

# Agenda

- ① What is Web-Scraping
- ② Understanding Website Structure
- ③ request module
- ④ beautiful Soup for Scraping
- ⑤ Fixing the format

## What is Web-Scraping

### Goal

- ① To extract data from internet

Ex : ① E-commerce (Classification)  
② Stack-overflow (Coding LLM)  
③ Quora (QnA)

① Is it legal?

② Non commercial use-case  
Legal

(GDPR)

Issues when scraping:

- ① Legal Laws
- ② Website requiring Login

Look out for "robots.txt" or

ToS  
Terms  
&  
Service

# Understanding Website Structure

HTML: Hyper-text markup Language

<tag>

</tag>

Every component on a web page has a tag, class and id

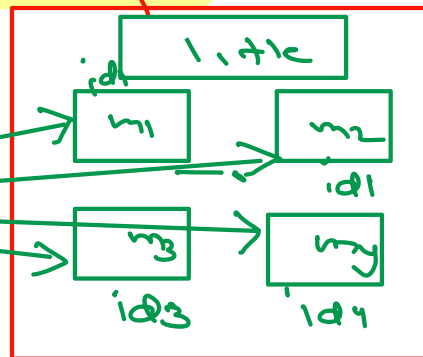
\* class

o Used for styling

o Multiple components can have same class

\* id is always Unique

class 1234



<div>

<div> — </div>

<div> — </div>

</div>

## Problem Statement:

- We'll be scraping this website <http://books.toscrape.com/index.html> for collecting book data like, name of book, genre, price, reviews, etc.
- Before scraping, it is recommended to check the website and understand its structure, what data it has.
- As per the current task, we have around 1000 books, distributed on multiple pages. We will scrape the data of all of books.

<http://books.toscrape.com/>

# Steps to Scrape :

- ① Send request to <http://books.toscrape.com/> and parse the Home-Page
- ② Scroll through all the search result pages
- ③ For Every search page, Scrape the Url of all the Book present on the page:
  - ① For each book, visit the Book's page and Extract information From Table
  - ② Parse and store table into Dictionary
  - ③ Post process to convert info into right Data-type.

# request module

Python module that enables interacting with web-services and APIs directly from code

HTTP methods for interacting:

• Get

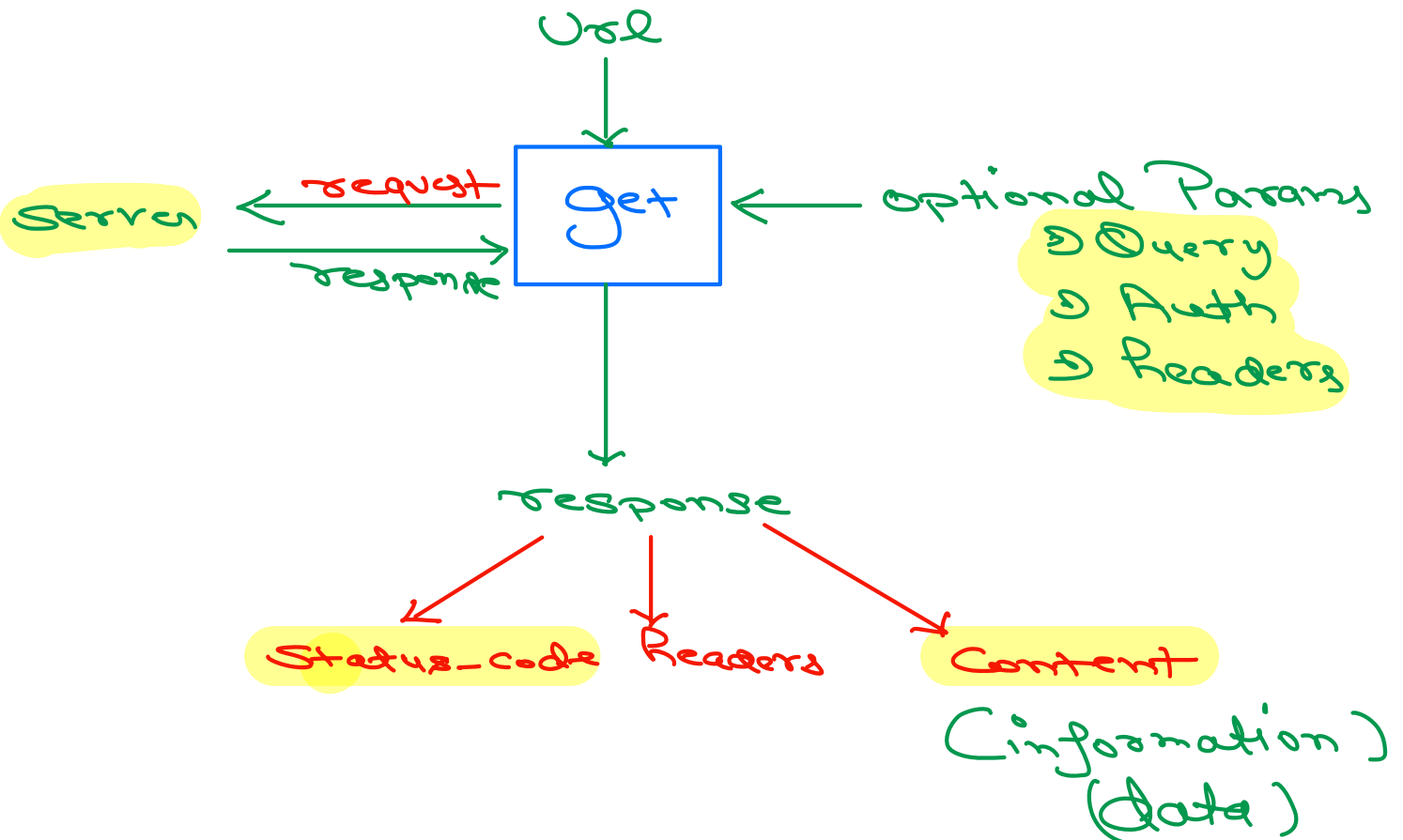
• Post

• Put

• Delete

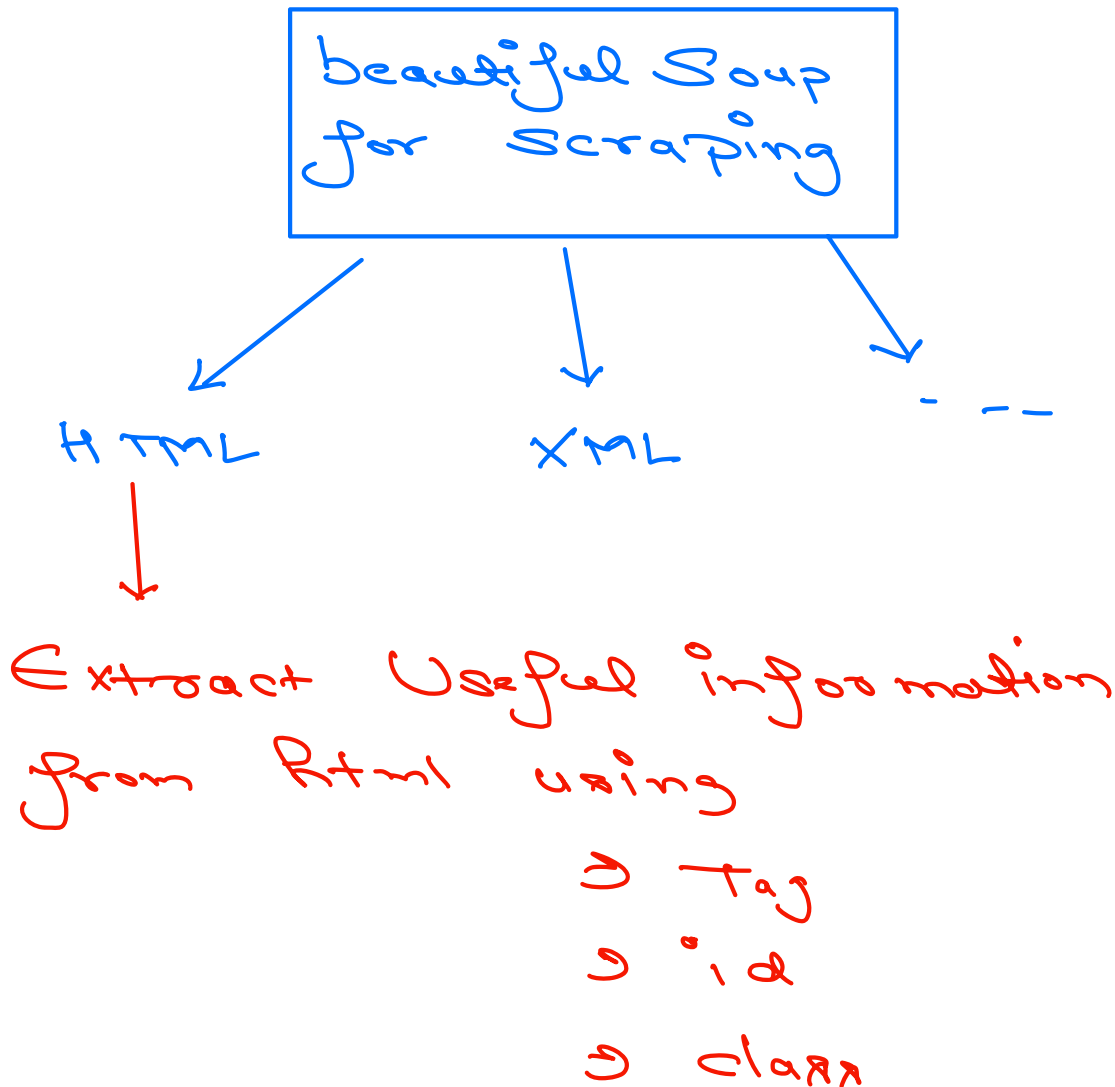
Explore

etc.



## Information about some common status codes :

1. **200 OK:** The request has been successfully processed, and the server returns the requested content. ✓
2. **404 Not Found:** The requested resource or page could not be found on the server.
3. **403 Forbidden:** Access to the requested resource is forbidden or not allowed for the client.
4. **500 Internal Server Error:** The server encountered an unexpected error while processing the request.
5. **302 Found (or 301 Moved Permanently):** The requested resource has been temporarily (or permanently) moved to a different URL, and the client should follow the redirection.



```
{'name': 'Tipping the Velvet',  
'UPC': '90fa61229261140a',  
'Product Type': 'Books',  
'Price (excl. tax)': '£53.74',  
'Price (incl. tax)': '£53.74',  
'Tax': '£0.00',  
'Availability': 'In stock (20 available)',  
'Number of reviews': '0',  
'url': 'http://books.toscrape.com/catalogue'}
```

str

53.74

pl. at

available

stock

20  
134  
0