大數據分析

作業一

(Programming Assignment1)

410921202 資工四 林芷萱

1. The problem description

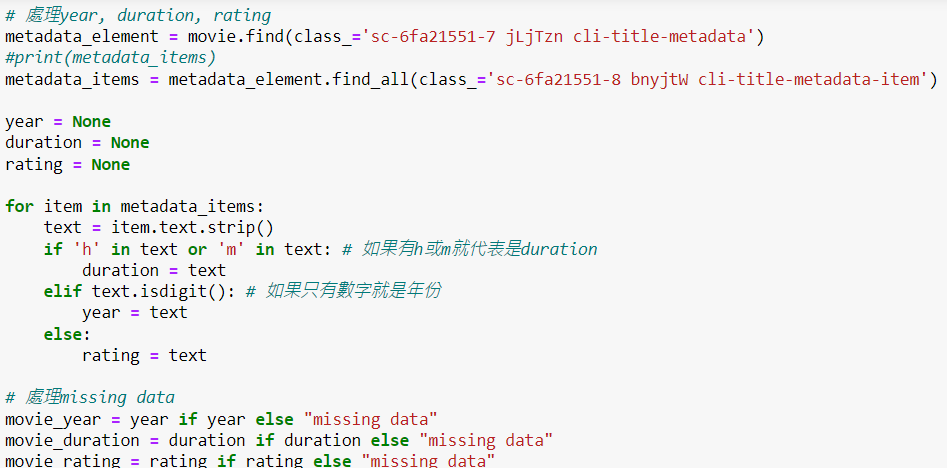
此次作業的要求為使用python寫一個程式爬取<https://www.imdb.com/chart/top/?ref_=nv_mv_250>的資料，分為兩個部分；Part A為爬取前250個電影的rank、name、year、duration、rating、rate並存入csv檔，檔名為films250.csv；Part B為將前250個電影的duration全部加起來。

1. Highlight of the way you write the program

Part A:

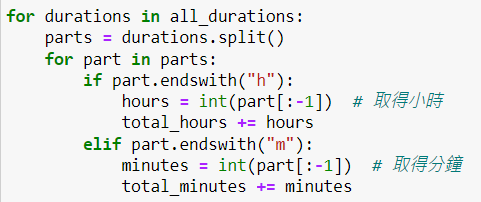
這部分的重點在於怎麼處理missing data，我先爬取year、duration、rating，然後先將每一個都設為None，接著再判斷裡面的值。duration的資料會出現h或m，所以如果有h或m的資料就是duration；觀察year跟rating兩個資料的差異會發現rating的資料是英文、數字、符號混和，而year全部由數字組成，因此只要資料只有數字就是年份，否則則為rating。

接著再將三個值依序放入對應的位置，如果沒有這個值，他就是missing data。

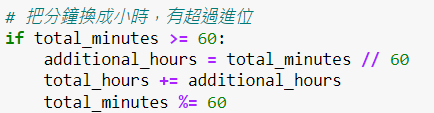


Part B:

這部分我的作法是分別把h跟m前面的數字加起來，如下:

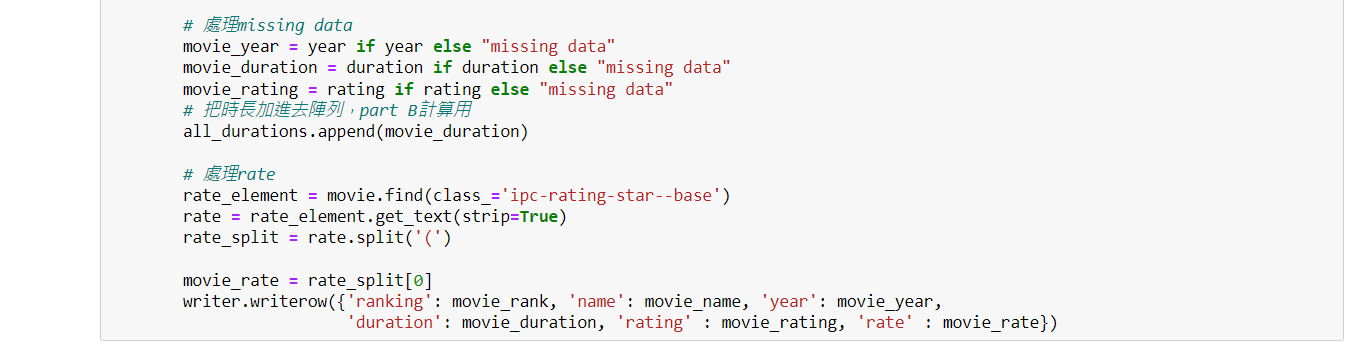


接著再做進位:



1. The program listing

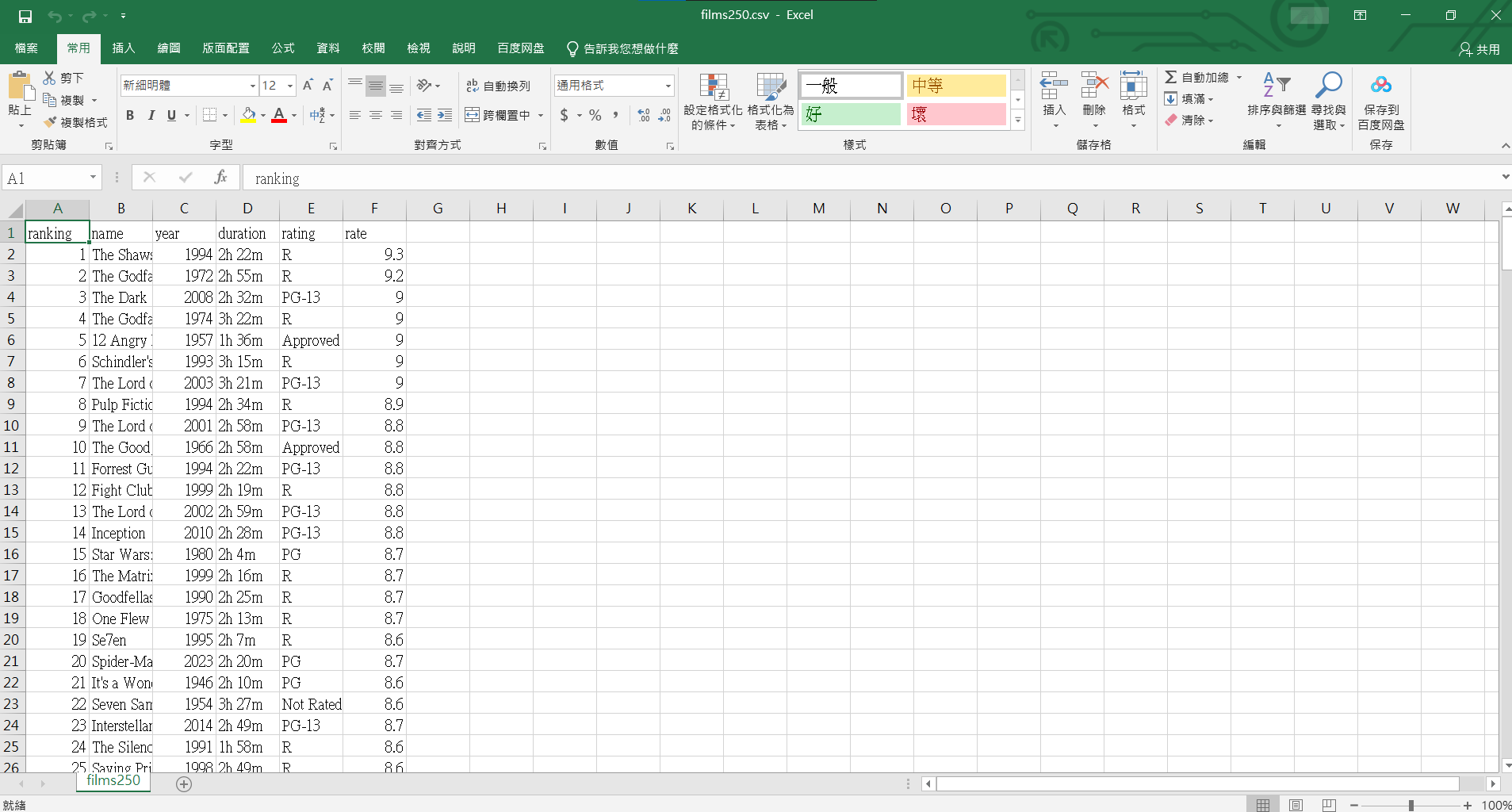
 





1. Test run results

Part A:



Part B:



1. Discussion

在這次的作業當中我學習到了如何使用爬蟲爬取需要的資料以及要怎麼將檔案存進csv檔，在進行的過程中遇到比較大的問題是，我發現在爬取時資料會自動爬取成中文，這會導致csv檔打開來是亂碼，儘管我將encoding設為utf-8也一樣，最後我在headers加入了這兩行程式碼:

'Cookie': 'language=en\_US',

'Accept-Language': 'en-US,en;q=0.5'

這使得爬蟲爬取該網頁時會強制使用英文，就沒有出現爬取到中文資料的問題了。