資訊檢索期中報告–網路爬蟲

組別 : 第九組

組員 : A1045516 蔡湘俊

A1045505 施彥廷

指導老師 : 黃健峰

1. 網路的誕生

在早期的世代當中，網路是不存在的，那時資料的流傳與記載都是以紙張的形式給記錄下來，所以在那時根本談不上有任何的資訊流通，而這也促使了個人更甚至是各國的資料與資料間的交流變得十分的困難與複雜，也因為這樣的訴求的存在，促使了網際網路這一個名詞的誕生，而也因為網際網路的誕生，才能夠促使現在的科技與社會蓬勃的發展起來，直到現今可以說網際網路包覆了整個世界，網際網路所帶來的便利性是人人所知曉的，而且如今的我們也很難離開網際網路的懷抱，網際網路可以說成了整個世界的核心之最，所有的事物都和網際網路有著密不可分的關聯性。



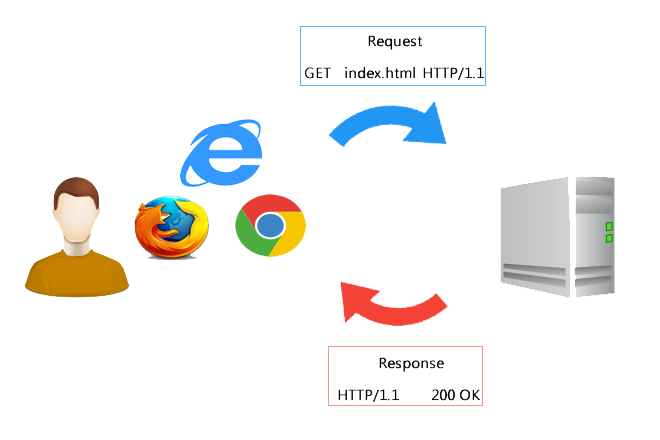
1. 爆炸性世代

而網際網路的誕生對全人類來講最明顯的影響就是造就出了一個新世代的誕生，那就是資訊爆炸世代，一個人擁有了網路就彷彿是擁有了全世界，所有自己想看的、想查的全部都找的到，網際網路就如同是一個虛擬的圖書館，在裡頭擁有無盡的資料存放在那，而且資料量每天是不斷地以好幾倍的速度在成長，沒有人是能夠將這些資料給完完全全的瀏覽過一次，而且這也造就了非結構化的資料量比結構化的資料量還要多上了好幾倍，在以往雖然網際網路是誕生了，可是那時的網際網路的技術沒有現如今來的那麼的成熟，所以那時候的資料量也沒有來的那麼多，再加上那時候網際網路並不是人人都能使用的，只有一些相關的機構單位才能夠去使用，所以那時在網際網路上流傳著的資料，結構化的是遠多於非結構化的，而在到後之所以非結構化的資料比結構化的資料還要來的多上許多的原因就在於網際網路在民間的普及，這也造就了人人都能在網際網路上提供許多各式各樣的資料，也因為如此就促使了資訊爆炸這一個名詞的誕生。



1. 網頁與瀏覽器

在要探討到網路爬蟲之前，就不得不提到網頁這個關鍵元素的存在，因為它是網路爬蟲的核心之處，在前頭我們也探討到現在的我們正處於一個資訊爆炸的世代，在網際網路上散佈著數不清的資料，但這些資料總也該有個顯示的媒介吧，否則我們要如何去看到這些資料呢?那這也正是瀏覽器與HTML誕生的原因所在，我們在網路上看到的網頁其實它的每一個部分都是由HTML給一一建構出來的，而我們之所以能看到一個網頁完整的呈現就在於瀏覽器這一個幕後功臣，瀏覽器的存在就是負責將這些HTML的編碼給進行一個解析的動作，之後再將整個網頁的內容給呈現在瀏覽器上，所以我們所看到的網頁其實是已經經過了一連串複雜的解析與轉化之後所呈現出來的結果，而不是它原本就是長成那樣的形式。

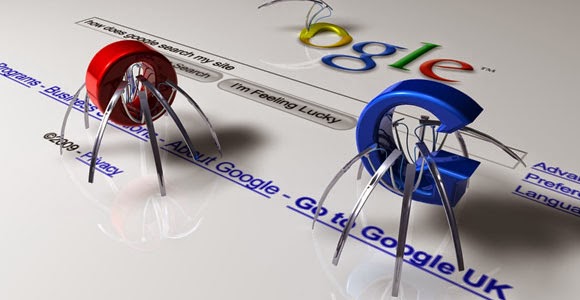


1. 網路爬蟲

那網路爬蟲就是靠這些HTML的Tag來去做一個網頁資料的抓取，那之所以會有網路爬蟲這一個名詞的誕生跟我們前面所提到的資訊爆炸的世代是有著十分大的關連性存在，面對網際網路上頭這些無盡的資料，我們不可能用人力的方式將我們所想找的內容給一一的查詢出來，這不但是一個十分浩大的工程，所需花的成本可以說是遠遠的超乎我們的想像，但更嚴格的說起來這根本就是一個不可能的任務，所以就有人把腦筋給動到電腦身上，竟然人力成本的花費這麼高，那何不讓電腦自動幫我們找尋呢?而也因為這樣的念頭的萌生，造就了網路爬蟲這項技術的誕生，網路爬蟲可以說是一項十分強大的技術，幾乎在任何地方都能夠看到網路爬蟲這項技術的存在，而也因為它強大的存在，造就了有越來越多人使用網路爬蟲這項技術來去抓取網頁上的內容。

雖然網路爬蟲十分的強大，但對於網頁經營者而言是一個十分頭痛的存在，網路爬蟲這項技術之所以強大的原因就在於只要是它爬的到的地方，他都能將在那網頁上所存在的資料給完完全全的抓取，並且只要使用者想爬隨時隨地都能爬，而且想重複爬多少次都沒有任何的關係，也因為網路爬蟲有這樣的特性的存在，使得對網頁經營者所架構的server產生了極大的負擔與危險，在平時正常的瀏覽下，瀏覽的人數可以說就已經十分的龐大了，如果又多了網路爬蟲的造訪，那可以說整個伺服器是呈現滿載的情況，並且造成伺服器所能提供的頻寬是大大的縮減，造成當有其他用戶要來造訪同樣一個網頁的時候會產生卡頓，更甚至直接導致連都連不上的情況發生，所以絕大部分的網站都是不允許網路爬蟲來去爬網站上的資料，一來是為了避免影響用戶的瀏覽體驗，二來是為了避免有心人士的惡意攻擊，但有些網頁還是允許網路爬蟲來去爬他們的網頁，可是他們都有做一個特別的規範，並且這些規範都會統一寫在robots.txt這一個文件裡頭，來去規範網路爬蟲的使用者哪些內容是可以爬的，而哪些內容是不允許爬的都詳細的被填寫在文件裡頭，而會這樣做也是為了避免有心人士將網站內部的機密給流傳出去。

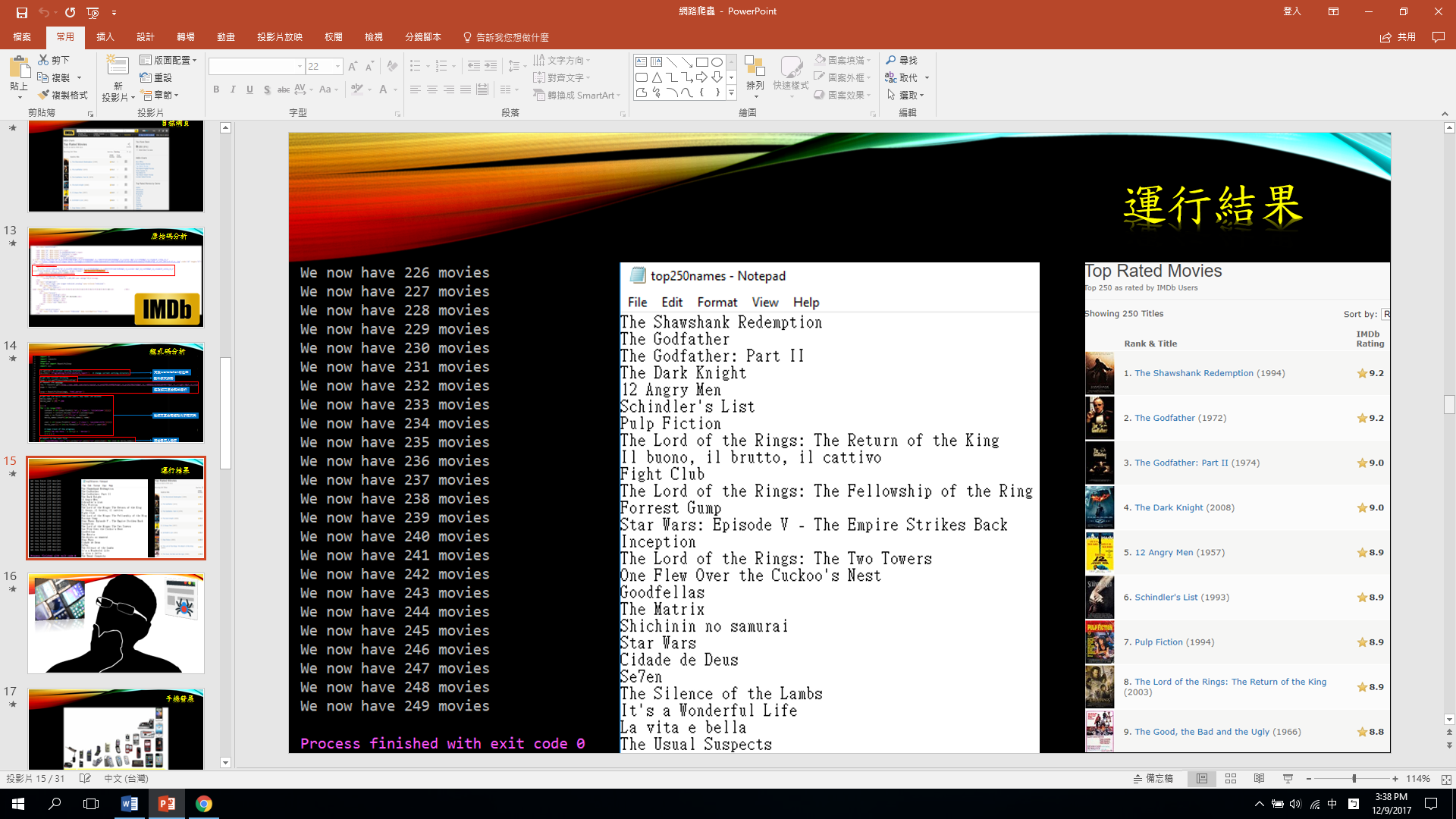
所以網路爬蟲技術是強大的，可是它可以被用在好的地方，那它當然也可以被用在不好的地方，所以才會產生出這麼多的防護機制，就是為了避免這些有心人士的惡意攻擊，導致自己的網站被別人給弄到出問題，因為一旦網頁出了問題，那其對網頁經營者所造成的損失絕對不是我們一般人所能夠想像的



1. 網路爬蟲實作

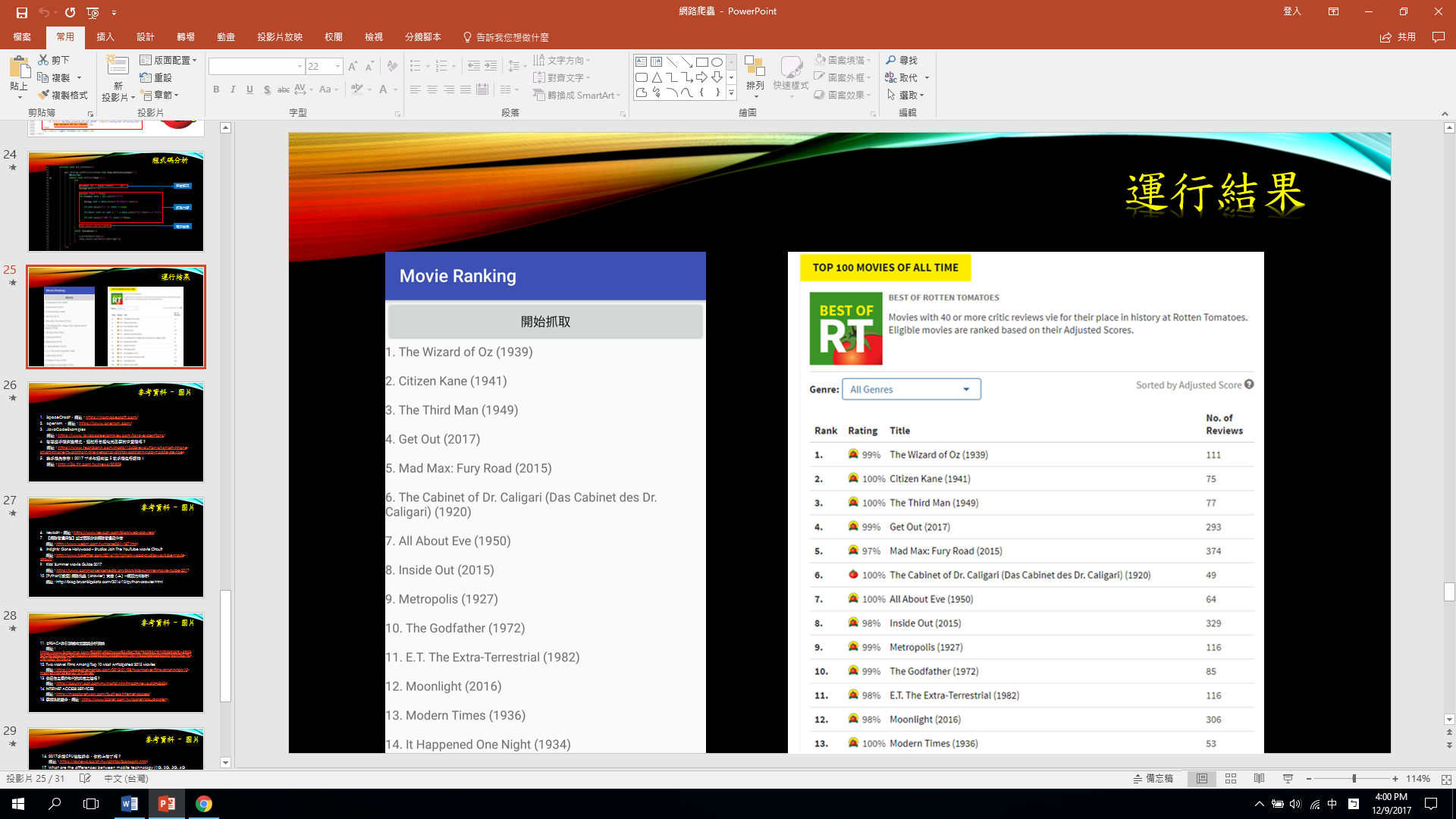
而我們組做的實作分成了兩大平台，一個是電腦端上，另一個是手機端上的，而我們爬的內容就是電影的排行榜，我們之所以會想去爬電影的排行榜的原因就在於說我們組的組員本身很喜歡看電影，但每次都煩惱於不知道要看哪一部電影，所以每次都會去查電影的排行榜來去挑選一部自己喜歡看的電影，那為了省去每次都要手動去查找的麻煩，所以我們就想到說用網路爬蟲這項技術來去幫我們爬電影的排行榜。

那電腦端的部份我們是用Python這個程式語言來去實作網路爬蟲這項技術，那下面這張圖就是我們爬出來的結果



我們所爬的網頁就是IMDb的Top Rated Movies這一個網頁，他們綜合了所有的評價來去票選出前250名的電影，那在Python裡頭我們用到了Requests和BeautifulSoup這兩個module，Requests這一個module它的功能最主要就是和目標網頁建立連結，而BeautifulSoup這一個module就是去將網頁的原始碼給存放到這個module裡頭，並且產生出了一個這個module下的物件，而這個產生出來的物件的內容本身是一個樹狀的結構，分別利用HTML的Tag下去進行建樹的動作，然後每一個HTML Tag下的內容就是對應到此Tag所呈現的資料，那我們就是利用這一個產生出來的物件來去利用HTML的Tag下去抓取出我們所想要的內容，然後一路的抓取直到抓到了所有電影的排行榜才結束爬蟲程式的執行。

而手機端的部份我們是用Java這一個程式語言來去實作網路爬蟲這項技術，那為什麼我們組會想要在手機端實作爬蟲的緣故是在於說在網路上大部分都是利用電腦下去進行爬蟲的動作，可是極少數的資料是用手機端下去進行爬蟲，那我們是想說現如今的智慧型手機其所搭載的處理器的效能可以說是一年比一年還要來的強大，更甚至產生了效能過剩的問題的存在，當然也不單單因為這一個因素的影響，另一個影響因素就是行動上網，從一開始1G的誕生到現如今4G的世代，手機端的網速可以說是有著大幅度的提升，更甚至觀看4K視頻都不是什麼問題，那基於這兩點的考量，所以我們組才會認為說手機端是可以實作爬蟲的，那下面這張圖就是我們爬出來的結果



而我們所爬的網頁就是Rotten Tomatoes的TOP 100 MOVIES OF ALL TIME這一個網頁，那在Java裡頭所用到的module就是Jsoup、Document和Element，Jsoup這個module可以說是Python的Requests和BeautifulSoup這兩個module的結合體，它既能和目標網頁進行連結的動作，也會利用HTML的Tag來去進行建樹的動作，而這些Tag下所存放的內容也會一併放在這個Tag的底下，然後再將其建出來的結果給放到Document這個類別來去進行管理與內容提取的動作，也就是我們可以取出特定Tag裡頭所存放的內容，之後再利用Element這個類別將提取出的內容給儲存起來以便作為之後資料的存取與使用，那藉由這三者的互相搭配就能夠將電影排行榜給抓取出來。

1. 結論

網路爬蟲的技術在現如今已經成為了一個不可或缺的技術，我們處處都可以看到網路爬蟲這項技術的應用與身影，並且網路爬蟲這項技術也是變得越來越成熟，而這項技術的成熟也促使我們所能爬到的內容是越來越多，但這項技術的強大所帶來的就是更加可怕的危害，這項技術被用在好的地方勢必可以帶來極大的效益，可是如果是被用在壞的地方的話，那麼對網頁經營者所造成的傷害將會十分的巨大，一項技術的進步不為了別的，只為了能夠帶來更多的便利性，而不是為了拿來當成做壞事的工具，作為技術的使用者，我們應該要將技術導向更光明的道路，並且讓這項技術可以造福更多的人、事、物。