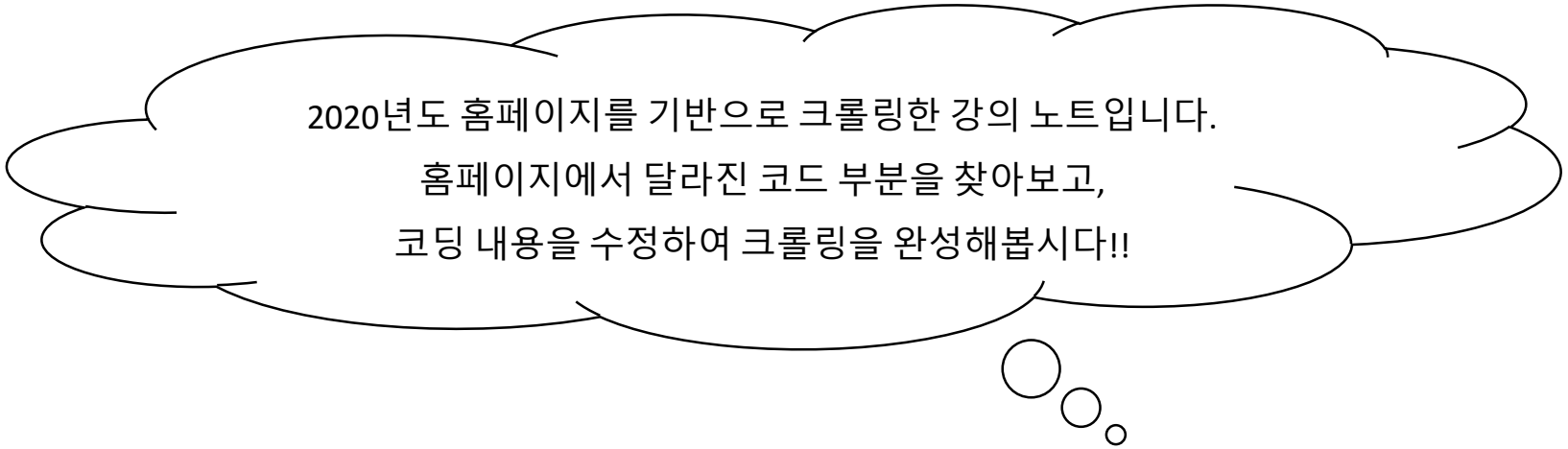
H.S. Kim
Christina



Web Crawling based on Text data

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

실습 노트 참고

- 09_Web Crawling based on Text data (한국관광공사_텍스트 형식으로 저장).ipynb

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

Step 1. 필요한 모듈과 함수를 로딩하고 검색어 입력받기

```
1 from bs4 import BeautifulSoup
2 from selenium import webdriver
3 import time
4 import sys
5
6 query_txt = input('크롤링할 키워드는 무엇입니까?: ')
```

```
크롤링할 키워드는 무엇입니까?: 제주도
```

Step 2. 크롤링 결과 데이터를 저장할 디렉토리와 파일명 지정하기

```
1 f_name = input('검색 결과를 저장할 파일경로와 이름을 지정하세요(예: D:₩₩ai₩data₩₩visit_ko_1.txt)')
```

```
검색 결과를 저장할 파일경로와 이름을 지정하세요(예: D:₩ai₩data₩visit_ko_1.txt)D:₩ai₩data₩visit_ko_1.txt
```

Step 3. 크롬 드라이버를 사용해서 웹 브라우저 실행하기

```
1 path = "c:/temp/chromedriver_89/chromedriver.exe"
2 driver = webdriver.Chrome(path)
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

Step 4. 한국 관광 공사 홈페이지 들어가기

```
1  driver.get("https://korean.visitkorea.or.kr")
2  time.sleep(2)  # 위 페이지가 모두 열릴 때 까지 2초 기다립니다.
```

Step 5. 팝업 창 닫기

```
1  driver.find_element_by_id("chkFrm01").click()
```

Step 6. popup 배너 창 닫기

```
1  driver.find_element_by_class_name("close").click()
```

Step 7. 입력 창 클릭하여 검색어 전달하기

```
1  element = driver.find_element_by_id("inp_search")
2  element.send_keys(query_txt)
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

Step 8. 검색 버튼 눌러 실행하기

```
1  driver.find_element_by_link_text("검색").click()
2
3  # class name 사용 (결과는 위와 동일)
4  #driver.find_element_by_class_name("btn_search").click()
5
6  # xpath 사용 (결과는 위와 동일)
7  #driver.find_element_by_xpath('//*[@id="gnbMain"]/div[1]/div/div[2]/span/a').click()
8
9  time.sleep(1)
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

Step 9. BeautifulSoup 활용하여 파싱하기

```
1 full_html = driver.page_source
2 soup = BeautifulSoup(full_html, 'html.parser')
3 content_list = soup.find('ul', class_ ='list_thumType flnon')
```

```
1 content_list
```

```
<ul class="list_thumType flnon"><li class="bdr_nor"><div class="photo"><a href="ja
vascript:" onclick='goDetail("25","80cbd7c5-3aa6-41ce-beb1-b949927f0b98")'><img al
t="화산섬 제주의 속살을 만나다." src="https://cdn.visitkorea.or.kr/img/call?cmd=VI
EW&amp;id=78079893-415a-4f5d-9d17-c35cd1bb1ea3"/></a></div><div class="area_txt">
<div class="tit"> <a href="javascript:" onclick='goDetail("25","80cbd7c5-3aa6-41ce
-beb1-b949927f0b98")'>화산섬 <!--HS-->제주<!--HE-->의 속살을 만나다.</a> </div> <p
>제주도</p> <p class="tag"><span>#제주세계자연유산센터</span><span>#휘닉스아일랜드
</span><span>#따라비오름</span><span>#조랑말체험공원</span><span>#해녀박물관</span
```

```
1 type(content_list)
```

```
bs4.element.Tag
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

< 참고 >

content_list = soup.find('ul', class_ ='list_thumType flnon')

위의 코드 결과 content_list는 태그를 가지고 있고, for 문을 수행하면 하위 태그에 해당하는 '< li class = bdr_nor >'단위로 처리한다.

```
1 content_list
```

```
<ul class="list_thumType flnon"><li class="bdr_nor"><div class="photo"><a href="javascript:" onclick='goDetail("25","80cbd7c5-3aa6-41ce-beb1-b949927f0b98")'><img alt="화산섬 제주의 속살을 만나다." src="https://cdn.visitkorea.or.kr/img/call?cmd=VIEW&amp;id=78079893-415a-4f5d-9d17-c35cd1bb1ea3" /></a></div><div class="area_txt"> <div class="tit"> <a href="javascript:" onclick='goDetail("25","80cbd7c5-3aa6-41ce-beb1-b949927f0b98")'>화산섬 <!--HS-->제주<!--HE-->의 속살을 만나다.</a> </div> <p>제주도</p> <p class="tag"><span>#제주세계자연유산센터</span><span>#휘닉스아일랜드</span><span>#따라비오름</span><span>#조랑말체험공원</span><span>#해녀박물관</span><span>#박물관</span><span>#레포츠</span><span>#등산</span><span>#등산로</span><span>#조깅</span><span>#아이와함께</span><span>#봄방학</span><span>#가족체험여행</span><span>#이색체험</span><span>#섬여행</span><span>#당일코스</span><span>#1박2일</span><span>#2박3일</span><span>#추천코스</span> </p> <p class="ar_tag"> <span>제주 세계자연유산센터</span> <span>휘닉스제주섭지코지(숙박)</span> <span>따라비오름</span> <span>조랑말체험공원</span> <span>해녀박물관</span> </p></div><button class="btn_view" title="내용 더보기" type="button">더보기</button><div class="pop_subMenu"> <ul> <li class="btn_far"><a href="javascript:" onclick='setFavoContent("80cbd7c5-3aa6-41ce-beb1-b949927f0b98")'>즐겨찾기</a></li> <li class="btn_share" id="80cbd7c5-3aa6-41ce-beb1-b949927f0b98"><a href="javascript:" onclick='getShareInfo("%ED%99%94%EC%82%B0%EC%84%AC%20%3C!HS%3E%EC%A0%9C%EC%A3%BC%3C!HE%3E%EC%9D%98%20%EC%86%8D%EC%82%B4%EC%9D%84%20%EB%A7%8C%EB%82%98%EB%8B%A4.","78079893-415a-4f5d-9d17-c35cd1bb1ea3","80cbd7c5-3aa6-41ce-beb1-b949927f0b98","25");'>공유하기</a></li> <li class="btn_cart" id="80cbd7c5-3aa6-41ce-beb1-b949927f0b98"><a href="javascript:" onclick='myCourseList("C","25","80cbd7c5-3aa6-41ce-beb1-b949927f0b98","")'>코스에 담기</a></li> </ul></div></li><li class="bdr_nor"><div class="photo"><a href="javascript:" onclick='goDetail("25","dd18f6b4-c48b-4ffe-8be4-d24bf6aa07da")'><img alt="있는 그대로 제주의 자연을 거닐어볼 코스" src="https://cdn.visitkorea.or.kr/img/call?cmd=VIEW&amp;id=26c3d46d-00bb-4831-85d7-b5b75a386f8c" /></a></div><div class="area_txt"> <div class="tit"> <a href="javascript:" onclick='goDetail("25","dd18f6b4-c48b-4ffe-8be4-d24bf6aa07da")'>있는 그대로 <!--HS-->제주<!--HE-->의 자연을 거닐어볼 코스</a> </div> <p>제주도 </p> <p class="tag"><span>#한라생태숲</span><span>#제주돌문화공원</span><span>#제주돌담길공원</span><span>#제주돌담길공원</span><span>#곶자왈박물관</span><span>#박
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

Step 10. for 반복문을 활용하여 텍스트 데이터 분리하기

```
1 for i in content_list:
2     print("●")
3     print(i.text.strip())
```

●
화산섬 제주의 속살을 만나다. 제주도 #제주세계자연유산센터#휘닉스아일랜드#따라비오름#조랑말체험공원#해녀박물관#박물관#레포츠#등산#등산로#조깅#아이와함께#봄방학#가족체험여행#이색체험#섬여행#당일코스#1박2일#2박3일#추천코스 제주 세계자연유산센터 휘닉스제주섭지코지(숙박) 따라비오름 조랑말체험공원 해녀박물관 더보기 즐겨찾기 공유하기 코스에 담기
●
있는 그대로 제주의 자연을 거닐어볼 코스 제주도 #한라생태숲#제주노루생태관찰원#제주돌문화공원#국립제주박물관#박물관#이색체험#자연좋은곳#전망좋은곳#힐링#바다풍경#숲체험원#가족여행#친구와함께#당일코스#1박2일#2박3일 한라생태숲 제주 노루생태관찰원 제주돌문화공원 국립제주박물관 더보기 즐겨찾기 공유하기 코스에 담기
●
감성 넘치는 인생 사진, 가을 제주 120% 즐기기 제주도 제주시 #가을여행#제주도가볼만한곳#제주도여행#인생샷#새

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

<참고>
bs4.element.Tag 타입의 경우,
text 메소드를 사용하여태그를 제거하고 텍스트만 추출 가능

```
1 type(content_list)
```

```
bs4.element.Tag
```

```
1 content_list.text
```

```
" 화산섬 제주의 속살을 만나다. 제주도 #제주세계자연유산센터#휘닉스아일랜드#따라비오름#조랑말체험공원#해녀박물관#박물관#레포츠#등산#등산로#조깅#아이와함께#봄방학#가족체험여행#이색체험#섬여행#당일코스#1박2일#2박3일#추천코스 제주 세계자연유산센터 휘
```

```
1 type(content_list.text)
```

```
str
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

Step 11. 텍스트 데이터 파일로 저장하기

```
1  # sys.stdout
2  # 표준 출력 상태를 확인할 수 있으며, 표준 출력은 모니터이다.
3
4  sys.stdout
```

```
<ipykernel.iostream.OutStream at 0x21124d90fa0>
```

```
1  # 표준 출력을 파일로 교체했다가 다시 모니터로 리셋하기 위해서
2  # orig_stdout 변수에 현재의 표준 출력 상태를 할당해 놓음
3
4  orig_stdout = sys.stdout
```

```
1  orig_stdout
```

```
<ipykernel.iostream.OutStream at 0x21124d90fa0>
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)


```
# 데이터를 추가하는 'a' 모드로 파일을 오픈
f = open(f_name, 'a', encoding='UTF-8')
```

```
# 표준 출력을 파일로 교체
sys.stdout = f
```

```
sys.stdout
```

```
<_io.TextIOWrapper name='D:₩₩ai₩₩data₩₩visit_ko_1.txt' mode='a' encoding='UTF-8'>
```

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

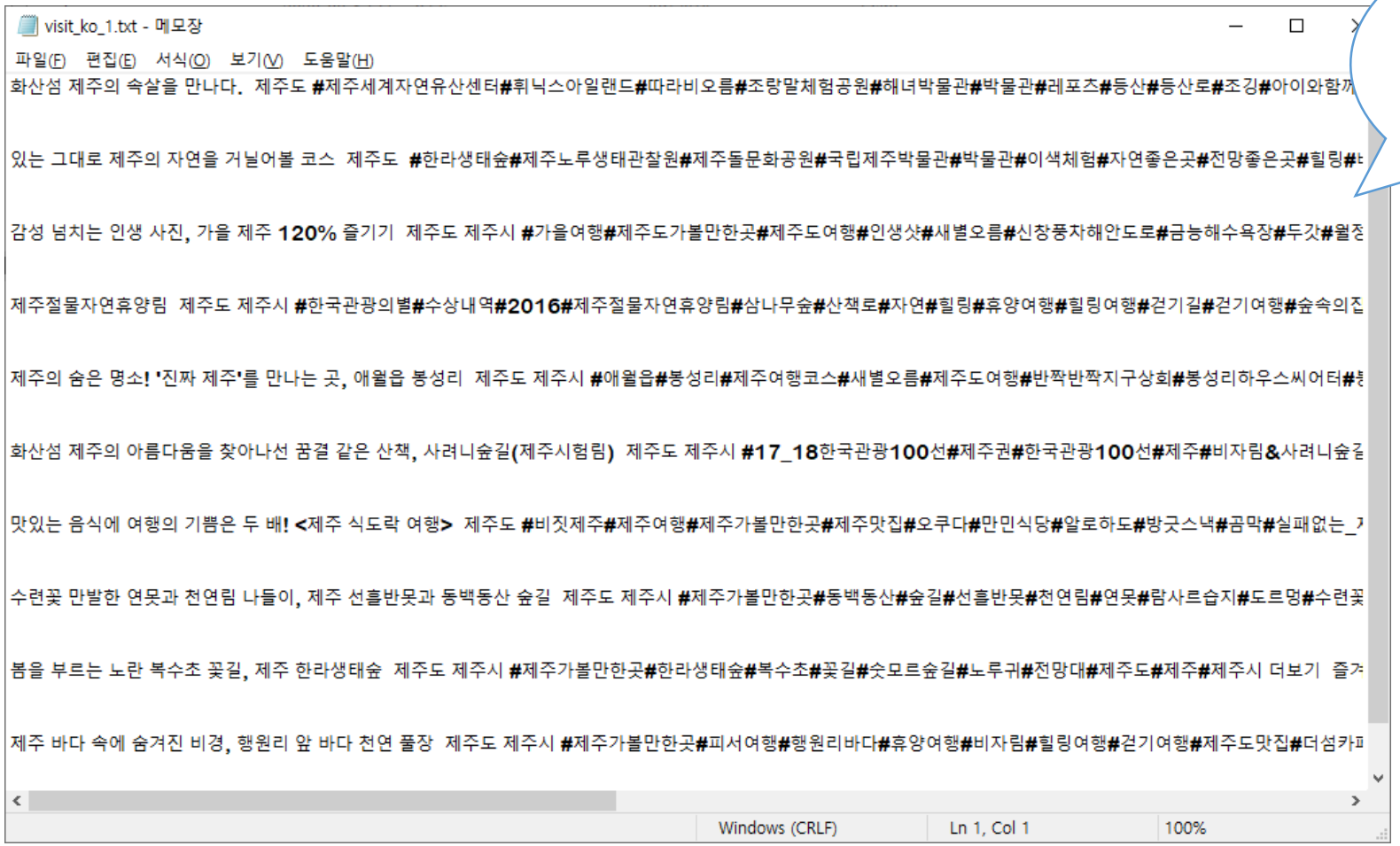
```
for i in content_list:
    print(i.text.strip())
    print('₩n')

sys.stdout = orig_stdout  # 표준 출력을 다시 모니터로 리셋
f.close()
```

검색 결과를 텍스트 형식으로 저장(한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성



D:\ai\DATA
visit_ko_1.txt

검색 결과를 텍스트 형식으로 저장(네이버 포스트 활용)

실습 노트 참고

- 09_Web Crawling based on Text data (네이버 포스트_텍스트 형식으로 저장).ipynb

검색 결과를 텍스트 형식으로 저장(네이버 포스트 활용)

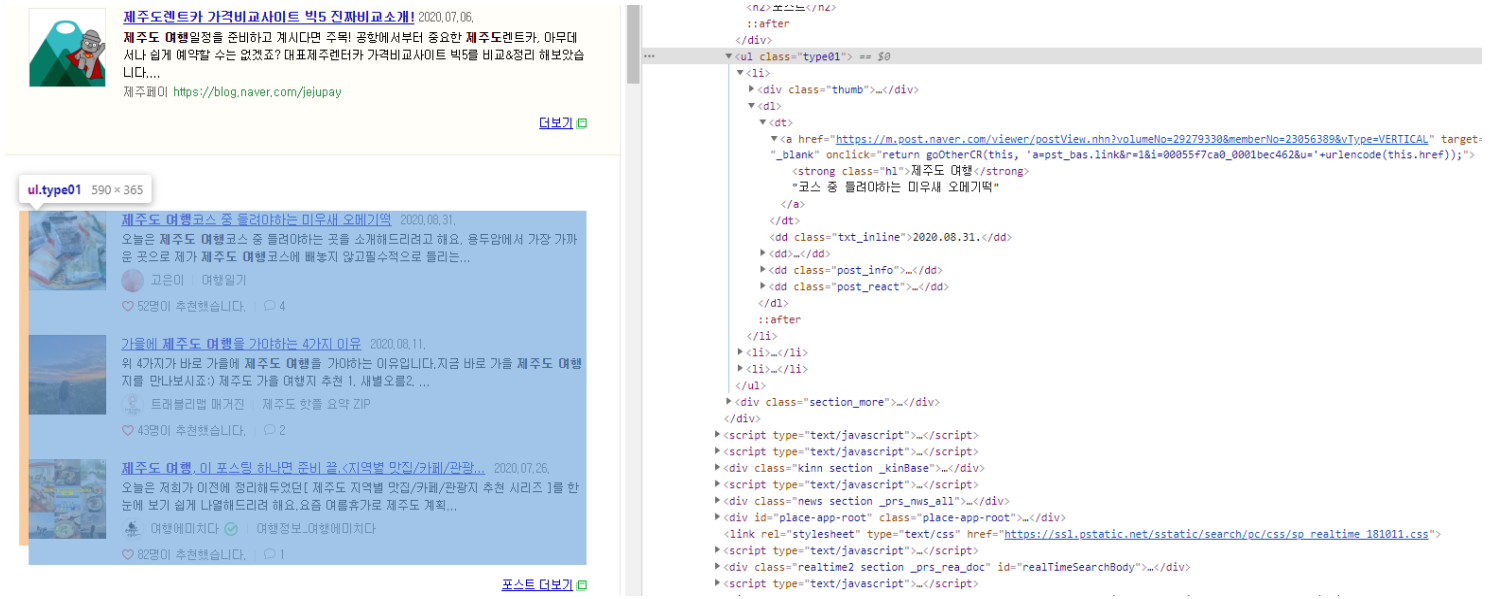
텍스트 데이터 웹 크롤링

1. **웹 크롤링 후 텍스트 형식 파일로 저장**
   - 한국관광공사
   - **네이버 포스트**
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

- 네이버 (www.naver.com) 검색창에서 **"제주도 여행"**으로 자동 검색하는 코드를 완성하고, 출력되는 결과 중에서 **"포스트"** 부분에 나타난 텍스트만 추출하여 크롤링하는 프로그램을 작성하시오.

- 크롤링 결과는 **D:\ai\DATA** 디렉토리에 **Naver_post.txt** 파일명으로 저장되도록 코드를 완성하시오.



검색 결과를 텍스트 형식으로 저장(네이버 포스트 활용)

크롤링 결과 예시

D:\ai\DATA
Naver_post.txt

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

실습 노트 참고

- 09_Web Crawling based on Text data (한국관광공사_다양한 형식으로 저장).ipynb

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

Step 1. 필요한 모듈과 함수 준비하기

```
1 from bs4 import BeautifulSoup
2 from selenium import webdriver
3 import time
4 import sys
```

```
1 query_txt = input('크롤링할 키워드는 ?: ')
2 f_name = input('검색 결과를 저장할 txt 파일경로와 파일명은? (예:D:₩₩ai₩₩data₩₩visit_ko_2.txt)')
3 fc_name = input('검색 결과를 저장할 csv 파일경로와 파일명은? (예:D:₩₩ai₩₩data₩₩visit_ko_2.csv)')
4 fx_name = input('검색 결과를 저장할 xls 파일경로와 파일명은? (예:D:₩₩ai₩₩data₩₩visit_ko_2.xlsx)')
```

```
크롤링할 키워드는 ?: 제주도 여행
검색 결과를 저장할 txt 파일경로와 파일명은? (예:D:₩ai₩data₩visit_ko_2.txt)D:₩ai₩data₩visit_ko_2.txt
검색 결과를 저장할 csv 파일경로와 파일명은? (예:D:₩ai₩data₩visit_ko_2.csv)D:₩ai₩data₩visit_ko_2.csv
검색 결과를 저장할 xls 파일경로와 파일명은? (예:D:₩ai₩data₩visit_ko_2.xlsx)D:₩ai₩data₩visit_ko_2.xlsx
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

Step 2. 크롬 드라이버를 사용해서 웹 브라우저를 실행한 후, 한국관광공사 홈피 들어가기

```
path = "c:/temp/chromedriver_240/chromedriver.exe"
driver = webdriver.Chrome(path)

driver.get("https://korean.visitkorea.or.kr/main/main.html")
time.sleep(2)
```

Step 3. 팝업 창, 배너 창 닫기

```
driver.find_element_by_id("chkForm01").click()
driver.find_element_by_class_name("close").click()
```

Step 4. 입력 창 클릭하여 검색어 전달하기

```
# driver.find_element_by_id("btnSearch").click()
element = driver.find_element_by_id("inp_search")
element.send_keys(query_txt)
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

Step 5. 검색 버튼 눌러 실행하기

```
driver.find_element_by_link_text("검색").click()

# class name 사용 (결과는 위와 동일)
#driver.find_element_by_class_name("btn_search2").click()

# xpath 사용 (결과는 위와 동일)
#driver.find_element_by_xpath('//*[@id="gnbMain"]/div/div/div[1]/div[1]/a').click()

time.sleep(1)
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

Step 6. BeautifulSoup 활용하여 파싱하기

```
1  html = driver.page_source
2  soup = BeautifulSoup(html, 'html.parser')
```

아래의 3개의 코드는 같은 의미를 가지고 있다.

```
1  # 1번 방식
2  content_list = soup.find('ul',class_='list_thumType flnon')
3  print(content_list)
4
5  # 2번 방식
6  #content_list = soup.find('ul','list_thumType flnon')
7  #print(content_list)
8
9  # 3번 방식
10 #img_src = soup.find ('ul', {'class' :'list_thumType flnon'})
11 #print(content_list)
```

```
<ul class="list_thumType flnon"><li class="bdr_nor"><div class="photo"><a href="javascript:
-41cf-aab7-fad6bf4d1441")'><img alt="가족 단위 추천 느린 여행지 2탄 - 제주시 동부" src="htt
IEW&amp;id=2ff9cdfa-f433-4859-b68c-8416bd675710"/></a></div><div class="area_txt"> <div cla
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

< 특정 항목들을 분리해서 추출하기 >

```
contents = content_list.find('div','tit').get_text( ) # find( ) 함수는 1건만 추출한다.
#contents = content_list.find('div','tit').text      # 위의 코드와 동일한 결과
contents
```

' 가족 단위 추천 느린 여행지 2탄 - 제주시 동부 '

```
#참고
re = contents.strip()
re
```

'가족 단위 추천 느린 여행지 2탄 - 제주시 동부'

```
print('● 내용')
print(contents.strip())
tag = content_list.find('p','tag').get_text()
print('● 태그')
print(tag.strip())
```

● 내용
가족 단위 추천 느린 여행지 2탄 - 제주시 동부
● 태그
#비짓제주#제주가볼만한곳#제주_여행법#가족여행#가족과함께#제주절물자연휴양림#제주돌문화공원#선흘분교#아부오름

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

< 각 항목별로 분리하여 추출하고 변수에 할당하기 >

```
no = 1
num_result =[ ]
contents_result=[ ]
tag_result=[ ]

for i in content_list:
    num_result.append(no)
    print(num_result)

    contents = i.find('div','tit').get_text( )
    contents_result.append(contents.strip())
    print(contents_result)

    tag = i.find('p','tag').get_text()
    tag_result.append(tag.strip())
    print(tag_result)
    print("-"*100)

    no += 1
```

```
[1]
['가족 단위 추천 느린 여행지 2탄 - 제주시 동부']
['#비짓제주#제주가볼만한곳#제주_여행법#가족여행#가족과함께#제주절물자연휴양림#제주돌문화공원#선흘분교#아부오름'
----------------------------------------------------------------------------------------------------
[1, 2]
['가족 단위 추천 느린 여행지 2탄 - 제주시 동부', "[국내여행 버킷리스트 1탄] '나도 제주 해녀가 되고 싶어요', 제주
해녀체험"]
['#비짓제주#제주가볼만한곳#제주_여행법#가족여행#가족과함께#제주절물자연휴양림#제주돌문화공원#선흘분교#아부오름'
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

```
1 num_result
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

```
1 contents_result
```

```
['가족 단위 추천 느린 여행지 2탄 - 제주시 동부',
 "[국내여행 버킷리스트 1탄] '나도 제주 해녀가 되고 싶어요', 제주 하도어촌체험마을 해녀체험",
 '엄마 아빠에겐 쉼표, 아이에겐 느낌표! 아이와 떠나는 제주 여행',
 '한국관광공사 추천 2월 걷기여행길, 이른 봄을 맞이하는 제주도 걷기길',
 '맛있는 음식에 여행의 기쁨은 두 배! <제주 식도락 여행>',
 '[국내 트레킹 추천] 제주 올레길 걷기 여행, 준비부터 코스 선택까지 꿀팁 총정리!',
 "소소한 '제주여행' 꿀팁! 〈겨울편〉",
 '가족 단위 추천 느린 여행지 1탄 - 제주시 서부',
 '반려견과 떠나는 여행 <제주 반려견 숙소&카페>',
 '오지게 멋진 제주 건축물 여행']
```

```
1 tag_result
```

```
['#비짓제주#제주가볼만한곳#제주_여행법#가족여행#가족과함께#제주절물자연휴양림#제주돌문화공원#선
 '#제주도가볼만한곳#이색체험#섬여행#친구와함께#수상레포츠#바다여행#레포츠여행#하도어촌체험마을#
 '#영유아추천코스#제주가볼만한곳#에코랜드#아쿠아플라넷제주#세계자동차제주박물관#헬로키티아일랜드
테지움#제주1박2일코스#제주2박3일코스#제주3박4일코스#무장애여행지#체험여행#화산송이#아이와함께#0
 '#걷기여행길#2월걷기여행길#제주도#제주도걷기길#트레킹코스',
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

< 분리된 데이터를 데이터 프레임으로 만들어서 csv , xls 형식으로 저장하기 >

```
# 출력 결과를 데이터 프레임 형태로 만들기
import pandas as pd

korea_trip = pd.DataFrame()
korea_trip['번호']=num_result
korea_trip['내용']=contents_result
korea_trip['태그']=tag_result
```

```
korea_trip
```

	번호	내용	태그
0	1	가족 단위 추천 느린 여행지 2탄 - 제주시 동부	#비짓제주#제주가볼만한곳#제주_여행법#가족여행#가족과함께#제주절물자연휴양림#제주돌문...
1	2	[국내여행 버킷리스트 1탄] '나도 제주 해녀가 되고 싶어요', 제주 하도어촌체험마...	#제주도가볼만한곳#이색체험#섬여행#친구와함께#수상레포츠#바다여행#레포츠여행#하도어촌...
2	3	엄마 아빠에겐 쉼표, 아이에겐 느낌표! 아이와 떠나는 제주 여행	#영유아추천코스#제주가볼만한곳#에코랜드#아쿠아플라넷제주#세계자동차제주박물관#헬로키티

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

```
1 korea_trip['번호']
```

```
0     1
1     2
2     3
3     4
4     5
5     6
6     7
7     8
8     9
9    10
Name: 번호, dtype: int64
```

```
1 korea_trip['내용']
```

```
0                                  가족 단위 추천 느린 여행지 2탄 - 제주시 동부
1    [국내여행 버킷리스트 1탄] '나도 제주 해녀가 되고 싶어요', 제주 하도어촌체험
마...
2                          엄마 아빠에겐 쉼표, 아이에겐 느낌표! 아이와 떠나는 제주 여행
3                    한국관광공사 추천 2월 걷기여행길, 이른 봄을 맞이하는 제주도 걷기길
4                          맛있는 음식에 여행의 기쁨은 두 배! <제주 식도락 여행>
5          [국내 트레킹 추천] 제주 올레길 걷기 여행, 준비부터 코스 선택까지 꿀팁 총정
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

```
1  korea_trip['태그']
```

```
0      #비짓제주#제주가볼만한곳#제주_여행법#가족여행#가족과함께#제주절물자연휴양림#제주돌문...
1      #제주도가볼만한곳#이색체험#섬여행#친구와함께#수상레포츠#바다여행#레포츠여행#하도어촌...
2      #영유아추천코스#제주가볼만한곳#에코랜드#아쿠아플라넷제주#세계자동차제주박물관#헬로키티...
3                                  #걷기여행길#2월걷기여행길#제주도#제주도걷기길#트레킹코스
4      #비짓제주#제주여행#제주가볼만한곳#제주맛집#오쿠다#만민식당#알로하도#방긋스낵#곰막#...
5      #제주올레길#제주올레걷기축제#제주올레#제주올레여행자센터#트레킹코스#트레킹#10월가볼...
6      #제주도#제주도여행코스#라프별빛축제#새별오름#카멜리아힐#향파두리유적지#제주가볼만한곳...
7      #비짓제주#제주여행#제주가볼만한곳#동명리#명월리#월령리#더럭초등학교#가족여행#가족과...
8      #비짓제주#제주여행#제주가볼만한곳#멍멍플레이스#브레이브독스#카페비글#카페3535.1...
9      #제주여행코스#제주가볼만한곳#제주건축물여행#국내건축물#제주건축여행#안도타다오#이타미...
Name: 태그, dtype: object
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - **한국관광공사**
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

```
# csv 형식으로 저장하기
korea_trip.to_csv(fc_name, encoding="utf-8-sig")
print(" csv 파일 저장 경로: %s" %fc_name)
```

```
csv 파일 저장 경로: D:₩ai₩data₩visit_ko_2.csv
```

```
# 엑셀 형식으로 저장하기
import xlwt   #conda 또는 pip install xlwt 설치되어 있어야 함
korea_trip.to_excel(fx_name)
print(" xls 파일 저장 경로: %s" %fx_name)
```

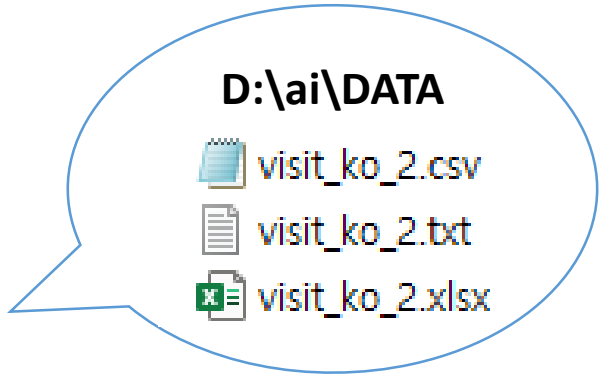
```
xls 파일 저장 경로: D:₩ai₩data₩visit_ko_2.xlsx
```

```
# txt 형식으로 저장하기
f = open(f_name, 'a',encoding='UTF-8')
f.write(str(num_result))  # 번호
f.write(str(contents_result))  # 내용
f.write(str(tag_result))  # 태그

f.close( )
print(" txt 파일 저장 경로: %s" %f_name)
```

```
txt 파일 저장 경로: D:₩ai₩data₩visit_ko_2.txt
```

```
type(korea_trip)
```

```
pandas.core.frame.DataFrame
```



검색 결과를 다양한 형식으로 저장 (네이버 블로그 활용)

실습 노트 참고

- 09_Web Crawling based on Text data (네이버 블로그_다양한 형식으로 저장).ipynb

검색 결과를 다양한 형식으로 저장 (네이버 블로그 활용)

웹 크롤링 코딩 지시사항

1. 네이버 (www.naver.com) 에서 아래 예시와 같이 사용자에게 검색할 키워드와 검색 결과를 저장할 디렉토리와 파일명을 입력 받도록 하시오.

   <예시>
   검색할 키워드는? 영화
   검색결과를 txt 타입으로 저장할 경로와 파일명은? D:₩ai₩DATA₩Naver_blog.txt
   검색결과를 csv 타입으로 저장할 경로와 파일명은? D:₩ai₩DATA₩Naver_blog.csv
   검색결과를 xlsx 타입으로 저장할 경로와 파일명은? D:₩ai₩DATA₩Naver_blog.xlsx

2. 위와 같이 입력하여 검색된 결과 중에서 **블로그**만 선택하여 처음 1페이지에 있는 결과 10건의 게시물을 txt 형식의 파일로 저장하시오. 그리고 10건의 검색된 결과를 데이터 프레임으로 구성한 후 xlsx , csv 형식으로 저장하시오.

검색 결과를 다양한 형식으로 저장 (네이버 블로그 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - 한국관광공사
   - **네이버 블로그**
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

1. 게시글 제목
2. 게시글 요약 내용
3. 게시글 작성 일자
4. 블로그 닉네임

[블로그 예시]

검색 결과를 다양한 형식으로 저장 (네이버 블로그 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - 한국관광공사
   - **네이버 블로그**
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

[txt 형식으로 저장 예시]

Naver_blog.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

1.번호:1
2.제목:영화 반도 후기 (강동원 위에 이정현 있었다!)
3.내용:영화 반도 후기 (강동원 위에 이정현 있었다!) 개봉일 : 2020. 07. 15 관람일 : 2020. 07. 15... 동원한 영화 부산행의 4년 후 스토리의 영화 반도가 개봉해서 개봉 당일 바로...
4.작성일자:2020.07.17.
5.블로그닉네임:실버린의 실버월드
==========
1.번호:2
2.제목:영화 침입자 - 아이고 이를 어쩔~
3.내용:참 오랜 시간 개봉이 미뤄졌던 영화 <침입자>가 바로 오늘 개봉했다. 솔직히 손꼽아 기다렸다고는 못하겠지만, 오랜만에 개봉하는 우리 상업 영화가 그저 반가워 별 고민...
4.작성일자:2020.06.04.
5.블로그닉네임:럽카키의 수다 공간
==========
1.번호:3
2.제목:영화 뮬란 - 디즈니 영화 역사상 최악의 재해석..
3.내용:'내 이름은 화 뮬란' 개봉전부터 큰 논란에 휩싸였던 디즈니 영화 <뮬란>. 그러나 특정 영화가 아닌 디즈니 자체에 대한 보이콧이나 중국 자체에 대한 보이콧이 더...
4.작성일자:2020.09.21.
5.블로그닉네임:단단한 달걀의 쉼터...
==========
1.번호:4

검색 결과를 다양한 형식으로 저장 (네이버 블로그 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. **웹 크롤링 후 다양한 형식 파일로 저장**
   - 한국관광공사
   - **네이버 블로그**
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성

[csv 형식으로 저장 예시]

Naver_blog.csv - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
번호,제목,내용,작성일자,블로그닉네임
1,영화 반도 후기 (강동원 위에 이정현 있었다!),영화 반도 후기 (강동원 위에 이정현 있었다!) 개봉일 : 2020. 07. 15 관람일 : 2020. 07. 15... 동원한 영화 부산행의 4년 후 스토리의 영화 반도가 개봉해서 개봉 당일 바로...,2020.07.17. ,실버린의 실버월드
2,영화 침입자 - 아이고 이를 어쩔~,"참 오랜 시간 개봉이 미뤄졌던 영화 <침입자>가 바로 오늘 개봉했다. 솔직히 손꼽아 기다렸다고는 못하겠지만, 오랜만에 개봉하는 우리 상업 영화가 그저 반가워 별 고민...",2020.06.04. ,럽카키의 수다 공간
3,영화 뮬란 - 디즈니 영화 역사상 최악의 재해석..,'내 이름은 화 뮬란' 개봉전부터 큰 논란에 휩싸였던 디즈니 영화 <뮬란>. 그러나 특정 영화가 아닌 디즈니 자체에 대한 보이콧이나 중국 자체에 대한 보이콧이 더...,2020.09.21. ,단단한 달걀의 쉼터...
4,129_영화 월드 워 Z (World War Z),"개봉영화 말고 뒤늦게 혹은 다시금 챙겨본 영화에 대한 짧은 저의 이야기를 담을게요! 월드워Z 감독 마크 포스터 출연 브래드 피트, 미레유 에노스 개봉 2013. 06. 20. 제목...",2020.08.21. ,시간의마법사
5,"영화 나를 찾아줘, 고생하는 이영애",나를 찾아줘 ▣ 스릴러 영화 후기 정연(이영애)과 명국(박해준)은 부부입니다.... 갈 길을 잃다 ▣ 무엇을 말하고 싶었는가 특급 영화 구독 서비스 캐치온의 무비...,2020.09.12. ,마시우가 영화 본다
6,담보-영화를 구원한 작은 보물,많은 분들이 JK필름의 영화들에 후한 점수를 주지 않는 것을 알지만 저는 대중들의... 어울리는 영화라는 생각도 들었습니다. 물론 너무도 안이한 개봉일 선정과 관객의...,6일 전 ,내 집으로 와요
7,영화 흥행에도 안전지대가 있다?,시리즈 영화가 아닐까 합니다. <쥬라기 공원> <배트맨> <캐리비안의 해적> <트랜스포머> <분노의 질주> <해리 포터> 등의 시리즈 영화들은 개봉하는 족족 연속적으로...,2020.07.24. ,씨네플레이
8,오케이 마담 - 코믹 영화,어느 순간부터 헐리우드 코믹 영화는 안 봐도 충무로에서 만든 코믹 영화는 본다. 코믹 영화 자체만 놓고 본다면 무적이나 유치하다. 그 유치함을 기꺼이 감수할...,2020.08.16. ,천천히 꾸준히(Slow...
9,난해한(어려운) 영화가 만들어지는 이유,"'테넷'이 몰고 온 어려운 영화의 여파. 하지만 이는 관객의 호불호를 낳기도 한다. 한 쪽에서는 대중 상업 영화는 무조건 쉬워야 한다는 주장을 펼치고, 다른 쪽에서는...",2020.09.11. ,가짜 영화평론가 님...
10,2020년 추석특선영화 - 공중파 & 종편,2020년 추석특선영화 - 공중파 & 종편 어느덧 추석 연휴가 오늘부터... 어떤 영화들이 방영될까 궁금해지기 마련이죠 설 연휴 공중파와 케이블(tvN)...,7일 전 ,한마루의 영화노트
```

[xlsx 형식으로 저장 예시]

	A	B	C	D	E
1	번호	제목	내용	작성일자	블로그닉네임
2	1	영화 반도 후기 (강동원 위에 이정현 있	영화 반도 후기 (강동원 위에 이정현 있었다!) 개봉일 : 2020. 07. 15	2020.07.17.	실버린의 실버월드
3	2	영화 침입자 - 아이고 이를 어쩔~	참 오랜 시간 개봉이 미뤄졌던 영화 <침입자>가 바로 오늘 개봉했	2020.06.04.	럽카키의 수다 공간
4	3	영화 뮬란 - 디즈니 영화 역사상 최악의	'내 이름은 화 뮬란' 개봉전부터 큰 논란에 휩싸였던 디즈니 영화 <	2020.09.21.	단단한 달걀의 쉼터...
5	4	129_영화 월드 워 Z (World War Z)	개봉영화 말고 뒤늦게 혹은 다시금 챙겨본 영화에 대한 짧은 저의 (	2020.08.21.	시간의마법사
6	5	영화 나를 찾아줘, 고생하는 이영애	나를 찾아줘 ▣ 스릴러 영화 후기 정연(이영애)과 명국(박해준)은 부	2020.09.12.	마시우가 영화 본다
7	6	담보-영화를 구원한 작은 보물	많은 분들이 JK필름의 영화들에 후한 점수를 주지 않는 것을 알지만	6일 전	내 집으로 와요
8	7	영화 흥행에도 안전지대가 있다?	시리즈 영화가 아닐까 합니다. <쥬라기 공원> <배트맨> <캐리비안	2020.07.24.	씨네플레이
9	8	오케이 마담 - 코믹 영화	어느 순간부터 헐리우드 코믹 영화는 안 봐도 충무로에서 만든 코믹	2020.08.16.	천천히 꾸준히(Slow...
10	9	난해한(어려운) 영화가 만들어지는 이유	'테넷'이 몰고 온 어려운 영화의 여파. 하지만 이는 관객의 호불호를	2020.09.11.	가짜 영화평론가 님...
11	10	2020년 추석특선영화 - 공중파 & 종편	2020년 추석특선영화 - 공중파 & 종편 어느덧 추석 연휴가 오늘부터	7일 전	한마루의 영화노트

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

실습 노트 참고

- 09_Web Crawling based on Text data (한국관광공사_검색 개수 기반).ipynb

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

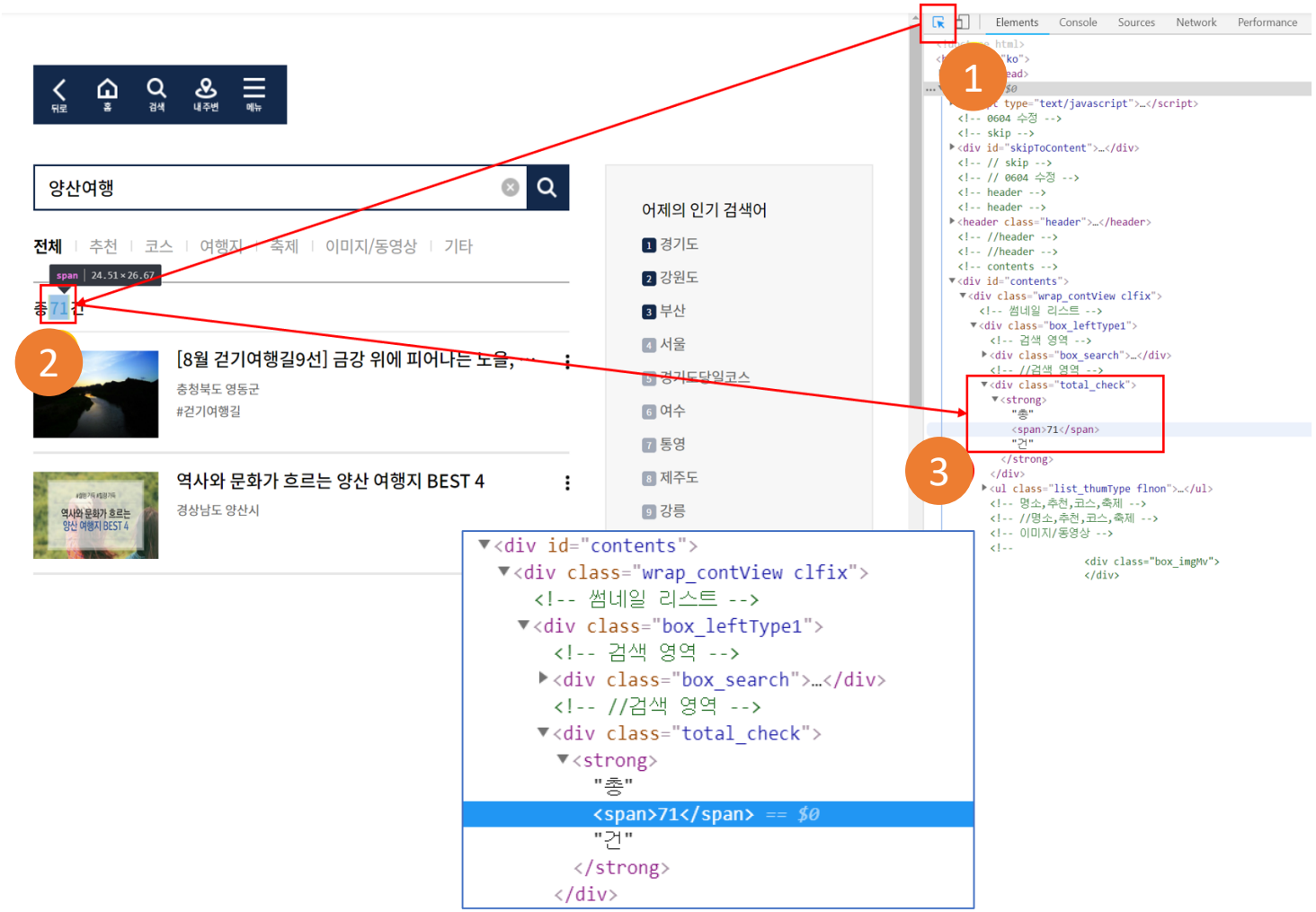
텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그

3. 검색 개수 기반의 웹 크롤링

- **한국관광공사**

Openpyxl 활용하여 엑셀 파일 생성



검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

검색 페이지와 게시글 접근을 위한 반복문 구조

100 건

re_cnt

(예) 75건

real_page_cnt

(예) 8 page

```
for x in range(1,real_page_cnt+1) :
    for i in content_list:
        검색된 건수 1개씩 처리
        if no == re_cnt
            break
        no += 1
    x += 1
    if x == real_page_cnt+1
        break
```

External for문
x는 검색 페이지 접근을 위한 변수

Internal for문
i는 한 페이지 내에서 10개의 게시글을 접근하기 위한 변수

no는 게시글 카운트 변수

re_cnt는 최종 검색할 게시글 개수

다음 페이지 접근을 위해서

real_page_cnt는 검색할 최종 페이지

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. **검색 개수 기반의 웹 크롤링**
   - **한국관광공사**

Openpyxl 활용하여 엑셀 파일 생성

[8월 걷기여행길9선] 금강 위에 피어나는 노을, …

충청북도 영동군

#걷기여행길

역사와 문화가 흐르는 양산 여행지 BEST 4

경상남도 양산시

해시 태그가 없음

양산의 슈퍼스타, 삼량빵과 울통치킨

경상남도 양산시

#양산가볼만한곳 #먹방여행 #친구와함께 #가족여행 #이색체험 #시장…

```
AttributeError                            Traceback (most recent call last)
<ipython-input-8-03f418459419> in <module>
     24         print('내용:',contents.strip())
     25
---> 26         tag = i.find('p','tag').get_text()
     27         tags2.append(tag)
     28         print('태그:',tag.strip())

AttributeError: 'NoneType' object has no attribute 'get_text'
```

검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. **검색 개수 기반의 웹 크롤링**
   - **한국관광공사**

Openpyxl 활용하여 엑셀 파일 생성

해시 태그 예외 처리를 프로그래밍

```
        try :
            tag = i.find('p','tag').get_text()
        except AttributeError :
            tag = '태그정보가 없습니다'
            tags2.append(tag)
            print('태그:',tag.strip())
            print("₩n")
        else :
            tags2.append(tag)
            print('태그:',tag.strip())
            print("₩n")

        if no == re_cnt :
            break

        no += 1

    time.sleep(2)        # 페이지 변경 전 2초 대기

    x += 1

    if x == real_page_cnt+1 :
        break

    if x == 6 :
        driver.find_element_by_xpath("""//*[@id="6"]""").click()
    else :
        driver.find_element_by_link_text("""%s""" %x).click()
```

```
● 번호: 8
제목: 무등산편백자연휴양림(안양산자연휴양림)
태그: #자연#무등산편백자연휴양림#안양산자연휴양림#힐링#휴양여행#전라권#관광지
● 번호: 9
제목: 옛 사람들의 자취를 만나며 걷는 길, 무등산옛길 2구간
태그: 태그정보가 없습니다
● 번호: 10
제목: 무등산 정상3봉 (무등산 국가지질공원)
태그: #자연#자연환경#자연관광지#자연발생#자연좋은곳#자연속으로#무등산#경치좋은곳#휴식하기좋은곳#
곳#국가지질공원#무등산권_국가지질공원#지질공원#지질명소
----------------------------------------------------------------------------------------------------
2 페이지 내용 크롤링을 시작합니다

● 번호: 11
제목: 의상봉 (무등산권 국가지질공원)
```

웹 페이지 번호

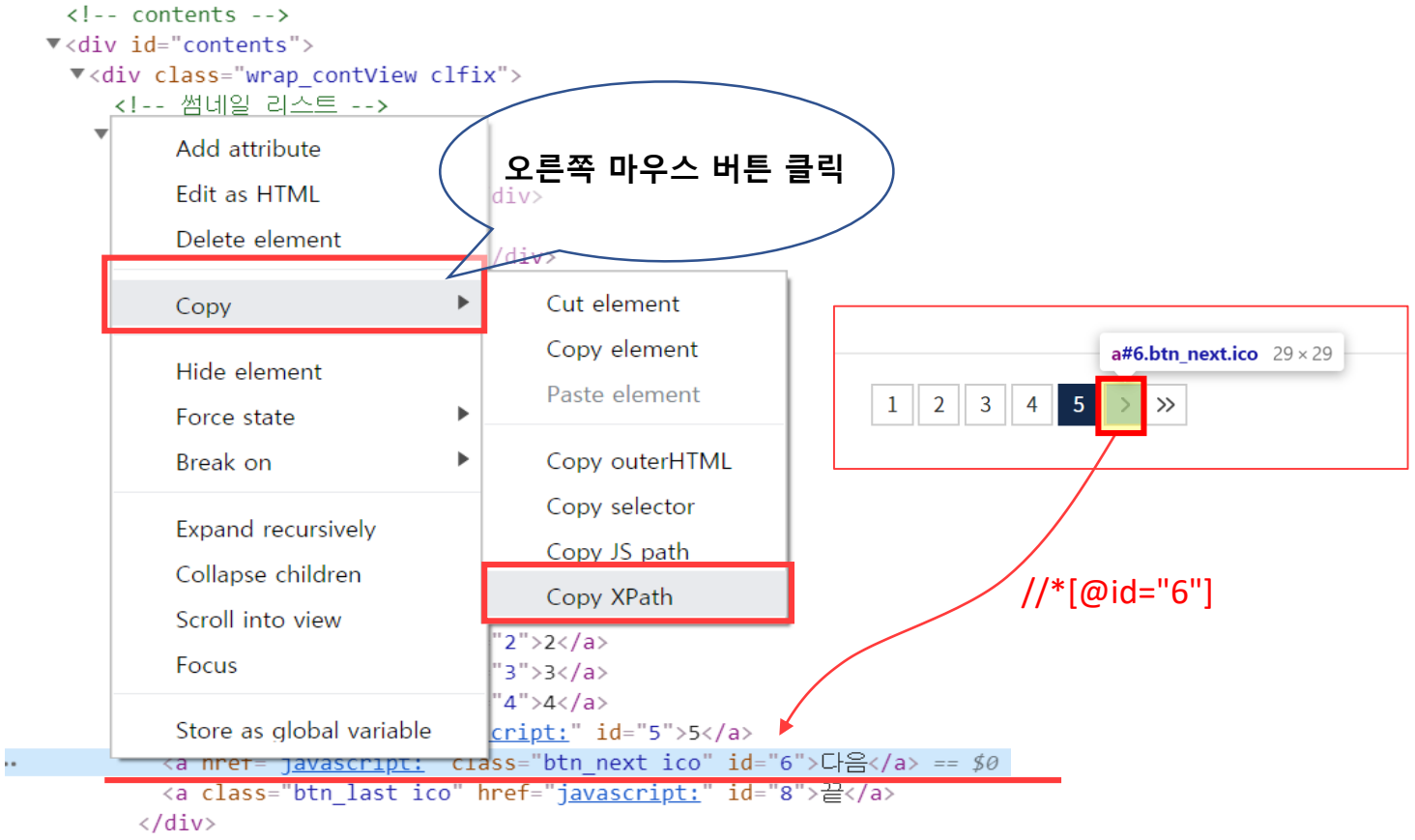
검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. **검색 개수 기반의 웹 크롤링**
   - **한국관광공사**

Openpyxl 활용하여 엑셀 파일 생성

[xpath 를 찾는 방법]



검색 결과를 다양한 형식으로 저장 (한국관광공사 홈페이지 활용)

실습 노트 참고

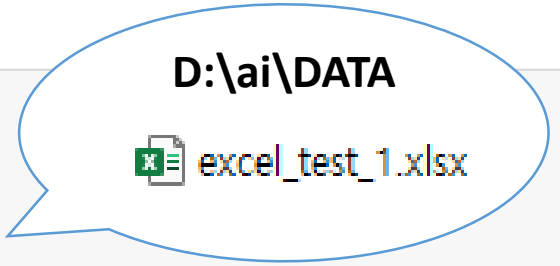
- 09_Web Crawling based on Text data (openpyxl 활용하여 엑셀 파일 생성).ipynb

Openpyxl 활용하여 엑셀 파일 생성

< openpyxl 패키지를 활용한 엑셀 형식의 파일 관리하기 >

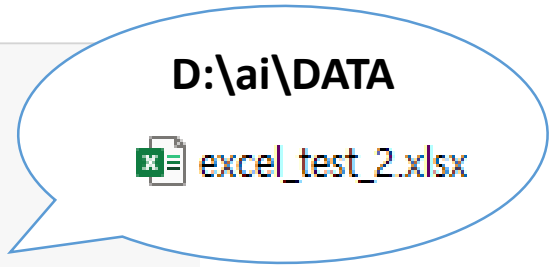
Step 1. 엑셀 파일 만들기

```
1  # openpyxl  패키지 설치되어 있어야 함
2  import openpyxl
3
4  wb = openpyxl.Workbook( )
5  wb.save("D:₩₩ai₩₩data₩₩excel_test_1.xlsx")
```



Step 2. 엑셀 파일 만들기 : 새로운 시트 생성하는 방법, 시트이름 변경하는 방법 포함

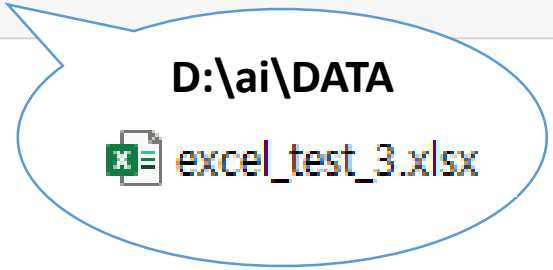
```
1  wb = openpyxl.Workbook( )
2  sheet_1 = wb.active    #활성화된 시트를 sheet_1에 할당
3
4  #또 다른 새로운 시트를 만들면서 시트 이름을 "매출현황"으로 변경
5  sheet_2 = wb.create_sheet("매출현황")
6
7  #기존 시트 이름 변경
8  sheet_1.title = '총매출현황'  #sheet_1에 할당된 시트 이름을 "총매출현황"으로 변경
9
10 wb.save("D:₩₩ai₩₩data₩₩excel_test_2.xlsx")
```



Openpyxl 활용하여 엑셀 파일 생성

Step 3. 엑셀 파일 만들기 : 기존 엑셀 파일 불러와서 cell에 데이터를 입력한 후, 새로운 이름으로 저장하는 방법 포함

```
1  import openpyxl
2
3  wb = openpyxl.load_workbook('D:₩₩ai₩₩data₩₩excel_test_2.xlsx')
4  sheet_1 = wb['총매출현황'] #시트 이름이 "총매출현황"인 시트를 의미함
5  sheet_1['A1'] = '첫번째 cell'
6  sheet_1['A2'] = '두번째 cell'
7
8  wb.save("D:₩₩ai₩₩data₩₩excel_test_3.xlsx")
```



Openpyxl 활용하여 엑셀 파일 생성

텍스트 데이터 웹 크롤링

1. 웹 크롤링 후 텍스트 형식 파일로 저장
   - 한국관광공사
   - 네이버 포스트
2. 웹 크롤링 후 다양한 형식 파일로 저장
   - 한국관광공사
   - 네이버 블로그
3. 검색 개수 기반의 웹 크롤링
   - 한국관광공사

Openpyxl 활용하여 엑셀 파일 생성