Intro To Scraping

Ciencia de Datos para la toma de decisiones en Economía

Ignacio Sarmiento-Barbieri

Motiviation Webscraping

- Web scrapin is a technique used to automate the process of extracting information from websites, such as tables, texts or links to other pages.
- Why web scrape?
  - Works better than copying and pasting information from the web.
  - It is fast and replicable.

Motiviation Webscraping

Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers†

By Alberto Cavallo*

Online prices are increasingly used for measurement and research applications, yet little is known about their relation to prices collected offline, where most retail transactions take place. I conduct the first large-scale comparison of prices simultaneously collected from the websites and physical stores of 56 large multi-channel retailers in 10 countries. I find that price levels are identical about 72 percent of the time. Price changes are not synchronized but have similar frequencies and average sizes. These results have implications for national statistical offices, researchers using online data, and anyone interested in the effect of the Internet on retail prices. (*JEL* D22, L11, L81, O14)

Motiviation Webscraping

Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health

SCOTT CUNNINGHAM
Baylor University
and
MANISHA SHAH
University of California, Los Angeles & NBER

First version received November 2015; *Editorial decision August* 2017; *Accepted November* 2017 (*Eds.*)

Most governments in the world, including the U.S., prohibit sex work. Given these types of laws rarely change and are fairly uniform across regions, our knowledge about the impact of decriminalizing sex work is largely conjectural. We exploit the fact that a Rhode Island District Court judge unexpectedly decriminalized indoor sex work to provide causal estimates of the impact of decriminalization on the composition of the sex market, reported rape offences, and sexually transmitted infections. While decriminalization increases the size of the indoor sex market, reported rape offences fall by 30% and female gonorrhoea incidence declines by over 40%.

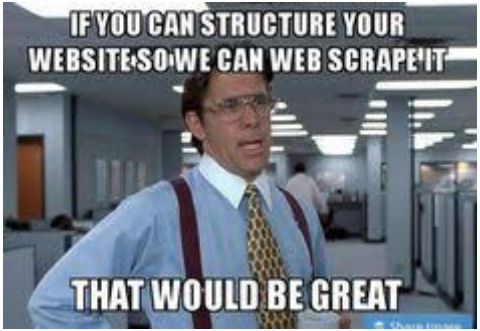
Key words: Regulation, Sex work, Public health, Crime.

JEL Codes: I18, J4, K42

Motiviation Webscraping

We also harvest data from an online review site called The Erotic Review. TER, a reputation website similar to Yelp.com, is one of the largest sex websites in the country and only covers indoor sex workers. Customers use it primarily to provide feedback on transactions with sex workers in a particular area. We collect approximately 90,000 records from TER database from 1999 to 2007 from all over the country. We identify Rhode Island-based sex workers by using phone number area codes. We primarily use the data to focus on the types of services provided, transaction prices, and provider race.

Webscraping basics



Webscraping basics

- How to get data, or "content", off the web and onto our computers.
- If you see it in your browser it exists somewhere
- To be "successful" one must have a working knowledge on:
  - how web pages display content (Hyper Text Markup Language or HTML)
  - where is the content "located"
    1. Server side
    2. Client side
  - The good news is that both server-side and client-side websites allow for web scraping

Caveat: ethical and legal limitations

- Just because you *can* scrape it, doesn't mean you *should*.
- Always check the terms and conditions and what they say about scraping. In Colombia, web scraping is legal **unless explicitly prohibited**
- Check `The Robots Exclusion Protocol` of a website, adding ``/robots.txt'' to the website's URL (see `http://www.robotstxt.org`)
- Remember the immortal words of uncle Ben: "with great power comes great responsibility"

About `robots.txt`

- The robots exclusion standard, also known as robots exclusion protocol or simply texttt“/robots.txt”,
- It is a standard protocol used by some websites to communicate with search engines and other robots on the web.
- This protocol tells search engines or web robots about the parts of that website that should/cannot be processed or scanned.

About `robots.txt`

- Example 1: Allow any robots to process/scan all elements on the page

```
# robots.txt for https://example.com/
User-agent: *
Disallow:
```

- Example 2: does not allow to render/scan any element of the page

```
# robots.txt for https://example.com/
User-agent: *
Disallow: /
```

About `robots.txt`

- Example 3: does not allow to process/scan any element of the /public/index.html file

```
# robots.txt for https://example.com/
User-agent: *
Disallow: /public/index.html
```

- Example 4: Do not allow the BadBot bot to process/scan any element on the page:

```
# robots.txt for https://example.com/
User-agent: BadBot
Disallow: /
```

- Let's check wikipedia's `https://en.wikipedia.org/robots.txt`

About Hyper Text Markup Language or HTML

- HTML is not a programming language, but rather a hypertext markup language.
- An HTML is written entirely with elements, which in turn are made up of tags, content and attributes.
- The elements are structured like a tree (trunk, branches, leaves).
- Therefore, to be able to extract an element (for example a leaf), the route of the node or label must be traced (indicating the trunk and the branch that contains the leaf).
- HTML is interpreted by web browsers displaying its content.

Elements in an HTML

- An element is composed of a
  - tag
  - attribute(s) (not always)
  - content

```
<p id="text"> Hello world </p>
```

Elements in an HTML

- Tags are used to delimit the start (<>) and end (< >) of an element.
- Here are some common element tags in HTML:
  - `<p>`: Paragraphs
  - `<head>`: Header of the page
  - `<body`: Body of the page
  - `<h1>, <h2>,...,<hi>`: Headings, Sections
  - `<a>`: links
  - `<li>`: Item in a list
  - `<table>`: Tables
  - `<td>`: A data cell in a table
  - `<div>`: Division. It is used to create sections or group content.
  - `<script>`: Used to insert or refer to a script

Elements in an HTML

- Attributes are used to configure or provide additional information to an element.
- They are always expressed in the start tag and are assigned a name and a value.

```
<a class="document-toc-link" col="red">Attribute List</a>
```

- Here the element label is `a` and it has two `class` attributes that indicate that the content is a link to another website and `col` that indicates that it should be displayed in red.

Server-side

- The website is ”static”, all the info is located in the HTML code that the host server sends
  - E.g. Wikipedia tables are already populated with all of the information - tables, numbers, dates, etc. - that we see in our browser.
- Challenges:
  - Finding the correct path CSS (or Xpath) ”selectors”.
  - Navigating dynamic webpages (e.g. ”Next page” and ”Show More” tabs).

Some useful tools

- CSS selectors:
  - SelectorGadget for Chrome
  - ScrapeMate for Firefox
  - Inspect Element
- Browsers: anything but explorer

Client-side

- The website contains an empty template of HTML and CSS.
  - E.g. It might contain a ”skeleton” table without any values.
- However, when we actually visit the page URL, our browser sends a *request* to the host server.
- If request is valid, then the server sends a response script, which our browser executes and uses to populate the HTML template with the specific information that we want.
- Challenges: Finding the ”API endpoints” can be tricky, since these are sometimes hidden from view.

Further Readings

- `Webscraping` tutorial from Prof. Grant McDermott.
- Web Scrapping slides from Fernandez Villaverde J., Guerrón P. & Zarruk Valencia, D.
- Wickham, H., & Wickham, M. H. (2016). Package 'rvest'. `https://cran.r-project.org/web/packages/rvest/rvest.pdf`.