https://gazpacho.xyz/

gazpacho is a web scraping library.
It replaces requests and
BeautifulSoup for *most* projects.

https://gazpacho.xyz/


y tho

X0,000 LOC
not on GitHub
lxml dependancy
learning curve
15 years old

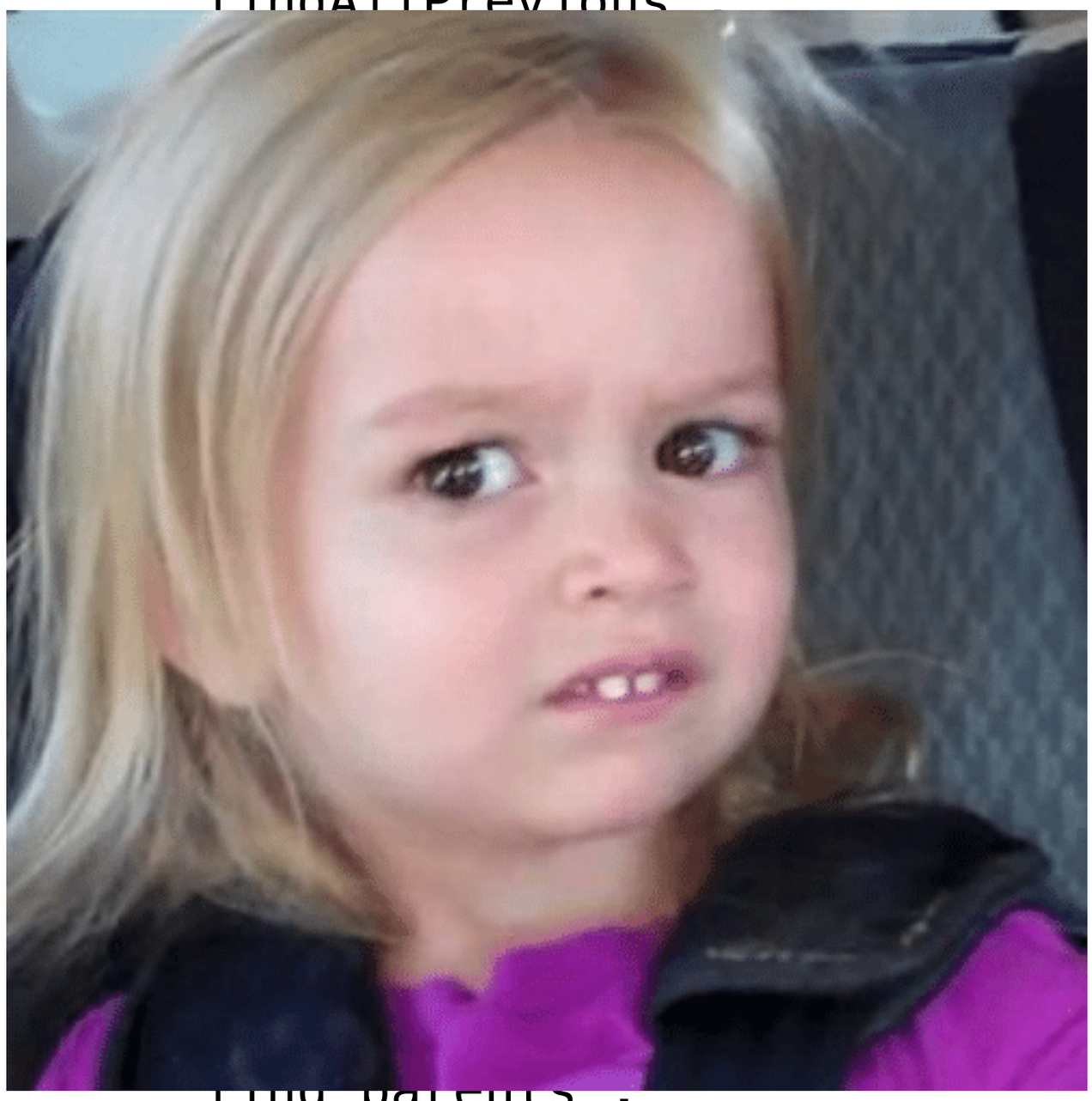
```
[‘find',
 'findAll',
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious',
 'findChild',
 'findChildren',
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious',
 'findChild',
 'findChildren',
 'findNext',
 'findNextSibling',
 'findNextSiblings',
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious',
 'findChild',
 'findChildren',
 'findNext',
 'findNextSibling',
 'findNextSiblings',
 'findParent',
 'findParents',
 'findPrevious',
 'findPreviousSibling',
 'findPreviousSiblings',
 'find_all',
 'find_all_next',
 'find_all_previous',
 'find_next',
 'find_next_sibling',
 'find_next_siblings',
 'find_parent',
 'find_parents',
 'find_previous',
 'find_previous_sibling',
 'find_previous_siblings']
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious'
```



```
 'find_parents',
 'find_previous',
 'find_previous_sibling',
 'find_previous_siblings']
```

find

Schedule - Day 2

Download schedule as CSV! Sign up for tutorials

DAY 1
Saturday, November 16

DAY 2
Sunday, November 17

SPRINTS
Monday, November 18 - Tuesday, November 19

8:30	Breakfast & Registration	30 mins

tr | 989.383 × 169

	Innovating in unusual places **Concert Hall** Sarah Sun	25 mins
	Operator overloading: you're doing it wrong **Sky Room** Greg Ward	25 mins

Inspector | Console | Debugger | Network | Style Editor | Performance | Memory | Storage | Accessibility

Search HTML

```
<tr> … </tr>
<tr>
  <td class="time" rowspan="4">
    11:00
  </td>
  <td class="schedule-table--col-description">
    <a href="../talks/talk-378/">
      Innovating in unusual places
    </a>
    <br>
    <b>Concert Hall</b>
    <br>
    Sarah Sun
  </td>
  <td class="duration">25 mins</td>
</tr>
<tr> … </tr>
```

html > body > div.container > div > div.container > table.schedule-table > tbody > tr > td.schedule-table--col-description > a

Filter Styles :hov .cls

```
element { inline
}

a { _reboot.scss:182
  color: #1284a3;
}

.font-weight-bold, a.blue-button _text.scss:39
span, a.invisible-button span, a.red-
button span, .h2, a, a.navmenu-item span,
a.navmenu-item-selected span, .section-header,
.footer-heading, .sponsor-title, .sponsor-button-
left, .sponsor-button-right, .sponsor-header,
.card-title {
  font-weight: 700 !important;
}

a { _reboot.scss:182
  color: #007bff;
  text-decoration: none;
  background-color: transparent;
}
```

```
<tr> … </tr>
<tr>
  <td class="time" rowspan="4">
    11:00
  </td>
  <td class="schedule-table--col-description">
    <a href="../talks/talk-378/">
      Innovating in unusual places
    </a>
    <br>
    <b>Concert Hall</b>
    <br>
    Sarah Sun
  </td>
  <td class="duration">25 mins</td>
</tr>
<tr> … </tr>
```

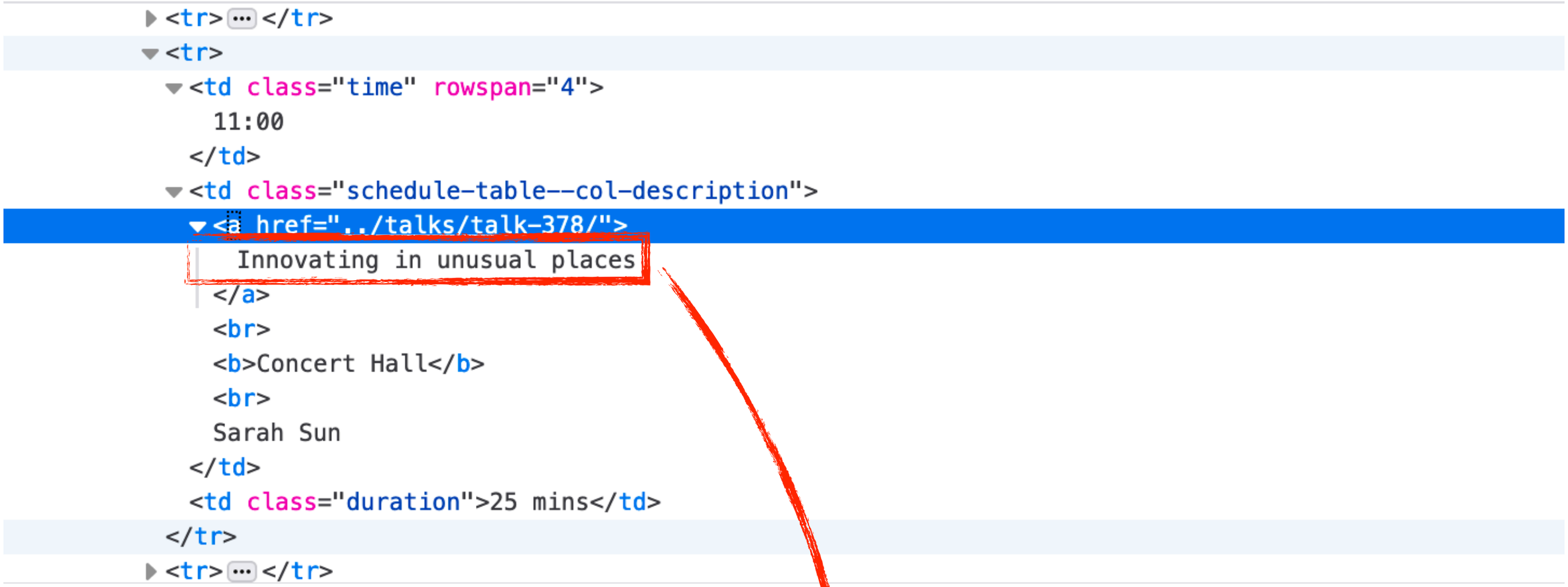
```
<tr> … </tr>
<tr>
  <td class="time" rowspan="4">
    11:00
  </td>
  <td class="schedule-table--col-description">
    <a href="../talks/talk-378/">
      Innovating in unusual places
    </a>
    <br>
    <b>Concert Hall</b>
    <br>
    Sarah Sun
  </td>
  <td class="duration">25 mins</td>
</tr>
<tr> … </tr>
```

```
find('td', {'class': 'time'})
```


▶ <tr> ··· </tr>
▼ <tr>
▼ <td class="time" rowspan="4">
11:00
</td>
▼ <td class="schedule-table--col-description">
▼
Innovating in unusual places

Concert Hall

Sarah Sun
</td>
<td class="duration">25 mins</td>
</tr>
▶ <tr> ··· </tr>
▶ <tr> ··· </tr>
find('a').attrs['href']


<tr> … </tr>
<tr>
<td class="time" rowspan="4">
11:00
</td>
<td class="schedule-table--col-description">

Innovating in unusual places

Concert Hall

Sarah Sun
</td>
<td class="duration">25 mins</td>
</tr>
<tr> … </tr>

find(‘a').text

```
<tr> … </tr>
<tr>
  <td class="time" rowspan="4">
    11:00
  </td>
  <td class="schedule-table--col-description">
    <a href="../talks/talk-378/">
      Innovating in unusual places
    </a>
    <br>
    <b>Concert Hall</b>
    <br>
    Sarah Sun
  </td>
  <td class="duration">25 mins</td>
</tr>
<tr> … </tr>
```

```
find('td', {'class': ‘duration’}).text
```

```
from gazpacho import Soup, get

html = get('https://2019.pycon.ca/schedule-day-2/')
soup = Soup(html)
trs = soup.find('tr')

def parse_tr(tr):
    title = tr.find('a').text
    link = tr.find('a').attrs['href'].replace('..', '')
    return {'title': title, 'link': link}

talks = []
for tr in trs:
    try:
        talks.append(parse_tr(tr))
    except AttributeError:
        pass
```

```
[{'title': 'Innovating in unusual places', 'link': '/talks/
talk-378/'},
 {'title': "Operator overloading: you're doing it wrong",
  'link': '/talks/talk-87/'},
 {'title': 'PySpark: avoiding common pitfalls and keeping your
sanity',
  'link': '/talks/talk-169/'},
 {'title': 'Python & Kubernetes a match made in the cloud',
  'link': '/talks/talk-T-603/'},
 {'title': 'Making multiple inheritance not work', 'link': '/ta
talk-211/'},
 {'title': 'Introduction to asynchronous programming',
  'link': '/talks/talk-25/'},
 {'title': 'Feature engineering: an apprentice’s guide to the d
art” of machine learning',
  'link': '/talks/talk-23/'},
 {'title': 'Fun with compilers: exploring languages one Python
time',
  'link': '/talks/talk-225/'},
 {'title': 'Put Your Data in a Box', 'link': '/talks/talk-112/'
 {'title': 'Energy indicators and Jupyter Notebooks at the Cana
Energy Regulator',
  'link': '/talks/talk-82/'},
```

```
import pandas as pd

pd.DataFrame(talks)
```

	link	title
0	/talks/talk-378/	Innovating in unusual places
1	/talks/talk-87/	Operator overloading: you're doing it wrong
2	/talks/talk-169/	PySpark: avoiding common pitfalls and keeping ...
3	/talks/talk-T-603/	Python & Kubernetes a match made in the cloud
4	/talks/talk-211/	Making multiple inheritance not work
	/talks/talk-	Introduction to asynchronous

✅ 0 dependancies
✅ one method (find)
✅ X00 LOC
✅ fast (30% & 300%)
✅ 1 package

https://gazpacho.xyz/

```
pip install gazpacho
```

https://learning.oreilly.com/live-training/

https://pythonbytes.fm/episodes/show/152/

hit me up