In [ ]:
```
DOM(HTML -> Tree)
root -> leaf find_
find, find_all -> RE, limit, recursive
```

In [ ]:
```
HTML태그, 속성(ID, class, href, src, data-..), 구조, 가상선택자
```

In [14]:
```
html = '''
    <html>
        <head></head>
        <body>
            <div id="a">
                <uL>
                    <li class="a"></li>
                    <li class="a b"></li>
            </DIV>
                </Ul>
        </body>
    </html>
'''
```

In [2]:
```
from bs4 import BeautifulSoup
```

In [15]:
```
dom = BeautifulSoup(html, 'html.parser')
```

In [9]:
```
# ' ' -> 자손, '>' -> 자식, '+' -> next sibling
dom.select('body > div > li'),\ # dom.find(div, recursive=False).find(ul)
dom.select('body div li')
```

Out[9]:
```
([],
 [<li>
  <li>
  </li></li>,
  <li>
  </li>])
```

In [13]:
```
dom.select('li:has(>li)')
dom.select('li li')
```

Out[13]:
```
[<li>
 </li>]
```

In [17]:
```
#ID, .class .class1.class2 [class=어쩌고]
dom.select('#a, div, div#a, div[id=a]')
```

Out[17]:
```
[<div id="a">
 <ul>
 <li class="a"></li>
 <li class="a b"></li>
 </ul></div>]
```

In [18]:
```
dom.select_one('#a') is dom.select_one('div')
```

Out[18]:
```
True
```

In [19]:
```
dom.select_one('#a') is dom.select_one('div#a')
```

Out[19]:
```
True
```

In [20]:
```
dom.select_one('#a') is dom.select_one('[id=a]')
```

Out[20]:
```
True
```

In [21]:
```
dom.select_one('#a') is dom.select_one('body > *')
```

Out[21]:
```
True
```

In [26]:
```
dom.select_one('#a') is dom.select_one('*:has(> ul)')
```

Out[26]:
```
True
```

In [29]:
```
dom.select_one('li'), dom.select_one('li + li')
```

Out[29]:
```
(<li class="a"></li>, <li class="a b"></li>)
```

In [ ]:
```
첫번째 li, ul > li, .a, li.a, li:has(+li), li:first-child
```

In [30]:
```
dom.select_one('li:has(+li)')
```

Out[30]:
```
<li class="a"></li>
```

In [31]:
```
dom.select_one('li:nth-of-type(1)')
```

Out[31]:
```
<li class="a"></li>
```

In [ ]:
```
두번째 li, ul > li + li, .b, .a.b, li:last-child, [class^$~=]
```

In [41]:
```
from requests import request

url = 'http://pythonscraping.com/pages/page3.html'
resp = request('GET', url)
dom = BeautifulSoup(resp.text, 'html.parser')
```

In [43]:
```
dom.select('table img[src$=jpg]') # 구조, <img src='.....jpg'> => src
```

Out[43]:
```
[<img src="../img/gifts/img1.jpg"/>,
 <img src="../img/gifts/img2.jpg"/>,
 <img src="../img/gifts/img3.jpg"/>,
 <img src="../img/gifts/img4.jpg"/>,
 <img src="../img/gifts/img6.jpg"/>]
```

In [55]:
```
import re

[re.search('[$][0-9.]+', td.get_text()).group()
 for td in dom.select('tr > td:nth-of-type(3)')]
```

Out[55]:
```
['$15.00', '$10', '$10', '$0.50', '$1.50']
```

In [56]:
```
url = 'https://pythonscraping.com/pages/wikipedia.html'
resp = request('GET', url)
dom = BeautifulSoup(resp.text, 'html.parser')
```

In [64]:
```
dom.select('div:has(> input[type=radio])')[0].get_text().split()
```

Out[64]:
```
['English',
 'French',
 'Chinese',
 'Cebuano',
 'Vietnamese',
 'Catalan',
 'German',
 'Italian',
 'Hungarian',
 'Esperanto',
 'Czech',
 'Finnish',
 'Danish',
 'Japanese',
 'Spanish',
 'Russian',
 'Dutch',
 'Malay',
 'Portuguese',
 'Norwegian',
 'Serbian',
 'Indonesian',
 'Korean',
 'Swedish',
 'Turkish',
 'Slovak',
 'Ukranian',
 'Polish']
```

In [121…
```
headers = {
    'user-agent':'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.3
}

url = 'https://www.google.com/search'
params = {'q':'한소희'}
resp = request('GET', url, params=params, headers=headers)
dom = BeautifulSoup(resp.text, 'html.parser')
```

In [73]:
```
for a in dom.select('a[href]:has(>.LC20lb)'):
    print(a.attrs['href'])
    print(a.select_one('.LC20lb').get_text().strip())
    print(a.get_text().strip())
```

```
https://namu.wiki/w/%ED%95%9C%EC%86%8C%ED%9D%AC
한소희 - 나무위키
한소희 - 나무위키namu.wikihttps://namu.wiki › 한소희
https://www.instagram.com/xeesoxee/
한소희 (@xeesoxee) • Instagram photos and videos
한소희 (@xeesoxee) • Instagram photos and videosinstagram.comhttps://www.instagra
m.com › xeesoxee
https://ko.wikipedia.org/wiki/%ED%95%9C%EC%86%8C%ED%9D%AC
한소희 - 위키백과, 우리 모두의 백과사전
한소희 - 위키백과, 우리 모두의 백과사전wikipedia.orghttps://ko.wikipedia.org › wiki
› 한소희
https://news.mt.co.kr/mtview.php?no=2023012509243970933
한소희, 만취해 大자로 뻗은 모습 '셀프 폭로'…"다시는 안 ...
한소희, 만취해 大자로 뻗은 모습 '셀프 폭로'…"다시는 안 ...mt.co.krhttps://news.mt.c
o.kr › mtview
https://news.sbs.co.kr/news/endPage.do?news_id=N1007054804
술 먹고 大자로 뻗은 한소희…"다시는 술을 먹지 않겠습니다"
술 먹고 大자로 뻗은 한소희…"다시는 술을 먹지 않겠습니다"sbs.co.krhttps://news.sbs.c
o.kr › 일반기사 › 연예
https://www.mk.co.kr/star/hot-issues/view/2023/01/70046/
한소희, 20대 여배우 중 독보적 '고혹美' - 매일경제
한소희, 20대 여배우 중 독보적 '고혹美' - 매일경제mk.co.krhttps://www.mk.co.kr › hot
-issues › view › 2023/01
https://mobile.newsis.com/view.html?ar_id=NISX20230124_0002168088
한소희, 만취해 길거리에 大자로…"다시는 술 먹지 않겠습니다"
한소희, 만취해 길거리에 大자로…"다시는 술 먹지 않겠습니다"newsis.comhttps://mobile.
newsis.com › view
https://m.youtube.com/watch?v=P4sx6CxE6dU
(꿀팁 뒤집어짐) 진짜 한소희 헤메 전담 선생님께 세상 고급진 ...
(꿀팁 뒤집어짐) 진짜 한소희 헤메 전담 선생님께 세상 고급진 ...youtube.comhttps://m.
youtube.com › watch
https://www.marieclairekorea.com/celebrity/2021/07/hansohee/
입덕을 부르는 한소희 | 마리끌레르
입덕을 부르는 한소희 | 마리끌레르marieclairekorea.comhttps://www.marieclairekorea.c
om › CELEB
```

In [123…

```python
url = 'https://search.naver.com/search.naver'
params = {
    'where':'nexearch',
    'query':'한소희'
}
resp = request('GET', url, params=params, headers=headers)
dom = BeautifulSoup(resp.text, 'html.parser')
```

In [80]:

```python
for a in dom.select('a.link_tit, a.news_tit, a.name_link, a.total_tit'):
    print(a.get_text().strip())
    print(a.attrs['href'])
```

```
한소희 - 나무위키
https://namu.wiki/w/%ED%95%9C%EC%86%8C%ED%9D%AC
한소희(@xeesoxee) • Instagram 사진 및 동영상
https://www.instagram.com/xeesoxee/
[얼마예요] “제니가고 한소희 왔다”…‘포스트 이효리’ 노리는 소주 모델들...
https://economist.co.kr/article/view/ecn202303060021
'처음처럼' 모델, 제니→한소희 교체
http://news.mt.co.kr/mtview.php?no=2023030309434084113
한소희 '자낳괴' 였네. 소주모델 발탁 후 '금주 선언' 급취소 폭소
http://www.sportsseoul.com/news/read/1202946?ref=naver
한소희, 금주 선언 한달만에.."취소" 선언 사연은?
https://www.starnewskorea.com/stview.php?no=2023030710265572973
한소희 - 씨네21
http://www.cine21.com/db/person/info/?person_id=104232
천륜 앞에 한소희는 당당하고, 한소희 엄마는 반성하길 [모이라 뷰] - 스타투데이
https://www.mk.co.kr/star/hot-issues/view/2022/03/214568/
한소희X차은우 만났다…美친 비주얼 케미 포착(악마리) - 조선비즈
https://biz.chosun.com/entertainment/tv/2022/09/27/MVYMIJ7JSJBAU674XDFI6FAPEM/
이것도 품절되는 거 아니야? 한소희 3만원대 코프 coap 니트
https://in.naver.com/nilla/contents/internal/533066569133344?areacode=ink*A&query
=%ED%95%9C%EC%86%8C%ED%9D%AC
회자되는 2020 여자헤어스타일 여다경머리 한소희 is 뭔들!
https://in.naver.com/dndkwls_/contents/internal/537643009016768?areacode=ink*A&que
ry=%ED%95%9C%EC%86%8C%ED%9D%AC
볼드한 하트 목걸이 트렌드 모아보기 f.한소희
https://in.naver.com/hongya/contents/internal/532124103089440?areacode=ink*A&query
=%ED%95%9C%EC%86%8C%ED%9D%AC
한소희 타투 디자인 종류! 문신 제거 (피팅모델 시절)
https://blog.naver.com/borareview/222930653320
한소희 눈썹 입술 피어싱, 옆구리 치골 타투, 블로그 주소는?
https://blog.naver.com/chois909/222899860624
한소희 드레스 영국 브리티시 패션 어워즈 가슴 몸매 라인 여신 미모
https://blog.naver.com/ryuri666/222948588408
천원짜리 변호사 김지은 프로필 나이 인스타 총정리 (한소희, 류진 닮은꼴)
https://blog.naver.com/qkrdmsdhr95/222890376163
더 글로리 배우들 차기작 정보 정리 송혜교X한소희 임지연X김태희 박성훈X김수현 정성일X
강동원 이도현...
https://blog.naver.com/okjoa012/223042966713
샬롯틸버리 음영 발렌타인데이 한소희 메이크업룩 섀도우팔레트 추천
https://blog.naver.com/bbekimha/222992899748
[차은우/한소희/민혁/이성경/저스디스] 연예인들의 2023 S/S 시즌 화보 알아보기★
https://post.naver.com/viewer/postView.naver?volumeNo=35387639&memberNo=42107144&v
Type=VERTICAL
```

In [81]:
```
url = 'https://comic.naver.com/webtoon/detail'
params = {
    'titleId':800770,
    'no':24
}
resp = request('GET', url, params=params, headers=headers)
dom = BeautifulSoup(resp.text, 'html.parser')
```

In [83]:
```
len(dom.select('img[id^=content_image_]'))
```

Out[83]: 78

In [84]:
```
len(dom.select('#sectionContWide > img'))
```

Out[84]: 78

In [85]:
```
len(dom.select('[alt="comic content"]'))
```

Out[85]:
```
78
```

In [88]:
```
len(dom.select('.wt_viewer [src^="https://image-comic.pstatic.net/webtoon"]'))
```

Out[88]:
```
78
```

In [94]:
```
len(dom.select('.wt_viewer [src*="webtoon"]'))
```

Out[94]:
```
78
```

In [ ]:
```
len(dom.select('.wt_viewer > img ~ img'))
# <tag class='wt_viewer ... ... ...'>
#   <img>
#   <img>
#   <img>
#   .....
```

In [99]:
```
len(dom.select('.wt_viewer > img:first-child, .wt_viewer > img ~ img'))
```

Out[99]:
```
78
```

In [92]:
```
from requests.compat import urljoin

nurl = urljoin(url, dom.select('[alt="comic content"]')[0].attrs['src'])
```

In [93]:
```
resp = request('GET', nurl, headers=headers)
resp.status_code, resp.reason, resp.headers, resp.request.headers
```

Out[93]:
```
(200,
 'OK',
 {'Server': 'nginx', 'Content-Type': 'image/jpeg', 'Content-Length': '116711', 'La
st-Modified': 'Fri, 16 Dec 2022 07:40:59 GMT', 'ETag': '"639c210b-1c7e7"', 'Accept
-Ranges': 'bytes', 'Cache-Control': 'max-age=1945580', 'Expires': 'Fri, 07 Apr 202
3 13:53:25 GMT', 'Date': 'Thu, 16 Mar 2023 01:27:05 GMT', 'Connection': 'keep-aliv
e'},
 {'user-agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.3
6 (KHTML, like Gecko) Chrome/110.0.0.0 Safari/537.36', 'Accept-Encoding': 'gzip, d
eflate, br', 'Accept': '*/*', 'Connection': 'keep-alive'})
```

In [104…
```
ext = {'jpeg':'jpg', 'jpg':'jpg', 'png':'png'}
media = {'image':ext}
```

In [110…
```
k1, k2 = resp.headers['content-type'].split('/')
if k1 in media.keys() and k2 in media[k1].keys():
    print('{}.{}'.format(resp.url.split('/')[-1], media[k1][k2]))
```

```
20221216164126_f2b84800b915d3e308935dcb994977da_IMAG01_1.jpg.jpg
```

In [111…
```
from os import listdir, mkdir
```

In [113…
```
params, listdir('.')
```

Out[113]:
```
({'titleId': 800770, 'no': 24},
 ['playlist.db',
  'Untitled1.ipynb',
  'Untitled3.ipynb',
  'Untitled.ipynb',
  'Untitled4.ipynb',
  'sns.db',
  'Untitled2.ipynb',
  '0313.ipynb',
  'orm1.db',
  'img1.jpeg',
  '0315.ipynb',
  '.ipynb_checkpoints',
  'test1.db',
  'test.db',
  '0316.ipynb'])
```

In [ ]:
```
List - Stack(LIFO,FILO), Queue(FIFO)
starting -> item - item - item - item <- push(), append()
            item - item - item - item <- starting, push(), append()
URL, depth, => Focused Crawling(웹툰, 뉴스, ...)
```

In [150…
```
from urllib.robotparser import RobotFileParser
from requests import get
from requests.compat import urlparse
from requests.exceptions import HTTPError
from time import sleep

URLs = list()
seens = list()
URLs.append('https://search.naver.com/search.naver?where=nexearch&query=%ED%95%9C%E

# Queue
# while URLs:
for _ in range(10):
    # 방문 할 주소 하나 꺼내기:Queue(첫링크)
    seed = URLs.pop()
    # 방문 한 주소 목록
    seens.append(seed)

    # robots.txt 확인
    # urljoin => host/robots.txt
#     rp = RobotFileParser(urljoin(seed, '/robots.txt'))
    # can_fetch('/path')
#     if rp.can_fetch(urlparse(seed).path):
#         print('OK')
#     else:
#         print('NO')
#         continue

    # HTTP request with Header(user-agent), GET
    resp = get(seed, headers=headers)

    # Status Code == 200
    try:
        resp.raise_for_status()
    except HTTPError as e:
        # 400-500
        print(e)
        # 만약 500 sleep() 하고 retry
```

```
        continue

#     resp 콘텐츠 타입 확인
#     resp.headers['content-type']
    if re.search('text|html', resp.headers['content-type']) is None:
        continue

#     DOM 생성
    dom = BeautifulSoup(resp.text, 'html.parser')

#     Hypyerlinks 추출
    for link in dom.select('a[href], iframe[src]'):
        # fragment 는 링크가 아니다, 동일 페이지 내 위치
        url = urljoin(seed, link.attrs['href']
                      if link.has_attr('href') else link.attrs['src'])
#         URL seen?
        if len(urlparse(url).fragment) == 0 and\
           urlparse(url).scheme in ['http', 'https']:
#             urlparse(url).scheme.startswith('http')
#             re.match('https?', urlparse(url).scheme)
            if url not in URLs and url not in seens:
                URLs.append(url)
    print(len(URLs))
```

```
202
280
292
312
373
379
396
459
458
459
```

In [ ]:
```
Queue: 411개, Stack: 459개
```

In [154…
```
URLs[-10:], seens
```

Out[154]:
```
(['http://www.webwatch.or.kr/Situation/WA_Situation.html?page=11&skey=&sval=&npp=&
MenuCD=110',
  'http://www.webwatch.or.kr/Situation/WA_Situation.html?page=1318&skey=&sval=&npp
=&m_id=&MenuCD=110',
  'http://www.msip.go.kr/',
  'http://www.nia.or.kr/',
  'http://www.ableforum.com/',
  'http://www.socialenterprise.or.kr',
  'https://www.w3.org/',
  'http://www.webwatch.or.kr/user_info/info_1.html?MenuCD=820',
  'http://www.webwatch.or.kr/user_info/info_2_2.html?MenuCD=830',
  'http://www.webwatch.or.kr/user_info/info_2_3.html?MenuCD=830'],
 ['https://search.naver.com/search.naver?where=nexearch&query=%ED%95%9C%EC%86%8C%E
D%9D%AC',
  'https://www.navercorp.com/',
  'https://www.naverfincorp.com',
  'https://post.naver.com/my.naver?memberNo=30633733',
  'http://www.navercorp.com/ko/index.nhn',
  'http://www.navercorp.com/policy/person',
  'https://ecrm.cyber.go.kr',
  'http://www.webwatch.or.kr/Situation/WA_Situation.html?MenuCD=110',
  'http://www.webwatch.or.kr/user_info/info_3.html?MenuCD=840',
  'http://www.webwatch.or.kr/user_info/info_2_1.html?MenuCD=830'])
```