Getting Data From Websites

Using Ruby and Python to Collect Data

What is web scraping?

Collecting data from a website is called “scraping.” When you scrape a website, you load a web page and then parse the page to get the data that you want. You can read more about web scraping at https://en.wikipedia.org/wiki/Web_scraping.

How can you scrape data from a website?

There are many ways to scrape data from a website. Some examples include the following.

Example	**Pros**	**Cons**
Load a web page in a browser and then copy and paste data into a document.	● Nothing to install. ● Almost anyone can do it.	● Slow and tedious and manual.
Use cURL to request a web page and parse the page with regular expressions.	● Not much to install. ● Perfect for simple requests and simple pages. ● Lots of resources available.	● Can get complicated fast. ● Can’t handle JavaScript. ● Breaks when the site’s content changes.
Use a programming language to request a web page and parse the page’s content.	● Can handle more web pages. ● Lots of tools available to handle different situations.	● Programming languages need to be installed. ● Breaks when the site’s content changes.
Use a programming language to access an API and parse the data returned.	● Usually the best option. ● Changes to the API are usually compatible.	● Programming languages need to be installed. ● Can use this only if the site makes it available.

When shouldn’t I scrape data from a website?

You should **not** scrape data from a website if you have been asked not to scrape data from a website. Make certain to read a websites Terms & Conditions to make certain that scraping is permissible.

What is an API?

An API, or an Application Programming Interface, is a set of methods for communicating between software systems. When you use an APIs to collect data, you request data in a structured way and then get it back in a standard format. You can read more about APIs at https://en.wikipedia.org/wiki/Application_programming_interface.

Resources

Tools

- cURL - https://curl.haxx.se/
- Regular Expressions - https://en.wikipedia.org/wiki/Regular_expression
- Chrome Dev Tools - https://developer.chrome.com/devtools
- Chrome Dev Tools Tutorial - https://www.codeschool.com/courses/discover-devtools

Python

- Mechanize - https://pypi.python.org/pypi/mechanize
- Mechanize Cheat Sheet - http://www.pythonforbeginners.com/cheatsheet/python-mechanize-cheat-sheet
- Beautiful Soup - https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Ruby

- Try Ruby - http://tryruby.org
- Typheous - https://github.com/typhoeus/typhoeus
- Mechanize - https://github.com/sparklemotion/mechanize
- Mechanize Tutorial - http://ruby.bastardsbook.com/chapters/mechanize/
- CSV Library - https://www.sitepoint.com/guide-ruby-csv-library-part/

Data

- Baltimore City Data - https://data.baltimorecity.gov/

Meetups

- Code for Baltimore - https://www.meetup.com/Code-for-Baltimore/
- Baltimore Legal Hackers Meetup - https://www.meetup.com/Baltimore-Legal-Hackers-Meetup/

Available online at https://goo.gl/ScgmCC