네이버 블로그에서 css_selector로 크롤링하기

0. css_selector(선택자)란

원래는 웹 디자인에서 웹 페이지 상의 특정 요소를 선택하여 스타일을 적용하는 스타일 범위적용 기능을 수행하는 것
우리는 웹 크롤링할 때 데이터 주소를 확인하는 용도로 사용합니다.

선택자 종류 :

* : HTML 페이지 내부의 모든 태그를 선택

ex) * { ~; ~; ~; }

태그명O : 태그명이 O인 특정 태그를 선택합니다. O은 h, p, a 등등 모든 특정 태그를 포함합니다.

ex) p { ~; ~; ~; }

.class : 클래스 속성값이 특정 classname으로 지정된 경우에 해당 요소를 선택합니다.

ex) .ft_accesstermd { ~; ~; ~; }

#id : id 값이 특정 id으로 지정된 경우에 해당 요소를 선택합니다.

ex) #ft_share_box { ~; ~; ~; }

아래 패턴에 대한 임시 예시 -> **부모** : section / **자식** : p / **자손 1** : li / **자손 2** : ul

부모요소 자손요소 : 부모요소의 자손 요소를 선택합니다.

ex) section ul { ~; ~; ~; }

부모요소>자식요소 : 부모요소의 자식인 F요소를 선택합니다.

ex) section>p { ~; ~; ~; }

요소1+요소2 : 요소1을 따르는 요소2를 선택합니다.

ex) h1+ul{ ~; ~; ~; }

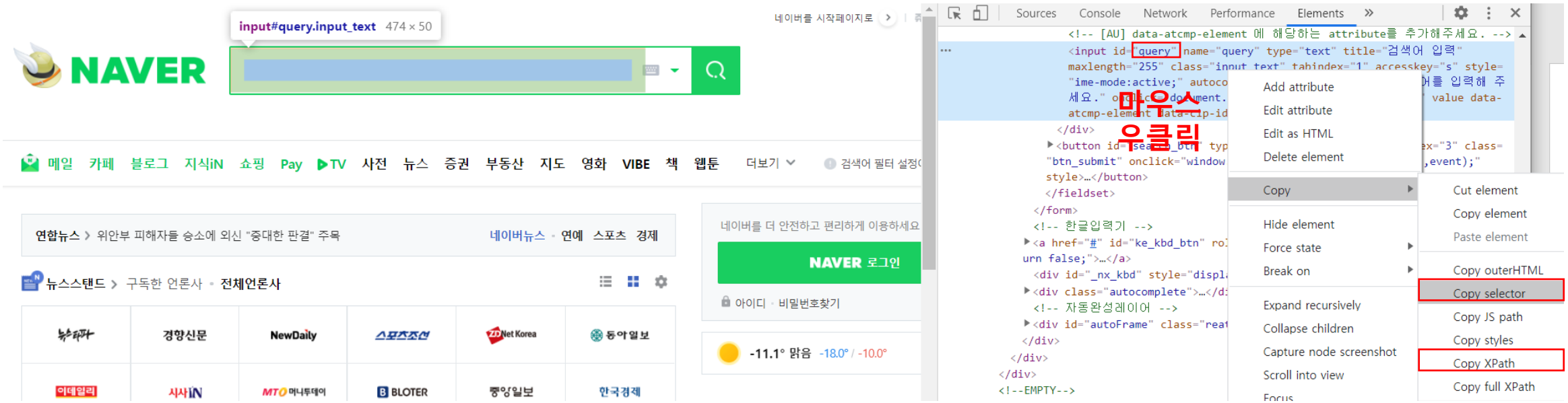
요소1-요소2 : 요소1가 앞에 존재하면 요소2를 선택합니다.

ex) h1+ul{ ~; ~; ~; }

0. css_selector(선택자)란

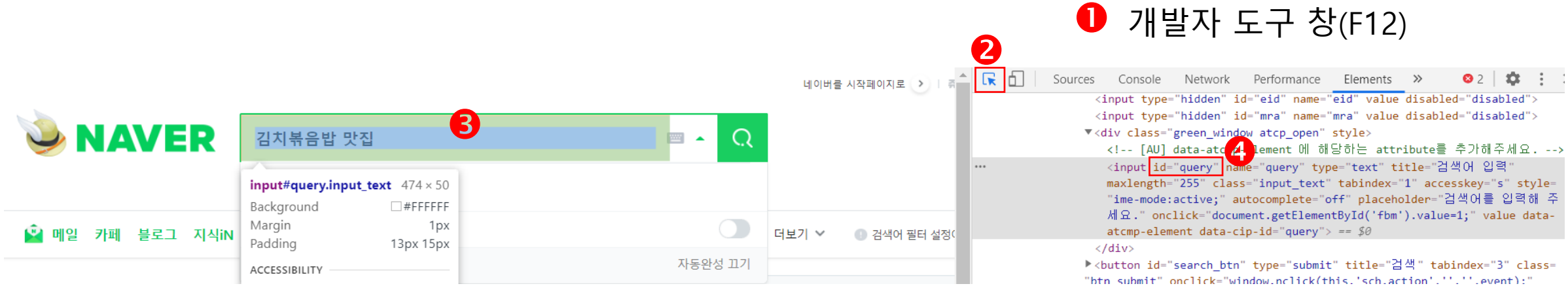
웹 페이지의 요소를 찾는 기본 공식
driver.find_element_by_css_selector(#id or .class)

기본은 find_element_by_css_selector를 활용해서 id, class 값으로 찾는다.



<응용 방법들>
driver.find_element_by_link_text(텍스트)
driver.find_element_by_xpath(xpath)
driver.find_element_by_id(id)
driver.find_element_by_name(name)

1. 네이버 검색창 주소를 찾아서 검색어 입력하기



네이버 크롬 창을 띄우고

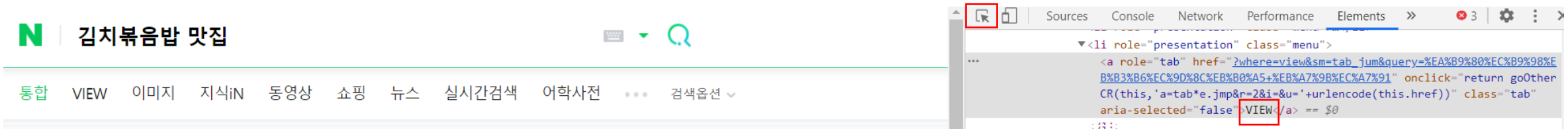
❶ F12를 눌러서 개발자 도구 창을 띄운다.
❷ 마우스로 화살표를 누른다.
❸ 마우스를 네이버 검색창에 가져다 댄다
❹ 검색창의 id값(주소값)을 확인한다.

-> 네이버 검색창 주소 id값은 'query'

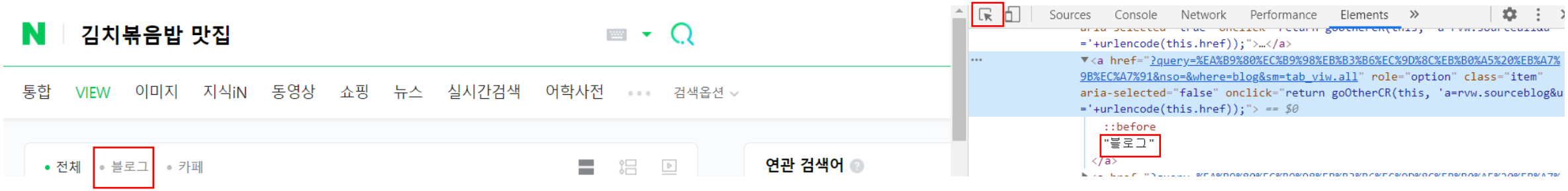
```
#Step 2. 네이버 검색창에 "검색어" 검색
element = driver.find_element_by_id("query")
```

2. 'VIEW' 버튼 누르고 'blog'버튼 누르기

여기서는 find_element_by_link_text 를 이용해서 찾는다.



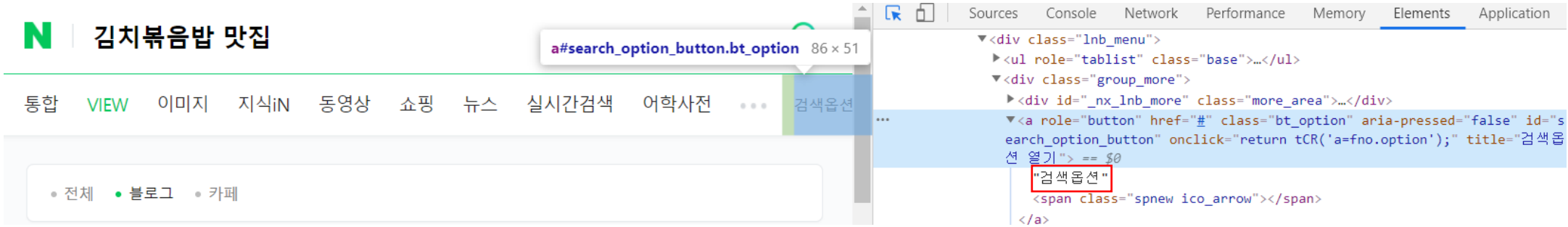

```
# 'VIEW' 클릭
driver.find_element_by_link_text("VIEW").click( )
```




```
# '블로그' 클릭
driver.find_element_by_link_text("블로그").click( )
```

3. "검색옵션" 누르기

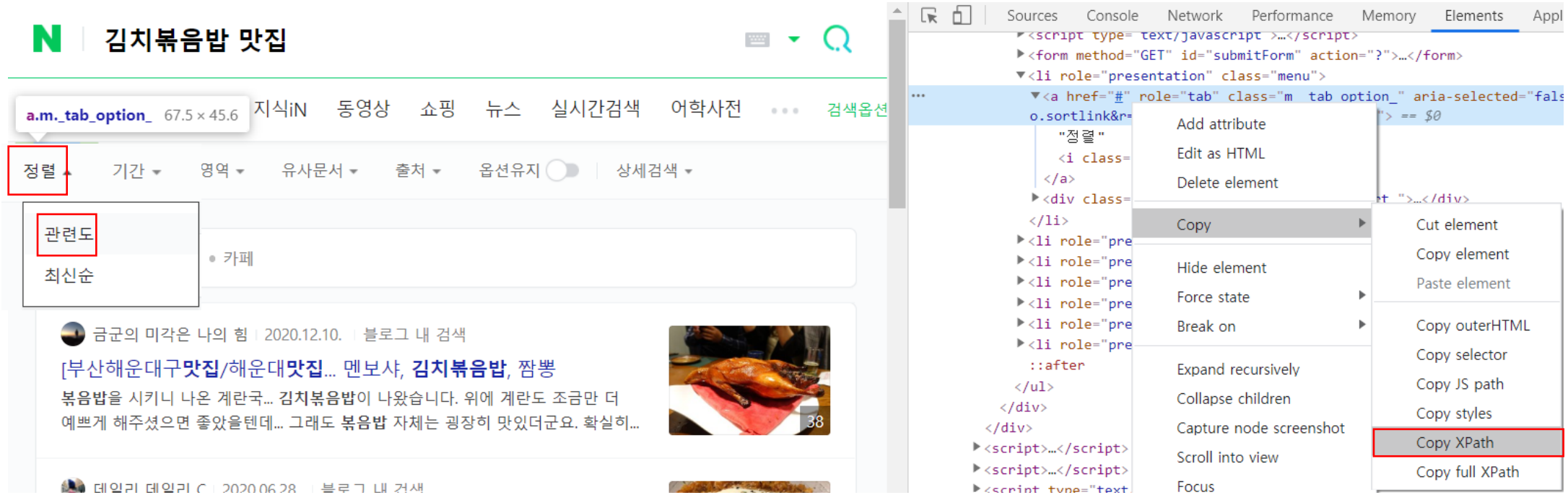
id, class 값이 없다. find_element_by_xpath를 이용해서 찾는다.




```
# '검색옵션' 클릭
driver.find_element_by_link_text("검색옵션").click( )
```

4. "정렬" 누르고 "관련도순" 누르기

id, class 값이 없다. find_element_by_xpath 를 이용해서 찾는다.



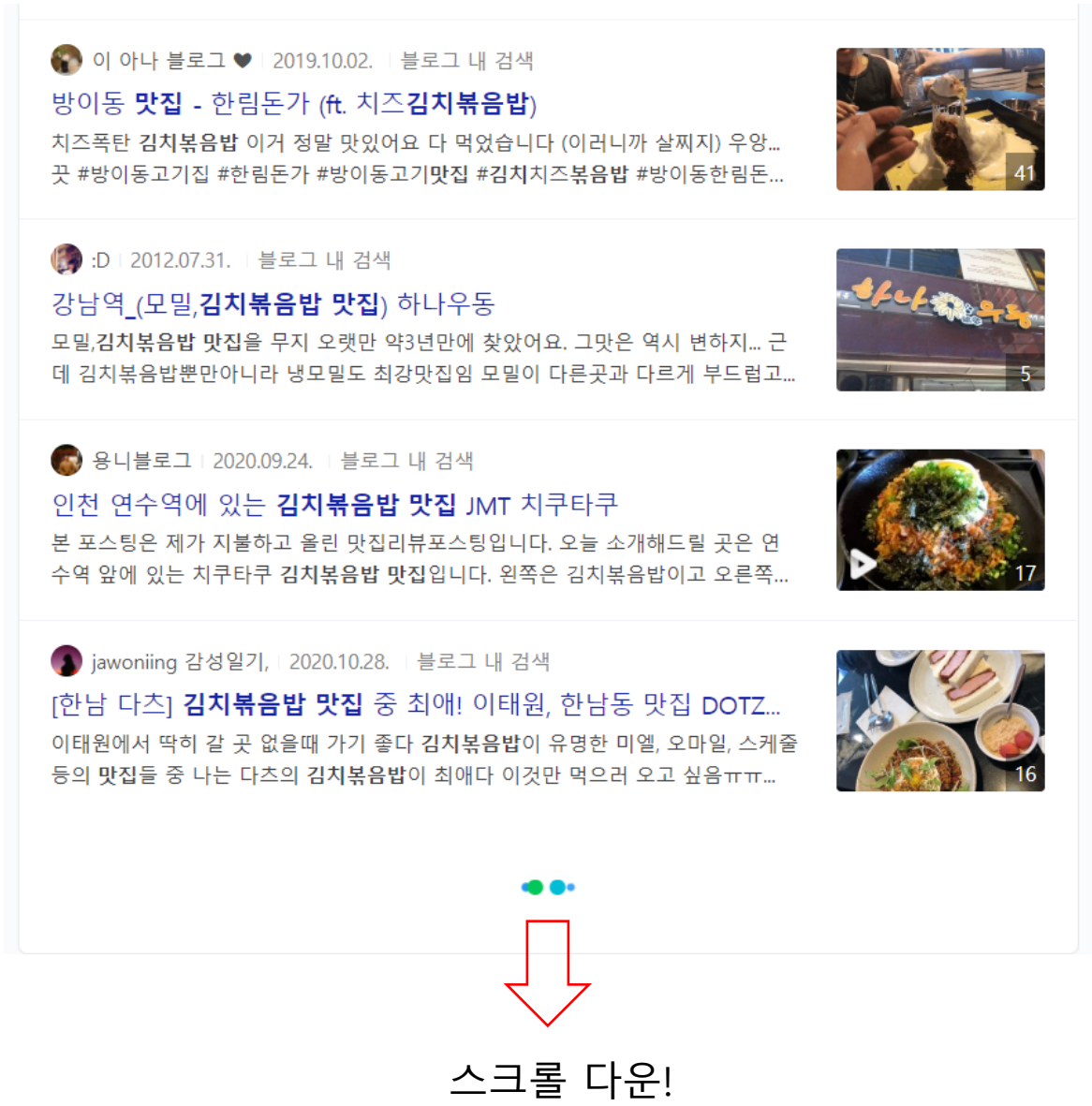

```
# Step 3. 정렬 : "관련도순"
# 개발자 도구에서 정렬 버튼의 id 가 보이지 않습니다.
# 이럴 경우 쉽게 사용할 수 있는 방법이 바로 xpath 를 이용하는 방법입니다.
# xpath는 개발자 도구에서 해당 메뉴 부분을 마우스 오른쪽 버튼을 누르고
# copy -> copy xpath 를 선택하면 됩니다
driver.find_element_by_xpath("""//*[@id="snb"]/div/ul/li[1]/a""").click( )  # 정렬 버튼의 xpath 클릭
driver.find_element_by_xpath("""//*[@id="snb"]/div/ul/li[1]/div/ul/li[1]/a""").click( ) # 관련도순 xpath
```

5. "기간" 입력 창 누르고 "시작 날짜", "종료 날짜" 입력하기

```
# Step 4. 날짜 입력
driver.find_element_by_link_text("기간").click()
time.sleep(2)

# 이 부분이 아주 중요합니다.
# 날짜 부분에 날짜를 입력할 때 입력 속도가 너무 빠를 경우 날짜가 입력이 되다가
# 오타가 나오는 경우가 많습니다.
# 그래서 날짜를 입력할 때 for 반복문을 사용해서 1 글자씩 입력하도록 코딩했습니다.

# 시작 날짜 입력하기
s_date = driver.find_element_by_id("date_from_input")
driver.find_element_by_id("date_from_input").click()
s_date.clear( )  # 날짜 입력 부분에 기존에 입력되어 있던 날짜를 제거합니다.
time.sleep(1)
# 아래 코드가 날짜를 for 반복문으로 1 글자씩 입력하는 부분입니다.
for c in start_date:
    s_date.send_keys(c)
    time.sleep(0.1)
time.sleep(1)

# 종료 날짜 입력하기
e_date = driver.find_element_by_id("date_to_input")
driver.find_element_by_id("date_to_input").click()
e_date.clear()
time.sleep(1)

for c in end_date:
    e_date.send_keys(c)
    time.sleep(0.1)

# Step5. 날짜 입력 "적용" 버튼을 클릭 합니다.
driver.find_element_by_class_name("tx").click()
time.sleep(3)
```

6. 웹 페이지 스크롤 다운

```python
# 스크롤 다운
def scroll_down(driver):
    driver.execute_script("window.scrollTo(0, 99999999)")
    time.sleep(1)

# n: 스크롤할 횟수 설정
n = 10
i = 0
while i < n:
    scroll_down(driver) # 스크롤 다운
    i += 1
```



스크롤 다운!

7. 블로그 글, url 수집 후 엑셀 파일로 저장하기

```
url_list = []
title_list = []

# URL 크롤링 시작
articles = ".api_txt_lines.total_tit"
article_raw = driver.find_elements_by_css_selector(articles)

# url 크롤링 시작
for article in article_raw:
    url = article.get_attribute('href')
    url_list.append(url)
time.sleep(1)

# 제목 크롤링 시작
for article in article_raw:
    title = article.text
    title_list.append(title)

    print(title)

print("")
print('url갯수: ', len(url_list))
print('title갯수: ', len(title_list))
```

```
홍대 배달 음식 맛집 [ 삼겹살김치볶음밥, 치킨마요덮밥 맛집...
[부산해운대구맛집/해운대맛집... 멘보샤, 김치볶음밥, 짬뽕
[맛집/고덕] 김치볶음밥이 맛있는 레트로 맛집 『차트타임,
[한식] 양림동 맛집 : 밥제작소
[병점 소고기 찐맛집] 진안동 낙원갈비집 시그니처 왕관갈비 후기
잠실 롯데월드타워 맛집 백프로 만족은
광화문 디티워 맛집 추천 조선기술 모임장소로 딱!
신촌 돈까스] "윤오므라이스" - 가성비·맛 모두 잡은 혼밥 맛집!
[부산 기장] 샌드위치 맛집
고가길 9공탄 삼각지 연탄구이 항정살 맛집
논현역 맛집- 오감
[추천 배달음식 : 한식] "맛집"
안양일번가 맛집, 안양 '코리안 스테이크 김볶스' 리뷰
단대오거리 맛집 - 남다른감자탕 세이브존 성남점 : 추운 날씨에...
북구 맛집 "황금리브런치"
대구 앞산 맛집 블루웨이브 일부러 찾아가는 맛집!
천안아산맛집 혜화동돈가스 양도푸짐하고 맛있네요.
원주카페추천 @원주 분위기좋은 카페 겸 맛집 ::원주 마이테라스
불로동맛집 충장로 하고스에서 양식 냠냠냠냠
우만동맛집 낙원갈비집 수원월드컵점 점심특선 혜자
Achoo ~
영통브런치 꼬메타포레스트 영통파스타맛집
마초갈비 여의도공원점 : 숯불 돼지왕갈비 맛집 (여의도 고기집)
[영국생활] 2020년 12월 런던일기 1, 런던맛집, 테이트브리튼
추천해주는 청주 맛집리스트/ 구미사람이 추천해주는 구미 맛집
평택 법원 맛집 육풍에서 삼겹살 푸짐하게!
[쏨쏨이 여행 맛집] 괌여행 추천 찐 맛집 (도스버거, 에그앤띵스...
[청담사진맛집]데일리청담/55도와인앤다인

url갯수:  330
title갯수:  330
```

```
df = pd.DataFrame({'url':url_list, 'title':title_list})

# 저장하기
df.to_excel("blog_url.xlsx")
```

8. 저장한 블로그 글, url 불러오기

```
import sys
import os
import pandas as pd
import numpy as np
```

```
# "url_list.csv" 불러오기
url_load = pd.read_excel("blog_url.xlsx") # 기본 모델

num_list = len(url_load)

print(num_list)
url_load
```

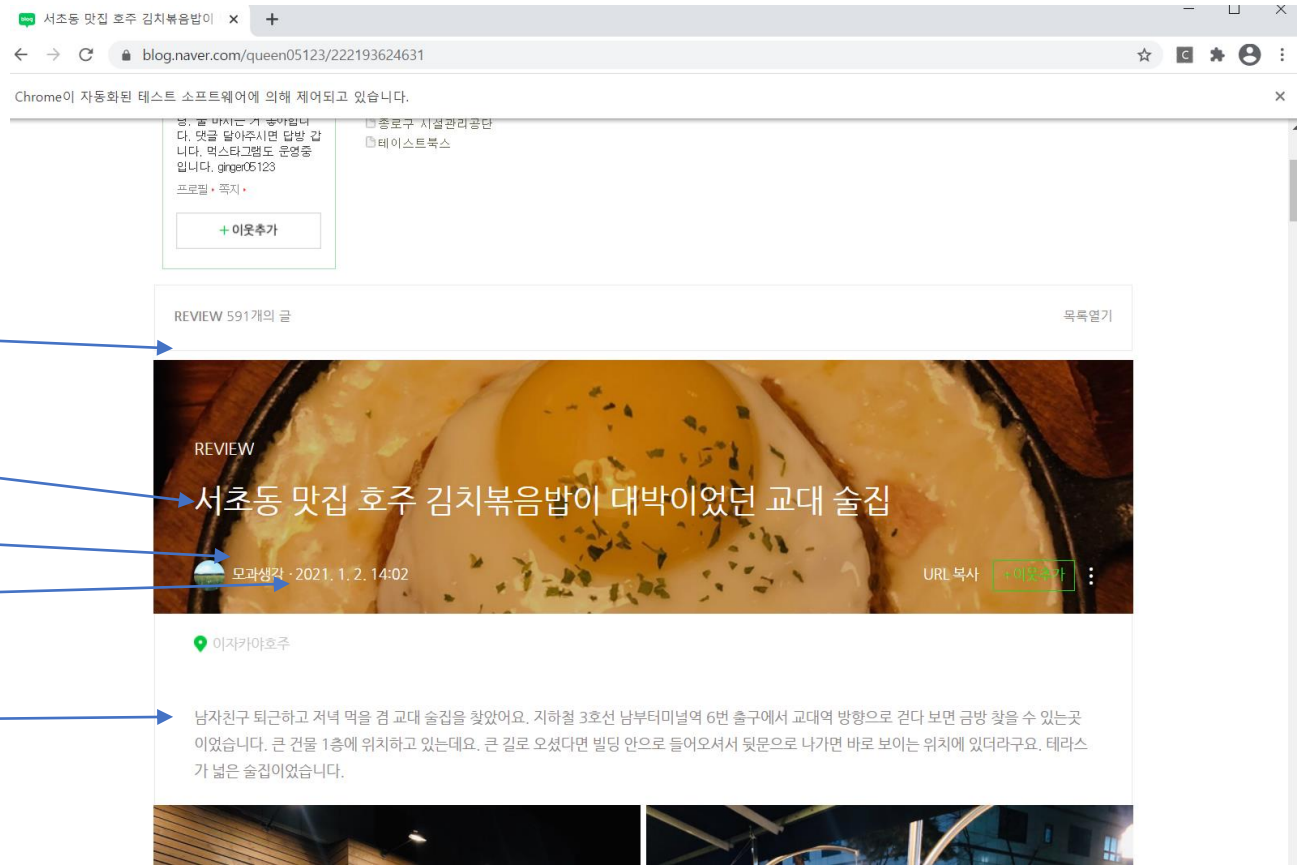
```
330
```

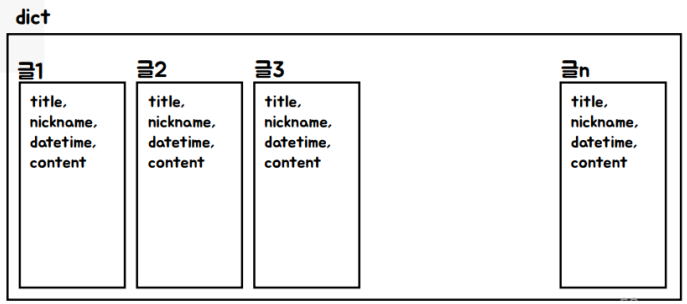
3]:

	Unnamed: 0	url	title
0	0	https://blog.naver.com/thearly?Redirect=Log&lo...	홍대 배달 음식 맛집 [삼겹살김치볶음밥, 치킨마요덮밥 맛집...
1	1	https://blog.naver.com/dnr6578?Redirect=Log&lo...	[부산해운대구맛집/해운대맛집... 멘보샤, 김치볶음밥, 짬뽕
2	2	https://blog.naver.com/i_believe28?Redirect=Lo...	[맛집/고터] 김치볶음밥이 맛있는 레트로 맛집『하트타임』
3	3	https://blog.naver.com/woonelee?Redirect=Log&l...	[한남동 맛집] 알프키친... 항정살&대파구이 김치볶음밥...
4	4	https://blog.naver.com/queen05123?Redirect=Log	서초동 맛집 호주 김치볶음밥이 대박이었던 교대 술집

9. 블로그 내용들 크롤링하기

```
dict = {}  # 전체 크롤링 데이터를 담을 그릇

# ★수집할 글 갯수 정하기
number = 30
for i in tqdm_notebook(range(0, number)):
    # 글 띄우기
    url = url_load['url'][i]
    driver = webdriver.Chrome("chromedriver.exe")  # 윈도우는 "chromedriver.exe"
    driver.get(url)   # 글 띄우기

    # 크롤링

    try :
        # iframe 접근
        driver.switch_to.frame('mainFrame')

        target_info = {}

        # 제목 크롤링 시작
        overlays = ".se-module.se-module-text.se-title-text"
        tit = driver.find_element_by_css_selector(overlays)          # title
        title = tit.text

        # 글쓴이 크롤링 시작
        overlays = ".nick"
        nick = driver.find_element_by_css_selector(overlays)         # nickname
        nickname = nick.text

        # 날짜 크롤링
        overlays = ".se_publishDate.pcol2"
        date = driver.find_element_by_css_selector(overlays)         # datetime
        datetime = date.text

        # 내용 크롤링
        overlays = ".se-component.se-text.se-l-default"
        contents = driver.find_elements_by_css_selector(overlays)

        content_list = []
        for content in contents:
            content_list.append(content.text)

        content_str = ' '.join(content_list)                         # content_str

        # 글 하나는 target_info라는 딕셔너리에 담기게 되고,
        target_info['title'] = title
        target_info['nickname'] = nickname
        target_info['datetime'] = datetime
        target_info['content'] = content_str

        # 각각의 글은 dict라는 딕셔너리에 담기게 됩니다.
        dict[i] = target_info
        time.sleep(1)

        # 크롤링이 성공하면 글 제목을 출력하게 되고,
        print(i, title)

        # 글 하나 크롤링 후 크롬 창을 닫습니다.
        driver.close()
```





9. 블로그 내용들 크롤링하기(이어서)

```
    # 에러나면 현재 크롤창을 닫고 다음 글(i+1)로 이동합니다.
    except:
        driver.close()
        time.sleep(1)
        continue

    # 중간,중간에 파일로 저장하기
    if i == 30 or 50 or 80:
        # 판다스로 만들기
        import pandas as pd
        result_df = pd.DataFrame.from_dict(dict, 'index')

        # 저장하기
        result_df.to_excel("blog_content.xlsx")
        time.sleep(3)

print('수집한 글 갯수: ', len(dict))
print(dict)
```


```
0 홍대 배달 음식 맛집 [ 삼겹살김치볶음밥, 치킨마요덮밥 맛집 - 돈떡 ]
2 [맛집/고터] 김치볶음밥이 맛있는 레트로 맛집 『하트타임』
3 [한남동 맛집] 알프키친 / 믹스드 프라이즈 / 버터밀크 오믈렛 라이스 / 항정살&대파구이 김치볶음밥 / 화이트라구 바질 페스토 파스타 / 화이트와인
4 서초동 맛집 호주 김치볶음밥이 대박이었던 교대 술집
5 여수 분위기 맛집 바다식탁 : 갓김치볶음밥, 로제 바다파스타
```

10. pandas로 만들고 저장하기

```
# 판다스로 만들기
import pandas as pd
result_df = pd.DataFrame.from_dict(dict, 'index')

result_df
```

	title	nickname	datetime	content
0	홍대 배달 음식 맛집 [삼겹살김치볶음밥, 치킨마요덮밥 맛집 - 돈떡]	EarLy\n(thearly)	2021. 1. 4. 20:57	요즘 사회적 거리두기\n적극 동참하며 \n음식은 댑부분 주문해서 먹고 \n나가질 ...
2	[맛집/고터] 김치볶음밥이 맛있는 레트로 맛집 『하트타임』	핑림	2020. 12. 10. 23:26	남부터미널 갈 일이 있어서 \n점심식사를 고터에서 하기로 했다!\n고터 가면 늘 들...
3	[한남동 맛집] 알프키친 / 믹스드 프라이즈 / 버터밀크 오믈렛 라이스 / 항정살&...	우넬리\n(woonelee)	2020. 12. 25. 10:45	썰랑 동생들과 함께 한남동맛집에 다녀왔지용~~\n코로나가 아니면 웨이팅이 어마무시하...
4	서초동 맛집 호주 김치볶음밥이 대박이었던 교대 술집	모과생강\n(queen05123)	2021. 1. 2. 14:02	남자친구 퇴근하고 저녁 먹을 겸 교대 술집을 찾았어요. 지하철 3호선 남부터미널역 ...
5	여수 분위기 맛집 바다식탁 : 갓김치볶음밥, 로제 바다파스타	논논이	2020. 12. 28. 22:30	안녕하세요 청춘블로그 논논이에요 !\n오늘은 여수 맛집이자 분위기 맛집으로\n유명한...
6	찐제주도민맛집 빨강코끼리에서 돈까스쫄면, 새우김치볶음밥 먹었음	CHCEKA	2021. 1. 2. 8:20	제주시 숙소에서 하루 자고 아침부터 뭘 먹을지 물색해봅니다\n부담스러운 건 딱히 ...
7	광화문 점심 맛집 도노베아토의 깔끔한 베이컨김치볶음밥!	광화문오피시아	2020. 12. 2.	안녕하세요~~\n광화문오피시아빌딩 지하1층\n오피시아부

```
# 엑셀로 저장하기
result_df.to_excel("blog_content.xlsx")
```