Final Project Report: NBA Search Engine

1. Overview:

In this project, we build a search engine related to NBA. At first, we analyze what kind of information is important to NBA games. Then, we decide to focus on the players' statistics and come up with four functionalities for the engine according to different aspect of players: (1) Get general statistics for a NBA player yearly, (2) Get career statistics for all NBA players, (3) Get similar players for a NBA player in playing style, and (4) Get league leaders statistics for one year. What's more, we apply the knowledge we have learnt in homework 1 to implement function (1) and (4) about information extraction, homework 4 to implement function (2) about web crawler and homework 2 and 3 to implement function (3) about similarity measurement. Finally, we use a main menu to combine all these functions together for users to use our search engine conveniently.

2. Run the Program

At first, change to the final_project directory. Then, run "Python3.6 nba_ir.py" command in the terminal since we use Python 3.6 to implement our project. If some libraries missing occur, please install them and try again. Finally, the main menu will appear as showed in the Fugure.1.

```
============================================================
=                                                          =
=            Welcome to the NBA search engine              =
=                                                          =
============================================================

  OPTIONS:
  1 = Get general statistics for a NBA player yearly
  2 = Get career statistics for all NBA players
  3 = Get similar players for a NBA player in playing style
  4 = Get league leaders statistics for one year
  5 = Quit

============================================================
```

Figure.1 the Main Menu for NBA Search Engine

3. Functionalities Explanations

3.1 Functionality 1: Get general statistics for a NBA player yearly

For this function, we let the user enter the first name and last name of a player. Then, we will search starting from the base URL "http://www.basketball-reference.com/players/" and try to get the link for the target player. With this URL, we extract the general statistics for the player throughout his whole career, such as average points, average assists, average rebounds and average steals for each year. Finally, we plot the four figures to give the user a better understanding of how the statistics tendency of the player.

For here, we use the knowledge of how web search works to get the URL for player by given the name. We also succeed to extract the information the user wants to know for a web page from the html decoded content. Finally, we call the "matplotlib", the plot library of python, to plot the statistics of the player to show if the player plays better or not.

As showed in Figure.2 and Figure.3, they demonstrate how this function works based on the example of “Kobe Bryan”.

```
Enter Option: 1
Enter the player first name: kobe
Enter the player last name: bryant
------------------------------------------------------------
The player's average statistics throughout all career years
------------------------------------------------------------
Average Points from:  1997 to 2016
7.6, 15.4, 19.9, 22.5, 28.5, 25.2, 30.0, 24.0, 27.6, 35.4, 31.6, 28.3, 26.8, 27.0, 25.3, 27.9, 27.3,
13.8, 22.3, 17.6
Average Assist from:  1997 to 2016
1.3, 2.5, 3.8, 4.9, 5.0, 5.5, 5.9, 5.1, 6.0, 4.5, 5.4, 5.4, 4.9, 5.0, 4.7, 4.6, 6.0, 6.3, 5.6, 2.8
Average Rebounds from:  1997 to 2016
1.9, 3.1, 5.3, 6.3, 5.9, 5.5, 6.9, 5.5, 5.9, 5.3, 5.7, 6.3, 5.2, 5.4, 5.1, 5.4, 5.6, 4.3, 5.7, 3.7
Average Steals from:  1997 to 2016
0.7, 0.9, 1.4, 1.6, 1.7, 1.5, 2.2, 1.7, 1.3, 1.8, 1.4, 1.8, 1.5, 1.5, 1.2, 1.2, 1.4, 1.2, 1.3, 0.9
```

Figure.2 the General Statistics for Kobe, Bryant in Plain Text

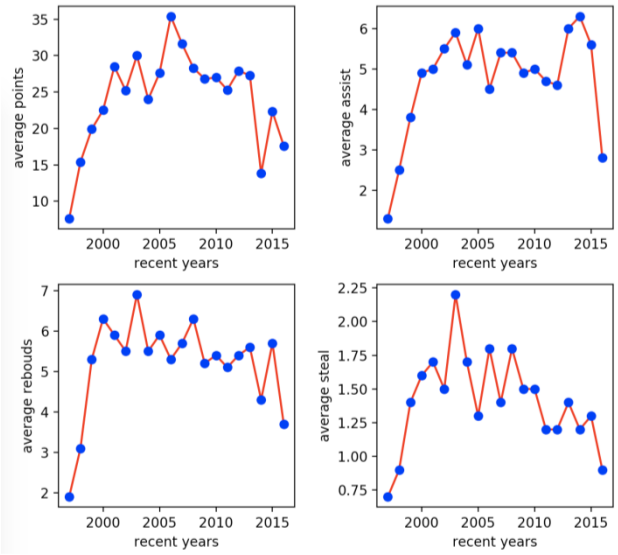

Figure.3 the General Statistics for Kobe, Bryant in Graph

3.2 Functionality 2: Get career statistics for all NBA players

In this function, we display the career statistics for all the NBA players, including name, position and five important career data.

In order to implement this function, we use a web crawler to go through the player pages according to the base URL “http://www.basketball-reference.com/players/”. We also need to use the regular expression knowledge to extract the URL link for each player and return a list of player’s URL. After we get this URL list, we can call the method in the function 1 to get the career statistics as the following.

For here, we let the user to input the initial of players’ last name for search since there are more than four thousands players in the website. If we go through all the players, it will take several hours. If we want all players’ statistics, uncomment the corresponding codes. As showed in Figure.4, this example displays the career statistics with initial of last name “z”.

```
Enter Option: 2
Enter the initial of players' last name for search: z
(We are displaying player 1 / 19 data)
Name: Max     Zaslofsky
Position: 2
Career Data: Points: 14.8  Blocks: 0.0  Steals: 0.0  Assits: 2.0  Rebounds: 2.8

(We are displaying player 2 / 19 data)
Name: Zeke    Zawoluk
Position: 3, 4
Career Data: Points: 6.8  Blocks: 0.0  Steals: 0.0  Assits: 1.2  Rebounds: 4.1

(We are displaying player 3 / 19 data)
Name: Cody    Zeller
Position: 4, 5
Career Data: Points: 8.0  Blocks: 0.8  Steals: 0.7  Assits: 1.3  Rebounds: 5.6

(We are displaying player 4 / 19 data)
Name: Dave    Zeller
Position: 1
Career Data: Points: 1.5  Blocks: 0.0  Steals: 0.0  Assits: 1.0  Rebounds: 0.4

(We are displaying player 5 / 19 data)
Name: Gary    Zeller
Position: 2
Career Data: Points: 3.2  Blocks: 0.0  Steals: 0.0  Assits: 0.4  Rebounds: 1.1
```

Figure.4 the Career Statistics with Initial of Last Name “Z” (Partially)

3.3 Functionality 3: Get similar players for a NBA player in playing style

Based on career statistics for all players, we want to compare players to get the similar players in playing style if we are given a target player for comparison.

For each player, we focus on the five important statistics, such as points, blocks, steals, assists and rebounds. We think that the more likely of the scale for these five dimension value of two players, the more similar of the playing style of this two players. The magnitude represents performance of the player so we don’t need to care when we want to compare playing style.

According this semantics, we use the cosine similarity function to get the similarity value between two players based on the five dimension value.

For here, we also restrict the players for comparison by initial of last name due to the time concern. As showed in Fugure.5, we find the top ten similar playing style players for Cody, Zeller with initial of last name “z”.

```
Enter Option: 3
Enter the target player first name: cody
Enter the target player last name: zeller
Enter the initial of players' last name for comparison: z
(We are collecting player 1 / 19 data)
(We are collecting player 2 / 19 data)
(We are collecting player 3 / 19 data)
(We are collecting player 4 / 19 data)
(We are collecting player 5 / 19 data)
(We are collecting player 6 / 19 data)
(We are collecting player 7 / 19 data)
(We are collecting player 8 / 19 data)
(We are collecting player 9 / 19 data)
(We are collecting player 10 / 19 data)
(We are collecting player 11 / 19 data)
(We are collecting player 12 / 19 data)
(We are collecting player 13 / 19 data)
(We are collecting player 14 / 19 data)
(We are collecting player 15 / 19 data)
(We are collecting player 16 / 19 data)
(We are collecting player 17 / 19 data)
(We are collecting player 18 / 19 data)
(We are collecting player 19 / 19 data)
(Finish collecting data!)

Top 10 similar players:
Similarity      FirstName       LastName
1.000           Cody            Zeller
0.998           Tyler           Zeller
0.994           Ivica           Zubac
0.992           Zeke            Zawoluk
0.992           George          Zidek
0.990           Paul            Zipser
0.984           Luke            Zeller
0.976           Phil            Zevenbergen
0.971           Tony            Zeno
0.969           Wang            Zhizhi
```

Figure.5 the Top Ten Similar Playing Style Players for Cody, Zeller with Initial of Last Name “Z”

3.4 Functionality 4: Get league leaders statistics for one year

For this function, we let the user enter a specific year. Then, we search starting from base URL "http://www.foxsports.com/nba/stats" and get the leader in each statistics, such as the leader in scoring, in rebound and in assist.

As showed in Figure.6, it demonstrates how this function works based on the example of the leaders' statistics for year 2000.

```
Enter Option: 4
Enter a year you want to check (From 2000 to 2017): 2000
-----------------------------------------------------
                    Leader Data
-----------------------------------------------------
The leader in scoring: Iverson, Allen       PPG: 31.1
The leader in rebound: Mutombo, Dikembe     RPG: 13.5
The leader in assist:  Kidd, Jason          APG: 9.8
```

Figure.6 the Leaders Statistics for Year 2000

4. Conclusion

Compared with other NBA statistics online data, there are following advantages of our application. At first, it is clean and compact to see the most important statistics for players, which fans like us are most concerned about. Secondly, our statistics range covers all the players who have ever played in NBA. Finally, we have our own similarity comparison method to compare the similarity between two players and provide useful feedback for users.