Agata Łysakowska
ID number: 394031

Web Scraping and Social Media Scraping Project

Topic and web page

Web page that I scrape is https://www.tennis.com/rankings/WTA/. It contains tour ranking of Women's Tennis Association (WTA). This page shows current ranking, players name, country and ranking points. Moreover if you click on the players name it takes you to the web page of a player. From this point you can go to tab 'STATS' and the page will show information about birthdate, height, weight and dominant hand (plays) of a player. In some cases you already see these information without going into 'STATS'.

Scrapers

In order to scrape information mentioned in previous paragraph I wrote three scrapers: one using Beautiful Soup, one using Scrapy and one using Selenium. All of them have Boolean parameter 'bool' that if it is True it will scrape only 100 pages (99 links + 1 main page).

Beautiful Soup

Program that uses Beautiful Soup firstly gets all the information (ranking, name, country and points) from the main web page. Then it connects them into one data frame. Next it finds all the links under the players' names. For each link it gets birthdate, height, weight, plays and name of a player and adds it to new data frame. At the end it merge two data frames into one on players name. It also saves result to .csv file.

This method is the second fastest in scraping all the information.

Scrapy

I wrote three scrapy programs in order to scrape the web page. First one extracts links to the player page from the players name tag. I used xpath to find tags and its attributes (@href) and then I created full web page address.

Second scrapy uses links scraped by the first scrapy in order to get name, birthdate, height, weight and plays information.

Third scrapy scrape the main web page (https://www.tennis.com/rankings/WTA/) in order to scrape ranking, name, country and points of a player.

The output you will get from these are two separate .csv files with the players information (and one with links). You can use them separately but I also wrote a program that merge them into one (it is not a scrapy). I use columns ‘name’ in both files to match the information with the player name.

This method is the fastest in terms of scraping the information. It may be not as convenient as the others because you have to run three separate scrapy program and then additional program to have everything in one .csv file.

Selenium

This method, similar to Beautiful Soup, firstly scrap the main page in order to find ranking, name, county and points. Then it finds links for the next drivers. For every link it clicks on it and then clicks on the ‘STATS’ tab if it is there. Next it gets birthdate, height, weight, plays and name and goes back to the main page. Afterward program combines data frames and merge them into one on players’ name and saves it to .csv file. At the end it closes the web page.

This is the slowest method. It takes a lot of time because it opens every web page and waits until it is load to scrap the information. An advantage of this method can be that you can see where your program exactly is.

Output

The output of these programs are .csv files with 8 columns. Column ‘rankings’ is numeric and it tells what place in ranking certain player has. Column ‘name’ is first and second name of a player. Column ‘countries’ is a country of a player. Column ‘points’ is numeric and tells how many points player has. Column ‘birthdate’ is a birthdate of a player. Column ‘height’ is height of a player in cm. Column ‘weight’ is a weight of a player in kg. Column ‘plays’ tells which hand is dominant hand of a player.

Data analysis

	points	height (without null and without outliers)	weight (without null and without outliers)
Mean	1 422	173	73
Min	352	155	50
Max	9 655	187	80

plays	Number
Left-handed	55
Right-handed	117
Unknown	2
Sum	174