



## 6. 웹 스크래핑 사례연구

### 6.1. 다음 뉴스

#### 6.1.1. 검색결과 가져오기

다음(<http://daum.net>)에서 "인공지능"으로 검색을 한다.



검색 결과에서 '뉴스', '최신순'을 클릭한다.



이제 해당 검색결과 URL을 `requests` 로 가져온다.

```
import requests
res = requests.get('http://search.daum.net/search?w=news&cluster=n&q=%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5&sort=recency')
```

#### 6.1.2. 기사 링크 모으기

다음 뉴스 검색 결과에서 기사 제목을 클릭하면 언론사 홈페이지로 이동한다. 언론사마다 홈페이지 디자인이 다르기 때문에 자동으로 기사 수집이 어렵다. 대신 주요 언론사의 기사는 기사 제목 옆에 "다음 뉴스"를 클릭하면 다음 내에서 확인할 수 있다.

```
import lxml.html
root = lxml.html.fromstring(res.text)

urls = []
for link in root.cssselect('a.f_nb'):
    urls.append(link.attrib['href'])
```

### 6.1.3. 기사 본문 수집하기

다음 뉴스의 기사 페이지에서 본문은 다음과 같이 수집한다.

```
articles = []
for u in urls:
    if not u.startswith('http'):
        continue
    res = requests.get(u)
    root = lxml.html.fromstring(res.text)
    body = root.cssselect('.article_view').pop()
    content = body.text_content()
    articles.append(content)
```

## 6.2. 네이버 카페

이번 강의에서는 네이버 카페의 게시판을 크롤링 해보자.

이번 예시에서는 의자 관련 카페인 '의자를 사용하는 사람들의 모임 - 의사모' 를 크롤링 해보자.

(<http://cafe.naver.com/duoin>)

필독 새해 복 많이 받으세요. 네이버 | 카페 | 가입카페 | 서이레

**카페정보** | **나의활동** | **★ 등업요청**

이 게시판의 새글을 **내이력.네오독**에서 쉽게 찾아 보시려면 **구독하기**

가입인사를 남겨 주시고 등업요청을 해주세요.

가져오기 25,003 | 즐겨찾는 멤버 2,606명 | 게시판 구독수 3583 | 우리카페글수 15회

초대 | 채팅하기

**카페 가입하기** | **검색**

전체글보기 (66,784) | 카페지식칼럼 | 우리카페지도

**Hot Board**

- 의자나눔 릴레이
- 의자 추천 요청
- 하차 / 결함신고
- T80 결함신고
- 합체어 부착신고
- 출발은 의자 선택법
- 리얼메시 특집
- 의사모 발전

**공지사항**

- 공지사항
- 회원등급
- 해피빈 모금해요

**카뮤니티**

**공지** **의사모 카페 개편 안내**

**공지** **의사모 카페회원의 승려 - '의자 1위업제' 시디즈 사과 이끌어낸 소비자의 열정 [26]**

**공지** **[카페 약용금지] 시디즈 불만글 삭제하신 회원님을 보세요. [16]**

**공지** **중은의자 구입을 위한 올바른 선택 방법 환경편 [76]**

**공지** **시디즈, 듀오백, 파트라, 리바트 어떤 의자가 좋을까? 국내 의자 브랜드 선택 가이드 [61]**

**공지** **와현만 봐도 좋은의자를 찾을 수 있다. [57]**

**공지** **의자를 사용하는 사람들의 건강체조 [77]**

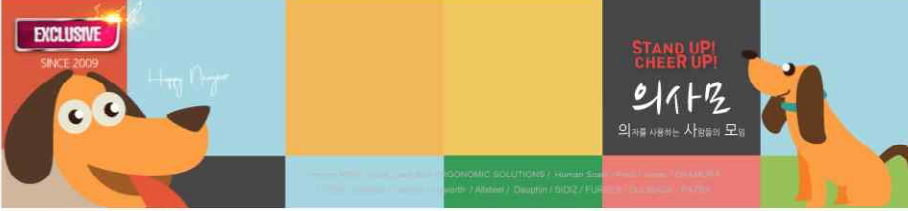
**공지** **의사모 카페 가입 및 강제탈퇴에 관하여 알라드립니다. [216]**

번호	제목	작성자	작성일	조회	응답
75586	등업신청이요	김성찬	07:49	0	0
75584	등업신청합니다.	VodkaTheCat	05:47	1	0
75582	등업신청합니다	Kayne	02:14	1	0
75580	등업신청합니다	DavidKey	02:06	0	0
75578	등업 부탁드립니다~	SOY	02:02	1	0
75575	등업요청 부탁드립니다^^	다고미	2018.01.29.	1	0
75570	등업부탁드립니다	손목아 나이라	2018.01.29.	1	0

메일 이미



Google 검색 또는 URL 입력



**카페정보** **나의활동** ☆

관리자 Manager | 디폴트  
since 2009.12.13. | [카테고리](#)

가시3년과 25,005

★ 즐겨찾는 멤버 2,608명  
☑ 게시물 구독수 358회  
☑ 우리카페업수 15회

[초대](#) [채팅하기](#)

**카페 가입하기**

[검색](#)

☐ 전체글보기 (66,704)  
☐ 카페자식활동  
☑ **우리카페지도**

▶ **Hot Board**

- ☐ 의자나눔 릴레이
- ☐ 의자 추천 요청
- ☐ 하자 / 결함신고
- ☐ T80 결함신고
- ☐ 입체머 부착신고
- ☐ 올바른 의자 선택법
- ☐ 리얼메이 특집
- ☐ 의자모 썰전

▶ **공지사항**

- ☐ 공지 사항
- ☐ 회원등급
- ☐ 해피빈 모금해요

▶ **커뮤니티**

**카페 멤버에게만 공개된 게시물입니다.**

**[의사모]듀오백/시디즈/퍼시스/파트라/해먼밀러/휴먼스케일/의자 카페에 가입해보세요!**

시디즈, 듀오백, 파트라, 퍼시스, 해먼밀러, 스틸케이스, 휴먼스케일, 놀, 바리에르 등 국내외 의자전문 커뮤니티  
카테고리: [가시3년과](#) | [카테고리](#) | [카테고리](#) | since: 2009.12.13.

[카페 가입하기](#)

많은 카페의 경우, 독립적으로 게시글에 접근하면 비회원에게 게시물을 공개하지 않는다.

NAVER

체형에 따른 이상적인 책상, 의자 높이



통합검색

블로그

카페

지식IN

이미지

동영상

어학사전

뉴스

더보기

검색옵션

정렬

기간

출처

영역

유사문서

카페글 게시판

상품등록 옵션

옵션유지

카페글

카페명

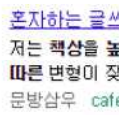
카페글 1-10 / 894건



**디스크환자가 알려드리는 책상,의자에 관해서(feat. 발발침대)** 2017.11.24.  
키에따른 책상/의자높이 여러 세부사항 세팅에 대해 도움을 줍니다 제 키의 경우에는 책상 높이는... 가더라고요 책상 높이에 따른 몸에 가해지는 부하도 달라지는데요...  
FPS관련 장비 대표카페 (embest) [대표 cafe.naver.com/embest...](http://cafe.naver.com/embest) | 카페 내 검색



**높이조절책상 어린이책상은 쌍비책상으로** 2015.07.07.  
줍니다. 높이조절책상 어린이책상 쌍비는 의자에 앉은 상태로 아이들도 손쉽게 높낮이를 조절할 수... 본인의 체형과 상황에 따른 최적의 위치 선정이 가능해졌습니다. 또... 책상&맘수다 체험단 공구카페... [대표 cafe.naver.com/bookc...](http://cafe.naver.com/bookc...) | 카페 내 검색



**혼자하는 글쓰기 연습** 2017.05.12.

저는 책상을 높이고 약간 낮은 의자를 사용해 적절한 팔각 환경을 꾸몄습니다 양 팔꿈치가 책상위에... 압력에 따른 변형이 갖게되고 어떤 펜은 미리듬미 적절히 위치하지 않기도 합니다 이상적인 것은 45...

문방삼우 [cafe.naver.com/fountainpentfriends/148181](http://cafe.naver.com/fountainpentfriends/148181) | 카페 내 검색



**높낮이 없는 의자 사면 책상과의 높이가 안 맞으면 어떡하죠??** **결론답변** 2012.09.26.  
요 청 내 용 : 본인의 키, 체중, 체형에 따른 내용도 작성해주시면 더욱 정확하게 제품을 추천해... 의자가 좋아도 책상과의 높이가 맞지 않으면 도루묵 아닌가요...  
[의사모]듀오백/시디즈/퍼시스/파... [cafe.naver.com/duoin...](http://cafe.naver.com/duoin...) | 카페 내 검색



카페정보

Manager 1 다들P

since 2009.12.13. 카페소개

가져오기

25,005

★ 즐겨찾는 카페

2,606명

게시판 구독수

358회

유인카페뷰수

15회

초대

재방하기

카페 가입하기

검색

전체글보기 (66,704)

카페지식활동

두리카페지도

Hot Board

의자나눔 팔레이

의자 추천 요청

하차 / 결함신고

T00 결함신고

합체어 부식신고

올바른 의자 신박법

리얼메이 특집

의사모 물건

광자사항

공지사항

회원등급

해피빈 모금해요

커뮤니티

작성자가 검색을 허용한 카페 글입니다. ?

이전글

다음글

목록

높낮이 없는 의자 사면 책상과의 높이가 안 맞으면 어떡하죠??

답변 1 | 의자 추천 요청

2012.09.26. 20:12

전환후사(3102\*\*\*\*)

11

<http://cafe.naver.com/duoin/32184>

주소불러

좀더 원활한 달빛을 위해 아래와 같은 형식으로 작성 하여주시길 부탁드립니다. 게시의 성격이 어긋난 광고, 비방성 글들은 삭제될 수 있습니다.

1. 구입 예정 : 예) 2010년 1월

2. 사용 중 도 : 예) 공부방용, 컴퓨터용

3. 사용 연 령 : 예) 20대

4. 희망 모델 : 예) DK-028A, DK-2900

5. 요청 내용 :

본인의 키, 체중, 체형에 따른 내용도 작성해주시면 더욱 정확하게 제품을 추천해 드릴 수 있습니다.

1. 당장

2. 공부용

3. 20대 중반

4. 시디즈 듀오백 의자없는 것으. 저렴하면서도 실용성 있는 것으로요

5. 176 예 65 키로입니다.

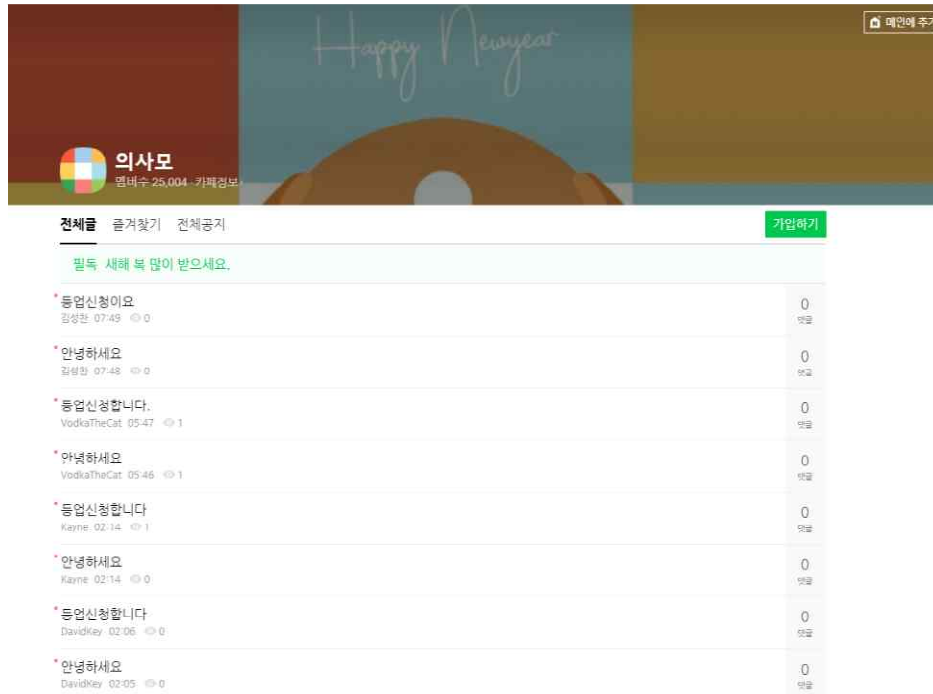
http://doc.mindscale.kr/km/unstructured/un06.html

2018-04-07



```
'https://m.cafe.naver.com/ArticleList.nhn?search.clubid=19773565&search.menuid=142&search.boardtype=L'
```

모바일 주소는 앞에 m 을 붙이면 된다.



모바일 버전의 의사모 게시판 주소를 `ur1` 이라는 변수에 저장한다.

```
ur1 = 'https://m.cafe.naver.com/ArticleList.nhn?search.clubid=19773565&search.menuid=142&search.boardtype=L'
```

후에 크롤링 결과를 확인하면, 최대 20개의 게시물만 크롤링해온다.

더 많은 양의 게시물을 가져오고 싶다면 `ur1` 주소에 페이지를 다르게 해서 가져오는 방법이 있다.

`ur1` 에 `param` 이라는 변수에 저장하면 다른 페이지의 정보도 가져올 수 있다.

```
naver_param = {'search.page': 1}
```

현재 있는 페이지가 1 페이지이다. `search.page` 의 숫자를 다르게하여 다른 페이지의 게시물을 가져올 수 있다.

`requests.get()` 에 `url` 주소와 `param` 이라는 옵션을 넘겨주면 된다.

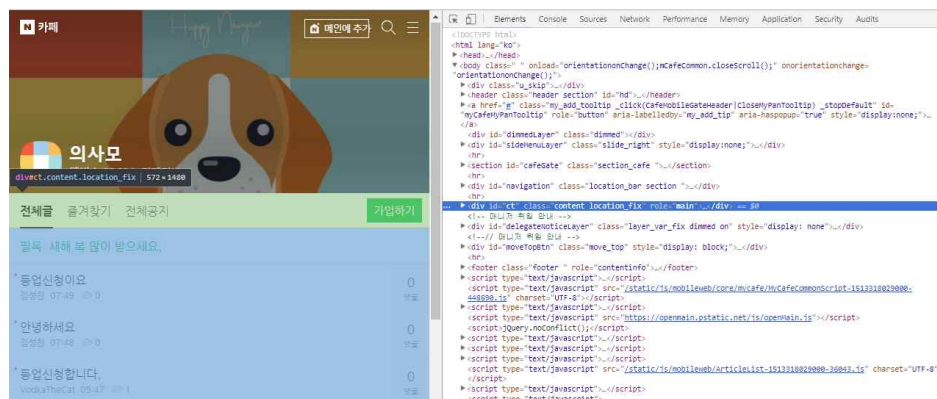


```
res = requests.get(url, param=naver_param)
```

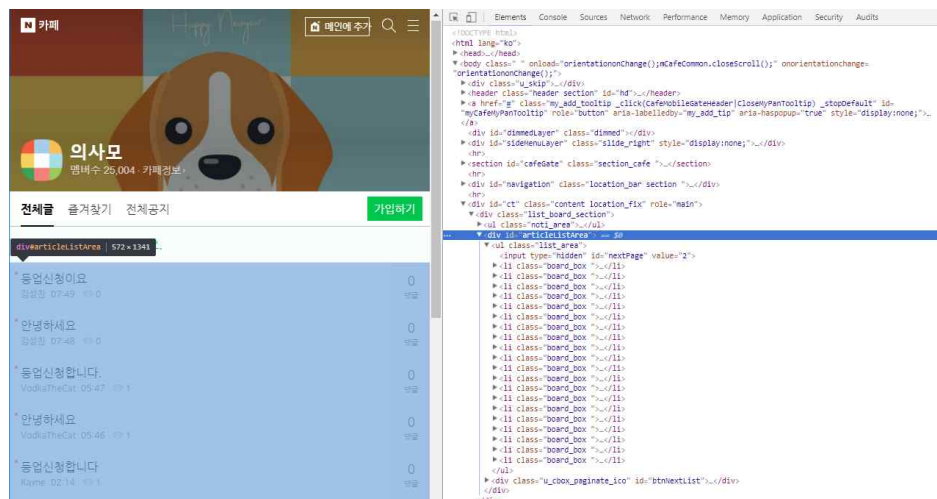
## 6.2.1. 게시판 url 가져오기

이제 게시판의 각 게시물의 url 주소를 가져와보자.

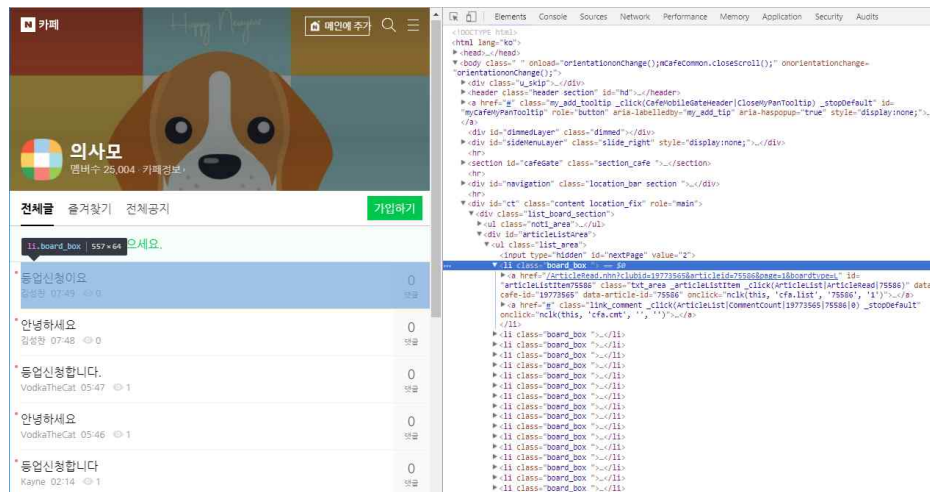
```
elements = lxml.html.fromstring(res.text)
```



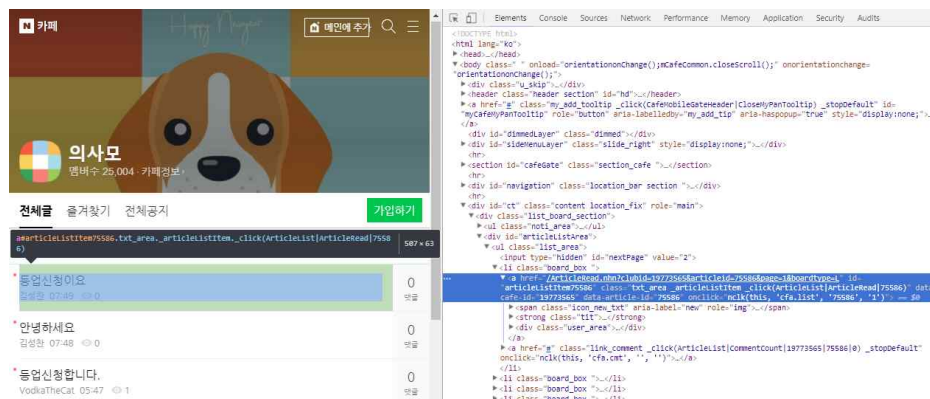
요소검사를 해서 게시글 요소를 찾아보자.



articleListArea 안에 게시물들이 들어있는 것을 확인할 수 있다.



각 게시물이 `li class="board-box"` 라는 태그 안에 `class="txt_area"` 안에 url 이 들어있는 것을 확인할 수 있다.



url 이 들어있는 부분을 선택하도록 하자.

```
postings = elements.cssselect('li.board_box .txt_area')
```

postings 에 있는 각 항목의 href 에 url 주소가 저장되어있으니, .attrib 사용 하여 주소를 뽑아온다.

```
href_list = [a.attrib['href'] for a in postings]
```

```
href_list # href_list 를 확인한다
```

```
[ '/ArticleRead.nhn?clubid=19773565&articleid=75127&page=1&boardtype=L&menuid=142',
  '/ArticleRead.nhn?clubid=19773565&articleid=75125&page=1&boardtype=L&menuid=142',
  ...]
```

각 게시물의 url 주소가 저장되어있다.

주소를 확인하면, `https://m.cafe.naver.com` 이 제외된 미완성 주소인 것을 확인할 수 있다.

`urljoin` 을 사용하여 `https://m.cafe.naver.com` 을 `href_list` 의 각 항목에 추가한다.

```
base_url = 'https://m.cafe.naver.com'
```

```
new_urls = [urljoin(base_url, i) for i in href]
```

`len` 를 통해 `new_urls` 의 개수를 세준다.

```
len(new_urls)
```

```
20
```

총, 20개로 주소가 잘 뽑힌 것을 확인할 수 있다.

### 6.2.2. 제목, 본문 뽑기

이제 각 게시물의 제목, 본문을 크롤링하도록 하자.

## 의사모

의의자정보 마당

## 좋은의자 구입을 위한 올바른 선택 방법 환경편

Manager I 더블P(soci\*\*\*\*) 2016.02.04. 15:01 조회 3,240 댓글출처 블로그>의자, 자연, 캠핑, 여행, 동물, 교육, 친환경 블로그 | Manager 더블P  
원문 [http://m.blog.naver.com/social\\_mkt/220618310225](http://m.blog.naver.com/social_mkt/220618310225)

## 좋은의자 구입을 위한 올바른 선택 방법

의자는 가구분야의 하나로써 교체주기가 빠른편에 속한다.

한사람이 평생사용하는 의자의 수는 생각보다 많지만 실제로 자신이 구입한 의자가 아니라면 실제 사용되어지는 의자들을 그렇게 기억하는 사람은 많지 않다.

서있거나 누워있는 시간을 제외하고는 모두 의자에서 보내는 시간이 대부분이라고 할 수 있다.

식사를 위해 집이나 음식점에서 앉는 의자, 이동을 위한 자동차나 대중교통수단의 의자, 사무실에서 사용하는 의자, 그리고 미팅장소에서의 의자등 의자의 디자인과 용도가 서로 다른 의자를 우리는 매일 반복해서 사용하고 있는 것이다.

타의에 의해 제공되는 의자에 대해서 우리는 그다지 큰 불만이 없다. 그이유는 잠깐동안 사용할 남의 의자라는 인식에서이다.

하지만, 본인이 사용하는 의자에 대해서는 민감하게 생각하는 경우가 많다.

의자를 구입할때 편안하고, 허리건강에 도움이되는 의자를 찾기위해 많은시간을 투자하는 경우가 점차 증가하고 있는 추세이다.

## Referer

앞서 말한대로, 그냥 접속하게 된다면 자료를 가져올 수가 없다.

그래서 네이버 검색창을 통해 접근한다고 생각하도록 해줘야한다.

그러기 위해선, referer 라는 옵션을 설정하면 된다.

먼저 네이버에서 아무 글자로 검색을 한다. 여기서는 숫자 1을 검색하였다.

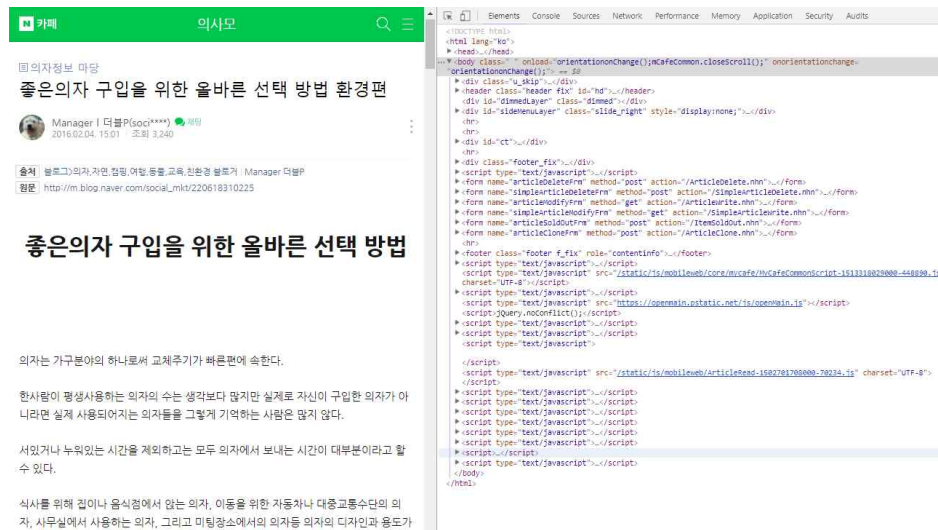
The screenshot displays the Naver search engine interface. At the top, the Naver logo is followed by a search bar containing the number '1'. Below the search bar, there are tabs for different search categories: 통합검색 (Unified Search), 블로그 (Blog), 이미지 (Image), 어학사전 (Language Dictionary), 카페 (Cafe), 지식IN (Knowledge IN), 동영상 (Video), 뉴스 (News), and 더보기 (More). The '어학사전' (Language Dictionary) tab is selected, showing results for the word '1' in various languages. The results include English, Korean, and other language dictionaries. The browser's address bar at the bottom shows the URL: [https://search.naver.com/search.naver?where=nexearch&sm=top\\_hjty&fbm=1&ie=utf8&query=1](https://search.naver.com/search.naver?where=nexearch&sm=top_hjty&fbm=1&ie=utf8&query=1).

주소를 확인하면 `https://m.search.naver.com` 뒤에 `search.naver?query=1` 부분을 `referer` 에 넘겨주면, 네이버 검색을 통한 유입이라고 흉내내면서 자료를 요청을 할 수 있다. `referer` 를 `headers` 라는 변수에 dictionary 형태로 저장한다.

```
headers = {'referer': 'https://m.search.naver.com/search.naver?query=1'}
```

`requests.get()` 에 `url` 과 `headers` 를 같이 넘겨주면 된다.

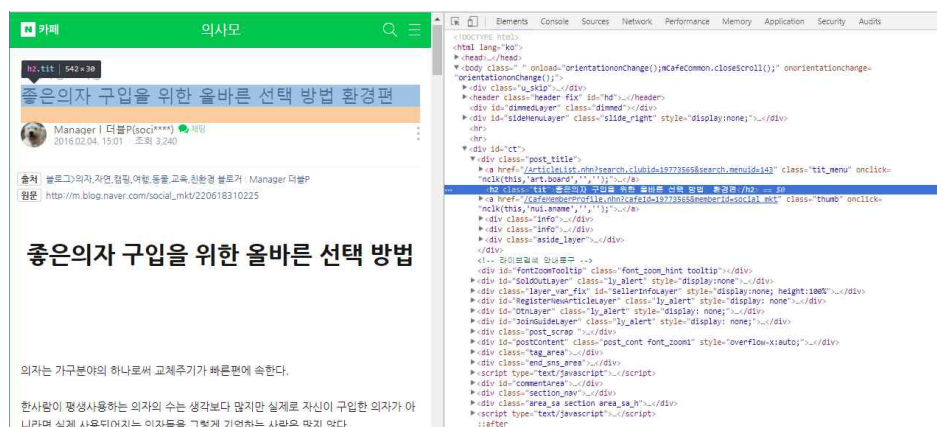
제목과 본문을 크롤링하기 위해 게시물을 하나 들어가보자.



```
ur12 = new_urls[1] # new_urls 목록 중, 두 번째 게시글에 접속한다.
# requests.get() 에 url 주소와 headers 를 같이 넘겨준다.
res2 = requests.get(ur12, headers=headers)
element2 = lxml.html.fromstring(res2.text)
```

## 제목

제목 요소 검사를 하여 내용을 뽑는다. 만약, 마우스 우클릭이 되지 않는다면, 크롬의 경우, 설정 밑에 도구를 클릭하면 개발자 도구를 통해 확인이 가능하다.



제목은 `h2 class="tit"` 아래 있는 것을 확인할 수 있다.

```
element2.cssselect('h2.tit')[0].text_content() # 해당 태그의 텍스트 정보를 가져온다.
```

'등업요청 부탁드립니다'

제목이 잘 뽑힌 것을 확인할 수 있다.





기존에 `new_urls` 에 저장한 게시물들의 제목과, 본문을 뽑아보도록 하자.

우선, 제목과 본문을 저장할 수 있는 빈 리스트를 생성한다.

```
titles = [] # 제목을 저장해둘 빈 리스트를 생성한다.
contents = [] # 본문을 저장할 빈 리스트를 생성한다.
```

그 후에 `for` 문을 통해, 위에서와 같은 방식으로 제목과 본문을 크롤링하면 된다.

간혹 크롤링이 안되는 게시물이 있는데, 그것은 작성자가 글을 올릴 때, 검색 설정을 허용하지 않았기 때문이다.

그럴때는 `referer` 로 흉내를 내도 정보를 가져올 수가 없다.

따라서 접속했을 때, 기존의 코드대로 제목을 뽑으면 빈 리스트가 추출이 된다.

그렇게 되면, `for` 문을 돌 때, 제목을 가져오는 부분에서 `IndexError` 가 발생한다.

그래서, `try` , `except` 문을 사용하여 제목과 본문을 추출한다..

`for` 문을 한번 확인하자.

```
for i in new_urls:
    res = requests.get(i, headers=headers)
    element = lxml.html.fromstring(res.text)

    try: # try 후에, 제목을 가져오는 코드를 실행한다.
        title = element.cssselect('h2.tit')[0].text_content()
    except IndexError: # 만약 인덱스 에러가 발생한다면, except 부분으로 넘어간다.
        continue # continue 를 하여, 해당 게시물을 건너 뛴다.
    titles.append(title) # 제목을 titles 에 append 한다.

    # 본문
    body = element.cssselect('div#postContent')[0].text_content().strip()
    contents.append(body) # 추출한 본문을 contents 에 append 한다.
```

`titles` 와 `contents` 에 제목과 본문들이 저장된 것을 확인할 수 있다. 이제 추출한 결과를 `pandas.DataFrame` 으로 만들어보자.

```
import pandas as pd
```



pd.DataFrame 에 titles 와 contents 를 넘겨주고 컬럼 이름으로 '제목'과 '본문'으로 지정한다.

```
pd.DataFrame({'제목': titles, '본문': contents})
```

제목	본문
0      등업요청~ 다~	일반회원 등업 요청입니
1      등업요청이에요 !~	등업부탁드려요~
2      등업요청 드립니다. 도와주세요^^	정말 좋은 의자 찾기 너무 어렵네요.
3      등업요청 부탁드립니다 잘부탁드리고 등업요청합니다	최근에 허리가 안 좋아서 의자에 관심이 많아졌네요
4      등업신청합니다^^ 최근들어 의자에 관심이 많이 생겨 검색하는 도중에 우연히 의사모를 알게되어 가입까지...	
5      가입인사 다.	안녕하세요 가입했습니
6      등업요청	등업요청이요
7      등업 신청합니다 등업 부탁 드립니다~	의자 구매가 쉽지가 않네요정보 공유를 위해
8      등업요청합니다^^ 부탁드려요^^	문의글 드리고 싶습니다등업
9      등업부탁드립니다 려요	등업부탁드
10      반가워요	등업
11      등업부탁드립니다	ㅎ
12      등업 부탁드려요 ~^^	등업 부탁드립니다
13      등업요청 드립니다. ~	부탁드려요
14      등업 부탁드립니다 ^..	등업 요청합니다^

게시물의 제목과, 본문이 데이터 프레임 형태로 잘 추출된 것을 확인할 수 있다.