="clickonly"):q={o:q;q.p.action:r[p]||"",event:null,v:!1};k=R(q,o,q
23 h||w(h,u)))k.action="";k.actionElement=null}else{m={};for(var t in c)"function"!=typeof c[t]&&"srcElement"!==t&&"target"!==t&&(m[t]=c[t]);m.type="mouseover"==c.type?"mouseenter":"mouseleave";m.target=m.srcElement+h;m.bubb
24 "A"!=h.actionElement.tagName||"click"!=h.eventType&&"clickmod"!=h.eventType||c.preventDefault?c.preventDefault():c.returnValue=!1),(c.a.i(h))&&f&&l.call(this,c,!1);else{if(f=e.document)&&f.createEvent&&f.createEventObjec
25 "mouseover":"mouseleave"==c&&(c="mouseout");if(d.addEventListener){if("focus"==c||"blur"==c||"error"==c||"load"==c)l=!0;d.addEventListener(c,f,l)}else d.attachEvent&&"focus"==c?c="focusin":"blur"==c&&(c="focusout"),f=y(d,
26 (0<a.j.length&&b(a.j),a.j=null)),fa=function(a){this.g=a;this.h=[]},S=function(a,b){for(a=a.g;a!=b&&b.parentNode;)b=b.parentNode;return a==b},W=function(a,b){for(var d=0;d<b.length;++d)if(b[d].g!=a.g&&S(b[d],a.g))return!0;
27 var h="function"==typeof Object.defineProperty?Object.defineProperty:function(a,b,c){if(a==Array.prototype||a==Object.prototype)return a;a[b]=c.value;return a},k=function(a){a=["object"==typeof globalThis&&globalThis,a,"
28 c))break;a=c[e];a=a.length-1;d=c[a];b=b(d);b!=d&&null!=b&&h(c,a,{configurable:!0,writable:!0,value:b})};m("String.prototype.startsWith",function(a){return a?a:function(b,c){if(null==this)throw new TypeError("The 'this'
29 var d=function(a){this.h=a.url;var b=?[&]dsh=1(&[&]).test(a);this.j=lb&&[?&]ae=1(&[&]).test(a);this.l=lb&&[?&]ae=2(&[&]).test(a);if((this.g=?[&]adur=1(&[&])*_.exec(a))&&this.g[1]){try{var c=decodeURIComponent(this.g[
30 function c(a){for(c=a&&a=document.documentElement;a.parentNode)return a;return null}function d(a){if(a=c(a).tagName){switch(a.getAttribute("data-agdh")){case "arwt":google.arwt(a);break;case "fvd3vc":w
31 try{var a=document.getElementById("cnt"),b=document.getElementById("searchform"),c="";if(window.gbar&&gbar.elr){var d=gbar.elr(.mo;c="md"=="d?"mdm":"lg"=="d?"big":"");else{var e=document.body&&document.body.offsetWidth:e&&
32 google.tick("load","set");}.call(this);</script><div id="easter-egg"></div><div id="dc"></div><style>.ygl5vc{background:#fff;height:58px;padding:0;position:relative;z-index:126;white-space:nowrap}.jdJdsb{display:inline-bl
33 var a=uer,b=Date.now();if(google.timers&&google.timers.load.t){for(var c=window.innerHeight||document.documentElement.clientHeight,d=window.pageYOffset,e=!1,f=!1,g=document.getElementById("fid"),h=g?g.getBoundingClientRect
34 var a=document.body.offsetWidth,b="rhstc4":a>c4&&(b=no5)||a<c5?"rhstc4":"rhstc5";var c=document.getElementById("rhs");c.className+=" "+b;if(spe&&window.innerWidth<=pws){var d=document.getElementById("Yf1RJc"),e=document.g
35 try{var a=new Image;google.zbi=a;a.onload=a.onerror=function(){delete google.zbi};a.src=zbu}catch(b){}.})();</script></div><script nonce="4R2FA4xPh8FnOk+AEMyLX0=">(function(){
36 function(){var c=Date.now();if(google.timers&&google.timers.load.t){for(var a=document.getElementById("img"),d=0,b=void 0;b=a[d++];)google.c.setup(b,!1,void 0);"hidden"==document.visibilityState&&google.c.e("load","l
37 function _setImagesSrc(e,d){function f(b){b.onerror=function(){b.style.display="none"};b.src=d}for(var g=0,a=void 0;a=g[+];){var c=document.getElementById(a)||document.querySelector('img[data-id="'+a+'"]');c?(a=void 0,(
38 var d=this|self,e=function(a){return a};var f;var h=function(a,b){this.g=b==g?"":h.prototype.toString=function(){return this.g+""};var g={};var l=null,m="/[#+/_-/+]=/{0,2}$/",n=function(a){return a.querySelector?(a=a,
39 function q(a){var b=document;var c="SCRIPT";application/xhtml+xml"==b.contentType&&(c="toLowerCase()");c=b.createElement(c);this.g=0==f?var k=d.trustedTypes;if(k&&k.createPolicy){try{b=k.createPolicy("goog#html
40 al=d?a=n(a.document):(null==l&&(l=n(d.document)),a=l);a&&c.setAttribute("nonce",a);google.timers&&google.timers.load&&google.tick&&google.tick("load","xjsls");document.body.appendChild(c);google.psa=!0};setTimeout(funcio
41 function _F_installCss(c){
42 (function(){google.jl=[blt:'none',dw:false,emtn:0,ine:false,ils:'default',pdt:0,snet:true,uwp:true];})();(function(){var pmc='{#x22ZyFBAe#x22:{},{#x22aa#x22:{},{#x22abd#x22:{#x22abd#x22:false,#x22deb#x22:false,#x22det#x22:fa
43 var a=m;window.W_jd=window.W_jd||{};for(var b=0;b<a.length;b+=2)window.W_jd[a[b]]={JSON.parse(a[b+1])};})();(function(){window.W_lz_global_data="GNIMoe":"","SO6Grb":"","LV1X0b":"","1","zChJod":"","@.l","epTZe":"","wizrcui"/Wizl
44 function f(){for(var c="&cshid="+window._cshid,d=document.querySelectorAll(["href=""/"],e=0,b=d[e++]);)var a=b.getAttribute("href");a.match(/[?#&]([e]ved)=/)&&-1==a.indexOf("&cshid=")&&(-1==a.search("#")?b.setAttribute
45 try{
46 /*
47
```

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 and CSS styles for the search results.

Search Results:

- 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
- 농심 신라면
- 오랜시간 변함없이 사랑받아 온 한국인의 매운맛, 사나이울리는 신라면, 세계인의 은 한국의 매운맛 농심 신라면.
- 신라면 | 브랜드관
- 신라면은 소고기를 주원료로 완전진공 농축 시킨 소고기 엑기스에 각종 천연 양념 만든, 전통 가정요리로부터 유래된 얼큰한 소고기국 맛의 제품 ...
- 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):

- Elements:** Shows the HTML structure of the search results. The selected element is a search result link for '신라면 - 나무위키'.
- Styles:** Shows the CSS styles applied to the selected element. The styles include padding-top, padding-bottom, margin-bottom, and font-size.

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>
```

```
<!DOCTYPE html>
<meta charset="utf-8">
<html lang="en">
  <!-- Head 브라우저 머신한테만 의미있는 내용 -->
  <head>
    <title>Script Language</title>
  </head>
  <!-- body 실제 브라우저 창에 표시될 내용 -->
  <body>
    <!-- h1 heading 1 가장 큰 제목, 사이즈는 1~6 까지 있음 -->
    <h1>2024 Script Language</h1>
    <!-- Paragraph 문단 -->
    <p>파이썬에 대한 다양한 내용을 다룹니다.</p>
    <h2>주요 내용</h2>
    <p>문법</p>
    <p>
      <!-- Anchor 외부 연결 링크 -->
      <a href="https://www.google.com">구글</a>
      접속이 필요할 수 있습니다.
    </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 describes Shin Ramyun as a premium ramen brand and mentions its popularity. The second result is from <http://www.shinramyun.com> titled '농심 신라면', which describes it as a Korean instant noodle brand. The third result is from <http://brand.nongshim.com> titled '신라면 | 브랜드관', which provides information about the brand's history and products. The fourth result is from <https://ko.wikipedia.org> titled '신라면 - 위키백과, 우리 모두의 백과사전', which provides a detailed overview of the brand.

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 various details: '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 for '관련 검색어' (Related searches) with 5개 이상 항목 더보기 (More than 5 items to view).

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
with open('alice.html') as f:
    soup = BeautifulSoup(f, 'lxml') # lxml 모듈 설치 필요.

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

soup.body
soup.body.string
soup.body.get_text()

soup.p
soup.p['class']

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

# string 은 엘리먼트 직속인 문자열만 표시
# get_text() 는 엘리먼트 아래 모든 문자열 합해서 표시

# soup.body.p 를 간소화, 여러개 있을 경우 가장 먼저 나오는 것.

# soup.select('a')
# soup.select('#link3')
# soup.select('.sister')
# 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

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

>>

8

Settings

More

Close

```
lang="ko" style="width:652px" jsaction="QyLbLe:OMITjf;ewaord:qsYrDe;xd28Mb:A6j43c" data-hveid="CBQQAA" data-ved="2ahUKEwiir4jOs07-AhUDQ94KHdD0AtEQF5gAegQIFBAA">
  <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> == $0
        <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">
```

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...
.LC201b {
  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:
  20px;
```

div.GLI88cUK95Uc div.Z26q7cUK95UcjGGQ5eVGXe8 div.yuRUBf a h3.LC201b.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 액세스하는 사용자의 정보
- 웹 서버는 이를 근거로 user agent 마다 다른 내용의 응답을 함.

HTTP에서의 사용 [\[편집 \]](#)

인간이 조작하는 웹 브라우저 형식 [\[편집 \]](#)

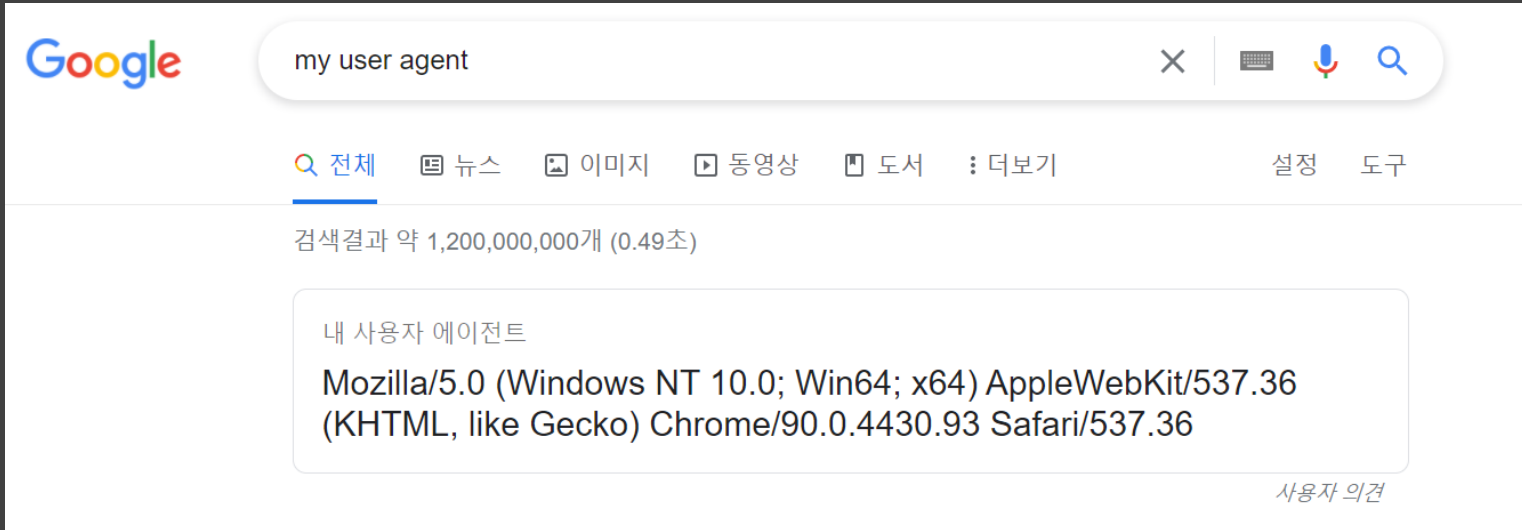
```
Mozilla/5.0 (iPad; U; CPU OS 3_2_1 like Mac OS X; en-us) AppleWebKit/531.21.10 (KHTML, like Gecko) Mobile/7B405
```

자동화된 에이전트(봇)의 형식 [\[편집 \]](#)

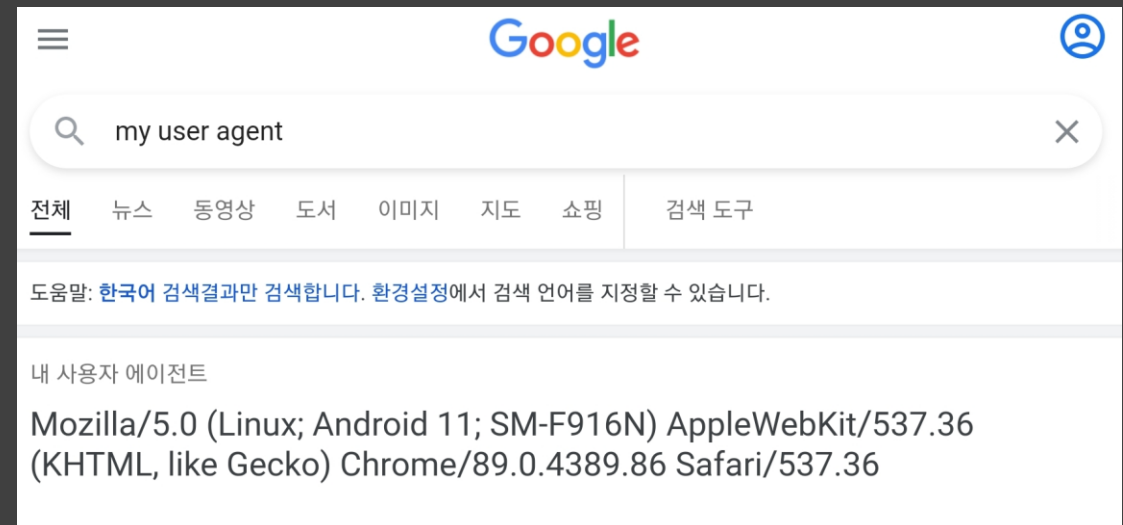
```
Googlebot/2.1 (+http://www.google.com/bot.html)
```


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="2ahUKEwi3t6OVwN73AA">
  ▼<div class="kwxlod" data-sokoban-container="ih6Jnb_LGyvfv"> 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="2ahUKEwi3t6OVwN73AA" data-cs="2" data-kind="parent">
            AoQAQ" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%25B4&ved=2ahUKEwi3t6OVwN73AAhVDA4KHTts8twQFnoECAoQAQ">
              <br>
              <h3 class="LC201b MBeu0 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="2ahUKEwjn5tzt_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="2ahUKEwjn5tzt_N_3
      AA">
        <div class="kwxLod" data-sokoban-container="ih6Jnb_LGyvfv"> 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="2ahUKEwjn5tzt
            AsQAQ" ping="/url?sa=t&source=web&rct=j&url=https://namu.wiki/w/%25EC%258B%25A0%
            5B4&ved=2ahUKEwjn5tzt_N_3AhXKCaYKHw5rBpMQGnoECAsQAQ">
              <h3 class="LCz01b MBeuO DKV0Md">신라면 - 나무위키:대문</h3> == $0
            <div class="TbwUpd NJjxre">...</div>
          </a>
        <div class="B6fmyf">

```

```

elms = soup.select('div[id="search"] a h3')
for i, e in enumerate(elms):
    print(f'{i} Title:{e.string}, URL:{e.parent["href"]}')

```

CSV 출력

```
import csv

elms = soup.select('#search a h3')
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')
```