11.3 Web Scraping

May 9, 2019

1 11.3 Web Scraping

HTML, which stands for "hypertext markup language", is an XML-like language for specifying the appearance of web pages. Each tag in HTML corresponds to a specific page element. For example:

- `<img>` specifies an image. The path to the image file is specified in the `src=` attribute.
- `<a>` specifies a hyperlink. The text enclosed between `<a>` and `</a>` is the text of the link that appears, while the URL is specified in the `href=` attribute of the tag.
- `<table>` specifies a table. The rows of the table are specified by `<tr>` tags nested inside the `<table>` tag, while the cells in each row are specified by `<td>` tags nested inside each `<tr>` tag.

Our goal in this section is not to teach you HTML to make a web page. You will learn just enough HTML to be able to scrape data programmatically from a web page.

2 Inspecting HTML Source Code

Suppose we want to scrape faculty information from the Cal Poly Statistics Department directory (`https://statistics.calpoly.edu/content/StatisticsDirectory%26Office%20Hours`). Once we have identified a web page that we want to scrape, the next step is to study the HTML source code. All web browsers have a "View Source" or "Page Source" feature that will display the HTML source of a web page.

Visit the web page above, and view the HTML source of that page. (You may have to search online to figure out how to view the page source in your favorite browser.) Scroll down until you find the HTML code for the table containing information about the name, office, phone, e-mail, and office hours of the faculty members.

Notice how difficult it can be to find a page element in the HTML source. Many browsers allow you to right-click on a page element and jump to the part of the HTML source corresponding to that element.

3 Web Scraping Using BeautifulSoup

`BeautifulSoup` is a Python library that makes it easy to navigate an HTML document. Like with `lxml`, we can query tags by name or attribute, and we can narrow our search to the ancestors and descendants of specific tags. In fact, it is possible to use `lxml` with HTML documents, but