Homework 6

Version 0.1

COS125 - Instructor: Zachary Hutchinson

Due: Friday November 3rd @12AM (midnight)

Submission Instructions

These must be followed or you will lose points.

1. Submit your coded solution as a .py file. Code in any other file format will be rejected.
2. Your .py file should be named: yourlastname_hwX.py. The 'X' in the filename is replaced by the homework number.
3. Your code must have as a comment at the top of the file: your name, the homework and the names of anyone from whom you received help. This includes students in the course as well as course staff. For example, if you went to Boardman 138 and received help from an MLA, put their name on your file. If a classmate helped you debug a bit of code, put their name down. This is for your benefit.
4. If you do not manage to squash all the bugs in your program, include a comment at the top of the file detailing the outstanding bugs. Acknowledging bugs is a sign of a mature programmer. Doing so will not eliminate point deductions but it might mitigate them. It shows you care about your work.

Learning Objectives

- Experience with reading from and writing to text files using Python.
- Understand the use of unique keys to coordinate different information stored in multiple files.

Code Tools Tested by This Homework

- File Input/Output

Problem Description

Unfortunately, you are a baseball fanatic. And even worse, you've a penchant for sports statistics. You have taken it upon yourself to improve your Python programming by writing a program to parse and combine information from several files containing baseball statistics from the history of baseball (up to 2020). As a first step, you want to create a program that will ask the user for a year and team, and from this information, the program will write to a file all the batters that played for that team and their slugging rating (which isn't included among the stats and must be calculated from several other statistics).

IMPORTANT: *You do not need to know anything about the game of baseball to complete this assignment.*

On Brightspace you will find three .csv files (or comma separated value files) containing baseball statistics. CSV files use commas to delineate between columns. Each line of the file is a row. You can think of this as a spreadsheet (in fact you can open a csv file in an spreadsheet program). The three files are:

- **Teams.csv**: This file contains a list of major league teams from 1871 to 2020. The file includes important data like: teamID, yearID and team name.
- **Batting.csv**: This file contains a list of batters who played in the major leagues between 1871 and 2020. The file contains a variety of statistics about their performance in a given year. NOTE: batters are identified by a playerID and yearID because players appear more than once in this file if they played for multiple years.
- **People.csv**: This file contains player personal information such as their first and last name. NOTE: players occur only once in this file.

Requirements

Using the information in the three files listed above, your program needs:

1. to ask the user for a year between 1871 and 2020.
2. present the user with a list of the major league teams in existence in the selected year.
3. ask the user to select a team using a number (index).
4. find all of the batters who played for the selected team.
5. calculate their slugging statistic (see below for formula).
6. write out to a file called YEAR_AND_TEAM_NAME.txt' the players first name, last name and slugging statistic in a neat, tabular format. YEAR_AND_TEAM_NAME should replaced by the year of the team and the team's name. The team's name should be in all lowercase. If there were spaces in the team's name, they should be replaced by underscores. For example, the 1877 Hartford Dark Blues would be stored in a file called: 1877_hartford_dark_blues.txt. **NOTE:** Some teams have punctuation in their names (e.g., the 1890 Brooklyn Ward's Wonders). This should be removed, if it exists.
7. In addition to the batter's information, the first line of this file should be: *The (WINS-LOSSES) YEAR TEAM_NAME roster:*.
    1. WINS: are the number of wins (W) the team had that year.
    2. LOSSES: the number of losses that year.
    3. YEAR: year of the team.
    4. TEAM_NAME: the name of the team.

Slugging Statistic (SLG) and Data

The slugging statistic is a measure of a batters effectiveness when they hit the ball. It is calculated using the following formula. See below for an explanation of each element of the SLG equation:

$$\mathrm{SLG} = \frac{\mathrm{H} + 2\mathrm{B} * 2 + 3\mathrm{B} * 3 + \mathrm{HR} * 4}{AB}$$

The statistics for the above equation are all found in the *Batting.csv* file.

- *H* - The number of singles for the given player.
- *2B* - The number of doubles.
- *3B* - The number of triples.
- *HR* - The number of home runs.
- *AB* - The number of at bats (or attempts).

In essence, SLG is a calculation of the number of bases touched by each batter divided by their number of times at bat. For example, 3B is a triple, meaning the batter, after hitting the ball, touches three bases;

therefore, the number of triples is multipled by 3 to represent this.

Output also requires that you write the wins and losses for the selected team (see example output file). These statistics are given by following columns in the *Teams.csv* file:

- *W* - The number of games won by the team that year.
- *L* - The number of games lost by the team that year.

Hints

- This assignment requires that you connect information spread across three different files. Carefully examine the three files. Certain columns in each file allow you to identify relevant information in another file. Part of the challenge of file IO is analyzing the data's format.
- Connecting data between files means that you must open a file, find data you need and store it in your program. Open the next file and use the stored data to extract more information from the next file. Etc.
- Ask yourself: how will you connect the information between files? The answer to this question will give you the order in which you should process the files.
- You will need to find the column numbers (indexes) of data stored in the three files. For example, the team's name is at index 40 of the *Teams.csv* file. It might be helpful to open these files in a spreadsheet program to see all the data lined up.
- Break the problem down into subproblems. For example, you are reading from three different files. Each file is a subproblem. Identify what information you need for each task and what information each task will produce.
- I encourage you to use functions to partition off different parts of the code so they can be more easily tested.

Example

I have uploaded an example output file for the 1927 New York Yankees so you can see what the file should look like. Do not just test your code against this one team. Try multiple teams while testing.

Here is a screenshot of the run that produced the 1927 NY Yankees file:

```
[zax@foley hw6]$ python3 hw6.py
Enter a year (1871-2020): 1927
0: Boston Red Sox
1: Brooklyn Robins
2: Boston Braves
3: Chicago White Sox
4: Chicago Cubs
5: Cincinnati Reds
6: Cleveland Indians
7: Detroit Tigers
8: New York Giants
9: New York Yankees
10: Philadelphia Athletics
11: Philadelphia Phillies
12: Pittsburgh Pirates
13: St. Louis Browns
14: St. Louis Cardinals
15: Washington Senators
Select a team to save roster: 9
The 1927 New York Yankees have been written to the file: 1927_new_york_yankees.txt.
[zax@foley hw6]$
```

Disclaimers

The files originate from: http://seanlahman.com/download-baseball-database/. I have not examined all the data in these files. It is possible some entries have unexpected text or errors. If you encounter a specific problem that seems due to data error, try other teams. If it's just one team with the issue, let me know.

No, I am not a baseball fan. But I like a nice dataset.