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/1qoOS8lRBp1OB9kV8_MmKdf2Nod0ixXoL?usp=sharing

---

Student ID: B0928002

Name: 林力

```
##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 th
score

---
```

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):

    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.c
{'ch_name': 'TÁR塔爾', 'en_name': 'Tár', 'movie_url': 'https://movies.yahoo.com.tw/movi
{'ch_name': '驚聲尖叫6', 'en_name': 'Scream VI', 'movie_url': 'https://movies.yahoo.com.
{'ch_name': '怪談比留子數位修復版', 'en_name': 'Hiruko The Goblin', 'movie_url': 'https://
{'ch_name': '天生一對2大電影：再續前緣', 'en_name': 'Love Destiny: The Movie', 'movie_url':
{'ch_name': '尋找第5味', 'en_name': 'Umami', 'movie_url': 'https://movies.yahoo.com.tw/m
{'ch_name': '超完美狗保姆', 'en_name': 'My Puppy', 'movie_url': 'https://movies.yahoo.com
{'ch_name': '蓋世棋蹟', 'en_name': 'The Royal Game', 'movie_url': 'https://movies.yahoo.
{'ch_name': '斷網', 'en_name': 'Cyberheist', 'movie_url': 'https://movies.yahoo.com.tw/
{'ch_name': '所有的美麗與血淚', 'en_name': 'All the Beauty and the Bloodshed', 'movie_url
```

```
import requests
import re
from bs4 import BeautifulSoup

Y_MOVIE_URL = "https://movies.yahoo.com.tw/movie_intheaters.html"

# browser =
def get_page():
  movies = []
  for i in range(1,9):
    try:
      resp = requests.get("https://movies.yahoo.com.tw/movie_intheaters.html"+"?page="+str(i
    except:
      resp = None

    if resp and resp.status_code == 200:
      print(resp.status_code)
      soup = BeautifulSoup(resp.text, 'html.parser')
      ch_list = soup.find_all('div', 'release_movie_name')
      # print(*h2_list, sep="\n")
      en_list = soup.find_all('div', 'en')
      hrefs = soup.find_all('a', 'gabtn')
      dates = soup.find_all('div', 'release_movie_time')
      intros = soup.find_all('div', 'release_text')
      # soup.find_element_by_class_name("nexttxt").click()

      c = intros
      # page_next = soup.find_element_by_class_name("nexttxt")
      for i in range(len(ch_list)):
        movie_info = {
            'ch_name':ch_list[i].text,
            'en_name':en_list[i].text,
            'movie_url':ch_list[i].a["href"],
            'release_date':dates[i].text,
            'intro':intros[i].text
        }
      movies.append(movie_info)
  return movies
```

```
        # for div in ch_list:
        #   print(div.a.text)
        # for div in en_list:
        #   print(div.a.text)
        # for href in ch_list:
        #   print (href.a["href"])
        # for date in dates:
        #   print (div.text)
        # for intro in intros:
        #   print (intro.text)

        # pages = soup.find_all('div', 'page_numbox')

        # print (pages)
        # page = []
        # page = pages
        # for pages in page:
        #   print (pages)

        # print (page.text)
        # print (c[1].text)



movies = get_page()
print(movies)
```

```
    200
    200
    200
    200
    200
    200
    200
    200
    [{'ch_name': '\n\n                    沙贊！眾神之怒\n\n\n                    Shazam! Fury
```

✓ 4 秒 完成時間：下午3:56