Universität Konstanz

Introduction to Computation for the Social Sciences

Assignment 5

Prof. Dr. David Garcia, Áron Sármány
Winter Term 2022 / 2023

Please solve the exercises below and commit your solutions to our GitHub Classroom until November 30, 23:59. Submit all your code in one executable file (*py* / *ipynb*). You can score up to 10 points (plus 2 bonus points) in this assignment. You will get individual feedback in your repository.

Exercise 1: Simple Web Crawler (7 Points)

Web Crawling, i.e. the automated access of web resources via software, and web scraping, i.e. the extraction of content from web resources, are essential techniques to gather data for many analysis tasks in the social sciences.
In this exercise, you need to collect data about the FIFA World Cup in Qatar from the following websites:

https://us.soccerway.com/
https://www.transfermarkt.us/

Implement the following steps:

- Open the following link and collect the names of the participating teams, and the links leading to the detailed information about the teams (the link is in the same element as the name of the country).
- Store the collected data in a pandas dataframe
- Open one of the extracted links in your browser – you will see a tab called 'Matches' on the page, which contains information about the previous results of the teams. Figure out how to change the previously extracted URL to access this information.
- Extract the results of the last 5 matches (before the World Cup for each team), and store the number of points collected by the team in these matches (victory: 3 points, draw: 1 point, loss: 0 points). E.g. a team which won 2 of the last 5 matches, and lost 3 would have 6 points. Add this value to your dataframe as a column.
- Also extract (from the same URL) the number of points achieved in the first match of the World Cup. Add this value to your dataframe as a column.
- Wait 1 second before accessing the next URL to reduce the load on the website

- Open the following link and extract the FIFA ranking score of each of the teams participating in the World Cup. Add this score to your dataframe as a column.
- Open the following link and extract the market value and the confederation (e.g. UEFA, CAF, etc.) of each of the teams participating in the World Cup. Add these to your dataframe as a column.
- Save the collected information (name the team, link to its page on *soccerway.com*, number of points collected in the last 5 matches before the World Cup, number of points collected on their first match of the World Cup, FIFA ranking score, confederation, market value) to a CSV file in your solution folder.

There are many correct solutions for how your crawler could be structured. As general advice, try to write as much code in a way that it can be reused and reapplied to different tasks.

Exercise 2: Visualization (3 Points)

Explore your data and visualize some aspect(s) of it you find interesting (e.g. average market value of national teams by continent). Label your plot(s) and make it as informative as possible.

Bonus Exercise (2 Points): Prediction

Use the data you collected previously and collect any additional data needed (except betting odds, other people's predictions, etc.) to train a simple model (e.g. a Linear regression) to predict the results of the matches on December 1 and 2 (Thursday and Friday). Try to achieve as high accuracy as you can.