Agenda

- What is Web-scraping
- Understanding Website structure
- request module
- beautiful Soup for scraping
- Fixing the format

What is Web-scraping

Goal

- To extract data from internet using Automated Script
  - Ex: - E-commerce (classification)
    - Stack-overflow (coding LLM)
    - Quora (QnA)

Is it legal?

- Non commercial use-case legal (GDPR)

Issues when scraping:

- Legal Laws
- Website requiring Login
  - Pass authentication as parameter
  - → Selenium

Look out for "robots.txt" or ToS (Terms of Service)

* A website is written using HTML
* Every component of website can be identified using following things
  * Tag
  * 'id'
  * Class

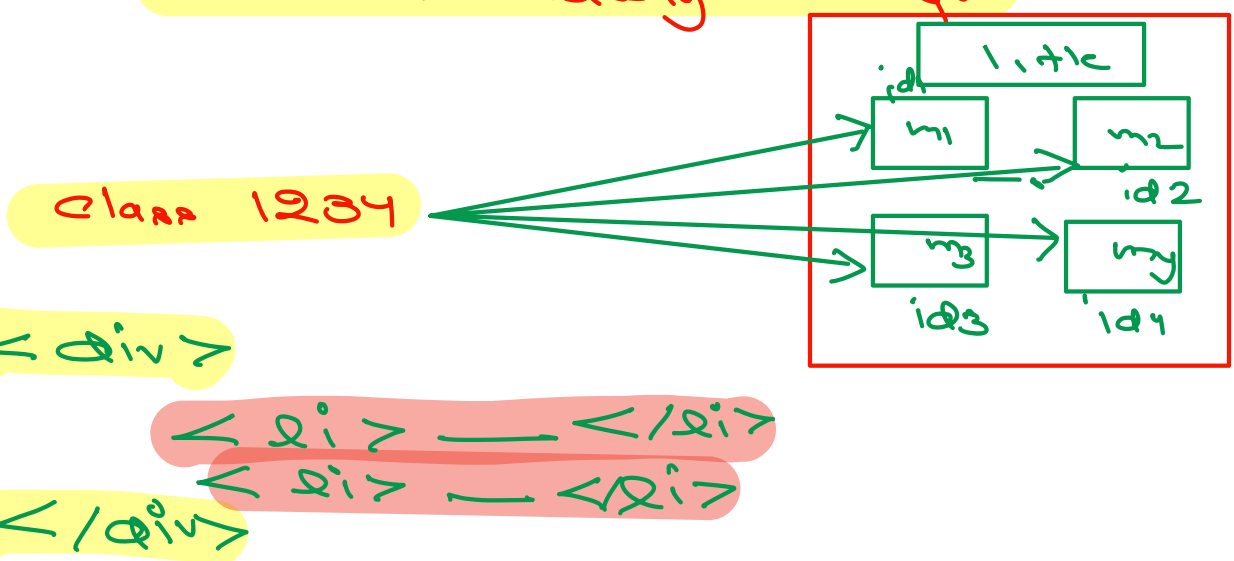
Understanding Website structure

HTML: Hyper-text markup Language

<tags> </tags>

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

* class
  - Used for styling
  - Multiple components can have same class
* ids are always Unique



 —
 —

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 it's 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

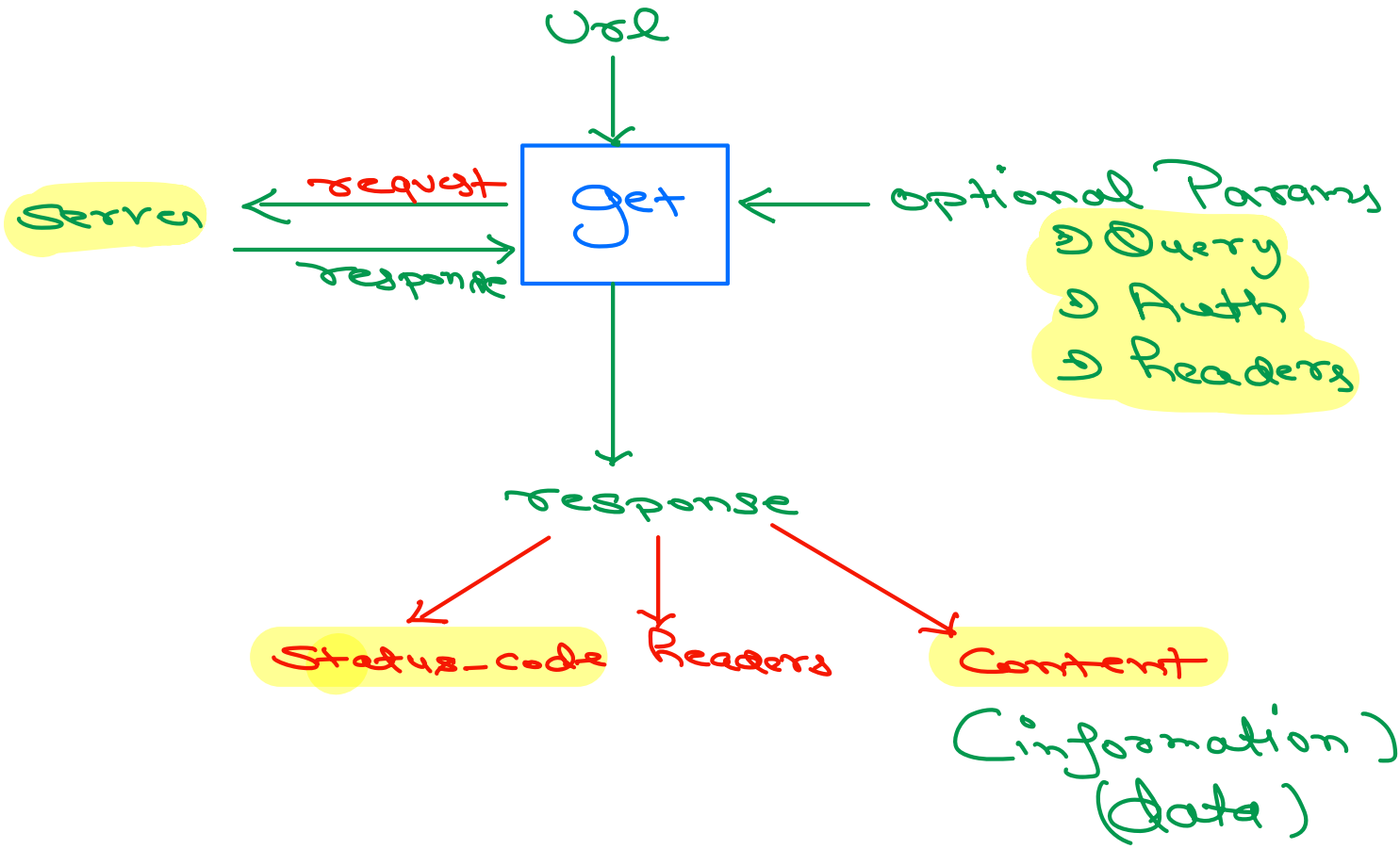
Python module that enables interacting with web-services and apis directly from code

HTTP methods for interacting:

- Get
- Post
- Put
- Delete

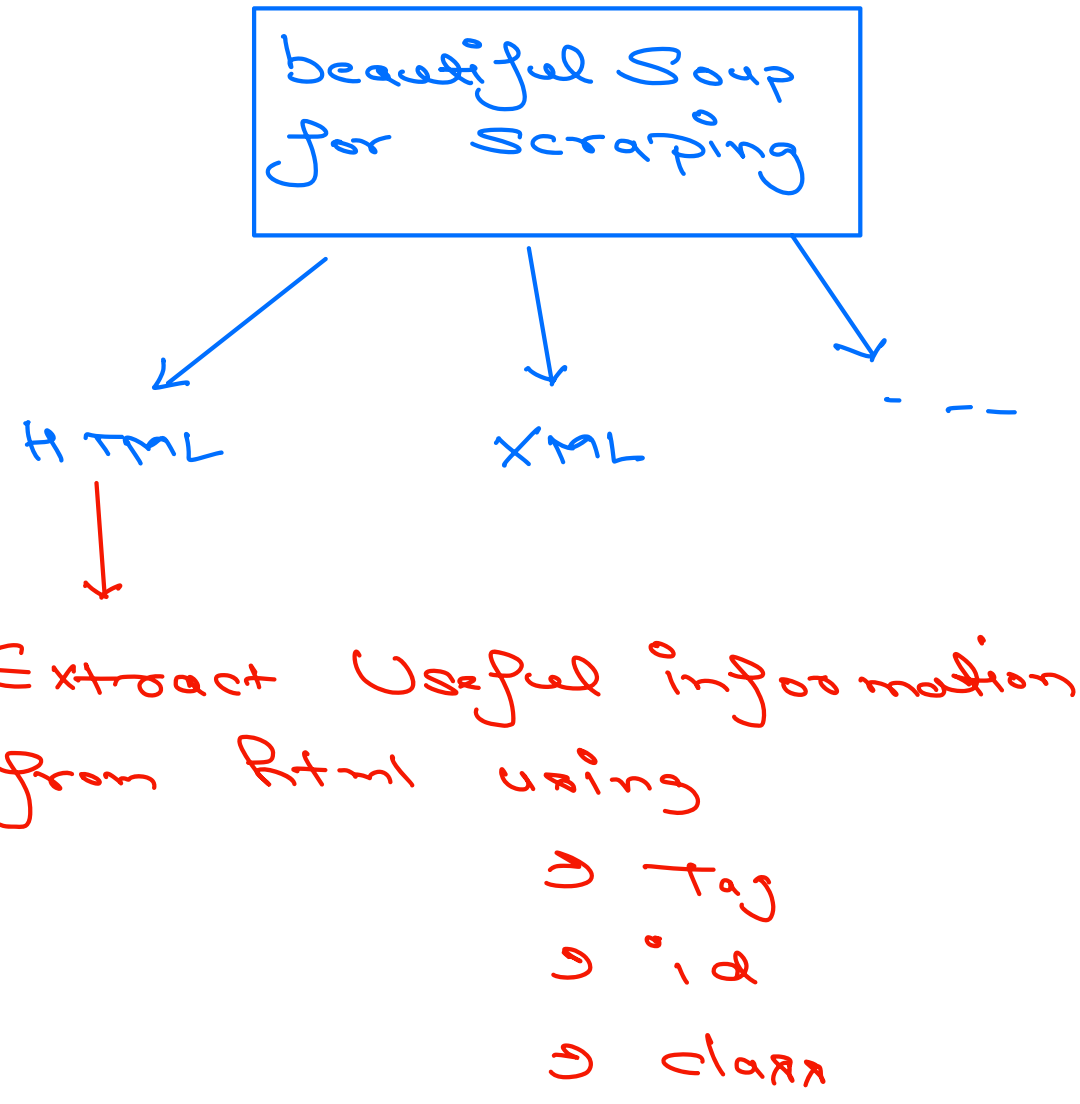
} Explore

etc.



Information about some common staus 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
```