數據分析作業一 : 電商資料爬蟲 – 以 momo 為例

作業內容 :

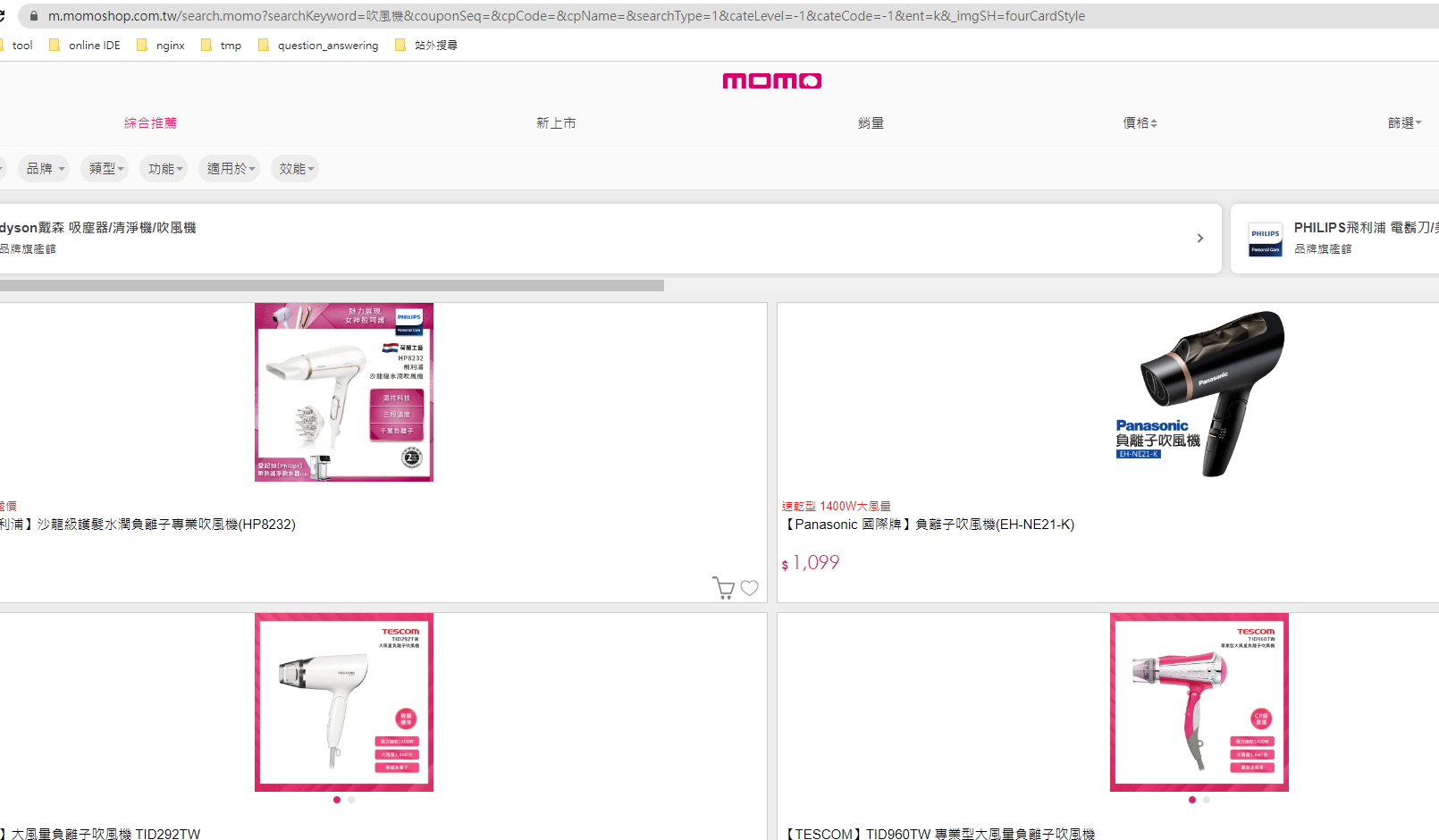
數據分析的第一步是蒐集資料，我們需要豐富的資料才能做進一步的分析，而現今，爬蟲技術是快速蒐集資料的方法之一，如何在資料豐富的網路世界裡蒐集自己需要的資料，成為了我們的一大課題。綜上所述，第一個數據分析的作業希望同學試著學習爬蟲的技術，請同學利用程式爬取momo網站上的商品資料，並依照規則蒐集資料，最後儲存成excel檔案。

爬蟲規則描述 :

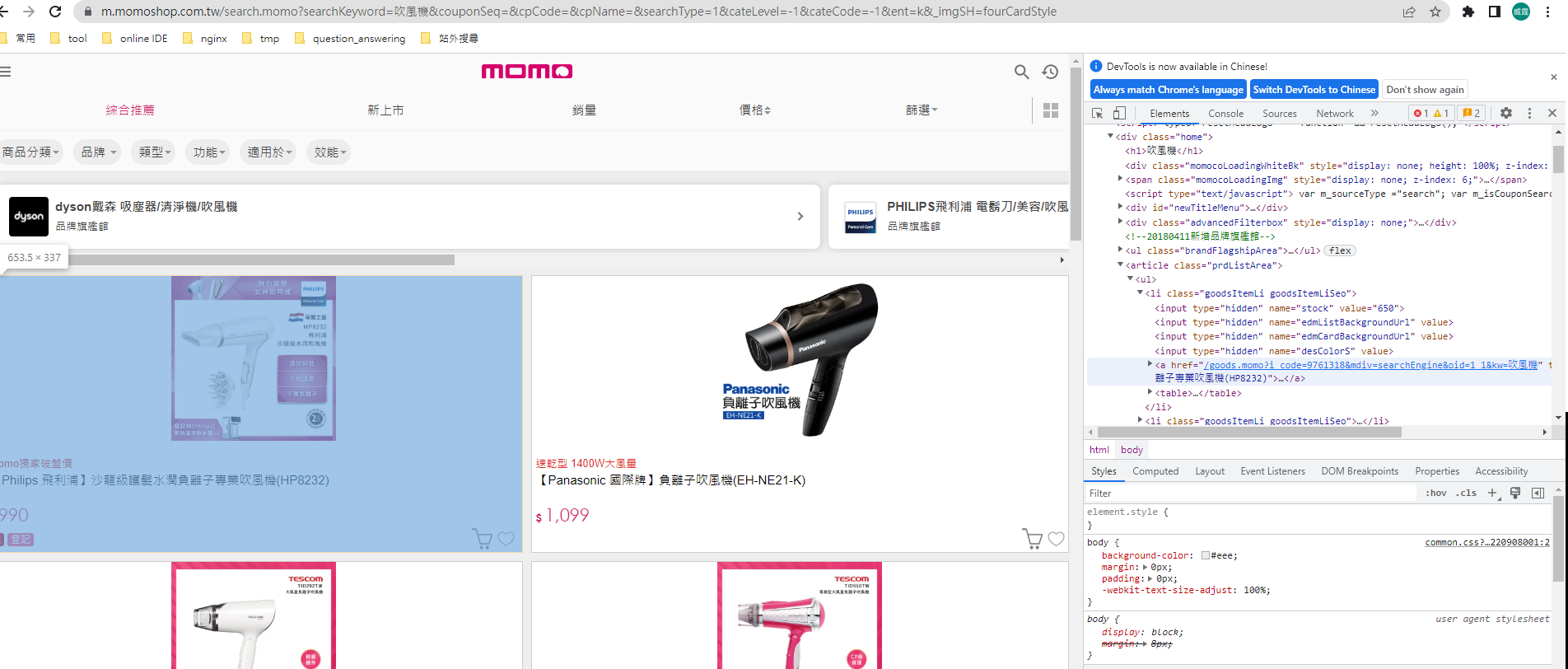
1. 在momo的首頁左方，有商品類別的14個分類，請根據各個分類的主題想出10個搜尋詞當作輸入，例如，「家電、電玩、視聽、廚三家」的分類可以用吹風機當作搜尋詞。



1. 我們可以發現在搜尋時，momo的網址其實有一定的規律，也就是在網址的searchKeyword 後面是帶著我們的搜尋詞的。我們可以利用此規則，對我們所想要搜尋的搜尋詞進行request，就可以得到以下搜尋結果。



1. 更進一步的，我們希望同學從搜尋到的頁面中，尋找每個商品的實際商品頁面，我們可以利用beautiful soup從request下來的頁面找到各個商品的實際網址，並再次利用request搜尋該網址。



3

2

1

1. 在商品頁面上我們希望同學們將以下紅色圈出的網頁資料利用beautiful soup爬取，並依照規定格式進行儲存(請參照範例excel)，紅色內的內容為每個商品都有的，橘色則不一定每個商品的敘述都一樣。



這是商品類別(cate)

這是價錢(price)

這是特價(amount)

這是商品名(title)

這是詳細的商品描述(desc)

1. 搜尋詞會找到許多的，請將前5頁的所有商品依照上述方式爬取下來。



1. 我們會另外給一份雲端的excel檔紀錄同學使用的搜尋詞，請同學在決定搜尋詞之前，先看一下有沒有與其他同學重複，並在決定之後填上自己的學號與搜尋詞。搜尋詞請發揮想像力，希望不要與其他人重複，並且搜尋的詞條可以廣泛一點，例如「智慧型手機」，而不是「華碩 ROG 智慧型手機 白色」。
2. 根據上述題目描述，我們有提供挖空的程式範本，請同學根據題意完成程式碼。
3. 作業繳交位置與時限 : 請至課程 moodle 作業一繳交區繳交作業，請繳交程式碼、14個分類的爬蟲結果 excel 檔案。作業繳交時限到10/4晚上23:59，逾期不收。
4. 備註 : 14個分類在電腦版網頁可以看到，實際爬蟲請在手機版網頁上爬蟲，手機版的html比較乾淨