

LG전자 H&A본부 - 한양대
DX Intensive Course

자연어처리 - 웹크롤링

김 종 우 교수



목차

- 웹크롤링과 HTML
- BeautifulSoup 사용하기
- BeautifulSoup을 이용한 웹크롤링 예제
- 셀레니움 사용하기
- 셀레니움을 이용한 웹크롤링 예제

- 웹크롤링과 HTML

목 차

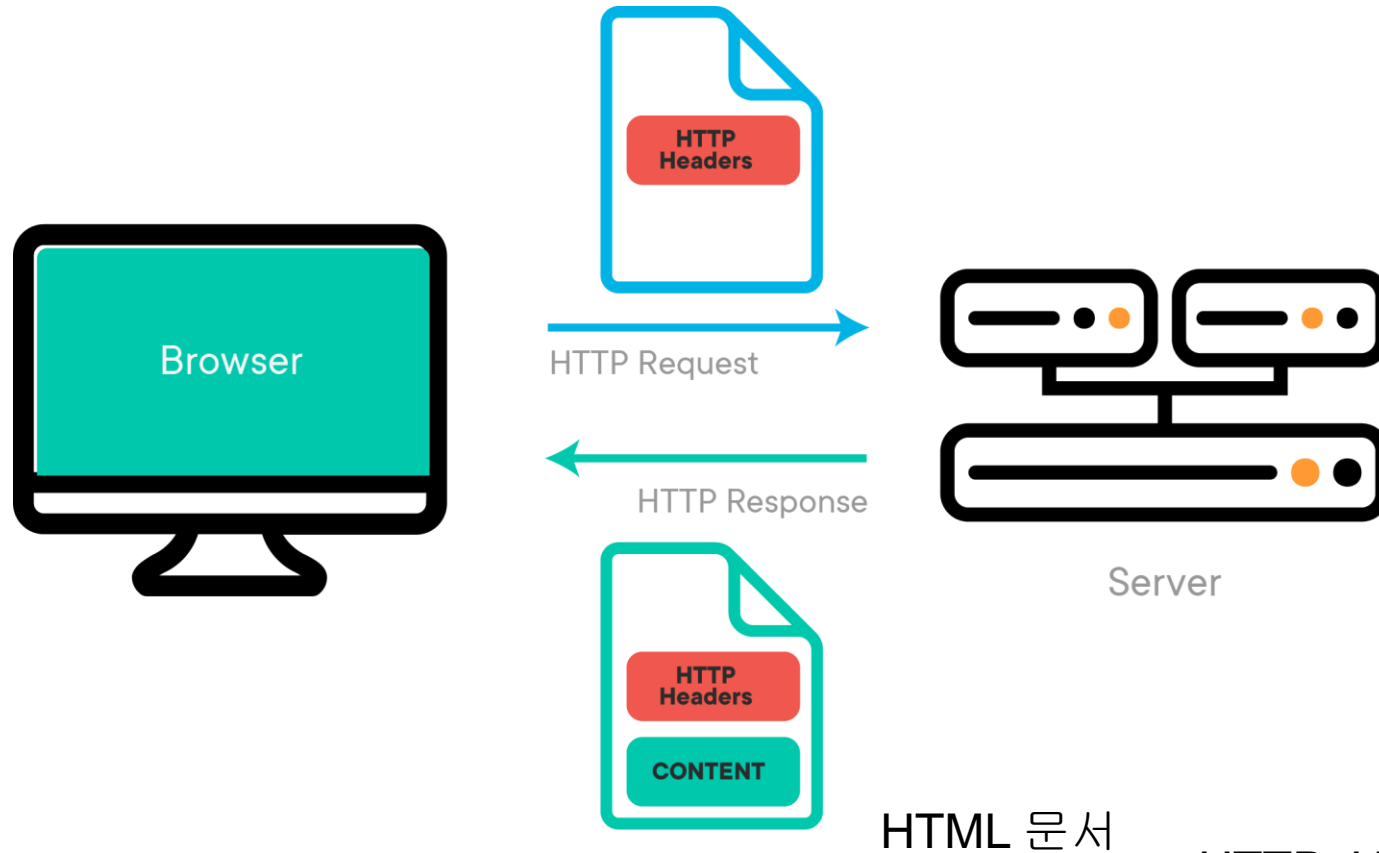
- 첫 번째 웹 스크레이퍼
- 웹 페이지 가져오기
- HTML

웹 크롤링

- 웹 스토레이핑



웹 페이지 가져오기



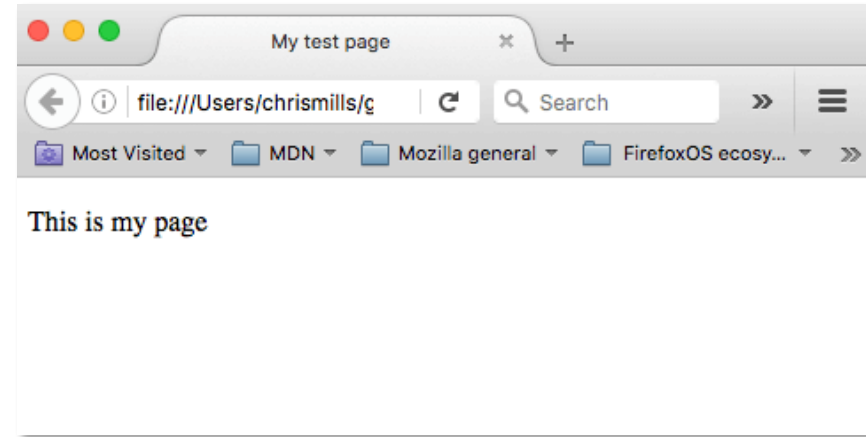
HTML 문서

HTTP: HyperText Transfer Protocol
HTML: HyperText Markup Language

HTML

- HTML 문서의 구조

```
<html>
  <head>
    <meta charset="utf-8">
    <title>My test page</title>
  </head>
  <body>
    <p>This is my page</p>
  </body>
</html>
```



웹브라우저에서 웹 페이지 가져오기

<https://pythonscraping.com/pages/page1.html>



파이썬에서 웹 페이지 가져오기

```
from urllib.request import urlopen  
html=urlopen("http://pythonscraping.com/pages/page1.html")  
print(html.read())
```

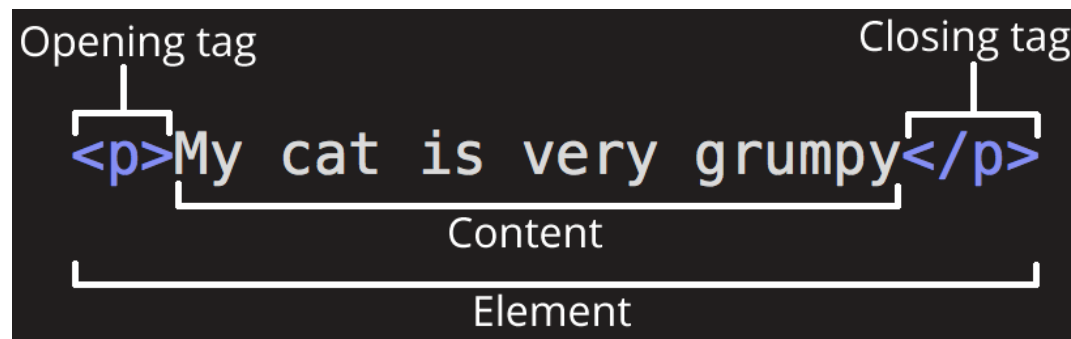
```
b'<html>\n<head>\n<title>A Useful Page</title>\n</head>\n<body>\n<h1>An  
Interesting Title</h1>\n<div>\nLorem ipsum dolor sit amet, consectetur adipisicing  
elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim  
ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea  
commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse  
cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non  
proident, sunt in culpa qui officia deserunt mollit anim id est  
laborum.\n</div>\n</body>\n</html>\n'
```



Lab 7-1

HTML

- HyperText Markup Language
- 웹페이지가 어떻게 구조화되어 있는 지 브라우저로 하여금 알 수 있도록 하는 마크업 언어
- Element들로 구성



HTML

- 중첩된 요소(Nesting elements)

<p>My cat is very grumpy.</p>

My cat is **very** grumpy.

HTML

- 블록 레벨 요소 vs 인라인 요소
 - Block versus inline element
 - 블록 레벨 요소(block-level element)
 - 웹 페이지 상에 블록을 만드는 요소
 - 블록 레벨 요소는 앞뒤 요소 사이에 새로운 줄을 만듦
 - 단락(paragraph), 목록(list), 네비게이션 메뉴(Navigation Menus) 등

HTML

- 블록 레벨 요소 vs 인라인 요소
 - 인라인 요소(**inline element**)
 - 항상 블록 레벨 요소 내에 포함
 - 문장, 단어 같은 작은 부분에만 적용
 - 새로운 줄(line)을 만들지 않음
 - <a>, , 등

HTML

- 블록 레벨 요소 vs 인라인 요소

`firstsecondthird`
`<p>fourth</p><p>fifth</p><p>sixth</p>`

Firstsecondthird

fourth

fifth

sixth

HTML

- 빈 요소(Empty elements)


- 모든 요소가 여는 태그, 내용, 닫는 태그 패턴을 따르지 않는
- 단일 태그(single tag)를 사용하는 요소도 있음

```

```

HTML

- 속성(Attribute)
 - 요소는 속성을 가질 수 있음
 - 속성은 요소에 실제로는 나타나고 싶지 않지만 추가적인 내용을 담고 싶을 때 사용



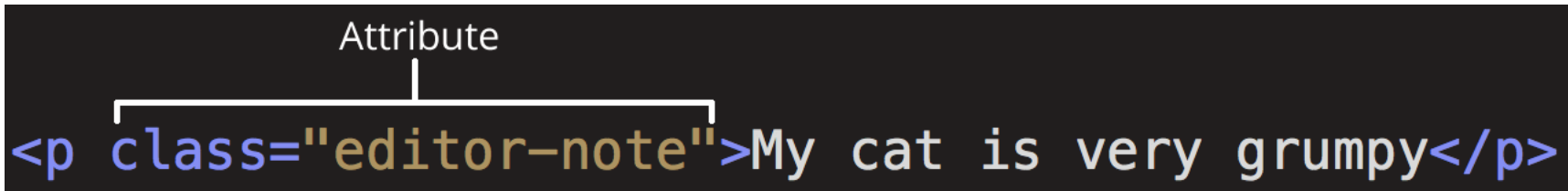
A diagram illustrating an HTML attribute. The word "Attribute" is written above a bracket that spans the `class="editor-note"` part of the HTML code snippet `<p class="editor-note">My cat is very grumpy</p>`. The code is displayed in a dark-themed editor with syntax highlighting: `<p` is blue, `class="editor-note"` is yellow, and `>My cat is very grumpy</p>` is white.

HTML

- 속성(Attribute)

- 속성을 사용할 때 지켜야 하는 규칙

- 요소 이름 다음에 오는 속성은 요소 이름과 속성 사이에 공백이 있어야 함
 - 하나 이상의 속성이 있는 경우 속성 사이에 공백이 있어야 함
 - 속성 이름 다음엔 등호(=)가 붙음
 - 속성 값은 따옴표로 감싸주어야 함



```
<p class="editor-note">My cat is very grumpy</p>
```

HTML

- 속성(Attribute) 예제

- <a> 요소
 - “anchor”
- href 속성
 - 연결하려는 웹 주소 지정
- title 속성
 - 링크에 대한 추가 정보
- target
 - 링크가 어떻게 열릴 지 지정
 - “_blank” 새 탭에서 보여줌

[A link to my favorite website.](#)

```
<p><a href="https://www.mozilla.org/" title="The Mozilla homepage"
target="_blank">A link to my favorite website.</a></p>
```

정리

- 웹페이지 가져오기
- HTML

BeautifulSoup 사용하기

목차

- BeautifulSoup 설치와 실행
 - BeautifulSoup 설치
 - BeautifulSoup 실행
- 신뢰할 수 있는 연결 만들기
- BeautifulSoup을 이용한 HTML 다루기
 - find()와 findAll()
 - 트리 이동
 - 정규 표현식
 - select()

Beautiful Soup 소개

160

THE LOBSTER

*“Beautiful Soup, so rich and green,
Waiting in a hot turcen!
Who for such dainties would not stoop?
Soup of the evening, beautiful Soup!
Soup of the evening, beautiful Soup!
 Beau—ootiful Soo—oop!
 Beau—ootiful Soo—oop!
Soo—oop of the e—e—evening,
 Beautiful, beautiful Soup!”*

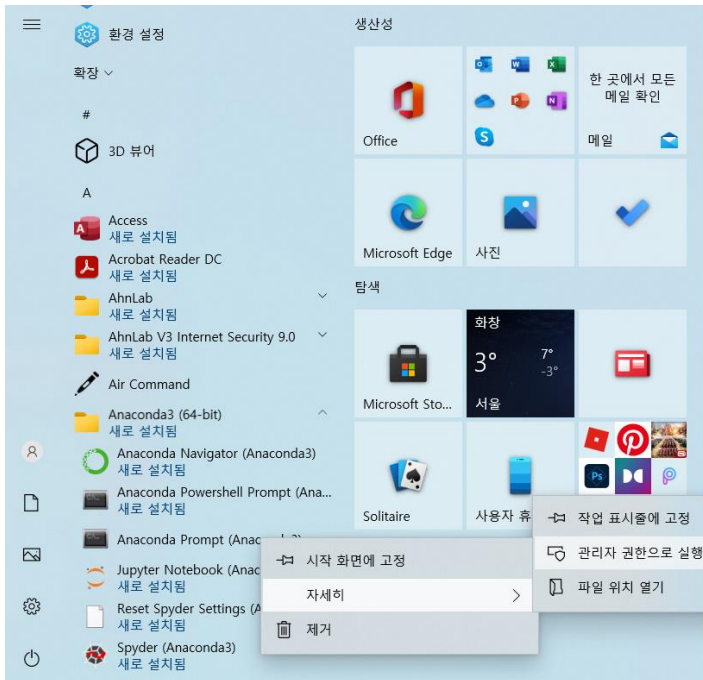
*“Beautiful Soup! Who cares for fish,
Game, or any other dish?
Who would not give all else for two p
ennyworth only of beautiful Soup?
Pennyworth only of beautiful Soup?
 Beau—ootiful Soo—oop!
 Beau—ootiful Soo—oop!
Soo—oop of the e—e—evening,
 Beautiful, beauti—FUL SOUP!”*

“Chorus again!” cried the Gryphon, and the Mock Turtle had just begun to repeat it, when



BeautifulSoup 설치

- Anaconda Prompt 실행
 - 관리자 권한으로 실행(오른쪽 마우스 클릭)
- `pip install beautifulsoup4`



```
관리자: Anaconda Prompt (Anaconda3)
(base) C:\Windows\system32>pip install beautifulsoup4
Requirement already satisfied: beautifulsoup4 in c:\program
9.3)
Requirement already satisfied: soupsieve>1.2; python_versio
a3\lib\site-packages (from beautifulsoup4) (2.0.1)
(base) C:\Windows\system32>
```

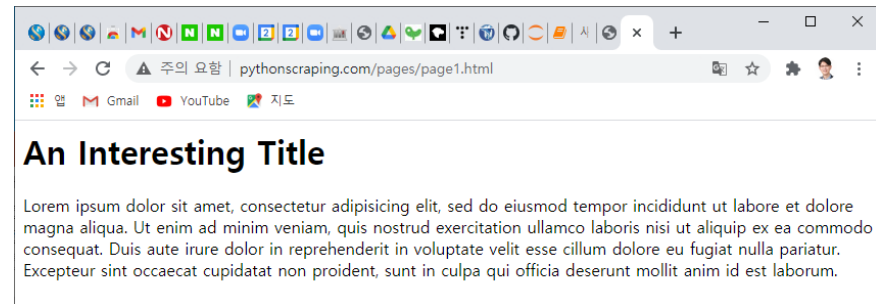
BeautifulSoup 실행

- BeautifulSoup 객체

```
from urllib.request import urlopen  
from bs4 import BeautifulSoup
```

```
html = urlopen('http://www.pythonscraping.com/pages/page1.html')  
bs = BeautifulSoup(html.read(), 'html.parser')  
print(bs.h1)
```

<h1>An Interesting Title</h1>



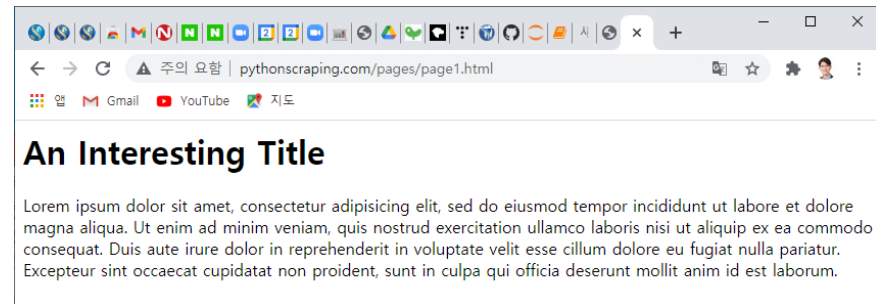
BeautifulSoup 실행

- BeautifulSoup 객체

```
print(bs.html.body.h1)
print(bs.body.h1)
print(bs.html.h1)
```

```
<html>
  <head>
    <title> ... </title>
  </head>
  <body>
    <h1> ... </h1>
    ...
  </body>
</html>
```

<h1>An Interesting Title</h1>



실패할 수 있는 연결

- 웹은 안정적이지 않음
- 가능한 오류들
 - 페이지를 찾을 수 없거나, URL 해석에서 에러가 생긴 경우
 - 서버를 찾을 수 없는 경우
 - 해당 태그가 존재하지 않는 경우
 - ..
- try, except 문 사용

실패할 수 있는 연결

- 서버가 존재하지 않는 경우

```
from urllib.request import urlopen
from urllib.error import HTTPError
from urllib.error import URLError
```

```
try:
    html = urlopen("https://pythonscrapingthisurldoesnotexist.com")
except HTTPError as e:
    print("The server returned an HTTP error")
except URLError as e:
    print("The server could not be found!")
else:
    print(html.read())
```

The server could not be found!

실패할 수 있는 연결

```
def getTitle(url):  
    try:  
        html = urlopen(url)  
    except HTTPError as e:  
        return None  
    except URLError as e:  
        return None  
    try:  
        bsObj = BeautifulSoup(html.read(), 'html.parser')  
        title = bsObj.body.h1  
    except AttributeError as e:  
        return None  
    return title
```

실패할 수 있는 연결

```
title = getTitle("http://www.pythonscraping.com/pages/page1.html")
```

```
if title == None:  
    print("Title could not be found")  
else:  
    print(title)
```

<h1>An Interesting Title</h1>



Lab 7-2

find()와 findAll()

- CSS

- Cascading Style Sheets
- HTML 요소들이 표시되는 방법을 기술하는 언어

- span과 div 태그

- 웹 페이지의 영역을 설정할 때 사용
- div 태그는 자동 줄 바꿈이 되지만, span 태그는 그렇지 않음

find()와 findAll()

- 예제

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen('http://www.pythonscraping.com/pages/warandpeace.html')
bs = BeautifulSoup(html, 'html.parser')
print(bs)
```

find()와 findAll()

- 예제

War and Peace

Chapter 1

"Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this means war, if you still try to defend the infamies and horrors perpetrated by that Antichrist- I really believe he is Antichrist- I will have nothing more to do with you and you are no longer my friend, no longer my 'faithful slave,' as you call yourself! But how do you do? I see I have frightened you- sit down and tell me all the news."

It was in July, 1805, and the speaker was the well-known Anna Pavlovna Scherer, maid of honor and favorite of the Empress Marya Fedorovna. With these words she greeted Prince Vasili Kuragin, a man of high rank and importance, who was the first to arrive at her reception. Anna Pavlovna had had a cough for some days. She was, as she said, suffering from la grippe; grippe being then a new word in St. Petersburg, used only by the elite.

All her invitations without exception, written in French, and delivered by a scarlet-liveried footman that morning, ran as follows:

"If you have nothing better to do, Count [or Prince], and if the prospect of spending an evening with a poor invalid is not too terrible, I shall be very charmed to see you tonight between 7 and 10- Annette Scherer."

"Heavens! what a virulent attack!" replied the prince, not in the least disconcerted by this reception. He had just entered wearing an embroidered court uniform, knee breeches, and

find()와 findAll()

- 예제

```
<html>
<head>
<style>
.green{
    color:#55ff55;
}
.red{
    color:#ff5555;
}
#text{
    width:50%;
}
</style>
</head>
<body>
<h1>War and Peace</h1>
<h2>Chapter 1</h2>
```

find()와 findAll()

- 예제

```
<body>
<h1>War and Peace</h1>
<h2>Chapter 1</h2>
<div id="text">
  "<span class="red">Well, Prince, so Genoa and Lucca are now just
  family estates of the
  Buonapartes. But I warn you, if you don't tell me that this means war,
  if you still try to defend the infamies and horrors perpetrated by
  that Antichrist- I really believe he is Antichrist- I will have
  nothing more to do with you and you are no longer my friend, no longer
  my 'faithful slave,' as you call yourself! But how do you do? I see
  I have frightened you- sit down and tell me all the news.</span>"
  <p></p>
  It was in July, 1805, and the speaker was the well-known <span
  class="green">Anna
  Pavlovna Scherer</span>, maid of honor and favorite of the <span
  class="green">Empress Marya
  Fedorovna</span>.
```

find()와 findAll()

- findAll(tag, attributes, recursive, text, limit, keywords)
- find(tag, attributes, recursive, text, keywords)

find()와 findAll()

- findAll(tagName, tagAttribute)
- 등장인물만 추출하기

```
nameList = bs.findAll('span', {'class': 'green'})  
for name in nameList:  
    print(name.get_text())
```

↑
태그를 제외하고 내용만 출력

Anna
Pavlovna Scherer
Empress Marya
Fedorovna
Prince Vasili Kuragin
Anna Pavlovna
St. Petersburg

find()와 findAll()

- 여러 태그 가져오기

```
headerList = bs.findAll({'h1', 'h2','h3','h4','h5','h6'})  
for header in headerList:  
    print(header.get_text())
```

War and Peace
Chapter 1

find()와 findAll()

- 여러 속성 가져오기

```
name_dialog_List = bs.findAll('span', {'class': {'green','red'}})
for name_dialog in name_dialog_List:
    print(name_dialog.get_text())
```

find()와 findAll()

- recursive
 - recursive = True
 - 자식, 자식의 자식을 검색
 - recursive = False
 - 문서의 최상위 태그만 찾을

find()와 findAll()

- text 매개 변수
 - 태그 속성이 아니라, 텍스트 콘텐츠에서 찾을
- 'the prince'가 몇 번 나왔는지 확인

```
princeList = bs.findAll(text='the prince')  
print(len(princeList))
```


트리 이동

- 예제

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
```

```
html = urlopen('http://www.pythonscraping.com/pages/page3.html')
bs = BeautifulSoup(html, 'html.parser')
```

```
for child in bs.find('table',{'id':'giftList'}).children:
    print(child)
```

자식만(vs. 자손(descendants))

제품 행 목록 출력

descendants를 사용했다면, 자식의 자식인 img, span, id 태그 포함

트리 이동

Totally Normal Gifts


Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our collection is hand-curated by well-paid, free-range Tibetan monks.

We haven't figured out how to make online shopping carts yet, but you can send us a check to:

123 Main St.

Abuja, Nigeria

We will then send your totally amazing gift, pronto! Please include an extra \$5.00 for gift wrapping.

Item Title	Description	Cost	Image
Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! <i>Now with super-colorful bell peppers!</i>	\$15.00	
Russian Nesting Dolls	Hand-painted by trained monkeys, these exquisite dolls are priceless! And by "priceless," we mean "extremely expensive"! <i>8 entire dolls per set! Octuple the presents!</i>	\$10,000.52	
Fish Painting	If something seems fishy about this painting, it's because it's a fish! <i>Also hand-painted by trained monkeys!</i>	\$10,005.00	
Dead Parrot	This is an ex-parrot! <i>Or maybe he's only resting?</i>	\$0.50	

트리 이동

- F12 - 소스보기



Totally Normal Gifts

Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our collection is hand-curated by well-paid, free-range Tibetan monks.

We haven't figured out how to make online shopping carts yet, but you can send us a check to:

123 Main St.

Abuja, Nigeria

We will then send your totally amazing gift, pronto! Please include an extra \$5.00 for gift wrapping.

Item Title	Description	Cost	Image
Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! Now with super-colorful bell peppers! Hand-painted by trained monkeys, these	\$15.00	

The screenshot shows the Chrome DevTools interface. The top panel displays the HTML source code, highlighting the `<body>` element. The `body` element contains a `<div id="wrapper">` which includes an `` and an `<h1>Totally Normal Gifts</h1>`. Below the `<div id="content">` is a paragraph of text. The bottom panel shows the 'Styles' tab, displaying the default user agent styles for the `body` element, such as `display: block;` and `margin: 8px;`. A visual box model diagram is overlaid on the page content, showing the margin, border, and padding of the selected element. The bottom right corner of the DevTools window shows a 'What's New' panel with updates from the Chrome 89 update.

트리 이동

- 실행 결과

```
<tr><th>  
Item Title  
</th><th>  
Description  
</th><th>  
Cost  
</th><th>  
Image  
</th></tr>
```

```
<tr class="gift" id="gift1"><td>  
Vegetable Basket  
</td><td>
```

```
This vegetable basket is the perfect gift for your health conscious (or overweight) friends!  
<span class="excitingNote">Now with super-colorful bell peppers!</span>
```

트리 이동

- 형제 다루기

```
for sibling in bs.find('table', {'id':'giftList'}).tr.next_siblings():  
    print(sibling)
```

표의 헤더가 빠짐!

```
<tr class="gift" id="gift1"><td>  
Vegetable Basket  
</td><td>  
This vegetable basket is the perfect gift for your health conscious (or overweight) friends!  
<span class="excitingNote">Now with super-colorful bell peppers!</span>  
</td><td>  
$15.00  
</td><td>  
  
</td></tr>
```

트리 이동

- 형제 다루기
 - next_siblings
 - previous_siblings
 - next_sibling
 - previous_sibling

트리 이동

- 부모 찾기
 - parent
 - parents

```
print(bs.find('img',  
             {'src':'../img/gifts/img1.jpg'})  
      .parent.previous_sibling.get_text())
```

\$15.00

트리 이동

Totally Normal Gifts


Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our collection is hand-curated by well-paid, free-range Tibetan monks.

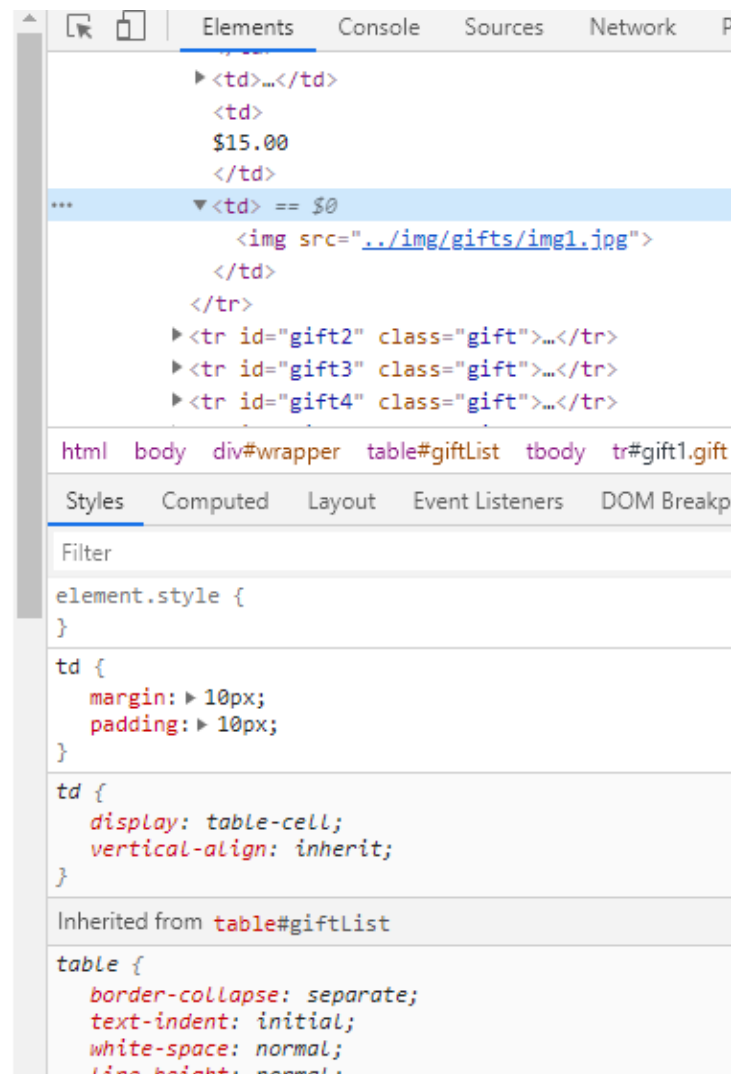
We haven't figured out how to make online shopping carts yet, but you can send us a check to:

123 Main St.

Abuja, Nigeria

We will then send your totally amazing gift, pronto! Please include an extra \$5.00 for gift wrapping.

Item Title	Description	Cost	Image
Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! <i>Now with</i>	\$15.00	



정규 표현식

- Regular Expression
- 문자열이 주어진 규칙에 일치하는지 여부를 판단
- 정규 문자열의 예
 - 글자 a를 최소한 한 번 쓰시오
 - 그 뒤에 b를 정확히 5개 쓰시오
 - 그 뒤에 c를 짝수 번 쓰시오
 - 마지막에 d가 있어도 되고 없어도 됨

aaaabbbbbccd
abbbbbccccc

aa*bbbb(cc)*(d|)

패턴	설명	예제
^	이 패턴으로 시작해야 함	^abc : abc로 시작해야 함 (abcd, abc12 등)
\$	이 패턴으로 종료되어야 함	xyz\$: xyz로 종료되어야 함 (123xyz, strxyz 등)
[문자들]	문자들 중에 하나이어야 함. 가능한 문자들의 집합을 정의함.	[Pp]ython : "Python" 혹은 "python"
[^문자들]	[문자들]의 반대로 피해야할 문자들의 집합을 정의함.	[^aeiou] : 소문자 모음이 아닌 문자들
	두 패턴 중 하나이어야 함 (OR 기능)	a b : a 또는 b 이어야 함
?	앞 패턴이 없거나 하나이어야 함 (Optional 패턴을 정의할 때 사용)	\d? : 숫자가 하나 있거나 없어야 함
+	앞 패턴이 하나 이상이어야 함	\d+ : 숫자가 하나 이상이어야 함
*	앞 패턴이 0개 이상이어야 함	\d* : 숫자가 없거나 하나 이상이어야 함
패턴{n}	앞 패턴이 n번 반복해서 나타나는 경우	\d{3} : 숫자가 3개 있어야 함
패턴{n, m}	앞 패턴이 최소 n번, 최대 m 번 반복해서 나타나는 경우 (n 또는 m 은 생략 가능)	\d{3,5} : 숫자가 3개, 4개 혹은 5개 있어야 함
\d	숫자 0 ~ 9	\d\d\d : 0 ~ 9 범위의 숫자가 3개를 의미 (123, 000 등)
\w	문자를 의미	\w\w\w : 문자가 3개를 의미 (xyz, ABC 등)
\s	화이트 스페이스를 의미하는데, [\t\n\r\f] 와 동일	\s\s : 화이트 스페이스 문자 2개 의미 (\r\n, \t\t 등)
.	뉴라인(\n) 을 제외한 모든 문자를 의미	.{3} : 문자 3개 (F15, 0x0 등)






정규 표현식과 BeautifulSoup

- 상품 이미지들만 가져오기

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
```

```
html = urlopen('http://www.pythonscraping.com/pages/page3.html')
bs = BeautifulSoup(html, 'html.parser')
images = bs.find_all('img')
for image in images:
    print(image['src'])
```

```
../img/gifts/logo.jpg
../img/gifts/img1.jpg
../img/gifts/img2.jpg
../img/gifts/img3.jpg
../img/gifts/img4.jpg
../img/gifts/img6.jpg
```

Totally Normal Gifts			
Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our c			
We haven't figured out how to make online shopping carts yet, but you can send us a check to: 123 Main St. Abuja, Nigeria We will then send your totally amazing gift, pronto! Please include an extra \$5.00 for gift wrapping.			
Item Title	Description	Cost	Image
Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! <i>Now with super-colorful bell peppers!</i>	\$15.00	
Russian Nesting Dolls	Hand-painted by trained monkeys, these exquisite dolls are priceless! And by "priceless," we mean "extremely expensive"! <i>8 entire dolls per set! Octuple the presents!</i>	\$10,000.52	
Fish Painting	If something seems fishy about this painting, it's because it's a fish! <i>Also hand-painted by trained monkeys!</i>	\$10,005.00	
Dead Parrot	This is an ex-parrot! <i>Or maybe he's only resting?</i>	\$0.50	
Mystery Box	If you love surprises, this mystery box is for you! Do not place on light-colored surfaces. May cause oil staining. <i>Keep your friends guessing!</i>	\$1.50	

정규 표현식과 BeautifulSoup

- 상품 이미지들만 가져오기

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import re
```

```
html = urlopen('http://www.pythonscraping.com/pages/page3.html')
bs = BeautifulSoup(html, 'html.parser')
images = bs.find_all('img', {'src':re.compile('\.\.VimgVgiftsVimg.*\.jpg')})
for image in images:
    print(image['src'])
```

[../img/gifts/img?????.jpg](#)

[../img/gifts/img1.jpg](#)
[../img/gifts/img2.jpg](#)
[../img/gifts/img3.jpg](#)
[../img/gifts/img4.jpg](#)
[../img/gifts/img6.jpg](#)



Lab 7-3

select()문

- select_one()
 - find()
- select()
 - find_all()

select()문

- 태그 검색

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen('http://www.pythonscraping.com/pages/page3.html')
bs = BeautifulSoup(html, 'html.parser')
```

```
title=bs.find('h1')
print(title.get_text())
```

```
title1=bs.select_one('h1')
print(title1.get_text())
```

```
title2=bs.select('h1')
print(title2[0].get_text())
```

select()문



Totally Normal Gifts

Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our collection is hand-curated by well-paid, free-range Tibetan monks.


We haven't figured out how to make online shopping carts yet, but you can send us a check to:

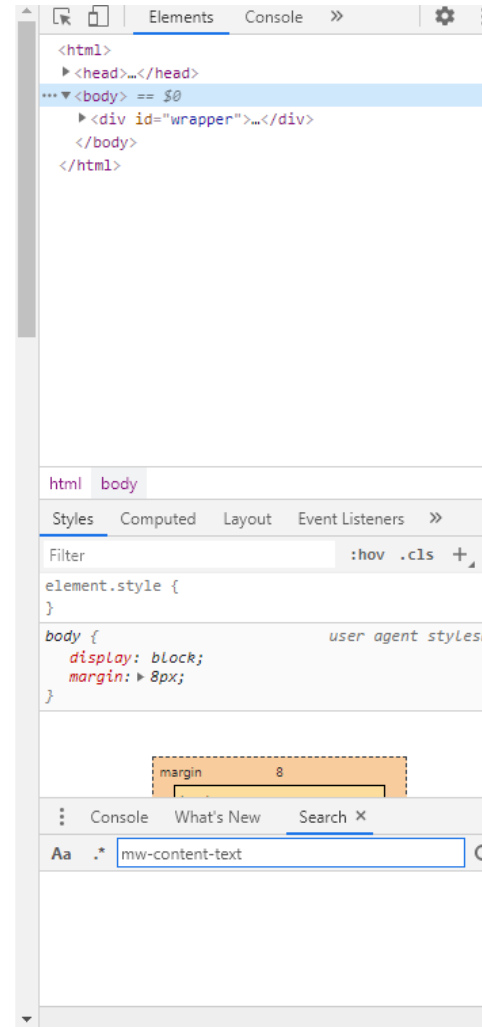
123 Main St.

Abuja, Nigeria

We will then send your totally amazing gift, pronto! Please include an extra \$5.00 for gift wrapping.

Item Title	Description	Cost	Image
------------	-------------	------	-------

Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! <i>Now with super-colorful</i>	\$15.00	
------------------	--	---------	---



select()문

- 하위 태그 검색
 - 띄어쓰기(직계 자식이 아니어도 됨)

```
table=bs.find('table').findAll('tr')
for tr in table:
    print("="*10)
    print(tr.get_text())
```

```
table1 = bs.select('table tr')
for tr in table1:
    print("="*10)
    print(tr.get_text())
```


select()문

- 하위 태그 검색
 - 띄어쓰기(직계 자식이 아니어도 됨)

```
table=bs.find('table').findAll('td')
for tr in table:
    print("="*10)
    print(tr.get_text())
```

```
table1 = bs.select('table td')
for tr in table1:
    print("="*10)
    print(tr.get_text())
```

select()문

- 직계 자식 태그 검색
 - > 사용

```
table1 = bs.select('table > tr')
for tr in table1:
    print("="*10)
    print(tr.get_text())
```

```
table2 = bs.select('table > td')
for tr in table2:
    print("="*10)
    print(tr.get_text())
```

select()문

- 태그의 class 탐색 시
 - . 사용

```
table=bs.find('table').findAll('tr',{'class':'gift'})
for tr in table:
    print("="*10)
    print(tr.get_text())
```

```
table1=bs.select('table tr.gift')
for tr in table1:
    print("="*10)
    print(tr.get_text())
```

select()문

- 태그의 id 탐색 시
 - # 사용

```
table=bs.findAll('table',{'id':'giftList'})  
for tr in table:  
    print("="*10)  
    print(tr.get_text())
```

```
table1=bs.select('table#giftList')  
for tr in table1:  
    print("="*10)  
    print(tr.get_text())
```

select()문

- 태그의 class, id 동시 탐색 시

```
table=bs.find('table').findAll('tr',{'id':'gift2','class':'gift'})
for tr in table:
    print("="*10)
    print(tr.get_text())
```

```
table1=bs.select('table tr.gift#gift2')
for tr in table1:
    print("="*10)
    print(tr.get_text())
```



Lab 7-4

BeautifulSoup을 이용한 웹크롤링 예제

목차

- 텍스트 정보 가져와서 저장하기
 - 상세 페이지 정보 가져와서 저장하기

상세 페이지 정보 가져와서 저장하기

• 네이버 쇼핑 주방가전 Best 100

BEST100 2021.09.18. SATURDAY

패션의류 패션잡화 화장품/미용 디지털/가전 가구/인테리어 식품 스포츠/레저 출산/육아 생활/건강

베스트100 홈 > 디지털/가전 > 주방가전 1/3

냉장고	밀치냉장고	냉동고	전용냉장고	가스레인지
가스레인지후드	전기밥솥	전자레인지	오븐	인덕션
하이라이트	하이브리드	한글라이트	식기세척/건조기	전기포트
결수기	커피머신	토스터기	친환경냉기	음식물처리기
커피메이커	커피자판기	계량기	거품/반죽기	

전기밥솥 인기검색어 인기브랜드 인기쇼핑몰 ✓ 가격비교 → 데이비데이 → 일반상품

① 최근 2달 7일 기준 네이버쇼핑을 통한 판매실적과 상품등록수를 반영하여 매일 업데이트 됩니다.

BEST 1

리하스 올스핀 KHD-16L

최저 129,000원

판매처 상품평 (1,204)

BEST 2

원머스트리빙 원머스트 쿡킹 A7 올스핀

최저 129,000원

판매처 상품평 (391)

BEST 3

원머스트리빙 원머스트 쿡킹 A9 올스핀

최저 199,000원

판매처 상품평 (2,586)

BEST 4

일리프원시스 Y3.3

최저 121,890원

판매처 상품평 (2,540)

BEST 5

삼성전자 비스포크 RF85A 910S AP (색상선택)

최저 1,370,000원

판매처 상품평 (718)

BEST 6

삼성전자 비스포크 R833A 3004AP (색상선택)

최저 1,212,800원

판매처 상품평 (1,207)

BEST 7

삼성전자 삼성 RS84T5081 5A

최저 1,212,800원

판매처 상품평 (1,621)

BEST 8

삼성전자 비스포크 RF85A 9001 AP (색상선택)

최저 1,704,040원

판매처 상품평 (320)

BEST 9

아이프릭 허가로 시그널

최저 169,000원

판매처 상품평 (12)

BEST 10

네스presso 버추오 플러스

최저 229,000원

판매처 상품평 (5,829)

BEST 11

바한 블루프리트

최저 69,000원

판매처 상품평 (1,334)

BEST 12

루루 CRP-LHTK0610PW 토윈프레스

최저 400,860원

판매처 상품평 (1,765)

BEST 13

LG전자 오보제압덕션 DUB J2EA

최저 1,320,700원

판매처 상품평 (395)

BEST 14

루루 CRP-DHP0610FD

최저 214,130원

판매처 상품평 (9,808)

상세 페이지 정보 가져와서 저장하기

- 제품명, 평점, 최저가, 배송정보

NAVER 네이버페이

쇼핑 | 쇼핑 NEW

검색

상세검색

카테고리 더보기

베스트 100 | 물 전체보기 | 쇼핑 MY

홈 | 백화점원도 | 아울렛원도 | 스타일원도 | 디자이너원도 | 뷰티원도 | 럭셔리 NEW | 리빙원도 | 푸드원도 | 장보기 NEW | 키즈원도 | 핏원도 | 플레이원도 | 아트원도 | MR. NEW | 핫딜 | 해외직구 | 기획전 | 트렌드상품

디지털/가전 > 주방가전 > 냉장고 > 양문형냉장고

LG전자 디오스 F873S11E

★★★★★ 4.8

브랜드 카탈로그 | 제조사 LG전자 | 브랜드 디오스 | 등록일 2020.07. | ❤️ 찜하기 3,300 | 📢 정보 수정요청


에너지효율: 1등급 | 용량: 870L | 품목: 4도어냉장고 | 냉장고, 냉동고set | 냉장실: 503L | 냉동실: 367L | 서랍: 신선야채실, 스마트디바이더 | 컴프레서: 리니어 | 할취: UV안심계균Plus
도어함들: 스웨이함들 | 스마트기능: 스마트진단 | 재질: 메탈 | 소비전력: 24.8kWh | 구조: 상냉장·하냉동 | 색상: 그레이

더보기 >

LG전자 브랜드스토어 [바로가기 >](#)

브랜드스토어에서 더 많은 브랜드 제품과 브랜드 행사 소식을 만나보세요.

1등급



이미지제공 | SSG닷컴

최저 1,596,400원

무로배송 | 옵션

판매가 **사라가기**

판매처별 최저가

인기순 | **최저가순** | 배송비포함 | 카드할인

기본구성

구매처

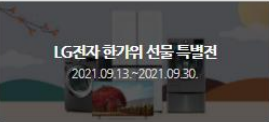
판매가

배송비

사라가기

A.육선	최저 1,596,400	무로배송	사라가기
타임팩스 TMON	1,596,500	무로배송	사라가기
오늘의집 <small>NEW</small>	1,596,600	무로배송	사라가기
11번가	1,597,890	무로배송	사라가기
INTERPARK <small>NEW</small>	1,642,550	무로배송	사라가기
치얼얼코리아주식회사 <small>NEW</small>	1,658,900	30,000원	사라가기
Gmarket	1,675,270	무로배송	사라가기


이벤트



LG전자 한가위 선물 특별전
2021.09.13~2021.09.30.

블로그리뷰

더보기 >



LG전자 디오스 F873S11E 상냉장 하냉동
냉장고 추천

블로그 | 경제적 행위를 통해 경제적 자유를 꿈꾸는 'free...

상세 페이지 정보 가져와서 저장하기

- 필요한 패키지 가져오기와 url 열기

```
from bs4 import BeautifulSoup
from urllib.request import urlopen
import requests
import csv
import os
import time
import random
```

```
save_file_path = 'D:\kjh\lecture\LG전자_21\week7_자연어처리'
#save_file_path = '크롤링 파일 저장 위치'
```

```
url = 'https://search.shopping.naver.com/best100v2/detail.nhn?catId=500000000'
result = requests.get(url)
```

상세 페이지 정보 가져와서 저장하기

- 제품명 가져오기

```
# 제품명 가져오기  
a = result.text  
soup = BeautifulSoup(a, 'html.parser')
```

```
name = soup.select('p.cont > a')
```

```
for i in name :  
    print(i['title'])
```

```
# Top-5  
for i in range(5) :  
    print(name[i]['title'])
```

상세 페이지 정보 가져와서 저장하기

- csv 파일에 저장하기

```
os.chdir(save_file_path)
save_file_name = 'naver_best_title.csv'
columns_list=[ ' 제품명']
f=open(save_file_name, 'w+', newline='')
cw = csv.writer(f)
cw.writerow(columns_list)

for i in name :
    temp_text=[]
    temp_text.append(i['title'])
    cw.writerow(temp_text)

f.close()
print('\n!!!! Best Product Title Web Crawling Complete !!!!\n')
```

상세 페이지 정보 가져와서 저장하기

P13											
	A	B	C	D	E	F	G	H	I	J	K
1	제품명										
2	리하스 올스텐 KHD-16L										
3	원더스리빙 원더스 쿡킹 A7 올스텐										
4	원더스리빙 원더스 쿡킹 A9올스텐										
5	일리 프란시스 Y3.3										
6	삼성전자 비스포크 RF85A9103AP										
7	삼성전자 비스포크 RB33A3004AP (색상선택)										
8	삼성전자 삼성 RS84T5081SA										
9	삼성전자 비스포크 RF85A9001AP (색상선택)										
10	아이프리 레가토 시그널										
11	네스프레소 버쥬오 플러스										
12	바란 분유포트										
13	쿠쿠 CRP-LHTR0610FW 트윈프레셔										
14	LG전자 오브제컬렉션 DUBJ2EA										
15	쿠쿠 CRP-DHP0610FD										
16	네스프레소 에센자 미니 C30										
17	쿠쿠전자 쿠쿠 CRP-CHP1010FD										
18	휴렉 HD-9000SD										
19	발뮤다 더 토스터 K05B (2021년형)										
20	스마트카라 PCS-400										
21	LG전자 LG퓨리케어 WD102AW(렌탈)										
22	LG전자 디오스 DUB22MA										
23	네스프레소 시티즈 D113										

상세 페이지 정보 가져와서 저장하기

- 상세 페이지 url 가져오기

```
product_page_urls = []  
for cover in soup.find_all('p',{'class':'cont'}):  
    link=cover.select_one('a').get('href')  
    product_page_urls.append(link)  
print(product_page_urls)  
len(product_page_urls)
```

상세 페이지 정보 가져와서 저장하기

- 상세 페이지에서 필요한 정보 가져와서 저장하기

```
save_file_name = 'naver_best_product_inf.csv'
columns_list=['제품명','평점','최저가','배송비']
f=open(save_file_name, 'w+', newline='')
cw = csv.writer(f)
cw.writerow(columns_list)
```

상세 페이지 정보 가져와서 저장하기

- 상세 페이지에서 필요한 정보 가져와서 저장하기

```
for index, product_page_url in enumerate(product_page_urls):
    html=urlopen(product_page_url)
    bsObject=BeautifulSoup(html,"html.parser")
    product_inf=[]
    title=bsObject.select('div.top_summary_title__15yAr > h2')[0].text
    product_inf.append(title)
    try:
        score=bsObject.select('div.top_grade__3jjdl')[0].text
    except:
        score=""
    product_inf.append(score)
```

스코어가 없는 경우 존재!!

상세 페이지 정보 가져와서 저장하기

- 상세 페이지에서 필요한 정보 가져와서 저장하기

```
min_price=bsObject.select('em.lowestPrice_num__3AIQ-')[0].text
product_inf.append(min_price)
delivery_inf=bsObject.select('div.lowestPrice_delivery_price__3f-2I')[0].text
product_inf.append(delivery_inf)

print(index+1,title, score, min_price, delivery_inf)
cw.writerow(product_inf)
time.sleep(random.uniform(1,3)) # 1~3 초 사이의 랜덤한 시간으로 쉬어 진행

f.close()
print('\n!!!! Product Information Web Crawling Complete !!!!\n')
```

상세 페이지 정보 가져와서 저장하기

	A	B	C	D	E	F	G	H
1	제품명	평점	최저가	배송비				
2	2021년형리하스 올스텐 KHD-16L	평점4.7	129,000	무료배송리하스				
3	원더스리빙 원더스 쿡킹 A7 올스텐	평점4.8	129,000	무료배송원더스리빙				
4	원더스리빙 원더스 쿡킹 A9올스텐	평점4.8	199,000	무료배송원더스리빙				
5	일리 프란시스 Y3.3	평점4.8	121,890	배송비 3,000원옥션				
6	삼성전자 비스포크 RF85A9103AP	평점4.9	1,370,000	무료배송하이마트쇼핑몰				
7	삼성전자 비스포크 RB33A3004AP (색상선택)	평점4.8	609,000	무료배송하이마트쇼핑몰				
8	삼성전자 삼성 RS84T5081SA	평점4.8	1,212,800	무료배송11번가				
9	삼성전자 비스포크 RF85A9001AP (색상선택)	평점4.8	1,704,020	무료배송11번가				
10	아이프리 레가토 시그널	평점5	169,000	무료배송아이프리 스토어				
11	네스프레소 버쥬오 플러스	평점4.8	229,000	무료배송쿠팡				
12	바란 분유포트	평점5	69,000	무료배송비비딕				
13	쿠쿠 CRP-LHTR0610FW 트윈프레셔	평점4.8	402,160	배송비 5,000원티몬				
14	LG전자 오브제컬렉션 DUBJ2EA	평점4.9	1,320,700	배송비 30,000원11번가				
15	쿠쿠 CRP-DHP0610FD	평점4.7	214,730	배송비 5,000원티몬				
16	네스프레소 에센자 미니 C30	평점4.9	125,600	무료배송오케이몰				
17	쿠쿠전자 쿠쿠 CRP-CHP1010FD	평점4.7	255,520	무료배송티몬				
18	휴렉 HD-9000SD	평점4.8	890,000	무료배송HULEC				
19	2021년형발뮤다 더 토스터 K05B (2021년형)	평점4.9	247,020	무료배송G9				
20	2020년형스마트카라 PCS-400	평점4.7	667,830	무료배송니즈라이프몰				
21	렌탈LG전자 LG퓨리케어 WD102AW(렌탈)	평점4.9	18,800	무료배송LG퓨리케어판매처				
22	LG전자 디오스 DUB22MA	평점4.8	1,193,000	무료배송11번가				
23	네스프레소 시티즈 D113	평점4.9	209,900	무료배송우리아가				
24	렌탈LG전자 LG퓨리케어 WD502AW(렌탈)	평점5	27,800	무료배송LG퓨리케어판매처				



Lab 7-5

셀레니움 사용하기

목차

- 셀레니움 사용하기
- 셀레니움 선택자 사용

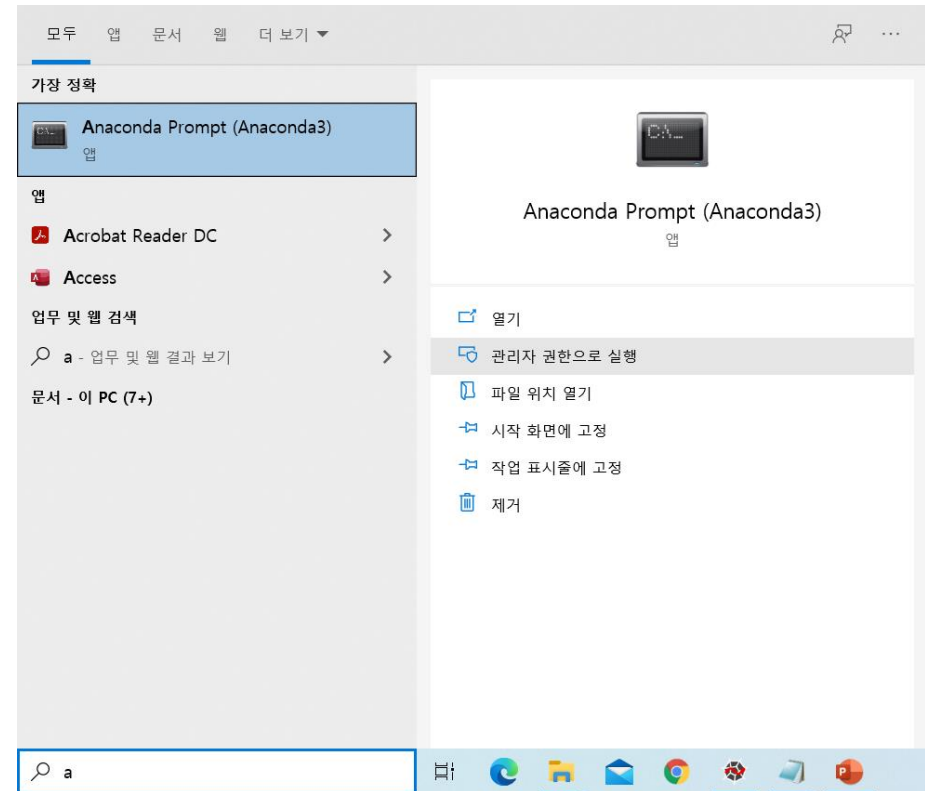
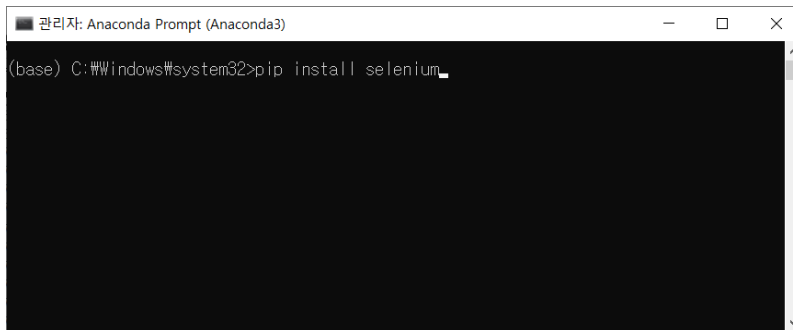
셀레니움

- 원래 웹사이트 테스트 목적으로 개발
- 강력한 웹 스크레이핑 도구로 사용할 수 있음
- 웹브라우저에서 수행하는 일을 자동화하는 것이 가능
- 자체 웹 브라우저가 있지 않음
 - 다른 웹 브라우저 사용
 - 백그라운드 실행을 원할 시
 - 팬텀JS 사용



셀레니움 설치하기

- selenium 설치
 - Anaconda prompt(관리자 권한)에서 다음 실행
 - pip install selenium



셀레니움 설치하기

- 크롬 드라이버 설치

- 다운로드

- <https://sites.google.com/a/chromium.org/chromedriver/downloads>
 - PC에 설치된 크롬 버전 확인 필요
 - 크롬 창에 chrome://version 입력
 - OS에 맞는 버전 다운로드 후 압축 풀기

셀레니움 설치하기

- 크롬 드라이버 설치

The image shows a Windows command prompt window on the left and the ChromeDriver website on the right. In the command prompt, the command `chrome --version` has been executed, and the output shows the Chrome version as 93.0.4577.82, which is circled in red. The ChromeDriver website on the right has a navigation menu on the left with 'Downloads' selected. The main content area features a large red warning message: 'Please note that we have migrated to a new ChromeDriver site. Current site will be deprecated soon.' Below this, under the 'Current Releases' section, there is a list of download links. The link for 'ChromeDriver 93.0.4577.63' is circled in red, matching the Chrome version in the command prompt.

ChromeDriver - WebDriver for Chrome

Search this site

Downloads

Please note that we have migrated to a new ChromeDriver site. Current site will be deprecated soon.

Current Releases

- If you are using Chrome version 94, please download [ChromeDriver 94.0.4606.41](#)
- If you are using Chrome version 93, please download [ChromeDriver 93.0.4577.63](#)
- If you are using Chrome version 92, please download [ChromeDriver 92.0.4515.107](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

Chrome: 93.0.4577.82 (공식 빌드) (a3a25d9b9e2d0b728e045ec87c heads/4577@1237)

OS: Windows 10 OS Version 2009

JavaScript: V8 9.3.345.19

사용자 에이전트: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/93.0.4577.82 Safari/537.36

명령줄: "C:\Program Files\Google\Chrome\Application\chrome.exe" --flag-switches-begin --flag-switches-end

실행 가능 경로: C:\Program Files\Google\Chrome\Application\chrome.exe

프로필 경로: C:\Users\USER\AppData\Local\Google\Chrome\User Data\Default

유사 버전: af81735d-ca7d8d80 4dc415b1-ca7d8d80 47722f0d-ca7d8d80 85bf39f3-ca7d8d80 19d59c16-377be55a 90a7075b-5dd5f5228 16b16054-377be55a 8e73c278-ca7d8d80 1fa5b2f3-2d4bb88b 59b6f412-377be55a 60d4b352-377be55a 5fff72eb-377be55a 152bdf52-ca7d8d80

CHROMEDRIVER

CAPABILITIES & CHROME OPTIONS

CHROME EXTENSIONS

CHROMEDRIVER CANARY

CONTRIBUTING

DOWNLOADS

VERSION SELECTION

GETTING STARTED

ANDROID

CHROME OS

LOGGING

PERFORMANCE LOG

MOBILE EMULATION

NEED HELP?

CHROME DOESN'T START OR CRASHES IMMEDIATELY

첫 번째 셀레니움 프로그램

- 구글에 검색하기

```
from urllib.parse import quote_plus  
from bs4 import BeautifulSoup  
from selenium import webdriver
```

```
baseUrl = 'https://www.google.com/search?q='  
plusUrl = input('무엇을 검색할까요? :')  
url = baseUrl + quote_plus(plusUrl) # quote_plus가 한글 변환
```

```
driver = webdriver.Chrome(executable_path=r'D:/kjuw/.../chromedriver.exe')  
driver.get(url)
```

첫 번째 셀레니움 프로그램

- 구글에 검색하기

```
html = driver.page_source
soup = BeautifulSoup(html, 'html.parser')

v = soup.select('div.yuRUbf')

for i in v:
    print(i.select_one('h3.LC20lb.DKV0Md').text)
    print(i.a.attrs['href'])
    print()

driver.close()
```

LG전자: LGE.COM
<https://www.lge.co.kr/>

스토어 - LG전자
<https://www.lge.co.kr/store>

LG그룹
<https://www.lg.co.kr/>

개인 | LG U+
<https://www.uplus.co.kr/>

LG전자 - 나무위키:대문
<https://namu.wiki/w/LG%EC%A0%84%EC%9E%90>

서울의 자존심 LG트윈스!
<https://www.lgtwins.com/>

셀레니움 선택자

- find_element_by_와 find_elements_by_
 - find_elements_by_tag_name
 - find_elements_by_xpath
 - find_elements_by_id
 - find_elements_by_class_name
 - find_elements_by_link_text
 - find_elements_by_partial_link_text
 - find_elements_by_css_selector

셀레니움 선택자

- xpath 사용하기

표현	설명
nodename	Node명이"nodename"인 Node선택
/	루트Node로 부터 선택
//	현재 Node로부터 문서상의 모든 Node를 조회
.	현재Node 선택
..	현재Node의 부모Node 선택
@	현재Node의 속성선택

셀레니움 선택자 사용

LG전자 디오스 F873S11E: 네이버 x

search.shopping.naver.com/catalog/23550945493?query=LG냉장고&NaPm=ct%3Dktpqedoo%7Cci%3Dfe4ac987d923bce5b3c8eaea30...

Chrome이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.

NAVER 네이버페이 강박구미 로그인

쇼핑 쇼핑 LIVE 상세검색 카테고리 더보기 베스트100 물 전체보기 쇼핑 MY

홈 백화점원도 아울렛원도 스타일원도 디자이너원도 뷰티원도 럭셔리 리빙원도 푸드원도 장보기 키즈원도 펫원도 플레이원도 아트원도

디지털/가전 > 주방가전 > 냉장고 > 양문형냉장고

LG전자 디오스 F873S11E ★★★★★ 4.8

브랜드 카탈로그 | 제조사 LG전자 | 브랜드 디오스 | 등록일 2020.07. | ❤️ 찜하기 3,309 | 📄 정보 수정요청

에너지효율: 1등급 | 용량: 870L | 품목: 4도어냉장고, 냉장고, 냉동고set | 냉장실: 503L | 냉동실: 367L | 서랍: 신선야채실, 스마트디바이더
컴프레서: 리니어 | 탈취: UV안식계균Plus | 도어함들: 스웨이함들 | 스마트기능: 스마트진단 | 재질: 메탈 | 소비전력: 24.8kWh

더보기 >

'LG전자' 브랜드스토어 바로가기 →
브랜드스토어에서 더 많은 브랜드 제품과 브랜드 행사 소식을 만나보세요.

1등급

최저 1,596,390원
무료배송 | 티몬

구매처별 최저가
인기순 최저가순 배송비포함 OFF 카드할인 OFF

기본구성

판매처	판매가	배송비	사라가기
타임커머스 TMON	1,596,390	무료배송	사라가기
Gmarket	최저 1,596,490	30,000원	사라가기
오늘의집 N Pay	1,596,600	무료배송	사라가기

TOP

셀레니움 선택자 사용

```
from selenium import webdriver
```

```
browser = webdriver.Chrome(executable_path=r'D:\...\week7_자연어처리\chromedriver.exe')  
browser.get("https://search.shopping.naver.com/...")
```

제품명 추출 xpath 사용 - 절대 경로

```
product_name = browser.find_element_by_xpath("/html/body/div/div/div[2]/div[2]/div[1]/h2")  
print(product_name.text)
```

제품명 추출 xpath 사용 - 상대 경로

```
product_name = browser.find_element_by_xpath("//div[@class='top_summary_title__15yAr']/h2")  
print(product_name.text)  
product_name = browser.find_element_by_xpath('//*[@id="__next"]/div/div[2]/div[2]/div[1]/h2')  
print(product_name.text)
```

셀레니움 선택자 사용

제품명 추출 css_selector 사용

```
#__next > div > div.style_container__3iYev > div.style_inner__1Eo2z > div.top_summary_title__15yAr > h2
```

```
product_name = browser.find_element_by_css_selector("div.top_summary_title__15yAr > h2")
```

```
print(product_name.text)
```

셀레니움 선택자 사용

- 리뷰 제목 가져오기

리뷰 제목들 가져오기 xpath 사용

```
review_titles = browser.find_elements_by_xpath('//*[ @id="section_review"]/ul/li[*]/div[2]/div/em')
for title in review_titles:
    print(title.text)
```

리뷰 제목들 가져오기 css_selector 사용

```
review_titles = browser.find_elements_by_css_selector("div.reviewItems_review__1eF8A > div > em")
for title in review_titles:
    print(title.text)
```



Lab 7-7

셀레니움 선택자 사용

- 여러 페이지 리뷰 제목 가져오기



LG전자 디오스 F873S11E
최저 1,596,390원 무료배송

최저가 사러가기

판매처 49

제품정보

블로그리뷰

쇼핑몰리뷰 1,370

AiTEMS추천

그리고 배송 당일.

리뷰펼치기 ▾

★★★★★ 5

11번가 · bl***** · 21.09.16.

최고예요

일요일에 구매했는데 월요일에 해피콜이 있었고 화요일 경오 쯤 설치가 완료 됐어요. 주택 2층이라 옮기는데 난관이 있었지만 잘 설치해 주셨고 흠이 없는지도 확인하고 냉장 고에 대한 설명도 잘 해주셨어요. 음식을 아이스박스에 다 옮겨 놓긴했지만 상할까 많이 걱정됐는데 다행히 일찍 배송이 돼서 정말 다행이었어요. 설치한지 3일 쯤인데 잘 가 동 되고 있고 이전에 쓰던 양문형 보다 상하4문이 정말 편하고 좋네요.

리뷰펼치기 ▾

1

2

3

4

5

6

7

8

9

10

다음 >

셀레니움 선택자 사용

- 페이지 내 리뷰 제목, 날짜, 내용 가져오기

```
from selenium import webdriver
```

```
browser =
```

```
webdriver.Chrome(executable_path=r'D:\kjh\lecture\LG전자_21\week7_자연어처리\chromedriver.exe')
```

```
browser.get("https://search.shopping.naver.com/catalog/23550945493?query=....")
```

```
# review title
```

```
review_titles = browser.find_elements_by_css_selector('div.reviewItems_review__1eF8A > div > em')
```

```
for title in review_titles:
```

```
    print("리뷰 제목 : " + title.text)
```

셀레니움 선택자 사용

- 페이지 내 리뷰 제목, 날짜, 내용 가져오기

```
# review date
```

```
#section_review > ul > li:nth-child(1) > div.reviewItems_etc_area__2P8i3 > span:nth-child(4)
```

```
review_dates = browser.find_elements_by_css_selector('div.reviewItems_etc_area__2P8i3 > span:nth-child(4)')
```

```
for date in review_dates:
```

```
    print("리뷰 날짜 : " + date.text)
```

```
# review contents
```

```
#section_review > ul > li:nth-child(1) > div.reviewItems_review__1eF8A > div > p
```

```
review_contents = browser.find_elements_by_css_selector('div.reviewItems_review__1eF8A > div > p')
```

```
for content in review_contents:
```

```
    print("리뷰 내용 : " + content.text)
```

셀레니움 선택자 사용

- 한 페이지 내용 저장하기

```
import pandas as pd
import os
```

```
df = pd.DataFrame(columns=["title", "date", 'content'])
for i in range(len(review_titles)):
    df = df.append({'title':review_titles[i].text,
                    'date':review_dates[i].text,
                    'content':review_contents[i].text}, ignore_index=True)
print(df)
```

셀레니움 선택자 사용

- 한 페이지 내용 저장하기

```
save_file_path = 'D:\kjl\lecture\LG전자_21\week7_자연어처리'  
os.chdir(save_file_path)  
df.to_csv('ref_review_page1.csv', index=False, encoding='utf-8-sig')
```

셀레니움 선택자 사용

- 여러 페이지 저장하기

페이지 번호 클릭하기

```
#section_review > div.pagination_pagination__2M9a4 > a.pagination_now__gZWGP.pointer # 첫 페이지
```

```
#section_review > div.pagination_pagination__2M9a4 > a:nth-child(2)
```

```
#section_review > div.pagination_pagination__2M9a4 > a:nth-child(3)
```

```
#section_review > div.pagination_pagination__2M9a4 > a.pagination_next__3ycRH.pointer # 다음
```

2번째 페이지로 이동

```
btn_page = browser.find_element_by_css_selector('#section_review >  
div.pagination_pagination__2M9a4 > a:nth-child(2)')  
btn_page.click()
```

셀레니움 선택자 사용

- 한 페이지 내 리뷰 제목, 날짜 내용 저장 함수 만들기

```
def get_review():  
    review_titles = browser.find_elements_by_css_selector('div.reviewItems_review__1eF8A > div > em')  
    review_dates = browser.find_elements_by_css_selector('div.reviewItems_etc_area__2P8i3 > span:nth-child(4)')  
    review_contents = browser.find_elements_by_css_selector('div.reviewItems_review__1eF8A > div > p')  
  
    df = pd.DataFrame(columns=['title','date','content'])  
    for i in range(len(review_titles)):  
        df = df.append({'title':review_titles[i].text,  
                        'date':review_dates[i].text,  
                        'content':review_contents[i].text}, ignore_index=True)  
    return(df)
```

셀레니움 선택자 사용

- 여러 페이지 크롤링하기

```
import pandas as pd
import os
import time
import random
from selenium import webdriver
```

```
browser = webdriver.Chrome(executable_path=r'D:\kjl...\chromedriver.exe')
browser.get("https://search.shopping.naver.com/catalog...")
```

```
all_review = pd.DataFrame(columns=['title', 'date', 'content'])
iter = 2 # 10 페이지씩 2번만 반복
```


셀레니움 선택자 사용

- 여러 페이지 크롤링하기

```
for i in range(iter):
    for j in range(1,11):
        review = get_review()
        all_review=pd.concat([all_review,review], ignore_index=True)
        if j < 10:
            btn_page = browser.find_element_by_css_selector('#section_review > div.pagination_pagination__2M9a4 > a:nth-child('+str(j+1)+')')
        else:
            btn_page = browser.find_element_by_css_selector('#section_review > div.pagination_pagination__2M9a4 > a.pagination_next__3ycRH.pointer')
            btn_page.click()
            time.sleep(random.uniform(1,3)) # 1~3 초 사이의 랜덤한 시간으로 쉬어 진행
        print("Iteration:"+str(i))
```

셀레니움 선택자 사용

- 여러 페이지 크롤링하기

```
print("Crawling Finished")
save_file_path = 'D:\kju\lecture\LG전자_21\week7_자연어처리'
os.chdir(save_file_path)
all_review.to_csv('ref_all_review.csv',index=False,encoding='utf-8-sig')
browser.close()
```

셀레니움 선택자 사용

- 여러 페이지 크롤링하기

	A	B	C	D	E	F	G	H	I	J	K	L
1	title	date	content									
2	실물이 예	21.08.14.	실물이 예상했던것 보다 훨씬 마음에 들었고 가격에 비해 굉장히 좋은 4도어 냉장고라고 느꼈습니다.									
			어머니댁 냉장고가 고장나서 검색중 평 이 좋아 주문 했습 니다. 제품은 제 가 보질 못했지만 어머니도 좋다고 하 시긴 한데 배송 과정 이 좀 아 쉽네요. 제품만은									
	어머니댁	21.08.10.										



Lab 7-8

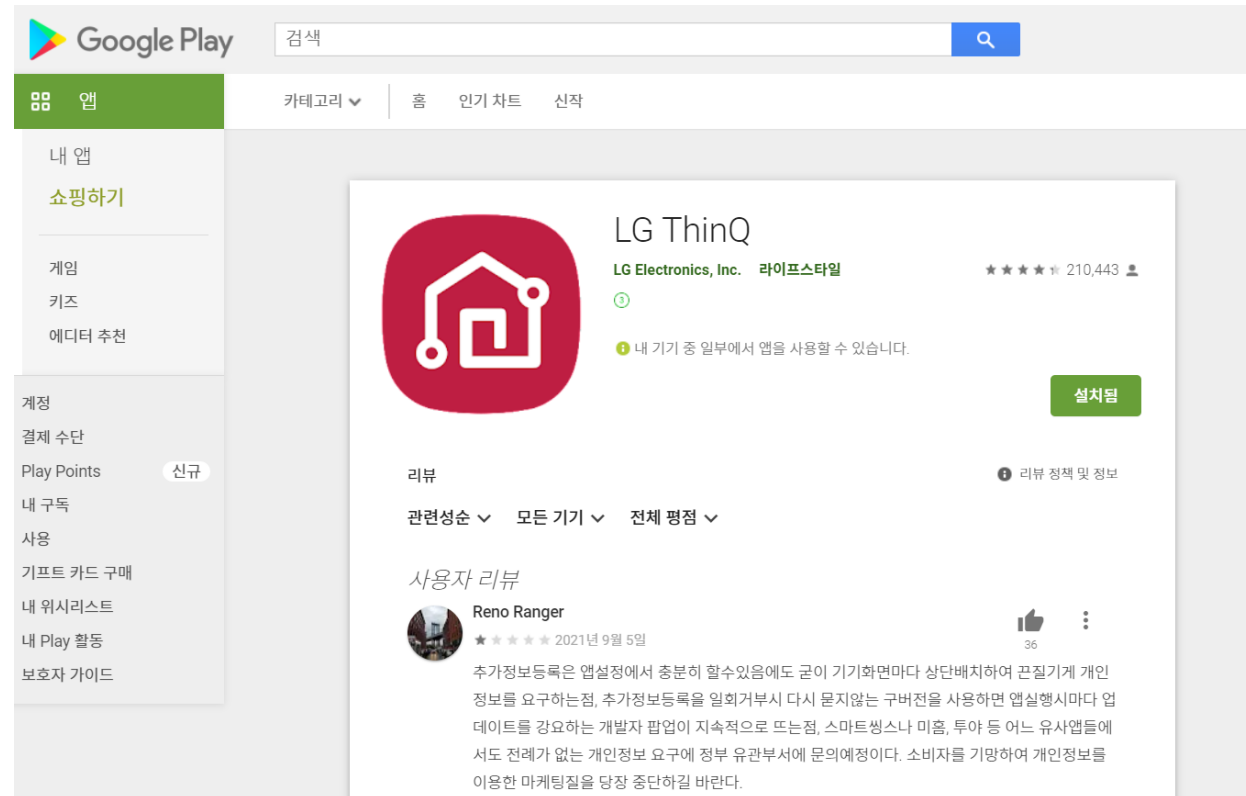
셀레니움을 이용한 웹크롤링 예제

목차

- Google Play 크롤링 예제

Google Play 크롤링 예제

- LG ThinQ 앱 리뷰
 - <https://play.google.com/store/apps/details?id=com.lgeha.nuts&showAllReviews=true>



Google Play 크롤링 예제

- driver 생성

```
from selenium import webdriver  
import pandas as pd  
import time
```

```
url = "https://play.google.com/store/apps  
      /details?id=com.lgeha.nuts&showAllReviews=true"  
driverPath = "D:/../chromedriver.exe"  
driver = webdriver.Chrome(driverPath)  
driver.get(url)
```

Google Play 크롤링 예제

- 개별 리뷰 살펴보기
 - 리뷰, 날짜, 좋아요 수, 별점 데이터

리뷰

```
comment = driver.find_element_by_xpath("//span[@jsname='bN97Pc']")  
comment.text
```

날짜

```
date = driver.find_element_by_xpath("//span[@class='p2TkOb']")  
date.text
```


Google Play 크롤링 예제

- 개별 리뷰 살펴보기

```
# 좋아요 수
```

```
like = driver.find_element_by_xpath("//div[@aria-label=  
    '이 리뷰가 유용하다는 평가를 받은 횟수입니다.']")
```

```
like.text
```

```
# 별점 데이터
```

```
star = driver.find_element_by_xpath("//span[@class='nt2C1d']/  
    div[@class='pf5lle']/div[@role='img']")
```

```
star.get_attribute('aria-label')
```

```
star.get_attribute('aria-label')[10]
```

Google Play 크롤링 예제

- 전체 리뷰 개수 살펴보기

```
comments = driver.find_elements_by_xpath("//span[@jsname='bN97Pc']")  
len(comments)
```

40

Google Play 크롤링 예제

- 스크롤 내리기와 더보기 누르기

```
driver.execute_script("window.scrollTo(0, document.body.scrollHeight)")
```

```
# 4번 스크롤 내리기
```

```
for i in range(4):
```

```
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight)")  
    time.sleep(2)
```

```
# 더보기 누르기
```

```
driver.find_element_by_xpath("//span[@class='RveJvd snByac']").click()
```

Google Play 크롤링 예제

- 스크롤과 더 보기 합하기

```
SCROLL_PAUSE_TIME = 3
last_height = driver.execute_script("return document.body.scrollHeight")
while True:
    # (1) 4번의 스크롤링
    for i in range(4):
        driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
        time.sleep(SCROLL_PAUSE_TIME)
    # (2) 더보기 클릭
    try:
        driver.find_element_by_xpath("//span[@class='RveJvd snByac']").click()
    except:
        break
    # (3) 종료 조건
    new_height = driver.execute_script("return document.body.scrollHeight")
    if new_height == last_height:
        break
    last_height = new_height
```

Google Play 크롤링 예제

- 편의상 한 번만 반복하기

```
for i in range(4):  
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")  
    time.sleep(SCROLL_PAUSE_TIME)
```

```
    # (2) 더보기 클릭  
    driver.find_element_by_xpath("//span[@class='RveJvd snByac']").click()
```

Google Play 크롤링 예제

- 데이터 가져오기

```
comments = driver.find_elements_by_xpath("//span[@jsname='bN97Pc']")
```

```
len(comments)
```

```
# 날짜
```

```
dates = driver.find_elements_by_xpath("//span[@class='p2TkOb']")
```

```
len(dates)
```

```
# 좋아요 수
```

```
likes = driver.find_elements_by_xpath(\
```

```
    "//div[@aria-label='이 리뷰가 유용하다는 평가를 받은 횟수입니다.']")
```

```
len(likes)
```

```
# 별점 데이터
```

```
stars = driver.find_elements_by_xpath(\
```

```
    "//span[@class='nt2C1d']/div[@class='pf5lle']/div[@role='img']")
```

```
len(stars)
```

Google Play 크롤링 예제

- 데이터 프레임에 저장하기

```
# Make a dataframe
res_dict = []
for i in range(len(comments)):
    res_dict.append({
        'date' : dates[i].text,
        'star' : stars[i].get_attribute('aria-label')[10],
        'like' : likes[i].text,
        'comment' : comments[i].text })
res_df = pd.DataFrame(res_dict)
res_df
```

Google Play 크롤링 예제

- csv 파일에 저장하기

```
import os
save_file_path = 'D:\kju\lecture\LG전자_21\week7_자연어처리'
os.chdir(save_file_path)
res_df.to_csv('thinq_review.csv', index=False, encoding='utf-8-sig')
```

	A	B	C	D	E	F	G	H	I	J
1	date	star	like	comment						
2	2021년 09월 05일	1	36	추가정보등록은 앱설정에서 충분히 할수있음에도 굳이 기기화면마다						
3	2021년 09월 06일	1	15	그냥 엘지제품은 리모컨만 사용하는것이 정신 건강에 좋을 듯합니다						
4	2021년 09월 05일	1	8	퓨리케어 공기청정기 1단짜리 사용 중인데 언제부터 어플에서 공기						
5	2021년 09월 06일	2	2	에너지모니터링 . 지난달 대비 사용량... 이전달하고 이번달하고 바꿨						
6	2021년 09월 07일	1		LG앱만.서비스오작동기능들이잘안되고짜증나네.통신사이동하고삼						
7	2021년 09월 08일	4		24시간매장운영중인데 에어컨을 새벽에 켜놓고 안고고 가는사람이						
8	2021년 09월 15일	2	1	청소구역설정에서 적어도 2이상의 도면을 설정할 수 있게 개선해야						
9	2021년 09월 16일	2		위젯 투명으로 만들어주세요! 요즘 누가 화이트배경 써요? 웬만한 어						



Lab 7-9

정리

- 웹크롤링과 HTML
- BeautifulSoup 사용하기
- BeautifulSoup을 이용한 웹크롤링 예제
- 셀레니움 사용하기
- 셀레니움을 이용한 웹크롤링 예제