



분석용 데이터

획득 전략

웹 리뷰 데이터 확보 기법

학습 목표

+ + +

학습 목표

- 결과를 저장할 폴더와 파일명을 자동으로 생성하도록 설정할 수 있다.
- 리뷰 데이터에서 원하는 항목을 선택하여 추출할 수 있다.
- 다양한 사이트별 리뷰 데이터 수집 방법을 활용할 수 있다.

학습 내용

- 다양한 형태의 리뷰 데이터 수집하기 - 이론
- 다양한 형태의 리뷰 데이터 수집하기 - 실습

인터넷 뉴스 리뷰 정보 수집하기 - 이론

1) 리뷰 수집용 인터넷 뉴스 예제

인터넷뉴스예제

2DNET Korea

PICK 1

"2층 전기버스다"...화성·용인 광역버스

입력 2022.11.22. 오후 1:11 · 수정 2022.11.22. 오후 1:38 · 기사원문

주문경 기자

11 36

| 화성시 7790번 10대·1006번 4대...용인시 5002번 11대

국토교통부 대도시권광역교통위원회는 화성시와 용인시 광역버스를 투입, 22일부터 연말까지 순차적으로 운행을 개시한다고 밝혔다.

협성대 정문을 기점으로 화성 봉담지구에서 의왕TG를 거쳐 사당에 22일부터 5대, 12월 6일 추가로 5대 등 총 10대를 투입한다.



KBS

구독

PICK 1

[취재후] 스모그 1300m 상공, 마침내 맑은 공기가 펼쳐졌다

입력 2019.01.15. 오후 3:53 · 기사원문

이정훈 기자

883 505 505

리뷰 건수



16년 만에 최악의 미세먼지가 닥친 어제(14일) 오전, 기자가 찾은 김포공항은 온통 스모그에 뒤덮여 있었습니다. '헬기 띄우면 그림 잘 나오겠군.' 주말 동안 헬기 취재를 준비한 보람을 느낀 것도 잠시, KBS 항공 1호기 조종사인 유태정 팀장 입에서 뜻밖의 얘기가 들렸습니다. "지금 못 떠요. 시정이 1/4마일밖에 안 나와. 오후에나 뜰 수 있으려나 모르겠는데요." 헬기에 함께 탑승해 관측하기로 한 강원대 대기질예측연구실 학생들의 표정에도 걱정이 묻어났습니다.

3

본 강의자료를 무단배포 및 복제하는 것을 금지하며
이를 위반하는 경우 저작권 침해로서 관련법에 따라 처벌될 수 있습니다.

인터넷 뉴스 리뷰 정보 수집하기 - 이론

2) 총 리뷰 건수 정보 수집

(1) 작업 순서

▪ 구성 및 변경 관리

- ① 사용자에게 **수집할 리뷰 건수** 물어보기
- ② 결과를 **저장할 폴더명** 물어보기

 위데이터를 입력받은 후 검색을 시작하여

- ➡ 전체 검색 결과 건수
- ➡ 실제 최종 출력 건수
- ➡ 총 검색 페이지 수 정보를 보여주고 수집 시작

주의사항

사용자가 요청한 건수와 실제 리뷰 건수와 비교하여 작업 조율이 필요

인터넷 뉴스 리뷰 정보 수집하기 - 이론

3) 수집할 상세 항목 정보

수집할 항목

01



리뷰작성자ID

02



리뷰작성일자

03



리뷰내용

04



공감/비공감수

4) 추가 리뷰 정보 보기

▪ 네이버 뉴스 리뷰



맨처음 페이지는 5개 리뷰만



'[댓글 더보기](#)>' 클릭하면 추가 리뷰 정보



1페이지에 20개의 리뷰



이후 리뷰를 더 보려면 아래 '[더보기](#)' 버튼 클릭

인터넷 뉴스 리뷰 정보 수집하기 - 이론

5) 리뷰 정보 저장하기

 아래와 같이 폴더와 파일을 자동으로 생성함



자동생성 **폴더** 이름 형식

[년도4자리-월-일-시간-분-초-뉴스기사댓글]



자동생성 **파일** 이름 형식

[년도4자리-월-일-시간-분-초-뉴스기사댓글.csv
년도4자리-월-일-시간-분-초-뉴스기사댓글.txt
년도4자리-월-일-시간-분-초-뉴스기사댓글.xls]

주의사항

파일 자동생성시 **파일 이름 중복으로 덮어쓰는 문제가 발생**

인터넷 뉴스 리뷰 정보 수집하기 - 실습

■ 검색어 지정

```
1 #Step 1. 필요한 모듈과 라이브러리를 로딩합니다.
2 from bs4 import BeautifulSoup
3 from selenium import webdriver
4 from selenium.webdriver.common.by import By
5 from selenium.webdriver.common.keys import Keys
6 from selenium.webdriver.chrome.service import Service
7 import time
8 import math
9 import numpy
10 import pandas as pd
11 import os
12
13
14 #Step 2. 사용자에게 검색어 키워드를 입력 받고 저장할 폴더와 파일명을 설정합니다.
15 print("=" * 80)
16 print("뉴스 기사의 댓글 정보 수집하기")
17 print("=" * 80)
18
19 query_txt = '뉴스기사댓글'
20 query_url = 'https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=102&oid=056&aid=0010661268'
21 try :
22     cnt = int(input('1.크롤링 할 뉴스 리뷰 건수는 몇건입니까?(기본값: 20건): '))
23 except ValueError :
24     cnt = 20
25
26 page_cnt = math.ceil(cnt / 20)
27
28 f_dir = input("2.결과를 파일을 저장할 폴더명만 쓰세요(기본값:c:\\py_temp\\):")
29 if f_dir=='':
30     f_dir='c:\\py_temp\\'
```

주의사항

리뷰 건수를입력안하고 실행할 경우 **ValueError**발생

인터넷 뉴스 리뷰 정보 수집하기 - 실습

■ 저장할 폴더 지정

```
23 except ValueError :
24     cnt = 20
25
26 page_cnt = math.ceil(cnt / 20)
27
28 f_dir = input("2. 결과를 파일을 저장할 폴더명만 쓰세요(기본값:c:\\py_temp\\):")
29 if f_dir=='':
30     f_dir='c:\\py_temp\\'
31
32 # 저장될 파일위치와 이름을 지정합니다
33 now = time.localtime()
34 s = '%04d-%02d-%02d-%02d-%02d-%02d' % (now.tm_year, now.tm_mon, now.tm_mday, \
35                                         now.tm_hour, now.tm_min, now.tm_sec)
36
37 os.makedirs(f_dir+s+'-'+query_txt)
38 os.chdir(f_dir+s+'-'+query_txt)
39
40 ff_name=f_dir+s+'-'+query_txt+'\\'+s+'-'+query_txt+'.txt'
41 fc_name=f_dir+s+'-'+query_txt+'\\'+s+'-'+query_txt+'.csv'
42 fx_name=f_dir+s+'-'+query_txt+'\\'+s+'-'+query_txt+'.xls'
43
44 #Step 3. 크롬 드라이버를 사용해서 웹 브라우저를 실행합니다.
45 s_time = time.time()
46
47 s = Service("c:/py_temp/chromedriver.exe")
48 driver = webdriver.Chrome(service=s)
49
50 driver.get(query_url)
51 driver.maximize_window()
52 time.sleep(5)
53
```

주의사항

폴더/파일 이름에 현재 날짜, 시간 등을 넣으면 덮어쓰지 않고 저장 가능

정보

time.localtime() 함수: 현재 시간을 년, 월, 일, 시 등의 형태로 보여주는 함수

주의사항

time.time() 함수: 현재 시간을 실수 형태로 반환하여 보여주는 함수

인터넷 뉴스 리뷰 정보 수집하기 - 실습

▪ 콤마(,) 기호 삭제

```
48 driver = webdriver.Chrome(service=s)
49
50 driver.get(query_url)
51 driver.maximize_window()
52 time.sleep(5)
53
54 # 현재 총 리뷰 건수를 확인하여 사용자의 요청건수와 비교 후 동기화합니다
55 html = driver.page_source
56 soup = BeautifulSoup(html, 'html.parser')
57
58 result= soup.find('div','media_end_head_info_variety_cmtcount _COMMENT_HIDE').get_text()
59
60 print("=" *80)
61 search_cnt = int( result.replace(",",""))
62
63 if cnt > search_cnt :
64     cnt = search_cnt
65
66 print("전체 검색 결과 건수 :",search_cnt,"건")
67 print("실제 최종 출력 건수 :",cnt)
68 print("총 페이지 수:" , page_cnt)
```

```
1 # 사용자가 요청한 건수가 많을 경우 리뷰 더보기 버튼을 클릭합니다
2 # 최초 10건 수집후 댓글 더보기 버튼 클릭
3 # 아래 버튼을 눌러 첫 화면에 총 20건의 댓글이 나오게 만들
4 driver.find_element(By.XPATH,'//*[@id="cbox_module"]/div/div[9]/a/span[1]').click()
5 time.sleep(3)
6
7 #Step 6. 20건 출력되어 있는 현재 페이지 리뷰와 점수 등 내용 수집
8 no2= [] # 리뷰 번호
9 writer_id2=[] # 리뷰 작성자 ID
10 review2=[] # 리뷰 내용
11 write_date2=[] # 리뷰 작성 일자
12 coord_nm=[] # 곳값 회수
```

정보

cnt:사용자가입력한값, **search_cnt**:실제 검색된건수

인터넷 뉴스 리뷰 정보 수집하기 - 실습

mment/056/0010661268?sid=102

입니다.

The screenshot shows a web interface for a news site. On the left, there's a list of comments with user avatars, names, timestamps, and content. On the right, there's a news article titled 'KBS 랭킹 뉴스' (KBS Ranking News) with a list of 5 items. The first item is about a fire at a university dormitory. Below the news article, there's a section for '같이 볼만한 뉴스' (News to watch together) with two video thumbnails. At the bottom, there's a progress bar and a timestamp '12:37 / 17:26'.

정보

작성자에 의해 삭제된 댓글은 공감/비공감 표시가 없음

인터넷 뉴스 리뷰 정보 수집하기 - 실습

```
3 # 아래 버튼을 눌러 첫 화면에 총 20건의 댓글이 나오게 만들
4 driver.find_element(By.XPATH, '//*[@id="cbox_module"]/div/div[9]/a/span[1]').click()
5 time.sleep(3)
6
7 # Step 6. 20건 출력되어 있는 현재 페이지 리뷰와 점수 등 내용 수집
8 no2= [] # 리뷰 번호
9 writer_id2=[] # 리뷰 작성자 ID
10 review2=[] # 리뷰 내용
11 write_date2=[] # 리뷰 작성 일자
12 gogam_0=[] # 공감 횟수
13 gogam_1=[] # 비공감 횟수
14 count = 0
15
16 for a in range(1, page_cnt+1) :
17     if a == page_cnt :
18         break
19     else :
20         driver.find_element(By.XPATH, '//*[@id="cbox_module"]/div/div[9]/a').click()
21         time.sleep(3)
22         print("%s페이지 이동 완료" % a)
23         time.sleep(random.randrange(1,3)) # 3-8 초 사이에 랜덤으로 시간 선택
24
25 print('이제 리뷰 정보를 수집합니다. 잠시만 기다려 주세요~~~~~')
26
27 # txt 파일에 저장하기 위해 파일 open하기
28 f = open(ff_name, 'a', encoding='UTF-8')
29
30 html = driver.page_source
31 soup = BeautifulSoup(html, 'html.parser')
32
33 slist = soup.find('ul', 'u_cbox_list').find_all('li')
34
35 for li in slist:
```