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/12kAOrHflnFHbTjrvlOMGK7pmTdFNwKJr

---

Student ID:B0928005

Name:湯嘉為

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")
```

者，以及500多首令人難忘的電影配樂的作曲家，本片則堪稱是他最完整肖像電影。奧斯卡導演托納多雷片中訪問了包括與莫利克奈合作過的王家衛、昆汀塔倫蒂訶
彡集《利器》）以及艾美獎得主瑪果麥汀達爾（諜報影集《冷戰諜夢》）和艾美獎得主雷里歐塔（《四海好傢伙》）。《熊蓋毒》一片的導演是伊莉莎白班克斯（
是以此為理由離開對方呢？'}

]的女友瑪麗安（珍妮佛康納莉 飾）也加入他們的計畫。這時四人還不知道，藥物將使他們的人生走向不可挽回的境地……。'}

-阿曼（《Second Chances》）飾演佩卓潘納；瑪莎米蘭絲（影集《白線》）飾演蘿莎范斯奎茲；庫柏安德魯斯（影集《陰屍路》）飾演維克多范斯奎茲；以及吉豪

惡行。第一個秘密過去，哇你真夠扯的令人髮指……第二個秘密過去，原來你不只壞而且連靈魂都壞光光了吧……第三個秘密……等等！這第三個秘密不及格吧？

二郎 X 導演：塚本晉也 X主演： 澤田研二 黃金聯手睽違32年全新修復版重返大銀幕，替影迷童年的噩夢還原上色《怪談比留子》電影原作，出自對日本漫畫界景
拉閣女士行為的文獻相同。難道馬希斯也是從未來穿越到過去的人嗎？命運又會帶領這三人面對怎樣的挑戰呢？'}
]情長片，找來曾兩度合作短片的法國傳奇影帝傑哈・德巴狄厄擔綱主演，並將舞台從法國拓展到日本北海道，集結日、法兩地影帝影后級演員：長塚京三、桑德琪
:對狗狗的唾液過敏，以至於每次跟民秀的黃金獵犬「魯尼」相處時，都必須先吞下抗過敏藥物…。面對愛犬與愛人必須二選一，民秀不得已只好幫魯尼尋找一個完

]紀實電影，榮獲奧斯卡最佳紀錄片提名、威尼斯影展最高榮譽金獅獎，由奧斯卡最佳紀錄片《第四公民》金獎導演蘿拉柏翠絲執導，記錄國際知名藝術家南戈丁的
通失能的家庭困局。以真摯沉穩的運鏡，道盡暗流湧動下心有千千結的家庭日常，更融入虛擬遊戲的科技元素，為關係修復提供另類解方。集結香港金像獎影后毛

八竿子都打不著關係的這兩個家庭，也迎來天翻地覆的變化……《玩具當家》翻拍自1976年的法國社會喜劇電影《The Toy》，由法國喜劇泰斗——賈梅德布茲飾演

，並奮力追擊這群敵兵…。【關於電影】真實故事改編，還原最震撼地下戰爭《火線埋伏》是資深影劇製作人馬克伯曼（Mark Burman）歷經多部監製作品之後，首
:林等待奇蹟出現…。遭受心理創傷的少女瑪瑞雅聽從醫師的建議，與閨蜜們一同來到百畝森林旁的木屋度假散心，木屋中女孩們的聲音吸引維尼跟小豬靠近，而女
某種神祕力量的吸引，鈴芽將手伸向了那扇門……。不久之後，日本各地開始出現類似的「門」，如果未能及時將門關上，存在於門另一端的災厄將降臨於現世。
:獎入圍者賈德赫希（《原鑽》、《凡夫俗子》）。根據史蒂芬史匹柏自己的童年所改編的半自傳電影《法貝爾曼》一片的編劇是史蒂芬史匹柏和曾經榮獲普立茲獎

不同往昔。'}

。【關於電影】《愛麗絲與藏六》今井哲也同名科幻漫畫改編！轟動國際各大影展的話題動畫電影《我們的黎明》，改編自漫畫家今井哲也的同名科幻漫畫作品，
族綁架…直到他遇見女科學家伊娃(莉蒂西亞杜希飾)，她完全改變了文森的價值觀，現在的他，將作出人生中最勇敢的決定...'}
行世界巡迴上映。因為同時收錄了《遊郭篇》的最高潮與《刀匠村篇》的第一集，才有辦法實現"上弦集結"，以大螢幕來描繪炭治郎等人的激戰與啟程。為了上映
引起日本觀眾廣泛的共鳴，鬼才導演白石和彌、齊藤工與北斗晶等各界名人都絕讚好評，許多觀眾一刷再刷，並評論這部電影是「今年最讓人感動落淚的電影」。
再度飾演黃蜂女老爸、前神盾局科學家皮姆博士，依然活躍於影視圈的資深金獎影后蜜雪兒菲佛也再度飾演皮姆妻子兼第一代黃蜂女珍娜，在熱門影集《美麗心計》

最佳男主角的許光漢、金馬最佳男配角的林柏宏、台北電影節影后王淨，三位人氣及實力兼具的演員主演，集結蔡振南、庹宗華、馬念先等實力派演員共同演出，
山林的記錄者，看見前人的戮力以赴。'}
:作公司「BAKKEN RECORD」打造《僕愛》！★《名偵探柯南》《魯邦三世》動畫製作公司「TMS」打造《君愛》！★人氣歌手須田景凪、知名樂團Saucy Dog為兩部
動畫製作公司「BAKKEN RECORD」打造《僕愛》！★《名偵探柯南》《魯邦三世》動畫製作公司「TMS」打造《君愛》！★人氣歌手須田景凪、知名樂團Saucy Dog為
困惑？蘇格蘭導演夏洛特威爾斯（Charlotte Wells）在家中找到兒時和父親的影像，發現自己已追上當年爸爸的歲數，她經由回憶的整理和翻閱，構思出這一部

改寫紀錄的女性桂冠指揮家。身為女性，瑪琳艾索普所取得的非凡成就得來並不容易。幼時隨著父親聆賞李奧納多•伯恩斯坦（Leonard Bernstein）演出的震撼體
名★ 2023 棕櫚泉國際影展 聚光燈獎 - 布蘭登費雪一名英語老師查理（布蘭登費雪 飾）中年出櫃後拋棄家庭與愛人私奔，如今他因為愛人過世而罹患暴食症，同