

中 華 大 學

學士專題

應用爬蟲技術抓取購物網站價格

學 系 別： 工業管理學系

學號姓名： B10603001 蔡 杰 哲

指導教授： 劉 光 泰 教 授

中 華 民 國 109 年 12 月

目錄

第一章 緒論.....	3
第一節 研究背景與動機.....	3
第二節 研究目的.....	5
第三節 研究限制.....	6
第二章 文獻探討	7
第一節 PYTHON 語言	7
介紹.....	7
特點.....	8
應用.....	9
第二節 網路爬蟲.....	10
什麼是網頁爬蟲.....	10
為甚麼要用爬蟲.....	10
如何開始爬蟲?.....	11
爬蟲應用.....	12
爬蟲相關套件.....	12
反爬蟲.....	12
第三節 HTML	13
介紹.....	13
結構.....	13
第四節 電子商務.....	14
第三章 研究方法	15
第一節 研究流程.....	15
第二節 研究對象及時間.....	16
第三節 研究環境.....	16
第四節 分析網頁架構.....	17
一、Momo 的網頁架構.....	17
二、Yahoo 購物中心的網頁架構.....	18
第五節 利用 PYTHON 的網路爬蟲爬取資料	19
一、套件.....	19
二、反爬蟲機制.....	19
三、抓取資料.....	21

四、資料封包.....	22
第四章 研究結果	24
第一節 通過反爬蟲系統.....	24
第二節 資料整合.....	25
第三節 分析價格變動.....	27
第五章 結論.....	30
第一節 結論.....	30
第二節 建議.....	31
第三節 未來展望.....	31
第六章 參考文獻	32

第一章 緒論

第一節 研究背景與動機

近年來網路成長幅度驚人，2019 年的網路趨勢報告《Internet Trends 2019》，指出網路人口數已達 38 億人，已經超越全球人口數的一半(2019 全球人數約 77 億人)，相較於 2010 年，世界人口 69 億，網路人口數為 20 億，約 30%，依資料分析結果，使用網路的人口增加了 19%之多。

隨著網際網路的發達，也帶給了人們許多的便利性，從前的報紙，電視，廣播能帶給我們的訊息，漸漸的被網際網路取代了，網路的便利性、迅速性，漸漸地改變了人們的生活型態，由從前的廣告看板，變為了現在的廣告影片置入，由補習班，變為了線上教學影片，傳統的股票也已被電子股票完全取代，這種便利也為購物領域，變得更為方便輕鬆，消費者只需要在家滑著手機電腦，即可尋找到自己想要的商品，省去了出門逛街的時間，商品的多樣性甚至也大於實體店面。

人們每天都會在網路產生上千上萬的資料訊息，可以說這些數據是跟著人們的每一天產生的，又稱為大數據，這些數據都可成為任何產業的工具，政府、商家、醫院等等...，政府可以用來分析人們的違法行為或消費方式，商家可以用來得知消費者最能接受的價位、目前流行的產品，醫院可用來分析藥物反應，加以研究出更佳的醫療方法等等...，這些數據都是在推進人類進步的重要資料。

也因為網際網路的蓬勃發展，電子商務已成為日常生活中經常會使用的網路應用。使用網路商城購物的人口數逐年增加，而網路商店也衝擊至傳統的實體商店，使得傳統商店的商業模式已漸漸轉換為網路商城模式，這種轉變的優點是在於能夠減少時間的運用、節省交通的成本以及網路商城的價格會遠比實體店面的價格來得低。使得許多消費者在購買商品時，會選擇在網路商城購買，而在眾多的網路商城平台下，為了吸引消費者購買，而打起價格戰、或是各類優惠，因此在網路上會看到相同商品，有多種不同的價格。

根據 Similarweb(2020)統計台灣每月網路商城之流量，可看出在台灣每個月使用網路商城的人數平均約為 194.705 佰萬人，每日約 6.28 佰萬人會至網路商城進行購物，其中以蝦皮購物之人數最多，平均每月人數約為 55.7 佰萬人，其次為 PChome 及露天拍賣，平均人數也有高達 30 佰萬人。由此可知，電子商務對網路資訊迅速發展的現今，成了不可或缺的角色，其流量如表 1-1 所示，

台灣每月網路商城流量表

網站	2020 年 5 月至 2020 年 10 月（單位為百萬）						總流量
	5 月	6 月	7 月	8 月	9 月	10 月	
蝦皮購物	51.84	52.16	56.4	58.55	55.44	59.83	334.22
PChome 24h 購物	30.47	31.91	31.71	33.57	28.67	30.29	186.62
露天拍賣	28.2	27.2	28.78	33.15	32.49	34.35	184.17
momo 購物網	29.56	29.8	30.12	31.37	28.85	28.84	178.54
博客來	14.43	14.54	14.79	15.28	14.37	14.05	87.46
yahoo 拍賣	14.1	13.2	13.5	13.5	13.3	13.6	81.2
Rakuten 台灣樂天	6.38	6.2	7.18	7.28	7.14	7.13	41.31
Pcone 松果購物	4.17	4.24	4.69	4.84	4.33	4.07	26.34
生活市集	4.12	4.09	4.4	4.71	4.12	4.04	25.48
ETMall 東森購物	3.44	3.66	3.75	4.13	3.87	4.04	22.89
總計	186.71	187	195.32	206.38	192.58	200.24	1168.23

參考資料：Similarweb。

一般消費者使用電子商城選購產品時，選擇的第一條件不外乎是選擇價格最優惠的產品，在不同商城亦或是不同的賣家，所販售的價格皆不同，因此往往在選購 3C 產品時，消費者會利用許多時間去一一比較。電子商務的商機一直持續地成長，使消費者能夠不出家門，在網路上便能購買自己所想購買的商品，電子商務的興起改變了許多人的生活模式，也使得市場上的競爭愈來愈激烈。對賣家而言，因

為有眾多的消費者族群加上不用負擔實體店面的租金成本及人力成本，故許多實體店家會轉戰在網路商城架設虛擬商店減少成本的支出，藉此降低價格，以吸引消費者購買。對消費者而言，除了能夠有更多的選擇，選擇購買的商品，也可減少交通成本以及購物的時間，能夠將大量花費在交通的時間去關注願意購買的商品，對商品做各種的比較。

第二節 研究目的

過去，要進行資料的採集與分析，多數以人工進行採集工作，必須一項一項慢慢收集，這樣的傳統方式不僅費時，且需要大量的人工成本，又或者有缺失的項目沒有蒐集到，因此，在現今的網路大數據時代，已不再適用這種傳統的資料蒐集方式。自動化搜尋大量資料並分析，可用一段程式，採取到大量網頁資料，再利用篩選功能挑出所需要的資料即可，網路爬蟲(Web Scarping)，又稱網路蜘蛛，就可做到上述所說的工作。

在這資訊眾多的時代，大家都可以利用網路上的開放資料去做分析，雖說資料眾多，但實際上所得到的資料或許是分散、不完整的，若能將資料加以統整、篩選、整合，會讓分析人員更加輕鬆的去辨別。

在台灣每月使用網路商城的人數平均約為 194.705 佰萬人，每日約 6.28 佰萬人會至網路商城進行購物，例如想搜尋一台電視，可以到許多購物網站搜尋型號及價位，但購物網站有上百種，一次僅能搜尋一個網站，及一種型號，這樣要依依比價是非常麻煩的。

本研究主要利用網路爬蟲技術，以 Momo 購物網、yahoo 商城為分析對象，知名手錶品牌 G-Shock，型號 GA-110GB-1A 為採取樣本，探討在 2 家電子商務平台的樣本價格波動程度，故此研究目的有以下三點。

1. 通過反爬蟲機制，抓取購物網站之資訊
2. 利用 Python 將取得的資訊做好排序

3. 分析兩家平台的價格關係

第三節 研究限制

- 一、一般網站的防爬蟲機制: 凡是有一定規模的網站，大公司的網站，或是盈利性質比較強的網站，都是有高階的防爬蟲措施的
- 二、Dos 阻斷式攻擊: 在網路世界中，使用者會與伺服器一直互相傳遞封包，一個動作，並向伺服器做出一個請求，若在爬蟲過程中請求數量超過負荷量，伺服器就會當機，不僅造成他人網站困擾，也有被鎖 IP 的可能性。
- 三、資料僅能抓取當日金額，因此本研究需每天開起一次此程式並執行，抓取當日金額，直到抓取樣本數足夠，才能開始做分析。

第二章 文獻探討

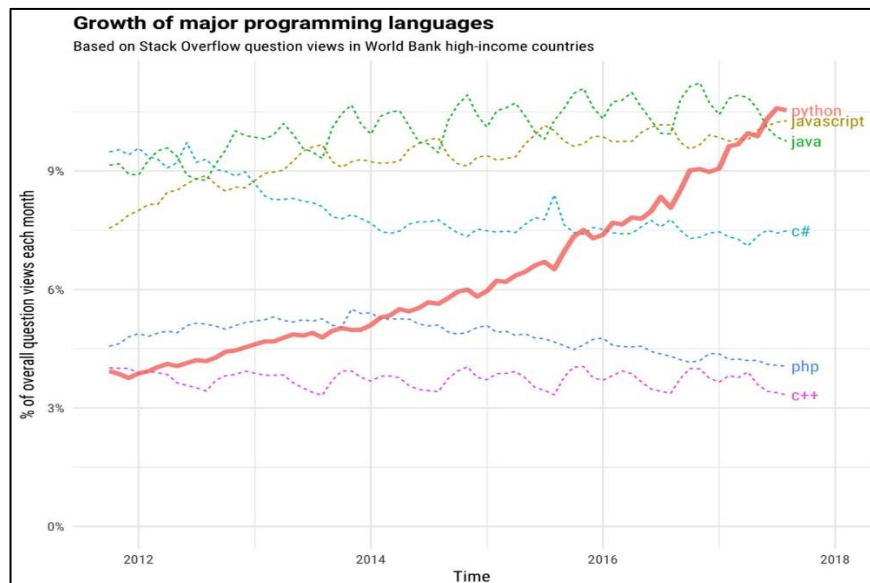
第一節 Python 語言

介紹

Python 由吉多·范羅蘇姆創造，第一版釋出於 1991 年，它是 ABC 語言的後繼者，也可以視之為一種使用傳統中綴表達式的 LISP 方言。其設計哲學強調代碼的可讀性和簡潔的語法，尤其是使用空格縮排劃分代碼塊。相比於 C 或 Java，Python 讓開發者能夠用更少的代碼表達想法。 [15]

Python 是一種易於學習、功能強大且被廣泛使用的高階程式語言，屬於通用型程式語言。最早於 1989 誕生最初版，在 1999 年排行僅為第 22 名的程式語言，在今年以爬升為全世界第 3 名，僅次於 Java 與 C 語言。[1][2]

簡單來說，Python 現在已經是世界上最流行的程式語言之一，也是未來的主流，能登上排行其中最大原因是下面將會講到的，容易撰寫，相對於其他程式語言，Python 語言的確是淺顯易懂，儘管沒有接觸過任何程式語言，初步的學習是相當簡單的。



圖片來源：stackoverflow.com

	<p>Stack Overflow 表示，在眾多程式語言中，有關 Python 問題的訪問者數量，增長得比其他任何語言的都快，這讓 Python 有資格聲稱它是世界上增長最快的主要編程語言之一。[17]</p>
特點	<p>能夠攀升為全球第 3 的程式語言，他佔據了甚麼優勢？</p> <p>1. 容易撰寫：以下圖片為例，相較於 Java 與 C，同樣的輸出結果，Python 的程式碼是最為簡潔的。</p> <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>Python 程式碼</p> <pre>Print("Hello!")</pre> </div> <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>C 程式碼</p> <pre>#include Int main(){ printf("Hello!"); }</pre> </div> <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p>Java 程式碼</p> <pre>public class Hello { public static void main(String[] args) { System.out.println("Hello!"); } }</pre> </div> <p>圖 2-1[2]</p> <p>2. 功能強大：Python 提供相當多套件供開發者使用，可從網路上抓去別人寫好的模組，應用於自己的程式中。一個.py 的檔案可視為一個模組，而在程式中匯入模組的好處是，讓使用中的程式碼不會過於攏長，後期會很難維護。匯入的方法也很簡單，像是：</p> <pre>import mod</pre> <pre>from mod import function</pre> <pre>import mod as 自定義名稱</pre>

	<p>3.跨平台：各種主流作業系統間相容，如：Mac、Windows、Linux。</p>
應用	<p>數據分析與處理：Python 通常被用來做數據分析，因為 Python 可直接進行調用，方便且靈活，可以根據數據分析與統計的需要靈活使用。Python 也是一個比較完善的數據分析生態系統，其中 matplotlib 經常會被用來繪製數據圖表，它是一個 2D 繪圖工具，有著良好的跨平台交互特性。[3]</p> <p>Web 開發應用：Python 是 Web 開發的主流語言，但不能說是最好的語言。畢竟，在 Web 開發中應用多使用 JavaScript，原因是已一套成熟的框架。Python 開發的 Web 項目雖小而精，但支持最新的 XML 技術，而且數據處理的功能較為強大。[3]</p> <p>人工智慧應用：Python 擁有強大而豐富的函式庫以及數據分析能力。而且 Python 是面向對象的動態語言，且適用於科學計算。不僅如此，Python 提供了大量的 API，這也正是因為 Python 當中包含著較多的適用於人工智慧的模塊[3]，近年來常聽見的機器學習、類神經傳遞也包含其中。</p> <p>網路爬蟲：簡而言之，爬蟲是模仿人類瀏覽之行為，我們在網頁上查看哪些資訊、點擊哪些按鈕和做哪些動作，都可以透過爬蟲進行模擬，爬蟲的兩大工作為下載檔案和分析內容，透過爬蟲程式下載網頁資料，並透過搜尋、字串處理和取代等各種技巧過濾出我們需要的資料。實際上，爬蟲所要處理的是沒有透過瀏覽器處理的 HTML 原始碼，我們平常瀏覽的網頁，其實是由 HTML、JavaScript 等程式語言撰寫而成，透過瀏覽器的處理後</p>

	<p>呈現成我們瀏覽的網頁，所以想透過 Python 程式語言撰寫爬蟲程式必須先有一點 HTML 網頁程式設計的基本知識[4]，此應用也將會是本研究的主题。</p>
--	--

第二節 網路爬蟲

什麼是網頁爬蟲	<p>是一個可以自動化抓取網頁內容的程式。[6]</p> <p>網頁爬蟲是指利用程式去自動獲取網頁資訊的技術，當你掌握了此技術之後，你就可以有源源不絕的資料進行各種應運。因此這是現今資料分析人員必定要掌握的技術。[5]</p> <p>一般使用者是以瀏覽器依據網址(url)向某一網站伺服器送出請求(request)，如果對方伺服器同意你的請求，就會做出回應(response)，將網頁的原始碼回傳給你的瀏覽器，瀏覽器再將原始碼轉換成圖文並茂的頁面。[5]</p> <p>而網路爬蟲即是以程式碼偽裝成一般的使用者，向對方伺服器送出請求，取得回應（原始碼）。再從原始碼中抽取出需要的資訊。[5]</p>
為甚麼要用爬蟲	<p>相信大家多少都遇過需要抓取網頁資訊的時候，也許是因為要做報告、或是出於興趣想研究，需要相關參考資料。最簡單的方法就是一筆一筆複製，然後貼到 excel 或是文字編輯器儲存，再做後續的分析。[6]</p> <p>相較於人工操作的，是將需要一頁頁的將某資訊複製貼上到你的資料集中，而網路爬蟲在拜訪網站時是以程式碼進行，因此可以重複的、自動的一再拜訪，這樣可以為你的省省下許多寶貴的時光。[5]</p>

<p>如何開始爬蟲?</p>	<p>Requests + 解析器(PyQuery 或 BeautifulSoup)：這是最簡單的做法，Requests 是 Python 的一個套件，它可以建立 HTTP 請求，也就是上面提到的 Request，接著我們收到的 HTML 在使用解析器去擷取出想要的資訊，比較常用的套件是 PyQuery 跟 BeautifulSoup，這種方式好處在於你不用裝太多東西。[6]</p> <p>使用爬蟲框架：有些網站對於資料的保護意識比較高，會對於爬蟲做一些防護，稱作為反爬蟲。這種網站可不是那麼好對付，框架通常都會幫你想些解決方案，你可以用框架提供的函示或是參數的調整來輕鬆達成，就不用什麼都自己寫了，最常聽到的 Scrapy 跟 Pyspider，兩者各有各的優缺點，簡單來說，Scrapy 自訂程度較高，而 Pyspider 則是有監控介面、易維護的優勢。[6]</p> <p>爬蟲流程：</p> <ol style="list-style-type: none"> 1. 鎖定目標：目標網站、所需資料? 2. 觀察網站結構：上述有提到須了解 HTML 結構原因就在此，需先看出結構，才能對程式下達指令。 3. 是否有反爬蟲：試圖獲得 html 原始碼。觀察一下，是否會被對方伺服器判定成惡意程式擋掉。 4. 剖析網頁原始碼：是否能將 html 原始碼中，含有你所要的資訊的片段，抓取下來？ 5. 重複爬取：當可以成功獲得一個分頁的特定資訊後，就可開始利用迴圈大規模的重複拜訪撈資料。這個過程中很容易會遇到例外狀況將你的爬蟲程式阻隔。
----------------	---

	<p>6. 打包資料，輸出。</p> <p>[5]</p>
爬蟲應用	<ol style="list-style-type: none"> 1. 自動下載新聞網的標題或內文 2. 自動抓取股票資訊 3. 自動抓取購物網站商品之名稱、價格、資訊 4. 抓取頁面所有圖片
爬蟲相關套件	<p>Request: 使用 Python 來下載網頁上的資料，最基本的作法就是以 requests 模組建立適當的 HTTP 請求，透過 HTTP 請求從網頁伺服器下載指定的資料。[5]</p> <p>Beautifulsoup：可以快速解析網頁 HTML 碼，從中取出使用者有興趣的資料。[7]</p> <p>time：負責處理時間上的運算，可以計算爬蟲花了多久，這樣可以有效推估一次爬蟲一萬筆、十萬筆大約多久。[7]</p>
反爬蟲	<p>我們知道爬蟲的目的是為了獲取網絡資源。爬蟲的本質就是抓取網站中有價值的數據；另外量變產生質變，當數據達到一定的量，我們就可以通過分析數據進而得到一些有用的結果。那麼反爬蟲的其中一個目的是為了保護某些數據不被過分地獲取，保證數據在服務方的控制範圍內使用。[16]</p> <p>有一些常見的反爬蟲操作，例如：限制訪問頻率及次數、圖形驗證碼、User-Agent 請求頭驗證。[16]</p>

第三節 HTML

介紹	<p>HTML 是 Hypertext Markup Language 的縮寫，也就是「超文本標記語言」，標記語言 (markup language)，而非一般熟知的程式設計語言；它會告訴瀏覽器該如何呈現你的網頁，HTML 包含了一系列的元素 (elements)，而元素包含了標籤 (tags) 與內容 (content)，我們用標籤來控制內容的呈現樣貌，例如字體大小、斜體、粗體、在文字或圖片設置超連結等。[8][9]</p> <p>所有的網頁都是 HTML 文件，說得更精確一點，所有的網頁內容，都必須透過 HTML 標記來定義。[8]</p>
結構	<p>1. 起始標籤：「<p>」。起始標籤代表這個元素從這裡開始。[9]</p> <p>2. 結束標籤：「</p>」。與起始標籤一樣，只是在元素名稱前面多了個前置斜線「/」。內容的最後加上結束標籤，代表這個元素的尾端。[9]</p> <p>3. 內容： 這個元素的內容，以下面的例子來說，內容就是這句文字。[9]</p> <p>4. 元素： 由起始標籤、結束標籤、內容所組成。元素還可以有「屬性 (Attribute)」。利用屬性，我們可以設定這個元素的色彩、對齊方式…等等。 [9]</p> <p>以一段句子為例，</p> <p><p style='font-weight:bold'>內容內容內容</p> [10]</p> <p>5. 文件標題 (heading)：讓你呈現這些內容的主題，就像一本書有書名、章節名稱和副標題，一份 HTML 文件也有類似的概念。HTML 最多可以有六層的 heading。</p> <p><h1>My main title</h1></p> <p><h2>My top level heading</h2></p>

	<pre><h3>My subheading</h3></pre> <pre><h4>My sub-subheading</h4> [9]</pre> <p>連結 (link)：連結對於網頁來說是非常重要的。要加上連結，我們需要用到這個元素 <code><a></code>，<code>a</code> 代表了「anchor」。要讓文字變成連結的步驟如下：</p> <ol style="list-style-type: none"> 1. 把他們包在這個<code><a></code> 元素裡：<code><a>超連結顯示的文字</code> 2. 在<code><a></code> element 中加上 href attribute 這個屬性： <pre>超連結顯示的文字</pre> 3. 添加網址：<code>超連結顯示的文字</code> [9]
--	--

第四節 電子商務

過去，你可能沒想過，原來像 PChome、Yahoo 這麼大的平台，也會面臨營收下降，勁敵出現的巨大挑戰，好比蝦皮、486 團購的異軍突起等。網路的無遠弗屆、科技的日新月異帶動交通運輸的便利以及線上安全交易平台的產生，讓「跨境銷售」不再困難重重。[13]

不論是身為行銷人員或是消費者，都會發現網購的商機仍持續地在增長，因為當代的使用者已經更習慣透過網路來找資料、搜集資訊，而人流動的地方就有商機，在網路上發生的交易簡單來說就稱之為「電商」。[13]

電子商務的本質是交易的場所，只是交易發生的地點轉移到了網路上，因此比起傳統交易方式可以更快速、更容易甚至是跨國界的連結買方與賣方。[13]

電子商務（簡稱電商）的發展從 1994 年第一筆在線上透過加密至今，也已經過了數十餘載。歷經了大大小小的市場考驗，再到如今百花齊放的盛況，不僅改變了世界的產業生態，也創造了無數的工作、市場發展機會。面對以驚人姿態蓬勃發展的電子商務市場，許多人不論是自創品牌，或是傳產轉型，也都紛紛投入電商經營的領域。然而隨著機會與可能的大增，市場競爭也越來越激烈。[14]

第三章 研究方法

第一節 研究流程

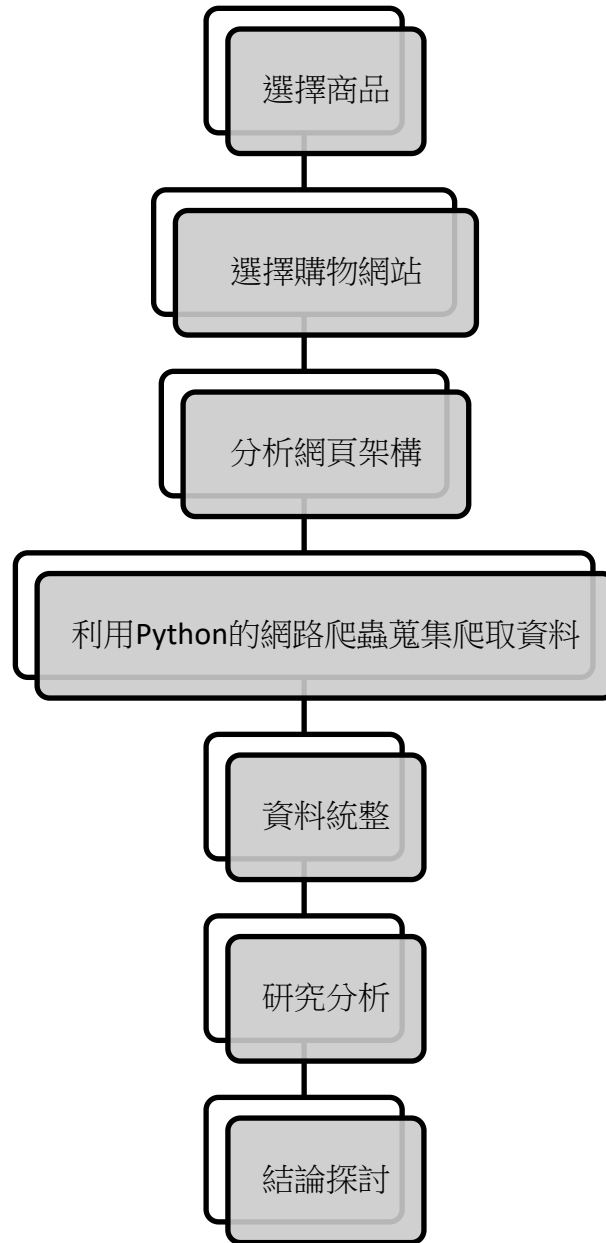


圖 3-1 研究流程圖

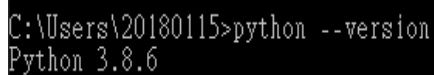
第二節 研究對象及時間

本研究的購物網站選擇 Yahoo 購物中心及 Momo 購物網，因為此二種購物網站所販售模式為 B2C，不會含有 C2C 之商品，因此商品皆為全新，且售價為市售價，而研究樣本選定為 G-Shock-GA-110GB-1A。

研究時間為 2020 年 11 月 27 日至 12 月 14 日，每日上午 10 點爬取樣本之價格，並輸出成 Json 檔，爬取資料會包含商品名稱、商品價格及商品網址。

第三節 研究環境

本研究使用 Python 3.8.6 版本，解譯器使用 PyCharm 2020.2.3 版本。



```
C:\Users\20180115>python --version
Python 3.8.6
```

圖 3-2 Python 版本



```
PyCharm 2020.2.3 (Community Edition)
内部版本号 #PC-202.7660.27, 构建于 October 6, 2020

运行时版本: 11.0.8+10-b944.34 amd64
JVM: OpenJDK 64-Bit Server VM by JetBrains s.r.o.
技术支持: 开源软件
```

圖 3-3 PyCharm 版本

第四節 分析網頁架構

一、Momo 的網頁架構

可以發現每個商品都包在 class 為 goodsItemLi 的 li 區塊裡面，而商品名稱在 h3 這個標籤(tag)裡面的文字(text)，商品價格在 class 為 price 的 b 區塊裡面，而網址均為'<http://m.momoshop.com.tw>'加上 a 區塊的 href。

```
<article class="prdListArea">
  <ul>
    <li class="goodsItemLi"> == $0
      <input type="hidden" name="stock" value="2">
      <input type="hidden" name="edmListBackgroundUrl" value>
      <input type="hidden" name="edmCardBackgroundUrl" value>
      <input type="hidden" name="desColor5" value>
      <a href="/goods.momo?i_code=7706608&mdiv=searchEngine&oid=1_1&kw=GA-110GB-1A" title="【CASIO 卡西歐】
      眼黑金重機造型雙顯手錶(GA-110GB-1A)">
        <div class="prdImgWrap">...</div>
        <div class="prdInfoWrap">
          <div class="edmbox">...</div>
          <p class="prdEvent">防水200米</p>
          <h3 class="prdName">
            【CASIO 卡西歐】G-SHOCK 耀眼黑金重機造型雙顯手錶(GA-110GB-1A)</h3>
          <p class="priceArea">
            <span class="priceSymbol">
              <b id="priceSymbol">$</b>
              <b class="price">3,230</b>
              <b class="priceText"></b>
            </span>
            <span class="discountArea">
              ...
            </span>
          </p>
        </div>
        <input type="hidden" id="viewProdId" name="viewProdId" value="7706608">
      </a>
      <table>...</table>
    </li>
    <li class="goodsItemLi">...</li>
  </ul>
</article>
```

圖 3-4 Momo 購物網 網頁架構 資料來源:[11]

二、Yahoo 購物中心的網頁架構

可以發現每個商品都包在 class 為 BaseGridItem__grid__2wuJ7
BaseGridItem__multipleImage__37M7b，標籤(tag) 為 li 的區塊，商品名稱在
claas 為 BaseGridItem__title__2Hwui，標籤(tag)為 span 區塊，商品價格在 em 這
個標籤(tag)裡面，商品網址為 a 區塊裡的 herf。

```
<li class="BaseGridItem__grid__2wuJ7 BaseGridItem__multipleImage__37M7b">...</li>
<li class="BaseGridItem__grid__2wuJ7 BaseGridItem__multipleImage__37M7b">...</li>
<li class="BaseGridItem__grid__2wuJ7 BaseGridItem__multipleImage__37M7b">
  <a href="https://tw.buy.yahoo.com/gdsale/CASIO-G-SHOCK-重機裝備-雙顯運動錶-GA-110GB-1A-黑x金-51mm-9089516.html">
    <div class="BaseGridItem__content__3LORP BaseGridItem__hover__3U1CS">
      <div>...</div>
      <span class="BaseGridItem__itemInfo__3E5Bx">
        <span class="BaseGridItem__title__2Hwui">CASIO G-SHOCK 重機裝備 雙顯運動錶(GA-110GB-1A)黑x金/51mm</span> == $0
        <em class="BaseGridItem__price__3ljkj">$3,528</em>
        <span class="BaseGridItem__tagList__29yes">...</span>
        <div class="BaseGridItem__bottomSpaceRight__3HRQ- BaseGridItem__bottom__2kCvU">...
      </div>
    </span>
  <::after
</div>
```

圖 3-5 Yahoo 購物中心 網頁架構 資料來源：[12]

第五節 利用 Python 的網路爬蟲爬取資料

一、套件

首先將網路爬蟲所需的套件匯入，本研究所用到的套件有「requests、time、json、os、bs4 的 BeautifulSoup」。程式碼如圖 3-6。

```
import requests
import time
import json
import os
from bs4 import BeautifulSoup
```

圖 3-6 套件程式碼

二、反爬蟲機制

本研究爬取的 Momo 購物網有反爬蟲機制，Yahoo 購物中心則沒有，因此 Momo 購物網的 requests.get 需要給他一個 headers，讓網頁認為程式為一般網頁瀏覽者，而 Yahoo 購物中心僅需要 get 程式所給他的網址即可。我們會將網址定義為 url，抓取的網站定義為 resp，並給定網頁編碼為 utf-8，再把網頁所抓取到的文字定義為 soup。程式碼如圖 3-7、3-8。

```

def search_momo():

    url = "http://m.momoshop.com.tw/search.momo?searchKeyword=" \
          "GA-110GB-1A&couponSeq=&cpName=&searchType=1&cateLevel=" \
          "-1&cateCode=-1&ent=k&_imgSH=fourCardStyle"

    headers = {'User-Agent': 'mozilla/5.0 (Linux; Android 6.0.1; '
                              'Nexus 5x build/mtc19t applewebkit/537.36 (KHTML, like Gecko) '
                              'Chrome/51.0.2702.81 Mobile Safari/537.36'}

    resp = requests.get(url, headers=headers)
    if not resp:
        return []
    resp.encoding = 'utf-8'
    soup = BeautifulSoup(resp.text, 'html.parser')

```

圖 3-7 Search-Momo 購物網

```

searchP =str(input("請輸入你想搜尋的產品"))
def search_yahoo():

    url = "https://tw.buy.yahoo.com/search/product?p="+searchP
    resp = requests.get(url)
    if not resp:
        return []
    resp.encoding = 'utf-8'
    soup = BeautifulSoup(resp.text, 'html.parser')

```

圖 3-8 Search-Yahoo 購物中心

三、抓取資料

程式抓取完給定的網址後，再來就是建造一個清單(list)，準備將抓取到的物件存取，要抓取網頁的資訊就會用到第四節所述的網頁架構碼，為了方便瀏覽者一眼看清所有資訊，此研究會將商品名稱、價格、網址丟到前面所建造的清單(list)。程式碼如圖 3-9、3-10。

```
items=[]
for product in soup.find_all("li", "goodsItemLi"):

    item_name = product.find('h3').text.strip()
    item_price = product.find('b', 'price').text.strip()
    if not item_price:
        continue
    item_url = 'http://m.momoshop.com.tw' + product.find('a')['href']

    item = {
        'name': item_name,
        'price': item_price,
        'url': item_url,
    }
    items.append(item)
return items
```

圖 3-9 Momo 購物網之商品名稱、價格、網址

```
items = []
for product in soup.find_all("li", "BaseGridItem__grid___2wuJ7 "
                             "BaseGridItem__multipleImage___37M7b"):

    item_name = product.find('span', "BaseGridItem__title___2HWui").text.strip()
    item_price = product.find('em').text.strip()
    if not item_price:
        continue
    item_url = product.find('a')['href']

    item = {
        'name': item_name,
        'price': item_price,
        'url': item_url,
    }
    items.append(item)
return items
```

圖 3-10 Yahoo 購物中心之商品名稱、價格、網址

四、資料封包

存取完網頁的資訊後，再來就是要做資料的封包，程式碼中的 print，是讓開發者在編譯器中先看到所抓取的資料是否正確，並將每天所抓取到的資料存取成 json 檔，裡面的內容會含有日期、爬取的網站和所抓取的資料，並判斷是否含有 momo、yahoo 的資料夾，若沒有，則建立一個資料夾，並把 json 檔存放。程式碼如圖 3-11、3-12。

```
if __name__ == '__main__':  
  
    items = search_momo()  
    today = time.strftime('%Y-%m-%d')  
    print('%s 搜尋 %s 共 %d 筆資料' % (today, "GA-110GB-1A", len(items)))  
    for i in items:  
        print(i)  
    data = {  
        'date': today,  
        'store': 'momo',  
        'items': items  
    }  
  
    path="momo"  
    if not os.path.isdir(path):  
        os.mkdir(path)  
  
    with open(os.path.join('momo', today+"momo爬蟲專案"), 'w', encoding='utf-8') as f:  
        json.dump(data, f, indent=2, ensure_ascii=False) #將json轉STR
```

圖 3-11 存取 Momo 購物網之資料

```

if __name__ == '__main__':

    items = search_yahoo()
    today = time.strftime('%Y-%m-%d')
    print('%s 搜尋 %s 共 %d 筆資料' % (today, searchP, len(items)))
    for i in items:
        print(i)
    data = {
        'date': today,
        'store': 'yahoo',
        'items': items
    }

    path="yahoo"

    if not os.path.isdir(path):
        os.mkdir(path)

    with open(os.path.join('yahoo', today+"yahoo爬蟲專案"), 'w', encoding='utf-8') as f:
        json.dump(data, f, indent=2, ensure_ascii=False) #將json轉STR

```

圖 3-12 存取 yahoo 購物中心之資料

第四章 研究結果

本章節分為三個階段，分別是通過反爬蟲系統、資料整合、分析價格變動。

第一節 通過反爬蟲系統

本研究以 Momo 購物網及 Yahoo 商城作為對比，前者有設置反爬蟲，因此在請求抓取網路前，需要設置一個 headers，而 headers 也可以在網頁開發者工具中尋找。

```
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
me/87.0.4280.88 Safari/537.36
```

圖 4-1 User-Agent

```
url = "http://m.momoshop.com.tw/search.momo?searchKeyword=" \
      "GA-110GB-1A&couponSeq=&cpName=&searchType=1&cateLevel=" \
      "-1&cateCode=-1&ent=k&_imgSH=fourCardStyle"

headers = {'User-Agent': 'mozilla/5.0 (Linux; Android 6.0.1; '
                        'Nexus 5x build/mtc19t applewebkit/537.36 (KHTML, like Gecko) '
                        'Chrome/51.0.2702.81 Mobile Safari/537.36'}

resp = requests.get(url, headers=headers)
```

圖 4-2 有給定 headers 之程式碼-Momo

```
url = "https://tw.buy.yahoo.com/search/product?p="+searchP
resp = requests.get(url)
```

圖 4-3 無給定 headers 之程式碼-Yahoo

由圖 4-2 所示，在 requests.get 之後，還有給定一個 headers，而 headers 是由網頁開發者工具中所取得的，如圖 4-1。若在爬取 Momo 購物網時，沒有給定 headers，程式則無法成功運行，如圖 4-4。

```
raise ConnectionError(err, request=request)
requests.exceptions.ConnectionError: ('Connection aborted.', ConnectionResetError(10054, '遠端主機已強制關閉一個現存的連線。', None, 10054, None))
```

圖 4-4 程式運行失敗

第二節 資料整合

檔名輸出結果為「年-月-日-商城名稱-爬蟲專案」副檔名為 json，如圖 4-5，

其文件內容如圖 4-6、4-7。

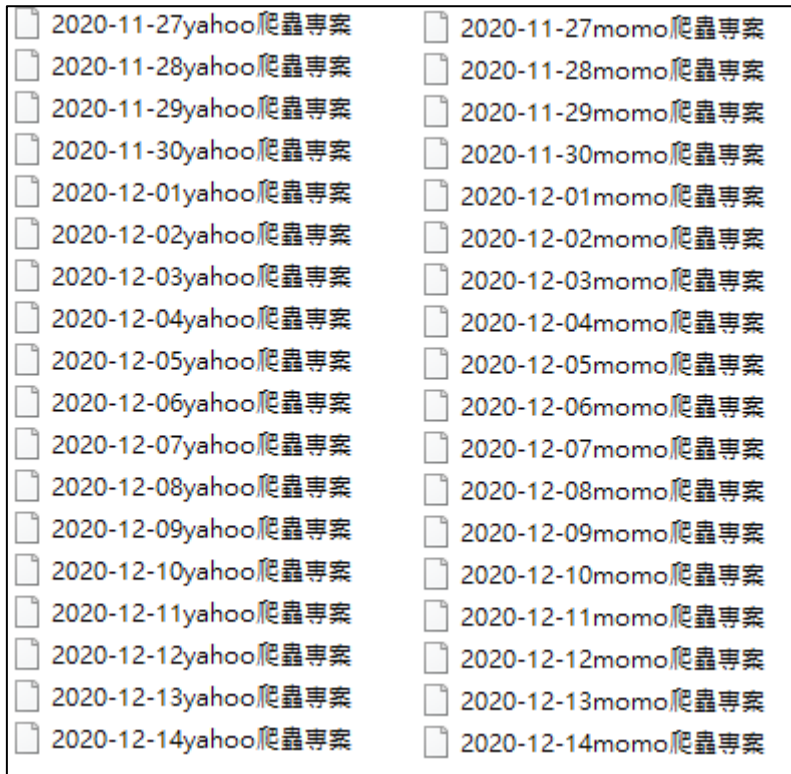


圖 4-5 檔名

程式輸出結果可由圖 4-6、4-7 得知，每個 Json 檔裡面包含了日期及爬取商城，物件含有名稱、價格及商品網址，且經由程式建立的清單，能把各項商品明確的區隔，方便瀏覽者一一查看。

```

"date": "2020-12-14",
"store": "momo",
"items": [
  {
    "name": "【CASIO 卡西歐】G-SHOCK 耀眼黑金重機造型雙顯手錶(GA-110GB-1A)",
    "price": "$3,230",
    "url": "http://m.momoshop.com.tw/goods.momo?i_code=7706608&mdiv=sea
  },
  {
    "name": "【CASIO 卡西歐】G-SHOCK 金燦重機雙顯手錶(GA-110GB-1A)",
    "price": "$4,900",
    "url": "http://m.momoshop.com.tw/goods.momo?i_code=6990561&mdiv=sea
  },
  {
    "name": "【CASIO 卡西歐】G-SHOCK 低調奢華男錶-黑X金(GA-110GB-1A)",
    "price": "$4,900",
    "url": "http://m.momoshop.com.tw/goods.momo?i_code=4400504&mdiv=sea
  },
  {
    "name": "【CASIO 卡西歐】經典黑金重機雙顯電子錶(黑/金 GA-110GB-1A)",
    "price": "$4,900",
    "url": "http://m.momoshop.com.tw/goods.momo?i_code=7601870&mdiv=sea
  },

```

```

"date": "2020-12-13",
"store": "yahoo",
"items": [
  {
    "name": "CASIO卡西歐 G-SHOCK 雙顯系列 GA-110GB-1A_51.2mm",
    "price": "$4,900",
    "url": "https://tw.buy.yahoo.com/gdsale/CASIO-卡西歐-GA-110GB-1A-8659791.html"
  },
  {
    "name": "G-SHOCK 變形金剛黑金重型休閒錶(GA-110GB-1A)-黑/51.2mm",
    "price": "$3,176",
    "url": "https://tw.buy.yahoo.com/gdsale/-4263206.html"
  },
  {
    "name": "CASIO卡西歐 經典黑金配色G-SHOCK系列(GA-110GB-1A)",
    "price": "$3,380",
    "url": "https://tw.buy.yahoo.com/gdsale/CASIO卡西歐-GA-110GB-1A-9037500.html"
  },
  {
    "name": "CASIO卡西歐 G-SHOCK 耀眼黑金重機造型雙顯手錶(GA-110GB-1A)",

```

圖 4-6、4-7 檔案內容

第三節 分析價格變動

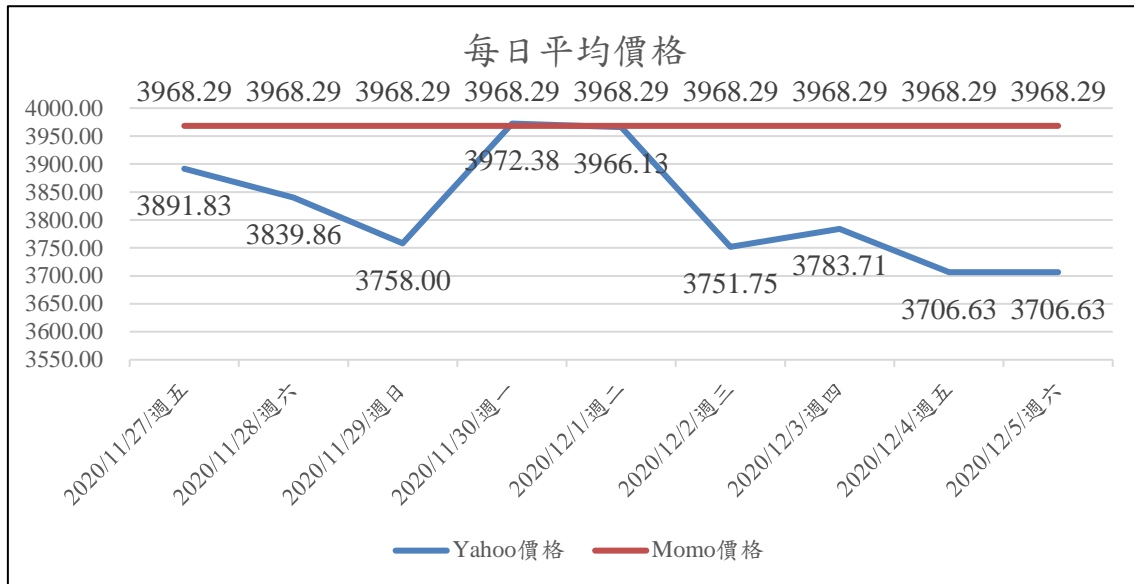


圖 4-8 平均價格折線圖

圖 4-8 是 2020 年 11 月 27 日到 2020 年 12 月 5 日的商品平均價格折線圖，可觀察出 Momo 購物網對商品沒有做出任何價格調整，平均價格皆維持為 3968，而 Yahoo 商城是每日都有變動，最低平均價達到了 3706.63 元，且除了 11/30 以外，其餘天數的價格皆低於 Momo 購物網。

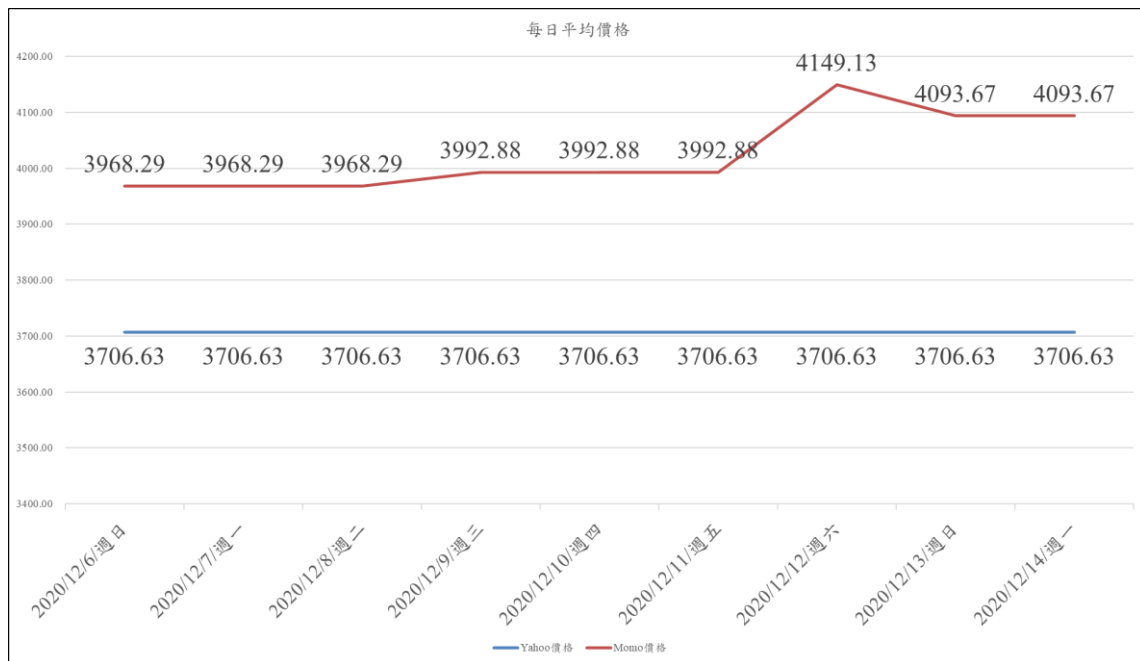


圖 4-9 平均價格折線圖

圖 4-8 是 2020 年 11 月 27 日到 2020 年 12 月 5 日的商品平均價格折線圖，可觀察出 Yahoo 商城，已經沒有做出任何價格變動，而是持續維持在最低點 3706.63 元，而 Momo 購物網則是開始做出些微的變動，但並非調降價格吸引顧客，而是做出些微的調漲，2020 年 12 月 12 日，是購物網站的大節慶，各大購物網站都會做出多種折扣、降價，但本研究發現在 2020 年 12 月 12 日當天，Yahoo 商城及 Momo 購物網的 G-Shock-GA-110GB-1A 不僅沒有做出價格調降的策略，Yahoo 商城的商品平均價格竟然是漲價到了最高點 4149.13 元。

組別統計量					
網站		個數	平均數	標準差	平均數的標準誤
價格	.00	140	3760.6071	550.11036	46.49281
	1.00	134	4000.3284	610.32352	52.72392

獨立樣本檢定										
		變異數相等的 Levene 檢定		平均數相等的 t 檢定						
		F 檢定	顯著性	t	自由度	顯著性 (雙尾)	平均差異	標準誤差異	差異的 95% 信賴區間	
									下界	上界
價格	假設變異數相等	7.072	.008	-3.418	272	.001	-239.72122	70.13532	-377.79830	-101.64413
	不假設變異數相等			-3.410	266.231	.001	-239.72122	70.29505	-378.12615	-101.31628

圖 4-10 獨立樣本 T 檢定

本研究利用 SPSS 對所抓取到的每日價格進行分析，其中 Yahoo 商城的資料個數有 140 項，平均價格為 3760.6071 元，標準差為 550.11036，平均數的標準差誤為 46.49281，而 Momo 購物網有 134 項，平均價格為 4000.3284 元，標準差為 610.32352，平均數的標準差誤為 52.72392，參考 T 檢定 $P < 0.05$ ，可得知此二樣本 F 檢定後的結果，顯著性 p 值為 0.008，小於 0.05，兩組變異數有顯著性差異，需要修正 T 統計值，故需參考不假設變異數相等，計算後的 t 統計值為 -3.418，雙尾顯著性 p 值 = 0.001 < 0.05，拒絕虛無假設，可得知 Yahoo 商城與 Momo 購物網的價格由顯著性差異。

第五章 結論

本章節將敘述本研究最終得知之結論與建議及未來展望。

第一節 結論

隨著網際網路及程式語言的卓越進步，網路已有非常多資料可供人們查閱，且有許多不同的方式可以將資料下載取用，在眾多的程式語言中，本研究選擇了相對容易撰寫的 Python 語言，在研究過程中發現，Python 不僅語言淺顯易懂，程式裡提供的套件也十分多元，使用官方提供之套件，便能做出許多變化，另外對於官方套件，使用者也能將其定義為自己習慣的變數名稱，方便使用者本人撰寫，此語言之使用者不僅在近年來快速竄升，在未來幾年依然會佔據眾多程式語言中的一席之地。

則本研究是利用 Python 的網路爬蟲來擷取 Momo 購物網及 Yahoo 商城的商品「G-Shock-GA-110GB-1A」。取得兩家商城每日的商品價格，並做出商品折線圖來觀察，依獨立樣本 T 檢定觀察，可得知 Yahoo 商城之價格有明顯的價格優勢，且依折線圖可觀察到價格變化量也相對較高，受限於購物網站之銷售量屬於商業機密，本研究無從得知二購物網站之消費人數為何，但若僅以價格來判斷，消費者是較有可能選擇在 Yahoo 商城來選擇購買 G-Shock-GA-110GB-1A。

一般消費者未使用爬蟲程式擷取逐日價格，則無法判斷商品價位均價為何，與賣家之間相對地處於弱勢，若能取得商品每日價格並匯出圖表，消費者便能推算何種價位屬於相對低點，有利於自身購買商品。

第二節 建議

- 一、本研究發現 Yahoo 商城相對於 Momo 購物網，價格變化量較大，或許會使對價格較為敏感之消費者產生較大的瀏覽興趣，因價格均無變化，便無法對消費者產生每日比價之舉動，若消費者逐日觀察相同商品之價格變化，便有機會使其瀏覽其餘商品，方能帶動網站瀏覽量並促進銷售。
- 二、2020 年 12 月 12 日，是購物網站的大節慶，各大購物網站都會做出多種折扣、降價，但此二購物網站在雙 12 節慶皆未對本研究的產品做出價格折扣，若能使其產品降價，勢必能帶來更高之銷售量。

第三節 未來展望

在未來熟悉初階的程式語言是必然，Python 的網路爬蟲，屬於相對入門的人工智慧，在未來人工智慧是極大的趨勢，而 Python 也是許多人撰寫人工智慧的工具之一，能使用 Python 撰寫的東西有常見的神經網路：迴歸網路、分類網路、迴圈神經網路（RNN），卷積神經網路(CNN)等等...

近年受 iPhone 智慧型手機的影響，人臉辨識的話題性也被廣泛討論，其 Face ID 也讓許多人注意到了人臉辨識這方面的技術，同時也為手機上游元件廠商有了極大的商機，在未來也能在醫療、金融、國安等領域機構應用人臉辨識來確認身分及提供服務，因此人臉辨識勢必在未來有著極大的商機，而 Python 也受惠於此。

第六章 參考文獻

Hello Python！Python 入門詳細介紹(MAPE Academy，2018) [1]：

<https://medium.com/python4u/hello-python-509eabe5f5b1>

Python 是什麼？不可不知的 Python 優缺點及發展前景(巨匠電腦，2020) [2]：

<https://www.pcschool.com.tw/blog/it/what-is-python>

HAPPY CODING 快樂學程式 (2018) [3]：

<https://www.happycoding.today/posts/22>

台糖通訊-資訊補給站-Python 爬蟲應用——以台糖易購網為例 (王柏宋，2020)

[4]：

<https://www.taisugar.com.tw/monthly/CPN.aspx?ms=1455&p=13387803&s=13387828>

python 網路爬蟲簡介 Even(2019) [5]：

<https://freelancerlife.info/zh/blog/python-web-scraping-overview/>

認識網路爬蟲：解放複製貼上的時間 (吳致賢，2016) [6]：

<https://pala.tw/what-is-web-crawler>

[Python 教學]Request 和 BeautifulSoup 爬蟲教學，初學者也可以馬上學會！

(Kevin，2020) [7]：

<https://medium.com/@zx2515296964/python-%E6%95%99%E5%AD%B8-%E7%B0%A1%E5%96%AE%E5%B9%BE%E6%AD%A5%E9%A9%9F-%E8%AE%93%E4%BD%A0%E8%BC%95%E9%AC%86%E7%88%AC%E8%9F%B2-928a816051c1>

HTML 語法教學，快速攻略網頁 HTML 標籤的基本元素 (ALPHA Camp，2020)

[8]：

<https://tw.alphacamp.co/blog/html-guide>

HTML 基礎 [9]：[https://developer.mozilla.org/zh-](https://developer.mozilla.org/zh-TW/docs/Learn/Getting_started_with_the_web/HTML_basics)

[TW/docs/Learn/Getting_started_with_the_web/HTML_basics](https://developer.mozilla.org/zh-TW/docs/Learn/Getting_started_with_the_web/HTML_basics)

網頁設計小知識：HTML 是什麼？（達格網頁設計公司，2019）[10]：

<https://www.targets.com.tw/%E7%B6%B2%E9%A0%81%E8%A8%AD%E8%A8%88%E6%96%B0%E7%9F%A5%E6%96%87%E7%AB%A0/41/%E7%B6%B2%E9%A0%81%E8%A8%AD%E8%A8%88%E5%B0%8F%E7%9F%A5%E8%AD%98%EF%B C%9AHTML%E6%98%AF%E4%BB%80%E9%BA%BC%EF%BC%9F>

MOMO 購物網 [11]：

<http://m.momoshop.com.tw/main.momo>

YAHOO 商城[12]：

<https://tw.buy.yahoo.com/>

電子商務是什麼？做電商前的 3 大思考關鍵和成功要素（Doris Lin）[13]：

<https://transbiz.com.tw/%E9%9B%BB%E5%AD%90%E5%95%86%E5%8B%99%E6%98%AF%E4%BB%80%E9%BA%BC%EF%BC%9Fecommerce-sucess/>

電子商務是什麼？開店前一定要知道的 3 個電商成功關鍵（SHOPLINE 電商教室 Josefin，2020）[14]：

<https://blog.shopline.tw/what-is-ecommerce-and-how-to-succeed/>

Python 維基百科 [15]：

<https://zh.wikipedia.org/wiki/Python>

python 系列之反爬蟲介紹（python 新視野，2019） [16]：

<https://kknews.cc/code/gm4xvvl.html>

真心話大公開！最符合世界趨勢的程式語言-Python（Uder，2018）[17]：

<https://www.coding543.com/%E7%9C%9F%E5%BF%83%E8%A9%B1%E5%A4%A7%E5%85%AC%E9%96%8B%E6%9C%80%E7%AC%A6%E5%90%88%E4%B8%96%E7%95%8C%E8%B6%A8%E5%8B%A2%E7%9A%84%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80-python/>

Stackoverflow：<https://stackoverflow.com/>