Lab#3, NLP@CGU Spring 2023

This is due on 2023/03/20 16:00, commit to your github as a PDF (lab3.pdf) (File>Print>Save as PDF).

IMPORTANT: After copying this notebook to your Google Drive, please paste a link to it below. To get a publicly-accessible link, hit the *Share* button at the top right, then click "Get shareable link" and copy over the result. If you fail to do this, you will receive no credit for this lab!

LINK: paste your link here

https://colab.research.google.com/drive/1JtNUYu3kF3eqEWIlpEzk_EKy3n1xxEFl?usp=sharing

---

Student ID:B0928012

Name:王晟翰

Question 1 (100 points)

Implementing Yahoo Movies Crawler.

1. Design a Yahoo! Movie Crawler.
2. Crawl all the movie information listed in movie_intheaters page
3. The more movie data crawled, the higher the score

按兩下 (或按 Enter 鍵) 即可編輯

```
import requests
import re
from bs4 import BeautifulSoup

Y_MOVIE_URL = "https://movies.yahoo.com.tw/movie_intheaters.html"

# YOUR CODE HERE!
# IMPLEMENTIG YAHOO MOVIES CRAWLER

class MovieCrawler(object):

    def __init__(self):
        self.headers = {
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.11
        self.movies = []
    def get_movies(self, page_url):
        res = requests.get(page_url, headers=self.headers)
        soup = BeautifulSoup(res.text, 'html.parser')
        movie_list = soup.find_all('div', class_='release_info')
        for movie in movie_list:
            # Get the Chinese name of the movie.
            ch_name = movie.find('div', class_='release_movie_name').a.text.strip()
            # Get the English name of the movie.
            en_name = movie.find('div', class_='release_movie_name').find('div', class_='en').a.text.strip()
            # Get the URL of the movie detail page.
            movie_url = movie.find('div', class_='release_movie_name').a['href']
            # Get the URL of the movie detail page.
            release_date = movie.find('div', class_='release_movie_time').text.strip()
            release_date = release_date.split(' ')[-1]
            ## Get the introduction of the movie.
            intro = movie.find('div', class_='release_text').text.strip().replace('\n', '').replace('\r', '')

            # Store the movie information in a dictionary and append it to the movies list.
            movie_dict = {
                'ch_name': ch_name,
                'en_name': en_name,
                'movie_url': movie_url,
                'release_date': release_date,
                'intro': intro
            }
            self.movies.append(movie_dict)
        # Find the link to the next page
        next_page = soup.find('a', rel='next')
        if next_page:
            # Construct the link to the next page and recursively call the get_movies() function.
            next_page_url = next_page['href']
```

```
                self.get_movies(next_page_url)

        return  self.movies


# # DO NOT MODIFY THE VARIABLES
crawler = MovieCrawler()
movies = crawler.get_movies(Y_MOVIE_URL)

# # THE RESULTS : AS THE FOLLOWING SECTION
# # {'ch_name', 'en_name', 'movie_url', 'release_date', 'intro'}
print(len(movies))
print(*movies, sep="\n")
```

```
77
{'ch_name': '配樂大師顏尼歐', 'en_name': 'Ennio: The Maestro', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%85%8D%E6%A8%82%E5
{'ch_name': '熊蓋毒', 'en_name': 'Cocaine Bear', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%86%8A%E8%93%8B%E6%AF%92-cocaine-
{'ch_name': '若愛重來', 'en_name': 'Marriages', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%8B%A5%E6%84%9B%E9%87%8D%E4%BE%86-
{'ch_name': '無人相信的真相', 'en_name': 'La syndicaliste', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%84%A1%E4%BA%BA%E7%9B
{'ch_name': '闇黑對決', 'en_name': "The Devil's Deal", 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%97%87%E9%BB%91%E5%B0%8D%E
{'ch_name': '噩夢輓歌 4K數位修復版', 'en_name': 'Requiem For A Dream', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%99%A9%E5%
{'ch_name': '人體動物圖鑑：烏龜的殼其實是肋骨', 'en_name': 'Turtle' s Shell is a Human' s Ribs', 'movie_url': 'https://movies.yahoo.com.tw/mov
{'ch_name': '流水落花', 'en_name': 'Lost Love', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B5%81%E6%B0%B4%E8%90%BD%E8%8A%B1-
{'ch_name': '聖蛛', 'en_name': 'Holy Spider', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%81%96%E8%9B%9B-holy-spider-14886',
{'ch_name': '沙贊！眾神之怒', 'en_name': 'Shazam! Fury of the Gods', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B2%99%E8%B4
{'ch_name': '夢遊樂園', 'en_name': 'Melody-Go-Round', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%A4%A2%E9%81%8A%E6%A8%82%E5
{'ch_name': '黑的教育', 'en_name': 'Bad Education', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%BB%91%E7%9A%84%E6%95%99%E8%8
{'ch_name': 'TÁR塔爾', 'en_name': 'Tár', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/T%C3%81R%E5%A1%94%E7%88%BE-tar-14393', 'rel
{'ch_name': '驚聲尖叫6', 'en_name': 'Scream VI', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A9%9A%E8%81%B2%E5%B0%96%E5%8F%A
{'ch_name': '怪談比留子 數位修復版', 'en_name': 'Hiruko The Goblin', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%80%AA%E8%AB
{'ch_name': '天生一對2大電影：再續前緣', 'en_name': 'Love Destiny: The Movie', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%A
{'ch_name': '尋找第5味', 'en_name': 'Umami', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B0%8B%E6%89%BE%E7%AC%AC5%E5%91%B3-u
{'ch_name': '超完美狗保姆', 'en_name': 'My Puppy', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B6%85%E5%AE%8C%E7%BE%8E%E7%8B
{'ch_name': '蓋世棋蹟', 'en_name': 'The Royal Game', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%93%8B%E4%B8%96%E6%A3%8B%E8%
{'ch_name': '斷網', 'en_name': 'Cyberheist', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%96%B7%E7%B6%B2-cyberheist-14809', '
{'ch_name': '所有的美麗與血淚', 'en_name': 'All the Beauty and the Bloodshed', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%8
{'ch_name': '過時·過節', 'en_name': 'Hong Kong Family', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%81%8E%E6%99%82-%E9%81%8E%
{'ch_name': '8釐米：詛咒影帶', 'en_name': '8MM: The Sinister Record', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/8%E9%87%90%E7%
{'ch_name': '屍蹤天使', 'en_name': 'Mindcage', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B1%8D%E8%B9%A4%E5%A4%A9%E4%BD%BF-
{'ch_name': '貓王艾維斯', 'en_name': 'Elvis', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B2%93%E7%8E%8B%E8%89%BE%E7%B6%AD%E
{'ch_name': '媽的多重宇宙', 'en_name': 'Everything Everywhere All at Once', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%AA%B
{'ch_name': '光影帝國', 'en_name': 'Empire Of Light', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%85%89%E5%BD%B1%E5%B8%9D%E5
{'ch_name': '金牌拳手3', 'en_name': 'Creed III', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%87%91%E7%89%8C%E6%8B%B3%E6%89%8
{'ch_name': '本日公休', 'en_name': 'Day Off', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%9C%AC%E6%97%A5%E5%85%AC%E4%BC%91-d
{'ch_name': '玩具當家', 'en_name': 'The New Toy', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%8E%A9%E5%85%B7%E7%95%B6%E5%AE%
{'ch_name': '驚爆點', 'en_name': 'Point Break', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A9%9A%E7%88%86%E9%BB%9E-point-br
{'ch_name': '火線埋伏', 'en_name': 'Ambush', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%81%AB%E7%B7%9A%E5%9F%8B%E4%BC%8F-am
{'ch_name': '小熊維尼：血與蜜', 'en_name': 'Winnie the Pooh: Blood and Honey', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B
{'ch_name': '鈴芽之旅', 'en_name': 'Suzume', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%88%B4%E8%8A%BD%E4%B9%8B%E6%97%85-su
{'ch_name': '法貝爾曼', 'en_name': 'The Fabelmans', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B3%95%E8%B2%9D%E7%88%BE%E6%9
{'ch_name': '人肉搜索2：失蹤搜救', 'en_name': 'Missing', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E4%BA%BA%E8%82%89%E6%90%9C
{'ch_name': '悲情城市', 'en_name': 'A City of Sadness', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%82%B2%E6%83%85%E5%9F%8E%
{'ch_name': '風再起時', 'en_name': 'Where The Wind blows', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A2%A8%E5%86%8D%E8%B5%
{'ch_name': '胡桃鉗與魔笛公主的奇幻冒險', 'en_name': 'The Nutcracker And The Magic Flute', 'movie_url': 'https://movies.yahoo.com.tw/movieinf
{'ch_name': '我們的黎明', 'en_name': 'Break of Dawn', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%88%91%E5%80%91%E7%9A%84%E9
{'ch_name': '不離職冒險王', 'en_name': 'Irreductible', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E4%B8%8D%E9%9B%A2%E8%81%B7%E
{'ch_name': '「鬼滅之刃」上弦集結，前進刀匠村', 'en_name': 'Demon Slayer Kimetsu No Yaiba To The Swordsmith Village', 'movie_url': 'https://m
{'ch_name': '追海豚的長崎夏日', 'en_name': 'Sabakan', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%BF%BD%E6%B5%B7%E8%B1%9A%E7
{'ch_name': '蟻人與黃蜂女：量子狂熱', 'en_name': 'Ant-Man and the Wasp: Quantumania', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_mai
{'ch_name': '超難搞先生', 'en_name': 'A Man Called Otto', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B6%85%E9%9B%A3%E6%90%9
{'ch_name': '關於我和鬼變成家人的那件事', 'en_name': 'Marry My Dead Body', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%97%9C
{'ch_name': '山椒魚來了', 'en_name': '', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B1%B1%E6%A4%92%E9%AD%9A%E4%BE%86%E4%BA%
{'ch_name': '僕愛君愛：致深愛妳的那個我', 'en_name': 'To me, The One Who Loved You', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/
{'ch_name': '僕愛君愛：致我深愛的每個妳', 'en_name': 'To Every You I' ve Loved Before', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_m
{'ch_name': '日麗', 'en_name': 'Aftersun', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%97%A5%E9%BA%97-aftersun-14693', 'rele
{'ch_name': '新世紀福音戰士新劇場版：終', 'en_name': 'Evangelion:3.0+1.0 Thrice Upon A Time', 'movie_url': 'https://movies.yahoo.com.tw/movie
{'ch_name': '瑪琳艾索普：首席女指揮', 'en_name': 'Conductor', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%91%AA%E7%90%B3%E8%
{'ch_name': '我的鯨魚老爸', 'en_name': 'The Whale', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%88%91%E7%9A%84%E9%AF%A8%E9%A
{'ch_name': '幻影', 'en_name': 'Phantom', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B9%BB%E5%BD%B1-phantom-14651', 'release
{'ch_name': '伊尼舍林的女妖', 'en_name': 'The Banshees of Inisherin', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E4%BC%8A%E5%B
{'ch_name': '鱷魚歌王', 'en_name': 'Lyle, Lyle, Crocodile', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%B1%B7%E9%AD%9A%E6%AD
```

✓ 4 秒 完成時間：下午3:59 ● ×