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

In [1]:
```
from urllib.request import urlopen # 서버 요청/응답 패키지
import bs4 # 파싱패키지
```

In [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')
```

In [3]:
```
# 출력이 길어서 일부분만 출력
print(bs_obj.prettify()[:1944])
```

```
<!DOCTYPE html>
<html class="fzoom" lang="ko">
 <head>
  <meta charset="utf-8"/>
  <meta content="origin" name="Referrer"/>
  <meta content="IE=edge" http-equiv="X-UA-Compatible"/>
  <meta content="width=1190" name="viewport"/>
  <title>
   NAVER
  </title>
  <meta content="NAVER" name="apple-mobile-web-app-title">
   <meta content="index,nofollow" name="robots">
    <meta content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요" name="description">
     <meta content="네이버" property="og:title"/>
     <meta content="https://www.naver.com/" property="og:url"/>
     <meta content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png" property="og:image"/>
     <meta content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요" property="og:description">
      <meta content="summary" name="twitter:card"/>
      <meta content="" name="twitter:title"/>
      <meta content="https://www.naver.com/" name="twitter:url"/>
      <meta content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png" name="twitter:image"/>
      <meta content="네이버 메인에서 다양한 정보와 유용한 컨텐츠를 만나 보세요" name="twitter:description">
       <meta content="uru5NJKa1Bfr5nv5AdQ26Qat7UrPU_02l-PIZRLzI-g" name="google-site-verification">
        <link href="/favicon.ico?1" rel="shortcut icon" type="image/x-icon"/>
        <link href="https://s.pstatic.net/static/www/nFavicon96.png" rel="apple-touch-icon-precomposed">
         <link href="https://s.pstatic.net/static/www/u/2014/0328/mma_204243574.png" rel="apple-touch-icon" sizes="114x114"/>
         <link href="https://s.pstatic.net/static/www/u/2014/0328/mma_20432863.png" rel="apple-touch-icon"/>
         <link href="https://ssl.pstatic.net/sstatic/search/pc/css/sp_autocomplete_251001.css" rel="stylesheet"/>
         <link href="https://www.naver.com/" rel="canonical"/>
         <link href="https://m.naver.com/" media="only screen and (max-width: 767px)" rel="alternate"/>
         <script>
```

In [4]:
```
test_menu = bs_obj.find('div',{'id':'wrap'})
# 출력이 길어서 일부분만 출력
print(test_menu.prettify()[1000:2950])
```

```
<input disabled="disabled" id="qdt" name="qdt" type="hidden" value=""/>
        <input id="ie" name="ie" type="hidden" value="utf8"/>
        <input disabled="disabled" id="acir" name="acir" type="hidden" value=""/>
        <input disabled="disabled" id="os" name="os" type="hidden" value=""/>
        <input disabled="disabled" id="bid" name="bid" type="hidden" value=""/>
        <input disabled="disabled" id="pkid" name="pkid" type="hidden" value=""/>
        <input disabled="disabled" id="eid" name="eid" type="hidden" value=""/>
        <input disabled="disabled" id="mra" name="mra" type="hidden" value=""/>
        <div class="search_input_box">
         <input aria-activedescendant="" aria-autocomplete="list" aria-controls="" aria-expanded="false" aria-haspopup="listbo
x" aria-owns="autoFrame" autocomplete="off" class="search_input" data-atcmp-element="" id="query" maxlength="255" name="query"
placeholder="검색어를 입력해 주세요." role="combobox" title="검색어를 입력해 주세요." type="search"/>
        </div>
        <button class="btn_search" onclick='window.ntm.push({event:"nclick",el:this,click_area:"sch.action"})' type="submit">
         <span class="ico_btn_search_svg">
          <svg id="search-btn" viewbox="0 0 50 50" xmlns="http://www.w3.org/2000/svg">
           <path d="M22.13.5c11.378 0 20.632 9.256 20.632 20.63 0 4.699-1.566 9.155-4.439 12.782l10.164 10.165a2.41 2.41 0 0 1-
1.706 4.115 2.412 2.412 0 0 1-1.706-.705L33.31 35.719a2.41 2.41 0 0 1 0-3.41 15.71 15.71 0 0 0 0 4.628-11.178c0-8.718-7.09-15.808
-15.807-15.808-8.718 0-15.808 7.09-15.808 15.808 0 7.15 4.817 13.43 11.714 15.273a2.41 2.41 0 0 1 1.705 2.954 2.41 2.41 0 0 1-
2.95 1.705C7.788 38.658 1.5 30.46 1.5 21.131 1.5 9.756 10.756.5 22.13.5zm4.716 34.746a3.483 3.483 0 1 1 0 6.966 3.483 3.483 0 0
1 0-6.966z">
           </path>
          </svg>
         </span>
         <span class="blind">
          검색
         </span>
        </button>
       </fieldset>
      </form>
```

수집한 데이터를 csv로 저장

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

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

In [5]:
```
# 네이버 뉴스 크롤링
# url은 기본 url 부터 사용
url = 'https://news.naver.com'
html = urlopen(url)
```

In [6]:
```
# html #<http.client.HTTPResponse at 0x1b6fa0cd0d0>
# html_text = html.read() # 바이너리 문자열로 반환
# # html_text
```

In [7]:
```
# bs4 객체 생성
bs_obj = bs4.BeautifulSoup(html, 'html.parser')
```

In [8]:
```
# 출력이 길어서 일부분만 출력
print(bs_obj.prettify()[:1000])
```

```
<!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() && isAbleApplyPrefersColorScheme()) {

                                document.querySelector("html").classList.add("DARK_THEME");
                        }
                } catch(e) {}

                function isAbleApplyPrefersColorScheme() {

                        if (window.matchMedia("(prefers-color-scheme)").matches === false) {
                                return false;
                        }

                        var userAgent = navigator.userAgent;

                        if (userAgent.indexOf("NAVER") > -1) {

                                if (/.*NAVER\([a-zA-Z]*;\s[a-zA-Z]*;\s([0-9]*);/.test(userAgent)) {
                                        return Number(RegExp.$1) >= 1000;
                                }
                        } else {

                                return document.cookie.indexOf("NSCS=1") > -1;
                        }

                        return false;
                }
        })();
```

```
</script>
<script>
 var g_ssc = 'news.v3_media' || null;
             var bSupportedIntersectionObserver = "IntersectionObserver" in window;
```

In [9]:
```
news_title = bs_obj.findAll('strong',{'class':'cnf_news_title'})
news_title2 = bs_obj.findAll('a',{'class':'cnf_news'})
news_titles = news_title + news_title2

# news_titles = bs_obj.findAll('strong', {'class': 'cnf_news_title'}) + bs_obj.findAll('a', {'class': 'cnf_news _cds_link _edi

# CSS 선택자 사용
# news_titles = bs_obj.select('strong.cnf_news_title, a.cnf_news')
```

In [10]:
```
len(news_titles)
```

Out[10]:
```
328
```

In [11]:
```
# 328개의 기사 중 상위 10개만 출력
news_titles[:10]
```

Out[11]:
```
[<strong class="cnf_news_title">“돈 없고 대출도 안되고 10평이라도 가야죠” 李 정부 출범 후, 중소형 12% 급등 [부동산360]</strong
>,
 <strong class="cnf_news_title">민주, 사법개혁안 공청회…국힘,구미서 장외 여론전</strong>,
 <strong class="cnf_news_title">[단독]이화여대 교내서 조류인플루엔자 검출… 학생·수험생에 ‘쉬쉬’</strong>,
 <strong class="cnf_news_title">[단독] 선익 8세대 증착기, BOE '최고등급'에 3·4라인 발주 착수 …'삼성·LGD' 공급망 위협</strong>,
 <strong class="cnf_news_title">청주서 50대 여성 40일 넘게 실종...경찰 '강력범죄' 가능성도</strong>,
 <strong class="cnf_news_title">“귀 계속 아프더니, ‘이 벌레’ 바글바글”… 20대 女, 무슨 일?</strong>,
 <strong class="cnf_news_title">법원행정처장, ‘법정모욕’ 김용현 변호인들 고발···“사법질서에 대한 부정행위”</strong>,
 <strong class="cnf_news_title">[속보] 여야, 대장동 국정조사 또 결론 못내…"추후 다시 논의키로"</strong>,
 <strong class="cnf_news_title">“1인당 10만원 또 드릴게요”…민생지원금 531억원 뿌린다는 ‘이 지역’</strong>,
 <strong class="cnf_news_title">비트코인 팔던 기요사키, 이젠 '이것' 산다?…"내년 4배로 뛸 것"</strong>]
```

In [12]:
```
for title in news_titles[:10]:
    print(title.text)
```

```
“돈 없고 대출도 안되고 10평이라도 가야죠” 李 정부 출범 후, 중소형 12% 급등 [부동산360]
민주, 사법개혁안 공청회…국힘,구미서 장외 여론전
[단독]이화여대 교내서 조류인플루엔자 검출… 학생·수험생에 ‘쉬쉬’
[단독] 선익 8세대 증착기, BOE '최고등급'에 3·4라인 발주 착수 …'삼성·LGD' 공급망 위협
청주서 50대 여성 40일 넘게 실종...경찰 '강력범죄' 가능성도
“귀 계속 아프더니, ‘이 벌레’ 바글바글”… 20대 女, 무슨 일?
법원행정처장, ‘법정모욕’ 김용현 변호인들 고발...“사법질서에 대한 부정행위”
[속보] 여야, 대장동 국정조사 또 결론 못내…"추후 다시 논의키로"
“1인당 10만원 또 드릴게요”…민생지원금 531억원 뿌린다는 ‘이 지역’
비트코인 팔던 기요사키, 이젠 '이것' 산다?…"내년 4배로 뛸 것"
```

네이버 뉴스섹션메뉴와 섹션별 url 추출

In [13]:
```
from urllib.request import urlopen
import bs4
import pandas as pd
```

In [14]:
```
url = 'https://news.naver.com'

html = urlopen(url)

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

In [15]:
```
# 네이버 뉴스 섹션메뉴 태그 확인(개발자도구)
# body > section > header > div.Nlnb._float_lnb > div > div > div.Nlnb_left._lnb_scroll > div > div > ul
# 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
```

Out[15]:
```
<ul class="Nlnb_menu_list" role="menu">
<li class="Nlist_item is_active"><a aria-selected="true" class="Nitem_link" href="https://news.naver.com/?viewType=pc" onclic
k="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_link" href="https://news.naver.com/section/100" onclick="nclk(ev
ent,'lnb.pol','','');" role="menuitem"><span class="Nitem_link_menu">정치</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/section/101" onclick="nclk(ev
ent,'lnb.eco','','');" role="menuitem"><span class="Nitem_link_menu">경제</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/section/102" onclick="nclk(ev
ent,'lnb.soc','','');" role="menuitem"><span class="Nitem_link_menu">사회</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/section/103" onclick="nclk(ev
ent,'lnb.lif','','');" role="menuitem"><span class="Nitem_link_menu">생활/문화</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/section/105" onclick="nclk(ev
ent,'lnb.sci','','');" role="menuitem"><span class="Nitem_link_menu">IT/과학</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/section/104" onclick="nclk(ev
ent,'lnb.wor','','');" role="menuitem"><span class="Nitem_link_menu">세계</span></a></li>
<li class="Nlist_item _isNew"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/main/ranking/popularDa
y.naver" onclick="nclk(event,'lnb.rank','','');" role="menuitem"><span class="Nitem_link_menu">랭킹</span></a></li>
<li class="Nlist_item _isNew"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/newspaper/home?viewTyp
e=pc" onclick="nclk(event,'lnb.paper','','');" role="menuitem"><span class="Nitem_link_menu">신문보기</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/opinion/home" onclick="nclk(e
vent,'lnb.opi','','');" role="menuitem"><span class="Nitem_link_menu">오피니언</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/main/tv/index.naver?mid=tvh"
onclick="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_link" href="https://news.naver.com/factcheck/main" onclick="nclk
(event,'lnb.fact','','');" role="menuitem"><span class="Nitem_link_menu">팩트체크</span></a></li>
<li class="Nlist_item"><a aria-selected="false" class="Nitem_link" href="https://media.naver.com/algorithm" onclick="nclk(eve
nt,'lnb.algo','','');" role="menuitem"><span class="Nitem_link_menu">알고리즘 안내</span></a></li>
<li class="Nlist_item _isNew"><a aria-selected="false" class="Nitem_link" href="https://news.naver.com/ombudsman/errorArticle
List" onclick="nclk(event,'lnb.correct','','');" role="menuitem"><span class="Nitem_link_menu">정정보도 모음</span></a></li>
</ul>
```

In [16]:
```
lis = ul.findAll('li')
len(lis)
```

Out[16]:
```
14
```

In [17]:
```
for li in lis :
    a_tag = li.find('a')
    print(a_tag.text," : ",a_tag['href'])
```

```
언론사별  :  https://news.naver.com/?viewType=pc
정치  :  https://news.naver.com/section/100
경제  :  https://news.naver.com/section/101
사회  :  https://news.naver.com/section/102
생활/문화  :  https://news.naver.com/section/103
IT/과학  :  https://news.naver.com/section/105
세계  :  https://news.naver.com/section/104
랭킹  :  https://news.naver.com/main/ranking/popularDay.naver
신문보기  :  https://news.naver.com/newspaper/home?viewType=pc
오피니언  :  https://news.naver.com/opinion/home
TV  :  https://news.naver.com/main/tv/index.naver?mid=tvh
팩트체크  :  https://news.naver.com/factcheck/main
알고리즘 안내  :  https://media.naver.com/algorithm
정정보도 모음  :  https://news.naver.com/ombudsman/errorArticleList
```

In [18]:
```
# li태그를 이용해서 추출
lis = bs_obj.findAll("li",{"class" : "Nlist_item"})
len(lis)
```

Out[18]: 14

In [19]:
```
for li in lis :
    a_tag = li.find('a')
    print(a_tag.text)
    print(a_tag['href'])
```

```
언론사별
https://news.naver.com/?viewType=pc
정치
https://news.naver.com/section/100
경제
https://news.naver.com/section/101
사회
https://news.naver.com/section/102
생활/문화
https://news.naver.com/section/103
IT/과학
https://news.naver.com/section/105
세계
https://news.naver.com/section/104
랭킹
https://news.naver.com/main/ranking/popularDay.naver
신문보기
https://news.naver.com/newspaper/home?viewType=pc
오피니언
https://news.naver.com/opinion/home
TV
https://news.naver.com/main/tv/index.naver?mid=tvh
팩트체크
https://news.naver.com/factcheck/main
알고리즘 안내
https://media.naver.com/algorithm
정정보도 모음
https://news.naver.com/ombudsman/errorArticleList
```

In [20]:
```
# 수집 데이터 df로 구성 후 저장
section = []
link = []
```

In [21]:
```
for li in lis :
    a_tag = li.find('a')
    section.append(a_tag.text)
    link.append(a_tag['href'])
```

In [22]:
```
col_dict = {'section':section, "link":link}
news_section_df = pd.DataFrame(col_dict)
```

```
news_section_df
```

Out[22]:

	section	link
0	언론사별	https://news.naver.com/?viewType=pc
1	정치	https://news.naver.com/section/100
2	경제	https://news.naver.com/section/101
3	사회	https://news.naver.com/section/102
4	생활/문화	https://news.naver.com/section/103
5	IT/과학	https://news.naver.com/section/105
6	세계	https://news.naver.com/section/104
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/factcheck/main
12	알고리즘 안내	https://media.naver.com/algorithm
13	정정보도 모음	https://news.naver.com/ombudsman/errorArticleList

In [23]:
```
# news_section_df.to_csv('./navew_news_section.csv', encoding='euc-kr')
news_section_df.to_csv('c:\\Myexam\\naver_news_section.csv', encoding='euc-kr')
```

In [ ]:
```

```