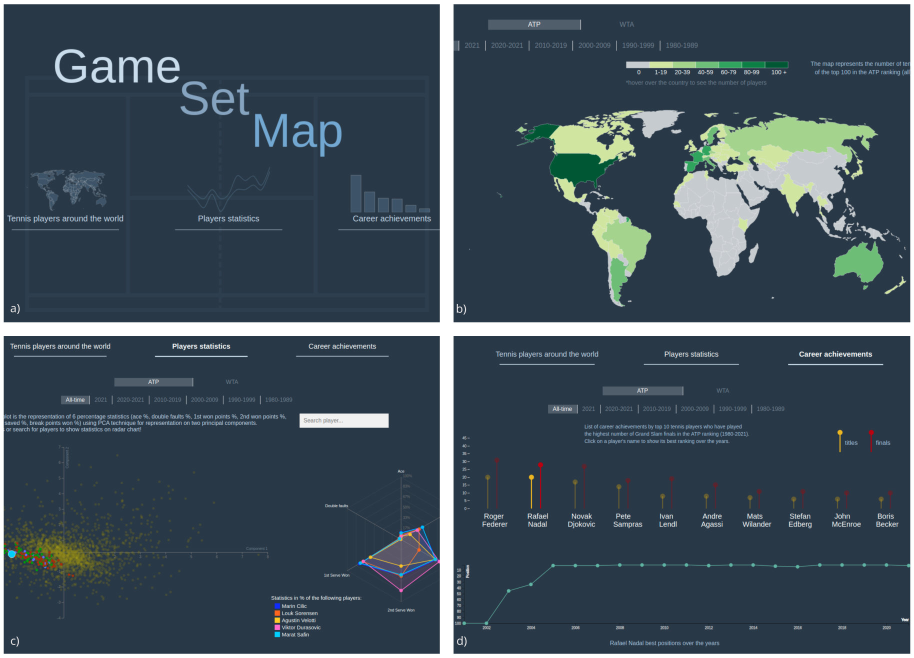
Game, Set, Map

Andrea Luca Antonini
”La Sapienza” University of Rome
`antonini.1707560@studenti.uniroma1.it`

May 28, 2021



The map represents the number of ter
of the top 100 in the ATP ranking (all

*hover over the country to see the number of players

plot is the representation of 6 percentage statistics (ace %, double faults %, 1st won points %, 2nd won points %,
saved %, break points won %) using PCA technique for representation on two principal components.
or search for players to show statistics on radar chart!

Statistics in % of the following players:

List of career achievements by top 10 tennis players who have played the highest number of Grand Slam finals in the ATP ranking (1980-2021). Click on a player's name to show its best ranking over the years.

Rafael Nadal best positions over the years

Figure 1: Framework Views: (a) Initial view, with three buttons, one for each sub-view; (b) Tennis players around the world view, with a map representing the tennis players nationalities distribution over the world; (c) Players statistics view, with a scatter plot and a radar chart; (d) Career achievements view, with a bargram showing the most winning champions of tennis and a relative connected scatter plot.

Abstract

The purpose of my application is to provide a framework able to graphically display data about tennis through a user-friendly interface. It’s possible to navigate through different views in order to get visual informations about this sport. The application allows to analyse various aspects of tennis statistics from 1980, based on the selected database, to nowadays (2021).

1 Introduction

Tennis is believed to have originated during 12th or 13th century in France. However, it was played by Major Walter C. Wingfield, in an hourglass shaped court in 1873 in Great Britain. It seems to have evolved from similar ancient sports like court tennis, squash racquets and badminton. This means that it is very difficult to navigate in its whole history, since it is so old that no found database contains every tennis data, from origins to nowadays. The goal of this project is to provide a way to navigate the history of tennis from 1980 to us, highlighting various aspects, such as particular statistics, nationalities, career achievements and sports career of a lot of great tennis players.

The framework is divided into four main views.

An initial view contains three highlightable buttons, each of which leads to a sub-view. The ”Tennis around the world” view shows the nationalities distribution over the world of tennis players in the top 100 ATP/WTA [1] [2] ranking over the years. In the ”Players statistics” view, a scatter plot shows the values of pre-calculated PCA over six statistics (ace %, double faults %, 1st won points %, 2nd won points %, break points saved %, break points won %), projecting the values of these statistics onto 2 principal components. Moreover, in this section there is a radar chart that, for every selected player, shows the percentages mentioned above, giving the possibility to compare up to 5 players. In the ”Career achievements” view, a bargram shows the 10 most winning champions, i.e. that have won more Grand Slams. Relatively to this, selecting a player it is possible to see his/her career best ranking in the ATP/WTA ranking over the years in the connected scatter plot below.

So, the framework is divided in two parts:

- an analytics part that computes the examination and exploration of tennis statistics;
- a visual representation of the features computed by the analytics part.

The project is a web application and it is developed using the d3.js [3] library and the Python3 language for implementing the preprocessing part on the chosen database.

2 Related works

There are other frameworks that deal with tennis statistics, in various forms. For example, Tennis Visuals [4] framework contains some graphs showing tennis statistics. The main difference is that in that project, specially in the first graphs, the attention is focused on the single matches, with every match analysed from the point of view of the single points. Another example could be a framework by Financial Times [5], called ”The greatest men’s tennis players ever”, in which there are a series of line charts, each representing an aspect of greatest champions careers, such as cumulative win percentage per age or per matches played and so on.

An interesting work based on the same dataset used for this project is ”Tennis match prediction using Machine Learning” [6], in which the dataset is used as base to run logistic regression on, in order to give a prediction on a tennis match with a quite high probability.

3 Dataset

The datasets used for this project are the Jeff Sackmann [7] ”tennis_atp” and ”tennis_wta” repositories available on Github. Each repository contains several *CSV* files, relative to ATP or WTA matches and rankings. For this project, I used only single players’ ATP or WTA matches, not considering doubles, futures and challengers. Since ranking and matches files before 1980 are resulted to be quite incomplete, I decided to take under consideration only files from 1980 on. Each ATP/WTA matches file is composed by several columns, but those used for this project are:

- winner_id: identification number of the match winner;
- winner_name: name of the match winner;
- winner_ioc: match winner nationality;
- loser_id: identification number of the loser;
- loser_name: name of the loser;
- loser_ioc: loser nationality;
- round: round of the tournament in which the match took place;
- w_ace/l_ace: number of aces, i.e. points directly from the service, of the winner/loser;
- w_df/l_df: number of double faults, i.e. points yield to the adversary with two consecutive service faults, of the winner/loser;
- w_svpt/l_svpt: number of winner service points by the winner/loser;
- w_1stIn/l_1stIn: number of first serves in the pitch from the winner/loser;
- w_1stWon/l_1stWon: number of points won with a first serve by the winner/loser;
- w_2ndWon/l_2ndWon: number of points won with a second serve by winner/loser;
- w_bpSaved/l_bpSaved: number of break points saved by winner/loser;
- w_bpFaced/l_bpFaced: number of break points faced by winner/loser.

Instead, in ATP/WTA ranking files it is possible to find the following information:

- ranking_date: date to which the corresponding rank refers to;
- rank: position in the ATP/WTA ranking;
- player: identification number of the player;

Finally, I used the ATP/WTA players file to bind the ID's in the various files to the corresponding names and nationalities.

4 Design Overview

The project is addressed to two kind of users: the simple tennis enthusiast or a person who works in this world (which can be a player but also a coach for example). Obviously these two categories of users can be interested in different information, for example a "simple" user can be interested in some statistics such as number of victories in the Grand Slams or nationalities distribution over the world; on the other side, a tennis insider can be interested in some other statistics, such as percentages of ace, double faults and other more detailed statistics.

4.1 Analytical Tasks

In order to satisfy these requirements, I have extracted the following high-level tasks for analysing the world of tennis:

T1 Analyse a single player performances. A single player performances are characterized by different aspects, some of them are: ace percentage (**T1.1**), double faults percentage (**T1.2**), percentage of points won with first server (**T1.3**), percentage of points won with the second serve (**T1.4**), break points saved percentage (**T1.5**) and break points won percentage (**T1.6**). Moreover, each selected player is compared with all the others on the basis of the number of victories on a surface (hard, clay or grass)

T2 Analyse single player career. Each player have a different history in terms of victories, and consequently in terms of ranking. So, analysing the best rankings from the start of career to nowadays, whether the player retired or not, can be a good index of the player's career trend.

T3 Analyse nationalities distribution. Analyse the nationalities distribution for every country in the world, in every time period from 1980 to nowadays.

T4 Analyse greatest champions. Analyse the greatest champions, i.e. those player who won more Grand Slams finals in both ATP/WTA, can give an idea of the magnificent exploits of those players.

4.2 Design Requirements

From the task requirements, I iteratively refined a set of design requirements to support the analysis. So I have mapped different analytic tasks to each design requirement:

Visualize nationalities distribution - D1. The framework should be able to provide to the user a way to visualize the countries that have at least one player in the ATP/WTA top 100 ranking (T3), in a determined period of time. For each country the user should be allowed to see how many players are in the top 100, referring to the selected period. Moreover, clicking on a specific country, the user should be able to see all the top 100 players coming from the selected country in the period of time of interest.

Visualize single player statistics - D2. The framework should allow the user to visualize specific statistics about a selected player (from T1.1 to T1.6) with the radar chart, or a global view about tennis statistics given by projecting these statistics on two principal components through PCA. Selecting a player among all top 100 players referring to a particular period, the user should be able to visualize all the players who have a similar number of victories on a determined surface. More in details, the dot colour corresponds to the surface on which the generic player has a number of victories more similar to the selected player.

Visualize career achievements - D3. The framework should be able to provide a way to visualize the most winning champions of the tennis history with their achievements (T4) and, for every player in this representation, his/her best rankings over the years (T2).

5 Visual Analytics Framework

The framework is composed, as anticipated in the introduction, by four main views. It is possible to navigate through the application using the upper left side arrow, that brings the user back the initial view, or by the three buttons in the header, each bringing the user to the selected view.

In the next sections each view will be explained more in depth, analysing especially the design.

5.1 Initial view

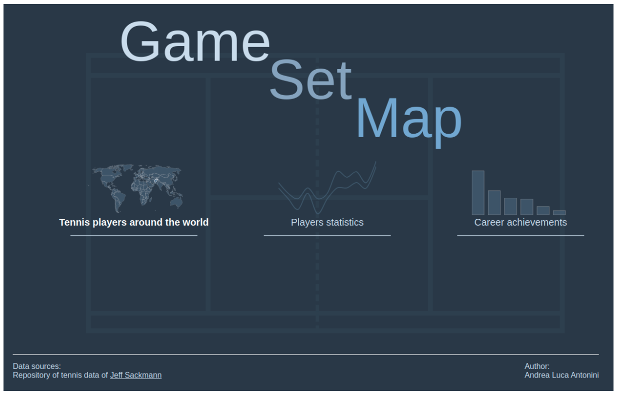
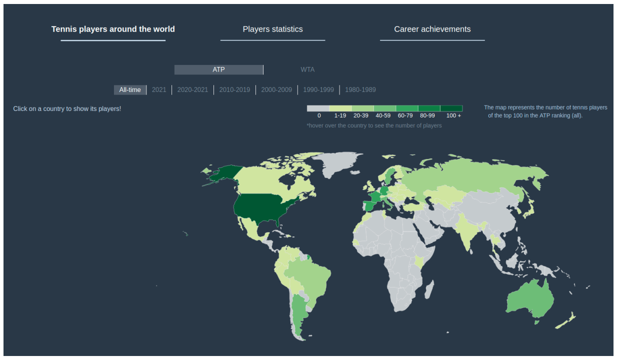


Figure 2: Initial view

This is the initial view. Here the user is able to select the desired section to visit, starting from ”Tennis players around the world”, passing from ”Players statistics” and ending with ”Career achievements”. In the footer, there is a link bringing the user to the dataset repositories.

5.2 Tennis players around the world view

In Figure 3 we can see the ”Tennis players around the world” section, in which a world map represents the nationalities distribution of the top 100 players of various time periods over the world. In particular, as we can see from the legend above the map, each country has a colour indicating that the number of players of that country is in the range of that colour. Indeed, colours follow a scale starting from light grey, passing from yellow and green, and finishing with a dark green indicating an important presence of top 100 players in that country.



Click on a country to show its players!

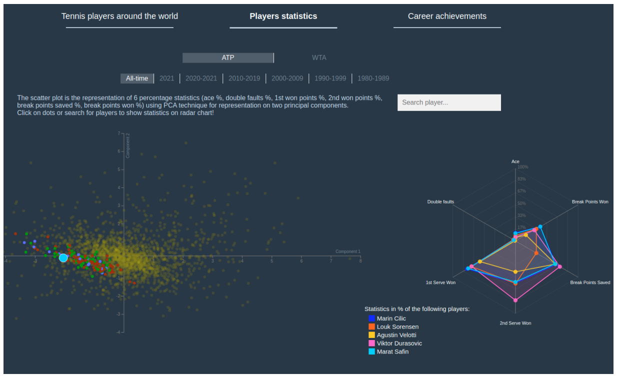
The map represents the number of tennis players of the top 100 in the ATP ranking (all).

*hover over the country to see the number of players

Figure 3: Tennis players around the world view

Through the radio buttons positioned above the legend, the user can choose between ATP or WTA rankings and, with the immediately below buttons, he/she can choose the time period for which visualize the nationalities distribution. The user can choose the period 1980-1989, 1990-1999, 2000-2009, 2010-2019, 2020, current rankings (2021) or all-time period. Passing the cursor over a country, the user should be able to see the name of that country and the corresponding number of players, causing its colour. Moreover, clicking on a country, the user will see on the left the list of all the top 100 players coming from that country in the selected period of time.
Finally, clicking on a player’s name, the user will be led to the ”Players statistics” view that will be discussed below.

5.3 Players statistics view



The scatter plot is the representation of 6 percentage statistics (ace %, double faults %, 1st won points %, 2nd won points %, break points saved %, break points won %) using PCA technique for representation on two principal components.
Click on dots or search for players to show statistics on radar chart!

Figure 4: Players statistics view

In this view, some statistics about single tennis players are shown.
There two sub-views: a scatter plot and a radar chart. The scatter plot contains all the players in the top 100 from 1980 to nowadays, representing them as initially yellow dots. Each dot coordinates are the result of the application of PCA technique on the statistics reported from T1.1 to T1.6: through this application it is possible to make a projection of six components onto two principal components (Component 1, Component2), that represents the coordinates.
The user can hover the dots to see the corresponding players’ name, and clicking on dot he/she can select

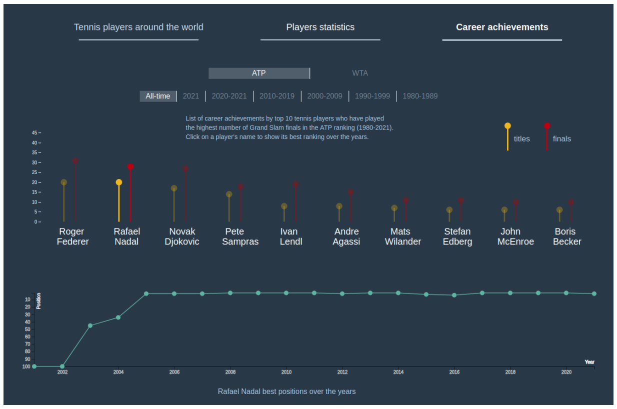
it to display its statistics on the radar chart. Moreover, the user can directly search for a player through the search bar above the graphs, and selecting a name from the search bar will cause the selection of the corresponding player on the scatter (red dot) and the display of detailed statistics on the radar chart.

As mentioned above, clicking on a dot will cause a change in the colour of all the other dots. More in details, each dot will have the colour corresponding to the surface (hard, clay or grass) on which the player has a number of victories ”similar enough” to the selected player. Those players who do not have ”similar enough” number of victories will remain in yellow. This classification can be useful for an hypothetical tennis player who wants to know which player has his/her same level of tennis and on which surface. Finally, the user can zoom whatever area of the scatter plot just selecting the region through clicking and dragging the cursor (this interaction is called *brush*). With a double click on an area of the scatter plot, the user can return to the original scatter plot view.

The second sub-view of this section is a radar chart, in which six percentage statistics relative to a player are shown. These statistics are, as reported from T1.1 to 1.6: ace %, double faults %, first serve points won %, second serve points won %, break points saved % and break points won %. These statistics, as will be explained in the Analytics section, are obtained considering all the matches played by the selected player from 1980 on.

The user is able to select up to 5 players from the scatter plot or from the search bar. These players are put in the radar chart, each with a different colour, as it is possible to see from the legend below the radar chart. In this way, the user is able to compare the statistics of the various players. Passing the cursor on a geometric figure relative to a player on the radar will cause the highlight of that figure; the same effect can be obtained clicking on a name in the legend.

5.4 Career achievements view



List of career achievements by top 10 tennis players who have played the highest number of Grand Slam finals in the ATP ranking (1980-2021).
Click on a player's name to show its best ranking over the years.

Rafael Nadal best positions over the years

Figure 5: Career achievements view

This view is designed to show the greatest achievements of tennis champions. In fact, the first graph in the section, a bargram, shows the 10 most winning champions in the ATP/WTA history, i.e. those who won more Grand Slams.

Each player has two bar, representing his/her Grand Slam finals number (red bar) and his/her Grand Slam titles (yellow bar).

Clicking on a name on the bargram x axis, the user can see the best positions in the ATP/WTA rankings from the beginning of his/her career to nowadays.

In fact, once the user has clicked on a name, the connected scatter plot below will be populated with data corresponding to that player. For each year in which the selected player has been in activity, there is a dot that tells the user the year and the best ranking in that year. All the dots are connected by a line to better showing the passage of time.

6 Analytics

This section is more focused on data analysis and how to obtain, starting from the dataset, the views described above.

Most of the views are produced by manipulating the dataset in different ways starting from an input given by the user. This manipulation is done using the *d3.js* [3] library, since it allows to do operations easily on *.csv* files, and using *Python 3* (in particular the *pandas* [8] and *scikit-learn* libraries [9]) for developing all the pre-processing operations done on the dataset. Quite every graph in this project has a part of pre-processing in Python3, since there is a big amount of data to take into consideration and, in some cases, there is a part of computation.

Examples could be the scatter plot radar chart in the ”Player statistics” view (5.3): in these cases I needed the calculation of all the percentages for every player in the top 100 from 1980 on. So, first of all I made a `JOIN` between the *id* attributes in the rankings files (from 1980 on) and the *id* attributes in all the matches files, `WHERE` the rank of the player is between 1 and 100, and then I updated the global statistics relative to each player taken under consideration. Then, I made up all the wanted percentages taken under consideration in this way:

- ace %: total number of aces / total number of service points;
- double faults %: total number of double faults / total number of service points;
- first serve points won %: total number of points won with first serve / total number of first services in the pitch;
- second serve points won %: total number of points won with second serve / total number of second services in the pitch;
- break points saved %: total number of break points saved / total number of break points faced;
- break points won %: 1 - (total number of break points saved by the opponent / total number of break points faced by the opponent)

At this point, concerning the scatter plot, I needed to represents all these statistics in two principal coordinates.

In order to do this, I have decided to use the PCA algorithm, that is a statistical procedure that allows to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analysed. The computation of PCA is done using Python3 and, in particular, using the implementation provided by the scikit-learn library [9]. I computed PCA over this new sub-dataset performing the following steps:

1. Standardize the features (such that the variance of the data is 1 while their mean is 0);
2. Perform the dimensionality reduction using the PCA algorithm (the data are projected on a 2D space).

Moreover, another analytics part in the scatter plot is that concerning the dots classification given a selected player. The selected player number of victories on each surface is compared with those of all the other players. Then, the surface on which there is the smallest difference between the number of victories is chosen to give the colour to each dot.

Concerning the interactive world map in ”Tennis around the world” view (5.2), what I did in the pre-processing phase is to make a `JOIN` between the players file elements and the ranking files, based on *ids*, `WHERE` the rank of the player is between 1 and 100. Then, every time a new player was found, a counter for the respective country was updated, as well as the list of the players incrementing the counter.

Similar operations were done for the bargram in ”Career achievements” section (5.4), in which instead of counting how many top 100 players there are in a country, I counted how many Grand Slam finals a player has played and how many of them have been won. Finally, I ordered the elements in a descending way, so that the first 10 most winning champions could be shown in the bargram.

7 Future Works

As the dataset provides a lot of data, it's possible to add more views to the application.
For example, it could be possible to add a view showing the results of a work similar to the above cited ”Tennis Match Prediction using Machine Learning” [6]. Another source of improvement for this project could be to add a part taking under consideration single matches, with point-to-point representation of all the matches of ATP/WTA tournaments.

8 Conclusion

The project is useful to be able to retrace the history of tennis easily, through intuitive graphics and insights. The application could be used for different purposes, for example by analysing the time over the years it is possible to guess how players characteristics and statistics have changed, while at the same time it can be used as a roll of honour of all times greatest tennis players.
At the same time, a tennis player or a coach can retrieve graphical useful information about himself/herself or his/her learner, studying the comparison with other tennis players from all over the world. Learning how to use the framework is very simple, as I have chosen to use immediate and intuitive graphics to give to the user a light but interesting experience.

References

[1] ATP: https://en.wikipedia.org/wiki/Association_of_Tennis_Professionals

[2] WTA: https://en.wikipedia.org/wiki/Women%27s_Tennis_Association

[3] Data-Driven Documents: https://d3js.org/

[4] Tennis Visuals: http://tennisviz.blogspot.com/

[5] Financial Times ”Greatest men’s players ever”: https://ig.ft.com/features/baseline/greatest-tennis-players-of-all-time/

[6] Tennis Match Prediction using Machine Learning: http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw

[7] Jeff Sackmann: https://github.com/JeffSackmann

[8] Pandas library: https://pandas.pydata.org/

[9] Scikit-learn library: https://scikit-learn.org/stable/