

WiFi LIVE ONLINE TRAINING

Advanced Web Scraping



MAX HUMBER



<https://resources.oreilly.com/binderhub/advanced-web-scraping>

March 12, 2020

9:00pm – 10:00pm +07

Soup - Wikipedia

https://en.wikipedia.org/wiki/Soup

• [Stone soup](#), a popular children's fable about a poor man who encourages villagers to share their food with him by telling them that he can make soup with a stone

• [Souperism](#), the practice of bible societies during the [Irish Great Famine](#) to feed the hungry in exchange for religious instruction. The expression 'took the soup' is used to refer to those who converted at the behest of these offers of food

• [Tag soup](#), poorly coded [HTML](#)

Inspector Console Debugger Network Style Editor Performance Layout Computed Changes Fo

Search HTML Filter Styles :hov .cls +

element { inline }
a:visited { load.php:1 @screen color: #0b0080; }
a:visited { load.php:1 @screen color: #0b0080; }
a { load.php:1 @screen text-decoration: none; color: #0645ad; background: none; }
a { load.php:1 @screen text-decoration: none; }

Flexbox
Select a Flex container or item to continue.

Grid
CSS Grid is not in use on this page

Box Model

margin 0
border 0
padding 0 56.8167x14.5 0 0 0

The screenshot shows the Firefox Developer Tools Inspector panel. The top part displays the page content with a callout pointing to the 'Tag soup' link. The bottom part shows the HTML tree on the left, the CSS styles on the right, and the box model details on the far right. The box model panel highlights the dimensions of the element as 56.8167x14.5.



A screenshot of a web browser window showing the Wikipedia page for "Soup". The page content discusses the fable of Stone soup and Souperism. A tooltip highlights a link to "Tag soup, poorly coded HTML". The browser's developer tools are open, specifically the Inspector tab, which displays the HTML structure and CSS styles for the highlighted element. The Layout panel shows the box model dimensions: margin 0, border 0, padding 0, and width 56.8167x14.5.

W Wikipedia

https://en.wikipedia.org/wiki/Soup

- [Stone soup](#), a popular children's fable about a poor man who encourages villagers to share their food with him by telling them that he can make soup with a stone
- [Souperism](#), the practice of bible societies during the [Irish Great Famine](#) to feed the hungry in exchange for religious instruction. The expression 'took the soup' is used to refer to those who converted at the behest of these

a | 56.8167 x 14.5 | offers of food

• Tag soup, poorly coded HTML

Inspector Console Debugger Network Style Editor Performance Layout Computed Changes Fo

Search HTML Filter Styles :hov .cls +

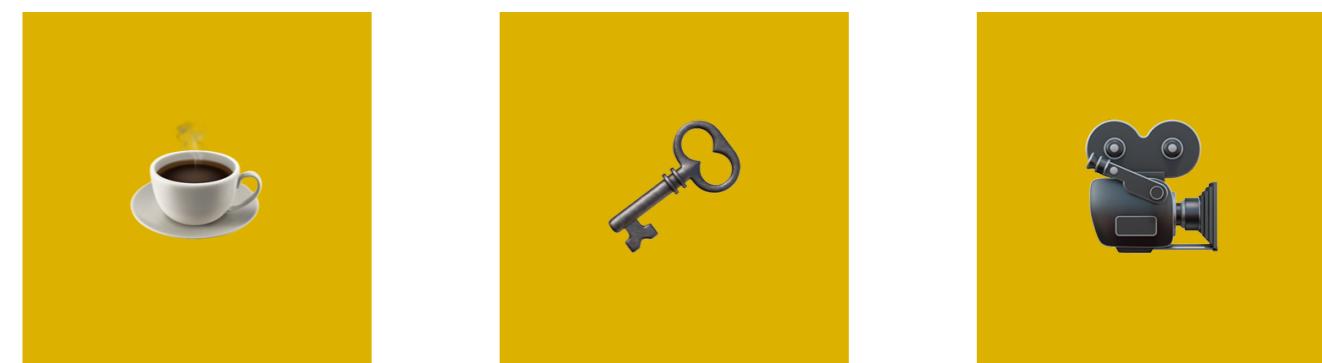
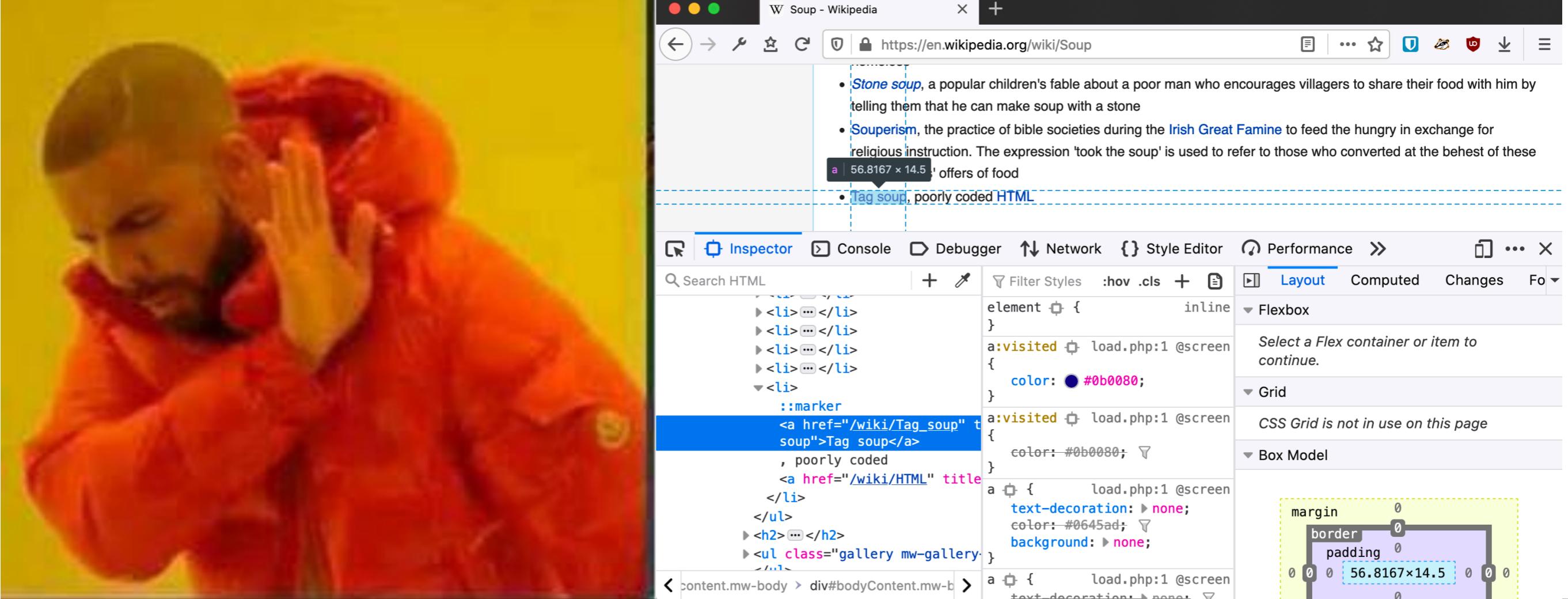
```
<ul>
  <li>...</li>
  <li>...</li>
  <li>...</li>
  <li>...</li>
  <li>
    ::marker
    <a href="/wiki/Tag_soup" title="Tag soup">Tag soup</a>
    , poorly coded
    <a href="/wiki/HTML" title="HTML">HTML</a>
  </li>
</ul>
<h2>...</h2>
<ul class="gallery mw-gallery">
  ...
</ul>
```

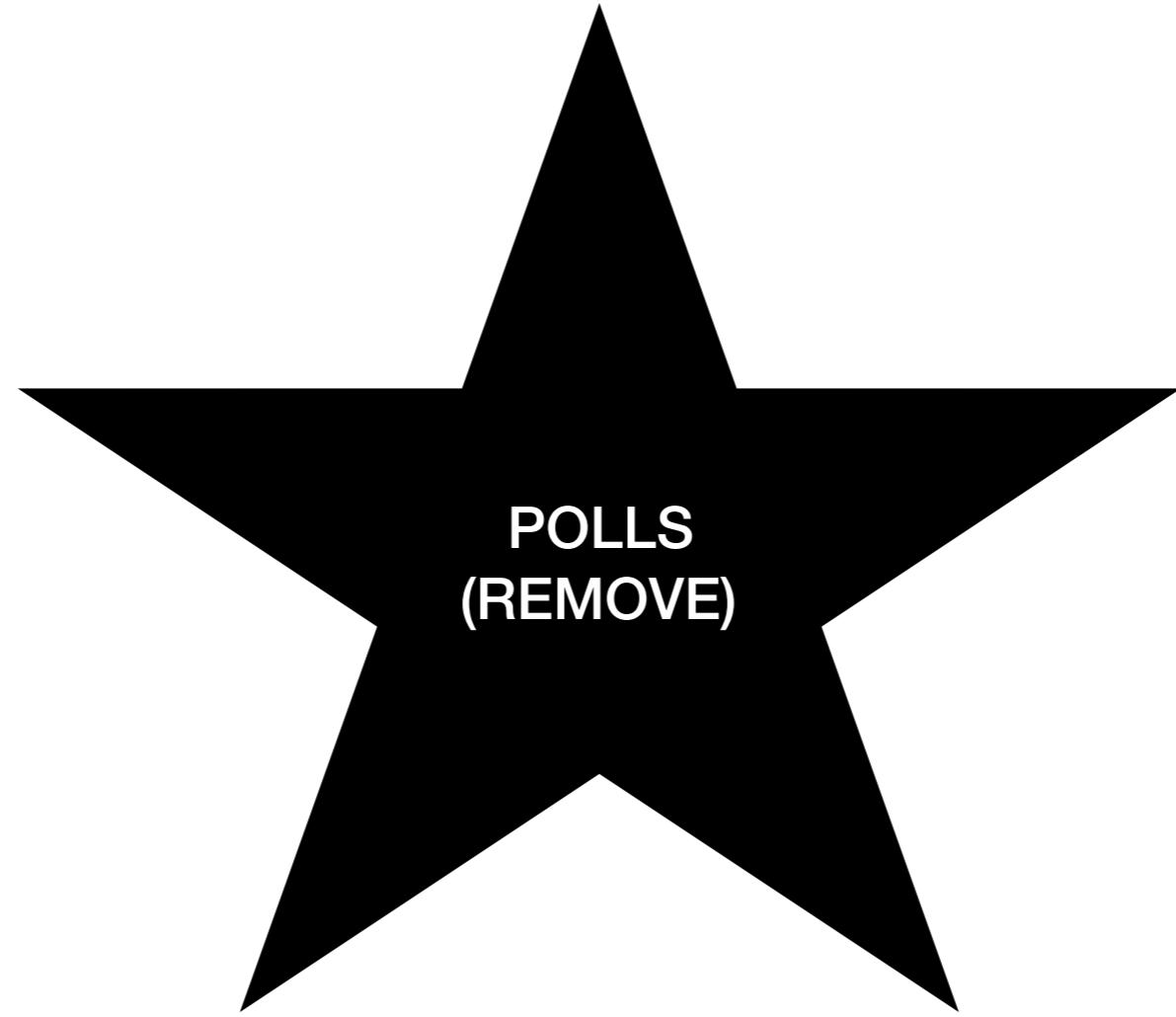
element { inline }
a:visited { load.php:1 @screen
color: #0b0080;
}
a:visited { load.php:1 @screen
color: #0b0080;
}
a { load.php:1 @screen
text-decoration: none;
color: #0645ad;
background: none;
}
a { load.php:1 @screen
text-decoration: none;
color: #0645ad;
background: none;
}

content.mw-body > div#bodyContent.mw-b >

margin 0
border 0
padding 0 56.8167x14.5 0 0 0

Flexbox
Select a Flex container or item to continue.
Grid
CSS Grid is not in use on this page
Box Model

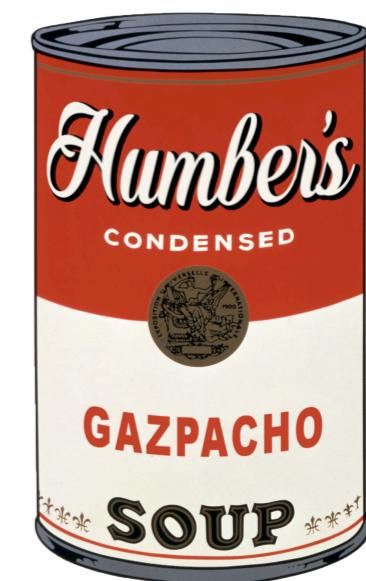




- Poll: How many websites have you scraped before? {0, 1, 10, 100+}
- Poll: Is your interest in web scraping professional or personal? {professional, personal}

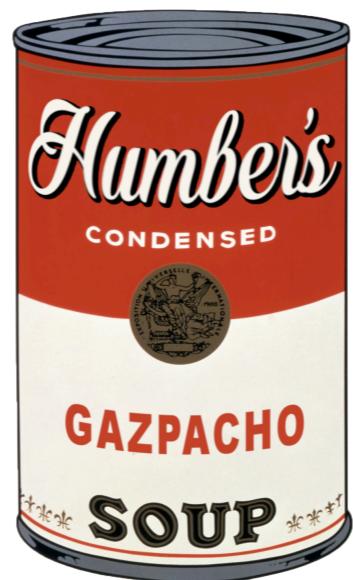




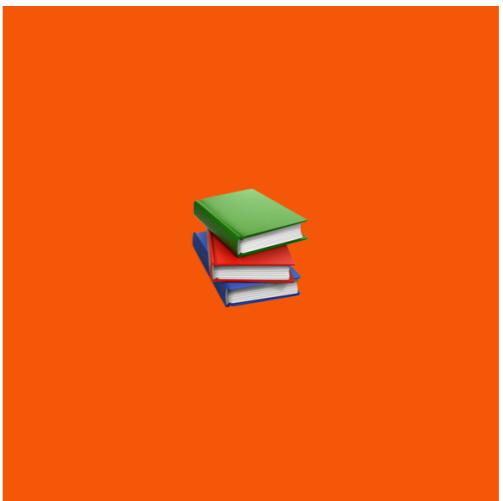
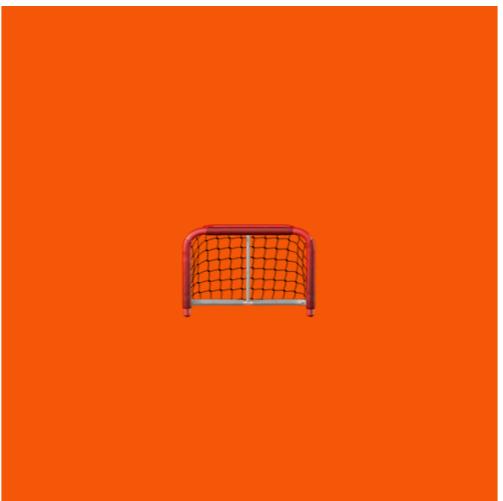
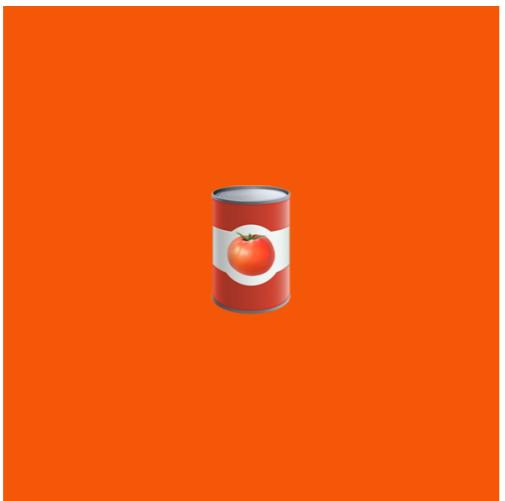




selenium-python.readthedocs.io



gazpacho.xyz

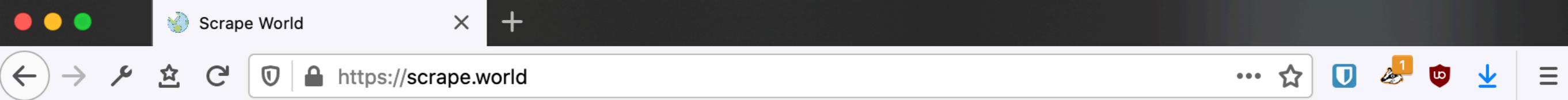


404





<https://www.scrape.world>



Home Challenges

Login

Welcome to Scrape World

This website is meant to be scraped.

The HTML on this website is garbage...

This is on purpose *wink*

Each page is a challenge.

To start a challenge click on the accordion:

Show Challenge

Try to complete each challenge *before* peeking at the solution.

Happy scraping!



LET'S GO

4TH 2:32 24

Q&A

That's all Folks!



April 2, 2020

TELL YOUR FRIENDS!

Web scraping in 60 minutes

April 6, 2020

First Steps: Visualization with Altair

April 9, 2020

Building Recommendation Engines in Python



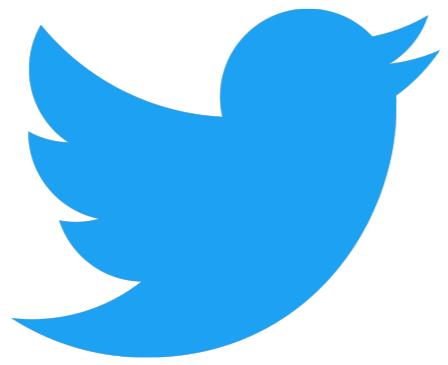
license MIT build passing pypi v1.0.2 downloads 2k

<https://github.com/maxhumber/gif>

MARC

dependencies zero build passing pypi v2.0 downloads 4k

<https://github.com/maxhumber/marc>



twitter.com/maxhumber



www.linkedin.com/in/maxhumber