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 = Y_MOVIE_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")
```

```
80
['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%
['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%
['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%
['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%
['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%
['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
```


✓ 4 秒 完成時間：下午3:53