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/1btirJ2gncIeKojDu5-BYQYdl7ir8TTya?usp=sharing

---

Student ID:B092800 **Name**:郭玉俊

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
import csv

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

        def get_movies(self, page_url):
                movies = []
                page = 1
                while True:
                        response = requests.get(url=page_url, params={'page': page})
                        soup = BeautifulSoup(response.text, 'lxml')
                        info_items = soup.find_all('div', 'release_info')
                        if not info_items:    # 如果找不到任何電影，就退出循環。
                                break
                        for item in info_items:
                                movie = {}
                                movie['ch_name'] = item.find('div', 'release_movie_name').a.text.strip()
                                movie['en_name'] = item.find('div', 'en').a.text.strip()
                                movie['movie_url'] = item.find('div', 'release_movie_name').a['href'].strip()
                                movie['release_date'] = item.find('div', 'release_movie_time').text.split(': ')[-1].strip()
                                movie['intro'] = item.find('div', 'release_text').span.text.strip()
                                movies.append(movie)
                        page += 1
                return 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")
```

{'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', 'releas
{'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
{'ch_name': '巴比倫', 'en_name': 'Babylon', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B7%B4%E6%AF%94%E5%80%AB-babylon-1401
{'ch_name': '詐團圓', 'en_name': 'Scamsgiving', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%A9%90%E5%9C%98%E5%9C%93-scamsgiv
{'ch_name': '天龍八部之喬峰傳', 'en_name': 'Sakra', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%A4%A9%E9%BE%8D%E5%85%AB%E9%8
{'ch_name': '屁屁偵探 : 天才惡人屁屁亞蒂', 'en_name': 'Butt Detective the Movie:Shiriarty', 'movie_url': 'https://movies.yahoo.com.tw/moviein
{'ch_name': '愛. 子', 'en_name': 'The Son', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%84%9B-%E5%AD%90-the-son-14054', 'rel
{'ch_name': '灌籃高手 THE FIRST SLAM DUNK', 'en_name': 'The First Slam Dunk', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%81
{'ch_name': '鞋貓劍客2', 'en_name': 'Puss in Boots: The Last Wish', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%9E%8B%E8%B2%
{'ch_name': '與你共舞', 'en_name': 'I Wanna Dance with Somebody', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%88%87%E4%BD%A0
{'ch_name': '想見你', 'en_name': 'Someday or One Day', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%83%B3%E8%A6%8B%E4%BD%A0-s
{'ch_name': '有你就是家', 'en_name': 'Tori and Lokita', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%9C%89%E4%BD%A0%E5%B0%B1%
{'ch_name': '阿凡達：水之道', 'en_name': 'Avatar: The Way of Water', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%98%BF%E5%87
{'ch_name': '決戰38度線', 'en_name': 'Devotion', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B1%BA%E6%88%B038%E5%BA%A6%E7%B7
{'ch_name': '王者製造', 'en_name': 'kingmaker', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%8E%8B%E8%80%85%E8%A3%BD%E9%80%A0
{'ch_name': '梵天神器：滅世開端', 'en_name': 'Brahmastra: Part One Shiva', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%A2%B5
{'ch_name': '骨肉的總和', 'en_name': 'Bones and All', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%AA%A8%E8%82%89%E7%9A%84%E7
{'ch_name': '黑豹2：瓦干達萬歲', 'en_name': 'Black Panther: Wakanda Forever', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%BB
{'ch_name': '角落小夥伴電影版：藍色月夜的魔法之子', 'en_name': 'Sumikkogurashi: The Little Wizard in the Blue Moonlight', 'movie_url': 'https
{'ch_name': 'PINA', 'en_name': 'PINA', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/PINA-pina-13278', 'release_date': '2022-06-03
{'ch_name': '郵差', 'en_name': 'Il Postino', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%83%B5%E5%B7%AE-il-postino-11303', '
{'ch_name': '寄生上流', 'en_name': 'Parasite', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%AF%84%E7%94%9F%E4%B8%8A%E6%B5%81-

Colab 付費產品 - 按這裡取消合約

✓ 4 秒 完成時間：下午3:57