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/1GUdWg0pv8H7Qd9KZAL7HZpeaiRx8cgsP?usp=sharing

Student ID:B0928016

Name:趙君熙

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.movies = []

    def get_movies(self, page_url):
        self.movies = []
        url = page_url
        for page in range(8):
            page_url = url + "?page=" + str(page+1)

            response = requests.get(url = page_url)
            soup = BeautifulSoup(response.text, 'lxml')
            info_items = soup.find_all('div', 'release_info')

            for item in info_items:
                ch_name = item.find('div', 'release_movie_name').a.text.strip()
                en_name = item.find('div', 'en').a.text.strip()
                movie_url = item.find('div', 'release_movie_name').a.get('href')
                release_date = item.find('div', 'release_movie_time').text.split('：')[-1].strip()
                intro = item.find('div', 'release_text').span.text.strip()
                self.movies.append([ch_name, en_name, movie_url, release_date, intro])
        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")
```

```
76
['配樂大師顏尼歐', 'Ennio: The Maestro', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%85%8D%E6%A8%82%E5%A4%A7%E5%B8%AB%E9%A1%8F%E5%B0%BC%E
['熊蓋毒', 'Cocaine Bear', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%86%8A%E8%93%8B%E6%AF%92-cocaine-bear-14462', '2023-03-17', '故事靈
['若愛重來', 'Marriages', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%8B%A5%E6%84%9B%E9%87%8D%E4%BE%86-marriages-14701', '2023-03-17', '
['無人相信的真相', 'La syndicaliste', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%84%A1%E4%BA%BA%E7%9B%B8%E4%BF%A1%E7%9A%84%E7%9C%9F%E7%9
['闇黑對決', "The Devil's Deal", 'https://movies.yahoo.com.tw/movieinfo_main/%E9%97%87%E9%BB%91%E5%B0%8D%E6%B1%BA-the-devils-deal-14846', '20
['噩夢輓歌 4K數位修復版', 'Requiem For A Dream', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%99%A9%E5%A4%A2%E8%BC%93%E6%AD%8C-4K%E6%95%B8
['人體動物圖鑑：烏龜的殼其實是肋骨', 'Turtle's Shell is a Human's Ribs', 'https://movies.yahoo.com.tw/movieinfo_main/%E4%BA%BA%E9%AB%94%E5%
['流水落花', 'Lost Love', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B5%81%E6%B0%B4%E8%90%BD%E8%8A%B1-lost-love-14874', '2023-03-17', '
```

```
['聖蛛', 'Holy Spider', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%81%96%E8%9B%9B-holy-spider-14886', '2023-03-17', '★2023 奧斯卡最佳國
['沙贊！眾神之怒', 'Shazam! Fury of the Gods', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B2%99%E8%B4%8A-%E7%9C%BE%E7%A5%9E%E4%B9%8B%E6%
['夢遊樂園', 'Melody-Go-Round', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%A4%A2%E9%81%8A%E6%A8%82%E5%9C%92-melody-go-round-14815', '202
['黑的教育', 'Bad Education', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%BB%91%E7%9A%84%E6%95%99%E8%82%B2-bad-education-14142', '2023-03
['TÁR塔爾', 'Tár', 'https://movies.yahoo.com.tw/movieinfo_main/T%C3%81R%E5%A1%94%E7%88%BE-tar-14393', '2023-03-10', '編劇/製片/導演陶德菲爾德
['驚聲尖叫6', 'Scream VI', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A9%9A%E8%81%B2%E5%B0%96%E5%8F%AB6-scream-vi-14557', '2023-03-10',
['怪談比留子 數位修復版', 'Hiruko The Goblin', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%80%AA%E8%AB%87%E6%AF%94%E7%95%99%E5%AD%90-%E6%
['天生一對2大電影：再續前緣', 'Love Destiny: The Movie', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%A4%A9%E7%94%9F%E4%B8%80%E5%B0%8D2%E5
['尋找第5味', 'Umami', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B0%8B%E6%89%BE%E7%AC%AC5%E5%91%B3-umami-14724', '2023-03-10', '東方與西
['超完美狗保姆', 'My Puppy', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B6%85%E5%AE%8C%E7%BE%8E%E7%8B%97%E4%BF%9D%E5%A7%86-my-puppy-1473
['蓋世棋蹟', 'The Royal Game', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%93%8B%E4%B8%96%E6%A3%8B%E8%B9%9F-the-royal-game-14796', '2023-
['斷網', 'Cyberheist', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%96%B7%E7%B6%B2-cyberheist-14809', '2023-03-10', '《斷網》故事描述駭客俱
['所有的美麗與血淚', 'All the Beauty and the Bloodshed', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%89%80%E6%9C%89%E7%9A%84%E7%BE%8E%E9%
['過時·過節', 'Hong Kong Family', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%81%8E%E6%99%82-%E9%81%8E%E7%AF%80-hong-kong-family-14826', '
['8釐米：詛咒影帶', '8MM: The Sinister Record', 'https://movies.yahoo.com.tw/movieinfo_main/8%E9%87%90%E7%B1%B3-%E8%A9%9B%E5%92%92%E5%BD%B1%E
['屍蹤天使', 'Mindcage', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B1%8D%E8%B9%A4%E5%A4%A9%E4%BD%BF-mindcage-14845', '2023-03-10', '★
['貓王艾維斯', 'Elvis', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B2%93%E7%8E%8B%E8%89%BE%E7%B6%AD%E6%96%AF-elvis-14907', '2023-03-10',
['媽的多重宇宙', 'Everything Everywhere All at Once', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%AA%BD%E7%9A%84%E5%A4%9A%E9%87%8D%E5%AE%
['光影帝國', 'Empire Of Light', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%85%89%E5%BD%B1%E5%B8%9D%E5%9C%8B-empire-of-light-14108', '202
['金牌拳手3', 'Creed III', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%87%91%E7%89%8C%E6%8B%B3%E6%89%8B3-creed-iii-14208', '2023-03-03',
['本日公休', 'Day Off', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%9C%AC%E6%97%A5%E5%85%AC%E4%BC%91-day-off-14569', '2023-03-03', '★ 金
['玩具當家', 'The New Toy', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%8E%A9%E5%85%B7%E7%95%B6%E5%AE%B6-the-new-toy-14581', '2023-03-03'
['驚爆點', 'Point Break', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A9%9A%E7%88%86%E9%BB%9E-point-break-14703', '2023-03-03', '★《危機
['火線埋伏', 'Ambush', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%81%AB%E7%B7%9A%E5%9F%8B%E4%BC%8F-ambush-14732', '2023-03-03', '★ 開春
['小熊維尼：血與蜜', 'Winnie the Pooh: Blood and Honey', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B0%8F%E7%86%8A%E7%B6%AD%E5%B0%BC-%E8
['鈴芽之旅', 'Suzume', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%88%B4%E8%8A%BD%E4%B9%8B%E6%97%85-suzume-14652', '2023-03-02', '★ 日本
['法貝爾曼', 'The Fabelmans', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B3%95%E8%B2%9D%E7%88%BE%E6%9B%BC-the-fabelmans-14024', '2023-02
['人肉搜索2：失蹤搜救', 'Missing', 'https://movies.yahoo.com.tw/movieinfo_main/%E4%BA%BA%E8%82%89%E6%90%9C%E7%B4%A22-%E5%A4%B1%E8%B9%A4%E6%90
['悲情城市', 'A City of Sadness', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%82%B2%E6%83%85%E5%9F%8E%E5%B8%82-a-city-of-sadness-14602',
['風再起時', 'Where The Wind blows', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A2%A8%E5%86%8D%E8%B5%B7%E6%99%82-where-the-wind-blows-14
['胡桃鉗與魔笛公主的奇幻冒險', 'The Nutcracker And The Magic Flute', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%83%A1%E6%A1%83%E9%89%97%
['不離職冒險王', 'Irreductible', 'https://movies.yahoo.com.tw/movieinfo_main/%E4%B8%8D%E9%9B%A2%E8%81%B7%E5%86%92%E9%9A%AA%E7%8E%8B-irreducti
['「鬼滅之刃」上弦集結，前進刀匠村', 'Demon Slayer Kimetsu No Yaiba To The Swordsmith Village', 'https://movies.yahoo.com.tw/movieinfo_main/-
['追海豚的長崎夏日', 'Sabakan', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%BF%BD%E6%B5%B7%E8%B1%9A%E7%9A%84%E9%95%B7%E5%B4%8E%E5%A4%8F%E
['蟻人與黃蜂女：量子狂熱', 'Ant-Man and the Wasp: Quantumania', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%9F%BB%E4%BA%BA%E8%88%87%E9%BB
['超難搞先生', 'A Man Called Otto', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B6%85%E9%9B%A3%E6%90%9E%E5%85%88%E7%94%9F-a-man-called-ot
['關於我和鬼變成家人的那件事', 'Marry My Dead Body', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%97%9C%E6%96%BC%E6%88%91%E5%92%8C%E9%AC%B
['山椒魚來了', '', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B1%B1%E6%A4%92%E9%AD%9A%E4%BE%86%E4%BA%86-14576', '2023-02-10', '★歷時17年
['僕愛君愛：致深愛妳的那個我', 'To me, The One Who Loved You', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%83%95%E6%84%9B%E5%90%9B%E6%84%
['僕愛君愛：致我深愛的每個妳', 'To Every You I' ve Loved Before', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%83%95%E6%84%9B%E5%90%9B%E6%
['日麗', 'Aftersun', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%97%A5%E9%BA%97-aftersun-14693', '2023-02-10', '★ 視與聽｜衛報｜觀察者報
['新世紀福音戰士新劇場版：終', 'Evangelion:3.0+1.0 Thrice Upon A Time', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%96%B0%E4%B8%96%E7%B4%
['瑪琳艾索普：首席女指揮', 'Conductor', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%91%AA%E7%90%B3%E8%89%BE%E7%B4%A2%E6%99%AE-%E9%A6%96%E
['我的鯨魚老爸', 'The Whale', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%88%91%E7%9A%84%E9%AF%A8%E9%AD%9A%E8%80%81%E7%88%B8-the-whale-14
['幻影', 'Phantom', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B9%BB%E5%BD%B1-phantom-14651', '2023-02-03', '★2023年開春懸疑動作最強檔！
['伊尼舍林的女妖', 'The Banshees of Inisherin', 'https://movies.yahoo.com.tw/movieinfo_main/%E4%BC%8A%E5%B0%BC%E8%88%8D%E6%9E%97%E7%9A%84%E5%
['鱷魚歌王', 'Lyle, Lyle, Crocodile', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%B1%B7%E9%AD%9A%E6%AD%8C%E7%8E%8B-lyle-lyle-crocodile-13
```

Colab 付費產品 - 按這裡取消合約

✓ 6 秒 完成時間：下午5:41