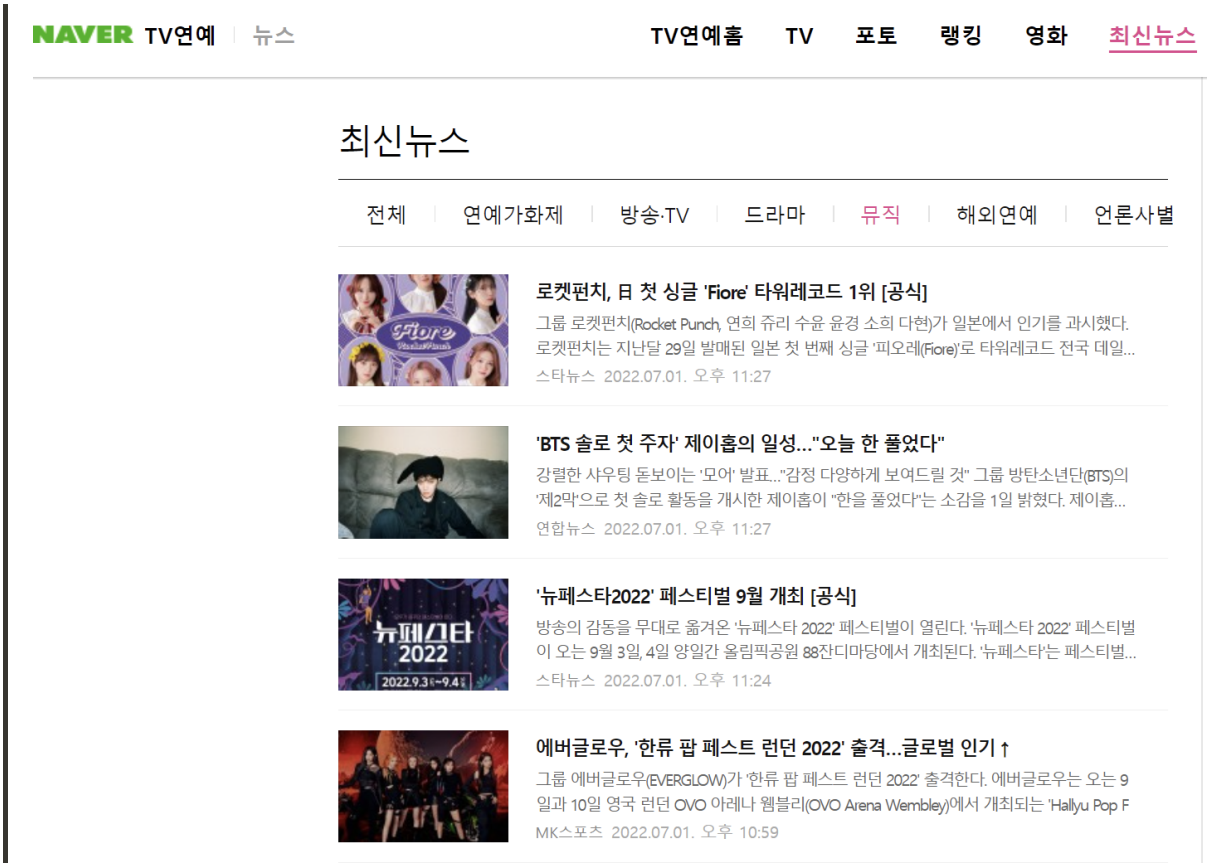
미션2 - 크롤링

> 해성, 잘했어.
>
> 이제 웹크롤링을 공부해보자.
>
> 아래 토닥토닥 파이썬- 데이터수집을 교재로 추천한다.
>
> https://wikidocs.net/book/2452
>
> 이책의 모든 챕터를 모두 공부할 필요는 없고, 1장, 2장, 4장 네이버 뉴스 수집 부분만 먼저 공부해보렴.
>
> 아래 스터디 자료도 쉽게 설명되어 있으니, 필요하면 참고하렴.
>
> 네이버 함께 정복하는 무적의 크롤링
>
> https://book.coalastudy.com/data-crawling/
>
> - beautifulsoup 공식 사이트도 북마킹했다가 사전처럼 참고하시고
>
> https://www.crummy.com/software/BeautifulSoup/
>
> https://www.crummy.com/software/BeautifulSoup/bs4/doc.ko/
>
> 이번 미션의 목표는 정적 웹사이트를 수집하는 것이고, 네이버 TV연예/최신뉴스/뮤직의 7월1일 전체 뉴스를 수집해보자꾸나.
>
> https://entertain.naver.com/now?sid=7a5#sid=7a5&date=2022-07-01&page=1
>
> 크롬에서 오른쪽 마우스 클릭하면 '검사' 메뉴가 나와. 그럼 현재 보고 있는 html 의 소스를 쉽게 볼 수 있단다. 소스위에서 마우스를 움직이면 현재 소스가 어느 부분을 커버하는지 뉴스의 해당 부분에 색이 바뀐단다. (이미 알고 있지?)
>
> 일단 7월1일자 첫페이지의 뉴스 25건부터 수집해서 출력해보고 성공하면 나에게 알려주렴.
>
> 굿럭!
>
> 최준연
>
> PS> 뮤직 말고 다른 카테고리의 뉴스를 수집해도 괜찮아. 다른 카테고리 수집해보고 싶으면 알려주렴.



웹 크롤링

> 위키백과
>
> **웹 크롤러**(web crawler)는 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램이다.
>
> 웹 크롤러가 하는 작업을 '웹 크롤링'(web crawling) 혹은 '스파이더링'(spidering)이라 부른다. 검색 엔진과 같은 여러 사이트에서는 데이터의 최신 상태 유지를 위해 웹 크롤링한다. 웹 크롤러는 대체로 방문한 사이트의 모든 페이지의 복사본을 생성하는 데 사용되며, 검색 엔진은 이렇게 생성된 페이지를 보다 빠른 검색을 위해 인덱싱한다. 또한 크롤러는 링크 체크나 HTML 코드 검증과 같은 웹 사이트의 자동 유지 관리 작업을 위해 사용되기도 하며, 자동 이메일 수집과 같은 웹 페이지의 특정 형태의 정보를 수집하는 데도 사용된다.

정적 웹사이트

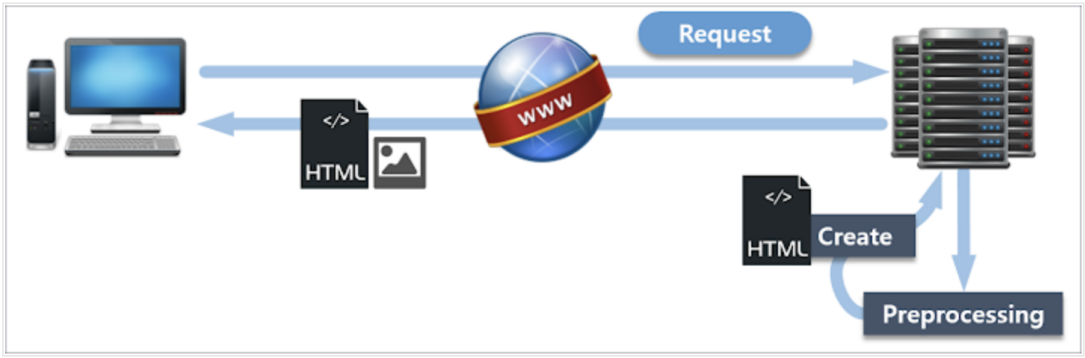
- 정적 웹사이트와 정적 웹사이트의 차이 (https://titus94.tistory.com/4)

정적 웹 페이지 (Static Web Page)

서버(웹 서버, Web Server)에 **미리 저장된 파일**(HTML 파일, 이미지, JavaScript 파일 등)이 그대로 전달되는 웹 페이지
서버는 사용자가 요청(Request)에 해당하는 저장된 웹 페이지를 보냄
사용자는 서버에 저장된 데이터가 변경되지 않는 한 고정된 웹 페이지를 보게 됨

동적 웹 페이지 (Dynamic Web Page)

서버(웹 서버, Web Server)에 있는 데이터들을 스크립트에 의해 **가공처리한 후 생성**되어 전달되는 웹 페이지
서버는 사용자의 요청(Request)을 해석하여 데이터를 가공한 후 생성되는 웹 페이지를 보냄
사용자는 상황, 시간, 요청 등에 따라 달라지는 웹 페이지를 보게 됨



- 정적 웹 페이지 : 서버에 미리 저장되 있던 파일이 그대로 전달되는 웹페이지
- 동적 웹 페이지 : 서버에 있는 데이터들을 스크립테에 의해 가공처리한 후 생성되어 전달되는 웹 페이지
- 뉴스 기사 각각의 페이지 처럼 DB에 저장된 내용이 그대로 전달되는 것은 정적 웹페이지라고 볼수 있을것 같다. 네이버 뉴스 메인 페이지는 동적 웹 페이지라고 보면 될것 같다. 사용자의 구독목록, 시간 등에 따라서 달라지기 때문이다.
- 마찬가지로 네이버 TV연예/최신뉴스/뮤직 페이지도 동적 웹 페이지라고 할 수 있을것 같다. 시간에 따라 최신의 뉴스를 보여주기 때문이다.
- 더 생각해보니 각각의 뉴스 기사 페이지 또한 동적 웹 페이지 인것 같다. 틀 형태로 된 html과 css가 존재하고 내용과 사진은 DB에 저장되어 있다가 사용자가 요청하는 파라미터를 보고 맞는 데이터를 가공하여 사용자에게 응답하는 방법일 것 같다. 이런 방법이 아니면 UI를 업데이트 해야할때 지금까지 만든 모든 html 파일들을 모두 수정해야하는 상황이 만들어 진다.
- 이번 과제는 이 뉴스페이지들이 정적 웹 사이트라고 가정하고 과제를 진행하면 될 것 같다.

뷰리플스프 공식문서 예제

```
from bs4 import BeautifulSoup

html_doc = """
<html><head><title>The Dormouse's story</title></head>

<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>

<p class="story">...</p>
"""

soup = BeautifulSoup(html_doc)
soup.prettify() # 들여쓰기된 html을 보여준다.

soup.title # 가장 먼저있는 title 테그를 가져온다.
# <title>The Dormouse's story</title>

soup.title.name # 타이틀 테그의 테그 이름을 가져온다.
# u'title

soup.title.string # 타이틀 테그안에 있는 글자
# u'The Dormouse's story'

soup.title.parent # 타이틀 테그의 부모 테그를 가져온다.
# <head><title>The Dormouse's story</title></head>
soup.title.parent.name # 타이틀 테그의 부모 테그의 이름을 가져온다.
# u'head'

soup.p # 가장 먼저 나오는 p 테그를 가져온다.
# <p class="title"><b>The Dormouse's story</b></p>
soup.p['class'] # 테그의 속성을 가져올때는 딕셔너리와 같은 문법을 사용하는듯 보인다.
# u'title'

soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

soup.find_all('a') # 모든 a 테그를 가져온다.
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find(id="link3") # find 함수에 id를 인자로 테그를 찾을 수 있다.
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>

for link in soup.find_all('a'): # 페이지에 모든 링크를 찾을 때 유용해 보인다.
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie

print(soup.get_text()) # 페이지의 모든 텍스트를 찾을 때 쓴다.
# The Dormouse's story
# . . .
# . . .
```

수프 만들기

문서를 해석하려면, 문서를 BeautifulSoup 구성자에 건네주자. 문자열 혹은 열린 파일 핸들을 건네면 된다:

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(open("index.html"))
soup = BeautifulSoup("<html>data</html>")
```

먼저, 문서는 유니코드로 변환되고 HTML 개체는 유니코드 문자로 변환된다:

```
BeautifulSoup("Sacr&eacute; bleu!")
<html><head></head><body>Sacré bleu!</body></html>
```

다음 뷰티플수프는 문서를 가장 적당한 해석기를 사용하여 해석한다. 특별히 XML 해석기를 사용하라고 지정해 주지 않으면 HTML 해석기를 사용한다. (XML 해석하기 참조.)

객체의 종류

뷰티플수프는 복합적인 HTML 문서를 파이썬 객체로 구성된 복합적인 문서로 변환한다. 그러나 객체의 종류를 다루는 법만 알면 된다.

태그

Tag 객체는 원래 문서의 XML 태그 또는 HTML 태그에 상응한다:

```
soup = BeautifulSoup('<b class="boldest">Extremely bold</b>')
tag = soup.b
type(tag)
# <class 'bs4.element.Tag'>
```

태그는 많은 속성과 메쏘드가 있지만, 그 대부분을 나중에 트리 항해하기 그리고 트리 검색하기에서 다룰 생각이다. 지금은 태그의 가장 중요한 특징인 이름과 속성을 설명한다.

이름

태그마다 이름이 있고, 다음 .name 과 같이 접근할 수 있다:

```
tag.name
# u'b'
```

태그의 이름을 바꾸면, 그 변화는 뷰티블수프가 생산한 HTML 조판에 반영된다:

```
tag.name = "blockquote"
tag
# <blockquote class="boldest">Extremely bold</blockquote>
```

속성

태그는 속성을 여러개 가질 수 있다. <b class="boldest"> 태그는 속성으로 “class”가 있는데 그 값은 “boldest”이다. 태그의 속성에는 사전처럼 태그를 반복해 접근하면 된다:

```
tag['class']
# u'boldest'
```

사전에 .attrs와 같이 바로 접근할 수 있다:

```
tag.attrs
# {u'class': u'boldest'}
```

태그의 속성을 추가, 제거, 변경할 수 있다. 역시 태그를 사전처럼 취급해서 처리한다:

```
tag['class'] = 'verybold'
tag['id'] = 1
tag
# <blockquote class="verybold" id="1">Extremely bold</blockquote>

del tag['class']
del tag['id']
tag
# <blockquote>Extremely bold</blockquote>

tag['class']
# KeyError: 'class'
print(tag.get('class'))
# None
```

값이-여럿인 속성

HTML 4에서 몇몇 속성은 값을 여러 개 가질 수 있도록 정의된다. HTML 5에서 그 중 2개는 제거되었지만, 몇 가지가 더 정의되었다. 가장 흔한 다중값 속성은 class이다 (다시 말해, 태그가 하나 이상의 CSS 클래스를 가질 수 있다). 다른 것으로는 rel, rev, accept-charset, headers, 그리고 accesskey가 포함된다. 뷰티플수프는 다중-값 속성의 값들을 리스트로 나타낸다:

```
css_soup = BeautifulSoup('<p class="body strikeout"></p>')
css_soup.p['class']
# ["body", "strikeout"]

css_soup = BeautifulSoup('<p class="body"></p>')
css_soup.p['class']
# ["body"]
```

속성에 하나 이상의 값이 있는 것처럼 보이지만, HTML 표준에 정의된 다중-값 속성이 아니라면, 뷰티플수프는 그 속성을 그대로 둔다:

```
id_soup = BeautifulSoup('<p id="my id"></p>')
id_soup.p['id']
# 'my id'
```

태그를 다시 문자열로 바꾸면, 다중-값 속성은 합병된다:

```
rel_soup = BeautifulSoup('<p>Back to the <a rel="index">homepage</a></p>')
rel_soup.a['rel']
# ['index']
rel_soup.a['rel'] = ['index', 'contents']
print(rel_soup.p)
# <p>Back to the <a rel="index contents">homepage</a></p>
```

문서를 XML로 해석하면, 다중-값 속성은 없다:

```
xml_soup = BeautifulSoup('<p class="body strikeout"></p>', 'xml')
xml_soup.p['class']
# u'body strikeout'
```

NavigableString

문자열은 태그 안에 있는 일군의 텍스트에 상응한다. 뷰티플수프는 NavigableString 클래스 안에다 이런 텍스트를 보관한다:

```
tag.string
# u'Extremely bold'
type(tag.string)
# <class 'bs4.element.NavigableString'>
```

NavigableString은 파이썬의 유니코드 문자열과 똑 같은데, 단 트리 항해하기와 트리 탐색하기에 기술된 특징들도 지원한다는 점이 다르다. NavigableString을 유니코드 문자열로 변환하려면 unicode()를 사용한다:

```
unicode_string = unicode(tag.string)
unicode_string
# u'Extremely bold'
type(unicode_string)
# <type 'unicode'>
```

문자열을 바로바로 편집할 수는 없지만, *replace_with()*을 사용하면 한 문자열을 또다른 문자열로 바꿀 수 있다:

```
tag.string.replace_with("No longer bold")
tag
# <blockquote>No longer bold</blockquote>
```

NavigableString은 트리 항해하기와 트리 탐색하기에 기술된 특징들을 모두는 아니지만, 대부분 지원한다. 특히, (태그에는 다른 문자열이나 또다른 태그가 담길 수 있지만) 문자열에는 다른 어떤 것도 담길 수 없기 때문에, 문자열은 .contents나 .string 속성, 또는 find() 메쏘드를 지원하지 않는다.

BeautifulSoup

BeautifulSoup 객체 자신은 문서 전체를 대표한다. 대부분의 목적에, 그것을 *Tag* 객체로 취급해도 좋다. 이것은 곧 트리 항해하기와 트리 검색하기에 기술된 메쏘드들을 지원한다는 뜻이다.

BeautifulSoup 객체는 실제 HTML 태그나 XML 태그에 상응하지 않기 때문에, 이름도 속성도 없다. 그러나 가끔 그의 이름 .name을 살펴보는 것이 유용할 경우가 있다. 그래서 특별히 .name에 “[document]”라는 이름이 주어졌다:

```
soup.name
# u'[document]'
```

주석과 기타 특수 문자열들

Tag, NavigableString, 그리고 BeautifulSoup 정도면 HTML이나 XML 파일에서 보게될 거의 모든 것들을 망라한다. 그러나 몇 가지 남은 것들이 있다. 아마도 신경쓸 필요가 있는 것이 유일하게 있다면 바로 주석이다:

```
markup = "<b><!--Hey, buddy. Want to buy a used parser?--></b>"
soup = BeautifulSoup(markup)
comment = soup.b.string
type(comment)
# <class 'bs4.element.Comment'>
```

Comment 객체는 그냥 특별한 유형의 NavigableString이다:

```
comment
# u'Hey, buddy. Want to buy a used parser'
```

그러나 HTML 문서의 일부에 나타나면, Comment는 특별한 형태로 화면에 표시된다:

```
print(soup.b.prettify())
# <b>
#  <!--Hey, buddy. Want to buy a used parser?-->
# </b>
```

뷰티플수프는 XML 문서에 나올만한 것들을 모두 클래스에다 정의한다: CData, ProcessingInstruction, Declaration, 그리고 Doctype이 그것이다. Comment와 똑같이, 이런 클래스들은 NavigableString의 하위클래스로서 자신의 문자열에 다른 어떤것들을 추가한다. 다음은 주석을 CDATA 블록으로 교체하는 예이다:

```
from bs4 import CData
cdata = CData("A CDATA block")
comment.replace_with(cdata)

print(soup.b.prettify())
# <b>
#  <![CDATA[A CDATA block]]>
# </b>
```

CSS 선택자

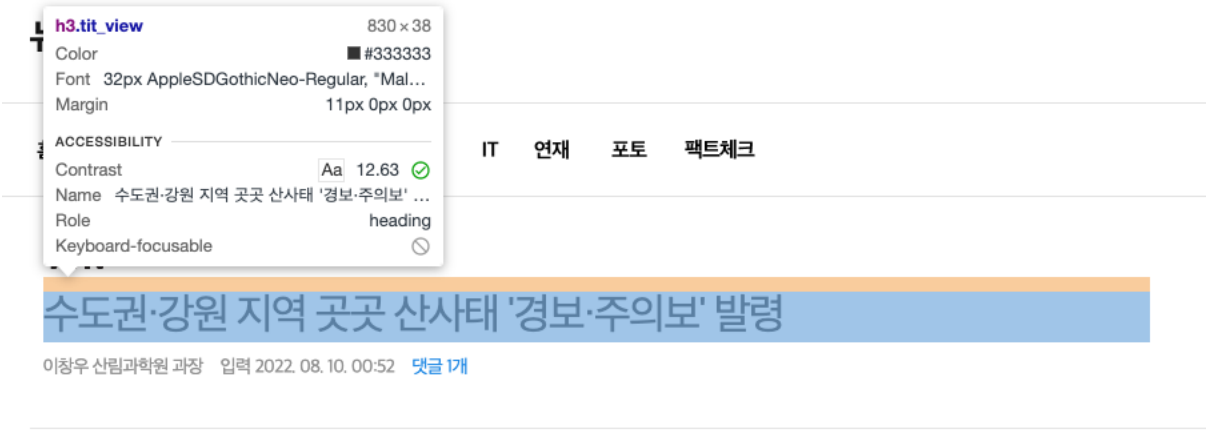
Tag, NavigableString, 그리고 BeautifulSoup 정도면 HTML이나 XML 파일에서 보게될 거의 모든 것들을 망라한다. 그러나 몇 가지 남은 것들이 있다. 아마도 신경쓸 필요가 있는 것이 유일하게 있다면 바로 주석이다:

```
soup.select("title")
# [<title>The Dormouse's story</title>]
```

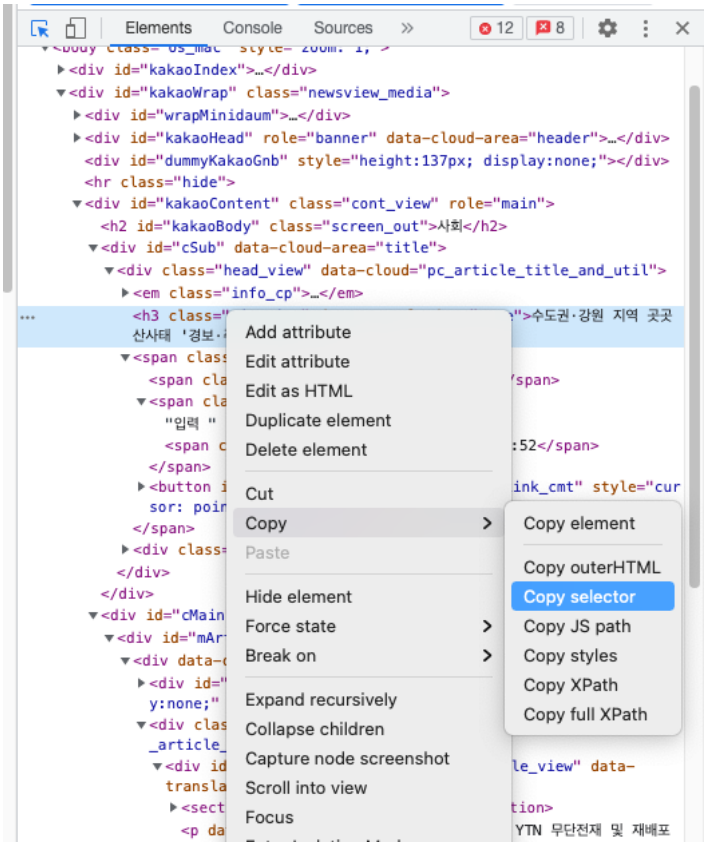
Pandas로 크롤링한 데이터 엑셀로 저장

```
import requests
from bs4 import BeautifulSoup
import pandas as pd # pandas 임포트

request = requests.get('https://news.daum.net/')

soup = BeautifulSoup(request.text, 'html.parser')

l = []
index = 0
for item in soup.find_all('div', class_ = 'item_issue'):
    link_txt = item.find('a')
    url = link_txt['href']
    print(index, url)
    l.append([index, url]) # 크롤링한 데이터 이중 리스트 형태로 저장
    index+=1

df = pd.DataFrame(l, columns=['index', 'url']) # 리스트를 데이터 프레임으로 변환
df.to_excel('daum_news.xlsx', index=False) # 엑셀로 변환
```

뉴스 상세 정보 크롤링



- 크롬 개발자 모드에서 돋보기 버튼을 누른후 원하는 부분을 클릭



- Copy → Copy selector 를 클릭
- select 또는 select_one의 인자로 사용

```
import requests
from bs4 import BeautifulSoup

request = requests.get('https://news.v.daum.net/v/20220810005210818')

soup = BeautifulSoup(request.text, 'html.parser')

title = soup.select_one('#cSub > div > h3').text
print(title)

date = soup.select_one('#cSub > div > span > span:nth-child(2) > span').text
print(date)

body = soup.select_one('#harmonyContainer > section').get_text().strip()
print(body)
```

과제 코드 - 오류 버전

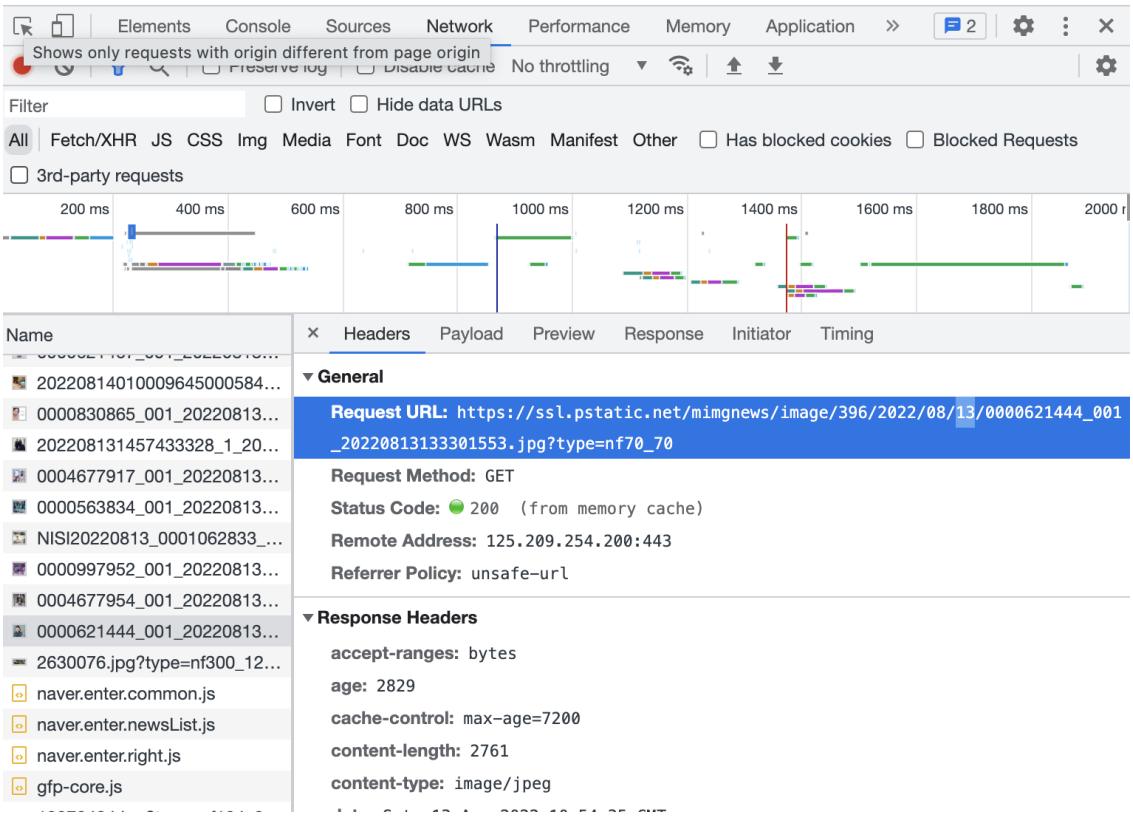
```
import requests
from bs4 import BeautifulSoup
import pandas as pd

def news_croller(url): # 페이지별 크롤링 코드
    request = requests.get(url)
    soup = BeautifulSoup(request.text, 'html.parser')

    # news_writing_time
    try:
        news_writing_time = soup.select_one('#content > div.end_ct > div > div.article_info > span:nth-child(1) > em').get_text()
```

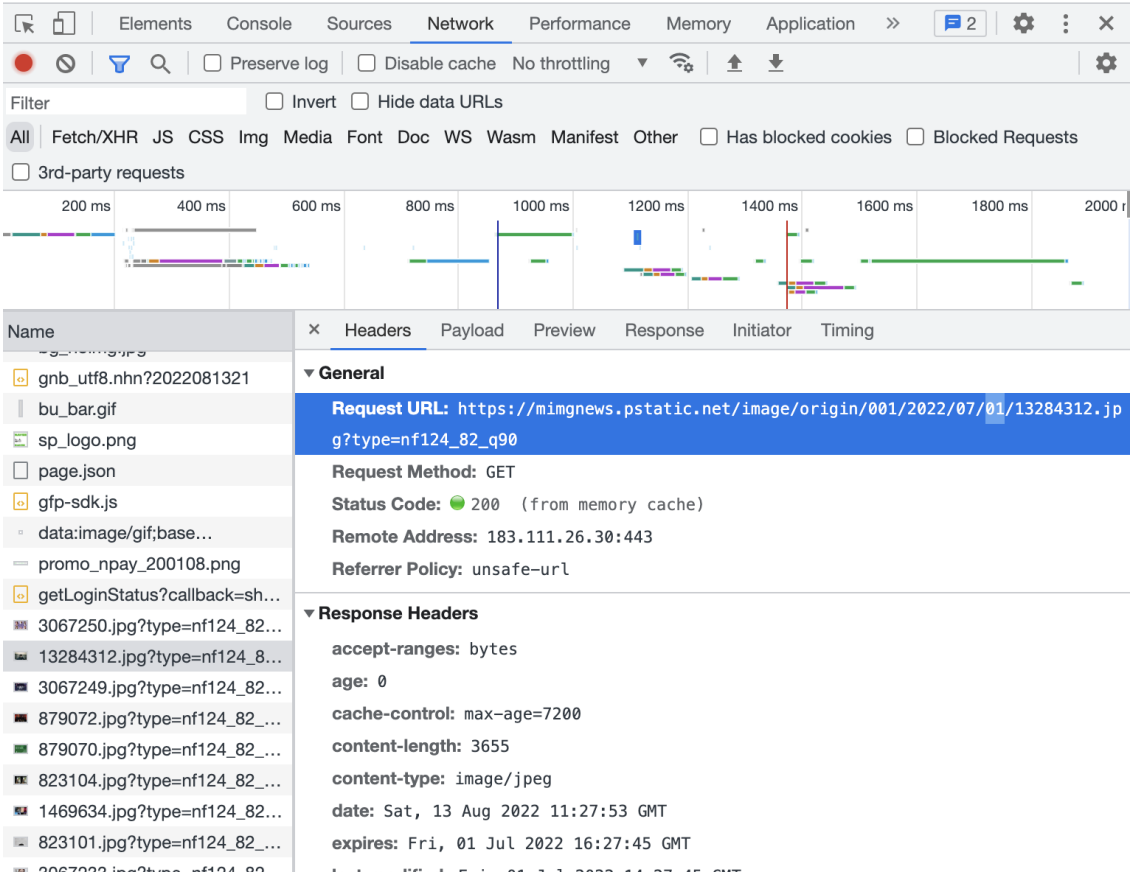
```
    except:
        news_writing_time = ''
    # news_img
    try:
        news_img = soup.select_one('#img1')['src']
    except:
        news_img = ''
    # news_body
    try:
        news_body = soup.select_one('#articeBody').get_text().strip()
    except:
        news_body = ''
    # news_author
    try:
        news_author = soup.select_one('#content > div.end_ct > div > div.article_journalist > div > div > div > div > div > div.journa
    except:
        news_author = ''

    return news_writing_time, news_img, news_body, news_author


base_url = 'https://entertain.naver.com'
request = requests.get(base_url + '/now?sid=7a5#sid=7a5&date=2022-07-01&page=1')
soup = BeautifulSoup(request.text, 'html.parser')

news_list = soup.find_all('a', class_ = 'tit')
l = []
index = 0

for news in news_list:
    news_title = news.get_text()
    news_url = base_url + news['href']
    news_writing_time, news_img, news_body, news_author = news_croller(base_url + news['href'])
    l.append([index, news_title, news_url, news_writing_time, news_img, news_body, news_author])
    index+=1
    print(news_title)
    print()


df = pd.DataFrame(l, columns=['index', 'news_title', 'news_url', 'news_writing_time', 'news_img', 'news_body', 'news_author'])
df.to_excel('naver_news.xlsx', index=False)
```

결과

	A	B	C	D	E	F	
1	index	news_title	news_url	news_writing_time	news_img	news_body	new
2	0	쪽우쪽염도 막지 못한 EDM 열기...	https://entertain.naver.com/now/re	2022.08.13. 오후 9:15	https://ssl.pstatic.net/mimgnews/i	월드 디제이 페스티벌 3년 만에 대면 재개...각국 유명 DJ 한자리에 2022 월	이태수 기자
3	1	'음악중심' 코요태, 싱그러운 'GO'	https://entertain.naver.com/now/re	2022.08.13. 오후 8:40	https://mimgnews.pstatic.net/ima	혼성그룹 코요태(김종민, 빽가, 신지)가 울산을 뜨겁게 달궜다.코요태는 13일	김나영 기자
4	2	'쇼! 음악중심' TAN, 조명 필요 없는	https://entertain.naver.com/now/re	2022.08.13. 오후 8:34	https://ssl.pstatic.net/mimgnews/i	그룹 TAN(티에이엔)이 시선 강탈 퍼포먼스를 선보였다.TAN(창선, 주안, 재준	김나영 기자
5	3	원호, 유니버스와 첫 콜라보..19일	https://entertain.naver.com/now/re	2022.08.13. 오후 7:14	https://ssl.pstatic.net/mimgnews/i	가수 원호 /사진제공=하이라인엔터테인먼트가수 원호가 돌아온다.13일 글로	안윤지 기자
6	4	아이브, 日 데뷔 전부터 열도 접수!	https://entertain.naver.com/now/re	2022.08.13. 오후 6:06	https://mimgnews.pstatic.net/ima	아이브(IVE, 안유진, 가을, 레이, 장원영, 리즈, 이서)가 일본 활동의 신호탄을	김혜영 기자
7	5	블랙핑크, 다크 카리스마...반전 매	https://entertain.naver.com/now/re	2022.08.13. 오후 6:04	https://mimgnews.pstatic.net/ima	선공개곡 'Pink Venom' 2차 티저 영상 공개 블랙핑크가 2차 티저 영상으로	정병근 기자
8	6	'음원 강자' 뉴진스, 데뷔곡 'Attenti	https://entertain.naver.com/now/re	2022.08.13. 오후 5:59	https://ssl.pstatic.net/mimgnews/i	뉴진스(NewJeans)의 데뷔곡 'Attention'이 스포티파이 글로벌 차트에 진입	김혜영 기자
9	7	'공주님의 티타임' 장원영, 휴가도	https://entertain.naver.com/now/re	2022.08.13. 오후 5:32	https://ssl.pstatic.net/mimgnews/i	아이브 장원영. 출처 | 장원영SNS[스포츠서울 | 박효실기자] 그룹 아이브 장	박효실 기자
10	8	생각의 여름, 정규 4집 '손' 발매 기	https://entertain.naver.com/now/re	2022.08.13. 오후 4:58	https://mimgnews.pstatic.net/ima	[서울=뉴시스] 생각의 여름. 2022.08.13. (사진 = 생각의 여름 제공) photo@	이재훈 기자
11	9	하성운, 타이틀곡은 'FOCUS'...새로	https://entertain.naver.com/now/re	2022.08.13. 오후 4:58	https://mimgnews.pstatic.net/ima	(엑스포츠뉴스 김노을 기자) 가수 하성운이 타이틀곡 'FOCUS'로 컴백한다.하	김노을 기자
12	10	정동원 열고 NCT드림 닫고, '음중'	https://entertain.naver.com/now/re	2022.08.13. 오후 4:50	https://ssl.pstatic.net/mimgnews/i	[OSEN=연휘선 기자] '쇼! 음악중심'에서 K팝 아티스트들이 울산을 뜨겁게 달	연휘선 기자
13	11	위아이 'Youth', 日 오리콘 일간 앨	https://entertain.naver.com/now/re	2022.08.13. 오후 4:49	https://ssl.pstatic.net/mimgnews/i	위아이(WEi)가 일본 데뷔 앨범으로 오리콘 차트 정상을 차지했다. 지난 12일	김혜영 기자
14	12	블랙핑크, 확 달라졌네...정규 2집 선	https://entertain.naver.com/now/re	2022.08.13. 오후 4:44	https://ssl.pstatic.net/mimgnews/i	[스포츠서울 | 심언경기자] 1년 10개월 만에 돌아오는 그룹 블랙핑크의 컴백	심언경 기자
15	13	블랙핑크, 감탄 부르는 美친 비주얼	https://entertain.naver.com/now/re	2022.08.13. 오후 4:44	https://mimgnews.pstatic.net/ima	블랙핑크의 변화무쌍한 비주얼 콘셉트와 흡인력 강한 음악적 서사를 기대케	김혜영 기자
16	14	코요태, '음악중심' 사로잡았다	https://entertain.naver.com/now/re	2022.08.13. 오후 4:35	https://mimgnews.pstatic.net/ima	MBC '쇼! 음악중심' 캡처국내 최장수 혼성그룹 코요태(김종민, 빽가, 신지)가	이예주 기자
17	15	'불후의 명곡' 베철수구창모 출연-	https://entertain.naver.com/now/re	2022.08.13. 오후 4:33	https://ssl.pstatic.net/mimgnews/i	[일간스포츠 김다은] 사진=KBS 제공 '불후의 명곡'이 송골매 편으로 다	김다은 기자
18	16	엑스지, 댄서 아이키와 역대급 댄스	https://entertain.naver.com/now/re	2022.08.13. 오후 4:12	https://ssl.pstatic.net/mimgnews/i	(엑스포츠뉴스 황수연 기자) 그룹 XG(엑스지)가 댄서 아이키와 역대급 컬래	황수연 기자
19	17	피원하모니 '둥두둥' 활동 마무리,	https://entertain.naver.com/now/re	2022.08.13. 오후 4:11	https://ssl.pstatic.net/mimgnews/i	[일간스포츠 김다은] 사진=FNC엔터테인먼트 제공 그룹 피원하모니가	김다은 기자
20	18	T1419, 조수미와 환상 듀엣 무대...3	https://entertain.naver.com/now/re	2022.08.13. 오후 4:07	https://ssl.pstatic.net/mimgnews/i	(엑스포츠뉴스 김노을 기자) 그룹 T1419(티일사일구)가 조수미와 환상의 듀	김노을 기자
21	19	뉴진스 'Attention' 스포티파이 글로	https://entertain.naver.com/now/re	2022.08.13. 오후 3:32	https://ssl.pstatic.net/mimgnews/i	/사진제공=어도어뉴진스(NewJeans, 민지 하니 다니엘 해린 혜린) 데뷔곡 'A	윤상근 기자
22	20	위아이 日 데뷔앨범 '유스', 오리콘	https://entertain.naver.com/now/re	2022.08.13. 오후 3:21	https://mimgnews.pstatic.net/ima	그룹 '위아이'. 뉴시스그룹 '위아이(WEi)'가 일본 데뷔 앨범으로 현지 오리콘	차트 정상에 올
23	21	뉴진스 '어텐션', 스포티파이 글로벌	https://entertain.naver.com/now/re	2022.08.13. 오후 3:04	https://ssl.pstatic.net/mimgnews/i	[서울=뉴시스] 뉴진스. 2022.07.25. (사진 = 어도어 제공) photo@newsis.cor	이재훈 기자
24	22	임영웅 '아버지' 뮤비 100만뷰 넘어	https://entertain.naver.com/now/re	2022.08.13. 오후 2:53	https://mimgnews.pstatic.net/ima	임영웅 '아버지' 뮤비 100만뷰 넘어섰다 '감동의 히어로'가수 임영웅의 '아버	윤상근 기자
25	23	임영웅, 7월 써클차트 가수별 점유	https://entertain.naver.com/now/re	2022.08.13. 오후 2:51	https://mimgnews.pstatic.net/ima	(엑스포츠뉴스 이정범 기자) 임영웅이 7월 써클차트 가수별 점유율 TOP400	이정범 기자
26	24	위아이, 日 데뷔앨범으로 오리콘 일	https://entertain.naver.com/now/re	2022.08.13. 오후 2:48	https://mimgnews.pstatic.net/ima	[서울=뉴시스] 위아이. 2022.08.13. (사진 = 위엔터테인먼트 제공) photo@n	이재훈 기자
27							
28							
29							

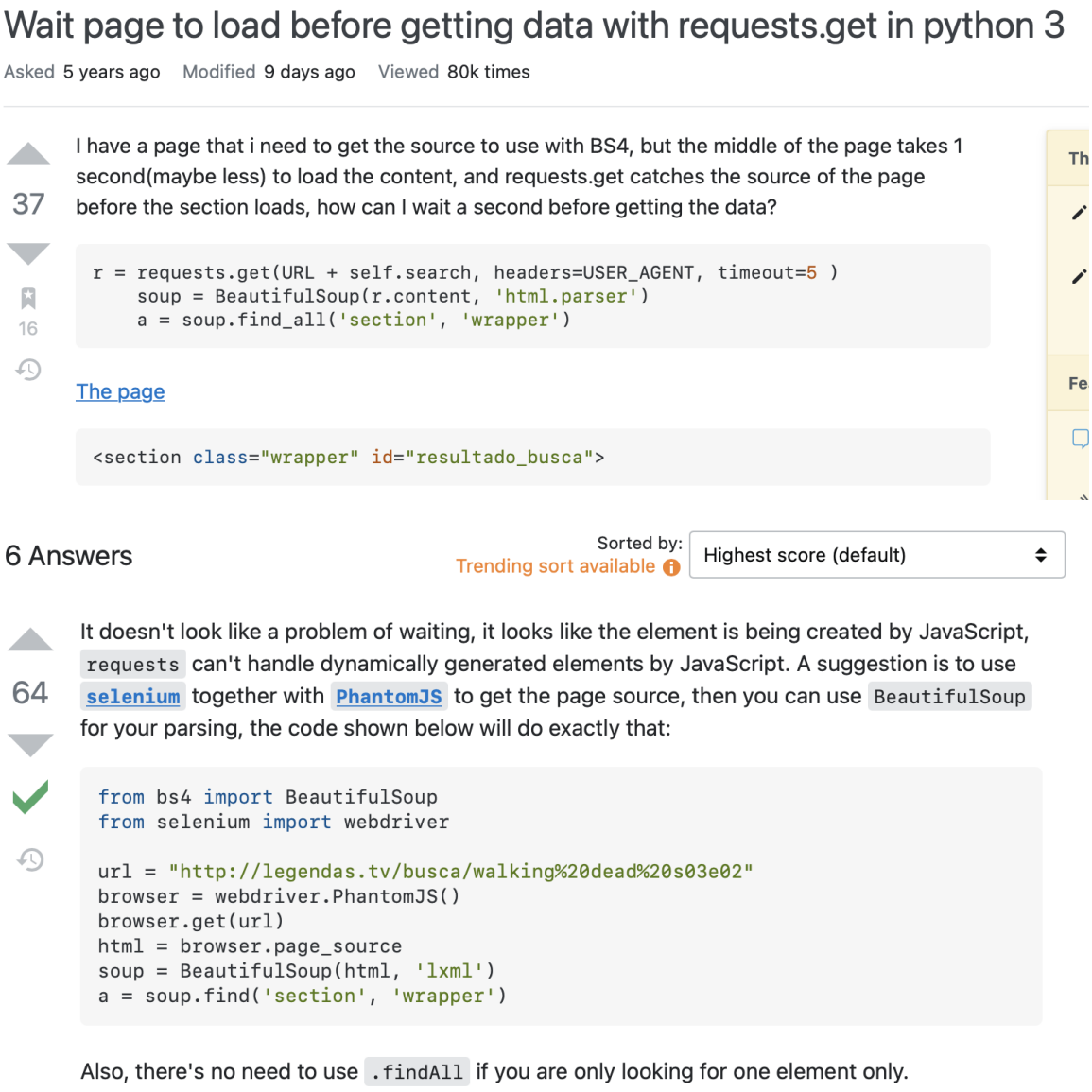
- news 별로 기자 이름이나 이미지 등이 없는 경우가 있어서 크롤링중 오류가 발생해 끊어지는 경우가 생겼다
  - → try-except 를 이용해서 항목이 없는경우 빈 문자열을 저장해줬다.
- 문제점 : url에 파라미터로 7월1일의 날짜를 지정해줬는데도 오늘의 뉴스를 크롤링 해왔다.
  - → 페이지를 로딩하는 방식이 우선 오늘 날짜의 뉴스를 먼저 보여주고 그 뒤에 7월 1일의 뉴스를 보여주는 방식인것 같다.





이렇게 0713의 사진이 0701의 사진보다 먼저 로드된다. 우리의 request 함수는 0701의 사진이 로드 되기전에 크롤링을 해오는 것으로 생각된다.

- 해결책 : requests함수는 로딩시간을 기다리는 옵션이 없어보인다. 웹서핑 결과 async 방식이나 셀레니움을 사용하는 방법을 사용해야 할듯 하다. (https://stackoverflow.com/questions/45448994/wait-page-to-load-before-getting-data-with-requests-get-in-python-3)



Wait page to load before getting data with requests.get in python 3

Asked 5 years ago Modified 9 days ago Viewed 80k times

37

I have a page that i need to get the source to use with BS4, but the middle of the page takes 1 second(maybe less) to load the content, and requests.get catches the source of the page before the section loads, how can I wait a second before getting the data?

```
r = requests.get(URL + self.search, headers=USER_AGENT, timeout=5 )
    soup = BeautifulSoup(r.content, 'html.parser')
    a = soup.find_all('section', 'wrapper')
```

16

The page

```
<section class="wrapper" id="resultado_busca">
```

6 Answers Sorted by: Highest score (default) Trending sort available

64

It doesn't look like a problem of waiting, it looks like the element is being created by JavaScript, `requests` can't handle dynamically generated elements by JavaScript. A suggestion is to use `selenium` together with `PhantomJS` to get the page source, then you can use `BeautifulSoup` for your parsing, the code shown below will do exactly that:

```
from bs4 import BeautifulSoup
from selenium import webdriver

url = "http://legendas.tv/busca/walking%20dead%20s03e02"
browser = webdriver.PhantomJS()
browser.get(url)
html = browser.page_source
soup = BeautifulSoup(html, 'lxml')
a = soup.find('section', 'wrapper')
```

Also, there's no need to use `.findAll` if you are only looking for one element only.

- 셀레니움 맥 사용법 정리 영상 (https://www.youtube.com/watch?v=7R5n0sNSza8)

최종 코드

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
from selenium import webdriver


def news_croller(url): # 페이지별 크롤링 코드
    request = requests.get(url)
    soup = BeautifulSoup(request.text, 'html.parser')
```

```
    # news_writing_time
    try:
        news_writing_time = soup.select_one('#content > div.end_ct > div > div.article_info > span:nth-child(1) > em').get_text()
    except:
        news_writing_time = ''
    # news_img
    try:
        news_img = soup.select_one('#img1')['src']
    except:
        news_img = ''
    # news_body
    try:
        news_body = soup.select_one('#articeBody').get_text().strip()
    except:
        news_body = ''
    # news_author
    try:
        news_author = soup.select_one('#content > div.end_ct > div > div.article_journalist > div > div > div > div > div > div.journa
    except:
        news_author = ''

    return news_writing_time, news_img, news_body, news_author


base_url = 'https://entertain.naver.com'
browser = webdriver.Chrome('./chromedriver')
browser.get(base_url + '/now?sid=7a5#sid=7a5&date=2022-07-01&page=1')
html = browser.page_source
soup = BeautifulSoup(html, 'html')

news_list = soup.find_all('a', class_ = 'tit')
l = []
index = 0

for news in news_list:
    news_title = news.get_text()
    news_url = base_url + news['href']
    news_writing_time, news_img, news_body, news_author = news_croller(base_url + news['href'])
    l.append([index, news_title, news_url, news_writing_time, news_img, news_body, news_author])
    index+=1
    print(news_title)
    print()


df = pd.DataFrame(l, columns=['index', 'news_title', 'news_url', 'news_writing_time', 'news_img', 'news_body', 'news_author'])
df.to_excel('naver_news.xlsx', index=False)
```

- 크롬 드라이버 라는게 필요한지 모르고 한참동안 헤맸다.
- 셀레니움을 처음써봤는데 실제로 브라우저를 열어 크롤링을 해온다. 따라서 앞의 문제를 해결 할 수 있었다.

결과

	A	B	C	D	E	F	G
1	index	news_title	news_url	news_writing_time	news_img	news_body	news_author
2	0	로켓펀치, 日 첫 싱글 'Fiore' 타워레코드	https://entertain.naver.com/now/read?	2022.07.01. 오후 11:27	https://ssl.pstatic.net/mimgnews/imag	/사진제공=올림엔터테인먼트그룹 로켓펀	김수진 기자
3	1	'BTS 솔로 첫 주자' 제이홉의 일성…"오	https://entertain.naver.com/now/read?	2022.07.01. 오후 11:27	https://ssl.pstatic.net/mimgnews/imag	강렬한 샤우팅 돋보이는 '모어' 발표…"	이태수 기자
4	2	'뉴페스타2022' 페스티벌 9월 개최 [공	https://entertain.naver.com/now/read?	2022.07.01. 오후 11:24	https://mimgnews.pstatic.net/image/1	'뉴페스타2022' 페스티벌 9월 개최 [공	김수진 기자
5	3	에버글로우, '한류 팝 페스트 런던 2022	https://entertain.naver.com/now/read?	2022.07.01. 오후 10:59	https://mimgnews.pstatic.net/image/4	그룹 에버글로우(EVERGLOW)가 '한류	진주희 기자
6	4	디원스 우진영, 미니 2집 'DELICIOUS'로	https://entertain.naver.com/now/read?	2022.07.01. 오후 10:49	https://ssl.pstatic.net/mimgnews/imag	디원스(D1CE) 멤버 우진영이 미니 2집	진주희 기자
7	5	'이별노래 장인' 이우, 싱어송라이터 최	https://entertain.naver.com/now/read?	2022.07.01. 오후 10:21	https://ssl.pstatic.net/mimgnews/imag	이우 SNS가수 이우가 동료 최낙타와의	손봉석 기자
8	6	태연→아이브, 2022년 상반기 인상적이	https://entertain.naver.com/now/read?	2022.07.01. 오후 10:12	https://ssl.pstatic.net/mimgnews/imag	(엑스포츠뉴스 이정범 기자) 2022년 상	이정범 기자
9	7	백찬, NFT IP 컬래버 프로듀싱 나선다	https://entertain.naver.com/now/read?	2022.07.01. 오후 10:04	https://ssl.pstatic.net/mimgnews/imag	디스이즈잇 레코즈 제공그룹 에이트(8e	안병길 기자
10	8	송가인, 안경 쓰고 깜짝 놀랄 귀여움.."E	https://entertain.naver.com/now/read?	2022.07.01. 오후 9:39	https://mimgnews.pstatic.net/image/1	/사진=송가인 인스타그램가수 송가인이	윤성열 기자
11	9	'이수♥' 린, 몸만 한 가방 들고 룰루랄	https://entertain.naver.com/now/read?	2022.07.01. 오후 9:23	https://mimgnews.pstatic.net/image/1	/사진=린 인스타그램가수 린이 몸만 한	이빛나리 기자
12	10	'뉴페스타2022' 페스티벌, 오는 7월 6일	https://entertain.naver.com/now/read?	2022.07.01. 오후 9:15	https://ssl.pstatic.net/mimgnews/imag	방송의 감동을 무대로 옮겨온 '뉴페스트	손봉석 기자
13	11	'공연기획사와 갈등' 김희재 소속사 "콘	https://entertain.naver.com/now/read?	2022.07.01. 오후 8:25	https://ssl.pstatic.net/mimgnews/imag	/사진제공=쇼플레이 /사진=임성균 기자	윤성열 기자
14	12	'유뱅' 버가부, 톡톡 튀는 여섯 소녀..달	https://entertain.naver.com/now/read?	2022.07.01. 오후 7:49	https://mimgnews.pstatic.net/image/1	/사진='뮤직뱅크' 방송 화면 걸 그룹 B	윤성열 기자
15	13	'유뱅' 라필루스, 틴크러시 폭발.'HIT YA	https://entertain.naver.com/now/read?	2022.07.01. 오후 7:42	https://mimgnews.pstatic.net/image/1	/사진='뮤직뱅크' 방송 화면신인 걸 그	윤성열 기자
16	14	진시몬, 새 소속사 루체엔터...설운도·조	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:52	https://mimgnews.pstatic.net/image/1	보약 같은 가수 진시몬이 새로운 소속사	강석봉 기자
17	15	'양현석 동생' 양민석, 3년 만에 YG 대표	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:27	https://ssl.pstatic.net/mimgnews/imag	블랙핑크 등 소속 대형 아티스트 컴백	이태수 기자
18	16	"현아♥던→AKMU·다듀 출격"..'슈퍼노바	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:27	https://ssl.pstatic.net/mimgnews/imag	[OSEN=이승훈 기자] '슈퍼노바 페스티	이승훈 기자
19	17	아이브, 4세대 최초 음원차트 월간 1위.	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:23	https://mimgnews.pstatic.net/image/1	걸그룹 아이브(IVE)가 5일 오후 서울 용	공미나 기자
20	18	'뮤직뱅크' 버가부, 독보적 걸크러쉬	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:13	https://ssl.pstatic.net/mimgnews/imag	버가부(bugAboo)가 완벽한 무대를 꾸	김나영 기자
21	19	방탄소년단, 美 케이팝 차트쇼 '케이팝	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:09	https://ssl.pstatic.net/mimgnews/imag	[마이데일리 = 박윤진 기자] 그룹 방탄	박윤진 기자
22	20	에스파, UN 간다..'지속가능발전 고위급	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:08	https://mimgnews.pstatic.net/image/0	그룹 에스파 /사진=SM엔터테인먼트 제	김수영 기자
23	21	DKZ 경윤, 숨겨뒀던 예능감 방출..김응	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:06	https://ssl.pstatic.net/mimgnews/imag	그룹 DKZ(디케이지) 멤버 경윤이 숨겨	김나영 기자
24	22	알렉사, 부러진 '아메리칸 송 콘테스트'	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:02	https://mimgnews.pstatic.net/image/0	기사내용 요약주한미국대사관이 전달드	이재훈 기자
25	23	그리웠던 이승기, 또 일냈다..'잊지마, 7	https://entertain.naver.com/now/read?	2022.07.01. 오후 6:01	https://mimgnews.pstatic.net/image/3	(엑스포츠뉴스 김예나 기자) 가수 이승	김예나 기자
26	24	적재X이진아, 입 맞췄다...초여름 감성	https://entertain.naver.com/now/read?	2022.07.01. 오후 5:58	https://mimgnews.pstatic.net/image/0	오늘(1일) 무신사 프로젝트 음원 공개	김수영 기자
27							
28							

- 7월 1일의 뉴스를 잘 가져온 모습이다.

결론

- requests 함수와 BeautifulSoup를 이용해 정적 웹페이지를 파싱하는법을 배웠다.
- 동적 웹페이지를 크롤링할때는 다른 패키지를 써야 하는 경우도 있는것 같다.
- 데이터가 모두 정형화 된게 아니므로 앞으로 try-except문을 사용해야 할 경우가 많을 것 같다.