

Web Scraper

網路爬蟲的網站地圖

C-1 爬取各類資訊

在書附範例的「Ch08\info」目錄下提供 13 個網路爬蟲的網站地圖，可以讓 Web Scraper 幫助我們爬取球賽、氣象、工作、Apps、美食和論文資料等各類資訊。

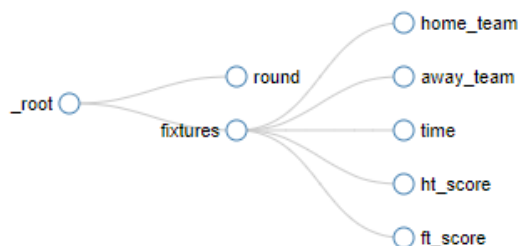
英國足球聯賽比賽成績和投注賠率：bet365_fixtures.txt、
oddsportal_premier_league.txt

在 bet365 網站可以查詢英國足球聯賽的比賽成績，其 URL 網址如下所示：

<https://s5.sir.sportradar.com/bet365/en/1/season/54571/fixtures>

Round 38						
12/05/19				HT	FT	
22:00	⚽	Leicester	-	Chelsea	0 : 0	0 : 0
22:00	⚽	Fulham	-	Newcastle	0 : 2	0 : 4
22:00	⚽	Crystal Palace	-	Bournemouth	3 : 1	5 : 3
22:00	⚽	Burnley	-	Arsenal	0 : 0	1 : 3
22:00	⚽	Brighton	-	Man City	1 : 2	1 : 4
22:00	⚽	Watford	-	West Ham	0 : 2	1 : 4
22:00	⚽	Liverpool	-	Wolverhampton	1 : 0	2 : 0
22:00	⚽	Man Utd	-	Cardiff	0 : 1	0 : 2

上述網頁是使用 HTML 表格顯示 18-19 年的 38 回合的比賽成績，因為 HTML 表格沒有完整的標題列，所以網站地圖 bet365_fixtures.txt 是使用 Element 類型來爬取 HTML 表格，如下圖所示：

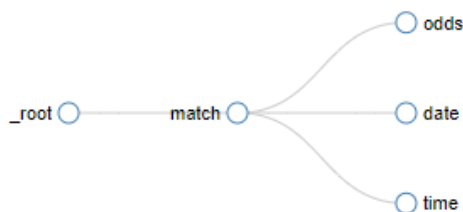


上述 `round` 節點是 `Text` 類型的回合數，`fixtures` 節點是 `Element` 類型的每一列比賽記錄，在之下有 5 個 `Text` 節點的欄位，依序取得主場、客場、時間、半場成績和全場成績。

網站地圖：`oddsportal_premier_league.txt` 是爬取 `OddsPortal.com` 網站的投注賠率，其 URL 網址如下所示：

<https://www.oddsportal.com/soccer/england/premier-league/>

上述網頁也是使用 `HTML` 表格來顯示各種比賽組合，每一種組合是一個超連結，點選後，可以顯示各大投注網站的賠率表格，其網站地圖如下圖所示：



Ch08-1-03.tif

上述 `match` 節點是 `Link` 類型的多筆超連結，`odds` 節點是 `Table` 類型的賠率表格，下方 2 個 `Text` 類型 `date` 和 `time` 節點是日期和時間。

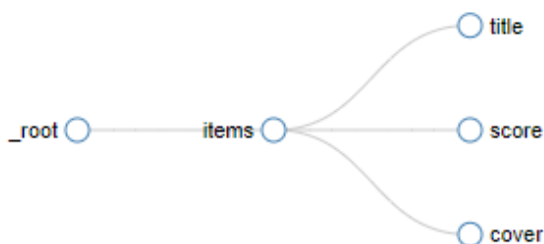
豆瓣電影的熱門影片：douban_movies_more.txt

豆瓣電影的熱門影片是使用表格方式來顯示影片劇照清單，其 URL 網址如下所示：

https://movie.douban.com/explore#!type=movie&tag=%E7%83%AD%E9%97%A8&sort=recommend&page_limit=20&page_start=0



上述網頁使用巢狀<div>標籤來顯示表格形式的影片清單，屬下一種更多按鈕巡覽的網站，在網站地圖是使用 Element click 類型 items 節點來爬取每一部影片的記錄，2 個 Text 類型 title 和 score 節點是影片名稱和分數，cover 節點是 Image 類型的劇照圖片，其網站地圖如下圖：



全國電子的各區門市資料：elifemall_shop.txt

在全國電子的官方網站可以查詢各縣市的門市資料，其 URL 網址如下所示：

<https://www.elifemall.com.tw/allnewweb/store.php?depno=02>

郵遞區號	門市編號	門市	住址	電話	營業時間	地圖	Google
100	02193	Digital City 和平店	台北市中正區和平西路一段20號	02-23658868	上午11:00 晚上10:00	查看	
100	02136	新生南門市	台北市中正區新生南路一段102號	02-23583578	上午11:00 晚上10:00	查看	
103	02002	酒泉門市	台北市大同區酒泉街111、113、115號	02-25997278	上午11:00 晚上10:00	查看	
104	02090	北安門市	台北市中山區北安路577號	02-85092099	上午11:00 晚上10:00	查看	
104	02180	民權門市	台北市中山區民權東路二段73號	02-25950300	上午11:00 晚上10:00	查看	

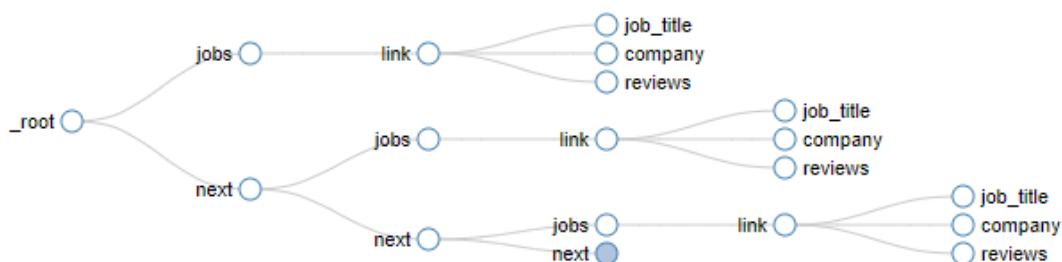
上述網頁是 HTML 表格的分頁資料，我們可以使用 Web Scraper 的 URL 參數範圍來爬取全國各區的門市資料，如下所示：

[https://www.elifemall.com.tw/allnewweb/store.php?depno=\[02-08\]](https://www.elifemall.com.tw/allnewweb/store.php?depno=[02-08])

上述參數範圍是第 2~8 頁的門市資料，在網站地圖是使用 Table 類型的 shops 節點來爬取 HTML 表格資料，其網站地圖如下圖所示：



上述網頁是加州洛杉磯資深專案經理工作的搜尋結果，這是頁碼和下一頁超連結的分頁網站，在網站地圖使用下一頁鈕來爬取所有工作的分頁資料，如下圖所示：

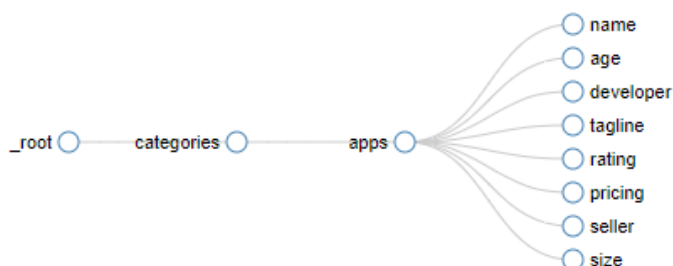


上述 jobs 節點是 Element 類型，next 節點是下一頁的 Link 類型，其父節點都有 2 個【_root, next】，Element 類型的記錄只有 1 個欄位的 Link 類型的 link 節點，用來切換至詳細頁面來爬取工作的相關資訊。

iTunes Preview 的熱門 Apps 清單：itunes_app_store.txt

<https://apps.apple.com/tw/genre/ios/id36>

上述網頁提供多層分類目錄的超連結，在點選後，可以檢視各分類下的熱門 Apps，只需再點選 App 名稱，即可看到 App 詳細資料，網站地圖是使用階層目錄方式來爬取所有 Apps 資料，如下圖所示：



上述 categories 節點是 Link 類型的第 1 層目錄；apps 節點是 Link 類型的第 2 層目錄，然後爬取 App 詳細頁面的應用程式資訊。

爬取 NBA 球員資料：nba_gsw.txt、nba_all.txt

在 Basketball Reference 網站可以查詢 NBA 球員資料，如下所示：

<https://www.basketball-reference.com/teams/GSW/2018.html>

上述網頁是使用 HTML 表格顯示 NBA 勇士隊 2018 年的球員資料，網站地圖：nba_gsw.txt 使用 Table 類型 roster 節點來爬取 HTML 表格資料，如下圖所示：



網站地圖：nba_all.txt 爬取所有 NBA 球隊的球員資料，其 URL 網址如下所示：

<https://www.basketball-reference.com/>

上述網頁是清單和詳細巡覽結構，在清單頁是使用表格顯示 NBA 球隊的超連結清單，點選超連結，可以用 HTML 表格顯示 NBA 球隊的球員資料，其網站地圖如下圖所示：

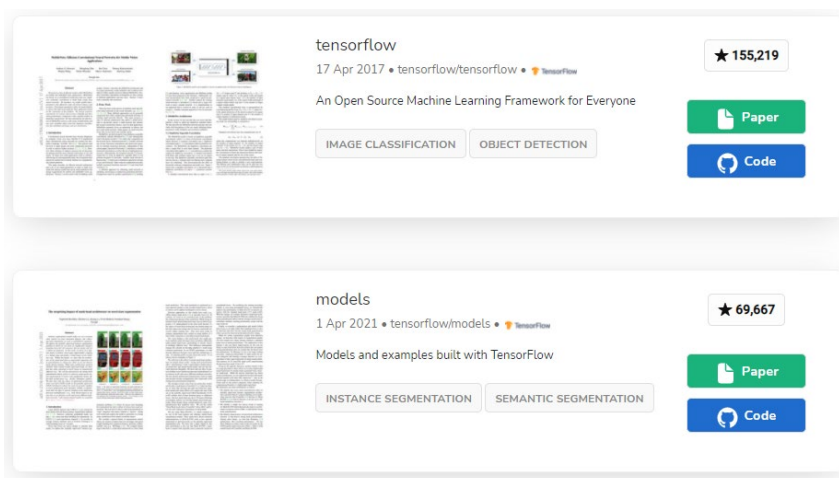


上述 teams 節點是 Link 類型的多筆超連結，member 節點是 Table 類型的球員表格資料。

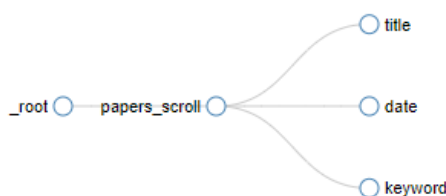
爬取 Paperswithcode 論文資料：paperswithcode.txt

在 Paperswithcode 網站可以搜尋包含程式碼的相關論文資料，其 URL 網址如下所示：

<https://paperswithcode.com/greatest>



上述網頁是<div>標籤的論文清單，屬於捲動頁面巡覽的網站，網站地圖是使用 Element scroll down 類型的 papers_scroll 節點來爬取每一篇論文記錄，3 個 Text 類型的 title、date 和 keyword 節點，可以分別擷取名稱、日期和關鍵字，其網站地圖如下圖所示：



網路的美食資訊和活動

網路美食資訊的 Web Scraper 網站地圖說明，如下表所示：

網站地圖	說明
tokyo_foods.txt	美食餐廳地圖的東京美食資訊，屬於頁碼分頁的資料爬取
yelp_events_nyc.txt	Yelp 網站的 New York 美食活動清單，屬於下一頁分頁的資料爬取

請注意！上述東京美食資訊的分頁超連結位置在左下角，剛好被 Web Scraper 選擇器工具列蓋住，請改用開發人員工具來取得分頁的 CSS 選擇器。

國內各城市的天氣資訊

天氣資訊的 Web Scraper 網站地圖說明，如下表所示：

網站地圖	說明
taiwan_weather.txt	中央氣象局的天氣資訊，點選地圖的超連結，就可以顯示此城市的天氣資訊，所以是使用 Element click 類型來點選地圖的城市
yahoo_weather.txt	Yahoo 網站的天氣資訊，屬於清單與詳細頁面巡覽結構