Homework 5: Regular Expressions

SI206 - Fall 2023

Introduction

In this assignment, you will work with a set of real-world data retrieved from Wikipedia. The data is stored in the file `"hw5_players.txt"`. In each function (except for get_raw_data) your task is to use regular expressions to extract useful information.

Sample data

The data from Wikipedia is in plain text with the following format:

```
Michael Jeffrey Jordan(born February 17, 1963), also known by his initials
MJ,[9] is an American former professional basketball player and
businessman. Widely considered the greatest basketball player in history,
his profile on the official National Basketball Association (NBA) website
states that "by acclamation, Michael Jordan is the greatest basketball
player of all time."[10] He played fifteen seasons in the NBA, winning six
NBA championships with the Chicago Bulls. He was integral in popularizing
the sport of basketball and the NBA around the world in the 1980s and
1990s,[11] becoming a global cultural icon.[12]
```

Your Tasks

Please implement the following functions in **`HW5.py`**:

- **`get_player_info(file_name)`**
  - This function reads the file with the file name and returns a list of strings. Each string should contain all the information for one player.

- **`create_bio_dict(player_info)`**
  - This function gets a list of strings with each player's information and returns a dictionary with player's full names (including any suffixes such as Jr) as keys and their birth year, month and date as values in tuples.
  - Sample output:

```
{
   "Michael Jeffrey Jordan": (1963, "February", 17),
   "Thomas Edward Patrick Brady Jr": (1977, "August", 3),
   ...
}
```

- **`create_short_bio(player_info)`**
  - This function gets a list of strings with each player's information and returns a list of short bios (the first sentence of each player's information). For example, if the data is:

    ```
    Michael Jeffrey Jordan (born February 17, 1963), also known by
    his initials MJ,[9] is an American former professional basketball
    player and businessman. Widely considered the greatest basketball
    player in history, his profile on the official National
    Basketball Association (NBA) website states that "by acclamation,
    Michael Jordan is the greatest basketball player of all
    time."[10] He played fifteen seasons in the NBA, winning six NBA
    championships with the Chicago Bulls. He was integral in
    popularizing the sport of basketball and the NBA around the world
    in the 1980s and 1990s,[11] becoming a global cultural icon.[12]
    ```

    Then the short bio should be:

    ```
    Michael Jeffrey Jordan (born February 17, 1963), also known by
    his initials MJ,[9] is an American former professional basketball
    player and businessman.
    ```

- **`get_valid_year(player_info)`**
  - This function finds all valid years (1980~2023 inclusive) and returns a list of lists containing valid years for each player. Note that `"2,000"` and `"2020s"` are not valid years.
  - Sample output:

    ```
    [
       [],
       ['2001', '2019', '2019'],
       ['1984', '2011', '2018'],
       ...
    ]
    ```

- **`Test Cases`**

  Make at least 2 test cases each for `get_player_info`, `create_bio_dict`, `create_short_bio` and `get_valid_year`.

Extra credit (6 points)

- **`get_abbr_dict(player_info)`**
  - This function finds all **abbreviations** and returns a dictionary with abbreviations as keys and full names as values.
    - Abbreviations are made of consecutive uppercase letters inside a set of parentheses.
    - Full names are consecutive words starting with an uppercase letter.
  - Sample format in data: `National Basketball Association (NBA)`
  - Sample output:

```
{
   "NBA": "National Basketball Association",
   "MVP": "Most Valuable Player",
   ...
}
```

Grading Rubric (60 points + 6 points)

This rubric does not show all the ways you can lose points. For each of the functions, you can earn:

- points for correctly implementing each of the 4 functions (44 points)
  - `get_player_info` (8 points)
  - `create_bio_dict` (16 points)
  - `create_short_bio` (10 points)
  - `get_valid_year` (10 points)
- points for creating tests for each function (16 points)
  - 2 points each * 8 test cases
- points for extra credits and test cases (6 points)
  - correctly implementing `get_abbr_dict` (4 points)
  - creating tests (2 points)

Submission:

Make at least 4 git commits and turn in your GitHub repo URL on Canvas by the due date to receive credit.