



Tech Talk: Webscrapping Crash Course

By Karim Zakir



Webscraping

- Webscraping is data extraction from websites
- It can be either manual or be automated
- Why am I talking about this?
- Why am I talking about this?
 - Valuable skill to have
 - Benefits for data scientists
 - Benefits for engineers

Basic Tools

- (HTTP) Requests - Generally and in Python
- Pandas - Python
- HTML

```
<div class="flex-section recursion-intro"> flex
  <div class="div-block-89">
    <h2 class="recursion-intro-text dark home">
      "We are a clinical-stage biotechnology company decoding biology by "
      <span class="text-span-35">integrating technological innovations</span>
      " across biology, chemistry, automation, machine learning and engineering "
      <span class="text-span-36">to industrialize drug discovery. </span>
      <br>
      <strong></strong>
    </h2>
    <div class="div-block-110">...</div> flex
  </div>
</div>
```

We are a clinical-stage biotechnology company decoding biology by **integrating technological innovations** across biology, chemistry, automation, machine learning and engineering **to industrialize drug discovery.**



Increased control over biology with tools such as CRISPR genome editing and synthetic biology



Reliable automation of complex laboratory research at unprecedented scale using advanced robotics



Iterative analysis of, and inference from, large, complex in-house datasets using neural network architectures



Increasing elasticity of high performance computation using cloud solutions



HTTP Requests

- Hypertext Transfer Protocol
- Requests contain headers and parameters
 - “Take a letter as example: the text written on the sheet is the PAYLOAD, while the stamp is the headers. Headers needs to delivery [sic] the letter, but does not contain the message inside (payload).” [SO](#)
- Responses contain the status code, along with the body of the response
- Multiple Types of Requests:
 - GET
 - POST
 - PUT
 - DELETE

HTML Breakdown

Tags

Attributes:

- class
- id
- href
- style
- ...

```
<div class="flex-section recursion-intro"> flex
  <div class="div-block-89">
    <h2 class="recursion-intro-text dark home">
      "We are a clinical-stage biotechnology company decoding biology by "
      <span class="text-span-35">integrating technological innovations</span>
      " across biology, chemistry, automation, machine learning and engineering "
      <span class="text-span-36">to industrialize drug discovery. </span>
      <br>
      <strong></strong>
    </h2>
    <div class="div-block-110">...</div> flex
  </div>
</div>
```

Text

Closing
Tags

Our favorite HTML Tag



```
<table>
  <tr>
    <th>Month</th>
    <th>Savings</th>
  </tr>
  <tr>
    <td>January</td>
    <td>$100</td>
  </tr>
  <tr>
    <td>February</td>
    <td>$80</td>
  </tr>
</table>
```

Month	Savings
January	\$100
February	\$80



Formula 1 Website

2022 Driver Standings

POS	DRIVER	NATIONALITY	CAR	PTS
1	Max Verstappen	NED	RED BULL RACING RBPT	233
2	Charles Leclerc	MON	FERRARI	170
3	Sergio Perez	MEX	RED BULL RACING RBPT	163
4	Carlos Sainz	ESP	FERRARI	144
5	George Russell	GBR	MERCEDES	143
6	Lewis Hamilton	GBR	MERCEDES	127
7	Lando Norris	GBR	MCLAREN MERCEDES	70
8	Esteban Ocon	FRA	ALPINE RENAULT	56
9	Valtteri Bottas	FIN	ALFA ROMEO FERRARI	46
10	Fernando Alonso	ESP	ALPINE RENAULT	37



Mid level tools

- Websites vary greatly in complexity and the data isn't always a table
- Requests
- BeautifulSoup4
 - Leverage the structure of the website
 - HTML or XML

```
<?xml version="1.0" encoding="UTF-8" ?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>https://www.formulal.com/en.html</loc>
    <lastmod>2022-07-19</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest.html</loc>
    <lastmod>2022-03-28</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017.html</loc>
    <lastmod>2017-01-03</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/mercedes-to-unveil-new-car-at-silverstone.html</loc>
    <lastmod>2017-03-07</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/force-india-confirm-launch-date-for-2017-car.html</loc>
    <lastmod>2017-01-05</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/mclaren--2017-cars-will-just-look-meaner.html</loc>
    <lastmod>2017-01-06</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/renault-confirm-earliest-car-reveal-date-so-far.html</loc>
    <lastmod>2017-01-06</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/f1-2017-car-launch-pre-season-testing-schedule.html</loc>
    <lastmod>2017-02-26</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/lowe-to-leave-mercedes.html</loc>
    <lastmod>2017-01-10</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/team-principal-vasseur-leaves-renault.html</loc>
    <lastmod>2017-01-11</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/f1-pirelli-tyre-choices-bahrain-russia.html</loc>
    <lastmod>2017-01-12</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/f1-mercedes-rosberg-earn-laureus-nominations.html</loc>
    <lastmod>2017-01-11</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/your-top-f1-overtake-of-2016-max-verstappen.html</loc>
    <lastmod>2017-01-17</lastmod>
  </url>
  <url>
    <loc>https://www.formulal.com/en/latest/headlines/2017/1/wehrlein-joins-sauber-for-2017-season.html</loc>
    <lastmod>2017-01-16</lastmod>
  </url>
</urlset>
```




Dynamic Pages

- Pages that update their content or don't receive all of the content at once
- Selenium
- Other tools/hacks



Applying this to Recursion

[Drugbank](#) + [GeneCards](#)



Honorable Mentions (Tools/Concepts)

- XML
- Postman

If I forget to mention them or don't get to them:

- XPath
- Frames

Code available [here](#)