

파이썬 크롤링 10~12

2017010715허지혜



목차

Chapter10. 웹 크롤링을 위한 환경 설정과 검색어 자동 실행하기

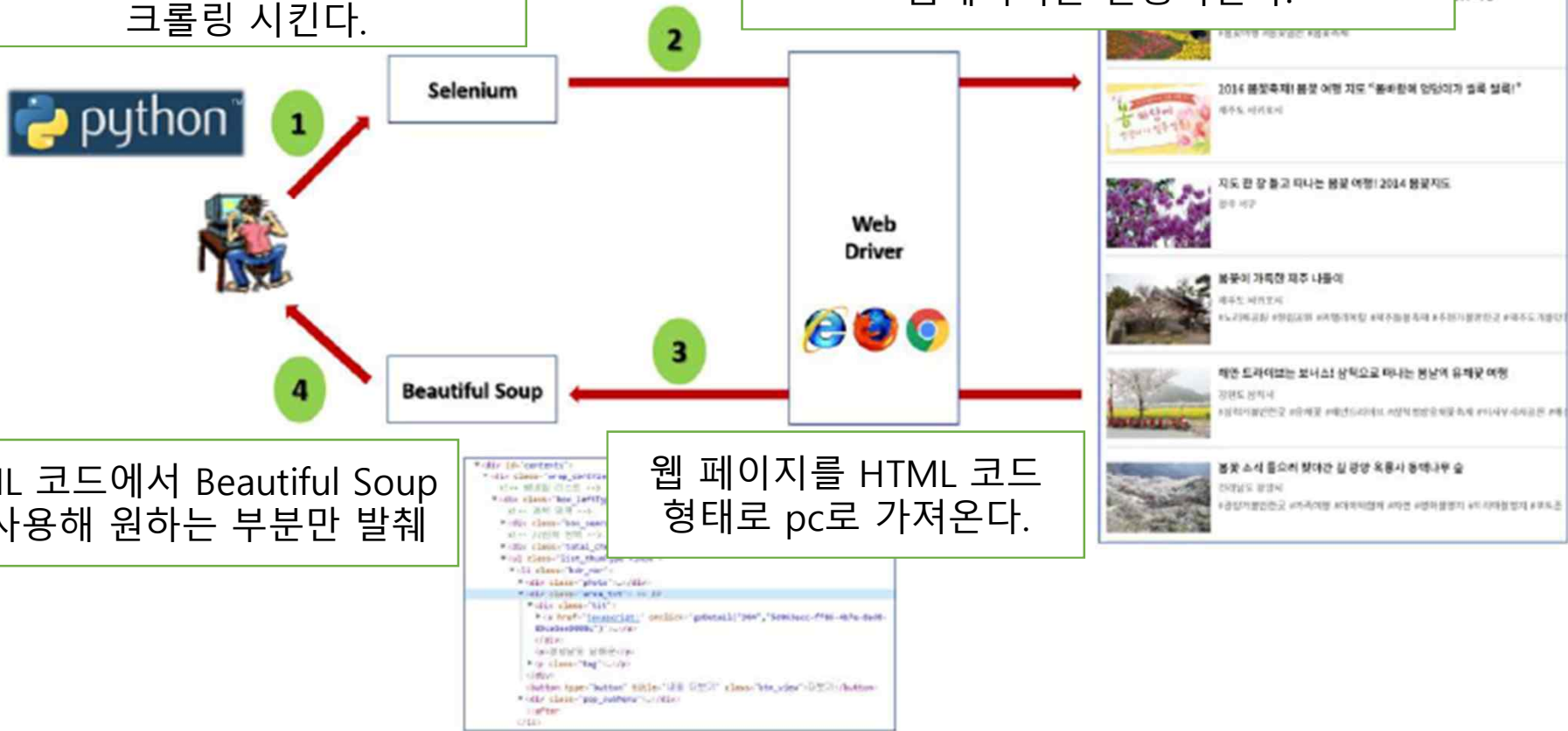
Chapter11. 검색된 결과에서 텍스트를 추출하여 저장하기

Chapter12. 검색된 결과를 다양한 형식의 파일로 저장하기

Chapter10-1) 웹 크롤링을 위한 환경 설정

파이썬 프로그래밍으로
Selenium에게 특정 웹 페이지를
크롤링 시킨다.

Selenium은 지정된 Web Driver를 실행해
웹페이지를 실행시킨다.



HTML 코드에서 Beautiful Soup 를 사용해 원하는 부분만 발취

웹 페이지를 HTML 코드
형태로 pc로 가져온다.

Selenium을 이용한 웹 크롤링 원리

Chapter10-1) 웹 크롤링을 위한 환경 설정

1단계. Beatiful Soup 설치

Beautiful Soup는 파이썬에서 HTML을 찾는 작업에 필수

```
Anaconda Prompt (anaconda3)

(base) C:\Users\stat>pip install bs4
Collecting bs4
  Using cached bs4-0.0.1.tar.gz (1.1 kB)
Requirement already satisfied: beautifulsoup4 in c:\users\stat\anaconda3\lib\site-packages (from bs4) (4.9.3)
Requirement already satisfied: soupsieve>1.2 in c:\users\stat\anaconda3\lib\site-packages (from beautifulsoup4->bs4) (2.1)
Building wheels for collected packages: bs4
  Building wheel for bs4 (setup.py) ... done
  Created wheel for bs4: filename=bs4-0.0.1-py3-none-any.whl size=1273 sha256=0356d42ad0b0bba2f76f241308753bf3eb7d15d516acd464957615b1e24d020d
  Stored in directory: c:\users\stat\appdata\local\pip\cache\wheels\0a\9e\ba\20e5bbc1afef3a491f0b3bb74d508f99403aabe76eda2167ca
Successfully built bs4
Installing collected packages: bs4
Successfully installed bs4-0.0.1
```

Anaconda Prompt 또는 cmd창에 pip install bs4

2단계. Selenium 설치

Selenium은 사람 대신 웹페이지를 열고 데이터를 수집

```
Anaconda Prompt (anaconda3)

(base) C:\Users\stat>pip install selenium
Collecting selenium
  Using cached selenium-3.141.0-py2.py3-none-any.whl (904 kB)
Requirement already satisfied: urllib3 in c:\users\stat\anaconda3\lib\site-packages (from selenium) (1.26.2)
Installing collected packages: selenium
Successfully installed selenium-3.141.0
```

Anaconda Prompt 또는 cmd창에 pip install selenium

Chapter10-1) 웹 크롤링을 위한 환경 설정

3단계. 웹 드라이버 설치

1

<https://sites.google.com/a/chromium.org/chrome-driver/downloads>

2

Chrome 버전
맞는 버전 들0

ChromeDriver 87.0.4280.20






Supports Chrome version 87

- Resolved issue 2421: Delete old port-forwarding channels on android adb-server
- Resolved issue 3474: Emulated mobile device list needs updating
- Resolved issue 3507: Implement "get computed role"
- Resolved issue 3508: Implement "get computed label"
- Resolved issue 3584: Rename ChromeDriver command line option --whitelisted-ips
- Resolved issue 3588: Bidi WebSocket connection
- Resolved issue 3594: Navigation completes prematurely if OOPIF loads before main page
- Resolved issue 3598: A command line option for devtools port to be forwarded to webview_devtools_remote socket

3

Chromedriver
win32.zip 클릭

Index of /81.0.4044.69/

	Name	Last modified	Size	ETag	
	Parent Directory		-		
	chromedriver linux64.zip	2020-03-17 16:16:51	4.73MB	11bc281b27db997b5045b376866b8ed5	
	chromedriver mac64.zip	2020-03-17 16:16:53	6.69MB	2d27f0b1b4cdc9e7f2e535a88223efbb	
	chromedriver win32.zip	2020-03-17 16:16:54	4.19MB	e6006040f914e704f591a6abdb3833ef	<input checked="" type="checkbox"/>
	notes.txt	2020-03-17 16:25:31	0.00MB	adf2a9dacb0ae755b7328973b31271d2	<input checked="" type="checkbox"/>

4

C:\Temp 폴더에 "chromedriver_240" 라는 폴더를 생성 , "chromedriver win32.zip" 를 풀어서 저장!

Chapter10-2) 웹 크롤링 검색어 자동 실행

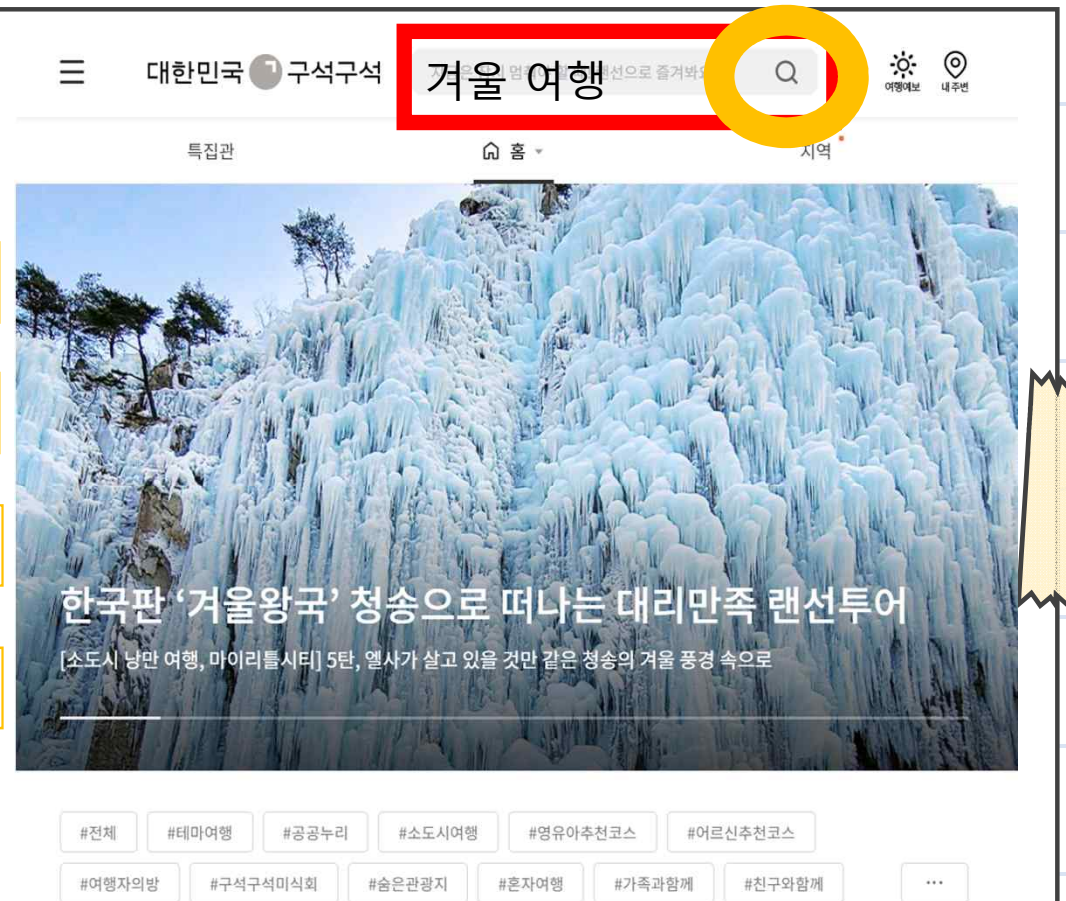
Selenium

웹페이지를 열어라

검색창을 찾아라

검색어를 입력해라

조회하라



참고 예제 URL : <https://korean.visitkorea.or.kr/main/main.do>

Chapter10-2) 웹 크롤링 검색어 자동 실행

```
In [20]: 1 # 1. 필요한 라이브러리 로드
2 from bs4 import BeautifulSoup
3 from selenium import webdriver
4 import time
5
6 #2. 검색어 입력 받기
7 query_txt = input('크롤링할 키워드는 무엇입니까?: ')
8
9 #3. 크롬 드라이버를 사용해서 웹 브라우저 실행.
10 path = "C:/Temp/chromedriver_240/chromedriver_win321/chromedriver.exe"
11 driver = webdriver.Chrome(path)
12
13 driver.get("https://korean.visitkorea.or.kr/main/main.html")
14 time.sleep(10) # 위 페이지가 모두 열릴 때 까지 10초 기다림.
15
16 #4. 팝업창(?) 지우기
17 driver.find_element_by_id("chkForm01").click()
18 #driver.find_element_by_xpath("//*[id='chkForm01']").click()
19 driver.find_element_by_xpath("//*[id='safetyStay1']/div[1]/div/div/button").click()
20
21 #5. 검색창의 이름을 찾아 검색어 입력
22 driver.find_element_by_id("gnbMain").click()
23 element = driver.find_element_by_id("inp_search")
24 element.send_keys(query_txt)
25
26 #6. 검색 버튼 실행
27 driver.find_element_by_link_text("검색").click()
28 #driver.find_element_by_class_name("btn_search2").click() # class name
29 #driver.find_element_by_xpath("//*[id='gnbMain']/div/div/div[1]/div[1]/a').click() # xpath
```

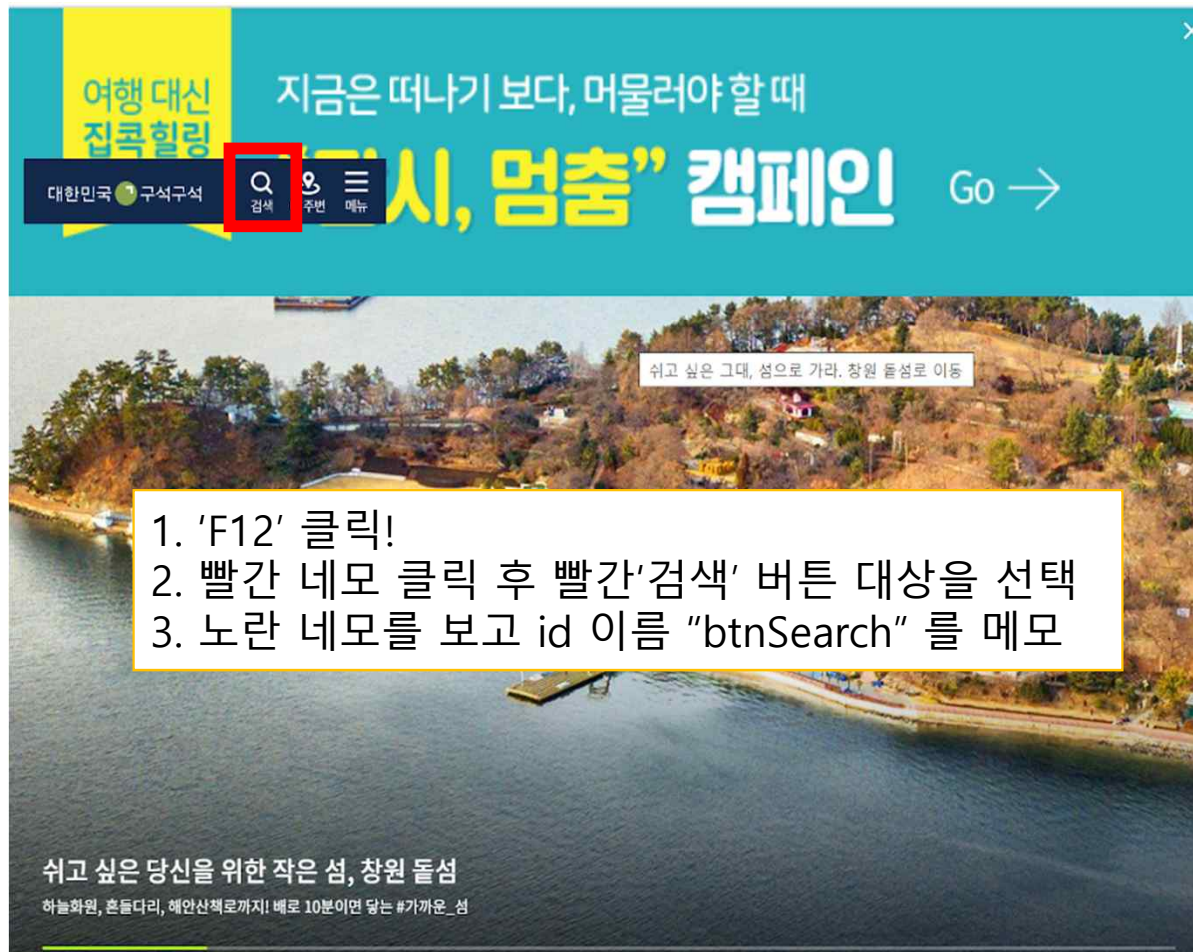
크롤링할 키워드는 무엇입니까?: 겨울 여행

Chapter10-2) 웹 크롤링 검색어 자동 실행

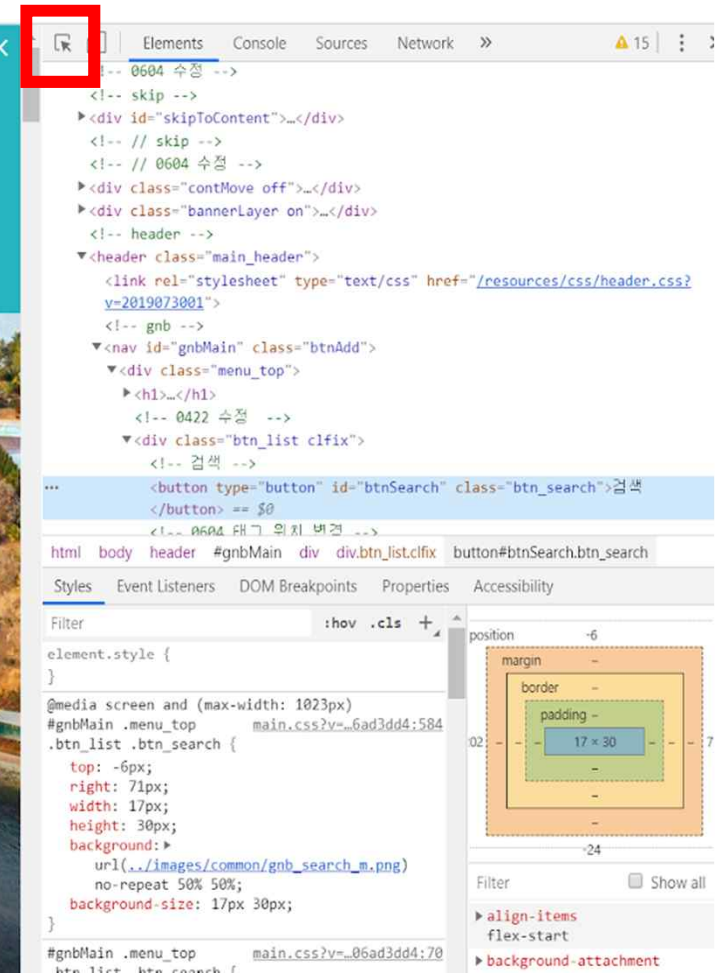
```
In [23]: 1 # 1. 필요한 라이브러리 로드
2 from bs4 import BeautifulSoup
3 from selenium import webdriver
4 import time
5
6 #2. 검색어 입력 받기
7 query_txt = input('크롤링할 키워드는 무엇입니까?: ')
8
9 #3. 크롬 드라이버를 사용해서 웹 브라우저 실행.
10 path = "C:/Temp/chromedriver_240/chromedriver_win32/chromedriver.exe"
11 driver = webdriver.Chrome(path)
12
13 driver.get("https://korean.visitkorea.or.kr/main/main.html")
14 time.sleep(10) # 위 페이지가 모두 열릴 때 까지 10초 기다림.
```

크롤링할 키워드는 무엇입니까?:

Chapter10-2) 웹 크롤링 검색어 자동 실행



1. 'F12' 클릭!
2. 빨간 네모 클릭 후 빨간 '검색' 버튼 대상을 선택
3. 노란 네모를 보고 id 이름 "btnSearch" 를 메모



Chapter10-2) 웹 크롤링 검색어 자동 실행

```
16 #4. 팝업창(?) 지우기
17 driver.find_element_by_id("chkForm01").click()
18 #driver.find_element_by_xpath("//*[@id="chkForm01"]").click()
19 driver.find_element_by_xpath("//*[@id="safetyStay1"]/div[1]/div/div/button").click()
20
```

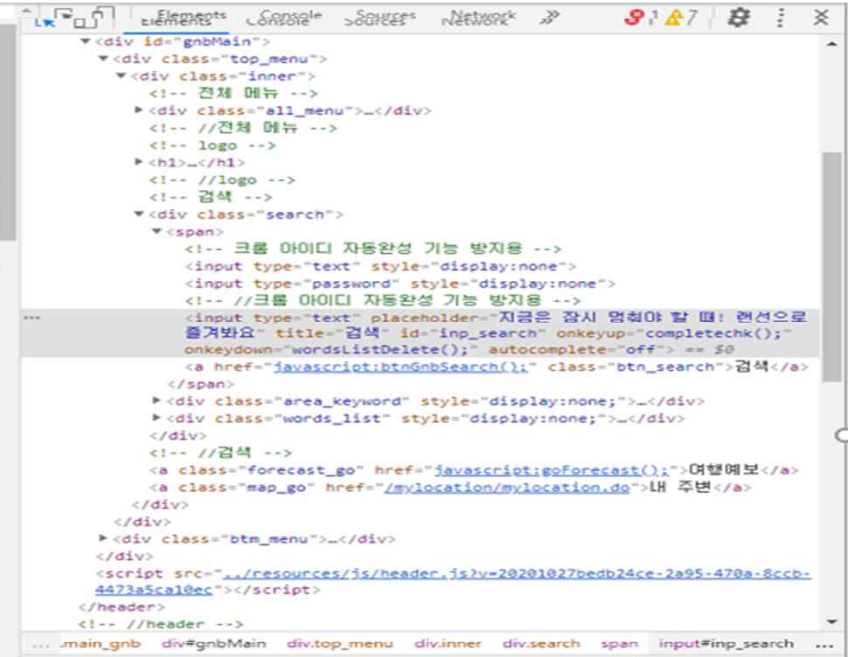


▼<div class="viewNone">

*** <input type="checkbox" tabindex="2" id="chkForm01" name=""> == \$0

<button type="button" tabindex="3" onclick="closeWin(1);">닫기
</button> == \$0

Chapter10-2) 웹 크롤링 검색어 자동 실행



```
20
21 #5. 검색창의 이름을 찾아 검색어 입력
22 driver.find_element_by_id("gnbMain").click()
23 element = driver.find_element_by_id("inp_search")
24 element.send_keys(query_txt)
25
26 #8. 검색 버튼 실행
27 driver.find_element_by_link_text("검색").click()
28 #driver.find_element_by_class_name("btn_search2").click() # class name
29 #driver.find_element_by_xpath('//*[@id="gnbMain"]/div/div/div[1]/div[1]/a').click() # xpath
```

크롤링할 키워드는 무엇입니까?: 겨울 여행

Chapter11-1) 웹 크롤링 검색 결과에서 텍스트 추출

```
In [4]: 1 # 1. 필요한 라이브러리 로드
2 from bs4 import BeautifulSoup
3 from selenium import webdriver
4 import time
5 import sys
6
7 #2. 검색어 입력 받기
8 query_txt = input('크롤링할 키워드는 무엇입니까?: ')
9 f_name = input('검색 결과를 저장할 파일경로와 이름을 지정하세요(예:c:\data\test.txt): ')
10
11 #3. 크롬 드라이버를 사용해서 웹 브라우저 실행.
12 path = "C:/Temp/chromedriver_240/chromedriver_win32/chromedriver.exe"
13 driver = webdriver.Chrome(path)
14 driver.get("https://korean.visitkorea.or.kr/main/main.html")
15 time.sleep(2) # 창이 모두 열릴 때 까지 2초 기다림.
16
17 #4. 팝업창(?) 지우기
18 driver.find_element_by_id("chkForm01").click()
19 #driver.find_element_by_xpath("//*[@id='chkForm01']").click()
20 driver.find_element_by_xpath("//*[@id='safetyStay1']/div[1]/div/div/button").click()
21
22 #5. 검색창의 이름을 찾아 검색어 입력
23 driver.find_element_by_id("gnbMain").click()
24 element = driver.find_element_by_id("inp_search")
25 element.send_keys(query_txt)
26
27 #6. 검색 버튼 실행
28 driver.find_element_by_link_text("검색").click()
29
30 #7. 텍스트 추출해 화면에 출력하기.
31 time.sleep(1)
32
33 full_html = driver.page_source
34
35 soup = BeautifulSoup(full_html, 'html.parser')
36
37 content_list = soup.find('ul', class_='list_thumType type1')
38
39 for i in content_list:
40     print(i.text.strip())
41     print("#n")
42
43 #8. 텍스트를 추출하여 txt 형식으로 저장하기.
44 orig_stdout = sys.stdout
45 f = open(f_name, 'a', encoding='UTF-8')
46 sys.stdout = f
47 time.sleep(1)
48
49 html = driver.page_source
50 soup = BeautifulSoup(html, 'html.parser')
51 content_list = soup.find('ul', class_='list_thumType type1')
52
53 for i in content_list:
54     print(i.text.strip())
55     print("#n")
56
57 sys.stdout = orig_stdout
58 f.close()
59
60 print(" 요청하신 데이터 수집 작업이 정상적으로 완료되었습니다")
```

Chapter11-1) 웹 크롤링 검색 결과에서 텍스트 추출

```
In [4]: 1 # 1. 필요한 라이브러리 로드
2 from bs4 import BeautifulSoup
3 from selenium import webdriver
4 import time
5 import sys
6
7 #2. 검색어 입력 받기
8 query_txt = input('크롤링할 키워드는 무엇입니까?: ')
9 f_name = input('검색 결과를 저장할 파일경로와 이름을 지정하세요(예 :c:\data\test.txt): ')
```

크롤링할 키워드는 무엇입니까?:

```
10 driver = webdriver.Chrome(executable_path=...)
14 driver.get("https://korean.visitkorea.or.kr/main/main.html")
15 time.sleep(2) # 대기 모두 열릴 때 까지 2초 기다림.
```

검색 결과를 저장할 파일경로와 이름을 지정하세요(예 :c:\data\test.txt):

```
20 driver.find_element_by_xpath("//*[id='safetyStay1']/div[1]/div/div/button").click()
21
22 #5. 검색창의 이름을 찾아 검색어 입력
23 driver.find_element_by_id("gnbMain").click()
24 element = driver.find_element_by_id("inp_search")
25 element.send_keys(query_txt)
26
27 #8. 검색 버튼 실행
28 driver.find_element_by_link_text("검색").click()
```


Chapter11-1) 웹 크롤링 검색 결과에서 텍스트 추출

텍스트 데이터 추출

현재 페이지에서 BeautifulSoup를 사용해 검색 결과가 들어있는 게시글의 HTML 코드의 태그를 추출

```
31 time.sleep(1)
32
33 full_html = driver.page_source
34
35 soup = BeautifulSoup(full_html, 'html.parser')
36
37 content_list = soup.find('ul', class_='list_thumType type1')
38
```

The screenshot shows a web browser displaying search results for '겨울 여행' (Winter Travel). A red box highlights a search result card. A red arrow points from the 'Elements' panel in the browser's developer tools to the highlighted card. The 'Elements' panel shows the HTML structure, with a red box highlighting the 'ul' element with class 'list_thumType type1'.

HTML structure shown in the Elements panel:

```
<ul class="list_thumType type1">
  <li></li>
  <li></li>
  <li></li>
  <li></li>
  <li class="Thuman_arrow"></li>
</ul>
```

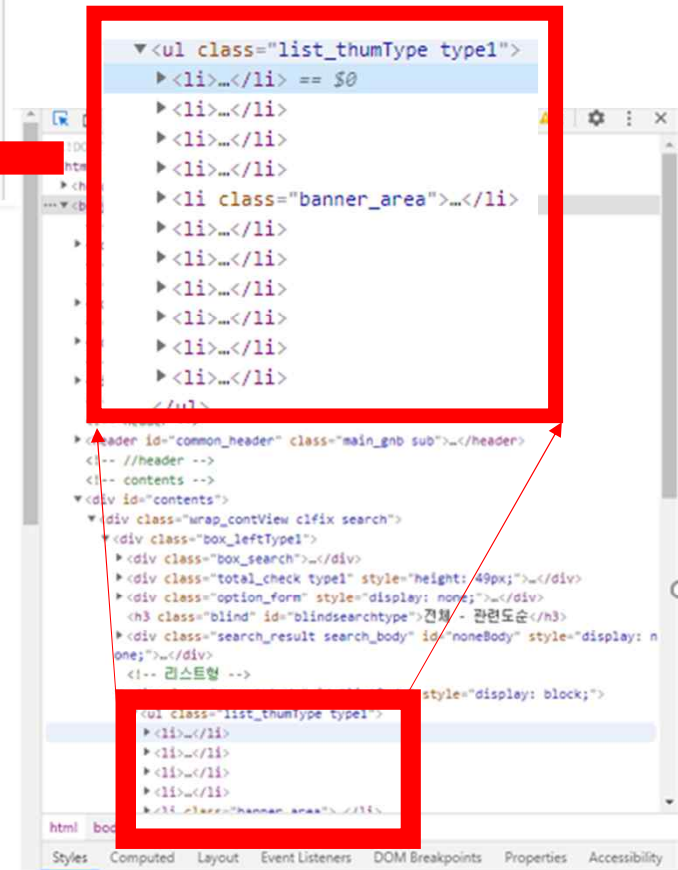
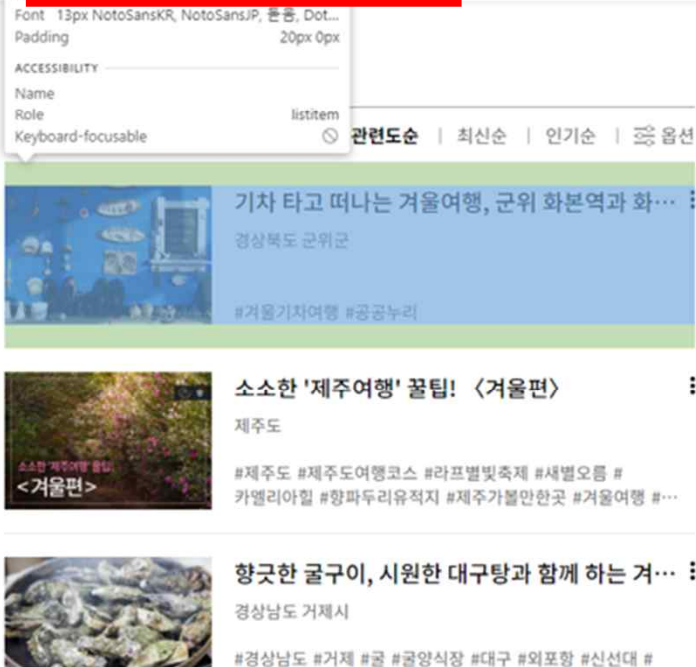

Chapter11-1) 웹 크롤링 검색 결과에서 텍스트 추출

텍스트 데이터 추출

현재 페이지에서 BeautifulSoup를 사용해 검색 결과가 들어있는 게시글의 HTML 코드의 태그를 추출

```
31 time.sleep(1)
32 full_html = driver.page_source
33
34 soup = BeautifulSoup(full_html, 'html.parser')
35
36 content_list = soup.find('ul', class_='list_thumType type1')
37
38
39 for i in content_list:
40     print(i.text.strip())
41     print("\n")
```

반복문
양쪽 끝 공백을 제거하고 출력



Chapter11-1) 웹 크롤링 검색 결과에서 텍스트 추출

크롤링할 키워드는 무엇입니까?: 겨울 여행

검색 결과를 저장할 파일경로와 이름을 지정하세요 (예 : c:\data\test.txt) : C:\Users\stat\Desktop\허지혜\수DA쟁이\완친파\test.txt
기차 타고 떠나는 겨울여행, 군위 화본역과 화본마을 경상북도 군위군 #겨울기차여행#공공누리 더보기 즐겨찾기 공유하기

겨울여행지 추천, 느린 만큼 매력적인 곳, 1박2일 안동 여행 코스 경상북도 안동시 #안동여행#겨울여행#새해맞이여행#안동하회마을#월영교#월영정#안동찜닭#안동간고등어#안동구시장#새해여행#겨울가볼만한곳#겨울소확행 더보기 즐겨찾기 공유하기

소소한 '제주여행' 꿀팁! <겨울편> 제주도 #제주도#제주도여행코스#라프별빛축제#새별오름#카멜리아힐#향파두리유적지#제주가볼만한곳#겨울여행#감성여행#겨울감성여행#동백꽃#불빛축제#오름#1월가볼만한곳 더보기 즐겨찾기 공유하기

새하얀 눈이 내리면 더 아름다워지는 부여 겨울여행 코스 충청남도 부여군 #부여가볼만한곳#부여여행#겨울여행#부소산성#백제문화단지#부여능산리고분군#정림사지#장원막국수#풍경여행#1월가볼만한곳 더보기 즐겨찾기 공유하기

향긋한 굴구이, 시원한 대구탕과 함께 하는 겨울 거제여행 경상남도 거제시 #경상남도#거제#굴#굴양식장#대구#외포항#신선대#바람의언덕#해금강테마박물관#몽돌해변#드라이브#바다#겨울#힐링#여행#가족#친구#연인#추천가볼만한곳#거제가볼만한곳#거제당일코스#거제1박2일코스#겨울제철음식#미식여행#굴무침#굴구이#대구회무침#대구탕#겨울먹거리#2015년12월추천가볼만한곳#12월추천가볼만한곳 더보기 즐겨찾기 공유하기

겨울 특집 용인여행, 전통과 현대가 공존하는 용인으로! 경기도 용인시 #용인여행#용인가볼만한곳#용인5일장#한국민속촌#장옥진고택#머로프슬라임스피스#특별한_겨울나들이 더보기 즐겨찾기 공유하기

"월 좋아할지 몰라서 다 준비했어!" 겨울방학 평창-강릉 가족 여행법 강원도 평창군 #평창가볼만한곳#강원도여행#가족여행#체험학습#이색체험#아이와함께#겨울방학가볼만한곳#강릉가볼만한곳#온가족_겨울여행 더보기 즐겨찾기 공유하기

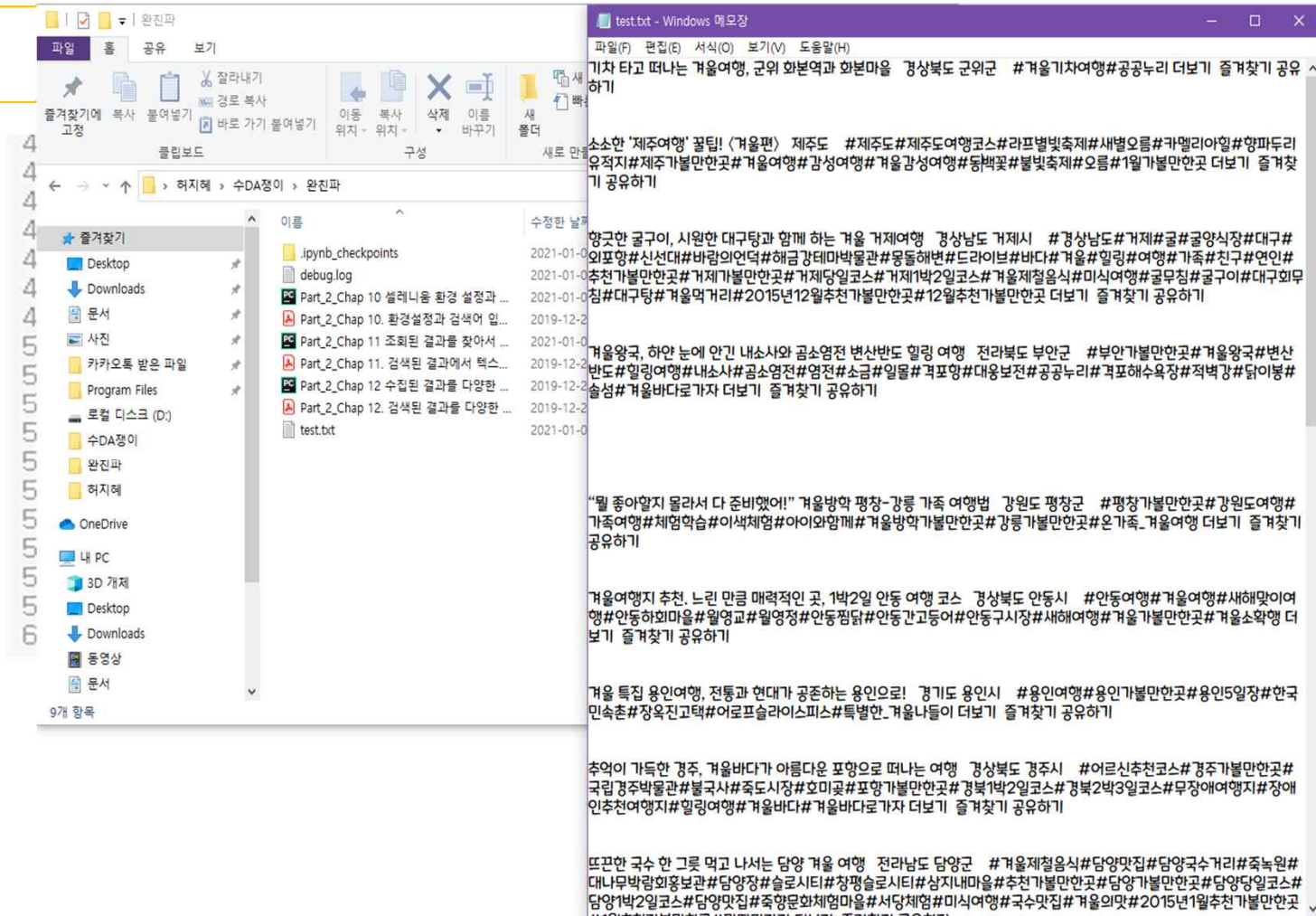
추억이 가득한 경주, 겨울바다가 아름다운 포항으로 떠나는 여행 경상북도 경주시 #머르신추천코스#경주가볼만한곳#국립경주박물관#불국사#죽도시장#호미곶#포항가볼만한곳#경북1박2일코스#경북2박3일코스#무장애여행지#장애인추천여행지#힐링여행#겨울바다#겨울바다로가자 더보기 즐겨찾기 공유하기

뜨끈한 국수 한 그릇 먹고 나서는 담양 겨울 여행 전라남도 담양군 #겨울제철음식#담양맛집#담양국수거리#죽녹원#대나무박람회#홍보관#담양장#슬로시티#찰평슬로시티#삼지내마을#추천가볼만한곳#담양가볼만한곳#담양당일코스#담양1박2일코스#담양맛집#죽향문화체험마을#서당체험#미식여행#국수맛집#겨울의맛#2015년1월추천가볼만한곳#1월추천가볼만한곳#맛따라가기 더보기 즐겨찾기 공유하기

겨울 여행의 추억 만들기 #한국관광품질인증#여행자의방#숲속의요정#전강원관광농원#소낭구펜션#옛마실펜션#꿈꾸는노마드#더세리리조트#메모리즈호텔#애플리트#목포마리나베미호텔#겨울여행#한국관광품질인증테마여행#우수숙소#인증숙소#평창숙소#거제숙소#제주숙소#숙초숙소#목포숙소#포항숙소#양양숙소 더보기 즐겨찾기 공유하기

요청하신 데이터 수집 작업이 정상적으로 완료되었습니다

Chapter11-2) 웹 크롤링 검색 결과에서 텍스트 저장



The image shows a Windows File Explorer window on the left and a Notepad window on the right. The File Explorer window displays the contents of a folder named '수DA정이' (Suda Jeong-i), which includes files like 'Part_2_Chap 10', 'Part_2_Chap 11', 'Part_2_Chap 12', and 'test.txt'. The Notepad window, titled 'test.txt - Windows 메모장', contains a large block of text that appears to be a list of search results or a log of web crawling data. The text is organized into several paragraphs, each starting with a date and time stamp (e.g., '2021-01-01 14:00:00'). The text describes various travel-related topics, including destinations like Jeju Island, Jeonnam, and Jeonbuk, and activities like hiking, sightseeing, and dining. The text is written in a casual, conversational style, typical of a travel blog or a personal diary.

수소한 '제주여행' 꿀팁! <겨울편> 제주도 #제주도#제주도여행코스#라프별빛축제#새별오름#카멜리아힐#양파두리
유적지#제주가볼만한곳#겨울여행#감성여행#겨울감성여행#동백꽃#불빛축제#오름#1월가볼만한곳 더보기 즐겨찾
기 공유하기

영국한 굴구이, 시원한 대구탕과 함께 하는 겨울 겨울여행 경상남도 거제시 #경상남도#거제#굴#굴양식장#대구#
이포항#신선대#바람의언덕#해금강테마박물관#몽돌해변#드라이브#바다#겨울#일링#여행#가족#친구#연인#
추천가볼만한곳#거제가볼만한곳#거제당일코스#거제1박2일코스#겨울제철음식#미식여행#굴무침#굴구이#대구회무
침#대구탕#겨울먹거리#2015년12월추천가볼만한곳#12월추천가볼만한곳 더보기 즐겨찾기 공유하기

겨울왕국, 하얀 눈에 안긴 내소사와 금소염전 변산반도 일링 여행 전라북도 부안군 #부안가볼만한곳#겨울왕국#변산
반도#일링여행#내소사#금소염전#염전#소금#일몰#격포항#대응보전#공공누리#격포해수욕장#적벽강#달이봉#
슬썹#겨울바다로가자 더보기 즐겨찾기 공유하기

"말 좋아할지 몰라서 다 준비했어!" 겨울방학 평창-강릉 가족 여행법 강원도 평창군 #평창가볼만한곳#강원도여행#
가족여행#체험학습#이색체험#아이와함께#겨울방학가볼만한곳#강릉가볼만한곳#온가족_겨울여행 더보기 즐겨찾기
공유하기

겨울여행지 추천. 느린 만큼 매력적인 곳, 1박2일 안동 여행 코스 경상북도 안동시 #안동여행#겨울여행#새해맞이여
행#안동하회마을#월영교#월영정#안동찜닭#안동간고등어#안동구시장#새해여행#겨울가볼만한곳#겨울소확행 더
보기 즐겨찾기 공유하기

겨울 특집 용인여행, 전통과 현대가 공존하는 용인으로! 경기도 용인시 #용인여행#용인가볼만한곳#용인5일장#한국
민속촌#장육진고택#어로프슬라이스피스#특별한_겨울나들이 더보기 즐겨찾기 공유하기

추억이 가득한 경주, 겨울바다가 아름다운 포항으로 떠나는 여행 경상북도 경주시 #여르신추천코스#경주가볼만한곳#
국립경주박물관#불국사#죽도시장#호미곶#포항가볼만한곳#경북1박2일코스#경북2박3일코스#무장애여행지#장애
인추천여행지#일링여행#겨울바다#겨울바다로가자 더보기 즐겨찾기 공유하기

뜨끈한 국수 한 그릇 먹고 나서는 담양 겨울 여행 전라남도 담양군 #겨울제철음식#담양맛집#담양국수거리#죽녹원#
대나무박물관#용보관#담양정#슬로시티#정평슬로시티#삼지내마을#추천가볼만한곳#담양가볼만한곳#담양당일코스#
담양1박2일코스#담양맛집#죽향문화체험마을#서당체험#미식여행#국수맛집#겨울의맛#2015년1월추천가볼만한곳
#12월추천가볼만한곳#12월추천가볼만한곳#12월추천가볼만한곳

Chapter12-1) 특정 항목 분리 후 추출

```
In [1]: 1 #1 1. 필요한 라이브러리 로드
2 from bs4 import BeautifulSoup
3 from selenium import webdriver
4 import time
5 import sys
6
7 # 2. 검색어 입력 받기
8 query_txt = input('크롤링할 키워드는 무엇입니까?: ')
9 f_name = input('검색 결과를 저장할 txt 파일경로와 이름을 지정하세요(예:c:\\data\\test_3.txt):')
10 fc_name = input('검색 결과를 저장할 csv 파일경로와 이름을 지정하세요(예:c:\\data\\test_3.csv):')
11 fx_name = input('검색 결과를 저장할 xls 파일경로와 이름을 지정하세요(예:c:\\data\\test_3.xls):')
12
13 # 3. 크롬 드라이버를 사용해서 웹 브라우저 실행
14 path = "C:/Temp/chromedriver_240/chromedriver_win32/chromedriver.exe"
15 driver = webdriver.Chrome(path)
16
17 driver.get("https://korean.visitkorea.or.kr/main/main.html")
18 time.sleep(2) # 네이버 창이 모두 열릴 때 까지 2초 기다림..
19
20 # 5. 검색창의 이름을 찾아 검색어 입력, 검색 버튼 실행
21 driver.find_element_by_id("btnSearch").click()
22 element = driver.find_element_by_id("inp_search")
23 element.send_keys(query_txt)
24 driver.find_element_by_link_text("검색").click()
25
26 # 8. 현재 페이지에 있는 내용 화면 출력
27 time.sleep(1)
28
29 html = driver.page_source
30 soup = BeautifulSoup(html, 'html.parser')
31 content_list = soup.find('ul', class_='list_thumType flnon')
32 print(content_list)
33
34 # 7. 특정 항목들을 분리해서 추출하기
35 contents = content_list.find('div', 'tit').get_text()
36 print('내용:', contents.strip())
37
38 tag = content_list.find('p', 'tag').get_text()
39 print('태그:', tag.strip())
40 print("\n")
41
42 # 8. 각 항목별로 분리하여 추출하고 변수에 할당하기
43 no = 1
44 no2 = [ ]
45 contents2 = [ ]
46 tags2 = [ ]
47
48 for i in content_list:
49     no2.append(no)
50     print('번호:', no)
51
52     contents = i.find('div', 'tit').get_text()
53     contents2.append(contents)
54     print('내용:', contents.strip())
55
56     tag = i.find('p', 'tag').get_text()
57     tags2.append(tag)
58     print('태그:', tag.strip())
59     print("\n")
60
61     no += 1
62
63
```

Chapter12-1) 특정 항목 분리 후 추출

```
65 # 9. 출력 결과는 데이터 프레임 형태로 만들기
66 import pandas as pd
67
68 korea_trip = pd.DataFrame()
69 korea_trip['번호'] = no2
70 korea_trip['내용'] = contents2
71 korea_trip['태그'] = tags2
72
73 # csv 형태로 저장하기
74 korea_trip.to_csv(fc_name, encoding="utf-8-sig")
75 print(" csv 파일 저장 경로: %s" % fc_name)
76
77 # 엑셀 형태로 저장하기
78 import xlwt # pip install xlwt 실행 후 수행
79 korea_trip.to_excel(fx_name)
80 print(" xls 파일 저장 경로: %s" % fx_name)
81
82 # 출력 결과를 txt 파일로 저장하기
83 f = open(f_name, 'a', encoding='UTF-8')
84 f.write(str(contents2))
85 f.write(str(tags2))
86 f.close()
87 print(" txt 파일 저장 경로: %s" % f_name)
```

```
88 # 10. openpyxl 패키지를 활용한 엑셀 형식의 파일 관리하기
89 import openpyxl
90
91 # 새로운 엑셀 파일을 1개 생성.
92 wb = openpyxl.Workbook()
93 wb.save("c:\\temp\\temp\\test1.xlsx")
94
95 # 새로운 시트 생성하고 시트이름 변경
96 import openpyxl
97 wb = openpyxl.Workbook()
98
99 sheet_1 = wb.active # 현재 활성화 된 sheet 가져오기
100
101 # 새로운 시트를 만들면서 시트 이름을 지정
102 sheet_2 = wb.create_sheet("매출현황")
103
104 # 시트 이름 변경
105 sheet_1.title = '총매출현황'
106
107 wb.save("c:\\temp\\temp\\test2.xlsx")
108
109 # Step 3. 기존 파일 불러와서 cell 에 내용 입력
110 import openpyxl
111
112 wb = openpyxl.load_workbook('c:\\temp\\temp\\test2.xlsx')
113 sheet_1 = wb['총매출현황']
114 sheet_1['A1'] = '첫번째 cell'
115 sheet_1['A2'] = '두번째 cell'
116
117 wb.save("c:\\temp\\temp\\test2.xlsx")
```

Chapter12-1) 특정 항목 분리 후 추출

```
In [1]: 1 #1 1. 필요한 라이브러리 로드  
2 from bs4 import BeautifulSoup
```

크롤링할 키워드는 무엇입니까?:

```
7 # 2. 검색어 입력 받기
```

검색 결과를 저장할 파일경로와 이름을 지정하세요(예 :c:\data\test.txt):

```
13 # 3. 크롤 드라이버를 사용해서 웹 브라우저 실행  
14 path = "C:/Tools/chromedriver_240/chromedriver_win32/chromedriver.exe"
```

검색 결과를 저장할 csv 파일경로와 이름을 지정하세요(예 :c:\data\test_3.csv):

```
20 # 5. 검색창의 이름을 찾아 검색어 입력, 검색 버튼 실행  
21 driver.find_element_by_id("btnSearch").click()  
22 element = driver.find_element_by_id("inp_search")
```

검색 결과를 저장할 xls 파일경로와 이름을 지정하세요(예 :c:\data\test_3.xls):

```
28  
29 html = driver.page_source  
30 soup = BeautifulSoup(html, 'html.parser')  
31 content_list = soup.find('ul', class_='list_thumType flnon')  
32 print(content_list)  
33
```


Chapter12-1) 특정 항목 분리 후 추출

```
33
34 # 7. 특정 항목들을 분리해서 추출하기
35 contents = content_list.find('div','tit').get_text( )
36 print('내용:',contents.strip())
37
38 tag = content_list.find('p','tag').get_text()
39 print('태그:',tag.strip())
40 print("#n")
41
42 # 8. 각 항목별로 분리하여 추출하고 변수에 할당하기
43 no = 1
44 no2 = [ ]
45 contents2=[ ]
46 tags2=[ ]
47
48 for i in content_list:
49     no2.append(no)
50     print('번호:',no)
51
52     contents = i.find('div','tit').get_text( )
53     contents2.append(contents)
54     print('내용:',contents.strip())
55
56     tag = i.find('p','tag').get_text()
57     tags2.append(tag)
58     print('태그:',tag.strip())
59     print("#n")
60
61     no += 1
62
63
```

Chapter12-1) 특정 항목 분리 후 추출

이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.

대한민국 구석구석 겨울 여행

전체 | 여행정보 | 여행기사 | 축제

관련도순 | 최신순 | 인기순 | 중 옵션

뜨끈한 국수 한 그릇 먹고 나서는 담양 겨울 ...

전라남도 담양군

#겨울제철음식 #담양맛집 #담양국수거리 #죽녹원 #대나무박람회홍보관 #담양장 #슬로시티 #창평슬로시티 #...

겨울 여행의 추억 만들기

#한국관광품질인증 #여행자의방 #숲속의요정 #정강원관광농원 #소낭구펜션 #옛마실펜션 #꿈꾸는노마드 #더세리리조트 #...

기차 타고 떠나는 겨울여행, 군위 화본역과 화...

경상북도 군위군

어제의 인기 검색어

- 1 겨울여행
- 2 겨울맛
- 3 드라이브
- 4 2021
- 5 공원
- 6 둘레길
- 7 함안
- 8 온천노천탕
- 9 속초
- 10 한국민속촌

찾으시는 정보가 없으신가요?

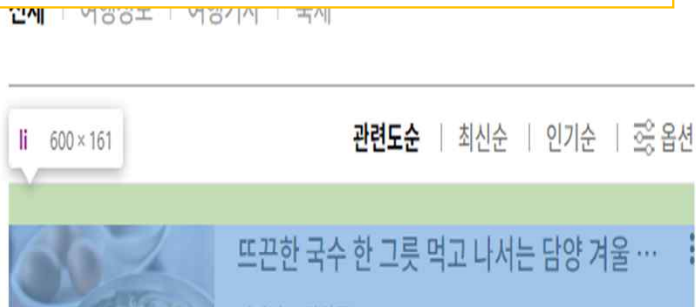
Elements Console Sources Network

```
<div class="wrap_contView clfix search">
  <div class="box_leftType1">
    <div class="box_search"></div>
    <div class="total_check type1" style="height: 49px;"></div>
    <div class="option_form" style="display: none;"></div>
    <h3 class="blind" id="blindsearchtype">전체 - 관련도순</h3>
    <div class="search_result search_body" id="noneBody" style="display: none;"></div>
    <!-- 리스트형 -->
    <div class="search_body" id="listBody" style="display: block;">
      <ul class="list_thumType type1">
        <li>
          <a href="javascript:goSearchDetail('fe4be6fc-76d1008d157c370');"> == $0
            "뜨끈한 국수 한 그릇 먹고 나서는 담양 "
            <!--HS-->
            "겨울"
            <!--HE-->
            <!--HS-->
            "여행"
            <!--HE-->
          </a>
        </li>
      </ul>
    </div>
  </div>
</div>
```

a태그 아래 제목 부분의 텍스트가 있음

Chapter12-1) 특정 항목 분리 후 추출

제목을 가져오기 위한 HTML 코드



어제의 인기 검색어

- 1 겨울여행
- 2 겨울맛
- 3 드라이브

```
<div class="box_search">...</div>
<div class="total_check type1" style="height: 49px;">...</div>
<div class="option_form" style="display: none;">...</div>
<h3 class="blind" id="blindsearchtype">전체 - 관련도순</h3>
<div class="search_result search_body" id="noneBody" style="display: none;">...</div>
<!-- 리스트형 -->
<div class="search_body" id="listBody" style="display: block;">
  <ul class="list_thumType type1">
    <li>...</li>
    <div class="photo">
      <a href="javascript:goSearchDetail('/fe4be6fc-76d1-4158-8f76-5008d157c370')";>
```

<li class="listBody" > -> <div class="area_txt"> -> <div class="tit"> -> <a> 안에 제목 내용이 있음



- 5 공원
- 6 둘레길
- 7 함안
- 8 온천노천탕
- 9 속초
- 10 한국민속촌

```
<div class="tit">
  <a href="javascript:goSearchDetail('/fe4be6fc-76d1-4158-8f76-5008d157c370')";>
    "뜨끈한 국수 한 그릇 먹고 나서는 담양 "
  <!--HS-->
  "겨울"
  <!--HE-->
  <!--HS-->
  "여행"
  <!--HE-->
</a>
</div>
<div class="service">
  <p>전라남도 담양군</p>
```

Chapter12-1) 특정 항목 분리 후 추출

전체 | 여행정보 | 여행기사 | 축제

어제의 인기 검색어

제목 아래는 p태그 밑에 내용이 있음



```
contents = content_list.find('div', 'tit').get_text()
print('내용:', contents.strip())

tag = content_list.find('p', 'tag').get_text()
print('태그:', tag.strip())
print("\n")
```

내용: 뜨끈한 국수 한 그릇 먹고 나서는 담양 겨울 여행

태그: #겨울제철음식#담양맛집#담양국수거리#죽녹원#대나무박람회홍보관#담양장#슬로시티#창평슬로시티#삼지내마을#추천가볼만한곳#담양가볼만한곳#담양당일코스#담양1박2일코스#담양맛집#죽향문화체험마을#서당체험#미식여행#국수맛집#겨울의맛#2015년1월추천가볼만한곳#1월추천가볼만한곳#맛따라가기



#한국관광품질인증 #여행자의방 #숲속의요정 #정강원관광농원
#소낭구펜션 #옛마실펜션 #꿈꾸는노마드 #더세리리조트 #...

9 속초

10 한국민속촌

```
<span>#국수맛집</span>
<span>#겨울의맛</span>
<span>#2015년1월추천가볼만한곳</span>
<span>#1월추천가볼만한곳</span>
<span>#맛따라가기</span>
</p>
```

Chapter12-2) 특정 항목 추출 후 변수 할당

1 # 8. 각 항목별로 분리하여 추출하고 변수에 할당하기

2 no = 1

3 no2 = []

4 contents2 = []

5 tags2 = []

6
7 for i in content_list:

8 no2.append(no)

9 print('번호:', no)

10
11
12 contents = i.find('div', 'tit').get_text()

13 contents2.append(contents)

14 print('내용:', contents.strip())

15
16 tag = i.find('p', 'tag_type').get_text()

17 tags2.append(tag)

18 print('태그:', tag.strip())

19 print("#n")

20
21 no += 1

22

23

24

번호: 1

내용: 뜨끈한 국수 한 그릇 먹고 나서는 담양 겨울 여행

태그: #겨울제철음식#담양맛집#담양국수거리#죽녹원#대나무박물관#홍보관#담양장#슬로시티#창평슬로시티#삼지내마을#추천가볼만한곳#담양가볼만한곳#담양당일코스#담양1박2일코스#담양맛집#죽향문화체험마을#서당체험#미식여행#국수맛집#겨울의맛#2015년1월추천가볼만한곳#1월추천가볼만한곳#맛따라가기

번호: 2

내용: 겨울 여행의 추억 만들기

태그: #한국관광품질인증#여행자의방#숲속의요정#정강원관광농원#소낭구펜션#옛마실펜션#꿈꾸는노마드#더세리리조트#메모리즈호텔#애플리트#목포마리나베이호텔#겨울여행#한국관광품질인증테마여행#우수숙소#인증숙소#평창숙소#거제숙소#제주숙소#속초숙소#목포숙소#포항숙소#양양숙소

번호: 3

내용: 기차 타고 떠나는 겨울여행, 군위 화본역과 화본마을

태그: #겨울기차여행#공공누리

번호: 4

내용: "월 좋아할지 몰라서 다 준비했어!" 겨울방학 평창-강릉 가족 여행법

태그: #평창가볼만한곳#강원도여행#가족여행#체험학습#이색체험#아이와함께#겨울방학가볼만한곳#강릉가볼만한곳#온가족_겨울여행

번호: 5

AttributeError

Traceback (most recent call last)

<ipython-input-11-bc15705d3e7f> in <module>

10

11

----> 12 contents = i.find('div', 'tit').get_text()

13 contents2.append(contents)

14 print('내용:', contents.strip())


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

이유는 ?


Chapter12-2) 특정 항목 추출 후 변수 할당

☰ ☱ ☲ ☳ ☴ ☵ ☶ ☷


관련도순 | 최신순 | 인기순 | ⚙ 옵션




뜨끈한 국수 한 그릇 먹고 나서는 담양 겨울 ... :
전라남도 담양군
#겨울제철음식 #담양맛집 #담양국수거리 #죽녹원 #
대나무박람회홍보관 #담양장 #슬로시티 #창평슬로시티 #...




겨울 여행의 추억 만들기 :
#한국관광품질인증 #여행자의방 #숲속의요정 #정강원관광농원
#소낭구펜션 #옛마실펜션 #꿈꾸는노마드 #더세리조트 #...




기차 타고 떠나는 겨울여행, 군위 화본역과 화... :
경상북도 군위군
#겨울기차여행 #공공누리




“뭘 좋아할지 몰라서 다 준비했어!” 겨울방학... :
강원도 평창군
#평창가볼만한곳 #강원도여행 #가족여행 #체험학습 #이색체험
#아이와함께 #겨울방학가볼만한곳 #강릉가볼만한곳 #온가족...



초보지만 괜찮아!
산림아를 위한 등산 가이드



색다른 즐거움이 있는
취향저격 캠핑장



겨울 특집 용인여행, 전통과 현대가 공존하는 ... :
경기도 용인시

1. 인기 검색어
2. 겨울맛
3. 드라이브
4. 겨울여행

2021 겨울맛

드라이브

겨울여행

전초전탕

국민속촌

시는 정보가 없으신가요?

신규요청바로그기

```
<span>#국수맛집</span>
<span>#겨울의맛</span>
<span>#2015년1월추천가볼만한곳</span>
<span>#1월추천가볼만한곳</span>
<span>#맛따라가기</span>
</p>
</div>
<button type="button" title="열기" class="btn_view">더보기</button>
</div>
<div class="pop_subMenu">...</div>
::after
</li>
</li>
<div class="photo">...</div>
<div class="area_txt"> == $0
</div>
<div class="tit">
<a href="javascript:goSearchDetail('aa29a379-b733-4f37-a993-e
f08b6e7745');">...</a>
</div>
<div class="service">...</div>
<p class="tag_type">
<span>#한국관광품질인증</span>
<span>#여행자의방</span>
<span>#숲속의요정</span>
<span>#정강원관광농원</span>
<span>#소낭구펜션</span>
<span>#옛마실펜션</span>
<span>#꿈꾸는노마드</span>
<span>#더세리조트</span>
<span>#메모리조호텔</span>
<span>#애플리트</span>
<span>#목포마리나베이호텔</span>
</p>
</div>
... rch div.box_leftType1 div#listBody.search_body ul.list_thumType.type1 li div.area_txt ...
Find by string, selector, or XPath
Filter
element.style {
}
.list_thumType.type1 > li .area_txt {
position: relative;
margin-left: 200px;
padding: 0;
height: 120px;
margin-top: 5px;
}
.list_thumType > li .area_txt {
min-height: 94px;
padding-left: 160px;
padding-right: 20px;
}
div {
box-sizing: border-box;
}
```


Chapter12-3) 데이터 데이터 프레임으로 만들기

수집한

1. 가상의 데이터 프레임을 생성한다.
2. 데이터 프레임에 각각의 컬럼을 지정한다.
3. 2번에서 만든 데이터 프레임을 xls 형식이나 csv 형식으로 저장한다.

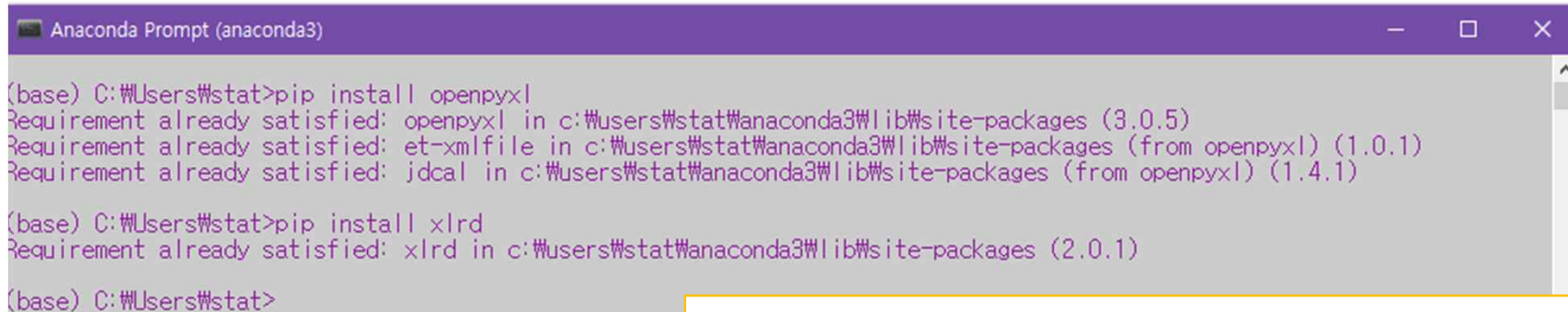
일로 저장

Chapter12-4) 데이터프레임 파일 저장

분리 수집된 데이터를 데이터프레임 형태로 만들어서 csv, xls파일 형식으로 저장하기

```
In [17]: 1 # 9. 출력 결과는 데이터 프레임 형태로 만들기
2 import pandas as pd
3
4 korea_trip = pd.DataFrame()
5 korea_trip['번호'] = no2
6 korea_trip['내용'] = contents2
7 korea_trip['태그'] = tags2
8
9 # csv 형태로 저장하기
10 korea_trip.to_csv(fc_name, encoding="utf-8-sig")
11 print(" csv 파일 저장 경로: %s" %fc_name)
12
13 # 엑셀 형태로 저장하기
14 import xlwt # pip install xlwt 실행 후 수행
15 korea_trip.to_excel(fx_name)
16 print(" xls 파일 저장 경로: %s" %fx_name)
17
18 # 출력 결과를 txt 파일로 저장하기
19 f = open(f_name, 'a', encoding='UTF-8')
20 f.write(str(contents2))
21 f.write(str(tags2))
22 f.close()
23 print(" txt 파일 저장 경로: %s" %f_name)
```

Chapter12-5) 패키지 를 이용한 엑셀 형식 파일 관리



```
Anaconda Prompt (anaconda3)

(base) C:\Users\stat>pip install openpyxl
Requirement already satisfied: openpyxl in c:\users\stat\anaconda3\lib\site-packages (3.0.5)
Requirement already satisfied: et-xmlfile in c:\users\stat\anaconda3\lib\site-packages (from openpyxl) (1.0.1)
Requirement already satisfied: jdcal in c:\users\stat\anaconda3\lib\site-packages (from openpyxl) (1.4.1)

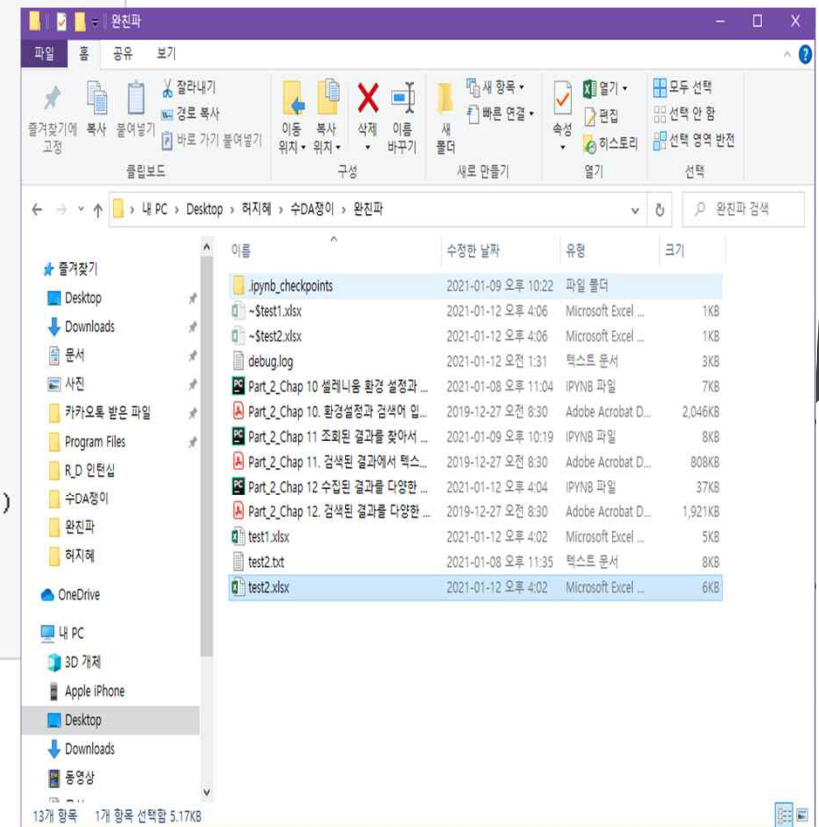
(base) C:\Users\stat>pip install xlrd
Requirement already satisfied: xlrd in c:\users\stat\anaconda3\lib\site-packages (2.0.1)

(base) C:\Users\stat>
```

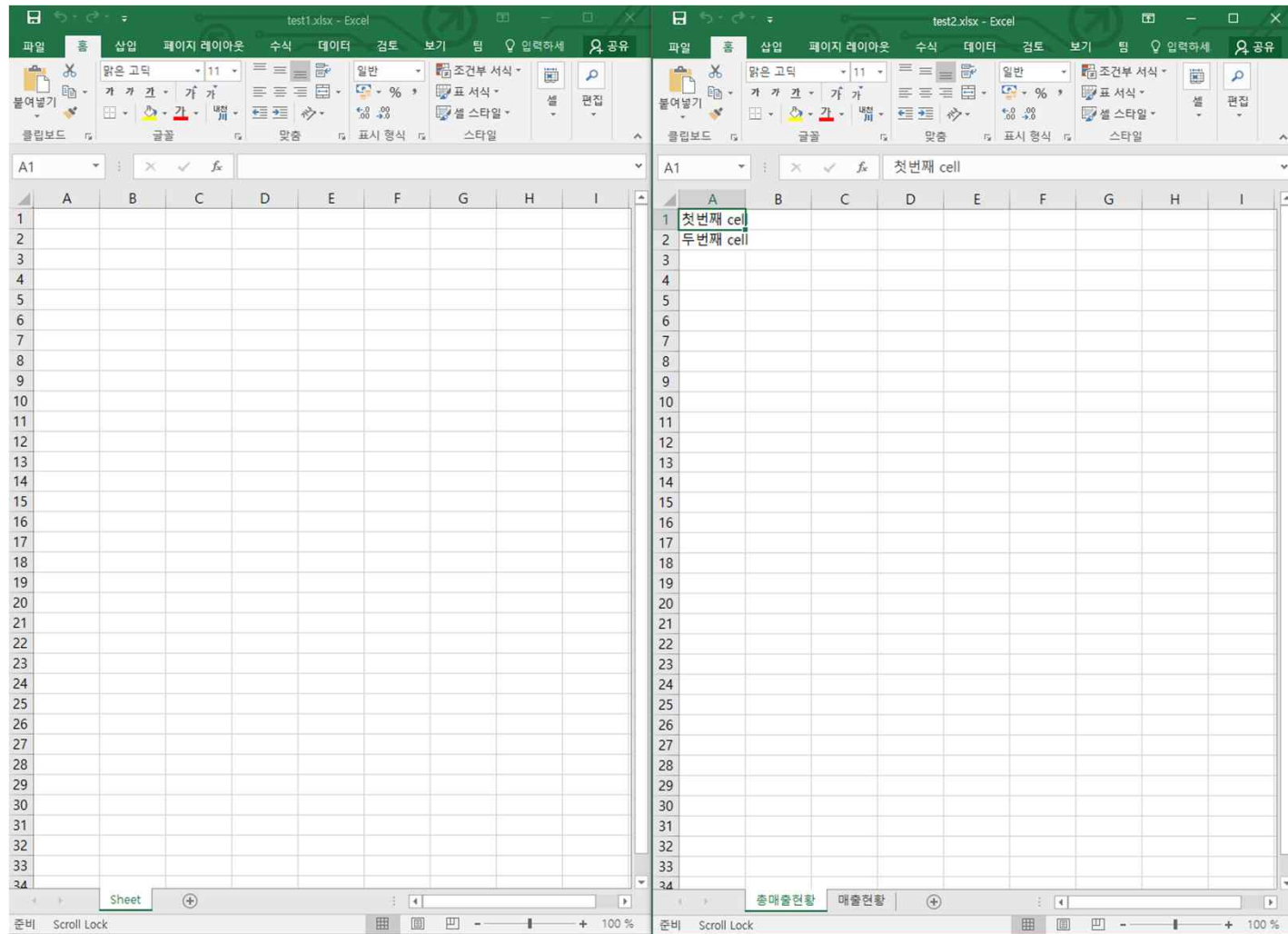
윈도의 cmd 창에서 pip install openpyxl 명령 실행
윈도의 cmd 창에서 pip install xlrd 명령 실행

Chapter12-5) 패키지를 이용한 엑셀 형식 파일 관리

```
In [20]: 1 # 10. openpyxl 패키지를 활용한 엑셀 형식의 파일 관리하기
2 import openpyxl
3
4 # 새로운 엑셀 파일을 1개 생성.
5 wb = openpyxl.Workbook()
6 wb.save('C:/Users/stat/Desktop/허지혜/수DA쟁이/완친파/test1.xlsx')
7
8
9 # 새로운 시트 생성하고 시트이름 변경
10 import openpyxl
11 wb = openpyxl.Workbook()
12
13 sheet_1 = wb.active # 현재 활성화 된 sheet 가져오기
14
15 # 새로운 시트를 만들면서 시트 이름을 지정
16 sheet_2 = wb.create_sheet("매출현황")
17
18 # 시트 이름 변경
19 sheet_1.title = '총매출현황'
20
21 wb.save('C:/Users/stat/Desktop/허지혜/수DA쟁이/완친파/test2.xlsx')
22
23 # Step 3. 기존 파일 불러와서 cell 에 내용 입력
24 import openpyxl
25
26 wb = openpyxl.load_workbook('C:/Users/stat/Desktop/허지혜/수DA쟁이/완친파/test2.xlsx')
27 sheet_1 = wb['총매출현황']
28 sheet_1['A1'] = '첫번째 cell'
29 sheet_1['A2'] = '두번째 cell'
30
31 wb.save('C:/Users/stat/Desktop/허지혜/수DA쟁이/완친파/test2.xlsx')
```



Chapter12-5) 파이썬을 이용한 엑셀 형식 파일 관리



END