In [70]:

```
pip install bs4
```

```
Requirement already satisfied: bs4 in c:\python311\lib\site-packages (0.0.1)
Requirement already satisfied: beautifulsoup4 in c:\python311\lib\site-packages
(from bs4) (4.11.2)
Requirement already satisfied: soupsieve>1.2 in c:\python311\lib\site-packages
(from beautifulsoup4->bs4) (2.4)
Note: you may need to restart the kernel to use updated packages.
```

In [71]:

```
pip install html5lib
```

```
Requirement already satisfied: html5lib in c:\python311\lib\site-packages (1.1)
Requirement already satisfied: six>=1.9 in c:\python311\lib\site-packages (from
html5lib) (1.16.0)
Requirement already satisfied: webencodings in c:\python311\lib\site-packages
(from html5lib) (0.5.1)
Note: you may need to restart the kernel to use updated packages.
```

In [72]:

```
pip install lxml
```

```
Requirement already satisfied: lxml in c:\python311\lib\site-packages (4.9.2)
Note: you may need to restart the kernel to use updated packages.
```

In [73]:

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

In [ ]:

```
movies = []
n = 0

for page in range(1, 15068):
    response = requests.get("https://movies.yahoo.com.tw/movieinfo_main/"+str(pa
    soup = BeautifulSoup(response.text, 'html.parser')
    doc_id = page

    try:
        info_items = soup.find('div', 'movie_intro_info_r')
        cname = info_items.find('h1').text
        ename = info_items.find('h3').text
        release_date = info_items.find('span', class_=None).text[5:]

        info_items = soup.find('div', 'level_name_box').find_all('div', 'level_r
        labels = []
        for label in info_items:
            labels.append(label.text.strip())

        info_items = soup.find('div', 'gray_infobox_inner')
        intro = str(info_items.select_one('span').text).strip().replace('\n', ''
    except:
        continue

    movies.append([doc_id, cname, ename, labels, intro, release_date])
    n += 1

with open('hw2.json', 'w', encoding='utf-8') as f:
    json.dump(movies, f, indent=4)
# print(n)
# print(*movies, sep="\n")
```

In [ ]:

In [ ]: