From WebSites to Datasets: Unleashing the Power of Data Harvesting with Python

September 9th, 2023

A workshop on python and its power on applied web scraping

Gonçalo Marques

Software Developer

Thoughtful Reminders

Here’s some quick reminders!

01 This is my first time organising a workshop

02 Wish me luck

03 Feel free to ask any questions

What will we learn?

Sumarisation of some topics we'll cover this afternoon

- What's webscraping and why is it important?
- Some examples on using webscrapers
- State of the art
- How to not get banned from the internet
- Limitations of what we are about to develop
- Design Patterns
- Data Structures

About Me

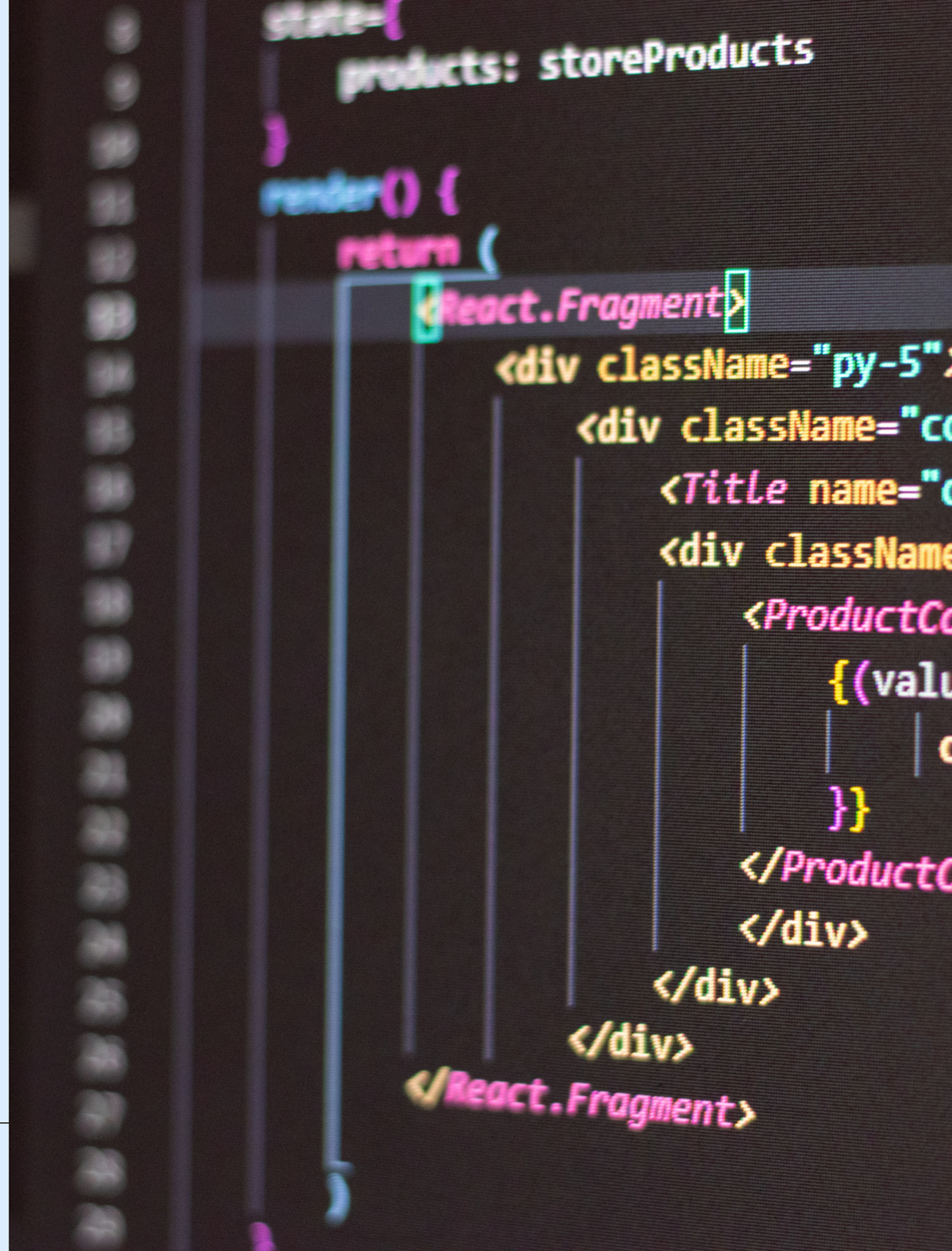
I am Gonçalo, from Aveiro

I am a student at University of Aveiro, I am currently enrolled on a Masters Degree on Computer Engineering.

I've been trying to contribute more and more to the open source community, and this year was my revelation year. My python library hit 2k downloads and a couple dozens of stars on GitHub.

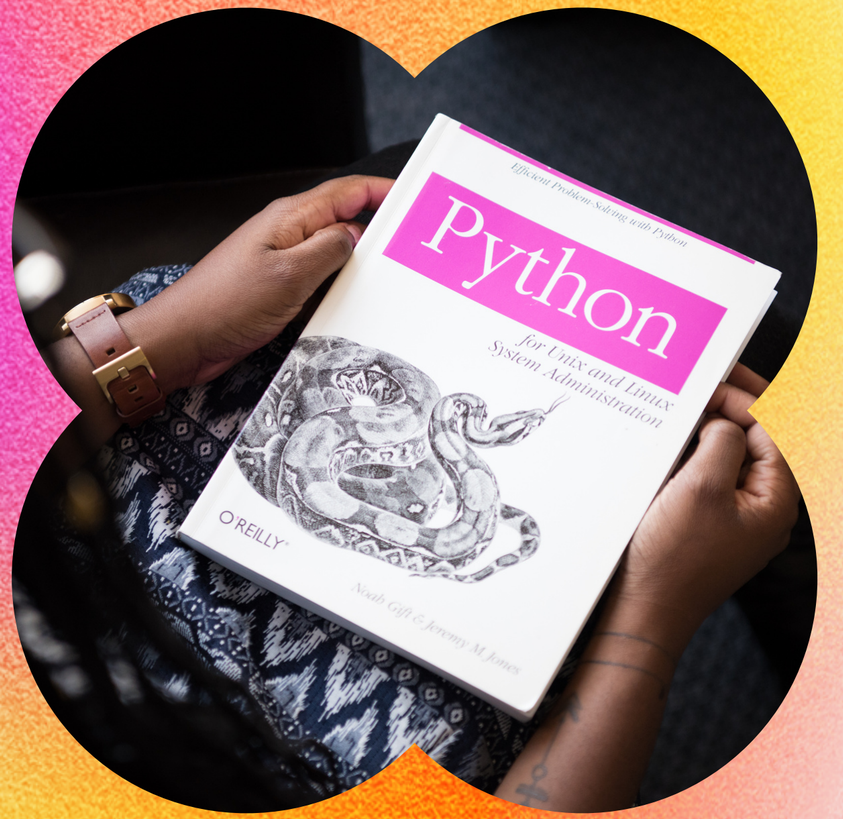
Webscraping, uh?

Involves knowing a bit of how the web works and how it’s structured.
It’s widely used for various purposes, like data collection, analysis, and automation.



But what can we do with them?

What’s your webscraping project about?



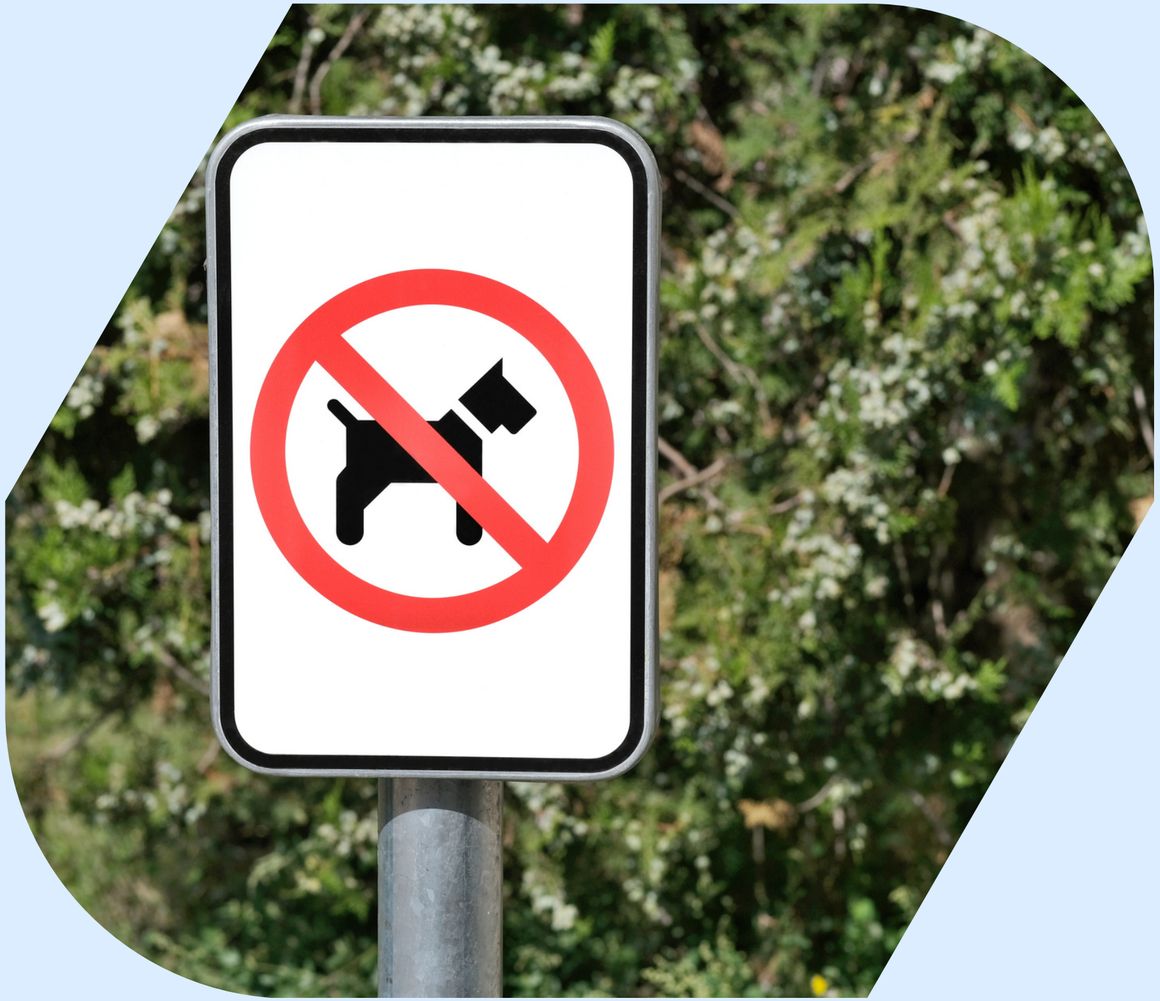
And what have we done?

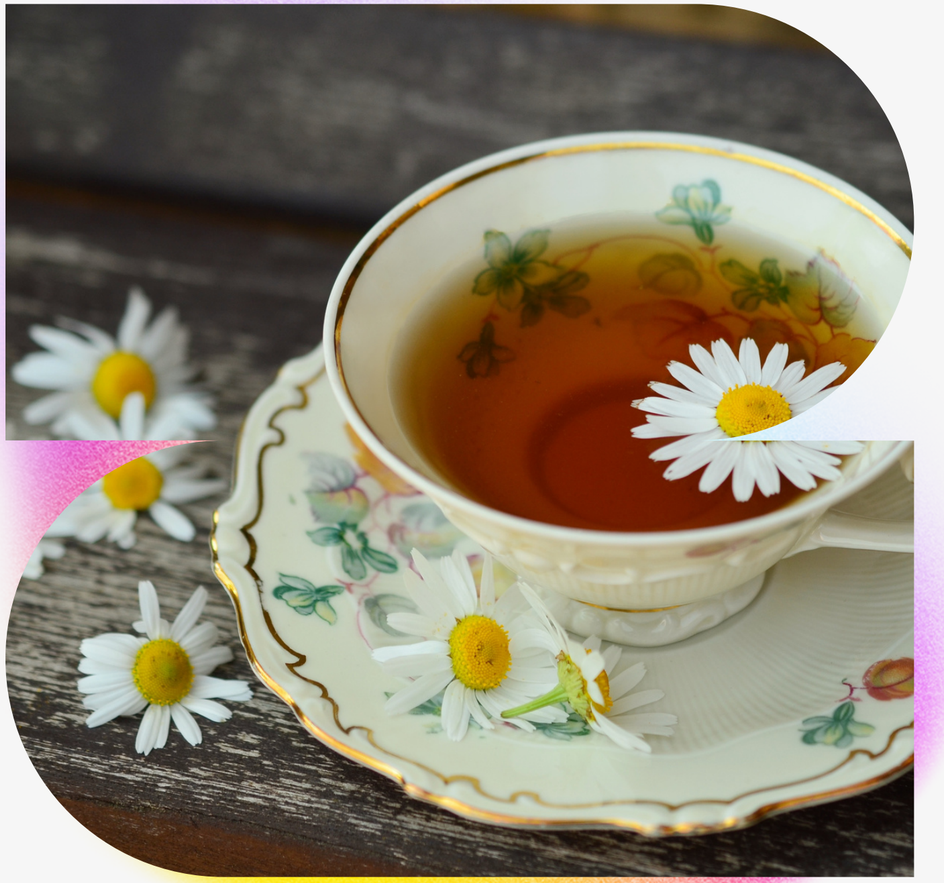
State of the art brings us to a couple of well known libraries

Selenium, BeautifulSoup4, Scrapy, etc...
What do they all have in common?

Nice, you got me banned from IMDb...

How can we avoid rate-limiting and also “respect” the website?





Quick Disclaimer

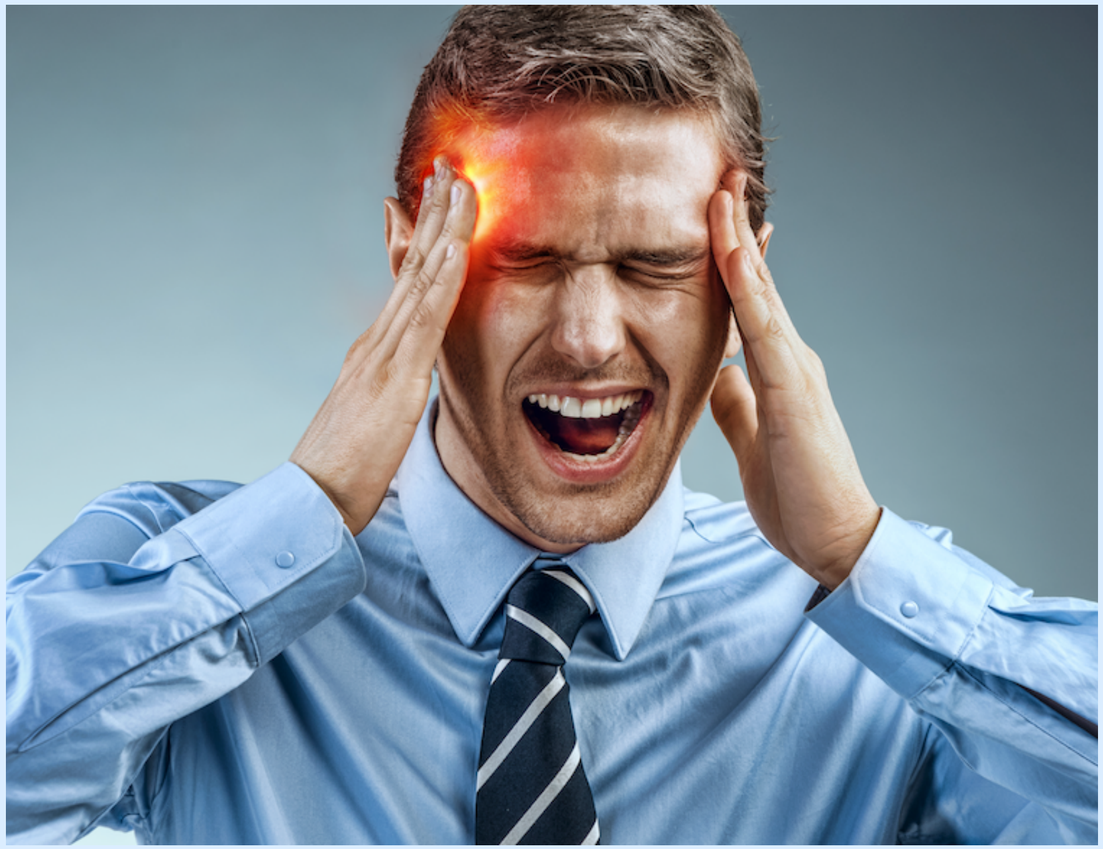
Time to be honest

We can’t make the best multiparadigm webscraper, crawler, most efficient, scalable product out there in 2 hours! But we can get you another excuse to code on your free time.

Liskov what?

My favorite is dependency inversion/injection

I do miss interfaces sometimes...



Can you relate?

LET'S GET SOME WORK DONE!

We will be building a config-based webscraper, using python dictionaries as our configuration provider.

Feel free to make your own, but on this session we will be focusing on making an audio classification dataset, specifically a bird song one.

The website is called Xeno-Canto

THANK YOU!

Feel free to connect on LinkedIn, GitHub, etc...