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/1lSbuX1YQsjrwj-QfkY9YKwDeABU9jiku?usp=share_link

---

```
# 此內容會顯示為程式碼
```

Student ID:B0928006

Name:廖采葳

Question 1 (100 points)

儲存成功！

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!
```

```
class MovieCrawler(object):

    def __init__(self):
          pass

    def get_movies(self, page_url):
        movies = []
        resp = requests.get(page_url)
        soup = BeautifulSoup(resp.text, 'html.parser')

        for row in soup.select('.release_list .release_info'):
            movie_dict = {}
            # movie_dict['ch_name'] = row.select_one('.release_movie_name').text.strip()
            movie_dict['ch_name'] =  re.sub(r'\s+', ' ', row.select_one('.release_movie_name').text

            movie_dict['en_name'] = row.select_one('.en').text.strip()
            movie_dict['movie_url'] = row.select_one('.release_movie_name a')['href']
            movie_dict['release_date'] = row.select_one('.release_movie_time').text.split('：')[-1]
            movie_dict['intro'] = row.select_one('.release_text').text.strip()
            movies.append(movie_dict)

        return movies


# # DO NOT MODIFY THE VARIABLES
crawler = MovieCrawler()
movies = crawler.get_movies(Y_MOVIE_URL)

# # THE RESULTS : AS THE FOLLOWING SECTION
                                                'release_date', 'intro'}
print('movies', sep='\n')
```

儲存成功！

```
10
{'ch_name': '配樂大師顏尼歐 Ennio: The Maestro 期待度 100% 網友想看 滿意度 0 綜合評分', 'en_n
{'ch_name': '熊蓋毒 Cocaine Bear 期待度 83% 網友想看 滿意度 0 綜合評分', 'en_name': 'Cocaine
{'ch_name': '若愛重來 Marriages 期待度 50% 網友想看 滿意度 0 綜合評分', 'en_name': 'Marriages
{'ch_name': '無人相信的真相 La syndicaliste 期待度 100% 網友想看', 'en_name': 'La syndicalist
{'ch_name': "闇黑對決 The Devil's Deal 期待度 100% 網友想看 滿意度 0 綜合評分", 'en_name': "T
{'ch_name': '噩夢輓歌 4K數位修復版 Requiem For A Dream 期待度 100% 網友想看 滿意度 0 綜合評分
{'ch_name': '人體動物圖鑑：烏龜的殼其實是肋骨 Turtle’s Shell is a Human’s Ribs 期待度 100%
{'ch_name': '流水落花 Lost Love 期待度 100% 網友想看 滿意度 0 綜合評分', 'en_name': 'Lost Lov
{'ch_name': '聖蛛 Holy Spider 期待度 100% 網友想看 滿意度 0 綜合評分', 'en_name': 'Holy Spide
{'ch_name': '沙贊！眾神之怒 Shazam! Fury of the Gods 期待度 67% 網友想看 滿意度 0 綜合評分',
```

Colab 付費產品 - 按這裡取消合約

儲存成功！