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/1rpYt8_SUsg4egGgC33oWy4iYYsKMRD5d?usp=sharing

---

Student ID: B0928022

Name: 杜云驊

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

儲存成功！ ×

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):
```

```
    def __init__(self):
        self.movies = []

    def get_movies(self, page_url):
        self.movies = []
        url = page_url
        for page_num in range(8):
            page_url = url + "?page=" + str(page_num + 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.strip
                intro = item.find('div', 'release_text').span.text.strip()
#                for symbol in [r'\r', r'\n']:
#                    intro.replace(symbol, '')
                self.movies.append([ch_name, en_name, movie_url, release_date, i

        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
['配樂大師顏尼歐', 'Ennio: The Maestro', 'https://movies.yahoo.com.tw/moviein
['熊蓋毒', 'Cocaine Bear', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%8
['若愛重來', 'Marriages', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%8B
['無人相信的真相', 'La syndicaliste', 'https://movies.yahoo.com.tw/movieinfo_
['闇黑對決', "The Devil's Deal", 'https://movies.yahoo.com.tw/movieinfo_main
['噩夢輓歌 4K數位修復版', 'Requiem For A Dream', 'https://movies.yahoo.com.tw/
['人體動物圖鑑：烏龜的殼其實是肋骨', 'Turtle’s Shell is a Human’s Ribs', 'https:
['流水落花', 'Lost Love', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B5
['聖蛛', 'Holy Spider', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%81%
['沙贊！眾神之怒', 'Shazam! Fury of the Gods', 'https://movies.yahoo.com.tw/m
['夢遊樂園', 'Melody-Go-Round', 'https://movies.yahoo.com.tw/movieinfo_main/
['黑的教育', 'Bad Education', 'https://movies.yahoo.com.tw/movieinfo_main/%E
['TÁR塔爾', 'Tár', 'https://movies.yahoo.com.tw/movieinfo_main/T%C3%81R%E5%
['驚聲尖叫6', 'Scream VI', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A
['怪談比留子 數位修復版', 'Hiruko The Goblin', 'https://movies.yahoo.com.tw/mo
['天生一對2大電影：再續前緣', 'Love Destiny: The Movie', 'https://movies.yahoo.
['尋找第5味', 'Umami', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B0%8B
['超完美狗保姆', 'My Puppy', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%
```

```
['蓋世棋蹟', 'The Royal Game', 'https://movies.yahoo.com.tw/movieinfo_main/%
['斷網', 'Cyberheist', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%96%B7
['所有的美麗與血淚', 'All the Beauty and the Bloodshed', 'https://movies.yahoo
['過時·過節', 'Hong Kong Family', 'https://movies.yahoo.com.tw/movieinfo_mai
['8釐米：詛咒影帶', '8MM: The Sinister Record', 'https://movies.yahoo.com.tw/
['屍蹤天使', 'Mindcage', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B1%
['貓王艾維斯', 'Elvis', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B2%93
['媽的多重宇宙', 'Everything Everywhere All at Once', 'https://movies.yahoo.c
['光影帝國', 'Empire Of Light', 'https://movies.yahoo.com.tw/movieinfo_main/
['金牌拳手3', 'Creed III', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%8
['本日公休', 'Day Off', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%9C%A
['玩具當家', 'The New Toy', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%
['驚爆點', 'Point Break', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A9
['火線埋伏', 'Ambush', 'https://movies.yahoo.com.tw/movieinfo_main/%E7%81%AB
['小熊維尼：血與蜜', 'Winnie the Pooh：Blood and Honey', 'https://movies.yahoo
['鈴芽之旅', 'Suzume', 'https://movies.yahoo.com.tw/movieinfo_main/%E9%88%B4
['法貝爾曼', 'The Fabelmans', 'https://movies.yahoo.com.tw/movieinfo_main/%E
['人肉搜索2：失蹤搜救', 'Missing', 'https://movies.yahoo.com.tw/movieinfo_main
['悲情城市', 'A City of Sadness', 'https://movies.yahoo.com.tw/movieinfo_mai
['風再起時', 'Where The Wind blows', 'https://movies.yahoo.com.tw/movieinfo_
['胡桃鉗與魔笛公主的奇幻冒險', 'The Nutcracker And The Magic Flute', 'https://m
['我們的黎明', 'Break of Dawn', 'https://movies.yahoo.com.tw/movieinfo_main/
['不離職冒險王', 'Irreductible', 'https://movies.yahoo.com.tw/movieinfo_main/
['「鬼滅之刃」上弦集結，前進刀匠村', 'Demon Slayer Kimetsu No Yaiba To The Sword
['追海豚的長崎夏日', 'Sabakan', 'https://movies.yahoo.com.tw/movieinfo_main/%E
['蟻人與黃蜂女：量子狂熱', 'Ant-Man and the Wasp: Quantumania', 'https://movies
['超難搞先生', 'A Man Called Otto', 'https://movies.yahoo.com.tw/movieinfo_ma
['關於我和鬼變成家人的那件事', 'Marry My Dead Body', 'https://movies.yahoo.com.
['山椒魚來了', '', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B1%B1%E6%A
['僕愛君愛：致深愛妳的那個我', 'To me, The One Who Loved You', 'https://movies.
['僕愛君愛：致我深愛的每個妳', 'To Every You I’ve Loved Before', 'https://movie
['日麗', 'Aftersun', 'https://movies.yahoo.com.tw/movieinfo_main/%E6%97%A5%E
['新世紀福音戰士新劇場版：終', 'Evangelion:3.0+1.0 Thrice Upon A Time', 'https:
['瑪琳艾索普：首席女指揮', 'Conductor', 'https://movies.yahoo.com.tw/movieinfo_
['我的鯨魚老爸', 'The Whale', 'https://movies.yahoo.com.tw/movieinfo_main/%E6
['幻影', 'Phantom', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B9%BB%E5
['伊尼舍林的女妖', 'The Banshees of Inisherin', 'https://movies.yahoo.com.tw/
['鱷魚歌王', 'Lyle, Lyle, Crocodile', 'https://movies.yahoo.com.tw/movieinfo
['巴比倫', 'Babylon', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B7%B4%
['詐團圓', 'Scamsgiving', 'https://movies.yahoo.com.tw/movieinfo_main/%E8%A9
['天龍八部之喬峰傳', 'Sakra', 'https://movies.yahoo.com.tw/movieinfo_main/%E5%
```

Colab 付費產品 - 按這裡取消合約

✓ 4 秒 完成時間：下午3:54 ● ✕