HW2-Problem6

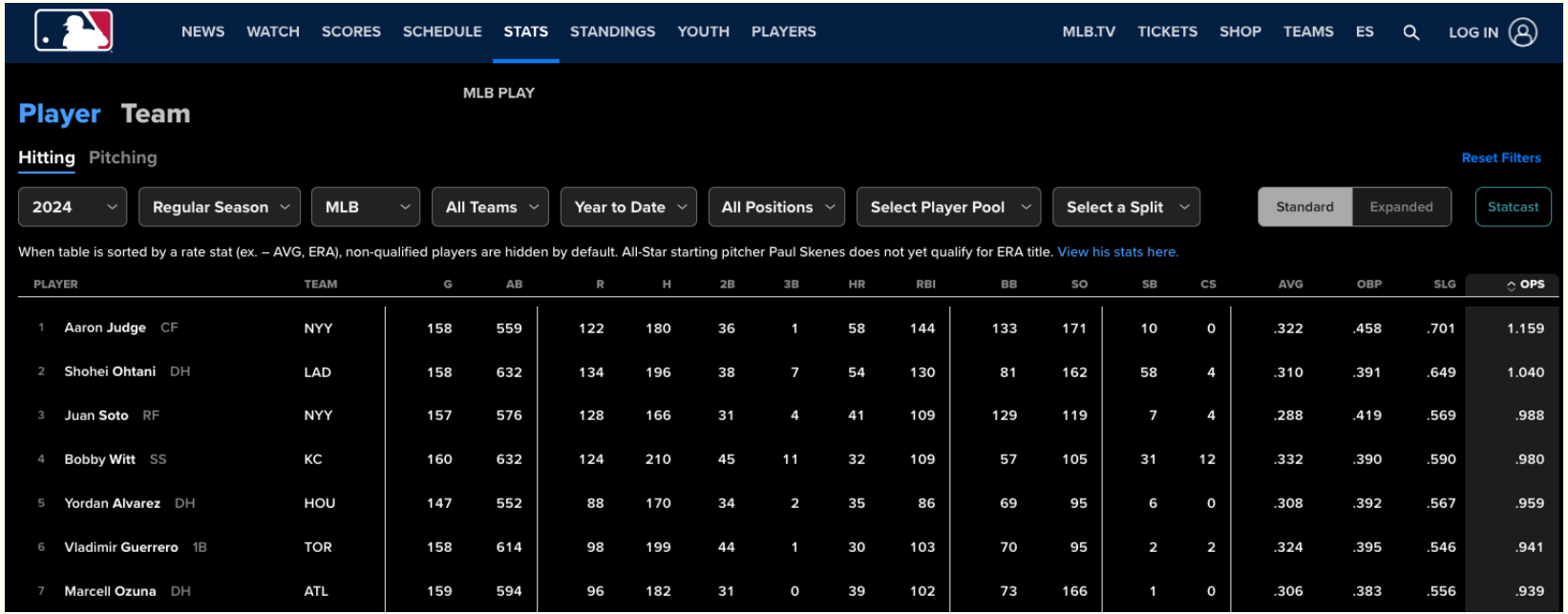
MLB Data Crawling and Analysis

Introduction to the problem6 of hw2

01

MLB 網頁爬蟲

MLB 網頁爬蟲



NEWS WATCH SCORES SCHEDULE STATS STANDINGS YOUTH PLAYERS MLB.TV TICKETS SHOP TEAMS ES LOG IN

MLB PLAY

Player **Team**

Hitting Pitching

Reset Filters

2024 | Regular Season | MLB | All Teams | Year to Date | All Positions | Select Player Pool | Select a Split | Standard | Expanded | Statcast

When table is sorted by a rate stat (ex. – AVG, ERA), non-qualified players are hidden by default. All-Star starting pitcher Paul Skenes does not yet qualify for ERA title. View his stats here.

	PLAYER	TEAM	G	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	AVG	OBP	SLG	OPS
1	Aaron Judge CF	NYY	158	559	122	180	36	1	58	144	133	171	10	0	.322	.458	.701	1.159
2	Shohei Ohtani DH	LAD	158	632	134	196	38	7	54	130	81	162	58	4	.310	.391	.649	1.040
3	Juan Soto RF	NYY	157	576	128	166	31	4	41	109	129	119	7	4	.288	.419	.569	.988
4	Bobby Witt SS	KC	160	632	124	210	45	11	32	109	57	105	31	12	.332	.390	.590	.980
5	Yordan Alvarez DH	HOU	147	552	88	170	34	2	35	86	69	95	6	0	.308	.392	.567	.959
6	Vladimir Guerrero 1B	TOR	158	614	98	199	44	1	30	103	70	95	2	2	.324	.395	.546	.941
7	Marcell Ozuna DH	ATL	159	594	96	182	31	0	39	102	73	166	1	0	.306	.383	.556	.939

MLB 網頁爬蟲

要求

- 使用 Scrapy 自己生成爬蟲框架（不提供程式碼）
- https://www.mlb.com/stats/

  爬取此網頁中所有球員的各項指標（6 頁都要抓到）
- 繳交整個爬蟲程式，不包含輸出 csv
- 使用網頁表格的表頭名稱作為輸出檔案的欄位名稱
- 球員名字後面的守備位置不用抓

評分方式

- TA 會執行同學的爬蟲程式，得到程式的輸出檔
- 而後與 TA 自己的 csv 檔案進行內容比較

MLB 網頁爬蟲

STEP 1: 安裝 Scrapy

STEP 2: 建立爬蟲專案

STEP 3: 建立爬蟲程式

pip + venv

```
pip install Scrapy # 安裝
scrapy version  # 確認安裝成功
```

conda

```
conda install conda-forge::scrapy # 安裝
scrapy version  # 確認安裝成功
```

MLB 網頁爬蟲

STEP 1: 安裝 Scrapy

STEP 2: 建立爬蟲專案

STEP 3: 建立爬蟲程式

使用 terminal 建立 scrapy 專案
完成後會看到如右圖的結構

scrapy startproject <project_name>

```
tutorial/
    scrapy.cfg              # 專案環境設定

    tutorial/               # 撰寫成 module
        __init__.py

        items.py            # 定義抓取的內容

        middlewares.py      # 中介流程，在各元件溝通之間作用

        pipelines.py        # items 的處理流程

        settings.py         # 設定檔案

        spiders/            # 爬蟲程式資料夾
            __init__.py
```

MLB 網頁爬蟲

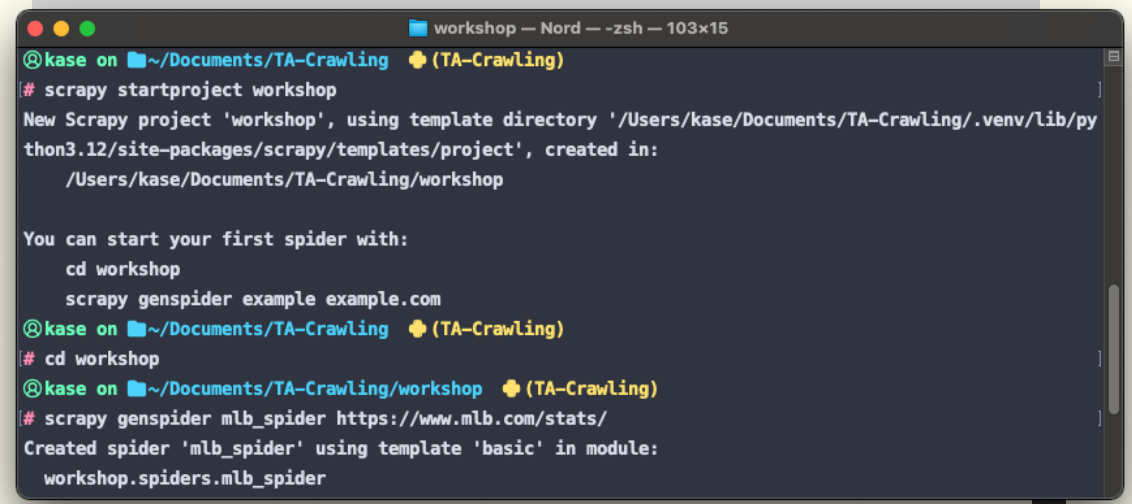
STEP 1: 安裝 Scrapy

STEP 2: 建立爬蟲專案

STEP 3: 建立爬蟲程式

使用 terminal 建立爬蟲程式
設定爬蟲名稱與爬取的網域

```
cd <project_name>/spider
scrapy genspider <spider_name> <domain>
```



MLB 網頁爬蟲

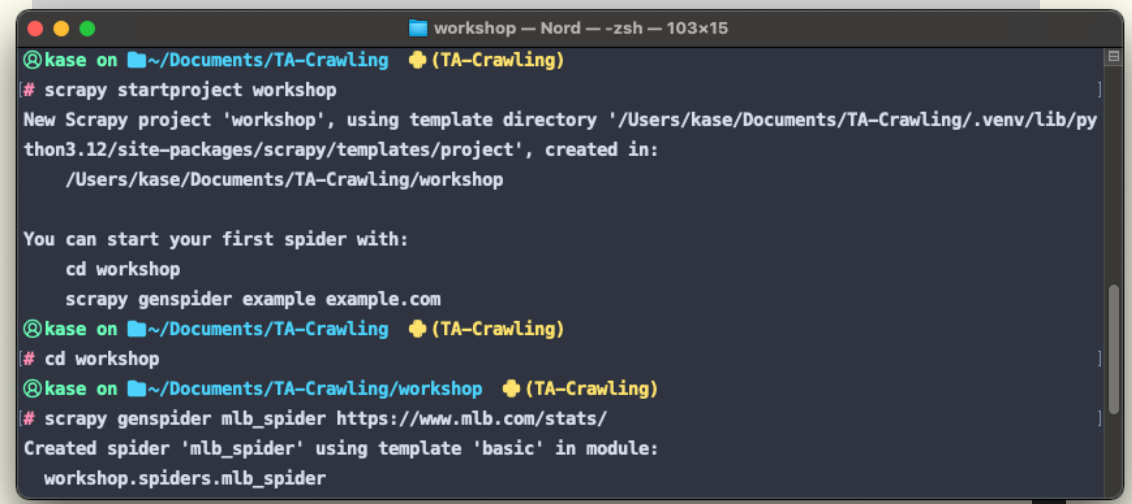
STEP 1: 安裝 Scrapy

STEP 2: 建立爬蟲專案

STEP 3: 建立爬蟲程式

此時你已成功建立爬蟲程式的框架！（開始寫作業吧）

```
cd <project_name>/spider
scrapy genspider <spider_name> <domain>
```



02

MLB 資料視覺化

MLB 資料視覺化

要求

- 使用提供的 hw2_vis_template.ipynb 模板
- 計算所有大聯盟球員的平均得分（Runs）
  將球員分為「高得分」與「低得分」兩組
- 將兩組球員的 *得分指標 全部相加
  畫出這兩組的 correlation matrix
- 解釋你對結果的看法

評分方式

TA 執行你的 .ipynb，檢驗是否畫出正確圖表

*得分指標

- 得分（Runs Scored, R）
- 打點（Runs Batted In, RBI）
- 安打數（Hits, H）
- 全壘打數（Home Runs, HR）
- 打擊率（Batting Average, AVG）
- 上壘率（On-Base Percentage, OBP）
- 長打率（Slugging Percentage, SLG）
- 盜壘數（Stolen Bases, SB）
- 打數（At Bats, AB）
- 四壞球率（Walks, BB）
- 被三振數（Strikeouts, SO）

Thanks!

Do you have any questions?

Please mail to Netai-2024@googlegroups.com