What are Web Scraping and HTML

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

HTML (Hyper Text Markup Language) is the de facto mechanism by which the web pages in the world wide web communicate with each other. It has ubiquitously been used since the beginning days of internet and browsers. Nowadays, though the advances in distributed computing combined with the unquenchable thirst for exchanging data of all sorts and types, the HTML paradigm has evolved accordingly. Web Scraping generally deals with extraction of data from the web, and as such the programs that do the scraping read HTML files.

In this paper, we will describe HTML in more detail and introduce web scraping, its applications and some technical details related to the techniques used for data extraction.

Keywords:

HTML, Web Scraping

What are Web Scraping and HTML

What is HTML

HTML is simply a standard markup language and is used to build web pages. The web pages that we see on the web are composed of text, pictures, videos, areas to enter data and areas that respond to input. We can even talk to web pages. All of this is possible first and foremost by the elements in the HTML document. The elements are themselves represented by tags. By convention, the tags in HTML are defined inside the angle brackets '<' and '>'. For example, the element for a header is represented as shown below:

<h1>Breaking News</h1>

This would appear on the web pages as shown below:

Breaking News

There are tags for title of the page (<title></title>), paragraph(<body></body>). The real power horse behind a web page is HTML, but by itself it can't display pictures or be interactive. The extensions to HTML are Cascading Style Sheets (CSS), and JavaScript. With CSS, we can control how the HTML elements such as header and body are displayed. We can control font size, color and location of elements. For example, the above element can be represented as blow:

<h1 style="color:blue;">Breaking News</h1>

Which would show this on the browser.

Breaking News

CSS also has its syntax, format and standards which must be followed. We can specify CSS inside the HTML tags or put them in a file. Put in file, the style can be shared by other web pages. This is more popular in large scale web pages which require constant updates. The snippet of CSS code shown below, make headers 1 through 6 to look like the specification given for font

size, font, color, margin and padding. This way, developers of HTML pages who use the CSS file in their page will share the same characteristics for the headers. CSS also offers techniques to override them as shown above or place them in the CSS file itself.

```
h1, h2, h3, h4, h5, h6
{ font: Monospace 175% 'century gothic', arial, sans-serif;
  color: #000;
  margin: 0 0 15px 0;
  padding: 15px 0 5px 0;}
```

JavaScript is the programming language for HTML. With JavaScript, we can make a web page interactive.

There are other components that make a web page On the server side, for example scripting languages such as PHP and AJAX perform the back-end tasks.

PHP hypertext preprocessor is a server-side programming language that is embedded into documents such as HTML files, which may contain DHTML, JavaScript, and Java. PHP is great for creating pages on the fly and can be used to make guest books, message boards, and other interactive pages.

Asynchronous JavaScript and XML (AJAX) is a technique that uses the JavaScript-based XMLHttpRequest object to retrieve responses from a web server in a dynamic way, allowing for instant, on page updating.

As far as web authoring tools go, the cheapest, easiest one is notepad. But notepad or any other text editors won't suffice a large-scale deployment. Commercial web authoring tools are widespread, and each is used for certain parts of a web page. FX BLOG (ref. 4) lists Firefox Developer, Photoshop, Panic Coda, Dreamweaver, and Fireworks as the top five web design

tools. There is also Drupal which is a free and open-source web content management framework written in PHP and distributed under the GNU General Public License.

What is Web Scraping

Web scraping is basically extracting data from the web. As we know, volumes of data are exchanged between similarly large number of computers and servers on the world wide web. This Data are increasingly becoming valuable to a wide variety of business applications.

The idea behind web scraping is that programming applications such as python can be developed to connect to a web site (i.e. http://www.amazon.com) and "scrape" what data it is programmed to "scrape". The program would then reformat the data and store them to the specification of the customer. Needless to say, web scraping is a quite a lucrative business that serves businesses from finance, marketing, real-estate and banking to scientific research, news and content monitoring.

The mechanism to connect to a web site and upload data is rather simple. The program first connects to the web site as shown below:

```
url = 'http://web.mta.info/developers/turnstile.html'
response = requests.get(url)
```

We can then use BeatifulSoup (or other HTML parsing library) to parse the HTML returned by the request.

```
soup = BeautifulSoup(response.text, "html.parser")
```

BeatifulSoup has many parsers that can be used to parse the response. From here, we can find tags:

```
soup.findAll('a') # Gets All <a> tags
```

```
one_a_tag = soup.findAll('a')[36] # Gets 36 <a> tags


link = one_a_tag['href'] # Gets the tag <href> which would then refer to another web page which can be scraped.
```

In large scale deployment of data scraping, the application would automatically and dynamically crawl through the web and scrape the data they are designed to collect and store them in the format and the specified location.

Web scraping is an alternative to APIs which involves programming but by sending API calls to the server and getting a reply which carries the information requested. The architecture of API is RESTful, which involves request/response in a standard way. Like web scraping, the data received in the API replies are parsed. Some case they are in JASON, XML, or even HTML. A typical example is getting the weather information form https://api.openweathermap.org.

Conclusions

HTML plays a huge role in web communication. It has given way to technologies such as CSS, JavaScript, Java, PHP, AJAX and so on. It has given way to web scraping. It is so standardized that the same HTML can appear on any web browser-windows explorer, google chrome, Firefox, etc. Those who don't comply or fail to adapt and adopt rapidly evolving technologies are doomed to extinction.

With Web Scraping a whole new and by far wide pasture of data has been made available to the data hungry data scientist.

References

1. https://www.w3schools.com/whatis/whatis_html.asp
2. https://www.w3schools.com/html/html_css.asp
3. https://www.w3schools.com/php/default.asp
4. https://www.webfx.com/blog/web-design/top-five-web-design-tools/
5. Web Scraping using Python - https://www.datacamp.com/community/tutorials/web-scraping-using-python
6. What is web scraping? - https://scrapinghub.com/what-is-web-scraping/
7. https://en.wikipedia.org/wiki/Drupal