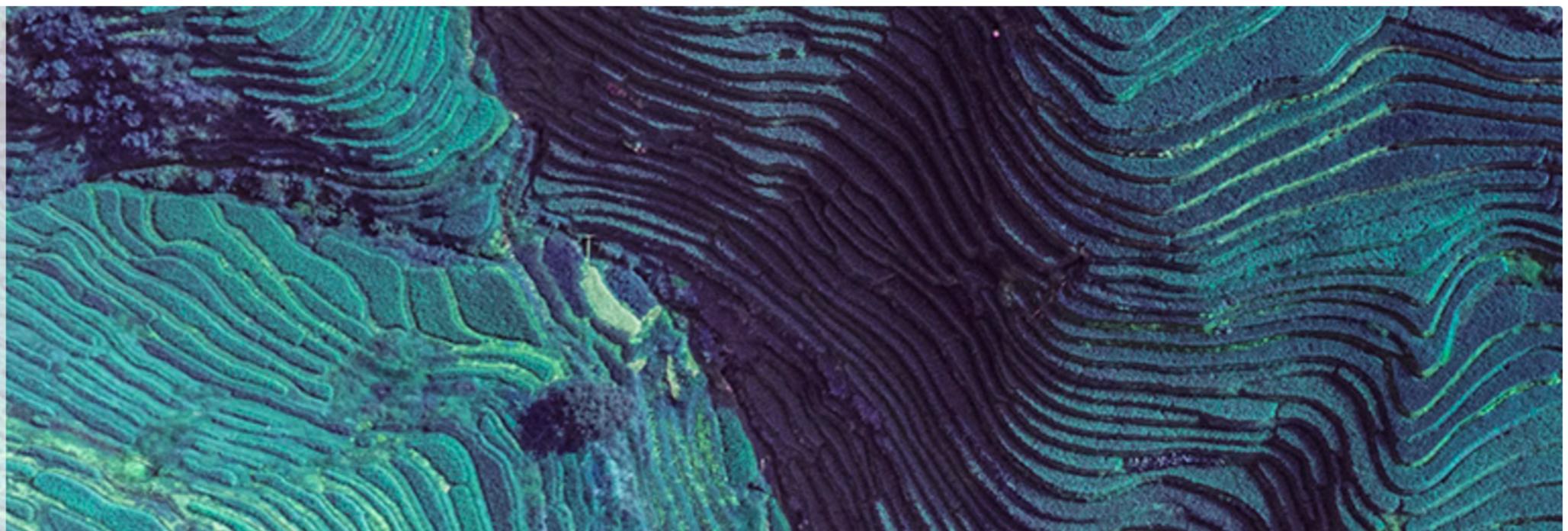


WiFi LIVE ONLINE TRAINING

# Advanced Web Scraping



MAX HUMBER



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

**November 10, 2020**

12:00pm – 1:00pm EST

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; }

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

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



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 +

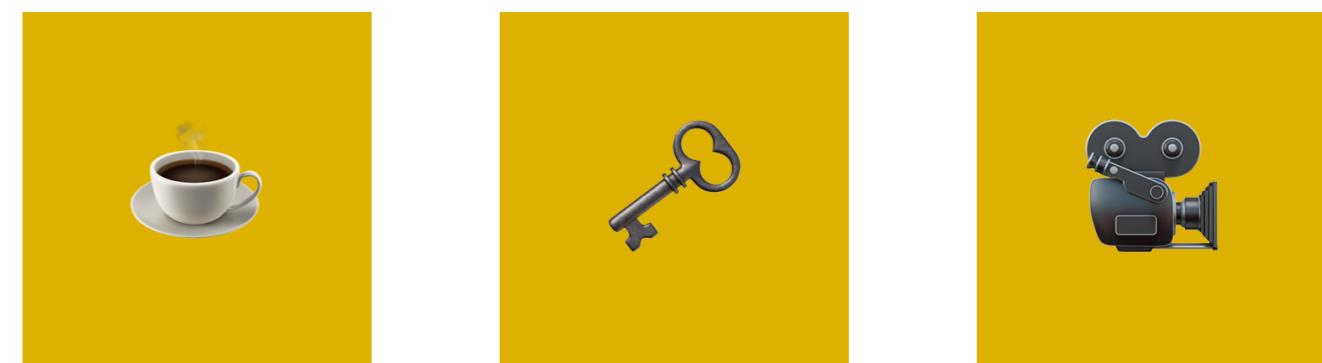
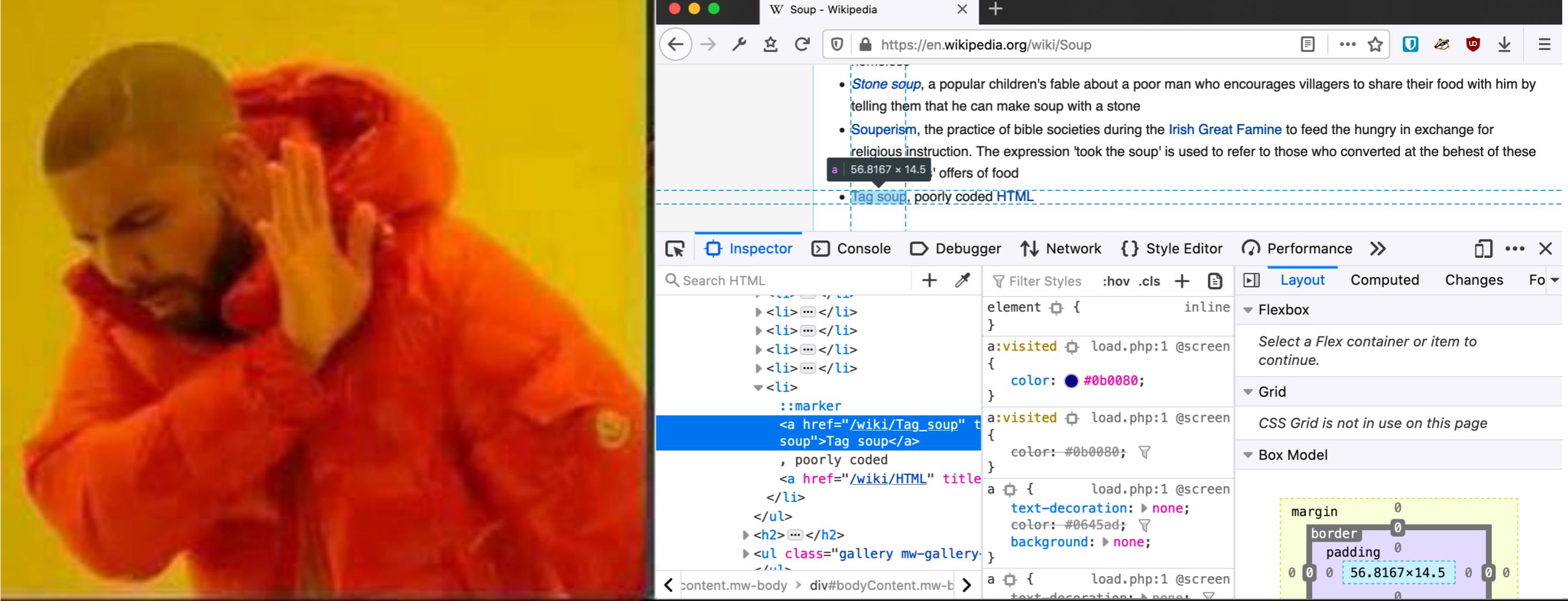
```
<ul>
  <li>...</li>
  <li>...</li>
  <li>...</li>
  <li>...</li>
  <li>
    ::marker
    <a href="/wiki/Tag_soup" title="Tag soup">Tag soup</a>
    , poorly coded
    <a href="/wiki/HTML" title="HTML">HTML</a>
  </li>
</ul>
<h2>...</h2>
<ul class="gallery mw-gallery">
  ...
</ul>
```

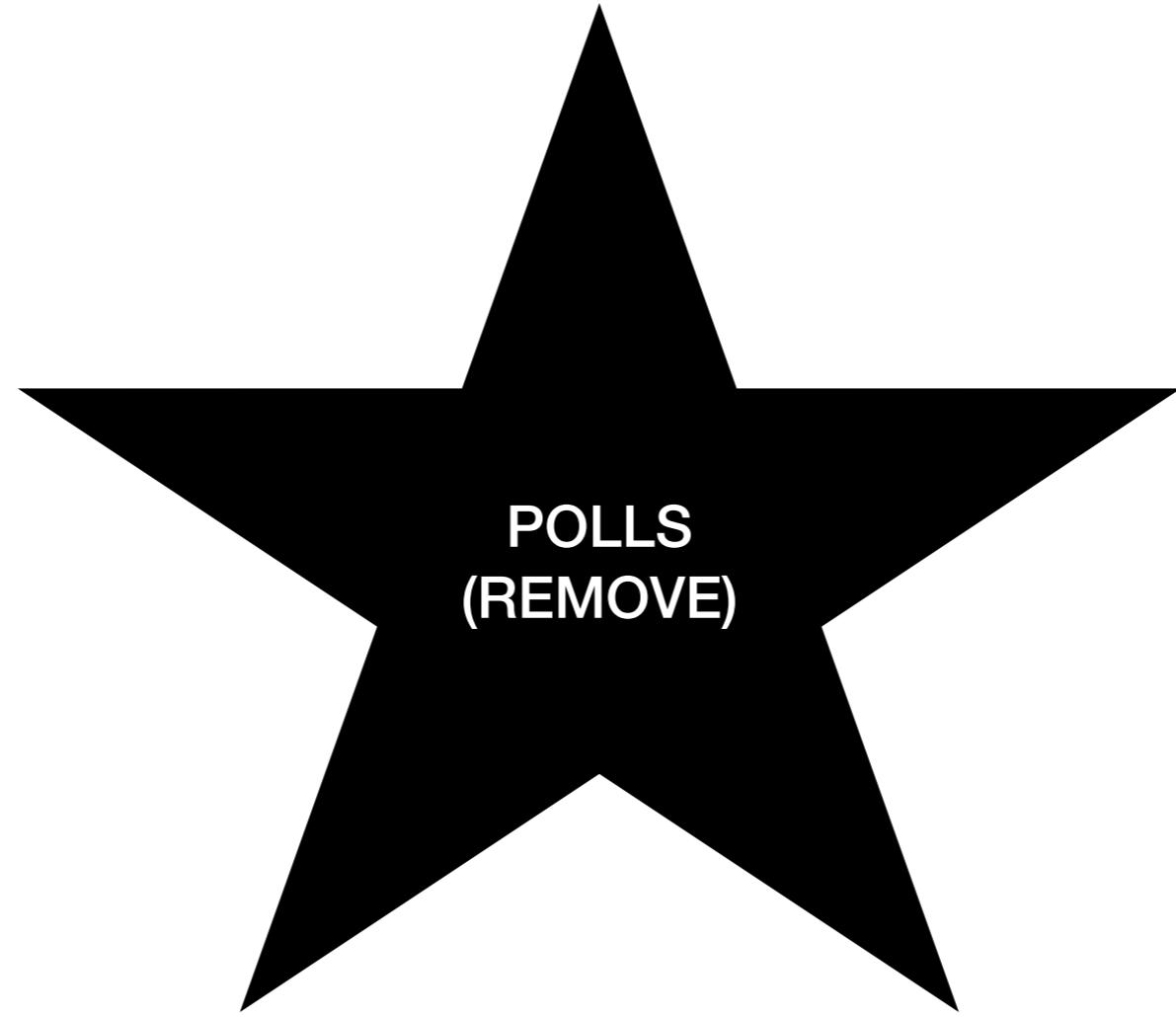
element { inline }
a:visited { load.php:1 @screen
color: #0b0080;
}
a:visited { load.php:1 @screen
color: #0b0080;
}
a { load.php:1 @screen
text-decoration: none;
color: #0645ad;
background: none;
}
a { load.php:1 @screen
text-decoration: none;
color: #0645ad;
background: none;
}

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

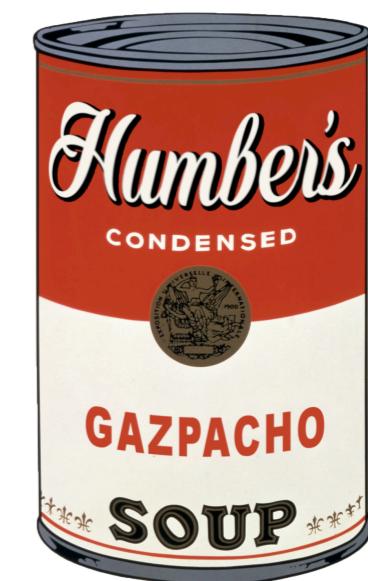




- Poll: How many websites have you scraped before? {0, 1, 10, 100+}
- Poll: Is your interest in web scraping professional or personal? {professional, personal}









33



5



57

...



# BeautifulSoup is so 2000-and-late: Web Scraping in 2020

#python

#webscraping

#gazpacho

#hacktoberfest

- 
1. No dependencies
  - 2. Batteries included**
  - 3. 1 find method**
  4. Production ready
  5. PEP 561 compliant
  6. Automatic formatting
  - 7. Speed (50% faster)**
  - 8. Partial matching**
  9. Debt-free (3 first)
  10. Open (and Friendly)

> 500/week  
> 450 now



**Max Humber**

Human

1mo •

...

🎉 gazpacho JUST TURNED 1 🎉

Some highlights from this year:

- ⬇️ 20,000 installs
- ⭐️ 300 stars
- 💻 Used by teams at Apple, Microsoft, and Facebook

To mark the occasion I've just bumped the library out of ZeroVer hell and released version 1.0!

Improvements include:

- ❓ type: hints = everywhere
- 🖼️ The ability to parse malformed void tags
- ➕ A new way to initialize a parser with: `Soup.get("url")`

Details: <https://gazpacho.xyz/>

And if you'd like to help build out the roadmap for gazpacho please fill out this quick (seven questions, all optional) survey:

<https://lnkd.in/gKWzRqM>



**Max Humber**

Human

2w •

...

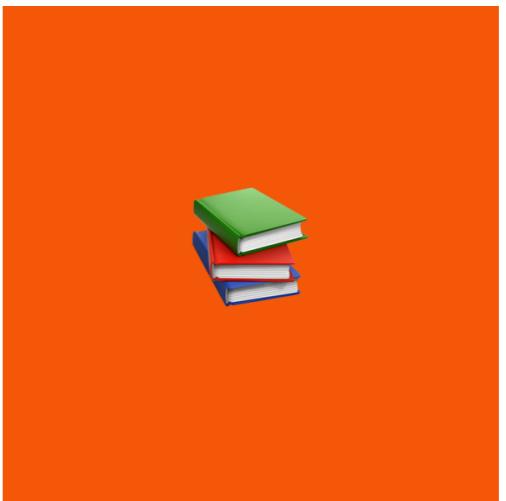
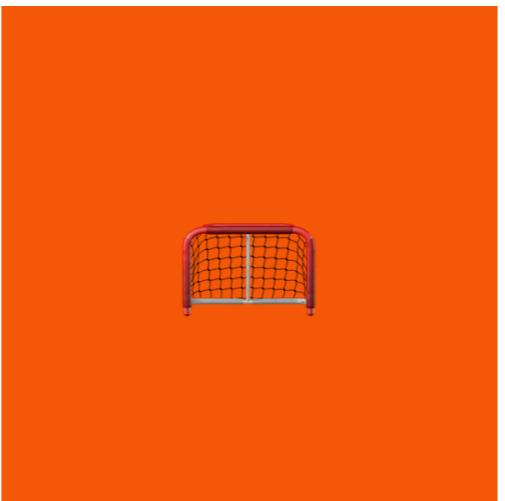
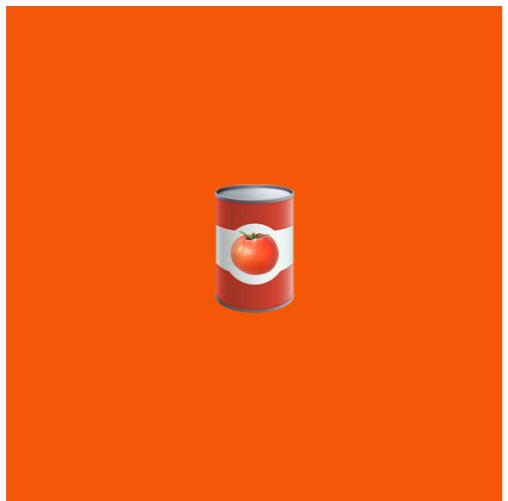
gazpacho is #1 on **#Hacktoberfest** right now

After a full year of development it finally feels like the \*most beautiful soup\* 😊 has some real momentum behind it

The screenshot shows a web browser window with the title bar "Hacktoberfest presented by DigitalOcean". The address bar displays the URL "https://hacktoberfest.digitalocean.com". The main content area is titled "Hacktoberfest projects" and contains the following text: "You can contribute to participating projects on GitHub. Here are a few looking for some help:". Below this, there is a dropdown menu labeled "Select language". The page displays six project cards arranged in two rows of three:

Project Name	Language	Description
gazpacho	Python	The simple, fast, and modern web scraping library
python-backgroundchanger	Python	A simple background changer using Unsplash API
LeetCode_Algorithms	Python	A collection of solutions for Medium/Hard LeetCode problems. Educational resource
jina	Python	An easier way to build neural search in the cloud
tmdb-app	JavaScript	An unofficial client app to The Movie Database made with React Native
github-readme-mediumizer	JavaScript	Post medium card in Github README

Each project card includes a small image, the project name, its language, a brief description, and a progress indicator at the bottom.

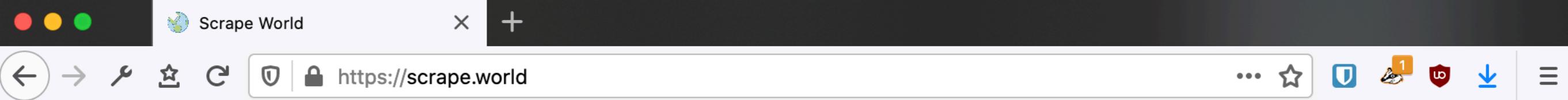


# 404





<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!*



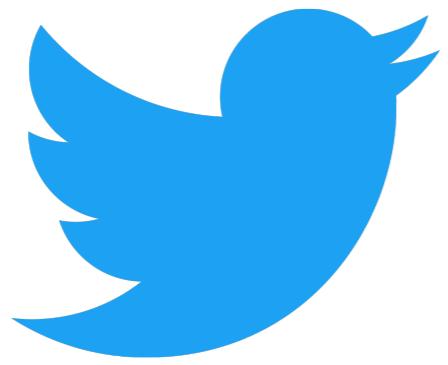
license MIT build passing pypi v1.0.2 downloads 2k

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

# MARC

dependencies zero build passing pypi v2.0 downloads 4k

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



[twitter.com/maxhumber](https://twitter.com/maxhumber)



[www.linkedin.com/in/maxhumber](https://www.linkedin.com/in/maxhumber)