

연금 데이터의 웹 크롤링 실전

2021년 8월

대진대학교 컴퓨터공학과 서혜선교수

jako403@daejin.ac.kr

연금 데이터의 웹 크롤링 실전

- 네이버 블로그 웹 크롤링
- 네이버 카페 웹 크롤링

1. 네이버 블로그 크롤링

- 네이버 블로그에서 연금 관련 크롤링하기
 - 기간 설정 (데이터가 많아 크롤링 하는데 시간이 많이 걸리므로 최근 1개월로만 지정함)

1. 필요한 라이브러리 import

```
: from selenium import webdriver
import requests
from bs4 import BeautifulSoup as BS
import pandas as pd
import time
import re
import pickle
```

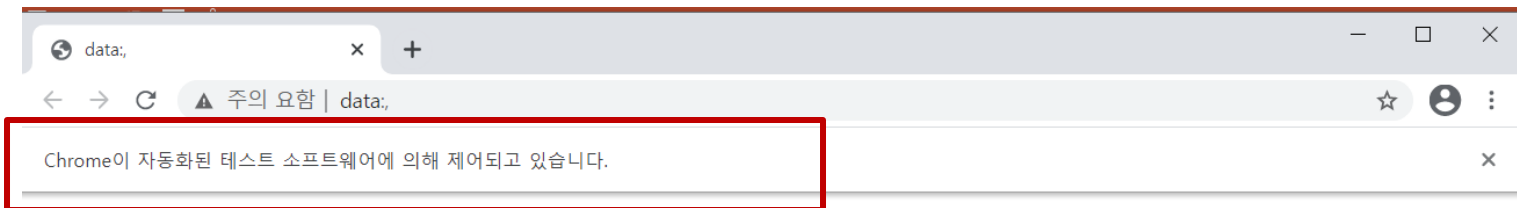
2. 검색어 지정

```
| # 다음이나 네이버의 기능상 시작일과 종료일을 별도로 지정해야 함
query_txt = '연금'
```

1. 네이버 블로그 크롤링

3. 필요 경로 지정 (크롬드라이버가 있는 경로 지정)

```
# file_path = 'C:\#chrome#chromedriver_win32\#' # 파일 경로 바꿀것  
file_path = 'C:\\\\chrome\\\\chromedriver_win32\\\\' # 파일 경로 바꿀것  
path = file_path + 'chromedriver.exe' #크롬드라이버  
driver = webdriver.Chrome(path)
```



새 크롬창이 뜨고 위와 같은 메시지가 나오면 크롤링할 수 있는 환경 완료

1. 네이버 블로그 크롤링

4. 검색하고자 하는 채널 지정 (네이버→ 블로그)

크롤링은 사람이 직접 네이버 검색창에서 진행하는 각 단계 단계를 프로그램으로 구성해 주어야 함

```
driver.get('https://www.naver.com') # 네이버를 열겠다
time.sleep(3)

#연금 검색 기능
element = driver.find_element_by_id('query') #네이버 검색창의 위치

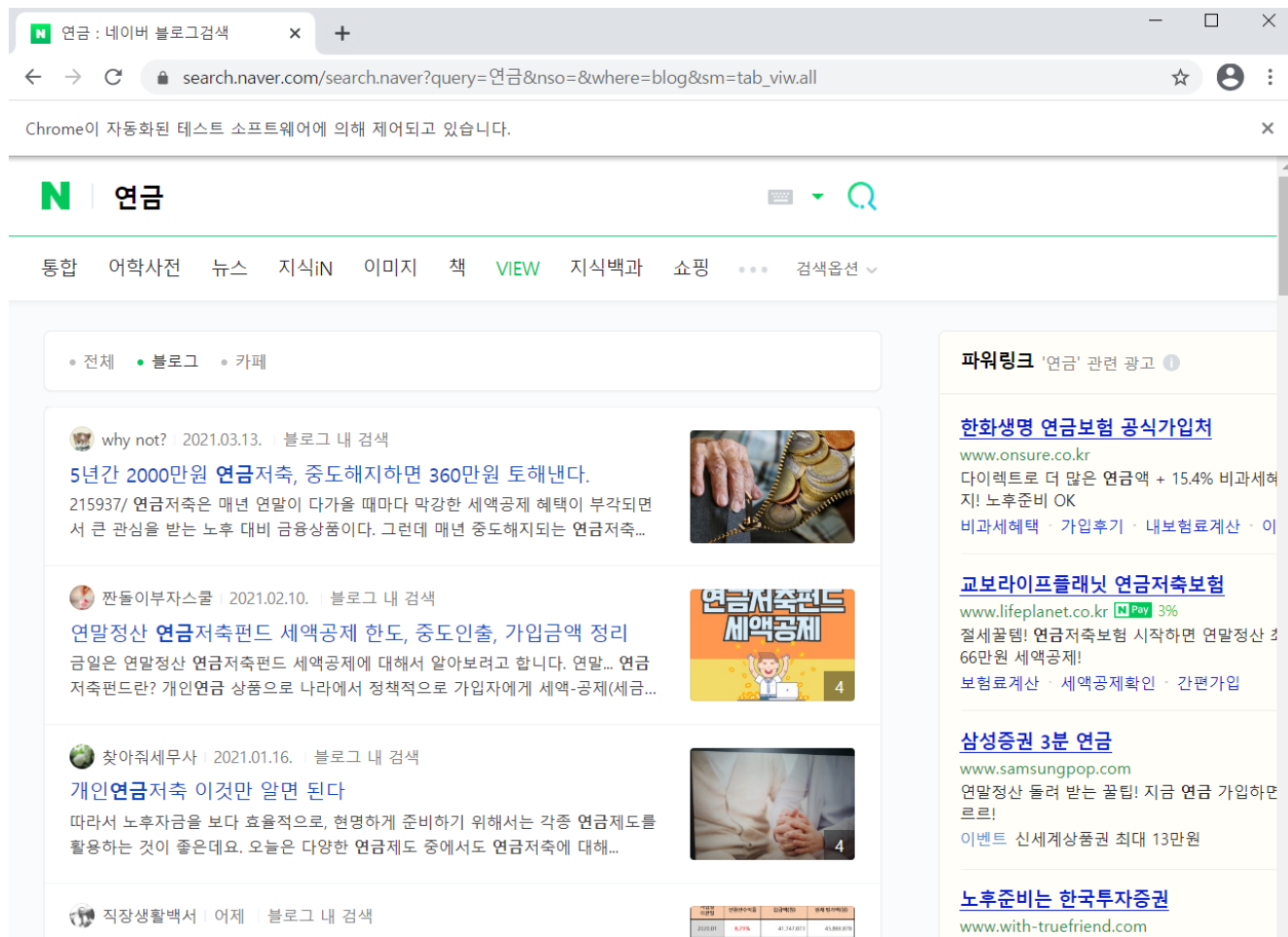
element.send_keys(query_txt) #검색창 위치에 연금을 입력
element.submit() # 검색버튼 누르기
# '연금' 검색

driver.find_element_by_link_text("블로그").click() # 네이버의 블로그 버튼 클릭
time.sleep(3)
```

네이버 검색창에 연금을 쓰고 클릭, 블로그 버튼 클릭까지의 과정


1. 네이버 블로그 크롤링

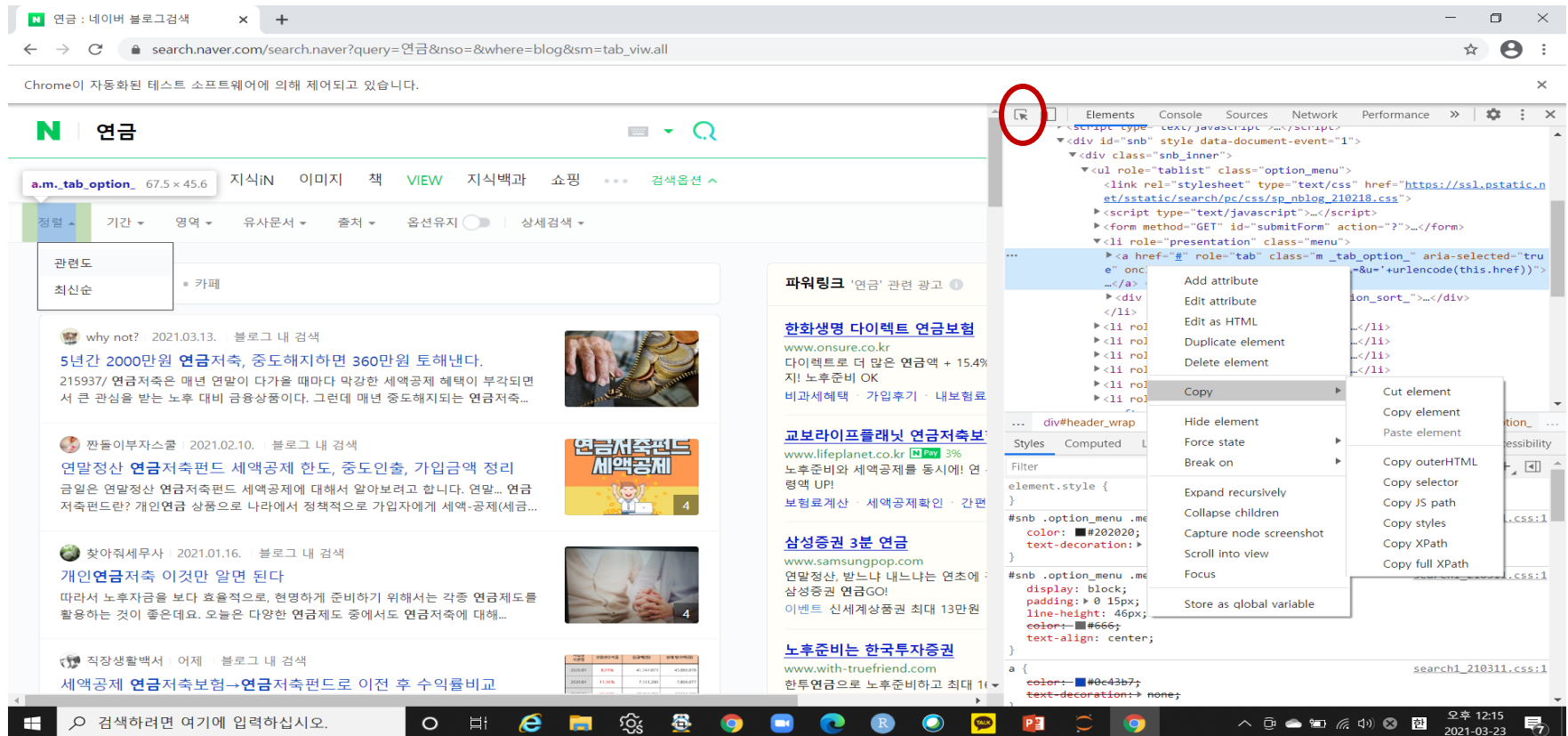
- 크롬창을 통해 앞의 과정이 어떻게 진행되고 있는지를 확인할 수 있음
- 즉 연금의 블로그까지 찾아 들어간 상태



1. 네이버 블로그 크롤링

5. 크롤링 검색어 기간 설정 방법 1 : 기간의 설정 등을 하나 하나 프로그램으로 지정하기

F12 >  안의 버튼 클릭하여 활성화 > 정렬 클릭 > html 음영부분에서 오른쪽 마우스 클릭
Copy > Copy Xpath



The screenshot shows a web browser window with the Naver search results for '연금' (pension). The search bar at the top shows the query '연금&nso=&where=blog&sm=tab_viw.all'. The search results page displays several articles related to pension, with a sidebar on the left containing filters like '정렬' (Sort), '기간' (Period), and '영역' (Area). A right-click context menu is open over the '정렬' button, showing options like 'Copy', 'Copy XPath', and 'Copy full XPath'. The browser's developer tools are also visible, showing the HTML structure of the page.

1. 네이버 블로그 크롤링

- 검색 옵션

```
driver.find_element_by_xpath("""//*[@id="snb"]/div[1]/div/div[2]/a""").click() # 네이버 메뉴중 검색옵션 클릭하기
time.sleep(3)
print("검색옵션 버튼")
```

- 기간 입력 버튼

```
driver.find_element_by_xpath("""//*[@id="snb"]/div[2]/ul/li[3]/div/div[1]/a[9]""").click()
# 오른쪽 버튼누르고 copy -> xpath(자동복사)
time.sleep(3)
print("기간의 직접입력 버튼")
```


1. 네이버 블로그 크롤링

■ 시작기간 설정

```
# 기간(date_from) 설정
driver.find_element_by_xpath("""//*[@id="snb"]/div[2]/ul/li[3]/div/div[2]/div[2]/div[1]/div/div/div/ul/li[18]""").click()
time.sleep(1)
print("년(Year)")
driver.find_element_by_xpath("""//*[@id="snb"]/div[2]/ul/li[3]/div/div[2]/div[2]/div[2]/div/div/div/ul/li[12]/a""").click()
time.sleep(1)
print("월(Month)")
driver.find_element_by_xpath("""//*[@id="snb"]/div[2]/ul/li[3]/div/div[2]/div[2]/div[3]/div/div/div/ul/li[1]/a""").click()
time.sleep(1)
print("일(Day)")
```

년(Year)
월(Month)
일(Day)

■ 시작기간 누르기

```
# 기간(date_to) 누르기
driver.find_element_by_xpath("""//*[@id="snb"]/div[2]/ul/li[3]/div/div[2]/div[1]/span[3]/a""").click()
time.sleep(1)
```

1. 네이버 블로그 크롤링

■ 종료기간 설정

```
# 기간(date_to) 설정
driver.find_element_by_xpath("//*[id='snb']/div[2]/ul/li[3]/div/div[2]/div[2]/div[1]/div/div/div/ul/li[18]/a").click()
time.sleep(1)
print("년(Year)")
driver.find_element_by_xpath("//*[id='snb']/div[2]/ul/li[3]/div/div[2]/div[2]/div[2]/div/div/div/ul/li[12]/a").click()
time.sleep(1)
print("월(Month)")
driver.find_element_by_xpath("//*[id='snb']/div[2]/ul/li[3]/div/div[2]/div[2]/div[3]/div/div/div/ul/li[31]/a").click()
time.sleep(1)
print("일(Day)")
```

년(Year)
월(Month)
일(Day)

■ 종료기간 누르기

```
# 수정 후
driver.find_element_by_xpath("//*[id='snb']/div[2]/ul/li[3]/div/div[2]/div[3]/button").click()
time.sleep(1)
print("적용")
```

적용

1. 네이버 블로그 크롤링

■ 적용 화면

Chrome이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.

N | 연금

통합 어학사전 VIEW 이미지 지식iN 동영상 쇼핑 뉴스 지도 책 ...

출처 • 전체 • 블로그 • 카페

정렬 • 관련도순 • 최신순

기간 • 전체 • 1시간 • 1일 • 1주 • 1개월 • 3개월 • 6개월 • 1년
• 2020.12.01. ~ 2020.12.31. ~

옵션 초기화

검색옵션 가이드

Hello hahaha:D | 2020.12.04.

암환자 장애연금 받기 (1년 6개월 경과시 / 금액 / 기간)

기억하실지 모르겠지만, 저희가 8월달에 '암환자 장애연금'을 신청한다고 했었던 글은 기억하시나요? 2개월이 지나가기가 끝나고 드디어 장애연금 지급이

한국투자증권 뱅키스 IRP

광고 truefriend-bankis.com

신규가입시 IRP 관리수수료 0원
펀드 보수 등 별도 발생
해외주식혜택 · 국내주식받기

한화생명 다이렉트 연금보

광고 www.onsure.co.kr

다이렉트로 더 많은 연금액 +
지! 노후준비 OK
비과세혜택 · 가입후기 · 내보

11

1. 네이버 블로그 크롤링

5. 크롤링 검색어 기간 설정 방법 2 : 검색어 기간 등을 직접 클릭하여 url로 지정하기

```
# file_path = 'C:\\\\Users\\samsung\\서울핀테크아카데미' # 작업한 파일을 저장할 경로를 미리 지정  
# file_path = 'C:\\\\Users\\samsng\\서울핀테크아카데미\\DATA' # 작업한 파일을 저장할 경로를 미리 지정
```

```
driver.get('https://search.naver.com/search.naver?where=blog&query=%EC%97%B0%EA%B8%88&sm=tab_
```



1. 네이버 블로그 크롤링

6. 해당 페이지 스크롤 정의 및 실행 (반복실행 3회만 진행)

- 1개월간 연금관련 블로그 내용은 상당히 많고, 네이버의 경우 페이지번호가 아닌 스크롤에 의해 그 다음 블로그 내용으로 이어져 감
- 따라서 스크롤로 마지막 페이지까지 내려가도록 지정하는 프로그램

```
import datetime
```

import datetime : 시간개념 라이브러리, 기본 내장 함수로 별도 설치 필요없음)

```
def doScrollDown(whileSeconds): # 스크롤 내리기
```

```
    start = datetime.datetime.now()
```

```
    end = start + datetime.timedelta(seconds=whileSeconds)
```

```
    while True:
```

```
        driver.execute_script('window.scrollTo(0, document.body.scrollHeight);')
```

```
        time.sleep(1)
```

```
        if datetime.datetime.now() > end:
```

```
            break
```

**스크롤바의 시작부터 마지막까지 실행 후 1초 뒤
반복 실행 (단 반복실행을 3회만 할것- 시간관계상)**

```
: doScrollDown(3)
```

1. 네이버 블로그 크롤링

7. 문서제목과 URL주소 리스트 가져오기

- 제목과 URL주소를 각 각 이미지가 있는 것과 없는 것으로 구분해서 가져오기
- 시간관계상 데이터를 1보다 크고 101까지(100)개만 읽는 것으로 지정.

```
title_list = []
url_list = []
for i in range(1,101):    1~100까지 총 100개 문서를 가져옴
    try:
        title = driver.find_element_by_xpath('//*[id="sp_blog_{0}"]/div/a[2]'.format(i)).text
        title2 = driver.find_element_by_xpath('//*[id="sp_blog_{0}"]/div/a[2]'.format(i))
        url = title2.get_attribute('href')

    except:
        title = driver.find_element_by_xpath('//*[id="sp_blog_{0}"]/div/div/a'.format(i)).text
        title2 = driver.find_element_by_xpath('//*[id="sp_blog_{0}"]/div/div/a'.format(i))
        url = title2.get_attribute('href')

    title_list.append(title)
    url_list.append(url)
    if (i % 10) == 0:
        print(i)
```

이미지 무

제목
url

이미지 유

제목
url

```
| len(url_list)|
```

1. 네이버 블로그 크롤링

 한화생명 블로그 | 2016.02.17. | 블로그 내 검색

대한민국의 노후를 책임질 주택연금의 잠재력!

주택연금만큼 매력적인 상품을 찾기 어렵습니다. 2014년 한국주택금융공사가 내놓은 여러 사례에서도, 높은 은행 대출 이자는커녕 기초연금에 가까운 결과가...



이미지가 있는 블로그 문서

 PNUt's SPACE | 2019.07.15. | 블로그 내

연금복권 하는법 쉽게설명

로또와 연금복권 로또 하는방법은 다들 아실테니 로또 보다 인기가 조금 떨어지지 만 많은 분들이 하고계시는 연금복권 하는법 알려드릴게요~ 연금복권은 2011년 ...



 Need not to know | 2020.12.10. | 블로그 내 검색

연금복권720 32회당첨번호입니다 (간단)

연금복권720 32 회 당첨번호를 확인해보겠습니다 32 회는 12월10일 목요일 12시20분 mbc 에서 추첨했습니다 12월 2 주차 추첨 결과를 바로 알아보겠습니다 연금...

이미지가 없는 블로그 문서

1. 네이버 블로그 크롤링

8. 새 탭 제어(doc 누적 저장)

- 클로링 과정중 가장 많은 시간이 소요되는 과정으로 임시탭을 만들어서 문서 하나하나를 저장하는 과정, 해당 문서(여기서는 100개 문서)를 다 읽어오면 종료되며 읽어오는 과정을 역시 크롬을 통해 확인 가능
- 네이버 블로그의 경우 iframe형태로 구성(페이지안에 페이지가 있는 경우)

1. 네이버 블로그 크롤링

```
new_doc = []
for i in range(len(url_list)):
    url_path = url_list[i]
    driver.switch_to_window(driver.window_handles[0]) # 탭 0번 전환 (기존탭)
    driver.execute_script("window.open('{}').format(url_path)") (url list를 순서대로 열기)
    driver.switch_to_window(driver.window_handles[1]) # 탭 1번 전환 (url list open 탭으로 전환)
    time.sleep(3)
    iframes = driver.find_elements_by_tag_name('iframe') (해당문서에 iframe이 있느냐?)
    if len(iframes) > 0:
        driver.switch_to_frame(0) # 첫번째 iframe
        html = driver.page_source # 개발자도구 불러옴
        soup = BS(html, 'html.parser') # 불러온 개발자 도구 사용 (개발자 도구 내용 보여줌)
        try:
            a = soup.find('div', class_='se-main-container').get_text() (본문에 해당되는 태그 불러오기)
        except :
            a = soup.find('div', id='postListBody') (본문에 해당되는 태그 불러오기)
            a = re.sub("[^ㄱ-ㅎㅌ-ㅣ가-힣 ]", "", str(a)) (한글만 불러오라고 지정)
        new_doc.append(a)
        driver.switch_to_default_content() (해당페이지의 frame에서 빠져나오기)

    else:
        new_doc.append(' ')
driver.close() # 탭 종료
time.sleep(3)
print(i)
```

1. 네이버 블로그 크롤링

```
0  
1  
2  
3  
4  
5  
6  
7
```

읽어온 문서의 숫자를 뿌려줌.
모두 100개 문서내용을 읽어오면 종료
new_doc

```
95  
96  
97  
98  
99
```

1. 네이버 블로그 크롤링

9. 타이틀, 문서내용, url주소, 채널 정보를 각 각의 데이터로 구성

```
raw_data = pd.DataFrame()
raw_data['title'] = title_list
raw_data['doc'] = new_doc
raw_data['url'] = url_list
raw_data['ch'] = 'naver' #네이버
raw_data['ch2'] = 'blog' #블로그
```

```
len(raw_data['title'])
```

100

```
len(raw_data['doc'])
```

100

1. 네이버 블로그 크롤링

10. raw_data를 데이터 파일로 저장하기 (확장자 : pkl(피클))

```
| #파일 저장
file_path='C:Userssamsung서울핀테크아카데미DATA'
f = open(file_path + "naver_blog.pkl", "wb")
pickle.dump(raw_data, f)
f.close()
```

pkl 파일 저장시 필수
wb : write binary
rb : read binary

*** 해당 폴더에 데이터파일이 생성되었는지 확인**

```
driver.quit() # 크롤 종료
```

1. 네이버 블로그 크롤링

11. raw_data 불러오기

```
f = open(file_path + "naver_blog.pkl", "rb")
temp_file = pickle.load(f)
f.close()

temp_file
```

12. 크롤링데이터 결과

22

2. 네이버 카페 크롤링

1. 필요 라이브러리 импорт

```
from selenium import webdriver
from bs4 import BeautifulSoup as BS
import pandas as pd
import time
import re
import pickle
```

2. 필요 경로 지정 (크롬드라이버가 있는 경로 지정)

```
#파일 경로설정
file_path = 'C:\\\\chrome\\\\chromedriver_win32 (2)\\\\' # chromedriver 있는 파일로 경로 설정하기
path = file_path + 'chromedriver.exe'
driver = webdriver.Chrome(path)
```

2. 네이버 카페 크롤링

3. 검색어, 검색기간 지정 방법 1 : 하나 하나 직접 지정하기

```
query_txt = '연금'
```

블로그에서와 동일하게 1개월간만 크롤링

- 드라이버로 네이버 열어 연금 검색하기

```
#드라이버로 네이버 열기
driver.get('https://www.naver.com')
time.sleep(3)

#연금 검색하기
element = driver.find_element_by_id('query')
element.send_keys(query_txt) #키 입력.
element.submit() # 제출(enter 기능)
```


2. 네이버 카페 크롤링

- 연금내용 중 카페 내용만 보기 위해 카페 클릭

```
#검색된 연금 내용중에 카페를 클릭  
driver.find_element_by_link_text("카페").click() #카페라고 설정되어진 링크 클릭  
time.sleep(3)
```

- 절대경로 Xpath를 이용하여 옵션 설정

```
driver.find_element_by_xpath("''/*[@id="snb"]/div[1]/div/div[2]/a''").click() # 네이버 메뉴중 검색옵션 클릭하기  
time.sleep(3)  
print("검색옵션 버튼")
```

2. 네이버 카페 크롤링

 연금

통합 어학사전 뉴스 VIEW 이미지 지식iN 동영상 쇼핑 지도 책 ...

- 전체
- 블로그
- 카페

응답 옵션

 교육공무직. 교육공무직원. 특수... | 4일 전

연금보험 추가납입 효율적인 가입방법!
그렇다보니 연금보험이나 각종 금융상품을 찾는분들이 늘어나고 있습니다. 특히 저 축성연금보험에 많은 분들이 가입하고 있는데요. 하지만 그냥 가입만해놓고 연금...

RE 연금보험 추가납입 확인해보니 꼭 준비해
놔야겠다는 생각이 생기네요. 연금보험 준비...

RE 연금보험 추가납입 알아보는 지인분들이
많이 있던데 보장범위, 보장한도, 보험료 보...

 텐인텐 대전세종 | 2일 전

혹시 국민연금 잘 아시는 분..계실까요??
올 6월달이 국민연금 끝이래요.. 근데 직장생활을 늦게 시작하셔서 이제 108개월 내셨고... 50개월 미납 된걸 추
가납입하면 국민연금을 65세 이후에 매월 45만원씩 받을 수 있다는데... 추가납입 금액이...

RE 국민연금공단에 문의하는게 가장 빠릅니
다. 한꺼번에 납부하지 않고 향후 남은 5년정...

RE 지금 60세 이시면 60세부터 받지 않나요?
현재 1969년생 이후가 65세부터 수령일텐데요

보험료계산 & 비교견적

**연금보험
분석하기** 4

26

2. 네이버 카페 크롤링

- Xpath를 이용한 기간설정

```
driver.find_element_by_xpath("//*[id='snb']/div[2]/ul/li[4]/div/div[1]/a[9]").click() # 오른쪽 버튼누르고 copy -> xpath(자동복사)
time.sleep(3)
print("기간의 직접입력 버튼")
```

기간의 직접입력 버튼

2. 네이버 카페 크롤링

- 시작일, 종료일 클릭 및 입력하고 적용 실행

```
# 기간(date_from) 설정
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[2]/div[1]/div/div/div/ul/li[31]/a").click()
time.sleep(1)
print("년(Year)")
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[2]/div[2]/div/div/div/ul/li[12]/a").click()
time.sleep(1)
print("월(Month)")
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[2]/div[3]/div/div/div/ul/li[1]/a").click()
time.sleep(1)
print("일(Day)")
```

년(Year)
월(Month)
일(Day)

```
# 기간(date_to) 누르기
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[1]/span[3]/a").click()
time.sleep(1)
```

```
# 기간(date_to) 설정
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[2]/div[1]/div/div/div/ul/li[18]/a").click()
time.sleep(1)
print("년(Year)")
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[2]/div[2]/div/div/div/ul/li[12]/a").click()
time.sleep(1)
print("월(Month)")
driver.find_element_by_xpath("//*[@id='snb']/div[2]/ul/li[4]/div/div[2]/div[2]/div[3]/div/div/div/ul/li[31]/a").click()
time.sleep(1)
print("일(Day)")
```

년(Year)
월(Month)
일(Day)

2. 네이버 카페 크롤링

출처

• 전체

• 블로그

• 카페

대상

• 전체글

• 거래글

• 일반글

• 카페명

정렬

• 관련도순

• 최신순

기간

• 전체

• 1시간

• 1일

• 1주

• 1개월

• 3개월

• 6개월

• 1년

• 2020-12-01 - 2020-12-31 ▾

결제방법

• 전체

• 안전거래

• 직접거래

거래상태

• 전체

• 판매중

• 판매완료

⌂ 옵션 초기화

검색옵션 가이드 ?

2. 네이버 카페 크롤링

4. 검색어, 검색기간 지정 방법 2 : 해당 url로 지정하기

```
driver.get('https://search.naver.com/search.naver?where=article&query=%EC%97%B0%EA%B8%88&ie=utf8&st=rel&date_option=99%')
```



2. 네이버 카페 크롤링

5. 마지막 문서까지 스크롤 내리는 함수

```
# 관련 문서가 많으므로 문서의 마지막까지 스크롤 내리는 함수(네이버 카페의 구조가 스크롤 형식)
import datetime

def doScrollDown(whileSeconds): #def는 함수를 정의하는 명령어 푸른색 : 사용자가 만든 함수
    start = datetime.datetime.now() # 현재 시간
    end = start + datetime.timedelta(seconds=whileSeconds) # 현재시간 + 입력한 초
    while True:
        # WINDOW.SCROLLTO(X, Y) X는 문서왼쪽상단부터 시작하는 가로축
        # y는 문서 왼쪽 상단부터 시작하는 세로축
        time.sleep(1) # 1초에 한번씩 스크롤다운하러
        driver.execute_script('window.scrollTo(0, document.body.scrollHeight);') # X, Y에 해당하는 범위를 스크롤
        if datetime.datetime.now() > end: #시간 지나면(120초가 지나가면 끝내라) 종료
            break

doScrollDown(60) #1분동안 스크롤 다운해라 |
```

2. 네이버 카페 크롤링

6. 제목리스트와 url 리스트 가져오기

(네이버 카페는 이미지 유무에 따른 xpath가 동일)

```
| title_list = [] #제목 리스트를 가져오기 위한 빈리스트 생성
  url_list = [] #주소를 가져오기 위한

  for i in range(1, 101): #100번 반복, 100개의 데이터만 가져와라

      title = driver.find_element_by_xpath('//*[@id="_view_review_body_html"]/div/more-contents/div/ul/li[{}]/div[1]/div/a'.format(i))
          # title 은 제목에 해당하는 텍스트만을 title에 저장하기
      title2 = driver.find_element_by_xpath('//*[@id="_view_review_body_html"]/div/more-contents/div/ul/li[{}]/div[1]/div/a'.format(i))
          # title2는 텍스트와 제목의 url주소인 하이퍼링크를 함께 가져와서 저장
      url = title2.get_attribute('href') #href : 하이퍼링크 연결할 주소 지정.

      title_list.append(title) # title 리스트에 title 순차적으로 저장
      url_list.append(url)    # url 리스트에 url 순차적으로 저장
      if (i % 10 == 0):
          print(i)
```


2. 네이버 카페 크롤링

7. 새 탭 제어하기 : 임시 탭을 만들어 그 안에 문서 본문내용들을 담기

```
# 새 탭 제어
new_doc = [] # 가져올 본문(doc) 내용을 담는 리스트 지정

for i in range(0, len(url_list)):
    url_path = url_list[i]
    driver.switch_to_window(driver.window_handles[0]) # 탭 0번 전환
    driver.execute_script("window.open('{}').format(url_path)") # url_path에 있는 주소 열기
    driver.switch_to_window(driver.window_handles[1]) # 탭 1번 전환
    # 탭 1 번 : 두 번째 탭으로 전환
    # 탭 0 번 : 최초의 탭으로 전환
    # 탭 -1 번 : 가장 최근 열린 탭으로 전환

    time.sleep(3)

# Selenium에서 페이지 내 페이지 내용을 가져오고자 할때 iframe을 사용
iframes = driver.find_elements_by_tag_name('iframe') #html에 태그가 iframe인거 찾기

if len(iframes) > 0: #iframe 있는 경우만
    driver.switch_to.frame('cafe_main')
    html = driver.page_source #개발자도구
    soup = BS(html, 'html.parser') #개발자도구
    try:
        a = soup.find('div', class_='article_viewer').get_text() #iframe 속 container에서 text만 추출
    except:
        a = 'null'
    new_doc.append(a) #new_doc 리스트에 저장
    driver.switch_to_default_content() #프레임 전환하기

else: #없는경우
    a = 'null'

driver.close()
time.sleep(3)
if i%10 == 0:
    print(i)
```

2. 네이버 카페 크롤링

8. 데이터 통합하기 (제목, url, 문서 내용 등) 하여 pickle 확장자로 저장하기

```
#데이터 모으기(종류별로)
raw_data = pd.DataFrame() #크롤링한 데이터를 판다스 데이터프레임형식으로 저장
raw_data['title'] = title_list #제목
raw_data['doc'] = new_doc #내용
raw_data['url'] = url_list #url
raw_data['ch'] = 'naver' #네이버
raw_data['ch2'] = 'cafe' #카페
```

2. 네이버 카페 크롤링

9. 데이터파일 저장하기

```
#파일 저장  
f = open(file_path + "naver_cafe.pkl", "wb")  
pickle.dump(raw_data, f)  
  
f.close()
```

```
#실행 종료 전 확인 요망
```

```
driver.quit() #크롬 드라이버 종료
```

2. 네이버 카페 크롤링

10. 저장된 파일 확인

```
f = open(file_path + "naver_cafe.pkl", "rb")
temp_file = pickle.load(f)
f.close()

temp_file
```

2. 네이버 카페 크롤링

11. 저장 데이터 결과

	title	doc	url	ch	ch2
0	퇴직연금 수령방법 dc형 db형 장단점 4가지	퇴직연금 수령방법 dc형 db형 장단점 4가지 퇴직연금을 수령하기 위해서는 dc형퇴...	https://cafe.naver.com/chokingwang/59132?art=Z...	naver	cafe
1	연금보험 어떻게 할까요?	보험이 안쓰면 제일좋은거고 말그대로 보험 인건데저는 부모님부터 보험을 잘 안드는 집...	https://cafe.naver.com/hotellife/1548526?art=Z...	naver	cafe
2	퇴직연금 DC형 DB형 차이 3가지 주의점	퇴직연금 DC형 DB형 차이 3가지 주의점고형화 시대가 점점더 가까워지면서 이젠나...	https://cafe.naver.com/yuasam/366956?art=ZXh0Z...	naver	cafe
3	국민연금 수령 즈음하여	이번 달 은퇴를 합니다. 이제 매월 받을 수 있는 근로소득이 끊긴 반면 다음 달부...	https://cafe.naver.com/dlxogns01/20451?art=ZXh...	naver	cafe
4	투석 후 신청할 것들(연금 및 행정업무)	안녕하세요 저는 아빠가 10년 넘게 신장이 안좋아서 약 드시다가 투석으로 넘어가게 ...	https://cafe.naver.com/tlswkd/171449?art=ZXh0Z...	naver	cafe
...
95	장애연금 신청이 되었습니다.	2월 14일 장애연금 신청했었는데 담당자가 대략 2개월 걸린다고 하더라 오늘 두달 ...	https://cafe.naver.com/tlswkd/177761?art=ZXh0Z...	naver	cafe
96	개인형 irp 퇴직연금 장단점 4가지	개인형 irp 퇴직연금 장단점 4가지최근 재테크에 관해서 관심을 갖고계시는 분들이...	https://cafe.naver.com/chokingwang/56176?art=Z...	naver	cafe
97	15년차 현직자가 직접 공개하는 퇴직연금 예상액(현재기준)	안녕하세요 이번에 새로 멘토로 합류하게된 서민입니다저는 모두의 관심사인 공무원연금에...	https://cafe.naver.com/m2school/2505920?art=ZX...	naver	cafe
98	연금복권720 첫회에 최대당첨자 발생했네요 ㄷㄷ	연금복권이 월 700씩 20년간 지급되는걸로 바뀌면서 2등당첨자는 1억 지급이 아니...	https://cafe.naver.com/xst/425770?art=ZXh0ZXJu...	naver	cafe
99	IRP와 개인연금 궁금한 사항이 있습니다!	1. IRP퇴직연금과 개인연금 모두제가 납입한 금액+이자에 대해서 일정기간 동안 수...	https://cafe.naver.com/onepieceholicplus/25189...	naver	cafe

100 rows × 5 columns