web_scraping_day01

January 22, 2024

Web scraping means extracting data from websites in a automated way. In 1998, Google was the first to do so; Search engines are using web scraping to retrieve HTML tags from public website in order to rank them. Web scraping is also used for collecting data when no official API is available. Be careful, web scraping may by illegal depending of the country and legislation.

0.1 EXERCICE 0

Perform a GET request to https://www.leboncoin.fr to retrieve the HTML homepage. You must use an user agent to do so, as basic web security prevent HTTP requests from unknown web browser. You can give this one a try : Mozilla/5.0 (Windows NT 10.0;Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/79.0.3945.74 Safari/537.36 Edg/79.0.309.43

0.2 EXERCICE 1 (10 PT)

With the help of the Exercice0, and the User-Agent: Create a function get_ps5_prices() that returns data from Playstation 5 game console sold on the website using BeautifulSoup python library. Route : /recherche?category=43&text=ps5 You must retrieve for each ads - **the title** - **the price of the article** (sellers may not have set a price for the article, put 0 instead) - **the date** of when it was posted as ISO8601 format - **the city** - **the postal code**

Store all the data into a pandas dataframe.

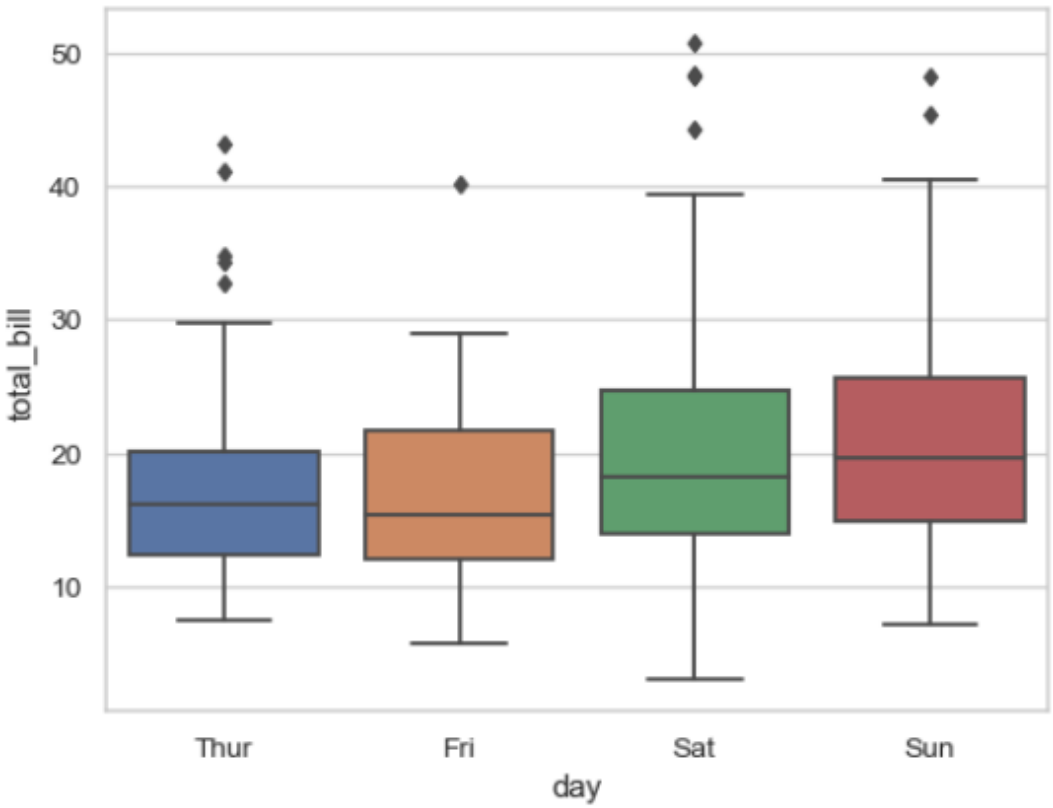
TIPS The HTML you receive from your HTTP request is a basically a snapshot of the website as if the search was done from a web browser such as Firefox. You should do this search from your web browser at the same time and use the web inspector to identify which HTML tags are relevant to get the data correctly.

0.3 EXERCISE 02

Rather than exporting the pandas dataframe into SQL database, export it as a file stored on your computer. This is called serialization. Export your data using pickle, name the file ps5-dataframe.pickle Create a loop that call get_ps5_prices function every 5 minutes using time.sleep(300) Only if a new article was published as compared to the previous iteration, no duplicate data, export it again.

0.4 EXERCISE 03

Create a new notebook call ex03 Create loop. Every 5 minutes, open ps5-dataframe.pickle in read-only mode in order to have your pandas dataframe back. Using Seaborn, box plot the price for each day.


50
40
30
20
10
total_bill
Thur
Fri
Sat
Sun
day