Request

Presenter: Weikun Ma

What is Web Scraper

- Data scraping used for sending requests to website and extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol(http), or through a web browser.

Web Scraping Procedure

Send Requests

Send a requests to target website with Headers, etc

Receive Response

The requests will bring us a responses with Json, HTML page,etc

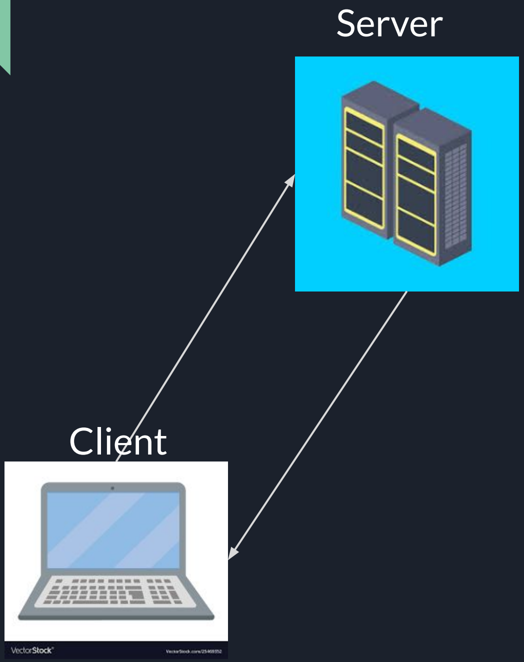
Analyze Contents

We can use Regex or other to analyze the pages and contents

Save the contents

Save the contents to local memory or disk

What is Requests and Response



1. Client send requests to Server called HTTP requests
2. Servers responses with HTTP response back to the browser
3. Browser analyze and shows the contents on the website page

Let’s see some real world requests

Baidu Website:

What is Request

- The Hypertext Transfer Protocol (HTTP) is designed to enable communications between clients and servers.

- Methods:
  - Mainly: GET POST methods, we also have PUT, DELETE, HEAD, etc
- GET requests: Send a requests with <u>info inside the URL link</u>
- POST requests: Send a requests with a <u>form data</u>
- Get & Post methods will return response body with information we need.

Request Continue

A request header is an HTTP header that can be used in an HTTP request, and that doesn't relate to the content of the message.

- User-Agent: software (a software agent) that is acting on behalf of a user.
- Request URL: Uniform Resource Locator that defines our requests location
- Cookies: Cookies were designed to be a reliable mechanism for websites to remember stateful information (such as items added in the shopping cart in an online store) or to record the user's browsing activity
- Form Data: in post request, we will have data stored in the form along with the requests body

User Agent

- In most cases, we need to include our own user agent in the requests body
- Otherwise, the website may not allow us to pass the requests
- But what is my User-Agent?
- Click mouse right bottom → inspect → network → select a request → go down to the bottom → you will get your User-Agent

Requests

```
Headers = {‘User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.116 Safari/537.36}

Response = requests.get(‘http://www.baidu.com’, headers=headers)
```

What is Response?

- Status Code
  - Check if the requests is successful or not
- Response Headers:
  - Cookies, server info, etc
- Response body:
  - The content we requests: including html pages, images, source, etc

What are some status_code

1. Informational responses (100–199)
2. Successful responses (200–299)
   a. 200 Ok
3. Redirects (300–399)
4. Client errors (400–499)
   a. 404 Not Found
   b. 403 Forbidden
5. and Server errors (500–599)

What kind of Data can we scrap?

- HTML
  - Typical Text file with a lot of sections
- JSON
  - Just like dictionary in Python
- Image
  - We can retrieve the info from the url