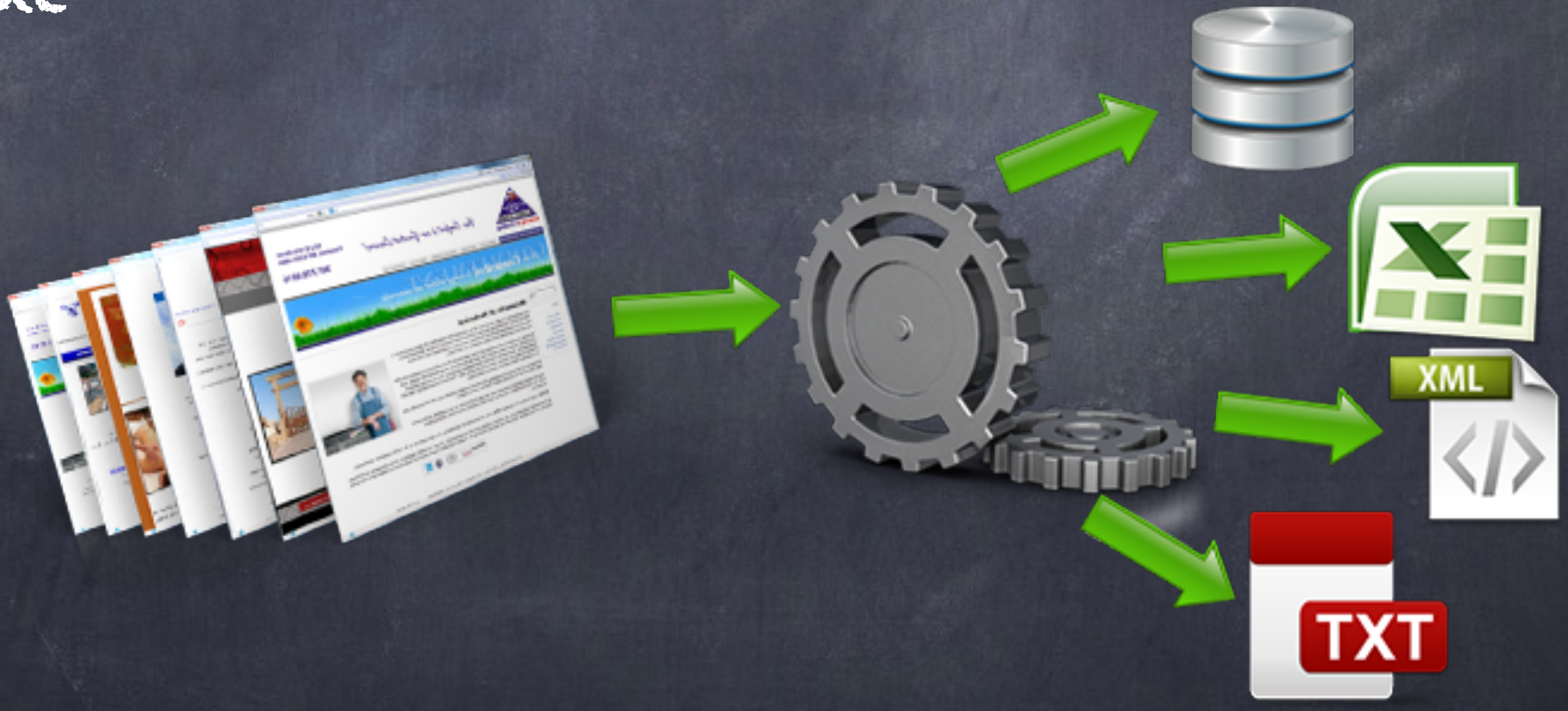
how to scrape data

in five minutes
(using Python)

felice ho
data team

what is web scraping

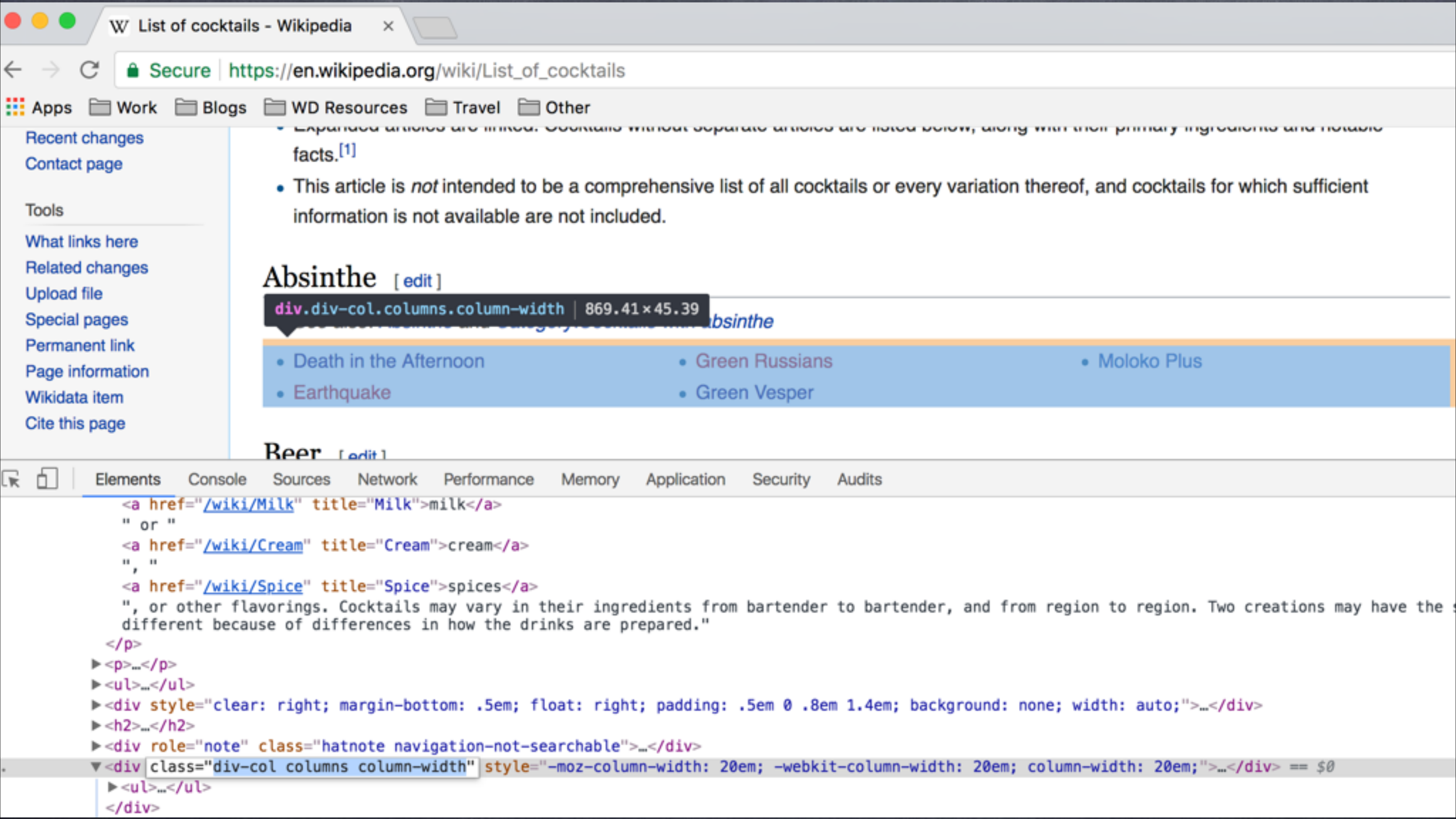
extract data from websites into a useful format



sample use cases

- monitor pricing
- track scores
- automate data collection
- conduct data experiment / research

step 1: know what you are scraping



- Expanded articles are linked. Cocktails without separate articles are listed below, along with their primary ingredients and notable facts.[1]
- This article is *not* intended to be a comprehensive list of all cocktails or every variation thereof, and cocktails for which sufficient information is not available are not included.

Absinthe [edit]

div.div-col.columns.column-width | 869.41 × 45.39

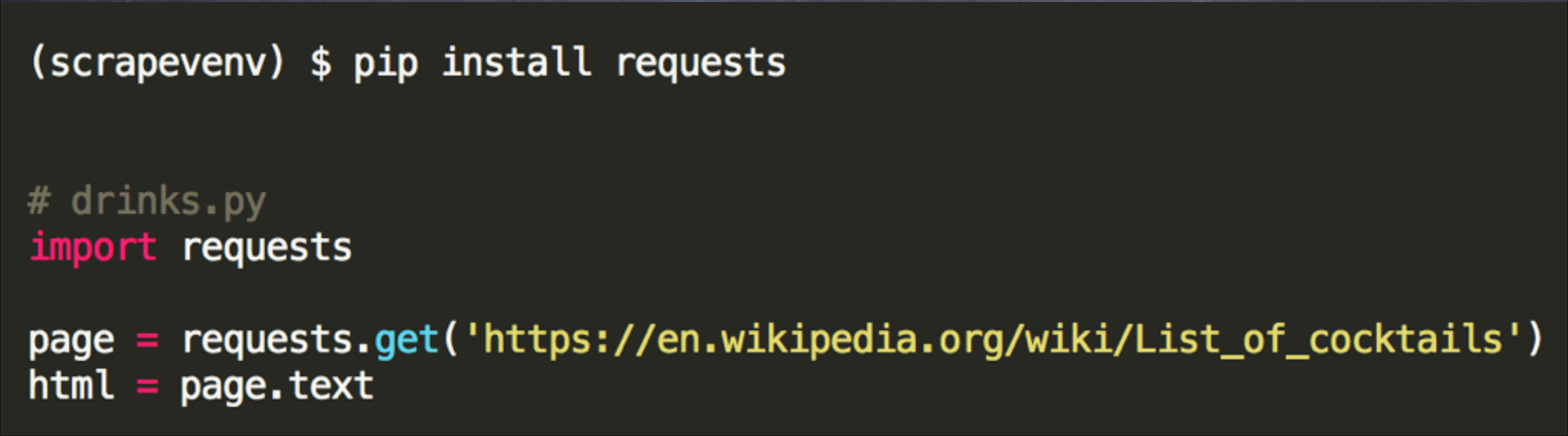
absinthe

- Death in the Afternoon
- Earthquake
- Green Russians
- Green Vesper
- Moloko Plus

Beer [edit]

```
      <a href="/wiki/Milk" title="Milk">milk</a>
      " or "
      <a href="/wiki/Cream" title="Cream">cream</a>
      ", "
      <a href="/wiki/Spice" title="Spice">spices</a>
      ", or other flavorings. Cocktails may vary in their ingredients from bartender to bartender, and from region to region. Two creations may have the s
      different because of differences in how the drinks are prepared."
    </p>
  ▶<p>…</p>
  ▶<ul>…</ul>
  ▶<div style="clear: right; margin-bottom: .5em; float: right; padding: .5em 0 .8em 1.4em; background: none; width: auto;">…</div>
  ▶<h2>…</h2>
  ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
  ▼<div class="div-col columns column-width" style="-moz-column-width: 20em; -webkit-column-width: 20em; column-width: 20em;">…</div> == $0
    ▶<ul>…</ul>
    </div>
```

step 2: connect to the page

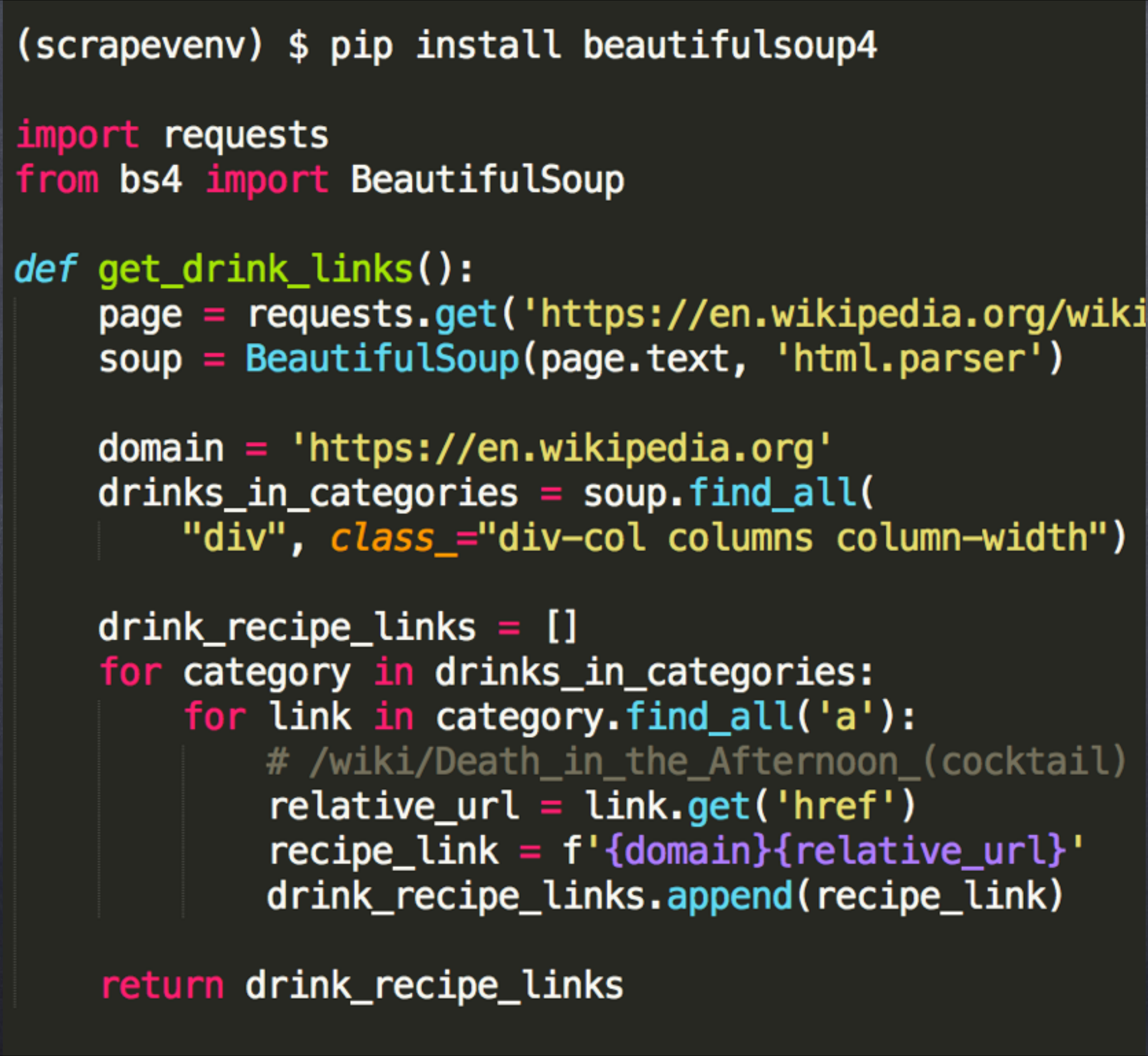


```
(scrapevenv) $ pip install requests
```

```
# drinks.py
import requests

page = requests.get('https://en.wikipedia.org/wiki/List_of_cocktails')
html = page.text
```

step 3: parse html



```
(scrapevenv) $ pip install beautifulsoup4

import requests
from bs4 import BeautifulSoup

def get_drink_links():
    page = requests.get('https://en.wikipedia.org/wiki
    soup = BeautifulSoup(page.text, 'html.parser')

    domain = 'https://en.wikipedia.org'
    drinks_in_categories = soup.find_all(
        "div", class_="div-col columns column-width")

    drink_recipe_links = []
    for category in drinks_in_categories:
        for link in category.find_all('a'):
            # /wiki/Death_in_the_Afternoon_(cocktail)
            relative_url = link.get('href')
            recipe_link = f'{domain}{relative_url}'
            drink_recipe_links.append(recipe_link)

    return drink_recipe_links
```

thank you!