# 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/1_Gw6MaB7HHa9Tor46aLUF9jPmyeoNe3n?usp=sharing#scrollTo=W9Fm6AQJQDFa---

**Student ID**:B09-2802 9

**Name**:毛謙鎧

# 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

---

編集するにはダブルクリックするか Enter キーを押してください

```
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 get_movies(self, page_url):


# # 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")
```

```
    68
    {'ch_name': '黑的教育', 'en_name': 'Bad Education', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%BB%91%E7%9A%84%E6%95%99%E8%82%B
    {'ch_name': 'TÁR塔爾', 'en_name': 'Tár', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/T%C3%81R%E5%A1%94%E7%88%BE-tar-14393', 'releas
    {'ch_name': '驚聲尖叫6', 'en_name': 'Scream VI', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E9%A9%9A%E8%81%B2%E5%B0%96%E5%8F%AB6-s
    {'ch_name': '怪談比留子數位修復版', 'en_name': 'Hiruko The Goblin', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%80%AA%E8%AB%87%E
    {'ch_name': '天生一對2大電影：再續前緣', 'en_name': 'Love Destiny: The Movie', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%A4%A
    {'ch_name': '尋找第5味', 'en_name': 'Umami', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E5%B0%8B%E6%89%BE%E7%AC%AC5%E5%91%B3-umami
    {'ch_name': '超完美狗保姆', 'en_name': 'My Puppy', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%B6%85%E5%AE%8C%E7%BE%8E%E7%8B%97%
    {'ch_name': '蓋世棋蹟', 'en_name': 'The Royal Game', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E8%93%8B%E4%B8%96%E6%A3%8B%E8%B9%
    {'ch_name': '斷網', 'en_name': 'Cyberheist', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%96%B7%E7%B6%B2-cyberheist-14809', 'rele
    {'ch_name': '所有的美麗與血淚', 'en_name': 'All the Beauty and the Bloodshed', 'movie_url': 'https://movies.yahoo.com.tw/movieinfo_main/%E6%89%8
```

```
import requests
import re
import json
from bs4 import BeautifulSoup


Y_MOVIE_URL = 'https://tw.movies.yahoo.com/movie_thisweek.html'

Y_MOVIE_INFO_URL = 'https://tw.movies.yahoo.com/movieinfo_main.htm'
Y_MOVIE_PHOTO_URL = 'https://tw.movies.yahoo.com/movieinfo_photos.html'
Y_MOVIE_TIME_URL = 'https://tw.movies.yahoo.com/movietime_result.html'
```

```
def get_web_page(url):
    resp = requests.get(url)
    if resp.status_code != 200:
        print('Invalid url: ', resp.url)
        return None
    else:
        return resp.text


def get_movies(dom):
    soup = BeautifulSoup(dom, 'html5lib')
    movies = []
    rows = soup.find_all('div', 'clearfix row')
    for row in rows:
        movie = dict()
        movie['expectation'] = row.find(id='ymvle').find('div', 'bd clearfix ').em.text
        movie['ch_name'] = row.find('div', 'text').h4.text
        movie['en_name'] = row.find('div', 'text').h5.text
        movie['movie_id'] = get_movie_id(row.find('div', 'text').h4.a['href'])
        movie['poster_url'] = row.find('div', 'img').img['src'].replace('mpost4', 'mpost')
        movie['release_date'] = get_date(row.find('div', 'text').span.text)
        movie['intro'] = row.find('div', 'text').p.text.replace(u'...詳全文', '').replace('\n', '')
        trailer_li = row.find('div', 'text').find('li', 'trailer')
        movie['trailer_url'] = get_trailer_url(trailer_li.a['href']) if trailer_li else ''
        movies.append(movie)
    return movies
```