Web Scraping with Python

Carlos Hurtado

Department of Economics
University of Illinois at Urbana-Champaign
hrtdmrt2@illinois.edu

Dec 5th, 2017

On the Agenda

1. Introduction
2. Installing Modules
3. HTML
4. HTML Tables in Python

On the Agenda

Introduction

- Much of what we do on the computer is really what we do on the Internet.
- It would be great if our programs could get online.
- The importance of extracting data from the web is becoming increasingly loud and clear.
- This lecture will guide you through the process of writing a Python script that can extract information from a web page.

On the Agenda

Installing Modules

- There are several modules that make it easy to scrape web pages in Python.
  - webbrowser: Comes with Python and opens a browser to a specific page
  - requests: Downloads files and web pages from the Internet
  - beautifulsoup: Parses HTML, the format that web pages are written in.
  - lxml: Processing XML and HTML in the Python language.
  - selenium: Launches and controls a web browser. Selenium is able to fill in forms and simulate mouse clicks in this browser.

Installing Modules

- The 'pip package manager' makes it easy to install open-source libraries that expand what you're able to do with Python.
- We will use it to install everything needed to create a working web application.
- pip package is already installed if you are using Python 2>=2.7.9 or Python 3>=3.4
- You can go to the pip web page for instructions on how to install it if you don't have it on your machine.
- In Windows, it's necessary to make sure that the Python Scripts directory is available on your system's PATH so it can be called from anywhere on the command line.
- Verify pip is installed with the following code on the console: pip -V

Installing Modules

open your terminal:

- If you only have one version of Python:

```
pip install request
pip install lxml
```

- If you have two versions of Python (e.g 2.7 and 3.4): To update your 2.X version use

```
pip2 install request
pip2 install lxml
```

- If you don't have pip2 installed, in Linux and iOs you can use

```
sudo apt install python-pip
```

On the Agenda

HTML

- HTML is a computer language devised to allow website creation.
- These websites can then be viewed by anyone else connected to the Internet.
- It is relatively easy to learn, with the basics being accessible to most people
- The definition of HTML is: HyperText Markup Language

HTML

- HyperText is the method by which you move around on the web âĂŤ by clicking on special text called hyperlinks
- Markup is what HTML tags do to the text inside them.
- It is relatively easy to learn, with the basics being accessible to most people
- HTML is a Language, as it has code-words and syntax like any other language

HTML

- How does it work?
- HTML consists of a series of short codes typed into a text-file by the site author - these are the tags.
- The text is then saved as a html file, and viewed through a browser
- This browser reads the file and translates the text into a visible form, hopefully rendering the page as the author had intended.
- Writing your own HTML entails using tags correctly to create your vision.
- You can use anything from a rudimentary text-editor to a powerful graphical editor to create HTML pages.

HTML

- The tags are what separate normal text from HTML code.
- You might know them as: the words between the <angle-brackets>.
- They give structure to the images, tables, text, etc, just by telling your browser what to render on the page.
- Different tags will perform different functions.
- The tags themselves don't appear when you view your page through a browser, but their effects do.

On the Agenda

HTML Tables in Python

- An HTML object consists of a few fundamental pieces: a tag.
- The format that defines a tag is

```
<tag property1="value" property2="value">
```

- It could have attributes which consists of a property and a value.
- A tag we are interested in is the table tag, which defined a table in a website.
- This table tag has many elements.

HTML Tables in Python

- An element is a component of the page which typically contains content.
- For tables in HTML, they consist of rows designated by elements within the tr tag, and then column content inside the td tag
- A typical example is

```
<table>
        <tr>
                <td> Hello! </td>
                <td> Table </td>
        </tr>
</table>
```

- Most sites organize data using tables, so we're going to learn to scrape those. (see the .py files!)