



香港浸會大學
HONG KONG BAPTIST UNIVERSITY

Web Technologies & Web Scraping

JOUR7280/COMM7780

Big Data Analytics for Media and Communication

Instructor: Dr. Xiaoyi Fu

Outline

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

What Is Web Scraping?

- Web scraping is the process of gathering information from the Internet.
 - Even copy-pasting the lyrics of your favorite song
- The construction of an agent to download, parse, and organize data from the web in an **automated** manner (Broucke, 2017).
- If you're scraping a page respectfully for **educational** purposes, then you're unlikely to have any problems.

Why Scrape the Web?

- How things work **without** web scraping


< HOME ONEZERO ELEMENTAL GEN ZORA FORGE HUMAN PARTS MARKER LEVEL HEATED DESIGN MO >

BASED ON YOUR READING HISTORY

Multicollinearity—How does it create a problem?

Understand in-depth how multicollinearity can affect the model's performance

Gagandeep Singh in Towards Data Science
Aug 9, 2019 · 3 min read

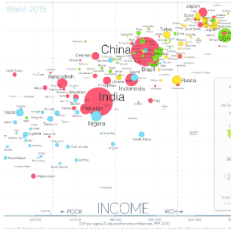


BASED ON YOUR READING HISTORY

Data Science with Python: Intro to Data Visualization with Matplotlib

Intro

Ventsislav Yordanov in Towards Data Science
Jul 22, 2018 · 10 min read

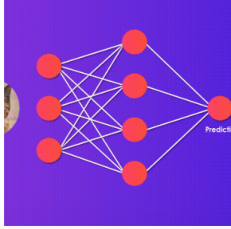


DATA SCIENCE *Suggested topic*

How to build your own Neural Network from scratch in Python

A beginner's guide to understanding the inner workings of Deep Learning

James Loy in Towards Data Science
May 14, 2018 · 7 min read



Popular on Medium

01

There's an Epidemic That's a Bigger Threat Than the Coronavirus

Dr. David L. Katz in Heated
Feb 29 · 8 min read ★

02

There's No Such Thing as 'Quality' Time

Ryan Holiday in Forge
Dec 16, 2019 · 5 min read ★

03

Three Magical Phrases to Comfort a Dying Person

Jenny Harrington in Human Parts
Mar 10, 2019 · 10 min read ★

04

Lessons I've Learned From Studying Money for 10 Years

Tim Denning in The Startup
Feb 26 · 6 min read ★

Why Scrape the Web?

- Manual web scraping can take a lot of time and repetition
 - when you try to get the information you want manually, you might spend a lot of time clicking, scrolling, and searching.
 - This is especially true if you need large amounts of data from websites that are regularly updated with new content.
- 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.

Challenges of Web Scraping

- Variety

- Every website is different.
- While you'll encounter general structures that tend to repeat themselves, each website is unique and will need its own personal treatment if you want to extract the information that's relevant to you.

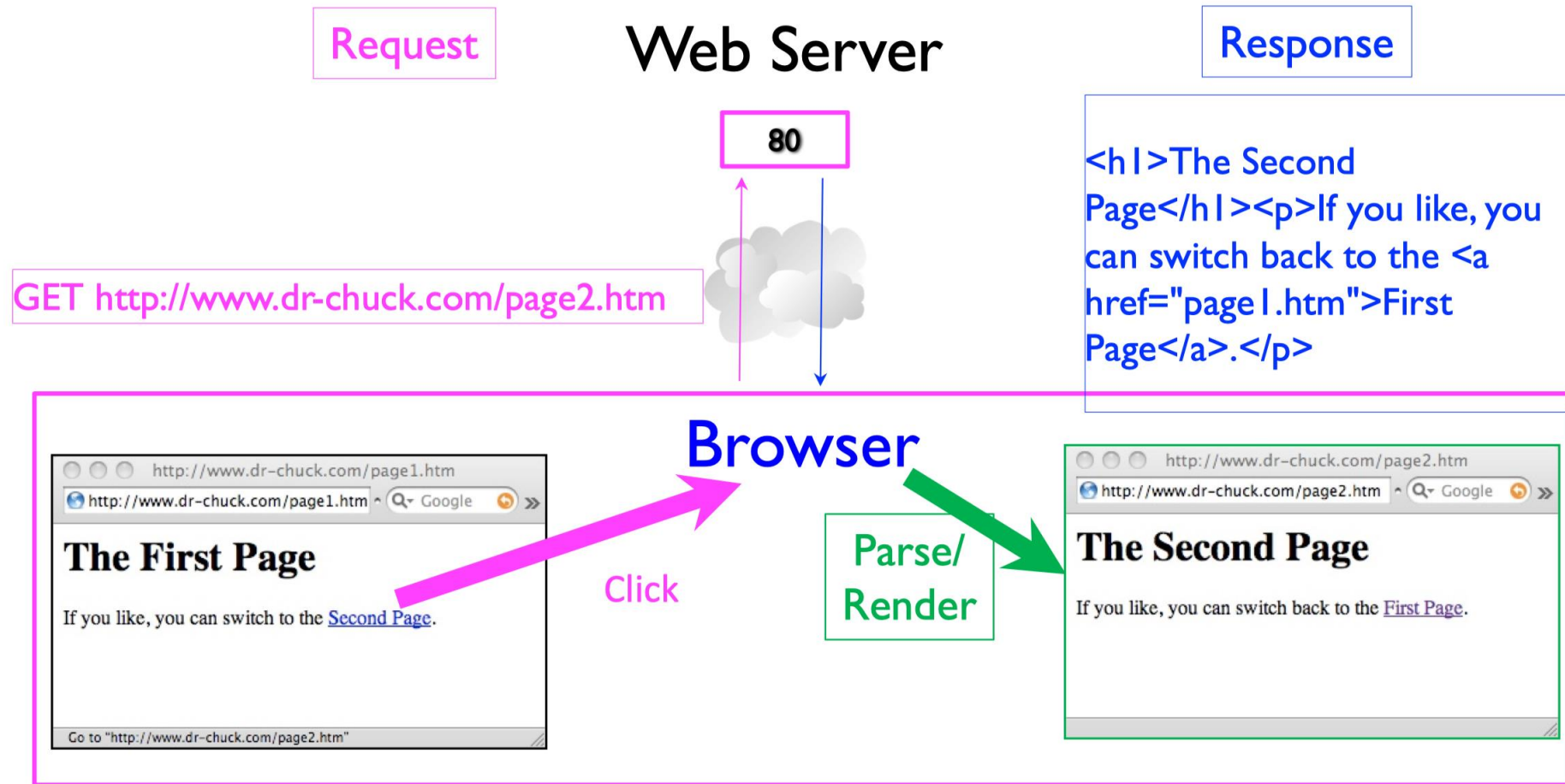
- Durability

- Many websites are in active development
- Once the site's structure has changed, your scraper might not work
- The good news is that many changes to websites are small and incremental, so you'll likely be able to update your scraper with only minimal adjustments.

The components of a web page

- When we visit a web page, our web browser makes a **GET** request to a web server.
 - Since we're getting files from the server. The server then sends back files that tell our browser how to render the page for us.
- The files fall into a few main types:
 - **HTML** — contain the main content of the page.
 - **CSS** — add styling to make the page look nicer.
 - **JS** — Javascript files add interactivity to web pages.
 - **Images** — image formats, such as JPG and PNG allow web pages to show pictures.
- After browser receives all the files, it renders the page and displays it to us.

The components of a web page



The Request Library

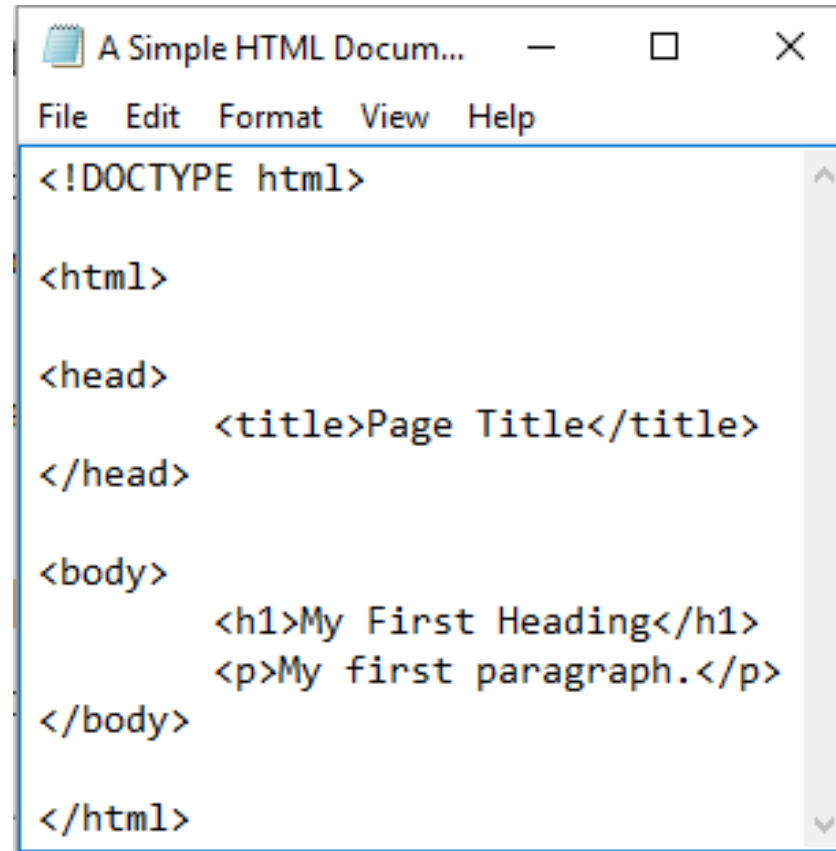
- To enable Python speaks and understands HTTP in order to “browse” the web
- URL (Uniform Resource Locator), a web address, is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it.
- The requests library will make a **GET** request to a web server, which will download the HTML contents of a given web page for us.

HTML

- HyperText Markup Language (HTML) is a language that web pages are created in
- It is used to “construct” and “structure” the webpages.
- Not a programming language
 - A markup language that tells a browser how to layout content

HTML

- The simplest HTML page looks like this:



```
<!DOCTYPE html>

<html>

<head>
    <title>Page Title</title>
</head>

<body>
    <h1>My First Heading</h1>
    <p>My first paragraph.</p>
</body>

</html>
```

HTML

- 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 Tag

- `<html> ... </html>`
 - tells the web browser that everything inside of it is HTML
- `<body>... </body>`
 - The main content of the web page goes into the body tag.
- `<head>... </head>`
 - contains data about the title of the page, and other information that generally isn't useful in web scraping
- `<p>...</p>`
 - defines a paragraph, and any text inside the tag is shown as a separate paragraph

HTML Tag

- `
` to set a line break
- `<h1>...</h1>` to `<h6>...</h6>` for headers
- `<a>...` for hyperlinks
 - tell the browser to render a link to another web page
- `<div>...</div>` to indicate a “division” in an HTML document, basically used to group a set of elements
- `...`, `...` for unordered and ordered lists respectively; inside of these, `...` is used for each list item

HTML Tags

- `class` and `id` attributes
 - These special attributes give HTML elements names, and make them easier to interact with when we're scraping.
 - One element can have multiple classes, and a class can be shared between elements.
 - Each element can only have one id, and an id can only be used once on a page. Classes and ids are `optional`, and not all elements will have them.

HTML Tags

- Using id and class in the web page

In the below case, the content has **class="entry-content"**

```
▼ <div id="primary">
  ▼ <div id="content">
    ▶ <nav id="nav-single">...</nav>
    ▼ <article id="post-188" class="post-188 post type-post status-
      ▶ <header class="entry-header">...</header>
      ... ▶ <div class="entry-content">...</div> == $0
          ▶ <footer class="entry-meta">...</footer>
        </article>
      </div>
    </div>
  </div>
  ▶ <footer id="colophon" role="contentinfo">...</footer>
</div>
```


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`
- For a full list of tags, look [here](#).

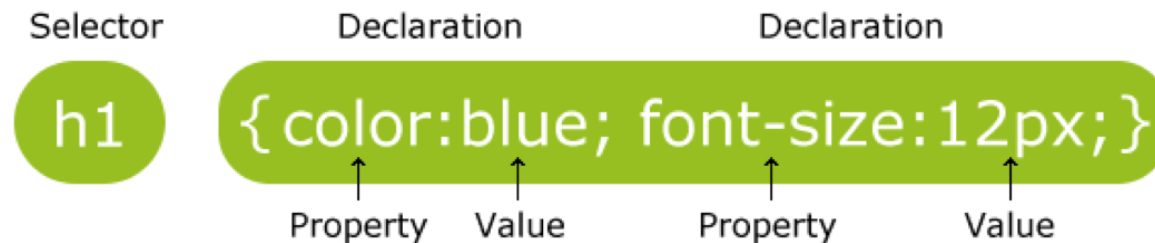
CSS

- **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;  
}
```

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.



Prelude about web scraping

- The “openness” spirit 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)

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

More preludes...

- 授人以鱼，考人以𩚑𩚒𩚓𩚔𩚕𩚖𩚗𩚘𩚙𩚚𩚛𩚜𩚝𩚞𩚟𩚠𩚡𩚢𩚣𩚤𩚥𩚦𩚧𩚨𩚩𩚪𩚫𩚬𩚭𩚮𩚯𩚰𩚱𩚲𩚳𩚴𩚵𩚶𩚷𩚸𩚹𩚺𩚻𩚼𩚽𩚾𩚿𩛀𩛁𩛂𩛃𩛄𩛅𩛆𩛇𩛈𩛉𩛊𩛋𩛌𩛍𩛎𩛏𩛐𩛑𩛒𩛓𩛔𩛕𩛖𩛗𩛘𩛙𩛚𩛛𩛜𩛝𩛞𩛟𩛠𩛡𩛢𩛣𩛤𩛥𩛦𩛧𩛨𩛩𩛪𩛫𩛬𩛭𩛮𩛯𩛰𩛱𩛲𩛳𩛴𩛵𩛶𩛷𩛸𩛹𩛺𩛻𩛼𩛽𩛾𩛿𩜀𩜁𩜂𩜃𩜄𩜅𩜆𩜇𩜈𩜉𩜊𩜋𩜌𩜍𩜎𩜏𩜐𩜑𩜒𩜓𩜔𩜕𩜖𩜗𩜘𩜙𩜚𩜛𩜜𩜝𩜞𩜟𩜠𩜡𩜢𩜣𩜤𩜥𩜦𩜧𩜨𩜩𩜪𩜫𩜬𩜭𩜮𩜯𩜰𩜱𩜲𩜳𩜴𩜵𩜶𩜷𩜸𩜹𩜺𩜻𩜼𩜽𩜾𩜿𩝀𩝁𩝂𩝃𩝄𩝅𩝆𩝇𩝈𩝉𩝊𩝋𩝌𩝍𩝎𩝏𩝐𩝑𩝒𩝓𩝔𩝕𩝖𩝗𩝘𩝙𩝚𩝛𩝜𩝝𩝞𩝟𩝠𩝡𩝢𩝣𩝤𩝥𩝦𩝧𩝨𩝩𩝪𩝫𩝬𩝭𩝮𩝯𩝰𩝱𩝲𩝳𩝴𩝵𩝶𩝷𩝸𩝹𩝺𩝻𩝼𩝽𩝾𩝿𩞀𩞁𩞂𩞃𩞄𩞅𩞆𩞇𩞈𩞉𩞊𩞋𩞌𩞍𩞎𩞏𩞐𩞑𩞒𩞓𩞔𩞕𩞖𩞗𩞘𩞙𩞚𩞛𩞜𩞝𩞞𩞟𩞠𩞡𩞢𩞣𩞤𩞥𩞦𩞧𩞨𩞩𩞪𩞫𩞬𩞭𩞮𩞯𩞰𩞱𩞲𩞳𩞴𩞵𩞶𩞷𩞸𩞹𩞺𩞻𩞼𩞽𩞾𩞿𩟀𩟁𩟂𩟃𩟄𩟅𩟆𩟇𩟈𩟉𩟊𩟋𩟌𩟍𩟎𩟏𩟐𩟑𩟒𩟓𩟔𩟕𩟖𩟗𩟘𩟙𩟚𩟛𩟜𩟝𩟞𩟟𩟠𩟡𩟢𩟣𩟤𩟥𩟦𩟧𩟨𩟩𩟪𩟫𩟬𩟭𩟮𩟯𩟰𩟱𩟲𩟳𩟴𩟵𩟶𩟷𩟸𩟹𩟺𩟻𩟼𩟽𩟾𩟿𩠀𩠁𩠂𩠃𩠄𩠅𩠆𩠇𩠈𩠉𩠊𩠋𩠌𩠍𩠎𩠏𩠐𩠑𩠒𩠓𩠔𩠕𩠖𩠗𩠘𩠙𩠚𩠛𩠜𩠝𩠞𩠟𩠠𩠡𩠢𩠣𩠤𩠥𩠦𩠧𩠨𩠩𩠪𩠫𩠬𩠭𩠮𩠯𩠰𩠱𩠲𩠳𩠴𩠵𩠶𩠷𩠸𩠹𩠺𩠻𩠼𩠽𩠾𩠿𩡀𩡁𩡂𩡃𩡄𩡅𩡆𩡇𩡈𩡉𩡊𩡋𩡌𩡍𩡎𩡏𩡐𩡑𩡒𩡓𩡔𩡕𩡖𩡗𩡘𩡙𩡚𩡛𩡜𩡝𩡞𩡟𩡠𩡡𩡢𩡣𩡤𩡥𩡦𩡧𩡨𩡩𩡪𩡫𩡬𩡭𩡮𩡯𩡰𩡱𩡲𩡳𩡴𩡵𩡶𩡷𩡸𩡹𩡺𩡻𩡼𩡽𩡾𩡿𩢀𩢁𩢂𩢃𩢄𩢅𩢆𩢇𩢈𩢉𩢊𩢋𩢌𩢍𩢎𩢏𩢐𩢑𩢒𩢓𩢔𩢕𩢖𩢗𩢘𩢙𩢚𩢛𩢜𩢝𩢞𩢟𩢠𩢡𩢢𩢣𩢤𩢥𩢦𩢧𩢨𩢩𩢪𩢫𩢬𩢭𩢮𩢯𩢰𩢱𩢲𩢳𩢴𩢵𩢶𩢷𩢸𩢹𩢺𩢻𩢼𩢽𩢾𩢿𩣀𩣁𩣂𩣃𩣄𩣅𩣆𩣇𩣈𩣉𩣊𩣋𩣌𩣍𩣎𩣏𩣐𩣑𩣒𩣓𩣔𩣕𩣖𩣗𩣘𩣙𩣚𩣛𩣜𩣝𩣞𩣟𩣠𩣡𩣢𩣣𩣤𩣥𩣦𩣧𩣨𩣩𩣪𩣫𩣬𩣭𩣮𩣯𩣰𩣱𩣲𩣳𩣴𩣵𩣶𩣷𩣸𩣹𩣺𩣻𩣼𩣽𩣾𩣿𩤀𩤁𩤂𩤃𩤄𩤅𩤆𩤇𩤈𩤉𩤊𩤋𩤌𩤍𩤎𩤏𩤐𩤑𩤒𩤓𩤔𩤕𩤖𩤗𩤘𩤙𩤚𩤛𩤜𩤝𩤞𩤟𩤠𩤡𩤢𩤣𩤤𩤥𩤦𩤧𩤨𩤩𩤪𩤫𩤬𩤭𩤮𩤯𩤰𩤱𩤲𩤳𩤴𩤵𩤶𩤷𩤸𩤹𩤺𩤻𩤼𩤽𩤾𩤿𩥀𩥁𩥂𩥃𩥄𩥅𩥆𩥇𩥈𩥉𩥊𩥋𩥌𩥍𩥎𩥏𩥐𩥑𩥒𩥓𩥔𩥕𩥖𩥗𩥘𩥙𩥚𩥛𩥜𩥝𩥞𩥟𩥠𩥡𩥢𩥣𩥤𩥥𩥦𩥧𩥨𩥩𩥪𩥫𩥬𩥭𩥮𩥯𩥰𩥱𩥲𩥳𩥴𩥵𩥶𩥷𩥸𩥹𩥺𩥻𩥼𩥽𩥾𩥿𩦀𩦁𩦂𩦃𩦄𩦅𩦆𩦇𩦈𩦉𩦊𩦋𩦌𩦍𩦎𩦏𩦐𩦑𩦒𩦓𩦔𩦕𩦖𩦗𩦘𩦙𩦚𩦛𩦜𩦝𩦞𩦟𩦠𩦡𩦢𩦣𩦤𩦥𩦦𩦧𩦨𩦩𩦪𩦫𩦬𩦭𩦮𩦯𩦰𩦱𩦲𩦳𩦴𩦵𩦶𩦷𩦸𩦹𩦺𩦻𩦼𩦽𩦾𩦿𩧀𩧁𩧂𩧃𩧄𩧅𩧆𩧇𩧈𩧉𩧊𩧋𩧌𩧍𩧎𩧏𩧐𩧑𩧒𩧓𩧔𩧕𩧖𩧗𩧘𩧙𩧚𩧛𩧜𩧝𩧞𩧟𩧠𩧡𩧢𩧣𩧤𩧥𩧦𩧧𩧨𩧩𩧪𩧫𩧬𩧭𩧮𩧯𩧰𩧱𩧲𩧳𩧴𩧵𩧶𩧷𩧸𩧹𩧺𩧻𩧼𩧽𩧾𩧿𩨀𩨁𩨂𩨃𩨄𩨅𩨆𩨇𩨈𩨉𩨊𩨋𩨌𩨍𩨎𩨏𩨐𩨑𩨒𩨓𩨔𩨕𩨖𩨗𩨘𩨙𩨚𩨛𩨜𩨝𩨞𩨟𩨠𩨡𩨢𩨣𩨤𩨥𩨦𩨧𩨨𩨩𩨪𩨫𩨬𩨭𩨮𩨯𩨰𩨱𩨲𩨳𩨴𩨵𩨶𩨷𩨸𩨹𩨺𩨻𩨼𩨽𩨾𩨿𩩀𩩁𩩂𩩃𩩄𩩅𩩆𩩇𩩈𩩉𩩊𩩋𩩌𩩍𩩎𩩏𩩐𩩑𩩒𩩓𩩔𩩕𩩖𩩗𩩘𩩙𩩚𩩛𩩜𩩝𩩞𩩟𩩠𩩡𩩢𩩣𩩤𩩥𩩦𩩧𩩨𩩩𩩪𩩫𩩬𩩭𩩮𩩯𩩰𩩱𩩲𩩳𩩴𩩵𩩶𩩷𩩸𩩹𩩺𩩻𩩼𩩽𩩾𩩿𩪀𩪁𩪂𩪃𩪄𩪅𩪆𩪇𩪈𩪉𩪊𩪋𩪌𩪍

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

This content is copyright protected and shall not be shared, uploaded or distributed.

Thank You

