爬取Yahoo!電影資料

In [15]:

```
import requests
from bs4 import BeautifulSoup
import json

def fetch_movie_data(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    cname = soup.find('div', {'class': 'movie_intro_info_r'}).find('h1').text.strip().re
    ename = soup.find('div', {'class': 'movie_intro_info_r'}).find('h3').text
    labels = soup.find_all('div', {'class': 'level_name'})
    class_labels = [label.text.strip() for label in labels if label.text.strip() not in 
    intro = soup.find('span', {'id': 'story'}).text.strip().replace('\n', '').replace('\
    spans = soup.find('div', {'class': 'movie_intro_info_r'}).find_all('span')
    for span in spans:
        if '上映日期' in span.text:
            released_date = span.text.strip().replace('上映日期：', '').replace('播出日期
            break
        elif '播出日期' in span.text:
            released_date = span.text.strip().replace('上映日期：', '').replace('播出日期
            break
    data = {
        'doc_id':0,
        'cname': cname,
        'ename': ename,
        'pagerank':0,
        'label': class_labels,
        'intro': intro,
        'released_date': released_date,
        'links':url
    }

    return data

base_url = 'https://movies.yahoo.com.tw/movieinfo_main/'
movies_data = []

for i in range(1, 15065):
    url = base_url + str(i)
    try:
        movie_data = fetch_movie_data(url)
        movie_data['doc_id']=i
        print(f"Movie {i}:")
        print(movie_data)
        movies_data.append(movie_data)
    except Exception as e:
        print(f"Error while fetching data for movie {i}: {e}")

# 將資料存成JSON檔案
with open('movies_data_1.json', 'w', encoding='utf-8') as f:
    json.dump(movies_data, f, ensure_ascii=False, indent=4)
```

```
Movie 1:
{'doc_id': 1, 'cname': '一世狂野', 'ename': 'Blow', 'pagerank': 0, 'labe
l': ['劇情', '犯罪', '歷史/傳記'], 'intro': '喬治戎格一生都在追求所謂的美國
夢，也就是享受美好富裕的生活，但是他卻不願像他父親那樣一輩子都只是個出賣勞力的
建築工人。於是他搬到陽光明媚的加州，靠著販賣大麻賺錢，起初，他販毒只是為了享受
自由自在的生活，但是當他野心越來越大，他的勢力也日益坐大之際，卻在此時被捕入
獄。他在牢裡認識一個能言善道，自稱熟識哥倫比亞販毒集團的牢友狄亞哥，他出獄後果
真把當時勢力最大的毒梟艾斯科巴介紹給喬治認識，艾斯科巴計畫將古柯鹼大量引進美國
的迪斯可舞廳，希望能引領一股吸毒狂歡的風潮。除了毒品供應商之外，狄亞哥也介紹了
一個美艷又狂野的女人瑪莎給喬治，他們瘋狂相愛，之後馬莎還替他生下一個可愛的女兒
克莉絲汀娜，也是喬治一生的最愛。喬治很快就靠著販毒發大財，他還得買一棟大房子專
門存放每天賺進來的大把鈔票，但是日進斗金卻整天提心吊膽的生活卻讓喬治開始省思，
到底他要繼續過著揮霍富裕的生活，還是為了自己心愛的女兒應該轉性投資正當的事業？
可是這時聯邦調查局的探員，也開始盯上毒源禍首的喬治……', 'released_date': '200
1-10-12', 'links': 'https://movies.yahoo.com.tw/movieinfo_main/1'}
Movie 2:
{'doc_id': 2, 'cname': '玩命關頭', 'ename': 'The Fast and the Furious',
'pagerank': 0, 'label': ['動作', '劇情', '犯罪', '懸疑/驚悚'], 'intro':
'唐米尼杜洛托是洛城街頭賽車界的老大哥，他身邊有一群忠心耿耿的手下，他白天忙著
組裝高性能跑車，晚上則是開著他的愛車，動輒以一次一萬美元的賭注和別人飆車，在
```

中文分詞後，建立 Inverted Index，利用 PageRank 演算法來排序

In [138]:

```python
import json
import jieba
from collections import defaultdict
import re
import nltk
from nltk.corpus import stopwords


# 讀取資料
with open('movies_data_2.json', 'r', encoding='utf-8') as f:
    movies_data = json.load(f)


nltk.download('stopwords')
stop_words = set(stopwords.words('chinese'))

# 分詞並過濾中文停用詞
for movie in movies_data:
    # 只對 cname, label, intro 欄位進行分詞
    for field in ['cname', 'label', 'intro']:
        if field == 'label':
            words = jieba.cut(' '.join(movie[field]))
        else:
            words = jieba.cut(movie[field])
        # 過濾中文停用詞
        words = [w for w in words if w not in stop_words]
        # 將分詞結果合併起來
        movie[field + '_words'] = ' '.join(words)
        movie[field + '_words'] = re.sub(r'[^\w\s]', '', ' '.join(words))

# 建立 Inverted Index
inverted_index = defaultdict(set)
for movie in movies_data:
    # Convert the 'label' field to string and concatenate its elements
    label_words = ' '.join(movie['label'])
    # Perform word segmentation on the three fields
    for field in ['cname', 'intro', 'label_words']:
        words = jieba.cut(movie[field])
        movie[field + '_words'] = ' '.join(words)

for movie in movies_data:
    for field in ['cname_words', 'label_words', 'intro_words']:
        for word in movie[field].split():
            inverted_index[word].add(movie['doc_id'])

import networkx as nx

# 創建有向圖
G = nx.DiGraph()

# 添加節點
for movie in movies_data:
    G.add_node(movie['doc_id'])

# 添加邊
for movie in movies_data:
    for incoming_movie_id in inverted_index[movie['cname_words']]:
        G.add_edge(incoming_movie_id, movie['doc_id'])
```

```
# 計算 PageRank
page_ranks = nx.pagerank(G, alpha=0.85)
for movie in movies_data:
    movie['pagerank'] = page_ranks[movie['doc_id']]
# 按照 PageRank 值進行排序
ranked_movies = sorted(movies_data, key=lambda x: page_ranks[x['doc_id']], reverse=True)

with open('movies_data_with_page_ranks.json', 'w', encoding='utf-8') as f:
    json.dump(movies_data, f, ensure_ascii=False, indent=4)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\iwin4\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [141]:

```
# 讀取JSON文件
with open('movies_data_with_page_ranks.json', 'r', encoding='utf-8') as f:
    movies_data = json.load(f)

# # 刪除cname_words和label_words_words欄位
# for movie in movies_data:
#     del movie['cname_words']
#     del movie['label_words']
#     del movie['label_words_words']
#     del movie['intro_words']

with open('hw2.json', 'w', encoding='utf-8') as f:
    json.dump(movies_data, f, ensure_ascii=False, indent=4)
```

輸入搜尋關鍵字，輸出搜尋結果呈現

In [136]:

```
import re

# 輸入搜尋關鍵字
keyword = '葉問'

# 將搜尋關鍵字變色的函數
def highlight_keyword(text, keyword):
    return re.sub(r'(' + keyword + ')', r'\033[1m\033[91m\1\033[0m', text, flags=re.IGNO

# 搜尋符合關鍵字的電影並輸出相關資訊
count = 0
search_engine = []
for movie in ranked_movies:
    if keyword in movie['cname_words'] or keyword in movie['intro_words'] or keyword in
        count += 1
        search_engine.append(f"{movie['doc_id']} ({page_ranks[movie['doc_id']]}): {highl

print(f"您的搜尋結果(Sorting by PageRank Value):")
print(f"共 {count} 筆，符合\"{keyword}\" --- 共 indexing {len(ranked_movies)} 筆電影資料")
for result in search_engine:
    print("-----------------------------------------------------------------------")
    print(result)
```

```
您的搜尋結果(Sorting by PageRank Value):
共 33 筆，符合"葉問" --- 共 indexing 12236 筆電影資料
-----------------------------------------------------------------------
3320 (0.0007836892534134854): 錦衣衛, 14 Blades, ★2010虎年賀歲，華語動作
片首選，與全亞洲同步上映★動作巨星甄子丹繼《葉問》後，帶您認識中國歷史上最神祕
殘酷的特務系統–「錦衣衛」違旨抗命…殺！干政弄權…殺！通敵叛國…殺！無孔不入，寧枉
毋縱「錦衣衛」人人聞之色變的恐怖特務機構！傳言中錦衣怒馬，白刃如雪，殺人於無形
的祕密警察組織，是明太祖朱元璋為施高壓統治，精挑禁衛軍內身懷絕技、忠心不二的大
內高手所組成。這班武功高強的死士，直接受皇命指揮，完全聽從於皇帝，佈下天羅地
網，以構陷冤獄等泯滅人性手段殘害忠良、以冷血殘虐的獨門暗器誅殺異己，無孔不入、
寧枉毋縱，來鞏固天子權威，造成朝野聞風喪膽，人人惶恐自危，寫下中國歷史上最黑暗
的一頁…。故事大綱：明朝末年，閹黨弄權亂政，「錦衣衛」被當朝司禮太監賈精忠所掌
控。賈精忠密謀造反，私通藩王。錦衣衛四大護法之首青龍（甄子丹　飾）奉師命盜取勤
王兵符卻遭閹黨構陷，曾經令人聞風喪膽的錦衣衛頭目，如今竟成同門追殺的目標。千鈞
一髮之際，青龍幸獲馭天鏢局總鏢頭喬永之女喬花（趙薇　飾）相助，攜手亡命江湖。兩
人一路遭遇錦衣衛和各路江湖高手。正當青龍逐漸查清賈精忠圖謀造反之陰謀真相之際，
行蹤卻被錦衣衛佈下的天羅地網所掌握，並派出神鬼莫測的淨衣派頭號女殺手脫脫（徐子
珊　飾）追殺，青龍在重重諜網，腹背受敵下，幸賴人稱大漠判官的江湖第一大幫主（吳
尊　飾），被青龍懲奸除惡的義行感動，決定拔刀相助。一場最後的正邪決戰，即將展
```

計算查詢後的Precision and recall值

In [137]:

```
x, y, z = 0, 0, 0
search_engine = []
for movie in ranked_movies:
    if keyword in movie['cname_words'] or keyword in movie['intro_words'] or keyword in
        x += 1
    if re.search(keyword, movie['cname_words'] + movie['intro_words'] + ' '.join(movie['
        y += 1
    if re.search(keyword, movie['cname'] + movie['intro'] + ' '.join(movie['label_words'
        z += 1
if y != 0:
    print("Precision: {:.2f}% -- {}/{}".format((x / y) * 100, x, y))
else:
    print("Precision: N/A -- {}/{}".format(x, y))
if z != 0:
    print("Recall: {:.2f}% -- {}/{}".format((x / z) * 100, x, z))
else:
    print("Recall: N/A -- {}/{}".format(x, z))
```

```
Precision: 100.00% -- 33/33
Recall: 100.00% -- 33/33
```

In [ ]:

```
```