Hyper-Text Markup Language (HTML) is a markup language which consists of a bunch of nested tags which describe the content of a web page. HTML is often combined with Cascading Style Sheets (CSS), which describes how the content should be displayed, and JavaScript, which allows interactivity and more advanced functionality.

In order to web scrape, you should have a basic understanding of the common HTML tags, and what the class and id attributes are. Every tag can have a “class” or “id” attribute. Classes and ids are often used to either stylize or programmatically interact with a webpage. You will often be searching by class or id to focus your web scraping.

COMMON TAGS:

* <a>
  + Defines a link.
  + Example: <a href=”about.html”>About Us!</a>
* <body>
  + Defines the document’s body. This contains everything the user sees.
* <div>
  + Defines a section of the document. Often used to organize items.
* <h1> to <h6>
  + Denotes that the text inside is a header.
  + Example: <h1>Cat Fancy</h1><h3>A site for cat lovers!</h3>
* <img>
  + Defines an image.
  + Example: <img src=”pics/crazy\_cat.jpg” />
* <input>
  + Defines items that the user can use to input information.
* <li>
  + Defines an item inside of a list
* <ol> and <ul>
  + Defines ordered (1, 2, 3) and unordered (bullets) lists. Contains <li> tags
  + Example: <h1>My Favorite Things</h1><ul><li>Raindrops on roses</li><li>Whiskers on kittens</li><li>Bright copper kettles</li><li>Warm woolen mittens</li></ul>
* <p>
  + Defines a paragraph. Basically just regular text.
* <span>
  + Defines a section that is “special” in some way.
* <table>, <tr>, <th>, <td>
  + Defines a table, a table row, a header cell, or a data cell
  + Example:

<table>

<tr><th>Name</th><th>Age</th></tr>

<tr><td>John</td><td>21</td></tr>

<tr><td>Chase</td><td>24</td></tr>

</table>