Web Scraping

Data Science Lab

Sukanya Nath

11.06.2024

About Me

Sukanya Nath

Education		
2018-2021	Ph.D. in Computer Science (Computational Linguistics)	@ University of Neuchatel
2016-2018	MSc. In Computer Science (Data Science Specialization)	@ University of Bern
2009-2013	BTech in Computer Science	@ National Institute of Technology, Silchar, India
Teaching		
Since 2023	Module Lecturer (CAS NLP)	@ University of Bern
Work Experience		
2023-present	Scientific Collaborator	@ University of Bern
2021-2023	Scientific Collaborator	@ Swiss Distance University of Applied Sciences Brig
2020-2020	Research Assistant	@ University of Applied Sciences Bern
2016-2017	Data Science Intern	@ BEDAG AG
2013-2015	Software Developer	@ Samsung Research Institute India

What is DSL?

Support
Consulting, scientific software support, etc.

support.dsl@unibe.ch

Training
Machine Learning, Data Science, Digital skills

Data Science Lab
Core facility working for the all faculties

Service Level agreements
software projects, tailored support, IT setup

Labs

Institutes

Faculties

Science Partners
Common applications for grants

CAS Programs, Trainings and Winter Schools

1. Training: Continuing Education Programs - Data Science Lab (unibe.ch)
2. Bern Winter Schools on Machine Learning (unibe.ch)
3. Training: Upcoming Training - Data Science Lab (unibe.ch)

Introduction to Data Scraping

What is Web Scraping?

- The process of extracting data from a website.
- While data can be requested via APIs, scraping can give access to non API data and also access dynamic content.

Introduction to Data Scraping

Why is Web Scraping necessary?

- While data can be requested via APIs, scraping can give access to non API data and also access dynamic content.
- Automating the process of extraction can speed up the process and help to gather more data.

Introduction to Data Scraping

What are some challenges to data scraping?

- Data is site specific.
- Dynamic content of web pages.
- Web pages are not single documents
- Web sites keep changing causing scraping pipeline to break.
- Unstructured data and additional website security features.

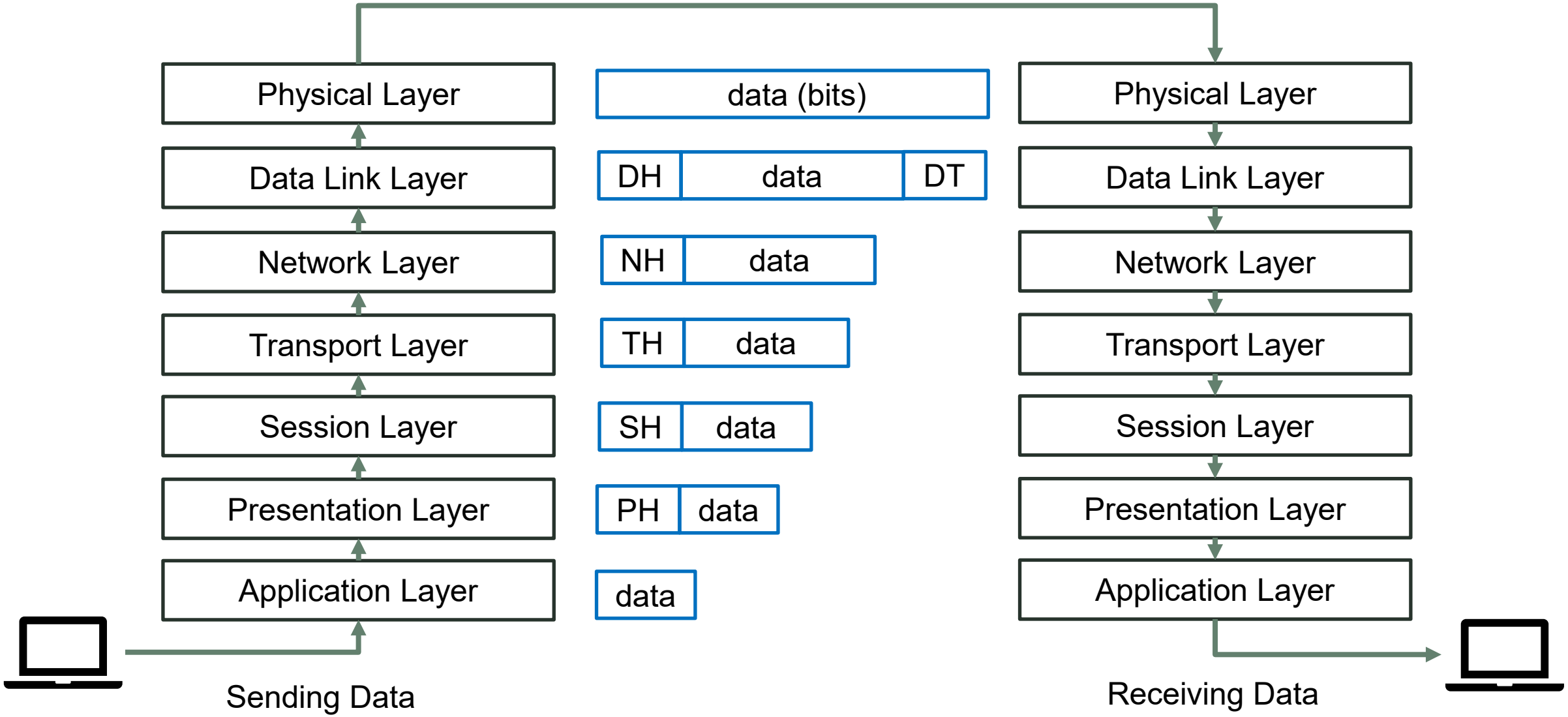
Ethics of Data Scraping

1. The act of scraping vs the application/purpose of collection.
2. Copyrighted Data cannot be published (as own) after scraping.
3. Check the T&Cs of a website.
4. If API is available, access data via API instead of scraping.
5. Check the Robots.txt file en.wikipedia.org/robots.txt.
6. Avoid being greedy and use time outs to avoid too frequent requests.
7. Protect self-identity as needed.

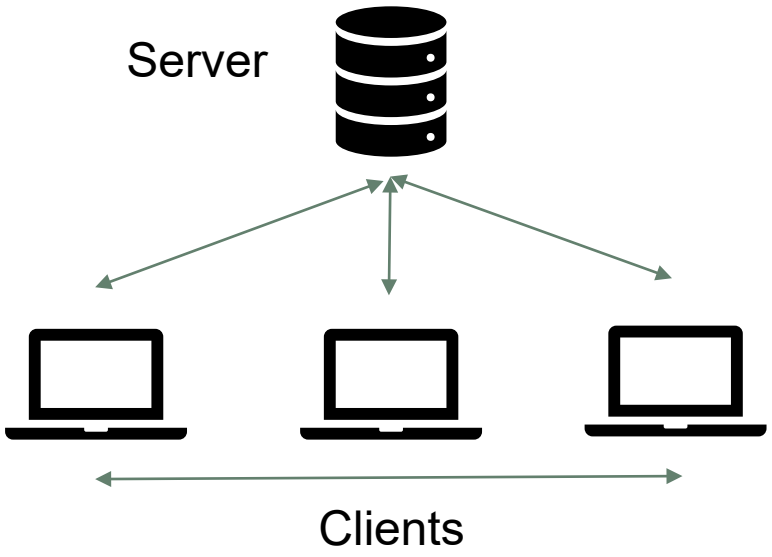
Steps in Web Scraping

1. Send a request to the web page using its URL.
2. On receiving an HTML response, parse the HTML to extract elements.
3. Retrieve relevant information from the elements and preprocess as needed.
4. Save the preprocessed data to a file for future use.

How does the internet work?

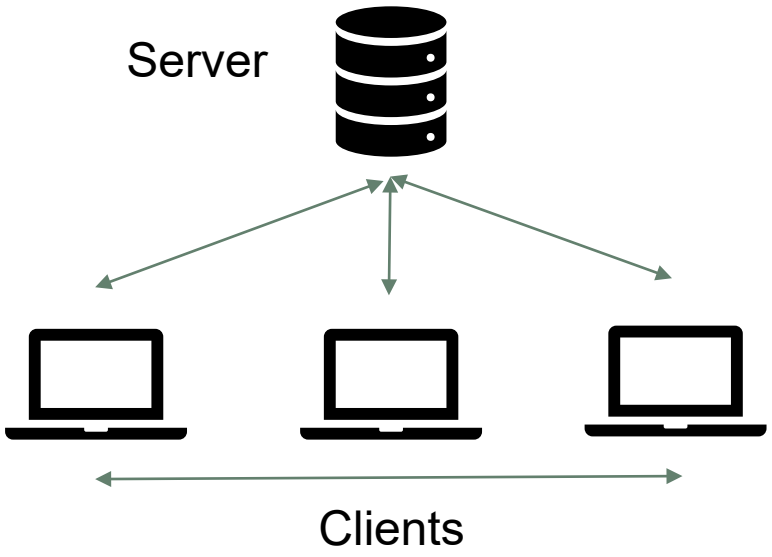


Client Server Model



- A client is a computer (host) which can receive information or request data or services from servers.
  - E.g. A web browser
- A server is a a remote computer which responds to client requests and provides data.
  - E.g. A web server

Client Server Model



- Web browser client requests DNS for the server's IP address by entering a URL.
- DNS server resolves the web server's IP address and responds to the web browser client.
- Browser sends an HTTP/HTTPS request to the server containing details such as destination IP and port number.
- Server responds with necessary files (HTML, CSS, etc.).
- Browser renders the website using DOM, CSS, and JS interpreters.

Server Response Status Codes

1xx	Informational	The server acknowledges the request, and more input is expected by client or server
2xx	Success	The client request was successful
3xx	Redirection	The server received the request, but the requested URL is located elsewhere, and further actions may be needed.
4xx	Client Error	Client-side error (404 Page not Found, 403 Forbidden)
5xx	Server Error	Server-side error (500 Internal Server Error, 503 Service Unavailable)

What is a Website Frontend made of?

HTML
(Defines structure)

CSS
(Defines style)

JavaScript
(Defines behaviour)

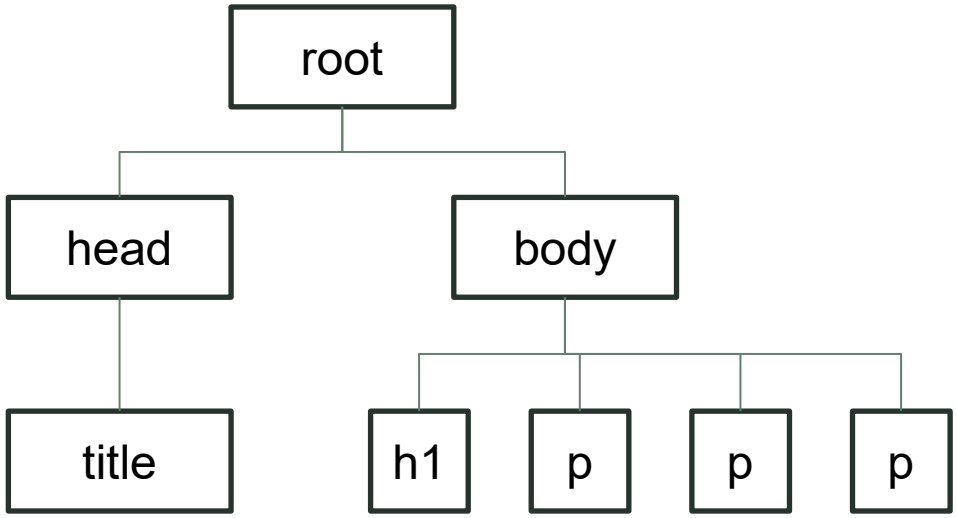
HTML (HyperText Markup Language)

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p class= “myclass”> Here is a very important paragraph </p>
    <p> Here is another very important paragraph </p>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit
W3Schools</a>
  </body>
</html>
```

1. HTML is a markup language that describes the structure of the web Page
2. HTML instructs the Web Browser how to display the content.
3. Contains a series of elements where each element can have
   - Tags (starting and ending the element)
   - Attributes (defining features of the element)
   - Content (e.g., text contained by the element)

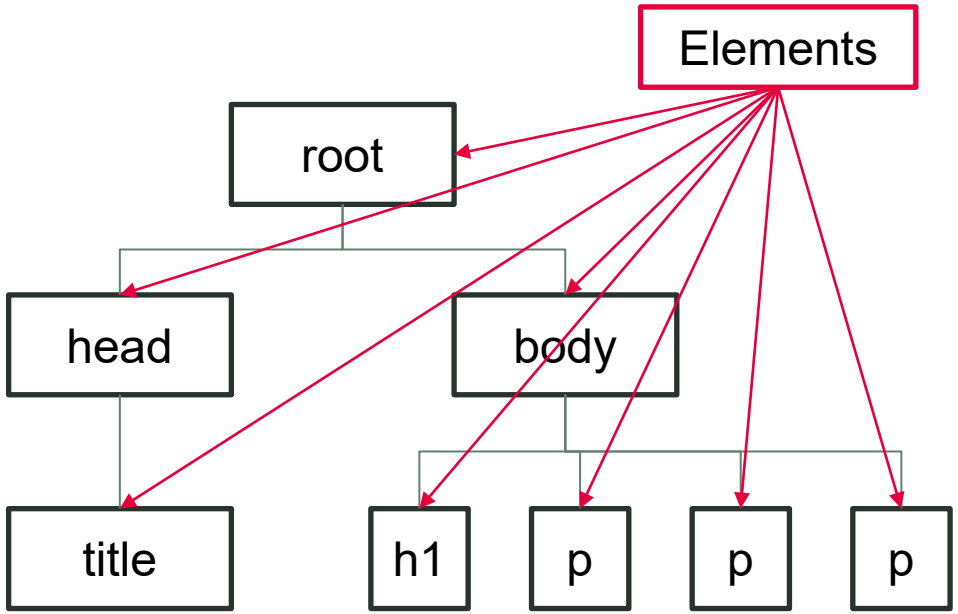
Document Object Model

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p> Here is a very important paragraph </p>
    <p> Here is another very important paragraph </p>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit
W3Schools</a>
  </body>
</html>
```



Document Object Model

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p> Here is a very important paragraph </p>
    <p> Here is another very important paragraph </p>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit
W3Schools</a>
  </body>
</html>
```



CSS (Cascading Style Sheets)

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
    <style>
        body {background-color: wheat;}
        h1   {color:blue;}
        p    {color:red;}
    </style>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p> Here is a very important paragraph </p>
    <p> Here is another very important paragraph </p>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit
W3Schools</a>
  </body>
</html>
```

1. Helps to format the layout of a page.
2. CSS elements may be added inline, internally in the header or as external files (most common).
3. Can be used to style multiple pages.
4. Can be used to control the layout in a variety of screen sizes.

Javascript

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
    <script>
       function changeContent() {
          document.getElementById("demo").innerHTML = "The very important paragraph is modified.";
       }
    </script>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p> Here is a very important paragraph </p>
    <p id="demo"> Here is another very important paragraph </p>
    <button type="button" onclick="changeContent()">Try it</button>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit W3Schools</a>
    <div class="content">Here is some high quality content.</div>
</body>
</html>
```

1. Javascript helps to make pages interactive and dynamic.
2. Can handle complex functions and features.

Extensible Markup Language (XML)

```
<classroom>
  <student>
    <name>John Doe</name>
    <subject><Maths</subject>
  </student>
  <student>
    <name>Mary Jane</name>
    <subject><History</subject>
  </student>
</classroom>
```

1. XML is a markup language which represents data in a tree-like structure of elements.
2. Like XML, elements can have attributes, other elements and content.
3. XML tags are self descriptive.
4. XML is platform independent.

XML Path Language (XPath)

XPath uses path expressions to select nodes or node-sets in an XML document.

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p> Here is a very important paragraph </p>
    <p> Here is another very important paragraph </p>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit
W3Schools</a>
  </body>
</html>
```

/html/body/p[1]

CSS Selectors

CSS Selectors are used by CSS to select elements.

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
  </head>
  <body>
    <h1>Introduction to Web Scraping</h1>
    <p> Here is a very important paragraph </p>
    <p> Here is another very important paragraph </p>
    <p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit
W3Schools</a>
  </body>
</html>
```

body > p:nth-child(2)

1. Selenium Tips: CSS Selectors (saucelabs.com)
2. Practice your CSS selector skills at https://flukeout.github.io

How to find the data to scrape?

View Page Source

view-source:file:///C:/Users/sukan/PycharmProjects/Data%20Scraping%20Workshop%202024/demo_js.html

ine wrap

```
<html>
  <head>
    <title>Web Scraping Workshop</title>
    <script>
        function changeContent() {
            document.getElementById("demo").innerHTML = "The very important paragraph is modified.";
        }
    </script>
  </head>
  <body>
<h1>Introduction to Web Scraping</h1>
<p> Here is a very important paragraph </p>
<p id="demo"> Here is another very important paragraph </p>
<button type="button" onclick="changeContent()">Try it</button>
<p/>Presented by <a href=https://www.dsl.unibe.ch/>Visit W3Schools</a>
<div class="content">
    Here is some high quality content.
</div>
</body>
</html>
```

How to find the data to scrape?

Page Inspector

File | C:/Users/sukan/PycharmProjects/Data%20Scraping%20Workshop%202024/demo_js.html

Introduction to Web Scraping

Here is a very important paragraph

The high quality content is now edited.

Try it

div.content 441.2 × 18.8 ools

Here is some high quality content.

Elements

```
<html>
<head> == $0
  <title>Web Scraping Workshop</title>
  <script> ... </script>
</head>
<body>
  <h1>Introduction to Web Scraping</h1>
  <p> Here is a very important paragraph </p>
  <p id="demo">The high quality content is now edited.</p>
  <button type="button" onclick="changeContent()">Try it</button>
  <p> ... </p>
  <div class="content"> Here is some high quality content. </div>
```

html head

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls

Console Issues Autofill

Include third-party issues Severity: Default levels Browser: Top browsers Group by kind

Accessibility

1 <html> element must have a lang attribute: The <html> element does not have a lang attribute — Warning

Compatibility

1 A 'viewport' meta element was not specified. — Error

Task

1. Open the local page in the project folder of demo_html.html in Chrome browser
2. Open the inspector -> Styles.
3. Select an element and change an aspect of style (e.g. Color).
4. Bonus Challenge: Can you change the background color for www.google.com?

Beautiful Soup

1. A Python library which extracts data out of HTML and XML files
2. It can be used to extract specific elements and their content.
3. It is simple and provides a higher level of abstraction and therefore easy to use even for beginners.
4. However, since it cannot handle Javascript, it is difficult to use for more complex, dynamic interactions.

Selenium-Python

1. Selenium Python allows one to access the Selenium WebDriver for automation tasks.
2. Selenium can work with a variety of web browsers such as Chrome, Firefox etc.
3. With appropriate actions, human like actions can be mimicked.
4. Useful for cases complex cases where popups, login etc are involved.

Scrapy

1. Scrapy is a full suite, high-level web crawling and web scraping framework which can be used to extract structured data.
2. Can execute multiple requests simultaneously.
3. Scrapy is efficient and fast however has a higher learning curve as compared to beautiful soup and Selenium.

Practice Sites for scraping

1. Fake Python (realpython.github.io)
2. freeCodeCamp/scrapepark.org: Source for scrapepark.org (github.com)
3. http://books.toscrape.com