COM5507 Social Media Data Acquisition and Processing
Week 5. Web Technologies & Web Scraping – Episode 1

Lectured by: Dr. Xinzhi ZHANG
Research Assistant Professor, Department of Journalism
Hong Kong Baptist University
3 Oct 2018 @ CityU M4003

Agenda

- Web scraping: an introduction
  - What & why
- The web technologies
  - The networking infrastructure & HTTP
  - HTML & CSS
- Prelude
- Web scraping: the first instances

Web scraping: an introduction

- Web scraping (also called “screen scraping,” “web harvesting,” “web data extraction,” or even “web data mining”), can be defined as “the construction of an agent to download, parse, and organize data from the web in an automated manner” (Broucke, 2017).

Why web scraping?

- The web is a rich data source of human's digital traces and social and cultural artifacts in many research domains;
- Web scraping is faster and more accurate than collecting data manually;
- It offers a solution for data collection when API (application program interface) is unavailable or is not free;
- It is a feasible empirical instance to learn a programming language.

Web scraping: an introduction

- Practitioners of web scraping
  - Academic researchers, esp. in digital humanities and social sciences
  - Business and finance analysts
  - Searching engine and web product developers
  - HR and employers
  - Digital marketing
  - Data-driven journalism professionals

The web technologies

- International Organization of Standardization (ISO) maintains the *Open Systems Interconnection (OSI) model*, in order to standardize the communication processes on the web.
- There are 7 layers

https://medium.com/@madhavbahl10/osi-model-layers-explained-ee1d43058c1f

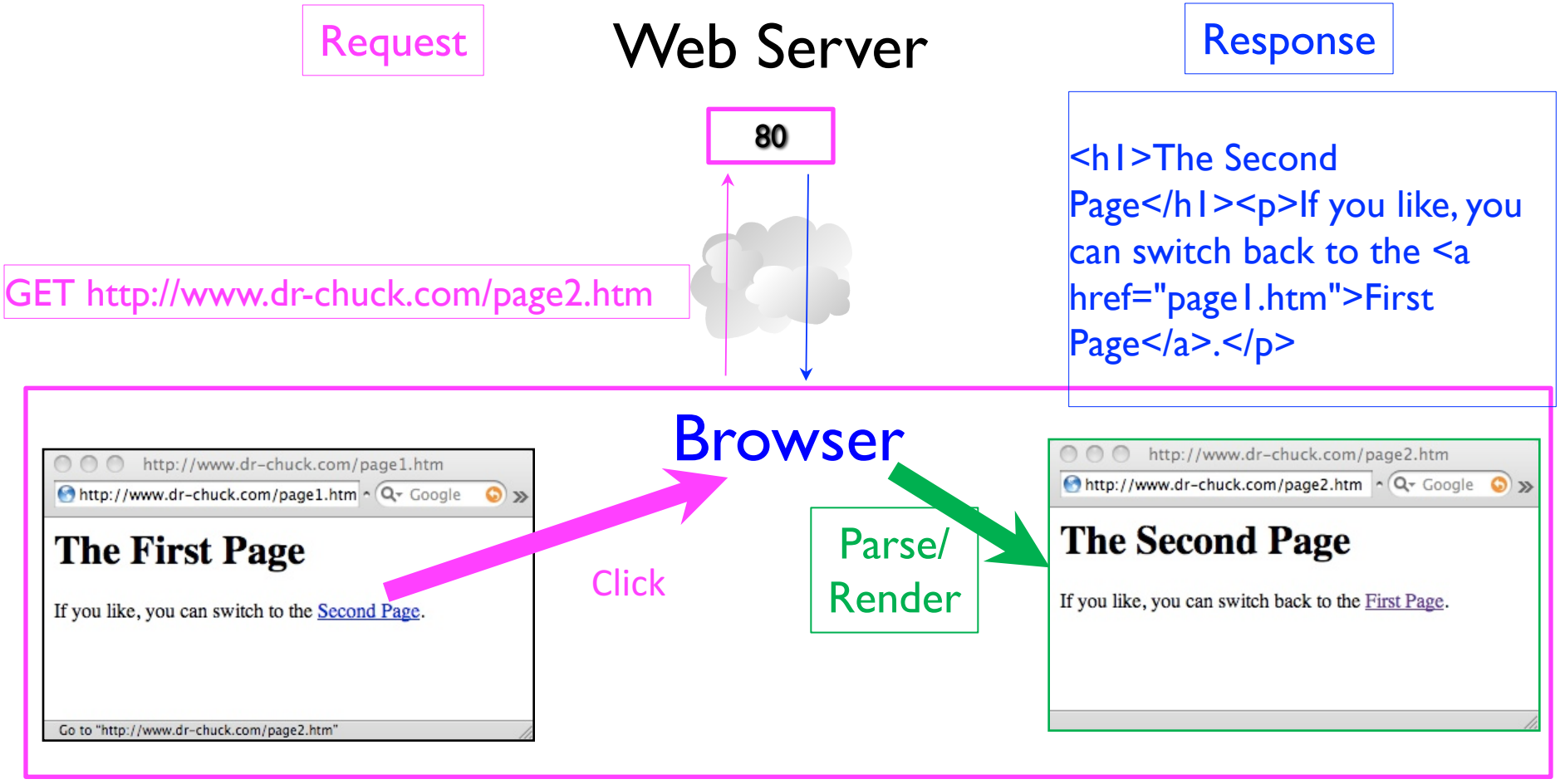
An introduction [URL]

HTTP - Hypertext Transfer Protocol

- The HyperText Transfer Protocol is the set of rules to allow browsers to retrieve web documents from servers over the internet.
- It is the dominant Application Layer Protocol on the internet.
- Basic concepts: Make a Connection - Request a document - Retrieve the Document - Close the Connection
- “The web speaks HTTP.”

“Getting data” from the server

- Each time the user clicks on an anchor tag with an href= value to switch to a new page, the browser makes a connection to the web server and issues a “GET” request - to GET the content of the page at the specified URL
- The server returns the HTML document to the browser, which formats and displays the document to the user



Picture source: Charles R. Severance (2010)

HTTP in Python

- The “Requests” library
- To enable Python speaks and understands HTTP in order to “browse” the web
- URL (Uniform Resource Locator), a web address, is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it.

http://www.cityu.edu.hk/com/

HTML

- HTML (Hypertext Markup Language) is the standard markup language for creating web pages and web applications.
- It is used to “construct” and “structure” the webpages.
  - [A beginner’s tutorial][Examples]

HTML

The simplest HTML page looks like this:

```
<html>
    <head>
        <title>Page Title</title>
    </head>
    <body>
        <h1>Page Title</h1>
        <p>This is a really interesting paragraph.</p>
    </body>
</html>
```

source: http://alignedleft.com/tutorials/d3/fundamentals/

HTML

- E.g.

- Tags
  - Opening tag and closing tag
  - ALL HTML tags should be closed
- Attributes
  - Attributes are extra bits of information
  - Attributes appear inside the opening tag and their value is always inside quotation marks

HTML tags

- ... to enclose a paragraph;
-
 to set a line break;
- <h1>...</h1> to <h6>...</h6> for headers;
- ... to indicate a “division” in an HTML document, basically used to group a set of elements;
- <a>... for hyperlinks;
- ..., <ol>... for unordered and ordered lists respectively; inside of these, <li>... is used for each list item.

HTML tags

- Tables
  - The <table> element defines the table.
  - The <tr> element defines a table row
  - The <th> element defines a table head
  - The <td> element defines a data cell. They must be enclosed in <t>r tags.

```
<table border="1">
 <tr>
    <th>Month</th>
    <th>Savings</th>
 </tr>
 <tr>
    <td>January</td>
    <td>$100</td>
 </tr>
</table>
```

HTML tags

- Images
-
- img tag does not have a closing tag, it closes itself, ending with "/>"

- Types of image that are supported on HTML
  - ***. jpg
  - ***. gif
  - ***. png

CSS

- Cascading Style Sheets are used to style the visual presentation of HTML pages.
  - Styles define how to display HTML elements
  - Styles were added to HTML 4.0 to solve a problem

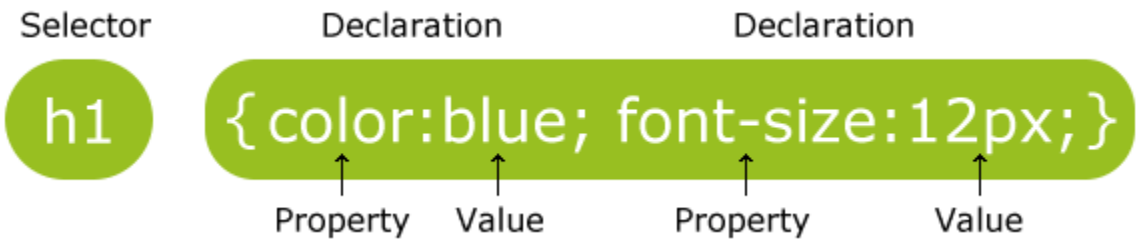
- A simplest CSS reads like this (selectors and rules):

```
body {
    background-color: white;
    color: black;
}
```


source: http://alignedleft.com/tutorials/d3/fundamentals/

CSS

- Selectors and rules (also called “declarations”)
  - The selector is normally the HTML element you want to style.
  - Each declaration consists of a property and a value.
  - The property is the style attribute you want to change.
  - Each property has a value.



Source: http://www.w3schools.com/js/default.asp

CSS

- Class
  - to define a group of elements
  - defined with .style
  - retrieved by: class="style":
  - can appear more than once in the same page
- id
  - to label individual, and unique element in the page
  - defined with #style
  - retrieved by: id="style"
  - can only appear once in the same page

CSS

- using id and class in the web page

class :

```
<style type="text/css">
  .footer{background:red;}
</style>
<div class="footer">footer</div>
```

id :

```
<style type="text/css">
  #footer{background:red;}
</style>
<div id="footer">footer</div>
```

- Priority in the browser:
  - style in the tag (attribute) > id > class > tag

Prelude about web scraping

- The “openness” sprit of the web
- The scientific progress
  - the on-going debate of data collection
- General purposes
  - Educational vs. academic vs. commercial

Prelude about web scraping

- Always be aware of laws and policies
  - The legal issues in the particular countries and societies
  - Terms and conditions
  - Copyright protection vs. public domain
  - “Computer Fraud and Abuse Act” (CFAA)
  - Robots.txt – the protocol set by the website for the “crawlers”

Prelude about scraping

- Revisiting the purpose
  - Governmental agencies
  - Academic institutions
  - Open data
  - Journalistic values
- Get written permission, if possible (apply for an API)
- Check the term of use
- Focusing on public information only
- Be respectful
  - “You are the guest in other people’s home”
  - Always set time intervals to avoid excessive amounts of requests
- Be open and transparent

Web scraping: the first instances

- The codes in Jupyter Notebooks

- Reading examples:
  - Basic ideas
  - Scraping multiple pages
  - Xpath
  - (preview) scraping Twitter without API (from Medium)

More preludes…

- 授人以鱼，考人以鱽鱾鲀鱿鲃鲂鲉鲌鲄鲆鲅鲇鲏鲊鲋鲐鲈鲍鲎鲝鲘鲙鲗鲓鲖鲞鲛鲒鲚鲜鲟鲔鲕鲑鲧鲬鲪鲫鲩鲣鲨鲡鲢鲤鲠鲥鲦鲺鲯鲹鲴鲶鲳鲮鲭鲵鲲鲰鲱鲻鲷鲸鳋鳊鳁鳀鲾鲼鳈鳉鳃鳄鲿鳇鳂鳆鳅鲽鳌鳒鳎鳏鳑鳐鳍鳘鳛鳕鳓鳙鳗鳚鳔鳖鳜鳟鳞鳝鳡鳠鳢鳣鳤

More preludes…

- “The web is messy.”
- There will never be an “one-size-fits-all” solution to all the web scraping.
- Solutions
  - Test & rework
  - Search online tutorials and try to replicate their codes
  - Compare the different outputs
  - Consult a friend