主題：中職球隊近況查詢器

選擇網站：中華職棒大聯盟全球資訊網 http://www.cpbl.com.tw/cpbl.html

為何選擇此網站進行爬蟲？

在台灣，觀看棒球比賽是許多人所享受的娛樂之一，許多人也會有著自己熱愛、支持的球隊。然而，若身為某隊的球迷，自然就會想要獲取更多有關球隊的資訊。

但在中職官網中，對於只想了解球隊近況的球迷充斥著太多不必要的資訊，若想要完整看到所支持球隊近況，及所支持球隊球員表現，更是需要點擊許多地方，才能零碎的蒐集完所有資料。因此，利用網路爬蟲，即可將當下的資料一網打盡，並整理好，完整的輸出。

透過哪些方法爬到資料？

資料的爬取主要透過 python 的兩個 library：

1. requests (取得網頁 html)
2. beautifulsoup (對 html 進行檢索)

針對網頁特定元素，找出對應 id、class 或 tag，並將所需資料存進合適的資料結構，以利結果輸出。

下頁將以範例說明。

● 爬取球隊球員成績

如下圖，在中職的網站中，可看到選手各項成績的排行榜，但除了聯盟排名，有時候球迷也會想了解，所支持球隊球員各成績的排名，並能隨著賽季的進行，而持續更新。

投手TOP5

ERA	W	SV	H	SO

1	索沙 (富邦)	1.72
2	羅力 (富邦)	1.89
3	萊福力 (中信兄弟)	2.63
4	尼克斯 (Lamigo)	2.69
5	潘威倫 (統一獅)	2.98

打擊TOP5

AVG	H	HR	RBI	SB

1	林泓育 (Lamigo)	0.367
2	朱育賢 (Lamigo)	0.361
3	藍寅倫 (Lamigo)	0.343
4	郭嚴文 (Lamigo)	0.325
5	胡金龍 (富邦)	0.319

對於各球隊球員，中華職棒提供所有相關數據(如下圖)，因此我們主要目標事將這些數據先爬下來，再找出各項目排行的前幾名。

NAME	TEAM	G	PA	AB	RBI	R	H	1B	2B	3B	HR	TB	SO	SB	OBP	SLG	AVG
蘇智傑	統一7-ELEVEn	44	182	157	24	22	40	19	12	1	8	78	31	5	0.352	0.497	0.255
陳重羽	統一7-ELEVEn	40	142	133	13	19	38	30	6	0	2	50	22	4	0.321	0.376	0.286
楊家維	統一7-ELEVEn	39	136	120	16	11	29	22	4	0	3	42	27	1	0.296	0.350	0.242
潘武雄	統一7-ELEVEn	39	150	134	21	17	40	28	6	1	5	63	25	1	0.373	0.470	0.299
唐肇廷	統一7-ELEVEn	36	93	81	2	10	19	15	3	0	1	25	15	1	0.333	0.309	0.235
高國慶	統一7-ELEVEn	33	98	84	10	10	20	13	4	0	3	33	18	0	0.327	0.393	0.238
陳鏞基	統一7-ELEVEn	32	116	104	12	21	27	17	5	0	5	47	22	4	0.336	0.452	0.260
林祐樂	統一7-ELEVEn	32	91	87	11	4	18	16	1	0	1	22	17	0	0.222	0.253	0.207
郭阜林	統一7-ELEVEn	28	97	90	7	7	15	9	4	0	2	25	30	0	0.216	0.278	0.167
林祖傑	統一7-ELEVEn	24	76	69	9	11	17	11	4	1	1	26	17	2	0.280	0.377	0.246
黃恩賜	統一7-ELEVEn	23	66	59	13	9	20	12	7	0	1	30	18	1	0.400	0.508	0.339
潘彥廷	統一7-ELEVEn	21	51	45	3	4	8	5	2	0	1	13	15	2	0.275	0.289	0.178
陳傑憲	統一7-ELEVEn	20	88	78	15	17	29	24	3	0	2	38	5	5	0.432	0.487	0.372
方昶詠	統一7-ELEVEn	12	21	18	0	0	2	2	0	0	0	2	9	3	0.158	0.111	0.111
江亮緯	統一7-ELEVEn	12	37	28	2	9	7	4	1	1	1	13	8	0	0.405	0.464	0.250
蔡奕玄	統一7-ELEVEn	11	23	19	2	2	4	3	0	0	1	7	6	0	0.318	0.368	0.211
鄭緯文	統一7-ELEVEn	11	27	23	2	4	5	2	1	1	1	11	8	1	0.333	0.478	0.217
郭峻偉	統一7-ELEVEn	11	26	22	1	2	4	3	1	0	0	5	3	0	0.280	0.227	0.182

透過檢視網頁的原始碼，我們可以清楚的發現只要讀取<table>中的每一個<tr>及<td>就能得到所有數據，但是這並不完全是最好的處理方式，因為我們的目標是各數據的前幾名。

因此，使用到 python 的 heapq 這個 library，對於所有要排名項目，各建立一個 Priority queue，每讀入一列資料，就將[該項目成績, 球員名稱] 這個結構存入 heapq 中，並計算出當時資料中的最大值，所有球員都讀完後，即能 pop 出前 3 名的球員，並有效率的得出結果，而不用進行所有資料的排序。

```
soup = BeautifulSoup(res.text, 'html.parser')
table = soup.table
row = table.find_all('tr')
c = 0
for r in row:
    c += 1
    if c > 1:
        data = r.find_all('td')
        j = 0
        for d in data:
            j += 1
            if j == 1:
                name = d.string
            if j == 9:
                hp.heappush(win,(-1*int(d.string),name))
            if j == 11:
                hp.heappush(save,(-1*int(d.string),name))
            if j == 13:
                hp.heappush(hld,(-1*int(d.string),name))
return win,save,hld
```

● 爬取球隊近況及最近一場球賽結果

相信許多人因為課業或是事業繁忙，沒辦法每場完整的收看整場球賽，且即使全程觀看，也未必能記得所有有興趣的數據，例如：所支持投手該場的三振數，或是某打者的打擊率。因此，除了排名的功能，這個工具也能提供，球隊最近一場球賽的詳細資料。

首先，我們必須先觀察到，中職網站中，存放各場球賽數據的網址，擁有甚麼樣的特性。

http://www.cpbl.com.tw/games/box.html?&game_type=01&game_id=94&game_date=2019-05-30&pbyear=2019

如下，我們能把這個網址，解析為下面這些字串所組成：

```
url = 'http://www.cpbl.com.tw/games/box.html?&game_type=01&game_id='
+ str(game_id) + '&game_date=' +str(game_date) + '&pbyear=2019
```

因此，我們必須先找出該場次的日期及球賽編號，我們便能成功得到，數據。

GAME NO.	STADIUM	DATE	TIME	AWAY TEAM	SCORE	HOME TEAM	SCORE	W
93	新莊	2019-05-29	18:35	統一7-ELEVEn	10	富邦	4	統一7-ELEVEn
89	台南	2019-05-26	17:05	中信兄弟	2	統一7-ELEVEn	6	統一7-ELEVEn
87	台南	2019-05-25	17:05	中信兄弟	6	統一7-ELEVEn	10	統一7-ELEVEn
85	台南	2019-05-24	18:35	中信兄弟	5	統一7-ELEVEn	2	中信兄弟
82	台南	2019-05-22	18:35	富邦	1	統一7-ELEVEn	0	富邦
81	台南	2019-05-21	18:35	富邦	11	統一7-ELEVEn	1	富邦
80	天母	2019-05-19	17:05	Lamigo	9	統一7-ELEVEn	10	統一7-ELEVEn
78	天母	2019-05-18	17:05	Lamigo	1	統一7-ELEVEn	8	統一7-ELEVEn

並依前一頁爬取表格的方式，爬取此表格的第一列，就能獲得關鍵的那兩個資料，並進入擁有該場球員成績的網站。

打擊成績

統一7-ELEVEn	AB	R	H	RBI	BB	SO	SB	AVG
黃恩賜 , 3B	3	1	1	2	0	2	0	0.339
林祖傑 , SS	2	1	1	1	2	1	0	0.254
蘇智傑 , LF	4	1	1	1	0	1	0	0.255
潘武雄 , DH	3	0	2	1	0	1	0	0.307
陳重羽 , RF	4	0	0	0	0	0	0	0.277
楊家維 , 2B	3	0	0	0	1	1	0	0.236
高國慶 , 1B	2	1	0	0	0	0	0	0.233
陳重廷 , (PH)	0	0	0	0	0	0	0	0.000
林祐樂 , C	3	0	1	0	0	0	0	0.211
方昶詠 , CF	2	1	1	0	1	0	1	0.150
Total	26	5	7	5	4	6	1	
more								

打擊成績

富邦	AB	R	H	RBI	BB	SO	SB	AVG
張正偉 , RF	4	1	2	0	0	0	0	0.341
王正棠 , 2B	5	0	0	0	0	0	0	0.303
胡金龍 , LF	2	1	1	0	2	0	0	0.321
林哲瑄 , (CF)	1	0	0	0	0	0	0	0.286
林益全 , 1B	3	1	0	0	1	0	0	0.305
蔣智賢 , DH	4	3	4	3	0	0	0	0.310
高孝儀 , CF (LF)	3	1	2	0	1	0	0	0.286
林威廷 , 3B	4	1	1	4	0	1	0	0.362
李宗賢 , SS	4	0	0	1	0	1	1	0.290
林宥穎 , C	4	0	1	0	0	1	0	0.250
Total	34	8	11	8	4	3	1	
more								

● 查詢球賽球員數據

進入 ptt 的棒球版，我們可以發現，每天大量出現 “今日XXX” 的文章也就是說，對於許多特定選手，許多人都希望能得到當天的比賽成績。但每個人所想要觀看的選手，可能是隨著支持球隊而改變，因此，同樣藉由爬蟲，能獲取這樣的資料。

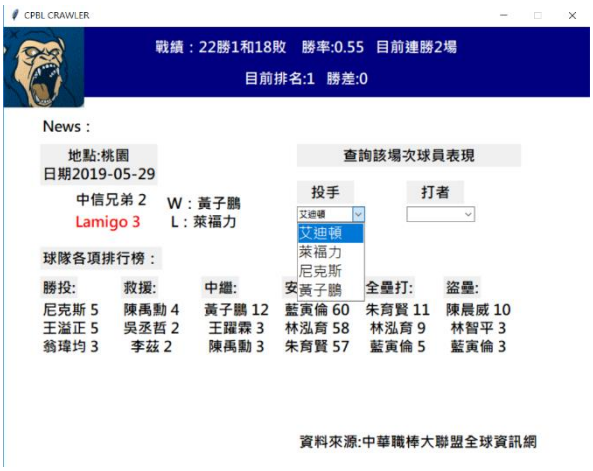
```
4 [分享] 今日 NPB & CPBL 人數
  cho840929
1 [分享] 今日王柏融
  fatman5566
5 [閒聊] 今日潘武雄
  jeff1013
2 [分享] NPB排名
  Ivers
  [分享] MLB排名
  Ivers
1 [分享] 今日蔣智賢
  asas115999
```

承上頁，我們同樣能以之前爬表格的方式，獲得這些資訊，不同的是，在這裡，我們需要以 dictionary 的資料結構來存這些資料，因為我們要支援的是查詢功能，因此將 key 及 value 在爬取時就存入資料結構中，即能達成查詢的效果。

● 輸出展示 (利用 tkinter 進行圖形化介面製作)

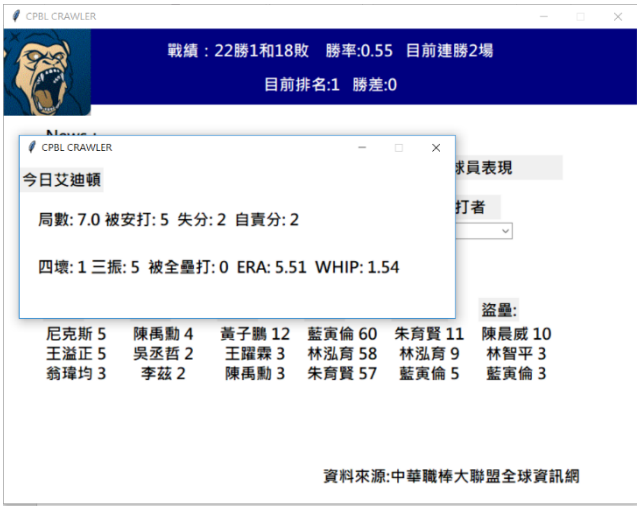
先選擇欲查詢球隊

即可得到球隊近況(上方)
最新球賽資訊，及排行榜



利用右方查詢功能選擇欲查詢球員

即會跳出今日球員表現



• 該資料能如何進一步使用？

就短期而言，這個程式提供了一個便利且清楚的查詢介面。若長期使用，每天固定讓電腦自動執行這個程式，並存入 excel 等表格(可輕易做到，並沒有包含在這次作業中的程式碼中)，即可獲取在不同時間點，各個球隊這些數據的變化趨勢，除了能幫助想進一步研究球隊的球迷，對運動科學分析有興趣的人，也能利用長期累積下來的數據做研究(球隊、球員不同月份，的表現等等)

• 未來展望

希望能將資料的爬取，擴展到日本、美國等職棒，除了方便查詢更多球隊資料，也希望能進一步分析不同國家職棒間的差異性，例如：美國職棒以長打得分著名，日本以小球戰術為主，是否在勝率較高的球隊中，犧牲打，或是全壘打的數量排名上，會不會在不同職棒間會有不同的結果。