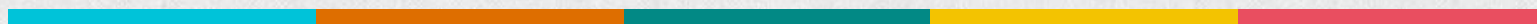




文字探勘X電影評論



05170107 巨資三A 侯聖恩

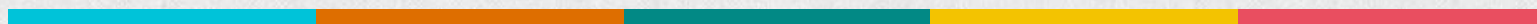




以《黑豹》為例



05170107 巨資三A 侯聖恩





目錄

1

專案簡介
動機&目的

2

資料集介紹
Yahoo電影評論&PTT movie版

3

實作方法與流程
模型&演算法

4

分析結果&結論
圖表&數據

An abstract graphic featuring a central cluster of overlapping circles in various colors including light blue, yellow, orange, pink, and teal. A large white number '1' is positioned in the center of these circles. Surrounding this central cluster are several smaller, isolated dots in the same color palette, scattered across the light gray background.

1

專案簡介

A horizontal bar at the bottom of the slide, composed of five distinct colored segments: light blue, orange, teal, yellow, and pink, arranged from left to right.

專案簡介

1 動機



1 動機



1 目的



1 目的

1

好評關鍵字

從Yahoo電影討論區抓取四星五星好評並找出評論中較常出現的關鍵字。

2

情感分析

用SnowNLP分析PTT上評論留言的情感平均分。

A decorative graphic featuring several overlapping circles in shades of blue, yellow, orange, and pink. A large white number '2' is centered within the overlapping circles. Surrounding the circles are numerous small dots in various colors (blue, yellow, orange, pink, green) scattered across the light gray background.

2

資料集介紹

A horizontal bar composed of five segments in different colors: blue, orange, green, yellow, and red.

2 資料集



Yahoo 電影討論區

一部電影的討論大約30-50頁
每頁10則留言

評分1-5顆星

一二星差評、四五星的好評

support較大的關鍵字

以一頁為一文本

PTT movie版

大約30頁，每頁20則討論，
每則可能有10到上百則留言不等

資料量較大

以周為區分，分上映後七週

好/負雷、新聞

只取留言，不取內文

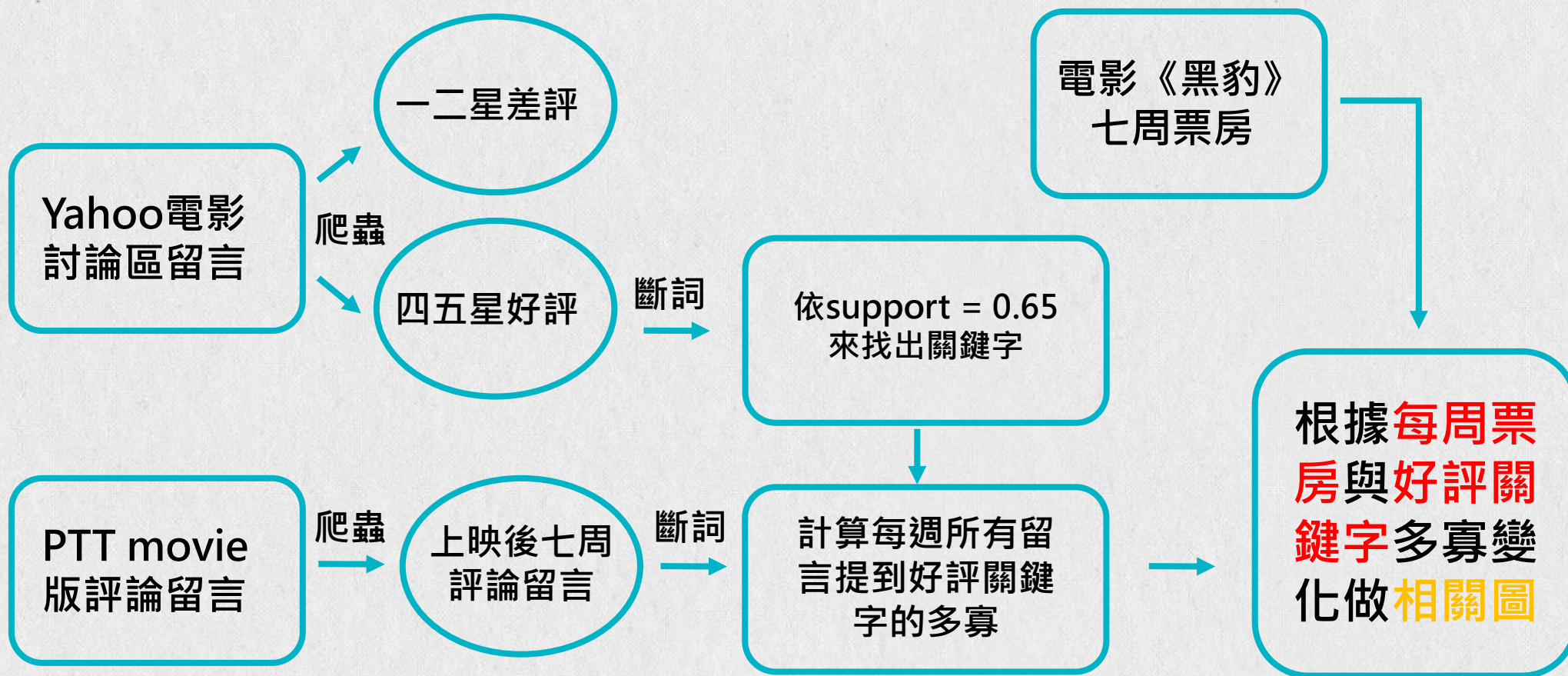
A decorative graphic featuring several overlapping circles in shades of blue, yellow, orange, and pink. A large white number '3' is centered within the overlapping circles. Surrounding the circles are numerous small dots in various colors (blue, yellow, orange, pink, green) scattered across the light gray background.

3

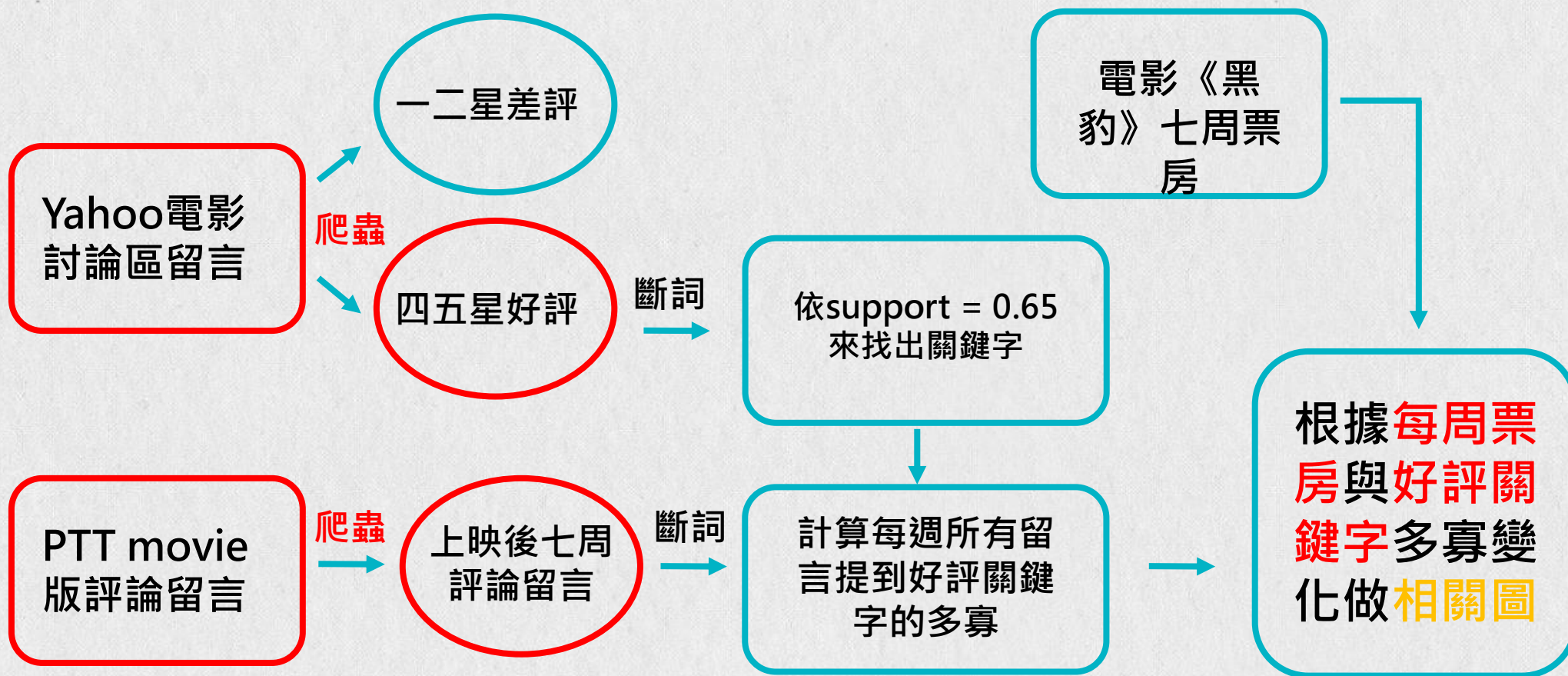
實作方法與流程

A horizontal bar composed of five colored segments: blue, orange, green, yellow, and red, arranged from left to right.

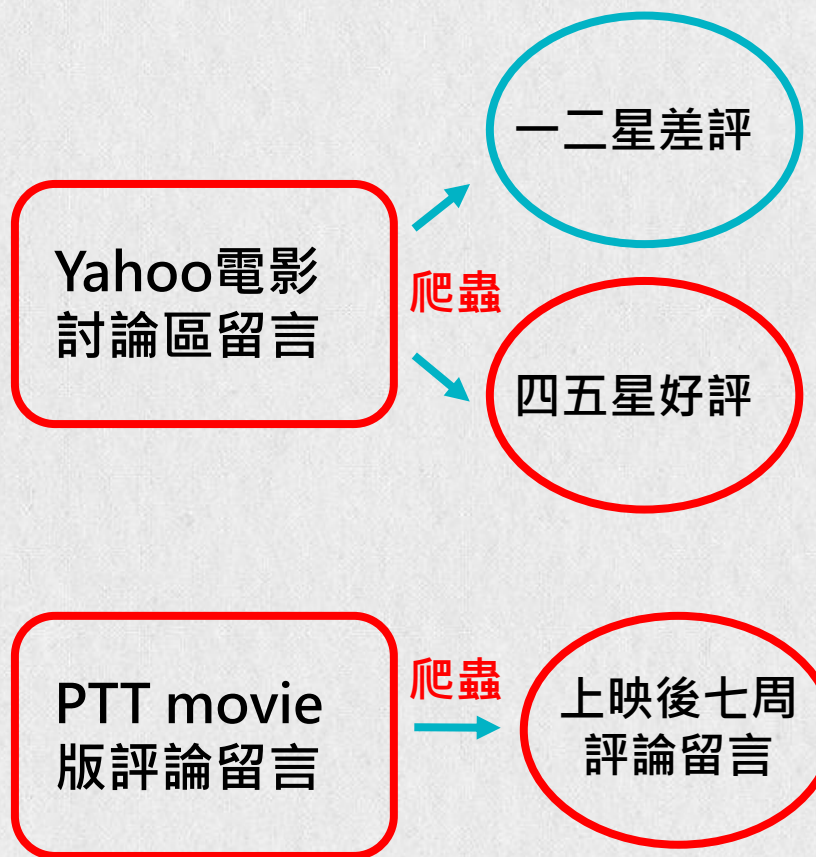
3 好評關鍵字實作流程



3 好評關鍵字實作流程



3 好評關鍵字實作流程



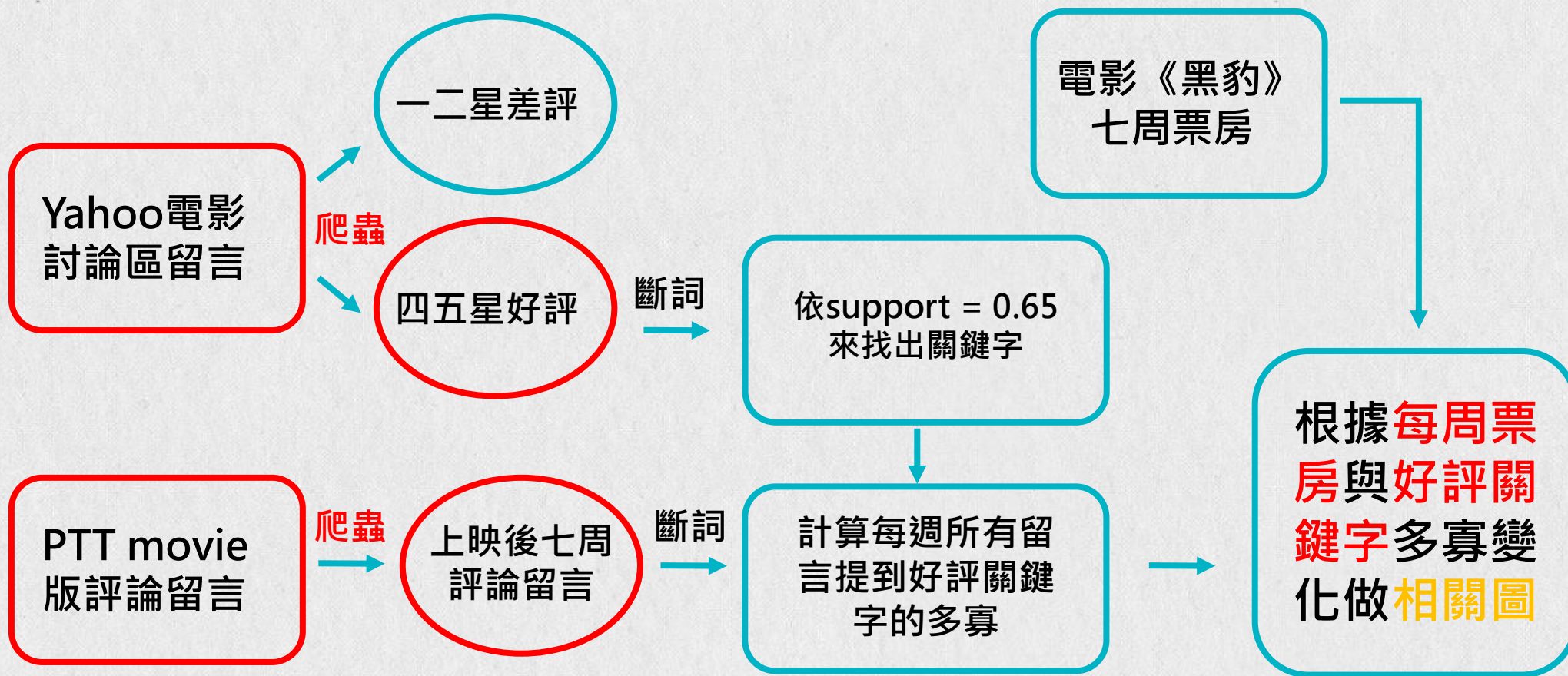
爬下四星五星好評資料

```
In [4]: url_set1 = ['https://movies.yahoo.com.tw/movieinfo_review.html?id=6953?sort=ra
```

```
In [5]: all_text = []
for i in url_set1:
    response = requests.get(i)
    soup = BeautifulSoup(response.text, 'lxml')
    articles = soup.find_all('div', 'usercom_inner_c')
    text = []
    for article in articles:
        messages = article.find('span', None).getText().replace(':', '').strip()
        text.append(messages)
    all_text.append(text)
print(all_text[0])
```

['沒有冷場，超好看的啊，評價怎麼只有3.9\r\n個人覺得第一集總是要鋪陳和講述起
棒，有別於以往Marvel風格，值得一看', '覺得很好看 不知道為什麼有人只給1顆星真
麼才是爽片，趕快去看吧！北美票房已經破史上紀錄了57億....', '去看就對了啦，廢
有台灣人給3.9分真的笑掉人家大牙' '節奏緊湊 動作頻繁 台式對白 會心一笑（可以

3 好評關鍵字實作流程




```
] cut_text = [' '.join(jieba.cut(w)) for w in t] for t in all_text]
document = [[s.split(' ') for s in i] for i in cut_text]
clean_documents = [[b for c in d for b in c] for d in document]
print(clean_documents[0])
```

Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\coco8\AppData\Local\Temp\jieba
Loading model cost 0.867 seconds.
Prefix dict has been built successfully.

['沒有', '冷場', ' ', ' ', '超', '好看', '的', '啊', ' ', ' ', '評價',
'鋪', '陳', '和', '講述', '起源', ' ', ' ', '不能', '期待', '絕無冷

Yahoo電影
討論區留言

爬蟲

四五星好評

斷詞

依support = 0.65
來找出關鍵字

PTT movie
版評論留言

爬蟲

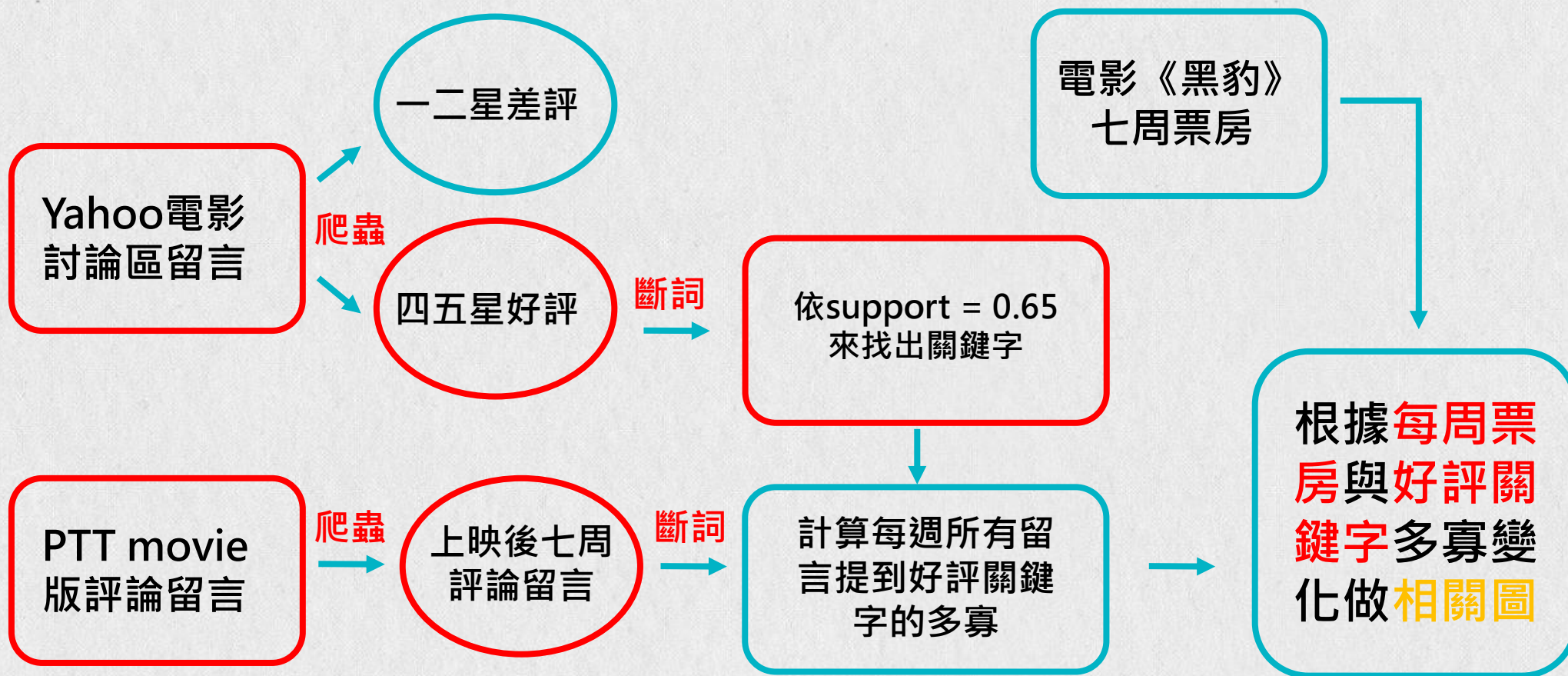
上映後七周
評論留言

斷詞

計算每週所有留
言提到好評關鍵
字的多寡

根據每周票
房與好評關
鍵字多寡變
化做相關圖

3 好評關鍵字實作流程



3

In [10]:

```
te = TransactionEncoder()  
te_ary = te.fit(clean_documents).transform(clean_documents)  
df = pd.DataFrame(te_ary, columns=te.columns_)  
frequent_itemsets = apriori(df, min_support=0.65, use_colnames=True)  
print(frequent_itemsets)
```

	support	itemsets
0	0.888889	(劇情)
1	0.666667	(喜歡)
2	1.000000	(好看)
3	0.833333	(漫威)
4	0.666667	(特效)
5	0.666667	(英雄)
6	0.777778	(覺得)
7	1.000000	(電影)
8	0.833333	(非洲)
9	0.833333	(風格)
10	0.722222	(黑豹)

版評論留言

AT 嘴 田 白

詞

新詞

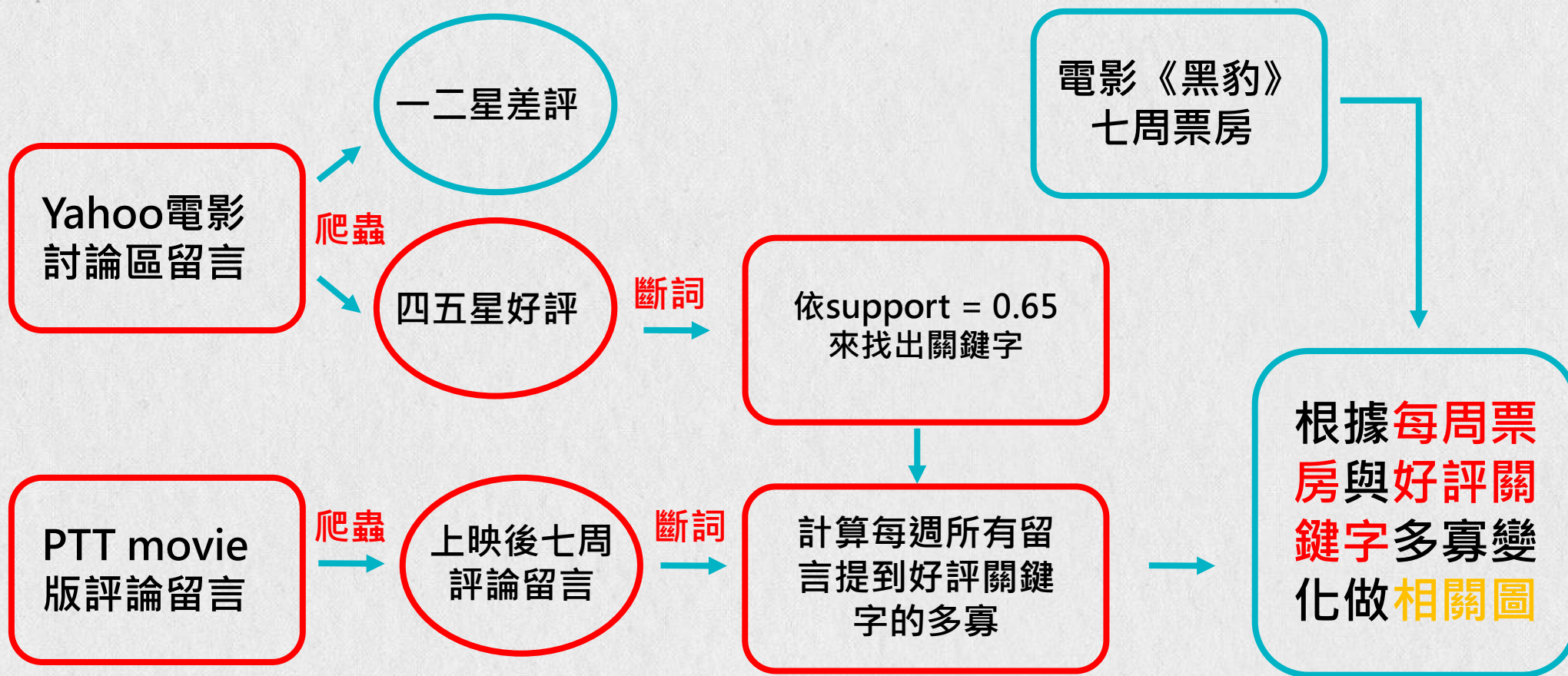
依support = 0.65
來找出關鍵字

計算每週所有留
言提到好評關鍵
字的多寡

電影《黑豹》
七周票房

根據每周票
房與好評關
鍵字多寡變
化做相關圖

3 好評關鍵字實作流程



3 好評

Yahoo電影
討論區留言

爬蟲

一二星差評

四五星好評

PTT movie
版評論留言

爬蟲

上映後七周
評論留言

斷詞

計算每週所有留
言提到好評關鍵
字的多寡

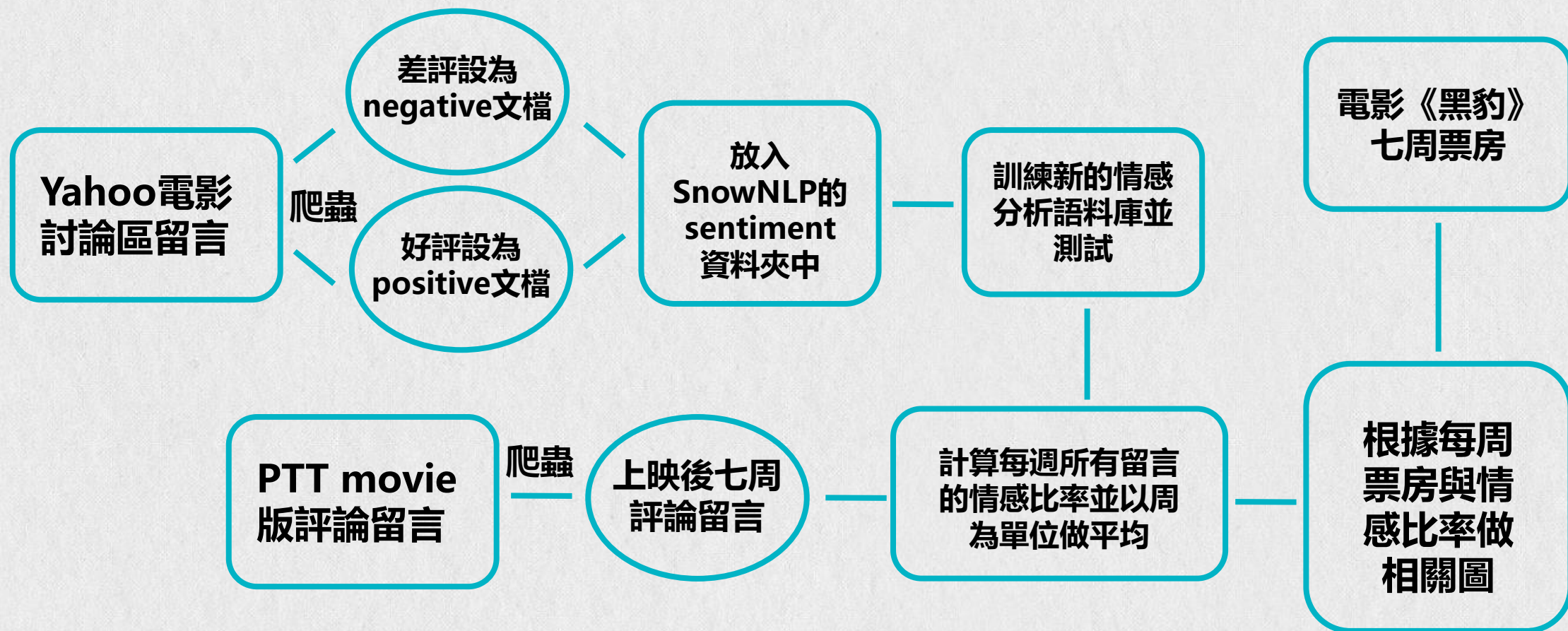
房與灯評關
鍵字多寡變
化做相關圖

```
In [53]: ptt_cut_text1 = [[' '.join(jieba.cut(w)) for w in t] for t in all_text_week1]
ptt_document1 = [[s.split(' ') for s in i] for i in ptt_cut_text1]
ptt_clean_documents1 = [[b for c in d for b in c] for d in ptt_document1]
new_ptt_clean_documents1 = []
for a in ptt_clean_documents1:
    for b in a:
        new_ptt_clean_documents1.append(b)

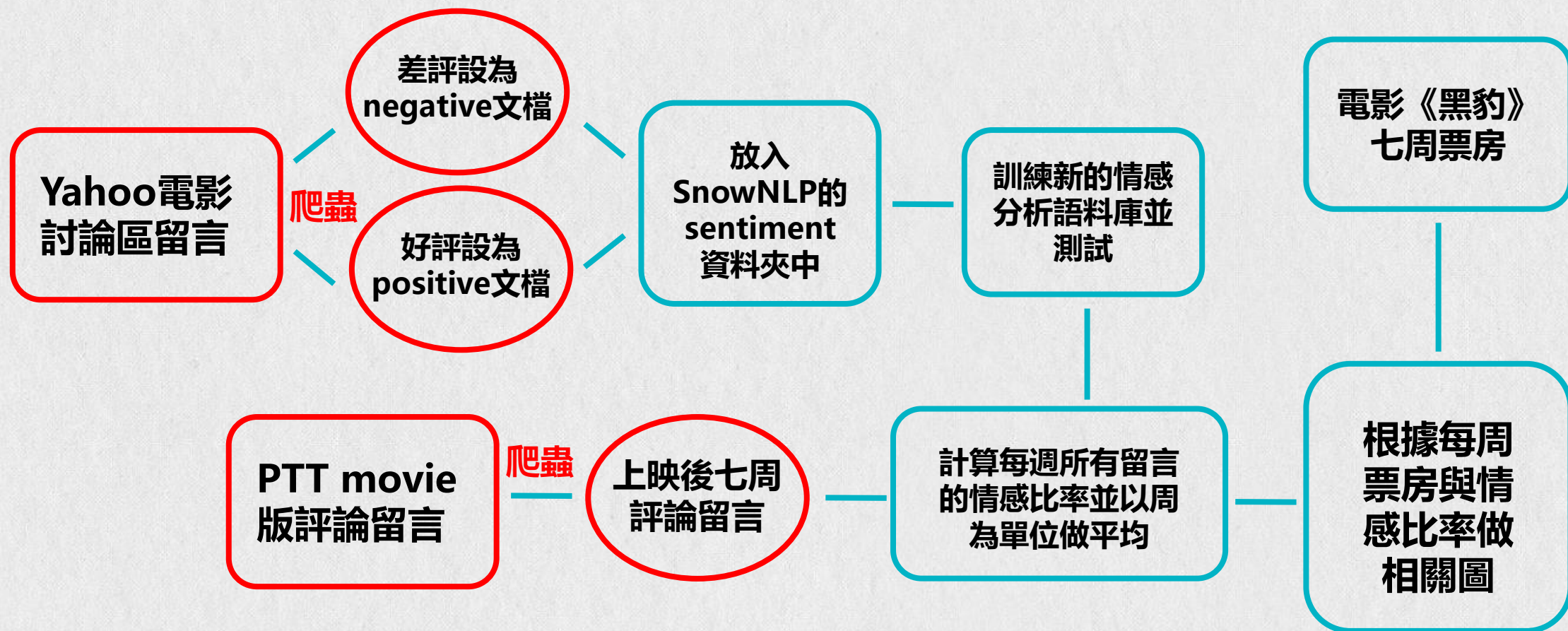
from collections import Counter
c_week1 = Counter(new_ptt_clean_documents1)
for a in c_week1.most_common(200):
    word, freq = a
    print(word, freq)

3108
的 1379
， 1045
是 612
了 410
黑豹 359
就 345
我 341
有 340
也 321
很 291
看 289
? 268
```


3 情感分析實作流程



3 情感分析實作流程



3

情

Yahoo電影
討論區留言

差評設為
negative文檔

好評設為
positive文檔

PTT movie
版評論留言

爬蟲

爬蟲

negative - Notepad

File Edit Format View Help

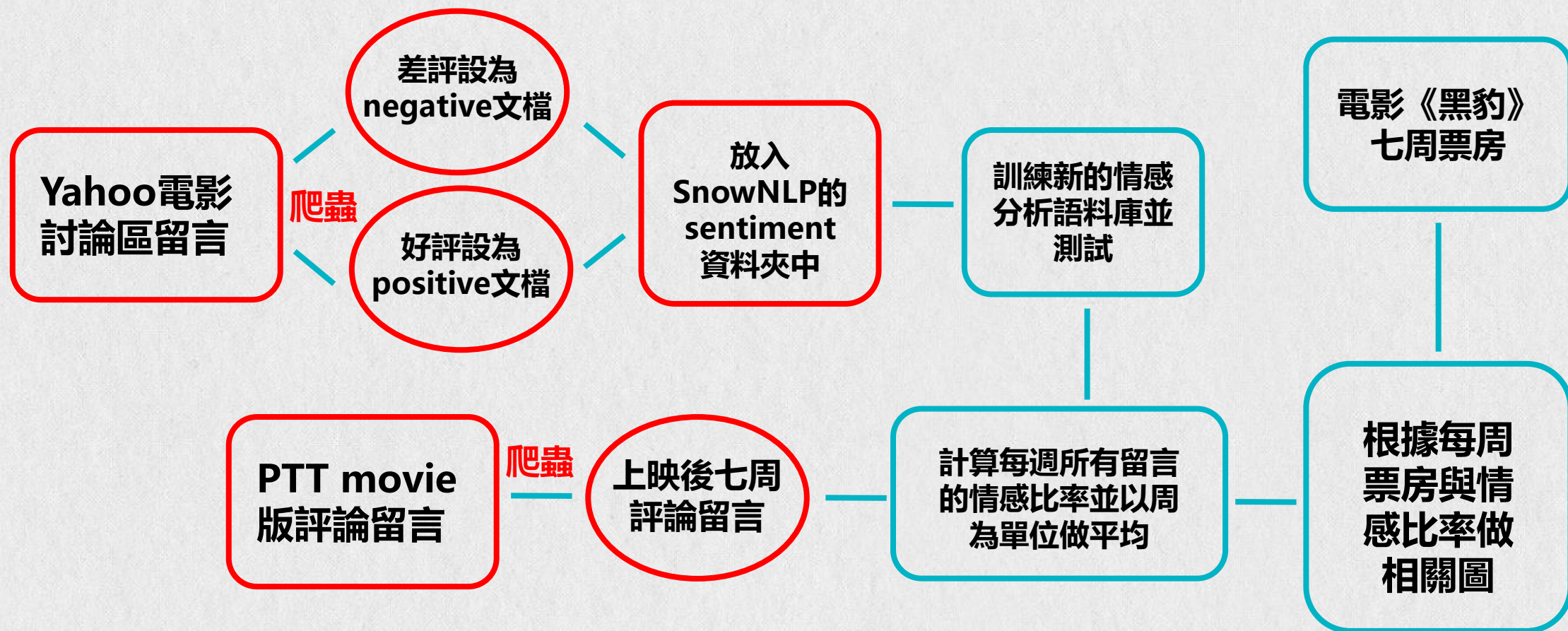
[['我是漫威迷漫畫跟電影都有在看我必須說真的蠻失望的特
的很帥沒話說但又沒把黑豹能力明顯做出來黑豹近戰能力明明
而且黑豹的個性是他是一個善良過頭的人影也沒有顯現出來以
說又不夠爽劇情滿多漏洞又頻繁多餘一些種族政治議題不難可

positive - Notepad

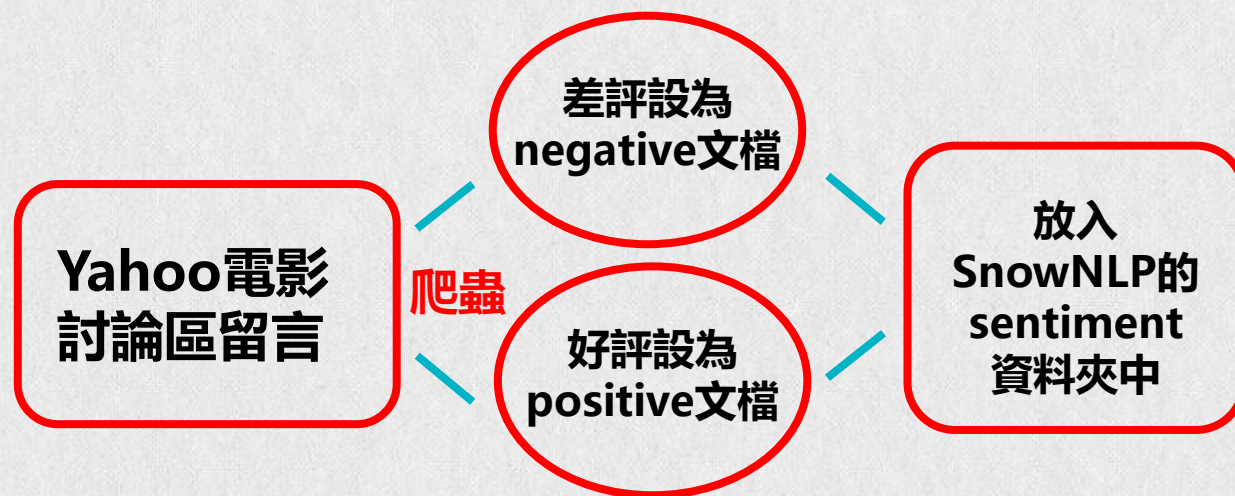
File Edit Format View Help

並?有拍出令人期
況是第一位黑人英雄電影
在有點且在王為爭總是要鋪陳和講述起源,不能期待絕無冷場,但我覺得拍
有點不之所綽屢好,很有非洲風情和音樂也很棒,有別於以往Marvel風格
讓人有所說服等等得一看', '覺得很好看 不知道為什麼有人只給1顆星真的
位黑人英雄電影', '我覺得很好看,值得去看,後面有兩個彩蛋', '這不
說真的不能完全片什麼才是爽片,趕快去看吧!北美票房已經破史上紀錄
角色的身世角度來意....', '去看就對了啦,廢話一大堆,不然過年要幹嘛
來的漫威電影銜接'還好沒看影評分數,國外一致好評,只有台灣人給3.9分
的有差距,劇情笑掉人家大牙', '節奏緊湊,動作頻繁,台式對白,會心一笑
劇中烏干達的科技可以開音樂嗎?又不是出殯)', '很好看,我以為會是雷片
達炫多了脫掉黑豹完之後不輸給其他英雄電影。', '看完後只能說怎麼可能
事張力不足劇情3.9!!!真的好好看!大家快去看', '雖然沒有以往的mar
，但這部真的很開打鬥的爽度，不過卻能將落後的非洲擁有世界上最高的和
這些時間在做什麼做一對比，真的有種不同的風氣，且片中穿插著許多非洲
這個電影一樣...的配樂，再結合壯麗的風景，真的很有fu，其實把它當作
!2個小時 坐的好，就不會失望啦~~']，['老實跟大家講，此片原先是毫不
趣的，雖然海外評價接近滿分，但是風格總覺得過於虛幻
是我的菜，沒想到最終走出戲院居然是驚訝又滿意！太讓
驚了。首先，漫威集團你們又做到了！真的不該懷疑你們
力，單看預告片或許會像我一樣，覺得過於科幻且不切實
不過實際上他們把風格的拿捏呈現得非常好，老實說還蠻
的，不會過於卡通化，而且景色和特效依舊很精緻，融合

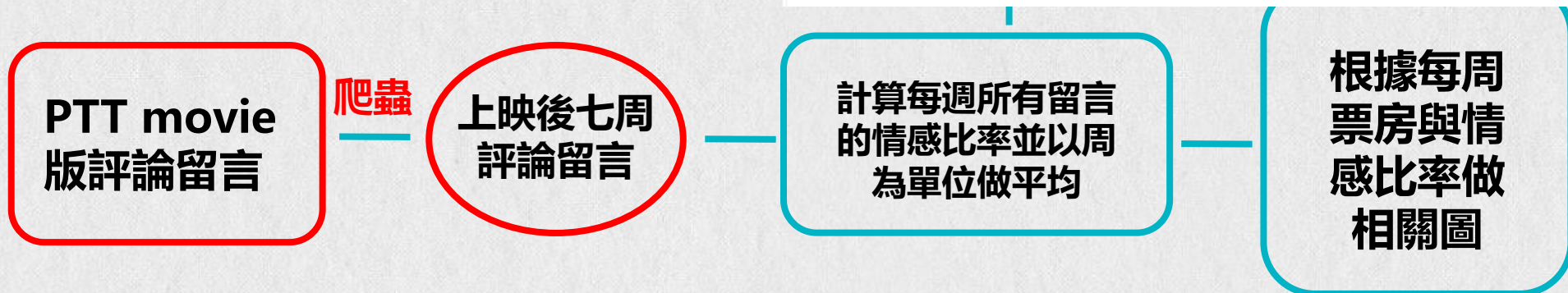
3 情感分析實作流程



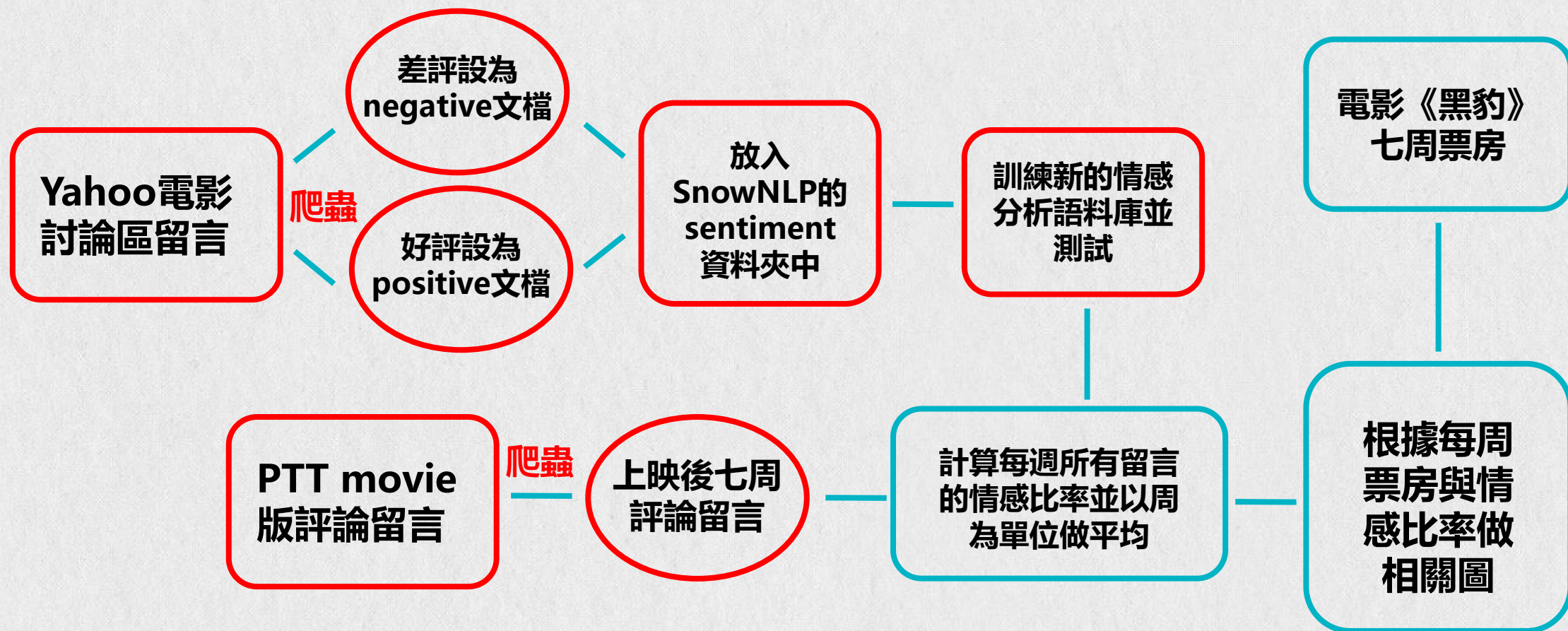
3 情感分析實作流程



Users > coco8 > Anaconda3 > Lib > site-packages > snownlp > sentiment		
Name	Date modified	Type
__pycache__	12/27/2018 10:32 AM	File folder
__init__.py	12/27/2018 10:32 AM	PY File
neg	12/29/2018 4:46 PM	Text Document
negative	12/31/2018 1:36 PM	Text Document
pos	12/27/2018 10:32 AM	Text Document
positive	12/29/2018 4:47 PM	Text Document
sentiment.marshal	12/27/2018 10:32 AM	MARSHAL File
sentiment.marshal.3	12/27/2018 2:04 PM	3 File
sentiment2.marshal.3	12/27/2018 1:52 PM	3 File
sentimentHW.marshal.3	12/30/2018 9:11 PM	3 File

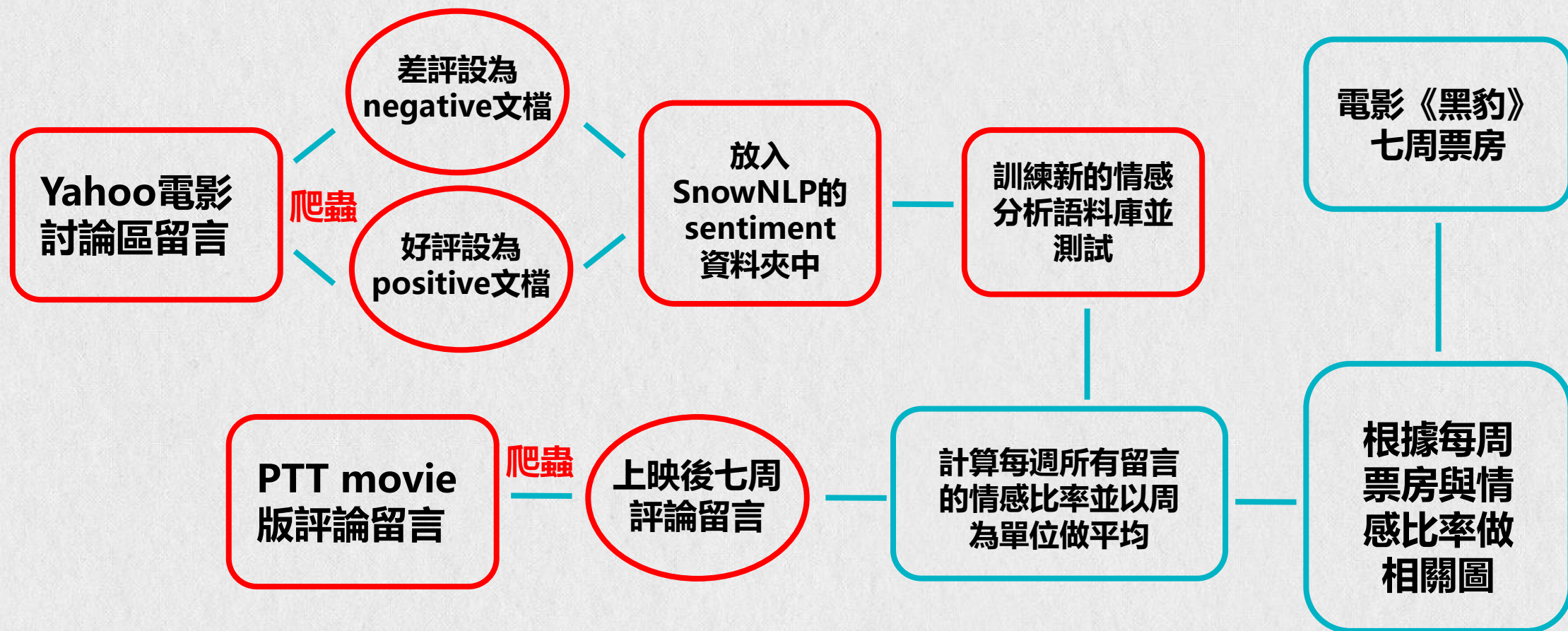


3 情感分析實作流程




```
In [30]: from snowlp import sentiment
```

```
In [31]: sentiment.train("C:/Users/coco8/Anaconda3/Lib/site-packages/snownlp/sentiment/negative.txt",  
                        "C:/Users/coco8/Anaconda3/Lib/site-packages/snownlp/sentiment/positive.txt")
```




```
In [30]: from snownlp import sentiment
```

```
In [31]: sentiment.train("C:/Users/coco8/Anaconda3/Lib/site-packages/snownlp/sentiment/negative.txt",  
                        "C:/Users/coco8/Anaconda3/Lib/site-packages/snownlp/sentiment/positive.txt")
```



train 完 儲存到新的model

```
In [21]: sentiment.save("C:\\Users\\coco8\\Anaconda3\\lib\\site-packages\\snownlp\\sentiment\\sentimentHW.marshall")
```




```
In [30]: from snowlp import sentiment
```

```
In [31]: sentiment.train("C:/Users/coco8/Anaconda3/Lib/site-packages/snowlp/sentiment/negative.txt",  
                        "C:/Users/coco8/Anaconda3/Lib/site-packages/snowlp/sentiment/positive.txt")
```



train 完 儲存到新的model

```
In [21]: sentiment.save("C:\\Users\\coco8\\Anaconda3\\lib\\site-packages\\snowlp\\sentiment\\sentimentHW.marshall")
```

設定等等要使用的情感分析model => 指定路徑

```
In [22]: sentiment.data_path = "C:\\Users\\coco8\\Anaconda3\\lib\\site-packages\\snowlp\\sentiment\\sentimentHW.marshall"
```


3 情感分析實作流程

Yahoo電影
討論區留言

爬蟲

差評
negat

好評
posit

PTT movie
版評論留言

```
In [23]: text1 = "《黑豹》是一部很好看的電影"
```

```
In [24]: t = SnowNLP(text1)  
t.sentiments
```

```
Out[24]: 0.7105212829671641
```

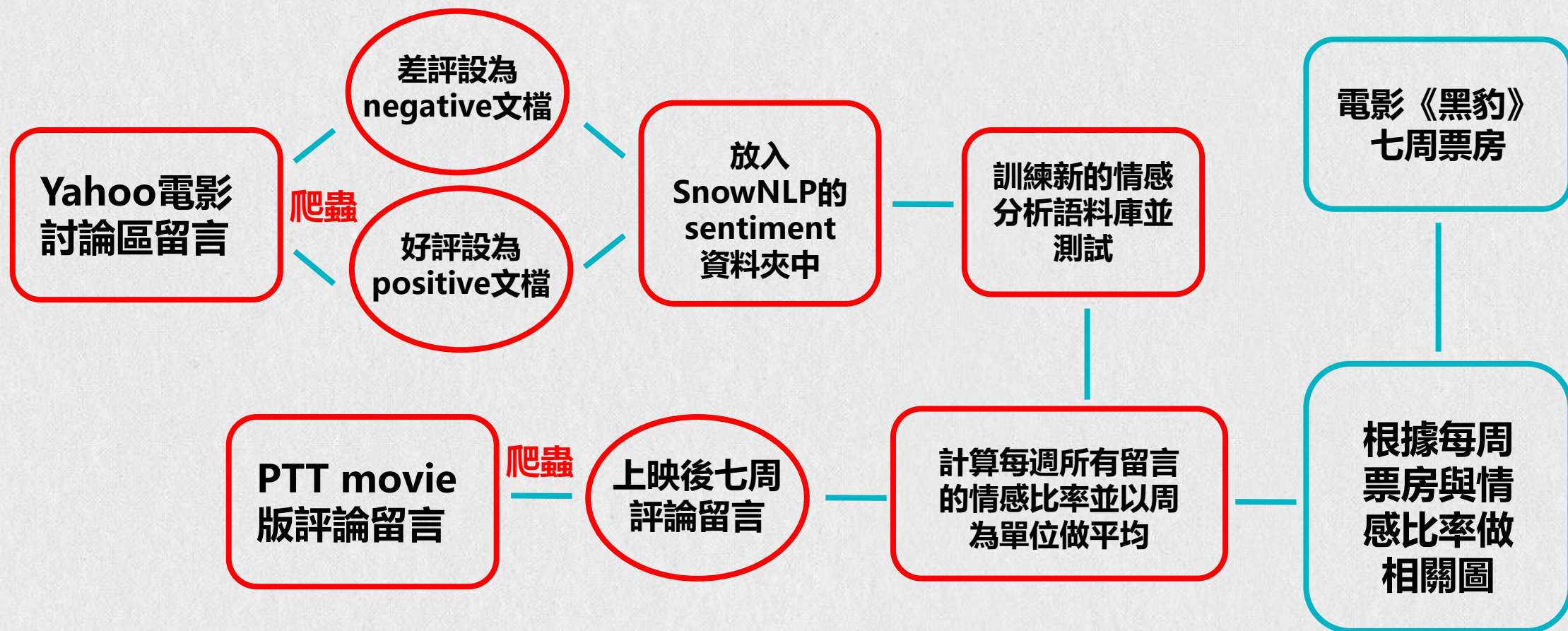
```
In [39]: text2 = '無聊、想睡'  
t2 = SnowNLP(text2)  
t2.sentiments
```

```
Out[39]: 0.046876957845756406
```

《黑豹》
票房

每周
情感
率做
圖

3 情感分析實作流程

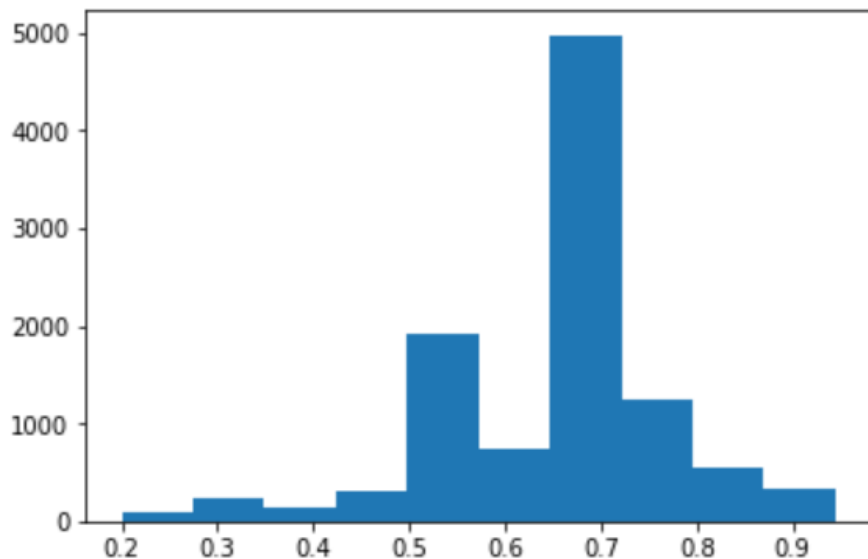


2

情感分析實作流程

```
t2 = []
for w in new_all_text_week2:
    for a in w:
        r = SnowNLP(a)
        t2.append(r.sentiments)
print(t2)
plt.hist(t2)
plt.show()
```

0.5, 0.8999999999999999, 0.6906900775826866, 0.6855670103092784, 0.99999999, 0.727272727272727, 0.3999999999999999, 0.6906900775826866, 0.775826866, 0.6666666666666666, 0.6666666666666666, 0.68556701030927, 0.1818181818, 0.8888888888888888, 0.9375, 0.8888888888888888, 0.83333



訓練新的情感
分析語料庫並
測試

計算每週所有留言
的情感比率並以周
為單位做平均

電影《黑豹》
七周票房

根據每周
票房與情
感比率做
相關圖

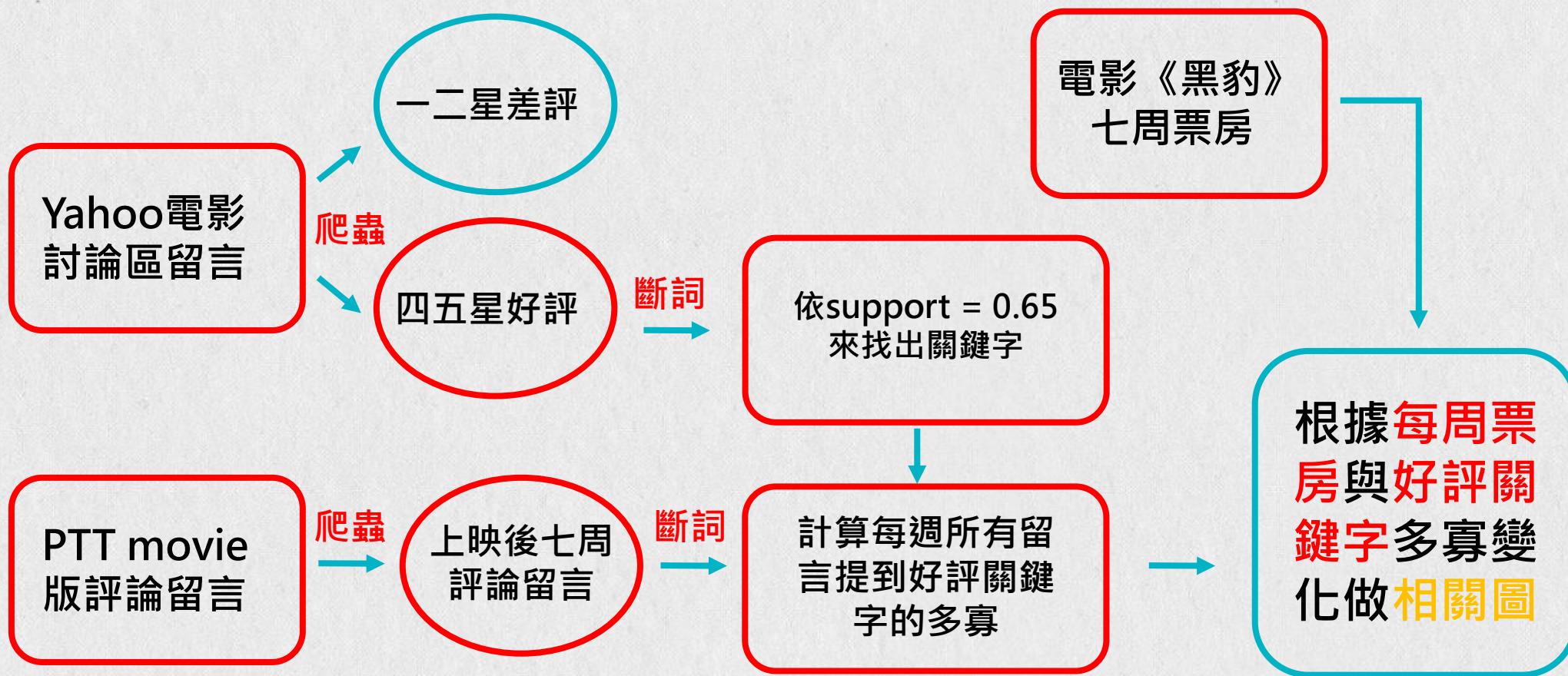
A decorative graphic featuring several overlapping circles in shades of blue, yellow, orange, and pink, with a large white number '4' in the center. Smaller dots in various colors are scattered around the circles.

4

結果分析&結論

A horizontal bar composed of five segments in blue, orange, teal, yellow, and red.

4 好評關鍵字實作流程

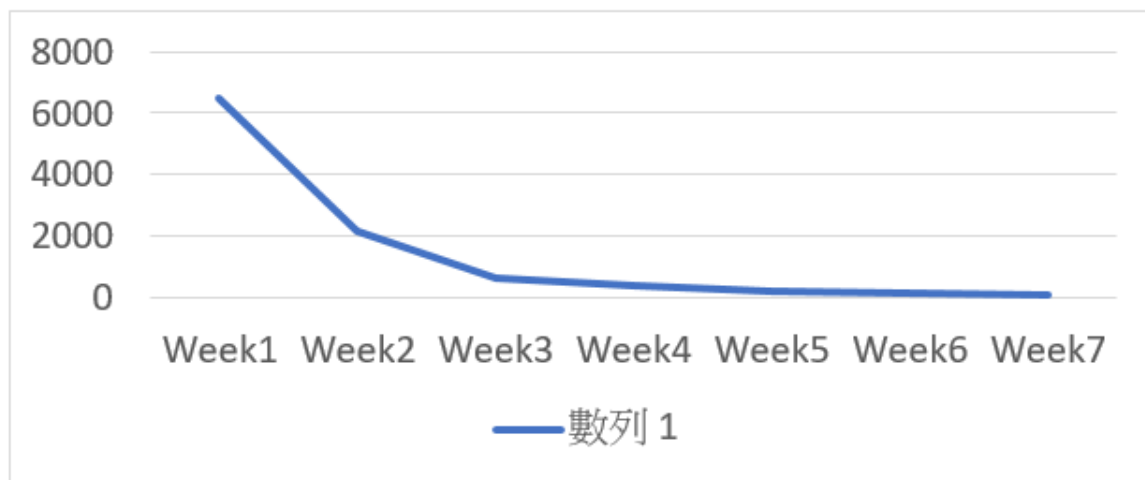


4 好評關鍵字實作流程

➤ 上映七周票房的變化

本專案中為了消除各家電影業者的價格落差與隨著時間漸長上映的電影院數越少這兩項變因，故對於票房做出以下調整：

$$\text{各週票房} = \frac{\text{各週銷售電影票數}}{\text{各週上映電影院數}}$$



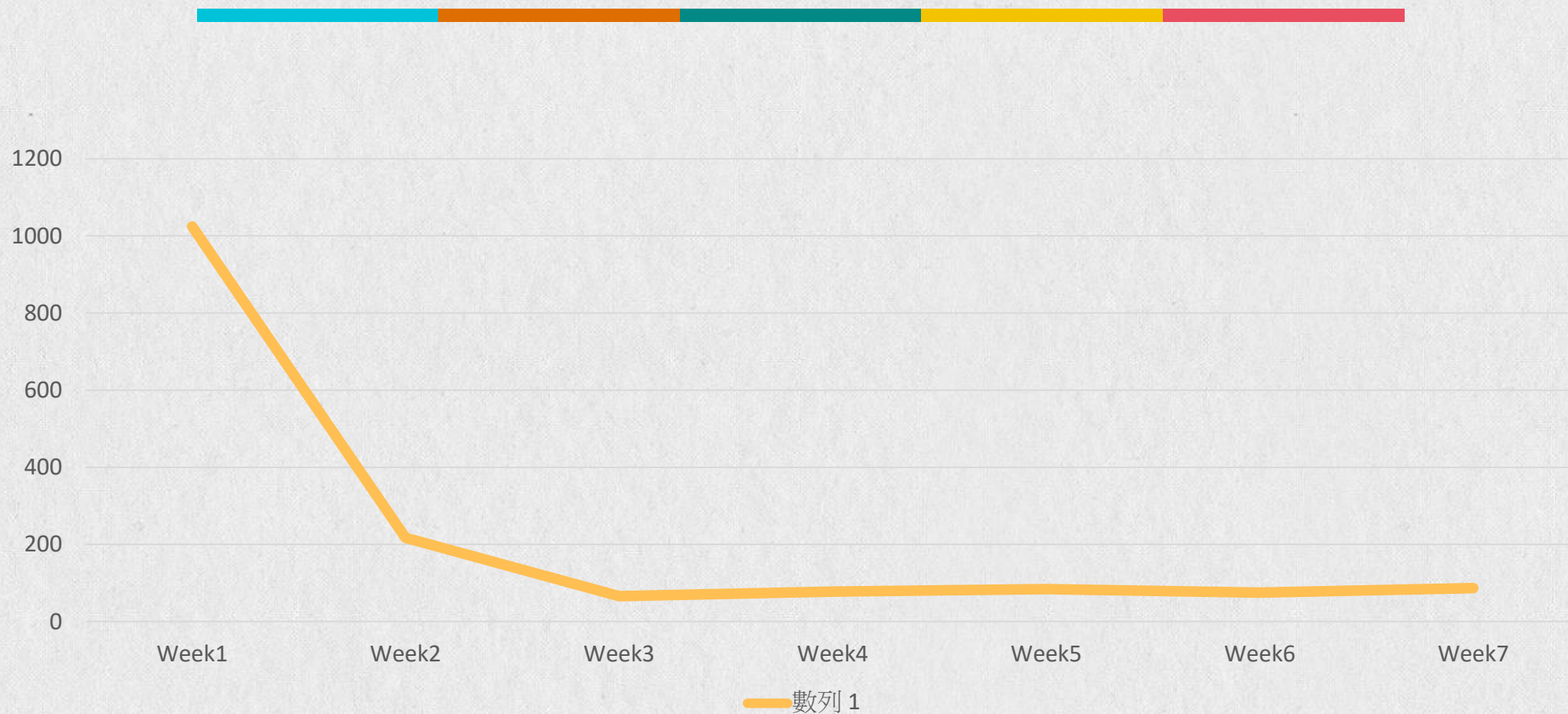
電影《黑豹》
七周票房

0.65

留
關鍵

根據每周票
房與好評關
鍵字多寡變
化做相關圖

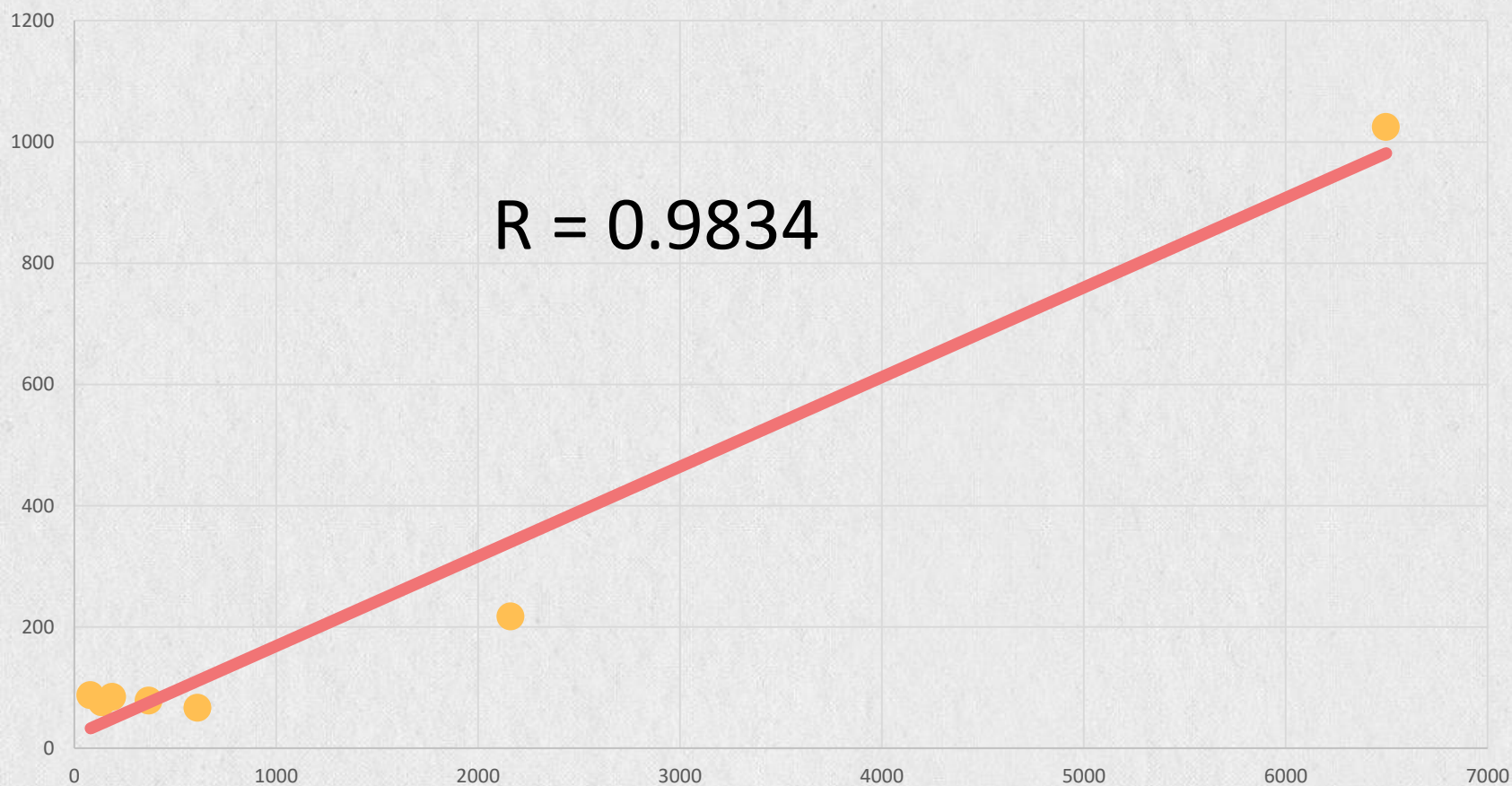
4 好評關鍵字七週變化



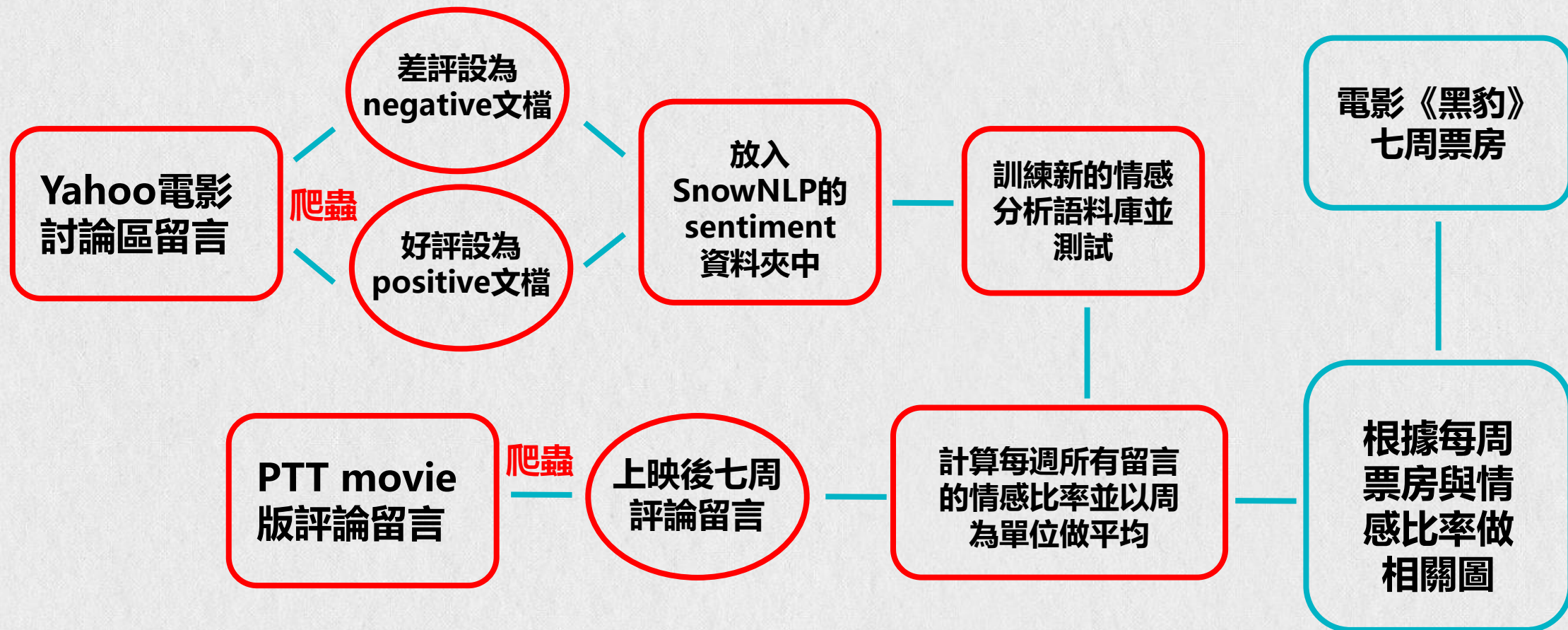
4

好評關鍵字與票房關係

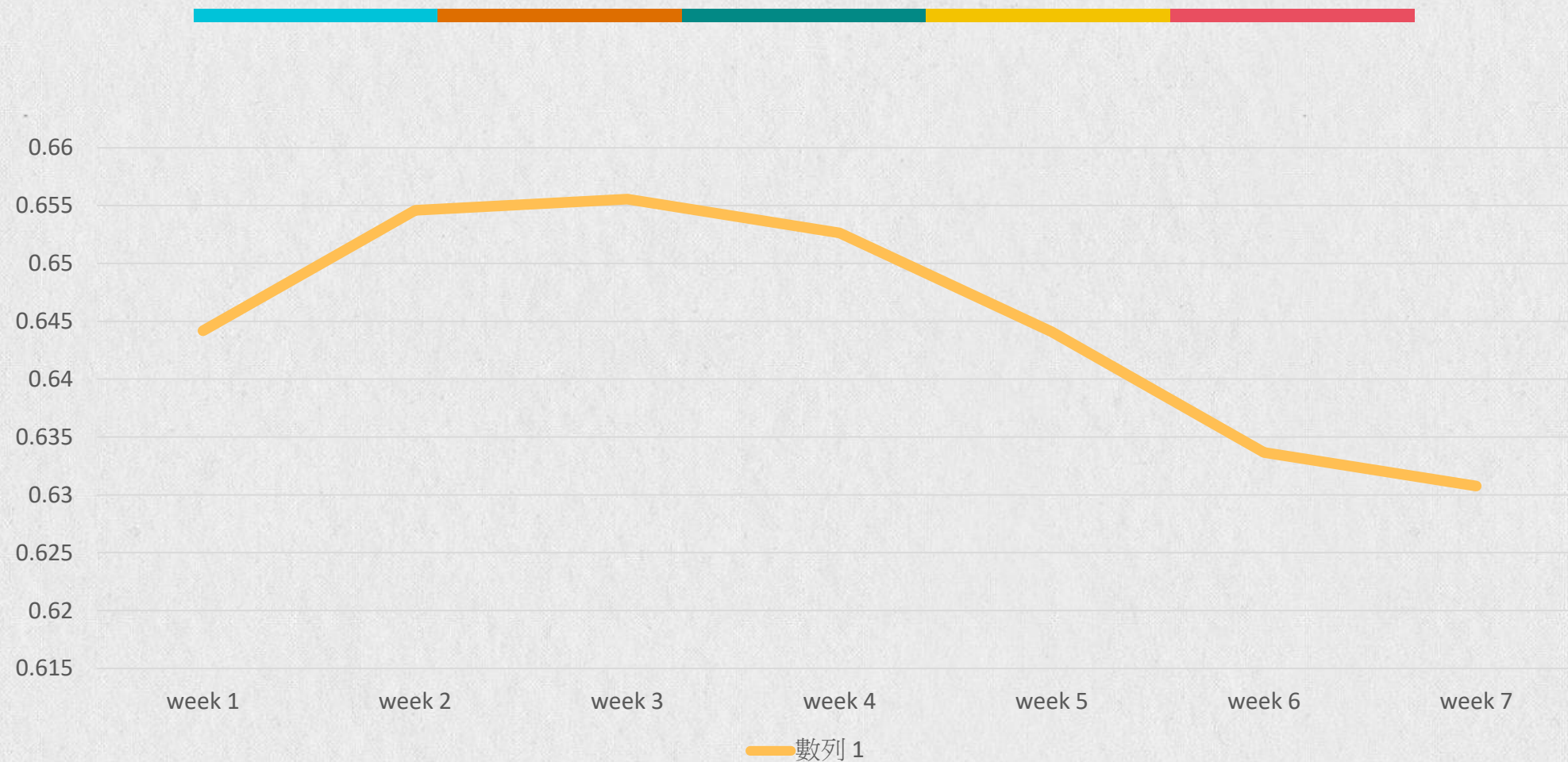
X = 票房 Y = 好評關鍵字



4 情感分析實作流程



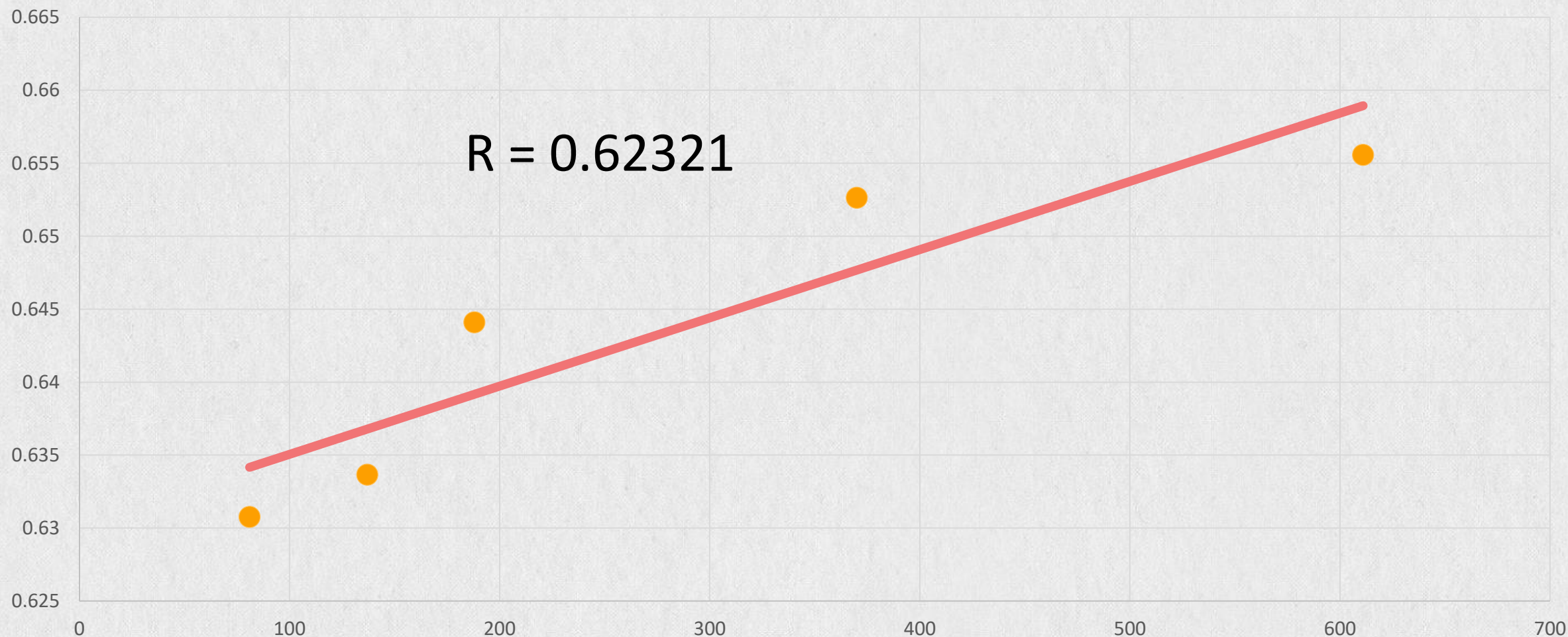
4 情感分析七週變化



4 情感分析與票房關係



4 情感分析與票房關係



4 結論&不足之處

關鍵詞減少≠好評
下降=>討論熱度?

相關圖的資料
過少=>偏誤

首映的票房熱度
≠好評上升

2

樣本資料過少,
模型不夠完善

4

1

3

An abstract graphic on a light gray background. In the center, several large, semi-transparent circles in various colors (teal, blue, green, yellow, orange, pink) overlap each other. The text "Thank You" is written in white, bold, sans-serif font across the middle of these circles. Surrounding the central cluster are numerous smaller, solid-colored dots in the same color palette, scattered across the background.

Thank You