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/1Ot0oUt6NdtPdYgnkBO71Dlcz0gf9OBSh?usp=sharing

---

Student ID: B0929038 **Name**: 吳安捷

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):
            movies_info={}
            resp = requests.get(page_url)
            text=BeautifulSoup(resp.text)
            #find all movie page
            links=text.find("div",class_="page_numbox")
            links=links.find_all("a")
            movie_pages=["https://movies.yahoo.com.tw/movie_intheaters.html"]
            for link in links:
                if link['href'] not in movie_pages:
                    movie_pages.append(link['href'])
            #find movies in one page
            mv_page="https://movies.yahoo.com.tw/movie_intheaters.html"
            resp = requests.get(mv_page)
            text=BeautifulSoup(resp.text)
            links=text.find_all("div",class_="en")
            movie_links=[]
            for link in links:
                link=link.find("a")
                if link["href"] not in movie_links:
                    movie_links.append(link['href'])
            #movie info
            movie_link=movie_links[0]
            resp = requests.get(movie_link)
            text=BeautifulSoup(resp.text)
            content=text.find("div",class_="movie_intro_info_r")
            ch_name=content.find("h1").text
            en_name=content.find("h3").text
            movie_url=movie_link
            spans=text.find_all("span")
            for span in spans:
                if not span.text.find("上映日期"):
                    release_date=span.text[5:-1]
                    break
            intro=text.find("span",id="story")
            intro=intro.text
```

```
            movies_info.add({ch_name:ch_name,en_name:en_name})
            #  print(ch_name,en_name,movie_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")
```

★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
★義大利國家影評協會年度銀絲帶獎
★義大利巴里國際影展最佳導演
★加爾各答國際邪典電影節最佳紀錄片

他是最具影響力的音樂家，
2座奧斯卡獎得主，500多首令人難忘的電影配樂家…

《配樂大師顏尼歐》是全球最知名和受人喜愛的音樂大師顏尼歐莫利克奈的傳記電影。電影製作歷時5年、橫跨歐美進行拍攝，透過音樂和檔案片段講述了一代音樂巨

顏尼歐莫利克奈是20世紀最具影響力和最多產的音樂家之一，兩項奧斯卡獎的獲得者，以及500多首令人難忘的電影配樂的作曲家，本片則堪稱是他最完整肖像電影。

✓ 1 秒 完成時間：下午3:50