Crawling with Python

2019.01.16

학부생 인턴 이지헌

멜론(Melon) 실시간 인기차트 TOP 100

- **Crawling with python Code**

```
import requests
from bs4 import BeautifulSoup
import pandas as pd


hdr = {'User-Agent': 'Mozilla/5.0'}
url = 'https://www.melon.com/chart/index.htm'
req = requests.get(url, headers=hdr)
soup = BeautifulSoup(req.content, 'html.parser')
lst_data = soup.select('.lst50, .lst100')

melon_lst = []
for i in lst_data:
    temp = []
    temp.append(i.select_one('.rank').text)
    temp.append(i.select_one('.rank01').a.text)
    temp.append(i.select_one('.rank02').a.text)
    temp.append(i.select_one('.rank03').a.text)
    melon_lst.append(temp)

melon_df = pd.DataFrame(melon_lst,
                        columns=['순위', '노래명', '아티스트', '앨범명'])
melon_df.to_csv('melon_100.csv', mode='w', encoding='utf-8-sig',
                header=True, index=False)
```

- **Melon DataFrame**

```
1 melon_df
```

	순위	노래명	아티스트	앨범명
0	1	아무노래	지코 (ZICO)	아무노래
1	2	METEOR	창모 (CHANGMO)	Boyhood
2	3	내게 들려주고 싶은 말 (Dear Me)	태연 (TAEYEON)	Purpose - The 2nd Album Repackage
3	4	Psycho	Red Velvet (레드벨벳)	'The ReVe Festival' Finale
4	5	다시 난, 여기	백예린	사랑의 불시착 OST Part 4
...	...	...	...	...
95	96	하루에도 열두 번	윤건	하루에도 열두 번
96	97	벗	닐로(Nilo)	벗
97	98	오늘도 (Day After Day)	EXO	OBSESSION - The 6th Album
98	99	샤넬 (Feat. 박봄)	MC몽	CHANNEL 8
99	100	Baby You Are	EXO	OBSESSION - The 6th Album

100 rows × 4 columns

- **Melon CSV**

	A	B	C	D	E
1	순위	노래명	아티스트	앨범명	
2	1	아무노래	지코 (ZICO)	아무노래	
3	2	METEOR	창모 (CHAN	Boyhood	
4	3	내게 들려주고 싶은 말	태연 (TAEYE	Purpose - The 2nd Album	
5	4	Psycho	Red Velvet (	'The ReVe Festival' Finale	
6	5	다시 난, 여기	백예린	사랑의 불시착 OST Part 4	
7	6	Blueming	아이유	Love poem	
8	7	아마두 (feat.우원재, 김	염따	Dingo X DAMOIM (Part 2	
9	8	늦은 밤 너의 집 앞 골목	노을	늦은 밤 너의 집 앞 골목길	
10	9	HIP	마마무(Mam	reality in BLACK	
11	10	Square (2017)	백예린	Every letter I sent you.	
12	11	너를 사랑하고 있어	백현 (BAEKH	낭만닥터 김사부 2 OST Pa	
13	12	흔들리는 꽃들 속에서	장범준	멜로가 체질 OST Part 3	
14	13	다시는 사랑하지 않고,	백지영	다시는 사랑하지 않고, 이별	
15	14	어떻게 이별까지 사랑하	AKMU (악동	항해	

네이버(Naver) 연극 월/일/주간/주말 별 연극

- **Crawling with python Code**

```
from urllib.parse import quote_plus    # 한글 텍스트를 퍼센트 인코딩으로 변환
from selenium import webdriver    # 라이브러리에서 사용하는 모듈만 호출
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait   # 해당 태그를 기다림
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException    # 태그가 없는 예외 처리
import time
import pandas as pd


user_input = quote_plus(input('''-월--일, -월, 이번주, 이번주말 중 선택하여 입력해주세요.
                                 (-은 숫자 입력, 이번년도만 가능) : '''))

url = f'https://search.naver.com/search.naver?where=nexearch&sm=tab_etc&query={user_inpu
chromedriver = 'C:/Users/LeeJiheon/Desktop/가천대학교/TEAMLAB/2019_winter_study/2주차/crawl

options = webdriver.ChromeOptions()
options.add_argument('headless')    # 웹 브라우저를 띄우지 않는 headlss chrome 옵션 적용
options.add_argument('disable-gpu')    # GPU 사용 안함
options.add_argument('lang=ko_KR')    # 언어 설정
driver = webdriver.Chrome(chromedriver, options=options)
```

```
driver.get(url)

try:    # 정상 처리
    element = WebDriverWait(driver, 3).until(
        EC.presence_of_element_located((By.CLASS_NAME, 'list_title'))
    )    # 해당 태그 존재 여부를 확인하기까지 3초 기다림
    theater_list = []
    pageNum = int(driver.find_element_by_class_name('_totalCount').text)

    for i in range(1, pageNum):
        theater_data = driver.find_elements_by_class_name('list_title')

        for k in theater_data:
            theater_list.append(k.text.split('\n'))

        driver.find_element_by_xpath("//a[@class='btn_page_next _btnNext on']").click()
        time.sleep(2)

    driver.quit()

except TimeoutException:    # 예외 처리
    print('해당 페이지에 연극 정보가 존재하지 않습니다.')

finally:    # 정상, 예외 둘 중 하나여도 반드시 실행
    driver.quit()

theater_df = pd.DataFrame(theater_list,
                          columns=['연극명', '기간', '장소'])
theater_df.index = theater_df.index + 1    # 인덱스 초기값 1로 변경
theater_df.to_csv('theater_df.csv', mode='w', encoding='utf-8-sig',
                  header=True, index=True)

print('웹 크롤링이 완료되었습니다.')
```

- **Theater DataFrame**

```
1 theater_df
```

	연극명	기간	장소
1	극적인 하룻밤 - 서울	19.03.01.~오픈런	대학로 드림아트센터 4관
2	옥탑방 고양이 - 서울	10.04.06.~오픈런	틴틴홀
3	행오버	16.03.08.~오픈런	휴먼시어터
4	쉬어매드니스	15.11.12.~오픈런	콘텐츠박스
5	수상한 흥신소 - 서울	19.11.01.~오픈런	수상한흥신소전용관
...	...	...	...
156	흑백다방	20.01.07.~20.04.12.	대학로스타시티 7층 후암 씨어터
157	노동풍경1: 실업	20.01.17.~20.02.02.	연우소극장
158	기적의 소년	20.01.15.~20.01.22.	SH아트홀
159	디벅	20.01.15.~20.12.31.	대학로 마당세실극장
160	룸넘버13	16.01.12.~오픈런	콘텐츠룸

160 rows × 3 columns

- **Theater CSV**

	A	B	C	D
1		연극명	기간	장소
2	1	극적인 하룻밤 - 서울	19.03.01.~오픈런	대학로 드림아트센터 4관
3	2	옥탑방 고양이 - 서울	10.04.06.~오픈런	틴틴홀
4	3	행오버	16.03.08.~오픈런	휴먼시어터
5	4	쉬어매드니스	15.11.12.~오픈런	콘텐츠박스
6	5	수상한 흥신소 - 서울	19.11.01.~오픈런	수상한흥신소전용관
7	6	작업의 정석 - 서울 대학로	12.06.29.~오픈런	대학로연극순위아트홀1관
8	7	오백에 삼십	18.09.04.~오픈런	대학로 아트포레스트 1관
9	8	한뼘사이	17.03.01.~오픈런	서연아트홀
10	9	라이어 1탄 - 서울	98.01.02.~오픈런	대학로 민송아트홀 1관
11	10	엘리펀트 송	19.11.22.~20.02.02.	예스24스테이지 3관
12	11	2호선 세입자	19.03.15.~오픈런	대학로 바탕골 소극장
13	12	꽃의 비밀	19.12.21.~20.03.01.	서경대학교 공연예술센터 스콘 2관
14	13	그남자 그여자 - 대구	19.11.01.~20.02.02.	여우별아트홀
15	14	그대를 사랑합니다 - 서울	19.11.22.~20.02.02.	서경대학교 공연예술센터 스콘1관