Python News Crawling TIL 3

2022-03-08

[3탄] 쉽게 따라하는 네이버 뉴스 크롤링 - 본문 가져오기

"본 포스팅은 네이버 뉴스의 title, url, 본문을 가져오는 크롤링을 설명하는 포스팅입니다 전 단계인 title, url 크롤링 방식을 확인하고 싶으신 분은 아래 링크(2탄)를 참고해주세요 :)" 히스토리 (2021.02.14)..

https://everyday-tech.tistory.com/entry/3%ED%83%84-%EC%89%BD%EA%B2%8C-%EB%94%B0%EB%9D%BC%ED%95%98%EB%8A%94-%EB%84%A4%EC%9D%B4%EB%B2%84-%EB%89%B4%EC%8A%A4-%ED%81%AC%EB%A1%A4%EB%A7%81-%EB%B3%B8%EB%AC%B8-%EA%B0%80%EC%A0%B8%EC%98%A4%EA%B8%B0?category=922285

TIL 3 목표

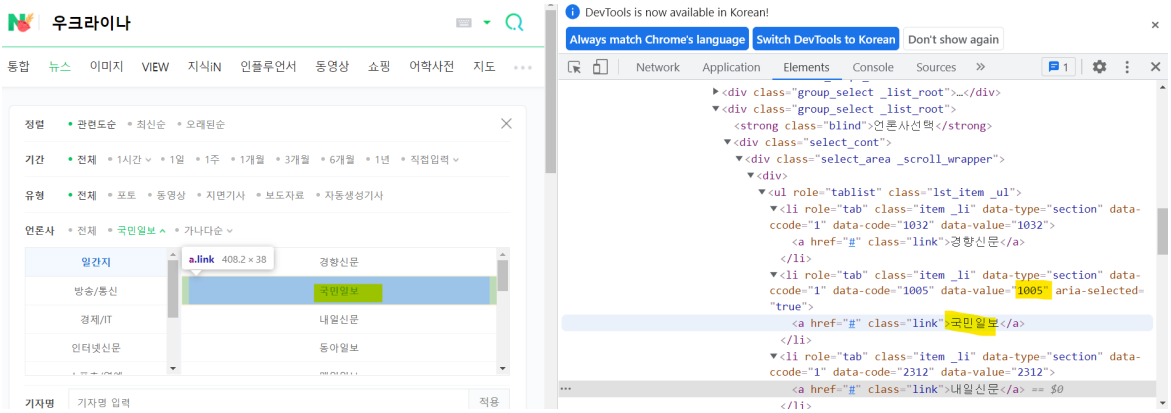
네이버 뉴스의 title, url, 본문을 가져오는 크롤링 방법 배우기

본문 크롤링이 어려운 이유와 해결 방법

- **본문 크롤링이 어려운 이유**
  - 언론사 마다 웹 페이지의 구성이 다름 (본문이 들어있는 tag, 속성 값 등등)
  - 모든 언론사의 웹 페이지 규칙을 알고 있다면 url 에 따라 어떤 언론사인지 확인하고 본문이 담긴 tag를 찾아 크롤링 가능하지만 너무 많은 시간이 들기 때문에 효율성이 떨어짐
- **해결 방법 1 : 네이버 뉴스 검색 시 제목 밑에 보이는 일부 기사 발췌**
  - 한계 : 본문 전체를 가져오지 못하기 때문에 큰 의미가 없음
- **해결 방법 2 : 일부 언론사 선택 후 기사 본문 크롤링**
  - 주로 보는 일부 언론사 선택 및 검색 후 해당 언론사의 title, url 만 크롤링
  - 언론사 별로 웹 페이지에서 본문이 위치한 tag 를 확인하고 해당 url 에 들어가서 본문 크롤링

해결 방법 2 에 관한 크롤링 방법 및 세부 수행 단계

- **크롤링 방법**
  - selenium 을 이용한 동적 크롤링을 사용하여 언론사를 선택하여 검색된 뉴스 기사를 크롤링 하는 방법 사용
- **URL 조건 확인**

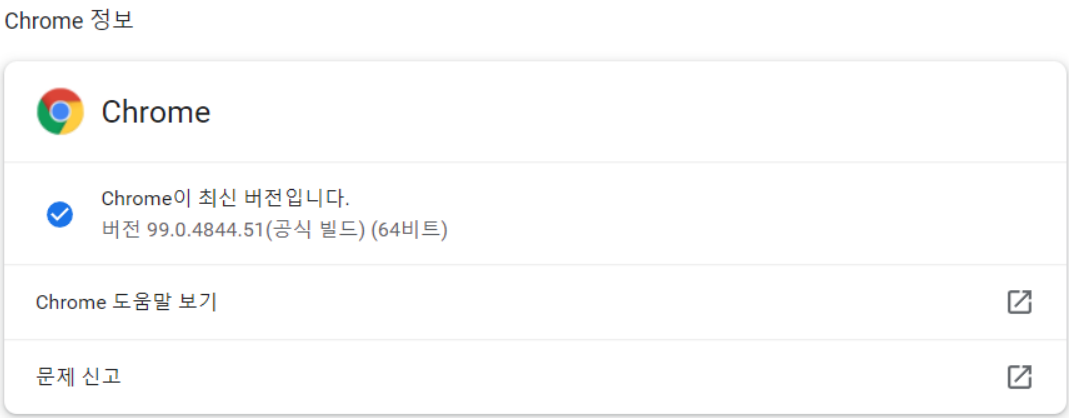


- “우크라이나” 키워드 검색 후 URL :
  https://search.naver.com/search.naver?where=news&ie=utf8&sm=nws_hty&**query=우크라이나**
- “국민일보” 언론사 선택 후 URL :
  https://search.naver.com/search.naver?where=news&query=우크라이나&sm=tab_opt&sort=0&photo=0&field=0&pd=0&ds=&de=&docid=&related=0&mynews=1&office_type=1&office_section_code=1&**news_office_checked=1005**&nso=&is_sug_officeid=0

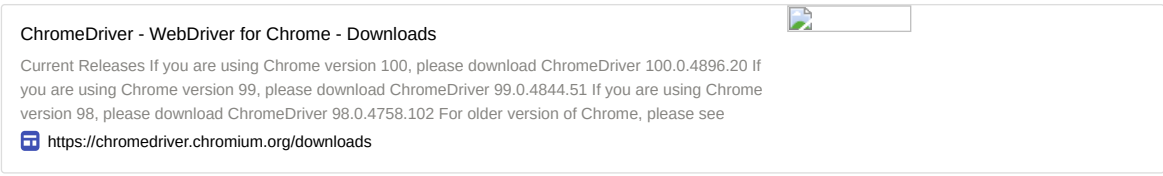
- **세부 수행 단계**
  - 환경 설정
  - STEP 1. 검색할 언론사 선택
  - STEP 2. 일부 언론사만 검색하는 기능
  - STEP 3. 선택한 언론사별 본문 tag 위치

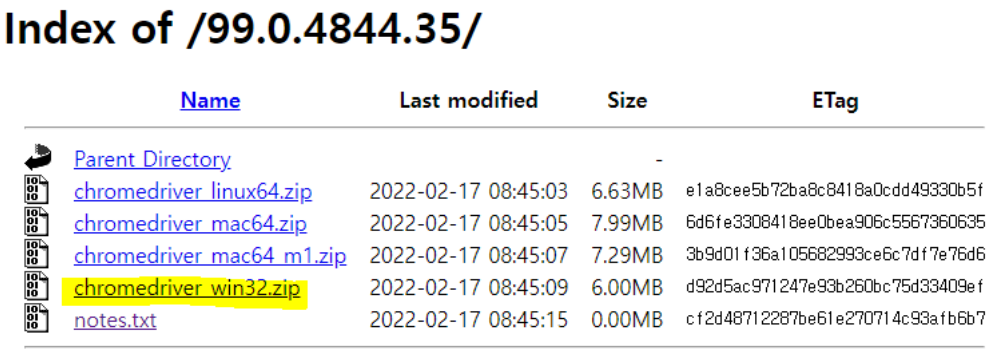
환경 설정

- **Chrome 버전 확인**



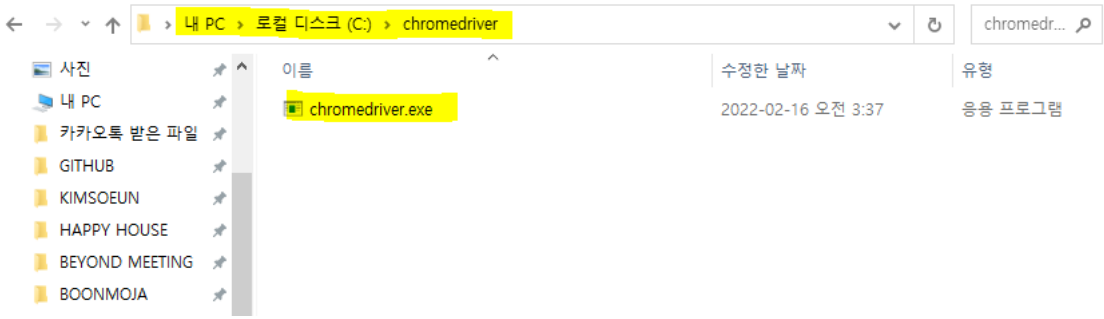
- **chromedriver.exe 설치**





Index of /99.0.4844.35/

Name	Last modified	Size	ETag
Parent Directory		-	
chromedriver_linux64.zip	2022-02-17 08:45:03	6.63MB	e1a8cee5b72ba8c8418a0cdd49330b5f
chromedriver_mac64.zip	2022-02-17 08:45:05	7.99MB	6d6fe3308418ee0bea906c5567360635
chromedriver_mac64_m1.zip	2022-02-17 08:45:07	7.29MB	3b9d01f36a105682993ce6c7df7e76d6
chromedriver_win32.zip	2022-02-17 08:45:09	6.00MB	d92d5ac971247e93b260bc75d33409ef
notes.txt	2022-02-17 08:45:15	0.00MB	cf2d48712287be61e270714c93afb6b7

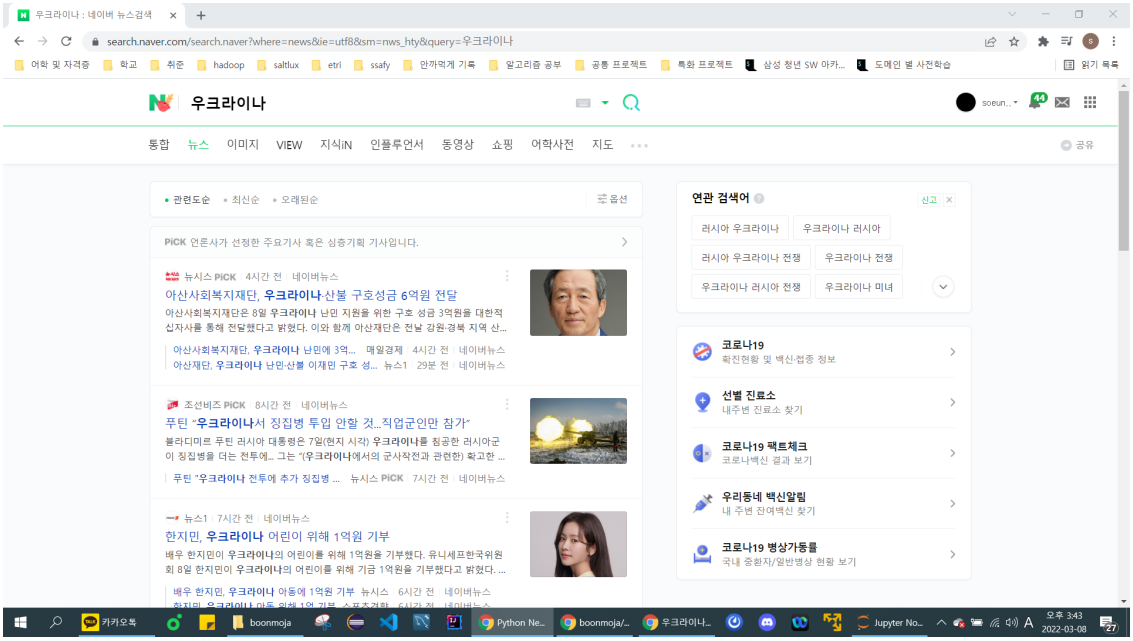


STEP 1. 검색할 언론사 선택

- **KBS**

STEP 2. 일부 언론사만 검색하는 기능

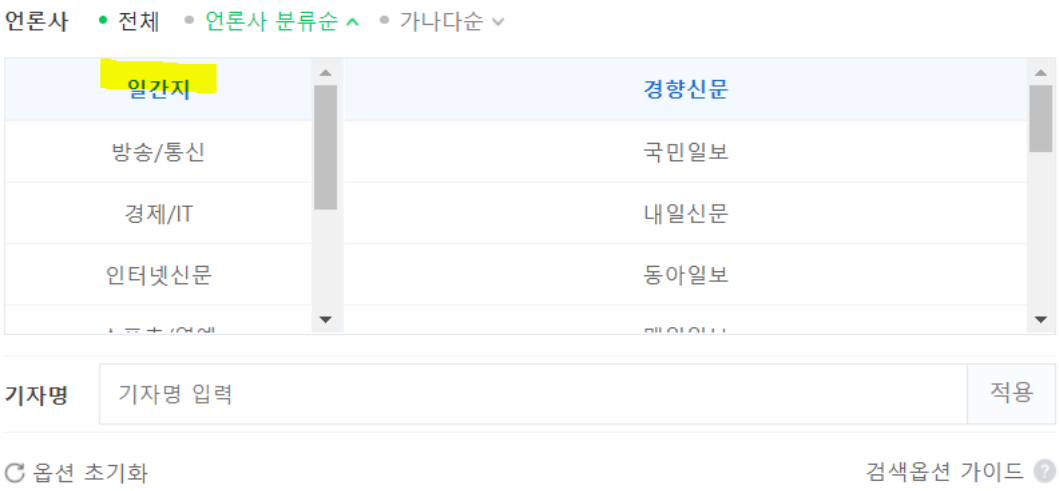
- **크롤링 방법**
  - selenium 패키지로 웹 브라우저를 제어하는 동적 크롤링을 사용
  - 언론사를 선택하여 뉴스 기사를 필터링
- **세부 수행 단계**
  - **url 에 키워드를 넣어 검색**
    - 검색 URL https://search.naver.com/search.naver?where=news&ie=utf8&sm=nws_hty&**query="키워드"**



  - **"옵션" bar 활성화**
    - 옵션 바 선택 xpath : a[@class="btn_option _search_option_open_btn"]

- **"언론사 분류순" bar 열기**
  - 전체 박스 선택 xpath : div[@role="listbox" and @class="api_group_option_sort _search_option_detail_wrap"]//li[@class="bx press"]
  - 언론사 분류순 바 선택 xpath : div[@role="tablist" and @class="option"]/a

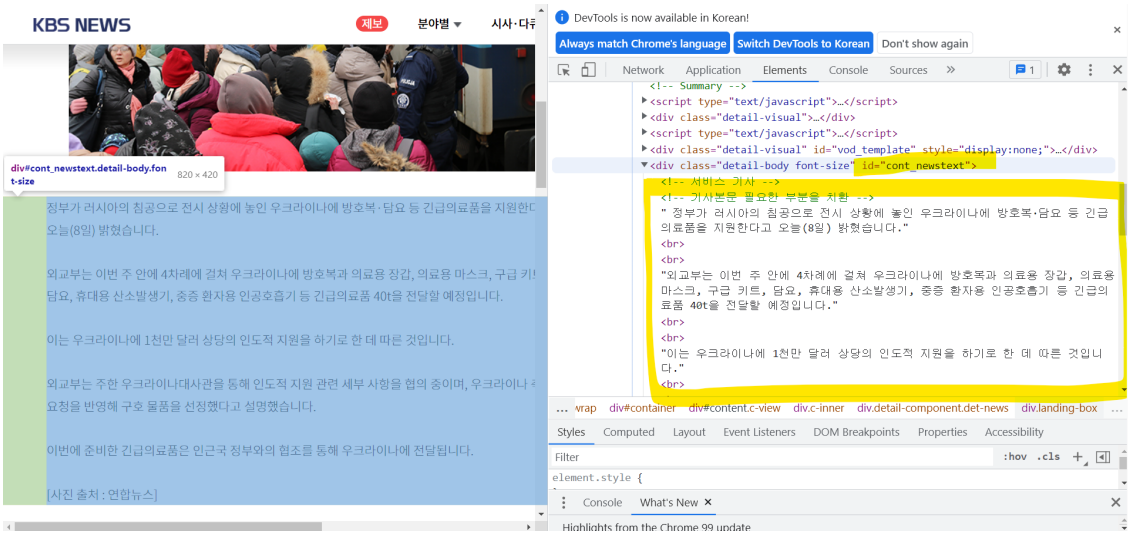
- **"언론사 종류" 선택**
  - 전체 박스 선택 xpath : div[@class="group_select _list_root"]
  - 언론사 종류 선택 xpath : ul[@role="tablist" and @class="lst_item _ul"]/li/a



- **원하는 언론사 선택**
  - 전체 박스 선택 xpath : div[@class="group_select _list_root"]
  - 언론사 선택 xpath : ul[@role="tablist" and @class="lst_item _ul"]/li/a

STEP 3. 선택한 언론사별 본문 tag 위치

- KBS
  - url : http://news.kbs.co.kr/
  - 본문을 담고 있는 태그 위치 : div[@id="cont_newstext"]



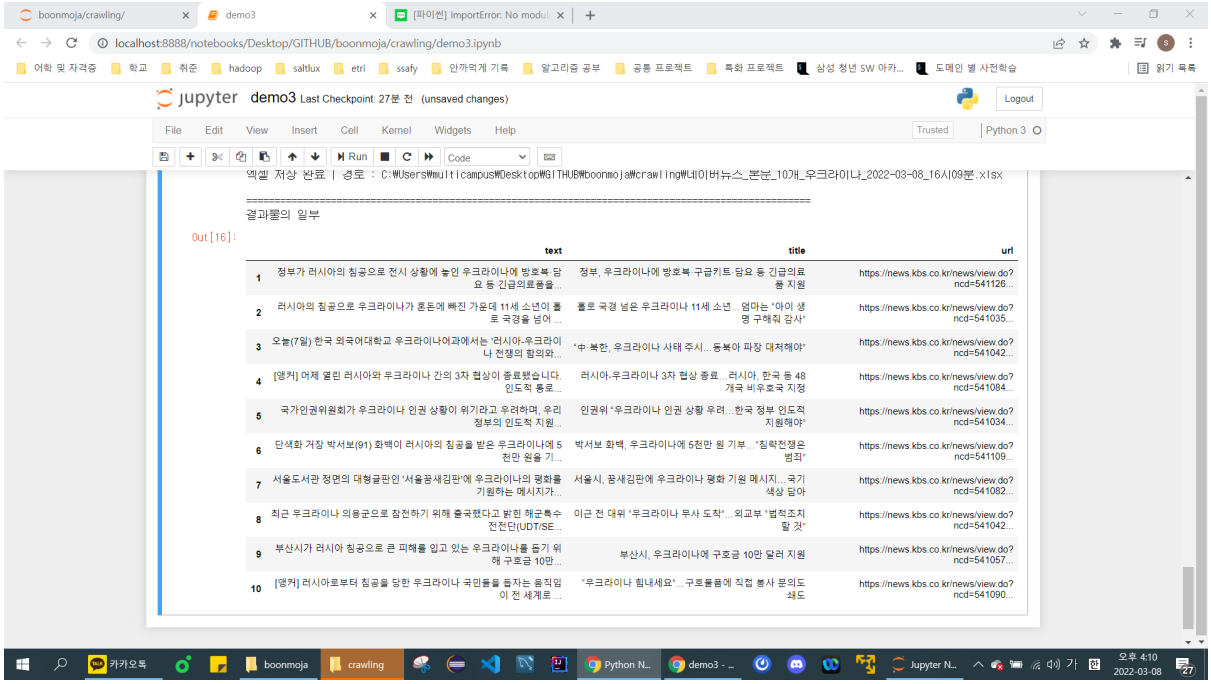
소스코드 및 결과

```
import sys, os
import requests
import selenium
from selenium import webdriver
import requests
from pandas import DataFrame
from bs4 import BeautifulSoup
import re
from datetime import datetime
import pickle, progressbar, json, glob, time
from tqdm import tqdm

###### 날짜 저장 ##########
date = str(datetime.now())
date = date[:date.rfind(':')].replace(' ', '_')
date = date.replace(':','시') + '분'

sleep_sec = 0.5


####### 언론사별 본문 위치 태그 파싱 함수 ###########
print('본문 크롤링에 필요한 함수를 로딩하고 있습니다...\n' + '-' * 100)
def crawling_main_text(url):

    req = requests.get(url)
    req.encoding = None
    soup = BeautifulSoup(req.text, 'html.parser')

    # 연합뉴스
    if ('://yna' in url) | ('app.yonhapnews' in url):
        main_article = soup.find('div', {'class':'story-news article'})
        if main_article == None:
            main_article = soup.find('div', {'class' : 'article-txt'})

        text = main_article.text

    # MBC
    elif '//imnews.imbc' in url:
        text = soup.find('div', {'itemprop' : 'articleBody'}).text

    # 매일경제(미라클), req.encoding = None 설정 필요
    elif 'mirakle.mk' in url:
        text = soup.find('div', {'class' : 'view_txt'}).text

    # 매일경제, req.encoding = None 설정 필요
    elif 'mk.co' in url:
        text = soup.find('div', {'class' : 'art_txt'}).text

    # SBS
    elif 'news.sbs' in url:
        text = soup.find('div', {'itemprop' : 'articleBody'}).text

    # KBS
    elif 'news.kbs' in url:
```

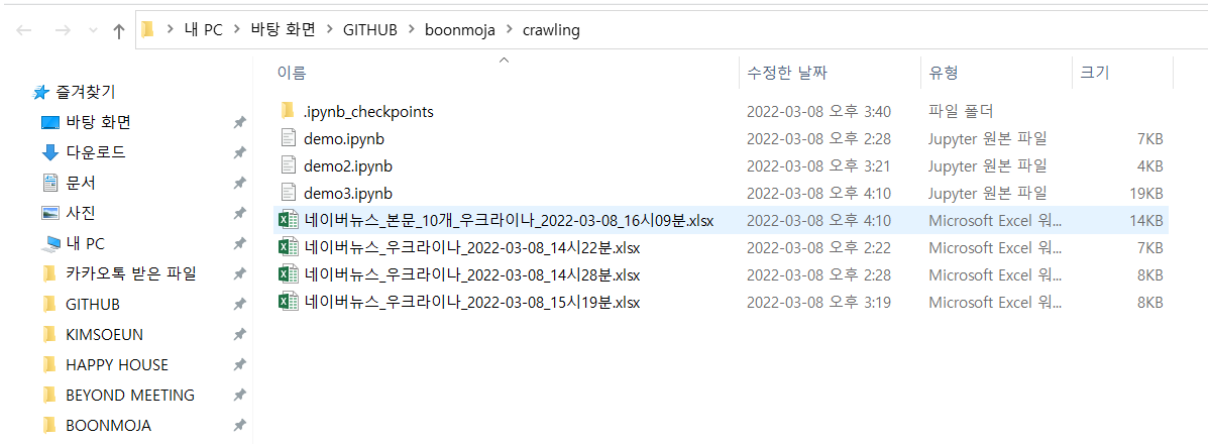
```
        text = soup.find('div', {'id' : 'cont_newstext'}).text

    # JTBC
    elif 'news.jtbc' in url:
        text = soup.find('div', {'class' : 'article_content'}).text

    # 그 외
    else:
        text == None

    return text.replace('\n','').replace('\r','').replace('<br>','').replace('\t','')


press_nm = 'KBS'

print('검색할 언론사 : {}'.format(press_nm))


############### 브라우저를 켜고 검색 키워드 입력 ####################
query = input('검색할 키워드  : ')
news_num = int(input('수집 뉴스의 수(숫자만 입력) : '))

print('\n' + '=' * 100 + '\n')

print('브라우저를 실행시킵니다(자동 제어)\n')
chrome_path = 'C:/chromedriver/chromedriver.exe'
browser = webdriver.Chrome(chrome_path)

news_url = 'https://search.naver.com/search.naver?where=news&query={}'.format(query)
browser.get(news_url)
time.sleep(sleep_sec)


######### 언론사 선택 및 confirm ####################
print('설정한 언론사를 선택합니다.\n')

search_opn_btn = browser.find_element_by_xpath('//a[@class="btn_option _search_option_open_btn"]')
search_opn_btn.click()
time.sleep(sleep_sec)

bx_press = browser.find_element_by_xpath('//div[@role="listbox" and @class="api_group_option_sort _search_option_detail_wrap"]//li[@cl

# 기준 두번 째(언론사 분류순) 클릭하고 오픈하기
press_tablist = bx_press.find_elements_by_xpath('.//div[@role="tablist" and @class="option"]/a')
press_tablist[1].click()
time.sleep(sleep_sec)

# 첫 번째 것(언론사 분류선택)
bx_group = bx_press.find_elements_by_xpath('.//div[@class="api_select_option type_group _category_select_layer"]/div[@class="select_wr

press_kind_bx = bx_group.find_elements_by_xpath('.//div[@class="group_select _list_root"]')[0]
press_kind_btn_list = press_kind_bx.find_elements_by_xpath('.//ul[@role="tablist" and @class="lst_item _ul"]/li/a')


for press_kind_btn in press_kind_btn_list:

    # 언론사 종류를 순차적으로 클릭(좌측)
    press_kind_btn.click()
    time.sleep(sleep_sec)

    # 언론사선택(우측)
    press_slct_bx = bx_group.find_elements_by_xpath('.//div[@class="group_select _list_root"]')[1]
    # 언론사 선택할 수 있는 클릭 버튼
    press_slct_btn_list = press_slct_bx.find_elements_by_xpath('.//ul[@role="tablist" and @class="lst_item _ul"]/li/a')
    # 언론사 이름들 추출
    press_slct_btn_list_nm = [psl.text for psl in press_slct_btn_list]

    # 언론사 이름 : 언론사 클릭 버튼 인 딕셔너리 생성
    press_slct_btn_dict = dict(zip(press_slct_btn_list_nm, press_slct_btn_list))

    # 원하는 언론사가 해당 이름 안에 있는 경우
    # 1) 클릭하고
    # 2) 더이상 언론사분류선택 탐색 중지
    if press_nm in press_slct_btn_dict.keys():
        print('<{}> 카테고리에서 <{}>를 찾았으므로 탐색을 종료합니다'.format(press_kind_btn.text, press_nm))

        press_slct_btn_dict[press_nm].click()
        time.sleep(sleep_sec)

        break



################ 뉴스 크롤링 ########################

print('\n크롤링을 시작합니다.')
# ####동적 제어로 페이지 넘어가며 크롤링
```

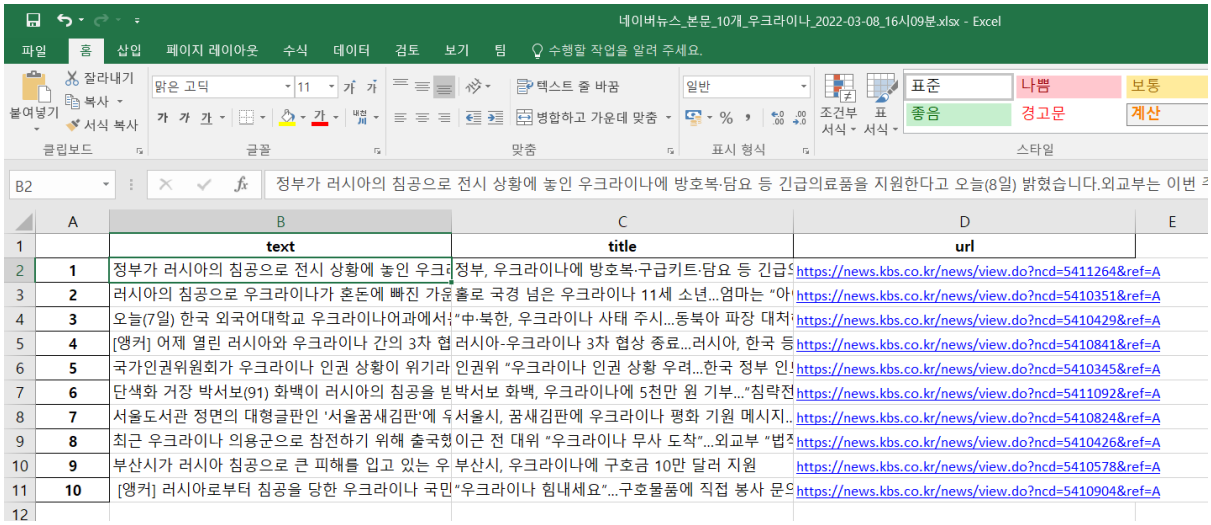
```
news_dict = {}
idx = 1
cur_page = 1

pbar = tqdm(total=news_num ,leave = True)

while idx < news_num:

    table = browser.find_element_by_xpath('//ul[@class="list_news"]')
    li_list = table.find_elements_by_xpath('./li[contains(@id, "sp_nws")]')
    area_list = [li.find_element_by_xpath('.//div[@class="news_area"]') for li in li_list]
    a_list = [area.find_element_by_xpath('.//a[@class="news_tit"]') for area in area_list]

    for n in a_list[:min(len(a_list), news_num-idx+1)]:
        n_url = n.get_attribute('href')
        news_dict[idx] = {'title' : n.get_attribute('title'),
                          'url' : n_url,
                          'text' : crawling_main_text(n_url)}

        idx += 1
        pbar.update(1)

    if idx < news_num:
        cur_page +=1

        pages = browser.find_element_by_xpath('//div[@class="sc_page_inner"]')
        next_page_url = [p for p in pages.find_elements_by_xpath('.//a') if p.text == str(cur_page)][0].get_attribute('href')

        browser.get(next_page_url)
        time.sleep(sleep_sec)
    else:
        pbar.close()

        print('\n브라우저를 종료합니다.\n' + '=' * 100)
        time.sleep(0.7)
        browser.close()
        break

#### 데이터 전처리하기 ######################################################

print('데이터프레임 변환\n')
news_df = DataFrame(news_dict).T

folder_path = os.getcwd()
xlsx_file_name = '네이버뉴스_본문_{}개_{}_{}.xlsx'.format(news_num, query, date)

news_df.to_excel(xlsx_file_name)

print('엑셀 저장 완료 | 경로 : {}\\{}\n'.format(folder_path, xlsx_file_name))

os.startfile(folder_path)

print('=' * 100 + '\n결과물의 일부')
news_df
```

헐 개신기해 미쳐따



	text	title	url
1	정부가 러시아의 침공으로 전시 상황에 놓인 우크라이나에 방호복·담요 등 긴급의료품을...	정부, 우크라이나에 방호복·구급키트·담요 등 긴급의료품 지원	https://news.kbs.co.kr/news/view.do?ncd=541126...
2	러시아의 침공으로 우크라이나가 혼돈에 빠진 가운데 11세 소년이 홀로 국경을 넘어 ...	홀로 국경 넘은 우크라이나 11세 소년...엄마는 "아이 생명 구해줘 감사"	https://news.kbs.co.kr/news/view.do?ncd=541035...
3	오늘(7일) 한국 외국어대학교 우크라이나어과에서는 '러시아-우크라이나 전쟁의 함의와...	"中·북한, 우크라이나 사태 주시...동북아 파장 대처해야"	https://news.kbs.co.kr/news/view.do?ncd=541042...
4	[앵커] 어제 열린 러시아와 우크라이나 간의 3차 협상이 종료됐습니다. 인도적 종료...	러시아-우크라이나 3차 협상 종료...러시아, 한국 등 48개국 비우호국 지정	https://news.kbs.co.kr/news/view.do?ncd=541084...
5	국가인권위원회가 우크라이나 인권 상황이 위기라고 우려하며, 우리 정부의 인도적 지원...	인권위 "우크라이나 인권 상황 우려...한국 정부 인도적 지원해야"	https://news.kbs.co.kr/news/view.do?ncd=541034...
6	단색화 거장 박서보(91) 화백이 러시아의 침공을 받은 우크라이나에 5천만 원을 기...	박서보 화백, 우크라이나에 5천만 원 기부...."침략전쟁은 범죄"	https://news.kbs.co.kr/news/view.do?ncd=541109...
7	서울도서관 정면의 대형글판인 '서울꿈새김판'에 우크라이나의 평화를 기원하는 메시지가...	서울시, 꿈새김판에 우크라이나 평화 기원 메시지...국기색상 담아	https://news.kbs.co.kr/news/view.do?ncd=541082...
8	최근 우크라이나 의용군으로 참전하기 위해 출국했다고 밝힌 해군특수전전단(UDT/SE...	이근 전 대위 "우크라이나 무사 도착"...외교부 "법적조치 할 것"	https://news.kbs.co.kr/news/view.do?ncd=541042...
9	부산시가 러시아 침공으로 큰 피해를 입고 있는 우크라이나를 돕기 위해 구호금 10만...	부산시, 우크라이나에 구호금 10만 달러 지원	https://news.kbs.co.kr/news/view.do?ncd=541057...
10	[앵커] 러시아로부터 침공을 당한 우크라이나 국민들을 돕자는 움직임이 전 세계로 ...	"우크라이나 힘내세요"...구호물품에 직접 봉사 문의도 쇄도	https://news.kbs.co.kr/news/view.do?ncd=541090...





	A	B	C	D
1		text	title	url
2	1	정부가 러시아의 침공으로 전시 상황에 놓인 우크라	정부, 우크라이나에 방호복·구급키트·담요 등 긴급의	https://news.kbs.co.kr/news/view.do?ncd=5411264&ref=A
3	2	러시아의 침공으로 우크라이나가 혼돈에 빠진 가운	홀로 국경 넘은 우크라이나 11세 소년...엄마는 "아이	https://news.kbs.co.kr/news/view.do?ncd=5410351&ref=A
4	3	오늘(7일) 한국 외국어대학교 우크라이나어과에서는	"中·북한, 우크라이나 사태 주시...동북아 파장 대처	https://news.kbs.co.kr/news/view.do?ncd=5410429&ref=A
5	4	[앵커] 어제 열린 러시아와 우크라이나 간의 3차 협	러시아-우크라이나 3차 협상 종료...러시아, 한국 등	https://news.kbs.co.kr/news/view.do?ncd=5410841&ref=A
6	5	국가인권위원회가 우크라이나 인권 상황이 위기라	인권위 "우크라이나 인권 상황 우려...한국 정부 인도	https://news.kbs.co.kr/news/view.do?ncd=5410345&ref=A
7	6	단색화 거장 박서보(91) 화백이 러시아의 침공을 받	박서보 화백, 우크라이나에 5천만 원 기부..."침략전	https://news.kbs.co.kr/news/view.do?ncd=5411092&ref=A
8	7	서울도서관 정면의 대형글판인 '서울꿈새김판'에 우	서울시, 꿈새김판에 우크라이나 평화 기원 메시지...	https://news.kbs.co.kr/news/view.do?ncd=5410824&ref=A
9	8	최근 우크라이나 의용군으로 참전하기 위해 출국했	이근 전 대위 "우크라이나 무사 도착"...외교부 "법적	https://news.kbs.co.kr/news/view.do?ncd=5410426&ref=A
10	9	부산시가 러시아 침공으로 큰 피해를 입고 있는 우	부산시, 우크라이나에 구호금 10만 달러 지원	https://news.kbs.co.kr/news/view.do?ncd=5410578&ref=A
11	10	[앵커] 러시아로부터 침공을 당한 우크라이나 국민	"우크라이나 힘내세요"...구호물품에 직접 봉사 문의	https://news.kbs.co.kr/news/view.do?ncd=5410904&ref=A
12				

자잘한 python setting

- **selenium**

```
선택 관리자: Anaconda Prompt

(base) C:₩Users₩multicampus>pip install -U selenium
Collecting selenium
  Using cached https://files.pythonhosted.org/packages/80/d6/4294f0b4bce4de0abf13e17190289f9d0613b0a44e5dd6a7f5ca9845985
3/selenium-3.141.0-py2.py3-none-any.whl
Requirement not upgraded as not directly required: urllib3 in c:₩program files (x86)₩microsoft visual studio₩shared₩anac
onda3_64₩lib₩site-packages (from selenium) (1.22)
distributed 1.21.8 requires msgpack, which is not installed.
Installing collected packages: selenium
Successfully installed selenium-3.141.0
You are using pip version 10.0.1, however version 21.3.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.

(base) C:₩Users₩multicampus>
```

```
In [9]: import selenium
        print(selenium.__version__)

        3.141.0

In [ ]:
```

- **progressbar**

```
(base) C:₩Users₩multicampus>pip install progressbar
Collecting progressbar
  Downloading https://files.pythonhosted.org/packages/a3/a6/b8e451f6cff1c99b4747a2f72
/progressbar-2.5.tar.gz
Building wheels for collected packages: progressbar
  Running setup.py bdist_wheel for progressbar ... done
  Stored in directory: C:₩Users₩multicampus₩AppData₩Local₩pip₩Cache₩wheels₩c0₩e9₩6b₩e
15206cda272ff0
Successfully built progressbar
distributed 1.21.8 requires msgpack, which is not installed.
Installing collected packages: progressbar
Successfully installed progressbar-2.5
You are using pip version 10.0.1, however version 21.3.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.

(base) C:₩Users₩multicampus>python -m pip install --upgrade pip
```

- **tqdm**

```
C:₩Users₩multicampus>pip install tqdm
Collecting tqdm
  Downloading tqdm-4.63.0-py2.py3-none-any.whl (76 kB)
     ---------------------------------------- 76.6/76.6 KB 4.1 MB/s eta 0:00:00
Collecting colorama
  Downloading colorama-0.4.4-py2.py3-none-any.whl (16 kB)
Installing collected packages: colorama, tqdm
Successfully installed colorama-0.4.4 tqdm-4.63.0

C:₩Users₩multicampus>
```

```
(base) C:₩Users₩multicampus>pip install tqdm
Collecting tqdm
  Using cached tqdm-4.63.0-py2.py3-none-any.whl (76 kB)
Collecting importlib-resources
  Downloading importlib_resources-5.4.0-py3-none-any.whl (28 kB)
Requirement already satisfied: colorama in c:₩program files (x86)₩microsoft visual studio₩shared₩anaconda3_64₩lib₩site-p
ackages (from tqdm) (0.3.9)
Collecting zipp>=3.1.0
  Downloading zipp-3.6.0-py3-none-any.whl (5.3 kB)
Installing collected packages: zipp, importlib-resources, tqdm
Successfully installed importlib-resources-5.4.0 tqdm-4.63.0 zipp-3.6.0

(base) C:₩Users₩multicampus>
```