社會科學程式設計
期末專題
2018.01.14

從社會議題分析奧斯卡入圍名單

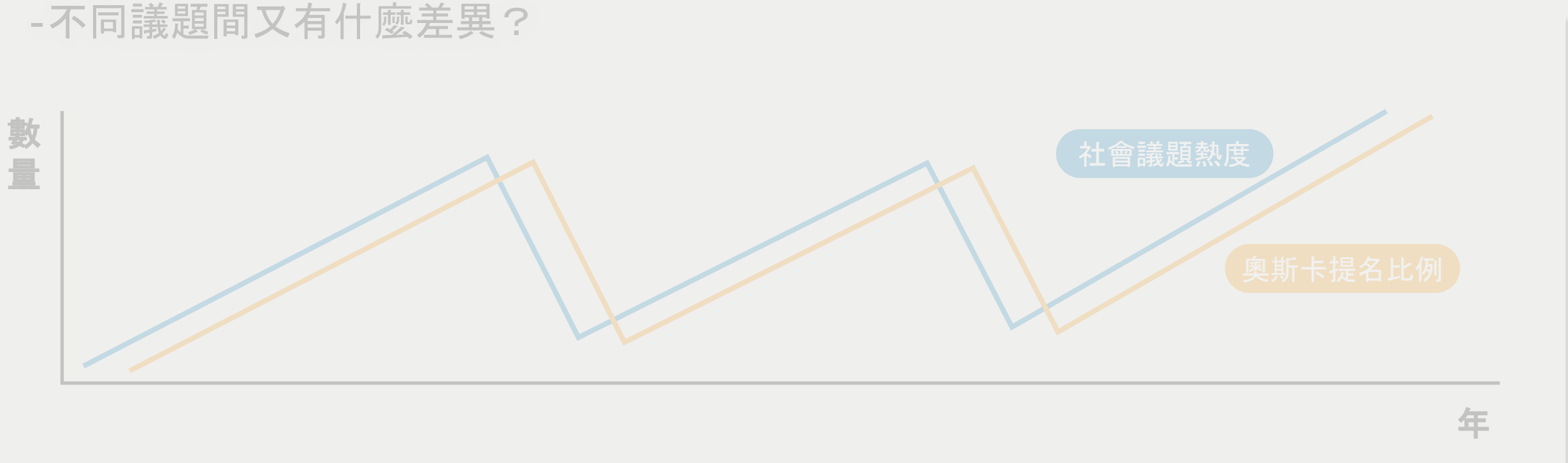
政論一 / 阮彥璋
社會一 / 王聖夫 劉哲愷 柯亮宇

問題意識：

-社會議題的討論熱度是否會影響奧斯卡入圍影片名單？

-所謂「政治正確」是否影響奧斯卡評審的選擇？

-不同議題間又有什麼差異？



簡報順序

電影分析

奧斯卡入圍名單

→

社會議題

LGBT、種族

A.電影分析(預計):

抓取片單 → 抓取資料 → 判斷類型

Crawl

從IMDb、boxofficereport、Wiki抓取:

1. Oscar:獎項入圍提名影片(最佳影片、原創劇本、改編劇本、長紀錄片)
2. LGBT-related films
3. Racism-related related films
4. top250 films (2004-2017)

Crawl

將電影丟入IMDb Search取得電影

1. Plot summary
2. Keywords

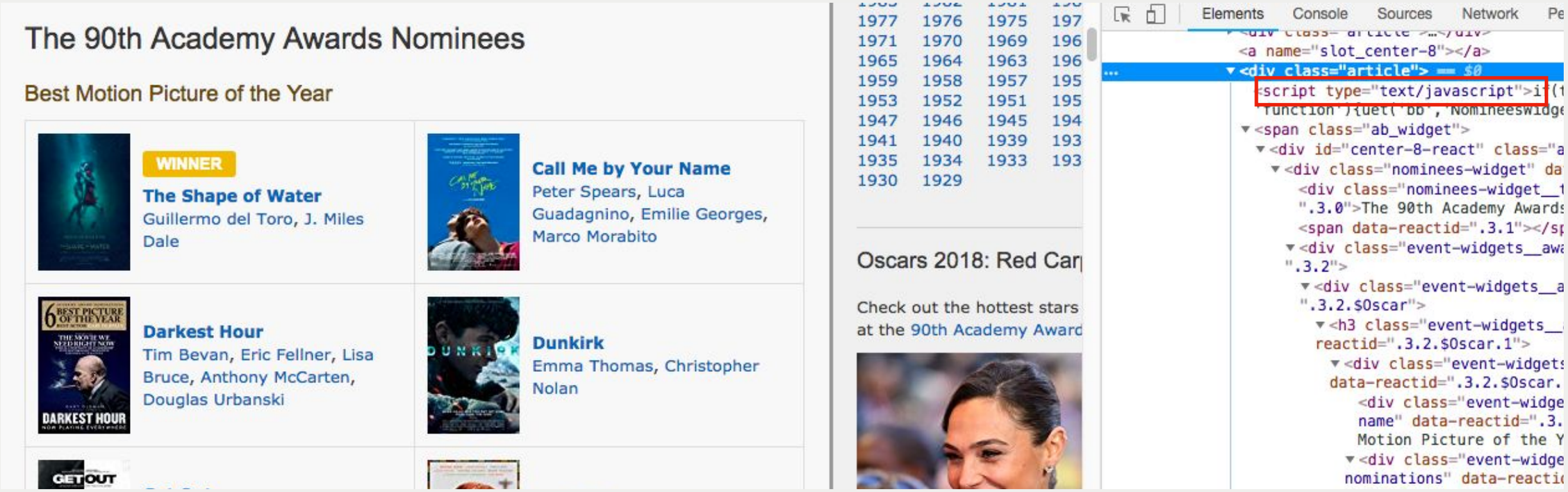
Train

Training & Predict

1. 用以上資料train出一套檢驗工具。
2. 判斷Oscar入圍影片是否是LGBT or Race-related film

1.抓取片單

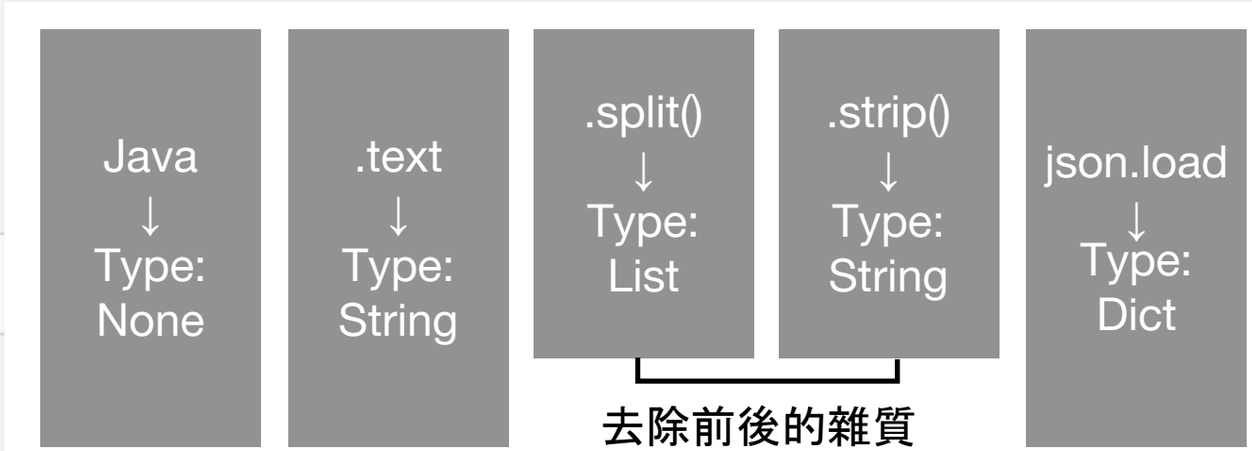
以從IMDb爬下奧斯卡四個獎項的提名名單為例




```
    list_1=[]
    for div in soup.find_all(class_="article"):
        list_1.append(div.find(class_="ab_widget"))
    string_1=list_1[0].text.strip()
    list_2=string_1.split("\n")
    string_2=list_2[2].strip("]);")
    list_3=string_2.split("',")
    string_3=list_3[1]
    list_a.append(string_3)
print(len(list_a))
```

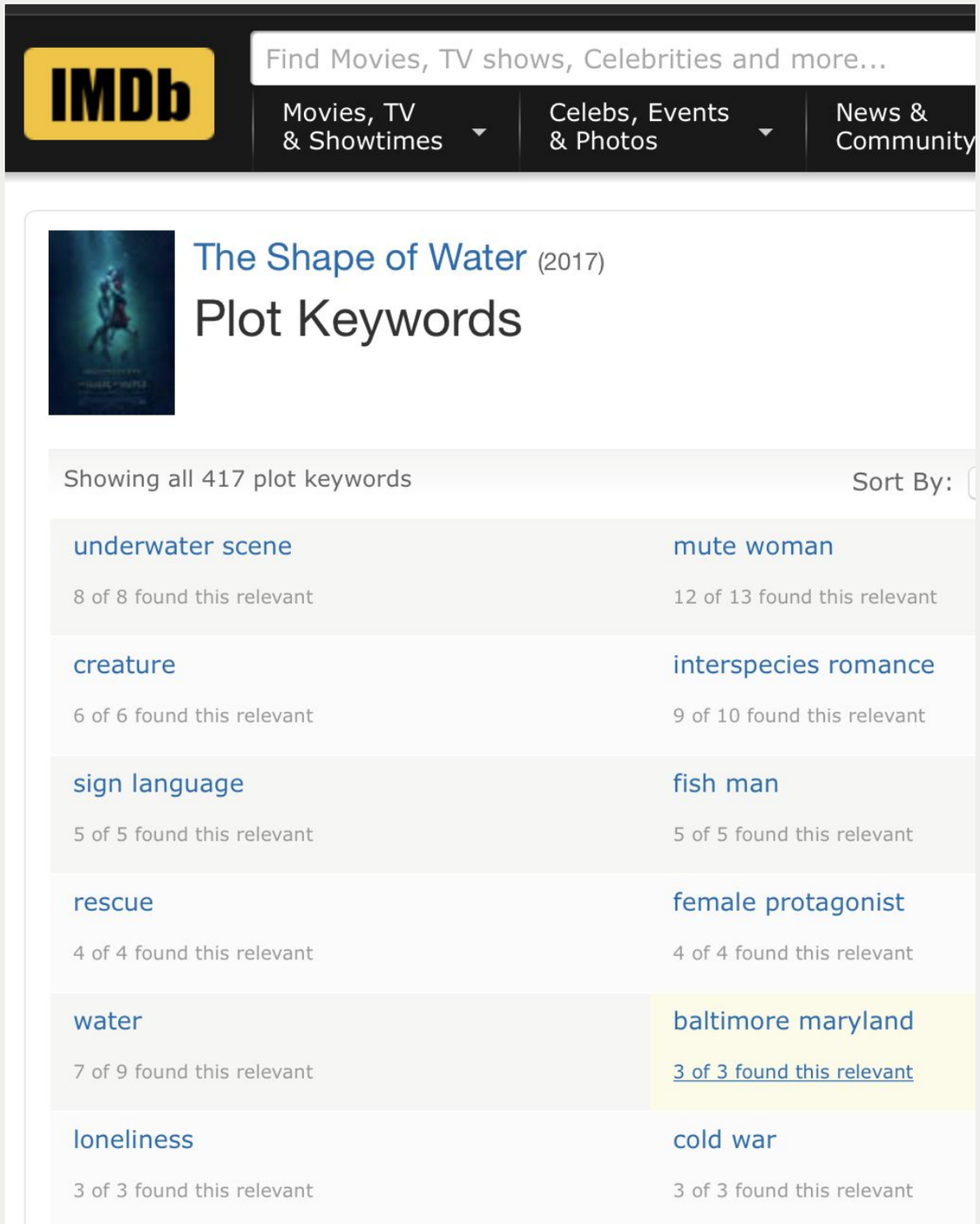
```
15
```

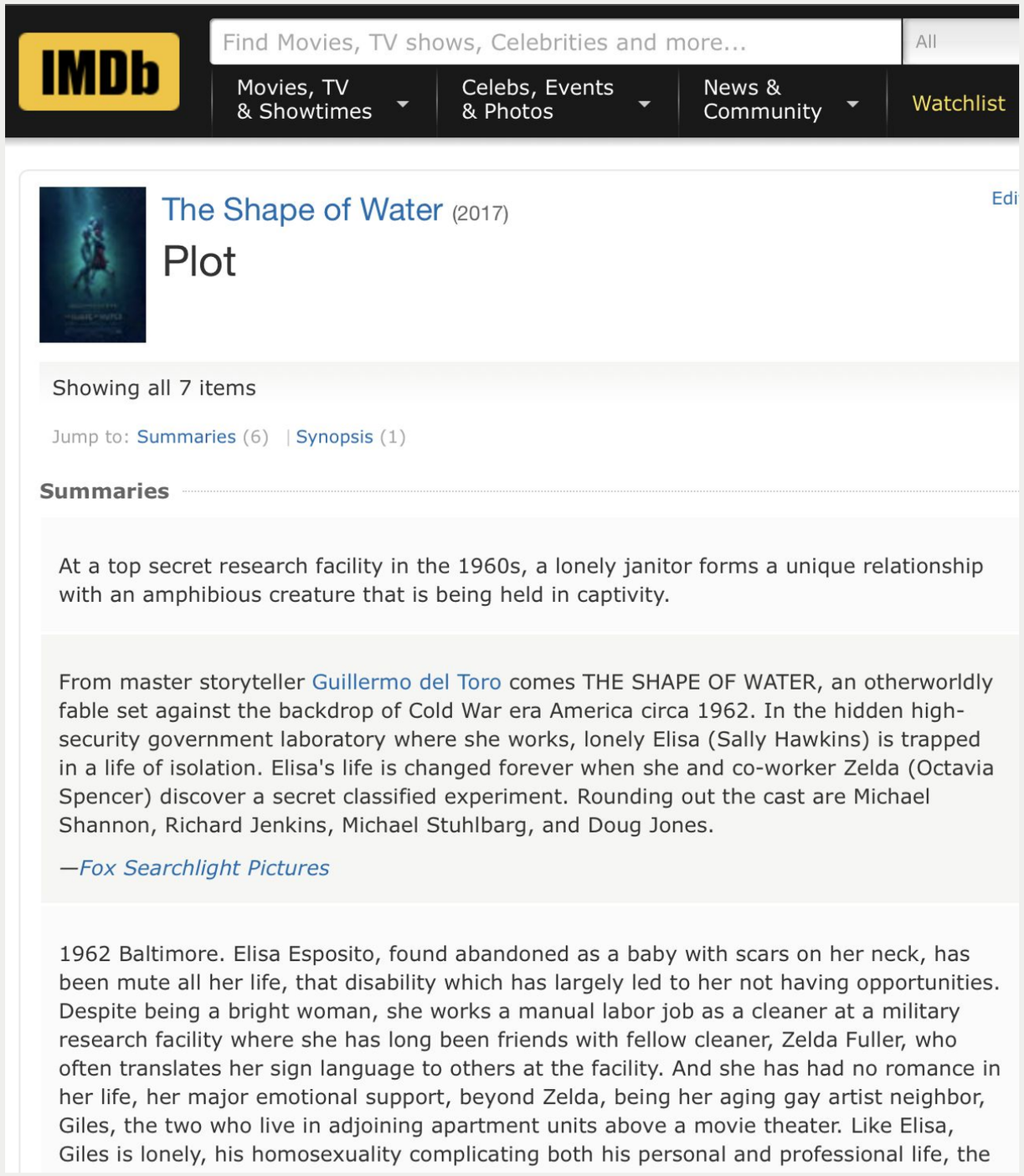
```
In [4]: import json
        list_b=[]
        for n in range(0,15):
            list_b.append(json.loads(list_a[n]))
        #list_b[0]為dict_2004;list_b[1]為dict_2005...list_b[14]為dict_2018
```



2.抓取影片資料

從IMDb爬下影片的關鍵字與劇情

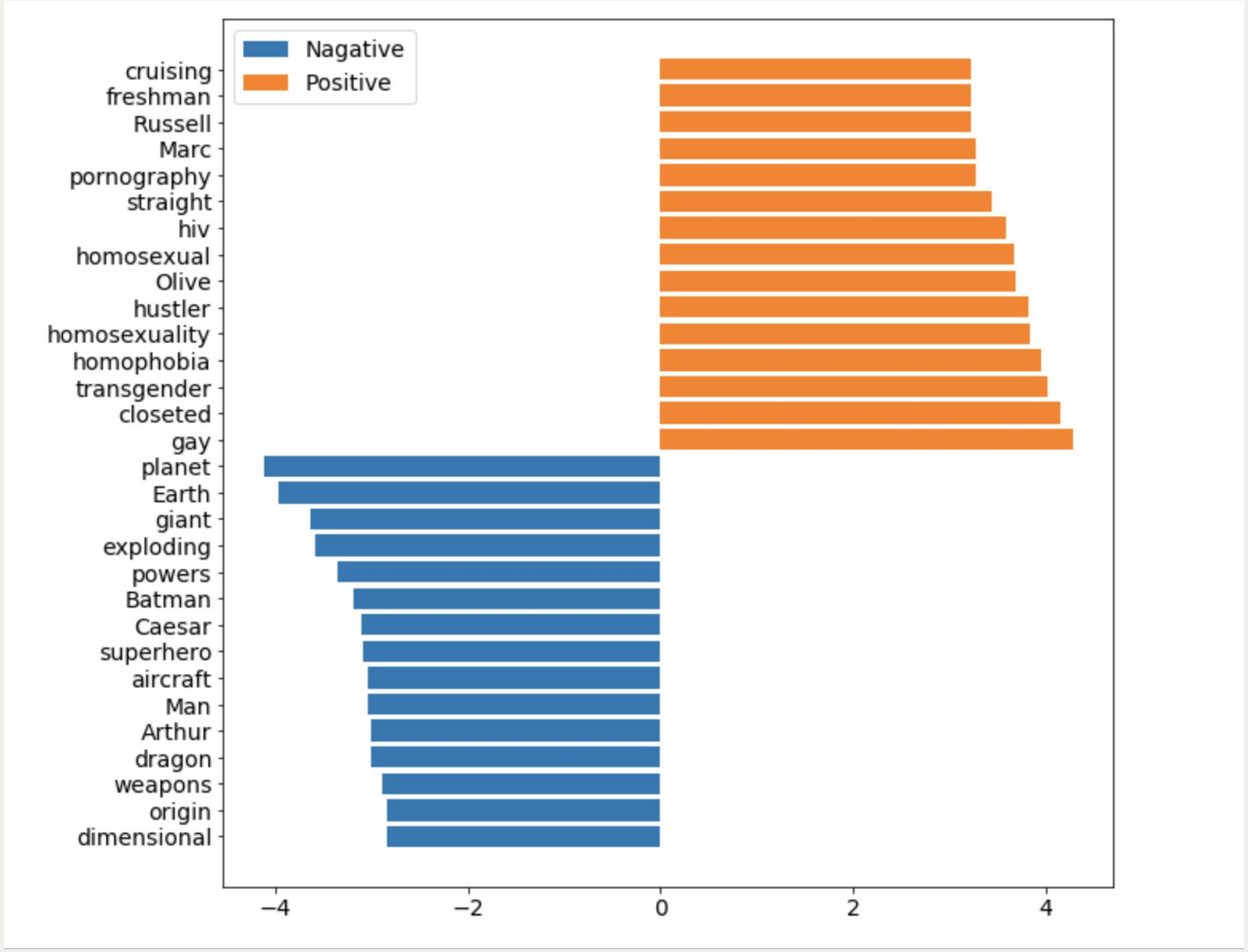




3.判斷影片類型

Training & Predict

log odd ratio、word embedding、LogisticRegression



結果

predict

```
with open('data/bestpicture_plotsummary.json') as f:
    data=json.load(f)
with open('data/bestpicture_keywords.json') as f:
    data2=json.load(f)
movielist=[]
x=[]
y=[]
result=[]
result_proba=[]
for movie in list(data.keys()):
    movielist.append(movie)
    for b in data[movie]:
        x.extend(word_tokenize(b))
    for b in data2[movie]:
        x.extend(word_tokenize(b))
    y.append(we_represent(x))
    x=[]

result=model.predict(y)
print(len(result))
result_counter=Counter()
for i in result:
    result_counter[i]+=1
print(result_counter)
```

```
111
Counter({0: 111})
```

A.電影分析：

抓取
片單

→

判斷
類型

從IMDb、boxofficereport、Wiki 抓取

1. Oscar：獎項入圍提名影片（最佳影片、原創劇本、改編劇本、長紀錄片)
2. LGBT-related films
3. Racism-related related films

A.電影分析：

抓取
片單

→

判斷
類型

從IMDb、boxofficereport、Wiki 抓取

1. Oscar：獎項入圍提名影片（最佳影片、原創劇本、改編劇本、長紀錄片）
2. LGBT-related films
3. Racism-related related films

判斷奧斯卡入圍影片是否在議題名單內

判斷類型

判斷奧斯卡入圍的電影是否在LGBT/Racism-related relate films

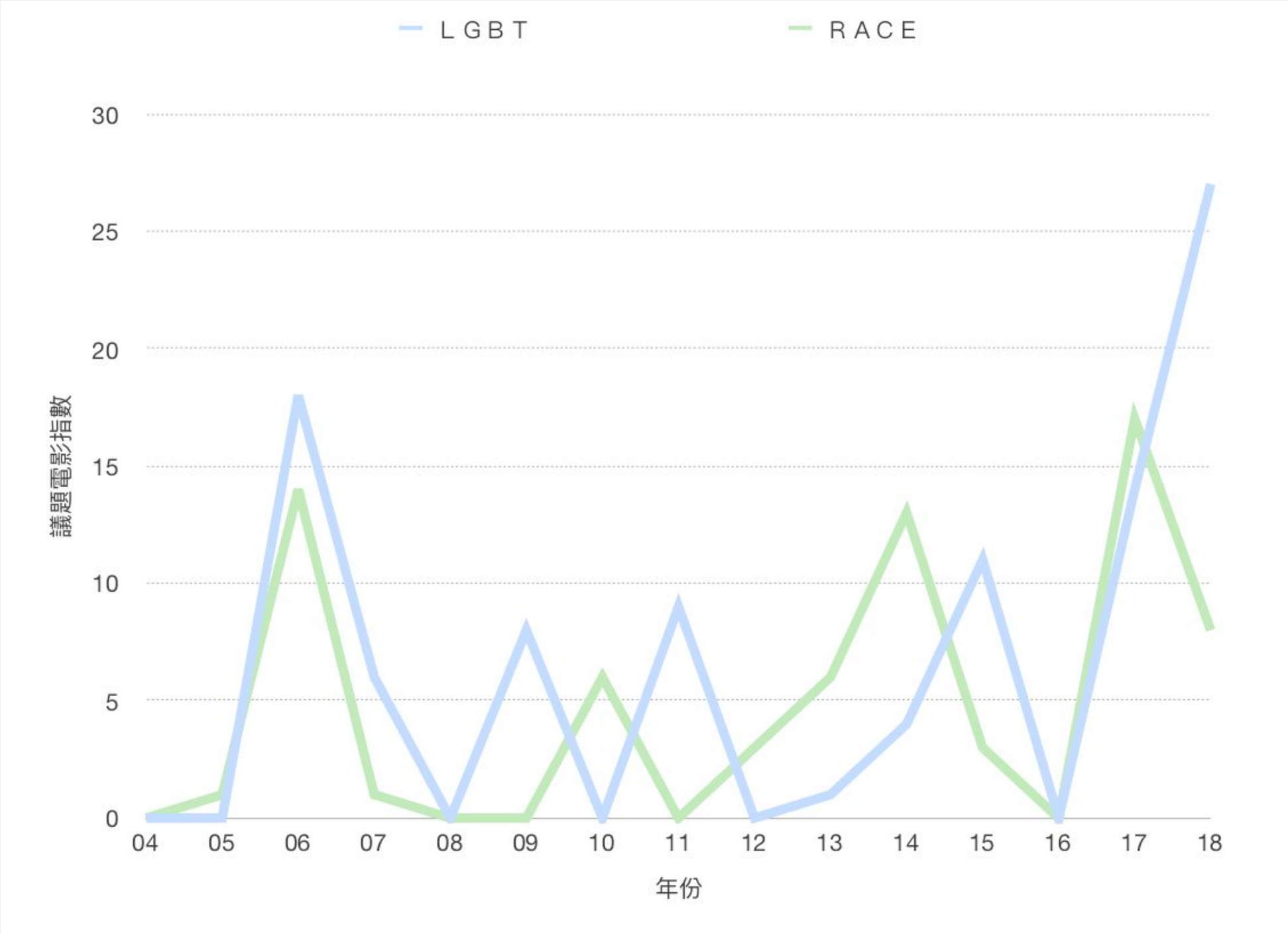
bestpicture_movielist = [[2004, ["The Lord of the Rings: The Return of the King", "Lost in Translation", …, "Mystic River", "Seabiscuit"], 2018,[...]]

LGBT-related_movielist= [[2004, ["The 24th Day", "Alexander", "Billy's Dad Is a Fudge-Packer", ..."True Love", "White Chicks", "Wild Things 2"],... 2018,[...]]

bestpicture
2006 : Brokeback Mountain
2006 : Capote
2007 : Little Miss Sunshine
2009 : Milk
2011 : Black Swan
2011 : The Kids Are All Right
2014 : Dallas Buyers Club
2015 : The Imitation Game
2017 : Moonlight
2018 : The Shape of Water
2018 : Call Me by Your Name
2018 : Lady Bird
[(2006, 2), (2007, 1), (2009, 1), (2011, 2), (2014, 1), (2015, 1), (2017, 1), (2018, 3)]
---
documentaryfeature
2013 : How to Survive a Plague
[(2013, 1)]

		權重
最佳影片	得獎	5
	入圍	3
最佳導演	得獎	4
	入圍	2
最佳原創劇本	得獎	2
	入圍	1
最佳改編劇本	得獎	2
	入圍	1
最佳紀錄長片	得獎	2
	入圍	1

資料結果


LGBT
RACE
議題電影指數
30
25
20
15
10
5
0
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
年份

B. 新聞分析

抓取新聞 → 文章分析 → 趨勢分析

用selenium,request從New York Times上抓取:
1.LGBT
2.Race
兩大議題在2004-2018年間所有的文章內容並存取為pickle

抓取新聞

```
link = []
url = 'https://www.nytimes.com/search?endDate={}1231&query=same%20sex%20marriage%2C%20gay%20right%2C%20ho
driver = webdriver.Chrome('/Users/Kai/Downloads/chromedriver')
driver.get(url)
soup = bs(driver.page_source, 'lxml')
while len(soup.select('.css-vsuiox'))>0:

    try:
        driver.find_element_by_xpath('//*[@id="site-content"]/div/div/div[2]/div[2]/div/button').click()
        time.sleep(1.7)

    except:
        time.sleep(1)
        for i in bs(driver.page_source).select('.css-138wel4 a'):
            link.append(prefix + i['href'])

        all_link.append(link)
        time.sleep(1)
        driver.close()
        break
```

B:新聞分析

抓取新聞 → 文章分析 → 趨勢分析

用selenium,request從New York Times上抓取:
1.LGBT
2.Race
兩大議題在2004-2018年間所有的文章內容並存取為pickle

1.將文章內容透過lemmatize以及詞性標記進行過濾並存至list中
2.訓練word2vec模型

訓練模型

原始
文章

```
ok  4200
ok  4201
ok  4202
ok  4203
ok  4204
ok  4205
```

文字
處理後

```
In [4]: len(clean_sent)
Out[4]: 121510
```

訓練
模型

```
In [13]: len(model.wv.vocab)
Out[13]: 26611
```

B:新聞分析

抓取新聞 → 文章分析 → 趨勢分析

用selenium,request從New York Times上抓取:
1.LGBT
2.Race
兩大議題在2004-2018年間所有的文章內容並存取為pickle

1.將文章內容透過lemmatize以及詞性標記進行過濾並存至list中
2.訓練word2vec模型

將過濾後的文章都丟入並比對與關鍵字之間的相關程度藉此計算出每篇文章的分數

趨勢分析

LGBT score trend:

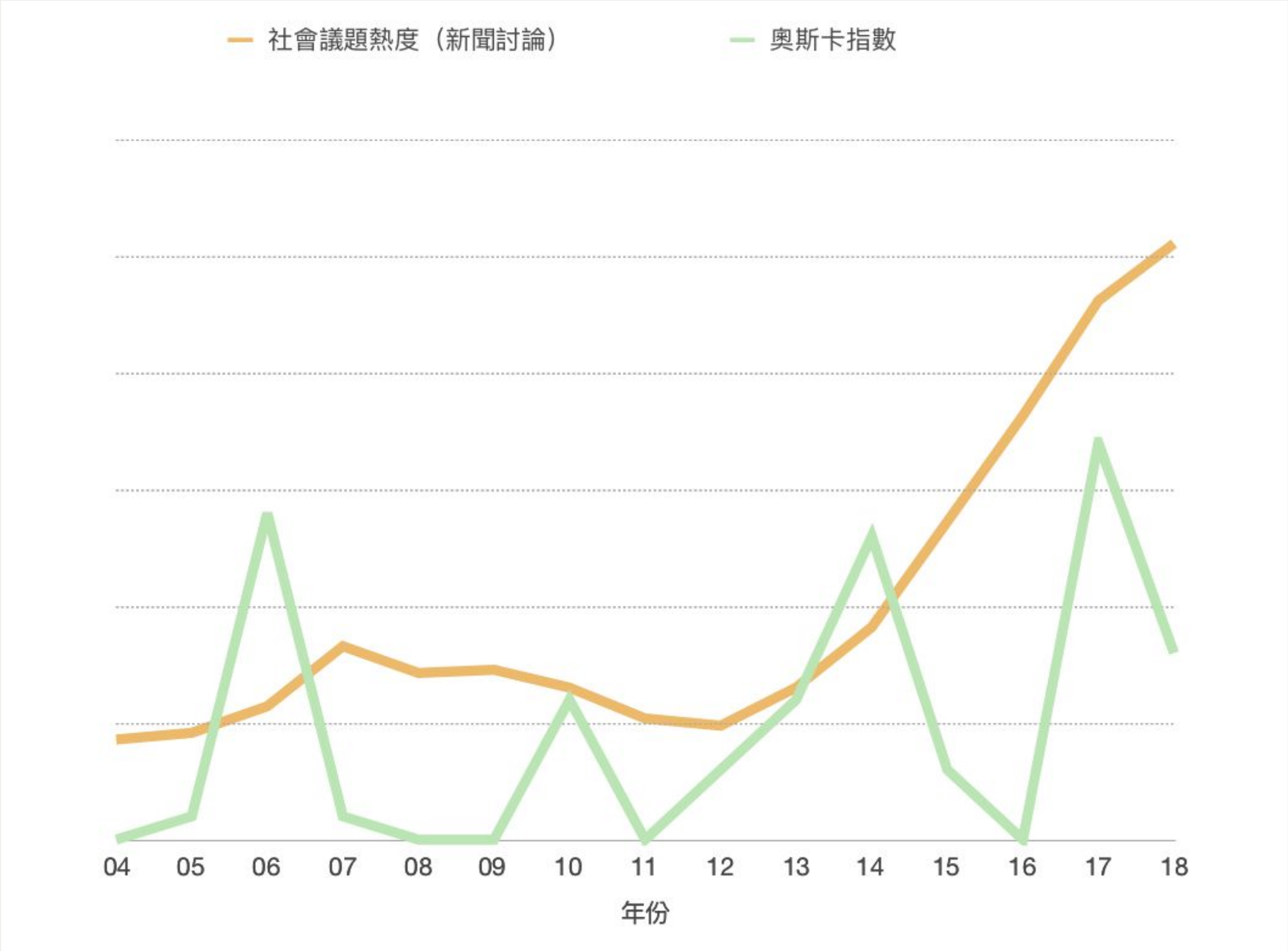
```
Out[58]: [0.0032003803880689 82,
          0.0029649266616669865,
          0.0027516169295359128,
          0.00412235853377 7343,
          0.0031613910120453,
          0.0029868513608570964,
          0.0022633815957704135,
          0.0019872309838907445,
          0.0016006616067974764,
          0.0019358804986200312,
          0.0016041966721701151,
          0.0014342737297882416,
          0.0020774516592979 28,
          0.0013289597198637 29,
```

2004

2018

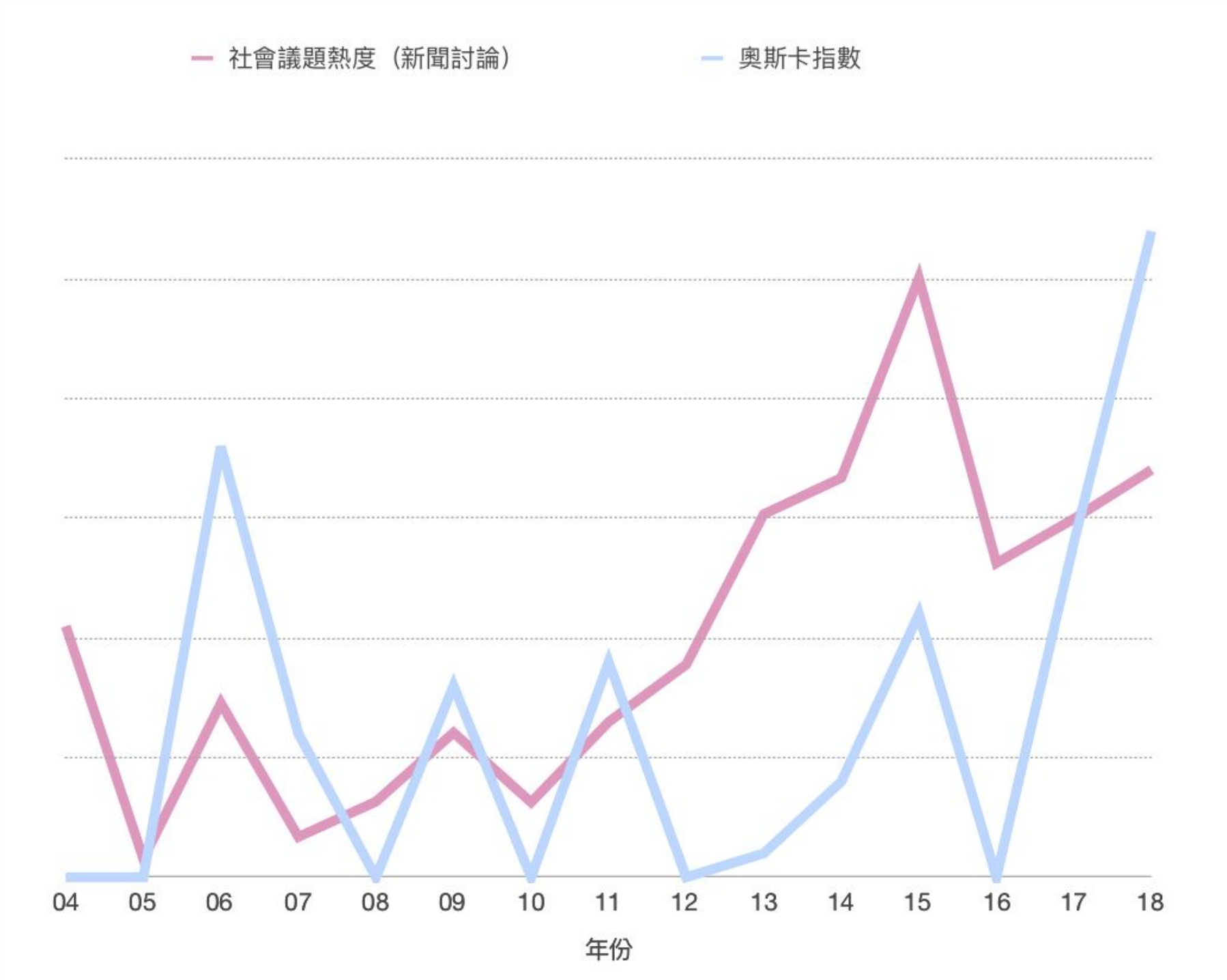
議題新聞佔比

結論：



RACE

結論：



LGBT

分工表

- 電影片單爬蟲:王聖夫、柯亮宇
- 電影分析:王聖夫
- 電影簡報製作:王聖夫、柯亮宇
- 新聞爬蟲:劉哲愷
- 新聞分析:劉哲愷
- 新聞簡報製作:柯亮宇、阮彥璋
- 口頭報告:柯亮宇、阮彥璋