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/1JtNUYu3kF3eqEWIlpEzk_EKy3n1xxEFl?usp=sharing

---

Student ID:B0928012

Name:王晟翰

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.parser = "html.parser"
                self.movies = []



        def get_movies(self, page_url):
                # 發送請求並獲取頁面內容
                res = requests.get(page_url, headers=self.headers)
                soup = BeautifulSoup(res.text, "html.parser")

                # 用 BeautifulSoup 提取電影信息
                movie_list = soup.find_all("div", class_="release_info")
                for movie in movie_list:
                        # 獲取中文名稱
                        ch_name = movie.find("div", class_="release_movie_name").a.text.strip()
                        # 獲取英文名稱
                        en_name = movie.find("div", class_="release_movie_name").find("div", class_="en").a.text.strip()
                        # 獲取電影網址
                        movie_url = movie.find("div", class_="release_movie_name").a["href"]
                        # 獲取上映日期
                        release_date = movie.find("div", class_="release_movie_time").text.split(": ")[-1].strip()
                        # 獲取簡介
                        intro = movie.find("div", class_="release_text").text.strip()
                        # 將電影信息添加到 movies 列表中
                        movie_dic={
                                "ch_name": ch_name,
                                "en_name": en_name,
                                "movie_url": movie_url,
                                "release_date": release_date,
                                "intro": intro
                        }
```

```
                self.movies.append(movie_dic)

            next_page = soup.find('a', rel='next')
            if next_page:
                next_page_url = next_page['href']
                self.get_movies(next_page_url)



            return 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")
```

```
10
{'ch_name': '配樂大師顏尼歐', 'en_name': 'Ennio: The Maestro', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%85%8D%E6%A8%82%E5%A4
{'ch_name': '熊蓋毒', 'en_name': 'Cocaine Bear', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%86%8A%E8%93%8B%E6%AF%92-cocaine-be
{'ch_name': '若愛重來', 'en_name': 'Marriages', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%8B%A5%E6%84%9B%E9%87%8D%E4%BE%86-ma
{'ch_name': '無人相信的真相', 'en_name': 'La syndicaliste', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E7%84%A1%E4%BA%BA%E7%9B%B8
{'ch_name': '闇黑對決', 'en_name': "The Devil's Deal", 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%97%87%E9%BB%91%E5%B0%8D%E6%B
{'ch_name': '噩夢輓歌 4K數位修復版', 'en_name': 'Requiem For A Dream', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%99%A9%E5%A4%
{'ch_name': '人體動物圖鑑：烏龜的殼其實是肋骨', 'en_name': 'Turtle's Shell is a Human's Ribs', 'movie_url': 'https://movies.yahoo.com.tw/movie
{'ch_name': '流水落花', 'en_name': 'Lost Love', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B5%81%E6%B0%B4%E8%90%BD%E8%8A%B1-lo
{'ch_name': '聖蛛', 'en_name': 'Holy Spider', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%81%96%E8%9B%9B-holy-spider-14886', 'r
{'ch_name': '沙贊！眾神之怒', 'en_name': 'Shazam! Fury of the Gods', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%B2%99%E8%B4%8A
```

Colab 付費產品 - 按這裡取消合約