포털사이트 크롤링

- 소스내에서 특정 문자열(data)을 지칭하는 선택자 얻기
  - 크롬 개발자도구 사용
- 전체 코드에서 수집하려고 하는 데이터(태그)의 위치를 찾고
  - 태그를 파싱한 후 필요데이터 추출

[1]:

```
from urllib.request import urlopen # 서버 요청/응답 패키지
import bs4 # 파싱패키지
```

[2]:

```
# 네이버 사이트의 기본 메뉴 문구 추출
url = 'https://www.naver.com'

# url로 요청후 응답
html = urlopen(url)

# 파서객체 생성 - bs4 객체로 변환
bs_obj = bs4.BeautifulSoup(html, 'html.parser')
```

[3]:

```
bs_obj
```

요" property="og:description"> <meta content="summary" name="t
witter:card"/> <meta content="" name="twitter:title"/> <meta c
ontent="https://www.naver.com/" name="twitter:url"/> <meta con
tent="https://s.pstatic.net/static/www/mobile/edit/2016/0705/m
obile_212852414260.png" name="twitter:image"/> <meta content
="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요" n
ame="twitter:description"> <link href="https://pm.pstatic.net/
dist/css/nmain.20221110.css" rel="stylesheet"/> <link href="ht
tps://ssl.pstatic.net/sstatic/search/pc/css/sp_autocomplete_22
0526.css" rel="stylesheet"/> <link href="/favicon.ico?1" rel
="shortcut icon" type="image/x-icon"> <link href="https://s.ps
tatic.net/static/www/u/2014/0328/mma_204243574.png" rel="apple
-touch-icon" sizes="114x114"> <link href="https://s.pstatic.ne
t/static/www/u/2014/0328/mma_20432863.png" rel="apple-touch-ic
on"/> <script>window.nmain=window.nmain||{},window.nmain.suppo
rtFlicking=!1;var nsc="navertop.v4",ua=navigator.userAgent,use
leJSFlag="1";window.nmain.isIE="0"===useleJSFlag,document.getE
lementsByTagName("html")[0].setAttribute("data-useragent",ua),
window.nmain.isIE&&(Object.create=function(n){function e(){}re

[4]:

```
1 test_menu = bs_obj.find('div',{'id':'gnb'})
2 test_menu
```

```
<div id="gnb" role="navigation">
<div class="gnb_inner" id="NM_FAVORITE">
<div class="group_nav">
<ul class="list_nav type_fix">
<li class="nav_item">
<a class="nav" data-clk="svc.mail" href="https://mail.naver.co
m/"><i class="ico_mail"></i>메일</a>
</li>
<li class="nav_item"><a class="nav" data-clk="svc.cafe" href="ht
tps://section.cafe.naver.com/">카페</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.blog" href="ht
tps://section.blog.naver.com/">블로그</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.kin" href="htt
ps://kin.naver.com/">지식iN</a></li>
<li class="nav_item"><a class="nav shop" data-clk="svc.shopping"
href="https://shopping.naver.com/"><span class="blind">쇼핑</spa
n></a></li>
<li class="nav_item"><a class="nav shoplive" data-clk="svc.shopp
inglive" href="https://shoppinglive.naver.com/home"><span class
="blind">쇼핑LIVE</span></a></li>
<li class="nav_item"><a class="nav" data-clk="svc.pay" href="htt
ps://order.pay.naver.com/home">Pay</a></li>
<li class="nav_item">
<a class="nav" data-clk="svc.tvcast" href="https://tv.naver.co
m/"><i class="ico_tv"></i>TV</a>
</li>
</ul>
<ul class="list_nav NM_FAVORITE_LIST">
<li class="nav_item"><a class="nav" data-clk="svc.dic" href="htt
ps://dict.naver.com/">사전</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.news" href="ht
tps://news.naver.com/">뉴스</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.stock" href="h
ttps://finance.naver.com/">증권</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.land" href="ht
tps://land.naver.com/">부동산</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.map" href="htt
ps://map.naver.com/">지도</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.vibe" href="ht
tps://vibe.naver.com/?from=naver_main&amp;utm_source=naver_main&
amp;utm_medium=naver_main_pcweb&amp;utm_campaign=naver_main_redi
rect">VIBE</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.book" href="ht
tps://search.shopping.naver.com/book/home">도서</a></li>
<li class="nav_item"><a class="nav" data-clk="svc.webtoon" href
="https://comic.naver.com/">웹툰</a></li>
</ul>
<ul class="list_nav type_empty" style="display: none;"></ul>
<a class="btn_more" data-clk="svc.more" href="#" role="button">
더보기</a>
<div class="ly_btn_area">
<a class="btn NM_FAVORITE_ALL" data-clk="map.svcmore" href="mor
e.html">서비스 전체보기</a>
```

[15]:
```
#NM_FAVORITE > div.group_nav > ul.list_nav.type_fix
ul = bs_obj.find("ul",{'class':'list_nav type_fix'})
ul
```

```
<a class="btn btn_set" data-clk="map.edit" href="#" role="butto
n">메뉴설정</a>
<a class="btn btn_reset" data-clk="edt.reset" href="#" role="but
ton">초기화</a>
<a class="btn btn_save" data-clk="edt.save" href="#" role="butto
n">저장</a>
</div>
<ul class="list_nav type_fix">
</div>
<li class="nav_item">
<div class="group_weather" id="NM_WEATHER">
<a class="nav" data-clk="svc.mail" href="https://mail.naver.co
<div>
m/" ><i class="ico_mail"></i>메일</a>
<a class="weather_area ico_w01" data-clk="squ.weat" href="http
</li>
s://weather.naver.com/today/09680101?cpName=KMA">
<li class="nav_item" ><a class="nav" data-clk="svc.cafe" href="ht
<div class="current_box">
tps://section.cafe.naver.com/">카페</a></li>
<strong aria-label="현재기온" class="current">-1.8° </strong><st
<li class="nav_item" ><a class="nav" data-clk="svc.blog" href="ht
rong class="state">맑음</strong>
tps://section.blog.naver.com/" >블로그</a></li>
</div>
<li class="nav_item"><a class="nav" data-clk="svc.kin" href="htt
<div class="degree_box">
ps://kin.naver.com/" >지식iN</a></li>
<span aria-label="최저기온" class="min">-8.0° </span><span aria-
<li class="nav_item" ><a class="nav shop" data-clk="svc.shopping"
label="최고기온" class="max">-1.0° </span>
href="https://shopping.naver.com/"><span class="blind">쇼핑</spa
</div>
n></a></li>
<span class="location">역삼동</span>
<li class="nav_item"><a class="nav shoplive" data-clk="svc.shopp
</a>
inglive" href="https://shoppinglive.naver.com/home"><span class
</div>
="blind">쇼핑LIVE</span></a></li>
<div>
<li class="nav_item"><a class="nav" data-clk="svc.pay" href="htt
<a class="air_area" data-clk="squ.dust" href="https://weather.na
ps://order.pay.naver.com/home">Pay</a></li>
ver.com/today/09680101?cpName=KMA">
<li class="nav_item">
<ul class="list_air">
<a class="nav" data-clk="svc.tvcast" href="https://tv.naver.co
<li class="air_item">미세<strong class="state state_good">좋음</
m/" ><i class="ico_tv"></i>TV</a>
strong></li>
</li>
<li class="air_item">초미세<strong class="state state_good">좋음
</ul>
</strong></li>
</ul>
<span class="location">역삼동</span>
</a>
</div>
</div>
</div>
<div class="ly_service">
<div class="group_service NM_FAVORITE_ALL_LY"></div>
<div class="group_service NM_FAVORITE_EDIT_LY" style="display: n
one;"></div>
</div>
</div>
```

[6]:

```
ul_test = bs_obj.select('#NM_FAVORITE > div.group_nav > ul.list_nav.ty
ul_test # list 형태로 반환
```

```
[<ul class="list_nav type_fix">
 <li class="nav_item">
 <a class="nav" data-clk="svc.mail" href="https://mail.naver.co
m/"><i class="ico_mail"></i>메일</a>
 </li>
 <li class="nav_item"><a class="nav" data-clk="svc.cafe" href="h
ttps://section.cafe.naver.com/">카페</a></li>
 <li class="nav_item"><a class="nav" data-clk="svc.blog" href="h
ttps://section.blog.naver.com/">블로그</a></li>
 <li class="nav_item"><a class="nav" data-clk="svc.kin" href="ht
tps://kin.naver.com/">지식iN</a></li>
 <li class="nav_item"><a class="nav shop" data-clk="svc.shoppin
g" href="https://shopping.naver.com/"><span class="blind">쇼핑</
span></a></li>
 <li class="nav_item"><a class="nav shoplive" data-clk="svc.shop
pinglive" href="https://shoppinglive.naver.com/home"><span class
="blind">쇼핑LIVE</span></a></li>
 <li class="nav_item"><a class="nav" data-clk="svc.pay" href="ht
tps://order.pay.naver.com/home">Pay</a></li>
 <li class="nav_item">
 <a class="nav" data-clk="svc.tvcast" href="https://tv.naver.co
m/"><i class="ico_tv"></i>TV</a>
 </li>
 </ul>]
```

[7]:

```
type(ul) # bs4.element.Tag
# 문자열이 아님
# 반복순환에 사용 가능
for u in ul :
    print(u)
```

```
<li class="nav_item">
<a class="nav" data-clk="svc.mail" href="https://mail.naver.co
m/"><i class="ico_mail"></i>메일</a>
</li>


<li class="nav_item"><a class="nav" data-clk="svc.cafe" href="ht
tps://section.cafe.naver.com/">카페</a></li>


<li class="nav_item"><a class="nav" data-clk="svc.blog" href="ht
tps://section.blog.naver.com/">블로그</a></li>


<li class="nav_item"><a class="nav" data-clk="svc.kin" href="htt
ps://kin.naver.com/">지식iN</a></li>


<li class="nav_item"><a class="nav shop" data-clk="svc.shopping"
href="https://shopping.naver.com/"><span class="blind">쇼핑</spa
n></a></li>


<li class="nav_item"><a class="nav shoplive" data-clk="svc.shopp
inglive" href="https://shoppinglive.naver.com/home"><span class
="blind">쇼핑LIVE</span></a></li>


<li class="nav_item"><a class="nav" data-clk="svc.pay" href="htt
ps://order.pay.naver.com/home">Pay</a></li>


<li class="nav_item">
<a class="nav" data-clk="svc.tvcast" href="https://tv.naver.co
m/"><i class="ico_tv"></i>TV</a>
</li>
```

[8]:

```
lis = ul.findAll('li') # 리스트 반환
for li in lis :
    print(li.find('a')['href'])
    print(li.text)
```

```
https://mail.naver.com/ (https://mail.naver.com/)

메일

https://section.cafe.naver.com/ (https://section.cafe.naver.co
m/)
카페
https://section.blog.naver.com/ (https://section.blog.naver.co
m/)
블로그
https://kin.naver.com/ (https://kin.naver.com/)
지식iN
https://shopping.naver.com/ (https://shopping.naver.com/)
쇼핑
https://shoppinglive.naver.com/home (https://shoppinglive.naver.
com/home)
쇼핑LIVE
https://order.pay.naver.com/home (https://order.pay.naver.com/ho
me)
Pay
https://tv.naver.com/ (https://tv.naver.com/)

TV
```

네이버 전체 메뉴 크롤링

[9]:

```
# 네이버 기본 메뉴는 ul > li.nav_item 형식으로 되어 있음
lis = bs_obj.findAll('li',{'class':'nav_item'})
len(lis) # class 속성값이 nav_item인 li 태그 추출 : 16개의 li 태그가 추
```

```
16
```

[10]:

```
for li in lis :
    a_tag = li.find('a')
    print(a_tag.text,":", a_tag['href'])
```

```
메일 : https://mail.naver.com/ (https://mail.naver.com/)
카페 : https://section.cafe.naver.com/ (https://section.cafe.nav
er.com/)
블로그 : https://section.blog.naver.com/ (https://section.blog.n
aver.com/)
지식iN : https://kin.naver.com/ (https://kin.naver.com/)
쇼핑 : https://shopping.naver.com/ (https://shopping.naver.com/)
쇼핑LIVE : https://shoppinglive.naver.com/home (https://shopping
live.naver.com/home)
Pay : https://order.pay.naver.com/home (https://order.pay.naver.
com/home)
TV : https://tv.naver.com/ (https://tv.naver.com/)
사전 : https://dict.naver.com/ (https://dict.naver.com/)
뉴스 : https://news.naver.com/ (https://news.naver.com/)
증권 : https://finance.naver.com/ (https://finance.naver.com/)
부동산 : https://land.naver.com/ (https://land.naver.com/)
지도 : https://map.naver.com/ (https://map.naver.com/)
VIBE : https://vibe.naver.com/?from=naver_main&utm_source=naver_
main&utm_medium=naver_main_pcweb&utm_campaign=naver_main_redirec
t (https://vibe.naver.com/?from=naver_main&utm_source=naver_main
&utm_medium=naver_main_pcweb&utm_campaign=naver_main_redirect)
도서 : https://search.shopping.naver.com/book/home (https://sear
ch.shopping.naver.com/book/home)
웹툰 : https://comic.naver.com/ (https://comic.naver.com/)
```

수집한 데이터를 csv로 저장

1. 항목별로 list에 저장
2. 항목들을 dict로 구성
3. dict를 데이터프레임으로 생성
4. 데이터프레임을 csv로 저장

[11]:

```
# 네이버 메뉴 정보 저장
# 빈 리스트 생성
menu = []
url = []
```

[12]:

```
for li in lis :
    a_tag = li.find('a')
    menu.append(a_tag.text)
    url.append(a_tag['href'])
```

[13]:

```
menu
```

```
['메일',
 '카페',
 '블로그',
 '지식iN',
 '쇼핑',
 '쇼핑LIVE',
 'Pay',
 'TV',
 '사전',
 '뉴스',
 '증권',
 '부동산',
 '지도',
 'VIBE',
 '도서',
 '웹툰']
```

[14]:

```

url
```

```
['https://mail.naver.com/',
 'https://section.cafe.naver.com/',
 'https://section.blog.naver.com/',
 'https://kin.naver.com/',
 'https://shopping.naver.com/',
 'https://shoppinglive.naver.com/home',
 'https://order.pay.naver.com/home',
 'https://tv.naver.com/',
 'https://dict.naver.com/',
 'https://news.naver.com/',
 'https://finance.naver.com/',
 'https://land.naver.com/',
 'https://map.naver.com/',
 'https://vibe.naver.com/?from=naver_main&utm_source=naver_main&
utm_medium=naver_main_pcweb&utm_campaign=naver_main_redirect',
 'https://search.shopping.naver.com/book/home',
 'https://comic.naver.com/']
```

[15]:

```
import pandas as pd
df = pd.DataFrame({'메뉴':menu, "URL":url})
df
```

	메뉴	URL
0	메일	https://mail.naver.com/
1	카페	https://section.cafe.naver.com/
2	블로그	https://section.blog.naver.com/
3	지식iN	https://kin.naver.com/
4	쇼핑	https://shopping.naver.com/
5	쇼핑LIVE	https://shoppinglive.naver.com/home
6	Pay	https://order.pay.naver.com/home
7	TV	https://tv.naver.com/
8	사전	https://dict.naver.com/
9	뉴스	https://news.naver.com/
10	증권	https://finance.naver.com/
11	부동산	https://land.naver.com/
12	지도	https://map.naver.com/
13	VIBE	https://vibe.naver.com/?from=naver_main&utm_so...
14	도서	https://search.shopping.naver.com/book/home
15	웹툰	https://comic.naver.com/

[16]:

```
#df.to_csv('./naver_menu.csv', encoding='euc-kr')
df.to_csv('c:\\Myexam\\naver_menu.csv', encoding='euc-kr')
```

네이버 뉴스 크롤링

네이버 뉴스는 네이버 정책에 따라 모든 언론사들의 뉴스가 랜덤하게 배치됨

- 단, 로그인 후 구독을 추가하면 구독한 언론사들의 뉴스가 나옴
- 헤드라인 뉴스는 표면적으로는 제공되지 않는다

[17]:

```
# 네이버 뉴스 크롤링
# url은 기본 url 부터 사용
url = 'https://news.naver.com'
html = urlopen(url)
```

[18]:

```
# html #<http.client.HTTPResponse at 0x1b6fa0cd0d0>
# html_text = html.read() # 바이너리 문자열로 반환
# # html_text
```

[19]:

```
# bs4 객체 생성
bs_obj = bs4.BeautifulSoup(html, 'html.parser')
```

[20]:

```
print(bs_obj.prettify())
```

```
<!DOCTYPE html>
<html lang="ko">
 <head>
  <title id="browserTitleArea">
   네이버 뉴스
  </title>
  <script>
   function isMobileDevice() {
                return /^.*(iPhone|iPod|iPad|Android).*/.test
(navigator.userAgent);
        }
  </script>
  <script>
   (function () {
                try {
                        if (isMobileDevice() && isAbleApplyPre
fersColorScheme()) {

                                document.querySelector("htm
l").classList.add("DARK_THEME");
```

[21]:

```
news_title = bs_obj.findAll('div',{'class':'cjs_t'})
```

[22]:

```
len(news_title)
```

```
82
```

[23]:

```
1 news_title
```

```
[<div class="cjs_t">캠핑 인구 700만 시대, 커져 가는 RV 시장</d
iv>,
 <div class="cjs_t">금융위기 때보다 '꽁꽁'…최악의 소비한파 온
다</div>,
 <div class="cjs_t">'한국인 2명  탑승' 네팔 여객기 추락, 수색
작업 재개</div>,
 <div class="cjs_t">[포착] 샤일라 쓴 김건희 여사…파병부대 방
문땐 '군복' </div>,
 <div class="cjs_t">손준성 재판에서 드러난 의혹, 검찰의 '김웅
봐주기' 꼬리 잡히나</div>,
 <div class="cjs_t"> '2시간 지각' 마이클 볼턴…관객들 여전한
분노 · 평점 2점대</div>,
 <div class="cjs_t">"알바생 비위 맞춰 돈XX?" 백화점 박살내고
드러누워</div>,
 <div class="cjs_t"> "전두환 시대에 나 건들면 지하실" 장제원
아들, 랩 가사 논란</div>,
 <div class="cjs_t">러 미사일 공격에 욕실로 피해 기적의 생존…
20대 우크라 여성에 무슨 일이?</div>,
 <div class="cjs_t">서로 얽힌 건설업자 · 집주인 · 부동산이 갖고
```

[24]:

```
for title in news_title :
    print(title.text)
```

캠핑 인구 700만 시대, 커져 가는 RV 시장
금융위기 때보다 '꽁꽁'…최악의 소비한파 온다
'한국인 2명 탑승' 네팔 여객기 추락, 수색 작업 재개
[포착] 샤일라 쓴 김건희 여사…파병부대 방문땐 '군복'
손준성 재판에서 드러난 의혹, 검찰의 '김웅 봐주기' 꼬리 잡히나
'2시간 지각' 마이클 볼턴…관객들 여전한 분노 · 평점 2점대
"알바생 비위 맞춰 돈XX?" 백화점 박살내고 드러누워
"전두환 시대에 나 건들면 지하실" 장제원 아들, 랩 가사 논란
러 미사일 공격에 욕실로 피해 기적의 생존…20대 우크라 여성에 무슨 일이?
서로 얽힌 건설업자 · 집주인 · 부동산이 갖고 놀았다
김성태 "이재명 때문에 인생 초토화"...귀국 후 작심 폭로하나
"유럽 시장 공략한다"...벨기에로 날아간 쌍용차
롯데케미칼 파키스탄 법인 1923억에 매각...PTA 사업 철수
마피 · 무피 속출하는 강원도 생활형숙박시설… "공급과잉 · 금리인상 탓"
"택배기사 승강기 사용료 내라" 황당 갑질에 뭇매 맞은 세종시 아파트
장제원 아들 노엘 "전두환 시대··나 건드리면 바로 지하실"
팬데믹 속에서 '사회적 가족' 인정 필요성 커졌다
김건희 여사, 순방길서 든 가방 화제....가격은 얼마?
"월급주기 겁나"… '나홀로 사장님 ' 426만명, 금융위기후 최고
건강하게 오래 사는 사람의 아침 습관 5가지
[단독]스타트업 키우는 '프랑스 재벌집 큰 딸' [스테파니]
野, 尹정부 강제징용 해법 총공세…이재명 "자해적 외교 중단하라"
샤일라 쓴 김건희 여사…파병부대 방문땐 '군복' [포착]
김대중 가둔 수의와 독방…한 사진가가 남긴 민주주의의 역사
서경덕 "온라인 쇼핑몰, '당나라 스타일 한복' 판매 어이없어"
"조울병은 뇌의 문제… 감정 기복 · 우울증과 전혀 달라"
네팔 실종자 한국인 2명은 여행 간 부자지간
[화보] 尹 · 김건희, 아크부대 장병들 만나 '손하트(♥)' 연발
외교부의 윤 대통령 바이든 MBC 소송에 쏟아지는 의문점
장제원 아들 "전두환 시대에 나 건들면 지하실"…가사 논란
격랑의 '재명이네 마을', 터져버린 개딸들…이낙연 · 박지원 · 박영선 '표적' 됐나
<주간 뉴스타파> 대장동 실체 담긴 '정영학 녹취록' 1,325쪽 전문 공개
화성 해병대사령부 PX에서 화재… "1시간여 만에 진화"
나경원, 'UAE 40조 투자 결정' "윤 대통령, 순방 이틀 만에 투자 유치... 가슴이 벅차오른다"
조희연 "초등학교 1~2학년 학폭법 제외"
국내 최초 美 FDA 승인 획득, 한국 제약산업 선도한 종근당
'월드컵 스타' 날개 꺾었다?…박지성, 조규성 이적 거절한 이유
김기현 "다음 대선 출마 안 해…윤 대통령과는 '부부관계'"
7% 할인이 어디냐…서울사랑상품권 3천억 풀린다
"한복이 中전통의복 한푸?"… 국내 온라인 쇼핑몰, 판매 논란
정부, 의대정원 확대 추진…농촌의료 이번엔 개선될까
"아직 소득 있다면 연금 받는 시기 늦춰보세요"
이재용 UAE와 남다른 인연…추가 투자 끌어낼까
[단독] '이태원 참사' 특수본, 한덕수 국무총리도 불송치
원전 확대 필수조건 '사용후핵연료' 처분...기술은 '성숙' 실증은 '아직'
윤 대통령, 김 여사와 UAE 아크부대 방문해 군 장병 격려

```
이번엔 현역의원 기득권 타파할까
유튜브 정치 점화되나...TBS 하차한 김어준, 5일만에 구독자 100만 돌파 [e라이프]
"라방이 눈앞에"…네이버 한가족된 美 '포시마크' 가보니
"한은, 금리인상 사실상 '종료'… 美, 인상 사이클도 마무리 단계"
삼성 vs 애플… "eSIM 폰 35억대 시장 잡아라"
김건희 순방길에 든 가방, 얼마인지 봤더니…'대반전'
네팔 추락기에 한국인들도 탑승…"68명 사망"
"이미 품절됐다" 김건희 UAE순방길 든 베이지색 가방 가격
장제원 아들 노엘, 이번엔 '전두환 시대' 가사 논란 "바로 지하실"
샤일라 쓰고, 사막 위장복 입고..김건희 여사 패션, 눈길 사로잡았다
국내 넘버원 일출 명소 강릉 정동진역 역사 신축 3월 첫 삽 뜬다
"차례상에 꼭 전 올리지 않아도 돼요"…성균관이 알려드립니다
[영상] 비닐봉지 안 준다고 침 뱉고 폭행…차량 몰고 편의점 박살 냈다
최저임금 6.6% 오를때 소비자물가는 7.7% 상승
이틀 새 14도 뚝…당분간 추위 이어져
나경원 "尹 UAE 40조 투자유치 가슴 벅차"…윤심 호소
이재명 "압수수색도 없이 이태원 참사 면죄부…국조 이후 진상규명"
[Herald Review] 'Phantom,' a colorful, classy spy action film set in 1930s
책 추천하다 책방까지 연다…'문재인식 소통법'
동탄 집값 5억~6억씩 '뚝뚝'…신규 분양에 선착순 계약 등장
UAE pledges $30 billion investment in Korea after summit
김여사 "사막여우 많나요" 尹 "별걸 다 알아"…아크부대서 '티키타카'
한국일보 '김만배 돈거래' 간부 해고
정기석 "겨울 코로나 유행 정점 지나…확연히 완화 추세"
은행 희망퇴직은 '돈 잔치'… 노조도 대상 확대 요구
[만남] 거리에서 여성 노숙인 안보이는 이유 아시나요?
'7년 방황' 옛 탐라대 부지 활용.. 이번에는 진짜?
정태영 부회장 며느리 리디아 고, 뉴질랜드 신혼여행서 '홀인원'
모두 발뺀 IPO 사막에 남은 '오아시스'의 앞날
[속보] 정기석 "실내마스크 조정하면 감염확산 가능성…고위험군 백신 접종 당부"
대구·경북 당분간 영하권 아침 추위 이어져…빙판길 주의
폭설·한파에 얼어붙은 도로…44중 연쇄추돌 등 사고 잇따라
아크부대 방문한 윤 대통령 "형제국 안보가 우리 안보"
[거칠부 다이어리] 안나푸르나에서 닭 백숙을 삶다
네팔 추락 여객기 탑승 한국인, 함께 여행 온 아빠와 아들
"신청에 2시간, 수령에 10일".. 여권 창구 '북새통'
```

[25]:

```
1 news_dec = bs_obj.findAll('p',{'class':'cjs_d'})
```

[26]:

```
1 len(news_dec)
```

82

[27]:

```
for dec in news_dec :
    print(dec.text)
```

‘700만 명.’ 2022년 한국의 캠핑 인구수다. 코로나19 사태 이후 200만 명 정도 늘었다. 캠핑 문화가 확산되면서 낚시인과 등산객 등이 쪽잠을 자던 차박(차 안에서 잠을 자는 캠핑)은 옛말이 됐다. 오히려
소매유통업체의 체감경기가 3분기 연속 큰 폭 하락하며 금융위기 때보다 심각한 소비한파가 올 것이라는 전망이 나왔다. 16일 대한상공회의소가 소매유통업체 500개 사를 대상으로 '1분기 소매유통업경기전망지수(RBSI)
68명 사망 · 4명 실종... 네팔 '애도의 날' 선포, 예티항공 모든 항공편 취소 ▲ 네팔 예티항공 여객기 추락 사고를 보도하는 AP통신 갈무리 ⓒ AP 네팔이 68명이 숨지고 4명이 실종된 여객기 추락 사고에 '애
아랍에미리트(UAE)를 국빈 방문한 윤석열 대통령과 부인 김건희 여사가 15일(현지시간) 그랜드 모스크를 찾아 UAE 초대 대통령 묘소를 참배했다. 뒤이어 현지 파병 중인 아크부대를 찾아 장병들을 격려했다. 윤 대통
고발 사주 의혹의 핵심 인물로 지목된 손준성 검사와 김웅 국민의힘 의원의 사법 판단은 밀접하게 연결된다. 손 검사와 김 의원이 공모 관계로 묶여 있어, 한쪽이 먼저 받은 사법처분은 다른 한쪽에

네이버 뉴스섹션메뉴와 섹션별 url 추출

[28]:

```
from urllib.request import urlopen
import bs4
import pandas as pd
```

[29]:

```
url = 'https://news.naver.com'

html = urlopen(url)

bs_obj = bs4.BeautifulSoup(html, 'html.parser')
```

[30]:

```
# 네이버 뉴스 섹션메뉴 태그 확인(개발자도구)
# body > section > header > div.Nlnb._float_lnb > div > div > div.Nlnb
# selector가 너무 길어서 유용하지 않아 보임
# 직접 확인한 태그와 클래스 속성 사용
# ul 태그의 class : Nlnb_menu_list
uls = bs_obj.findAll("ul",{"class":"Nlnb_menu_list"})
len(uls) # 원소가 1개이므로
ul = bs_obj.find("ul",{"class":"Nlnb_menu_list"})
ul
```

```
<ul class="NInb_menu_list" role="menu">
<li class="Nlist_item is_active"><a aria-selected="true" class
="Nitem_link" href="https://news.naver.com/?viewType=pc" onclick
="nclk(event,'lnb.pcmedia','','');" role="menuitem"><span class
="Nitem_link_menu">언론사별</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/main.naver?mode=LSD&amp;mid
=shm&amp;sid1=100" onclick="nclk(event,'lnb.pol','','');" role
="menuitem"><span class="Nitem_link_menu">정치</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/main.naver?mode=LSD&amp;mid
=shm&amp;sid1=101" onclick="nclk(event,'lnb.eco','','');" role
="menuitem"><span class="Nitem_link_menu">경제</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/main.naver?mode=LSD&amp;mid
=shm&amp;sid1=102" onclick="nclk(event,'lnb.soc','','');" role
="menuitem"><span class="Nitem_link_menu">사회</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/main.naver?mode=LSD&amp;mid
=shm&amp;sid1=103" onclick="nclk(event,'lnb.lif','','');" role
="menuitem"><span class="Nitem_link_menu">생활/문화</span></a></
li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/main.naver?mode=LSD&amp;mid
=shm&amp;sid1=105" onclick="nclk(event,'lnb.sci','','');" role
="menuitem"><span class="Nitem_link_menu">IT/과학</span></a></li
>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/main.naver?mode=LSD&amp;mid
=shm&amp;sid1=104" onclick="nclk(event,'lnb.wor','','');" role
="menuitem"><span class="Nitem_link_menu">세계</span></a></li>
<li class="Nlist_item _isNew"><a aria-selected="false" class="Ni
tem_link" href="https://news.naver.com/main/ranking/popularDay.n
aver" onclick="nclk(event,'lnb.rank','','');" role="menuitem"><s
pan class="Nitem_link_menu">랭킹</span></a></li>
<li class="Nlist_item _isNew"><a aria-selected="false" class="Ni
tem_link" href="https://news.naver.com/newspaper/home?viewType=p
c" onclick="nclk(event,'lnb.paper','','');" role="menuitem"><spa
n class="Nitem_link_menu">신문보기</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/opinion/home" onclick="nclk(even
t,'lnb.opi','','');" role="menuitem"><span class="Nitem_link_men
u">오피니언</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/tv/index.naver?mid=tvh" onc
lick="nclk(event,'lnb.tv','','');" role="menuitem"><span class
="Nitem_link_menu">TV</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_lin
k" href="https://news.naver.com/main/factcheck/main.naver" oncli
ck="nclk(event,'lnb.fact','','');" role="menuitem"><span class
="Nitem_link_menu">팩트체크</span></a></li>
</ul>
```

[31]:

```
lis = ul.findAll('li')
len(lis)
```

12

[32]:

```
for li in lis :
    a_tag = li.find('a')
    print(a_tag.text," : ",a_tag['href'])
```

언론사별 : https://news.naver.com/?viewType=pc (https://news.naver.com/?viewType=pc)
정치 : https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=100 (https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=100)
경제 : https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=101 (https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=101)
사회 : https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=102 (https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=102)
생활/문화 : https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=103 (https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=103)
IT/과학 : https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=105 (https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=105)
세계 : https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=104 (https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=104)
랭킹 : https://news.naver.com/main/ranking/popularDay.naver (https://news.naver.com/main/ranking/popularDay.naver)
신문보기 : https://news.naver.com/newspaper/home?viewType=pc (https://news.naver.com/newspaper/home?viewType=pc)
오피니언 : https://news.naver.com/opinion/home (https://news.naver.com/opinion/home)
TV : https://news.naver.com/main/tv/index.naver?mid=tvh (https://news.naver.com/main/tv/index.naver?mid=tvh)
팩트체크 : https://news.naver.com/main/factcheck/main.naver (https://news.naver.com/main/factcheck/main.naver)

[33]:

```
# li태그를 이용해서 추출
lis = bs_obj.findAll("li",{"class" : "Nlist_item"})
len(lis)
```

12

[34]:

```
for li in lis :
    a_tag = li.find('a')
    print(a_tag.text)
    print(a_tag['href'])
```

```
언론사별
https://news.naver.com/?viewType=pc (https://news.naver.com/?vie
wType=pc)
정치
https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=100
(https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=10
0)
경제
https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=101
(https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=10
1)
사회
https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=102
(https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=10
2)
생활/문화
https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=103
(https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=10
3)
IT/과학
https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=105
(https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=10
5)
세계
https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=104
(https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=10
4)
랭킹
https://news.naver.com/main/ranking/popularDay.naver (https://ne
ws.naver.com/main/ranking/popularDay.naver)
신문보기
https://news.naver.com/newspaper/home?viewType=pc (https://news.
naver.com/newspaper/home?viewType=pc)
오피니언
https://news.naver.com/opinion/home (https://news.naver.com/opin
ion/home)
TV
https://news.naver.com/main/tv/index.naver?mid=tvh (https://new
s.naver.com/main/tv/index.naver?mid=tvh)
팩트체크
https://news.naver.com/main/factcheck/main.naver (https://news.n
aver.com/main/factcheck/main.naver)
```

[35]:

```
# 수집 데이터 df로 구성 후 저장
section = []
link = []
```

[36]:

```
for li in lis :
    a_tag = li.find('a')
    section.append(a_tag.text)
    link.append(a_tag['href'])
```

[37]:

```
col_dict = {'section':section, "link":link}
news_section_df = pd.DataFrame(col_dict)
news_section_df
```

	section	link
0	언론사별	https://news.naver.com/?viewType=pc
1	정치	https://news.naver.com/main/main.naver?mode=LS...
2	경제	https://news.naver.com/main/main.naver?mode=LS...
3	사회	https://news.naver.com/main/main.naver?mode=LS...
4	생활/문화	https://news.naver.com/main/main.naver?mode=LS...
5	IT/과학	https://news.naver.com/main/main.naver?mode=LS...
6	세계	https://news.naver.com/main/main.naver?mode=LS...
7	랭킹	https://news.naver.com/main/ranking/popularDay...
8	신문보기	https://news.naver.com/newspaper/home?viewType=pc
9	오피니언	https://news.naver.com/opinion/home
10	TV	https://news.naver.com/main/tv/index.naver?mid...
11	팩트체크	https://news.naver.com/main/factcheck/main.naver

[38]:

```
# news_section_df.to_csv('./navew_news_section.csv', encoding='euc-kr
news_section_df.to_csv('c:\\Myexam\\navew_news_section.csv', encoding=
```

[ ]:

```

```

[ ]:

```

```