Python 班-爬蟲小作業

# 終於是最後一份作業了！

2020-06-07

**1** 前言

2019 年 12 月以來，中國湖北省武漢市發現多起病毒性肺炎案例，經過相關病毒分型檢測，正式被命名為「COVID-19」（武漢肺炎）。接下來的數個月中，肺炎先是席捲中國以及周圍國家如泰國、日本、南韓，後又傳播至歐美地區，造成全世界疫情大流行。面對嚴峻的情勢，各國政府紛紛祭出防疫手段，希望能夠有效防堵疫情傳播。

武漢肺炎除了衝擊各國的公衛體系，也對社會其他層面產生重大影響， 例如經濟、教育、政治等等。依據疫情嚴重程度的不同，有些國家實施封城，限制民眾外出。娛樂場所如電影院隨之關閉，電影票房一落千丈，世界三大影展之一的坎城影展取消實體放映，原本預計在今年上映的商業鉅片也因應疫情調整上映時間。

在這份小作業中，希望可以利用爬蟲與資料視覺化，來了解肺炎疫情與電影產業之間的關聯性。每一個大題之間沒有關聯性，所以可以依照你喜歡的順序來作答。

**2** 上映電影數量 **(30%)**

**2.1** 蒐集資料 **(15%)**

首先，請利用 IMDb 的[搜尋頁面](https://www.imdb.com/search/title/)，查詢每個月上映的電影數量。

1. 在 Title Type 的欄位選擇 Feature Film（劇情長片）。
2. 在 Release Date 輸入欲查詢的日期。例如想要知道 2019 七月上映的電影，則輸入2019-07-01 to 2019-07-31。其他欄位留空即可。
3. 查詢後，可以看到頁面左上角顯示電影總數。



再來，擷取 **2015** 年 **1** 月至 **2020** 年 **12** 月每個月的電影上映數量，

「日期 數量」（中間以空白分隔）的格式，寫入”number\_of\_films.csv” 當中。完成之後，檔案的內容格式如下：

以

|  |  |
| --- | --- |
| 2015-01 | 917 |
| 2015-02 | 833 |
| 2015-03 | 474 |
| . |  |
| . |  |
| . |  |
| 2020-11 | 291 |
| 2020-12 | 376 |

**2.2** 視覺化 **(10%)**

讀取上題 csv 檔案中的資料，以日期為橫軸，上映電影數量為縱軸， 畫出長條圖。

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

**2.3** 觀 察 **(5%)**

Q: 觀察上題圖表，電影上映的數量在不同月份之間以及武漢肺炎疫情延燒後是否具有規律或是顯著的趨勢？

A: 從圖中可以發現,電影上映的數量在不同月份之間以及武漢肺炎疫情延燒後具有顯著的趨勢，呈現明顯下滑的趨勢。

**3** 電影票房 **(40%)**

**3.1** 蒐集資料 **(15%)**

蒐集[澳洲](https://www.boxofficemojo.com/weekend/by-year/2020/?area=AU)、[日本](https://www.boxofficemojo.com/weekend/by-year/2020/?area=JP)、[德國](https://www.boxofficemojo.com/weekend/by-year/2020/?area=DE)以及[北美](https://www.boxofficemojo.com/weekend/?ref_=bo_nb_di_secondarytab)地區**2019**年 **6**月至**2020 5**月

的週末票房，也就是網頁中”Dates” 與”Overall Gross” 的兩個欄位，並且以

「日期 收益（」中間以空白分隔）的格式分別存入”Australia.csv”, ”Japan.csv”, ”Germany.csv”, NorthAmerica.csv” 當中，如下所示：

|  |  |
| --- | --- |
| 2019-06-06 | 8930319 |
| 2019-06-13 | 6436893 |
| 2019-06-20 | 7074798 |
| . |  |
| . |  |
| . |  |
| 2020-05-21 | 3083 |
| 2020-05-28 | 545 |

**3.2** 視覺化 **(15%)**

一張含有 畫畫 的圖片

自動產生的描述讀取上題輸出的四個檔案，以日期作為橫軸，收益作為縱軸，整合成一張折線圖（也就是圖上會有四條折線，分別表示澳洲、日本、德國與北美）

**3.3** 觀 察 **(10%)**

**Q**: 觀察上題圖表，比較不同國家的數據，並描述整體趨勢。

**A:** 從上圖可以看出美國的電影票房較其他影響更大，四個國家的票房皆在疫情爆發後呈現急遽下滑的趨勢。

4 **上映電影院數量 (40%)**

**4.1 蒐集資料 (15%)**

首先,利用Box Office Mojo來蒐集 2019 年 6 月至 2020 年 5 月的每部電影在幾家電影院上映。

1. 點進連結的頁面,可以看到 2019 年 6 月上映的電影清單。雖然左上角的選單顯示的是 2019 年 6 月,但是拉到清單的最下方會發現有一些 2019 年 7 月的電影,所以需要注意不要下載到重覆的資料。

2. 進入電影頁面之後,可以看見如下圖的表格,請擷取日期與該電影在多少電影院上映的數量,也就是表格中紅色框起的欄位。

3. 請將資料以「日期 數量」的格式(如 3.1 的範例),寫入”theaters.csv”。

4 上映電影院數量 (40%)

**4.2 視覺化 (15%)**

請讀取上題的檔案,繪製熱度圖 (heatmap)。以月分為橫軸,上映的電影院數量為縱軸,以 1000 為一個刻度,並在每個區塊內顯示該月共有多少電影落在對應的上映電影院數量區間。舉例來說,若二月時有 46 部電影上映的電影院數目在 1000 間以下,5部電影上映的電影院數目在 1000 到 2000 間之間,則應如下圖顯示。

一張含有 文字 的圖片

自動產生的描述

**4.3 觀察 (10%)**

Q:觀察上題圖表,比較不同月份數據之間的差異,並描述整體趨勢。

A:我們可以從圖中發現疫情爆發後，上映電影的戲院數量明顯下降

程式碼與資料庫

<https://drive.google.com/drive/folders/1Plmv0gV4osANBdYiJrVHsc-Czox5BwSv?usp=sharing>

<https://drive.google.com/drive/folders/1yEFo0_W5I47m0qV0qSsRbrgblqLsnpiL?usp=sharing>