2020 DB team proposal 02 組

以上內容來自 HackMD: https://hackmd.io/GFo9xYfMRNCeQgr_60Z1gg?view
請助教不要外流這連接喔^ ^

Github: https://github.com/brianchennn/Database_Team_Project

資源連結

組員 GitHub

洪瑋廷 hungdino
王昶淵 Channing-Wang
陳煜盛 brianchennn
李嘉盛 justwe8787

GitHub Repo
GitHub Repo BackUp

TODOs

Team Project Proposal

Spec

- 先討論出主題
- 填寫表單確認沒撞主題 # 看來有人用了最基本的空氣品質資料庫呢

主體要這個網站嗎: aniDB

- 再寫 Proposal

Deadline 4/29 23:59

討論共同時間

```
把你偏好的討論時間寫下來
陳煜盛：週一晚上  週二 H、晚上  週三晚上  週六   沒了 QQ
洪瑋廷：週二晚上、週四晚上、週五晚上
李嘉盛：週二、週四晚上以外都沒問題  我要去償還我的原罪(微積分)
王昶淵：一二三五晚上都行
```

git 遇到的問題

git add 後 後悔了 怎麼辦:

git reset HEAD [file_name]
Dino: https://gitbook.tw/chapters/using-git/reset-commit.html
猴子都懂的git: https://backlog.com/git-tutorial/tw/intro/intro21.html

4/21 晚上討論結果

本日先討論出要選什麼主題
一開始想做 aniDB（動畫推薦資料庫）
但是覺得這並非太複雜的資料庫
後來又參考了政府的網頁 覺得空氣品質的資料庫很棒 可惜其他組已經選了
然後因為我很喜歡看 MLB 於是就往體育賽事資料庫找 找到了
SortsDataIO https://sportsdata.io 可是因為這網站詳細的分數數據要付費
所以又網免費來源 kaggle 去找 找到了 2015-2018 pitch data
最後決定要分析 MLB 投手的相關資料

The description of the data

Source

Kaggle Pitch Data 2015-2018

five table:

- **games.csv**

    - **attendance** number of fans who attended (NOTE: for first game of doubleheaders, value is often erroneously 1 or 0. This comes directly from XML files. This data may not be recorded for those games; MLB gameday pages do not report attendance for these game)
    - **away_final_score** final score for the visiting team
    - **away_team** three letter abbreviation for away team; third letter often indicates league(national vs american)
    - **date** date of game
    - **elapsed_time** length of game, in minutes
    - **g_id** game ID. Matches with game_id in atbats.csv
    - **home_final_score** final score for the home team
    - **home_team** three letter abbreviation for home team; third letter often – - indicates league(national vs american)
    - **start_time** start time of game
    - **umpire_1B**
    - **umpire_2B**
    - **umpire_3B**
    - **umpire_HP**
    - **venue_name** name of stadium
    - **weather** description of weather
    - **wind** description of wind
    - **delay** length of delay before game, in minutes
- **ejections.csv**
    - **ab_id** foreign key for atbats.csv, may be unreliable (ejection happened before, after, during atbat
    - **des** Human readable, in format
    - **event_num** event number for ejection (from xml file; many event_nums are skipped)
    - **g_id** foreign key for games.csv
    - **player_id** foreign key for player_names.csv
    - **date** directly from games.csv
    - **BS** ‘Y’ if ejection was for arguing balls and strikes, empty otherwise
    - **CORRECT** Whether the ejection was correct (only for BS ejection). From closecallsports.com
    - **team** team for player ejected
    - **is_home_team** whether that team is the home team-

- **pitches.csv** (Pitch-level data, including lots of information about the trajectory of the pitch. Match up with atbats.csv for complete picture of game situation. Data comes from unlabeled xmls from MLB website, so the meaning of some fields is not clear.)
    - **px** x-location as pitch crosses the plate. X=0 means right down the middle
    - **pz** z-location as pitch crosses the plate. Z=0 means the ground
    - **start_speed** Speed of the pitch just as it's thrown
    - **end_speed** Speed of the pitch when it reaches the plate
    - **spin_rate** The pitch's spin rate, measure in RPM
    - **spin_dir** Direction in which pitch is spinning, measured in degrees
    - **break_angle**
    - **break_length**
    - **break_y**
    - **ax**
    - **ay**
    - **az**
    - **sz_bot**
    - **sz_top**
    - **type_confidence** Confidence in pitch_type classification. Goes up to 2 for some reason.
    - **vx0**
    - **vy0**
    - **vz0**
    - **x**
    - **x0**
    - **y**
    - **y0**
    - **z0**
    - **pfx_x**
    - **pfx_z**
    - **nasty**
    - **zone**
    - **code** Records the result of the pitch. See dataset description for list of codes and their meaning
    - **type** Simplified code, S (strike) B (ball) or X (in play)

  - **pitch_type** Type of pitch. See dataset description for list of pitch types
  - **event_num** event number, used for finding when exactly ejections happen.
  - **b_score** score for the batter’s team
  - **ab_id** at-bat ID. Matches up with atbats.csv
  - **b_count** balls in the current count
  - **s_count** strikes in the current count
  - **outs** number of outs (before pitch is thrown)
  - **pitch_num** pitch number (of at-bat)
  - **on_1b** True if there’s a runner on first, False if empty
  - **on_2b** True if there’s a runner on second, False if empty
  - **on_3b** I don’t know
- **atbats.csv** (This file lists the information that cannot change over the course of an at-bat)
  - **ab_id** at-bat ID. First 4 digits are year. Matches with ab_id in pitches.csv
  - **batter_id** player ID of the batter. Given by MLB, player names found in player_names.csv
  - **event** description of the result of the at-bat
  - **g_id** game ID. First 4 digits are year
  - **inning** inning number
  - **o** number of outs after this at-bat
  - **p_score** score for the pitcher’s team
  - **p_throws** which hand pitcher throws with. Single character, R or L
  - **pitcher_id** player ID of the pitcher. Given by MLB, player names found in player_names.csv
  - **stand** which side batter hits on. Single character, R or L
  - **top** True if it’s the top of the inning, False if it’s the bottom
- **player_names.csv** (Matches names with player’s ID)
  - **id** matches with batter_id and pitcher_id
  - **first_name** first name
  - **last_name** last name

other information of data

```
此資料庫不會再更新了喔><
```

User Interface:

```
1.我們想用website的方式呈現出結果
2.User 可以 insert 資料進去

3.讓使用者能夠查出每一球,每一場比賽,或是全賽季之球速,轉速,擊球初速,擊球仰角,
守備表現,edge %,wOBA,xwBOA,SRC+,這些現代棒球的數據
```

因為我們的csv檔案很大 所以需要兩個以上伺服器
且能限制client request的頻率次數
(使用freeBSD中的HAPROXY防止DDoS DNS大量攻擊之類的)