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/15AowHScZyYHdJzHFmpBTi9YxlitqO6zh?usp=share_link

Student ID:B0928018

Name:呂哲睿

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):
                try:
                        self.resp = requests.get(Y_MOVIE_URL)
                except:
                        self.resp = None

        def get_movies(self, page_url):
                soup = BeautifulSoup(self.resp.text, 'html.parser')
                movie_list = []
                for movie in soup.find_all("div", class_="release_info_text"):
                        ch_name = movie.find("div", class_="release_movie_name").a.text.strip()
                        en_name = movie.find("div", class_="release_movie_name").find("div", class_="en").a.text.strip()
                        movie_url = movie.find("div", class_="release_movie_name").a["href"]
                        release_date = movie.find("div", class_="release_movie_time").text.strip()
                        intro = movie.find("div", class_="release_text").text.strip()
                        movie_info = {"ch_name": ch_name, "en_name": en_name, "movie_url": movie_url, "release_date": release_date, "intro":
                        movie_list.append(movie_info)
                return movie_list

# # 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")
```

```
4413', 'release_date': '上映日期：\n                    2023-03-17', 'intro': '★義大利奧斯卡大衛獎最佳紀錄片、最佳剪輯、最佳音效三項大獎\r\n★義大利
3-03-17', 'intro': '故事靈感來自發生在1985年的真實事件，一名毒販的飛機失事墜機，遺失了一批古柯鹼，結果被一隻黑熊吃掉。這部劇情既荒誕又瘋狂的喜劇
  2023-03-17', 'intro': '★賽博台客搖滾樂團 美秀集團 ★重磅合作電影主題曲〈戀人〉\r\n★義大利奧斯卡金獎編劇《完美陌生人》導演 保羅克斯泰拉 最新力
  'release_date': '上映日期：\n                    2023-03-17', 'intro': '她是最大受害者\u3000也是唯一嫌疑犯\n\r\n本片改編自莫琳嘉內（伊莎貝雨蓓飾）
               2023-03-17', 'intro': '★《我只是個計程車司機》製作人X《犯罪都市》製作團隊攜手打造犯罪鉅獻！\r\n★《財閥家的小兒子》李星民X《無間
%E7%89%88-requiem-for-a-dream-14847', 'release_date': '上映日期：\n                    2023-03-17', 'intro': '★ 《黑天鵝》《我的鯨魚老爸》金獎名導成
91%91-%E7%83%8F%E9%BE%9C%E7%9A%84%E6%AE%BC%E5%85%B6%E5%AF%A6%E6%98%AF%E8%82%8B%E9%AA%A8-turtles-shell-is-a-humans-ribs-14856', 'release_date': '
```

```
 2023-03-17', 'intro': '★本屆香港電影金像獎最佳女主角、最佳服裝造型設計、最佳原創電影歌曲三項大獎提名！\r\n★第29屆《香港電影評論學會大獎》最佳
tro': '★2023 奧斯卡最佳國際影片 丹麥代表\r\n★入圍 第75屆坎城影展 主競賽\r\n★榮獲 第75屆坎城影展 最佳女主角獎\r\n★榮獲 第33屆斯德哥爾摩國際電
ds-13749', 'release_date': '上映日期：\n                    2023-03-16', 'intro': '《沙贊！眾神之怒》接續描述青少年比利貝特森的故事，只要他喊出神奇字
```