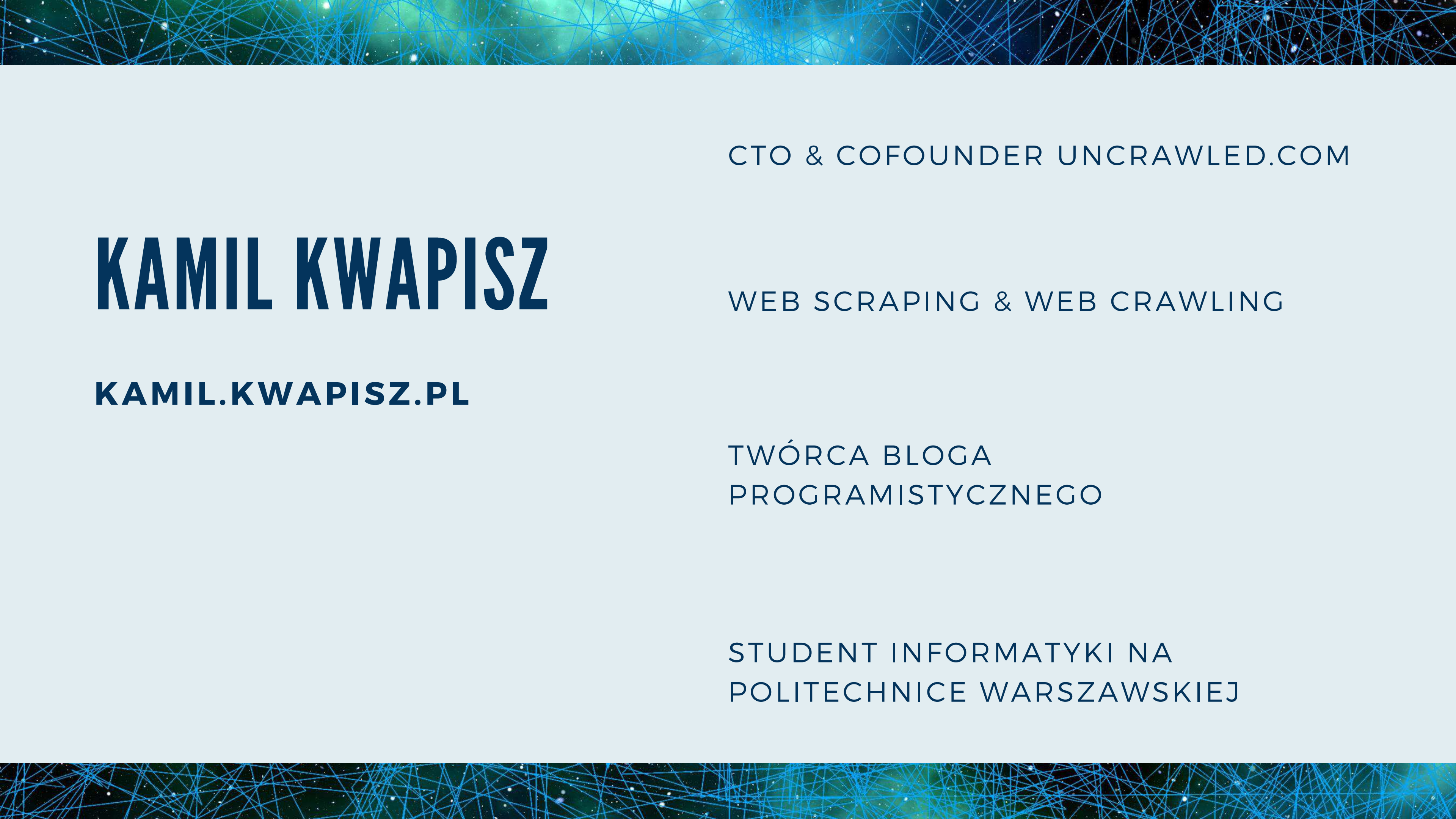


KAMIL KWAPISZ

WEB SCRAPING & WEB CRAWLING

PyWaw 2020



KAMIL KWAPISZ

KAMIL.KWAPISZ.PL

CTO & COFOUNDER UNCRAWLED.COM

WEB SCRAPING & WEB CRAWLING

TWÓRCA BLOGA
PROGRAMISTYCZNEGO

STUDENT INFORMATYKI NA
POLITECHNICE WARSZAWSKIEJ



Czym jest web scraping?

[KAMIL.KWAPISZ.PL/WEB-SCRAPING-PYTHON](https://kamil.kwapisz.pl/web-scraping-python)



Web Scraping - proces
przechodzenia botów po
stronie internetowej oraz
ekstrakcji konkretnych danych.

An abstract background featuring a dense, chaotic network of thin blue lines. A bright green light source in the upper left corner creates a radial glow, illuminating the lines and casting a greenish tint over the scene. The overall effect is one of digital connectivity and complexity.

**web
scraping**

VS

An abstract background featuring a dense, chaotic network of thin blue lines. A bright green light source in the lower right corner creates a radial glow, illuminating the lines and casting a greenish tint over the scene. The overall effect is one of digital connectivity and complexity.

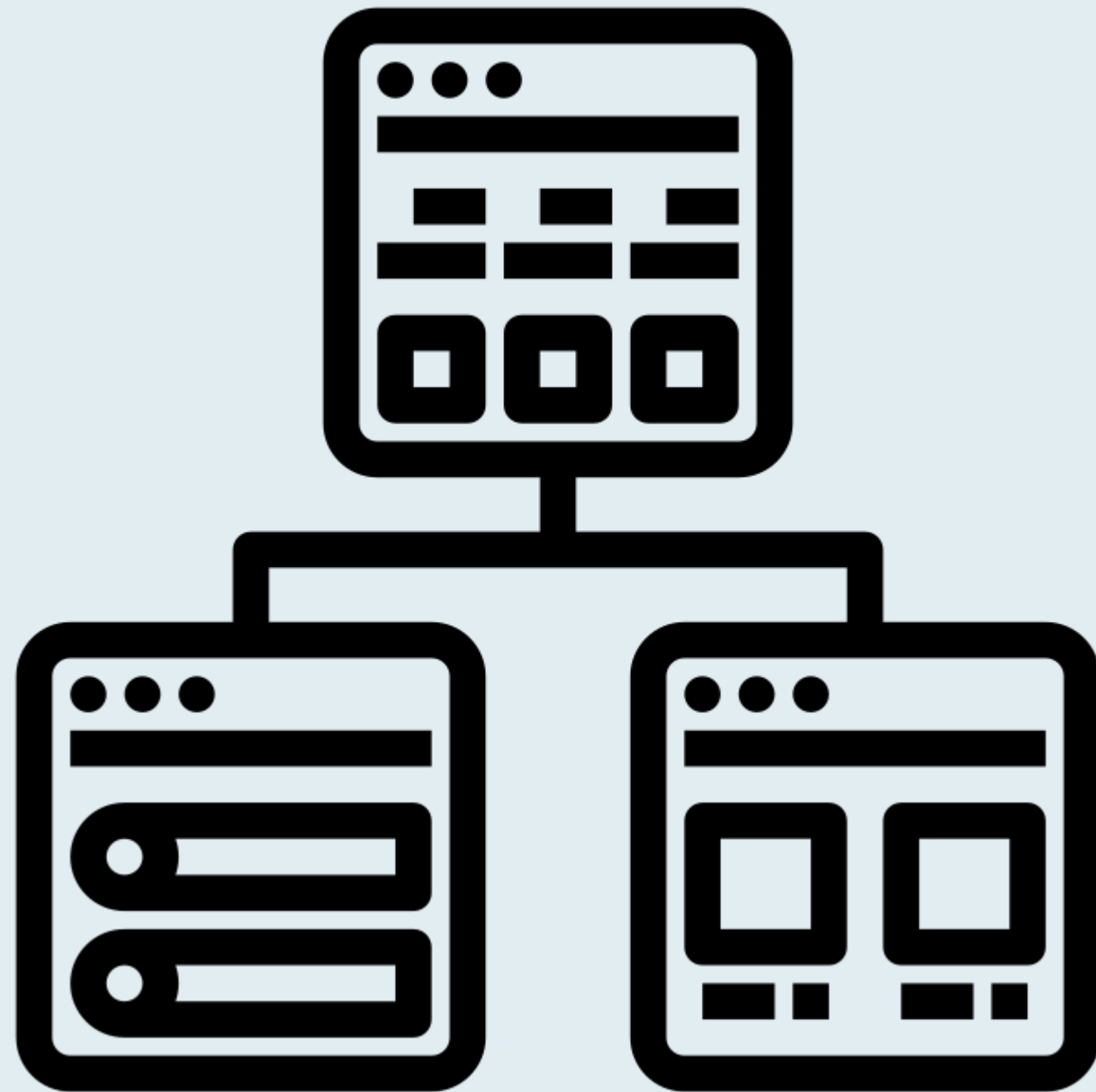
**web
crawling**

Crawlowanie:

Strategiczne

Ad hoc

Po sitemapie





Narzędzia w Pythonie

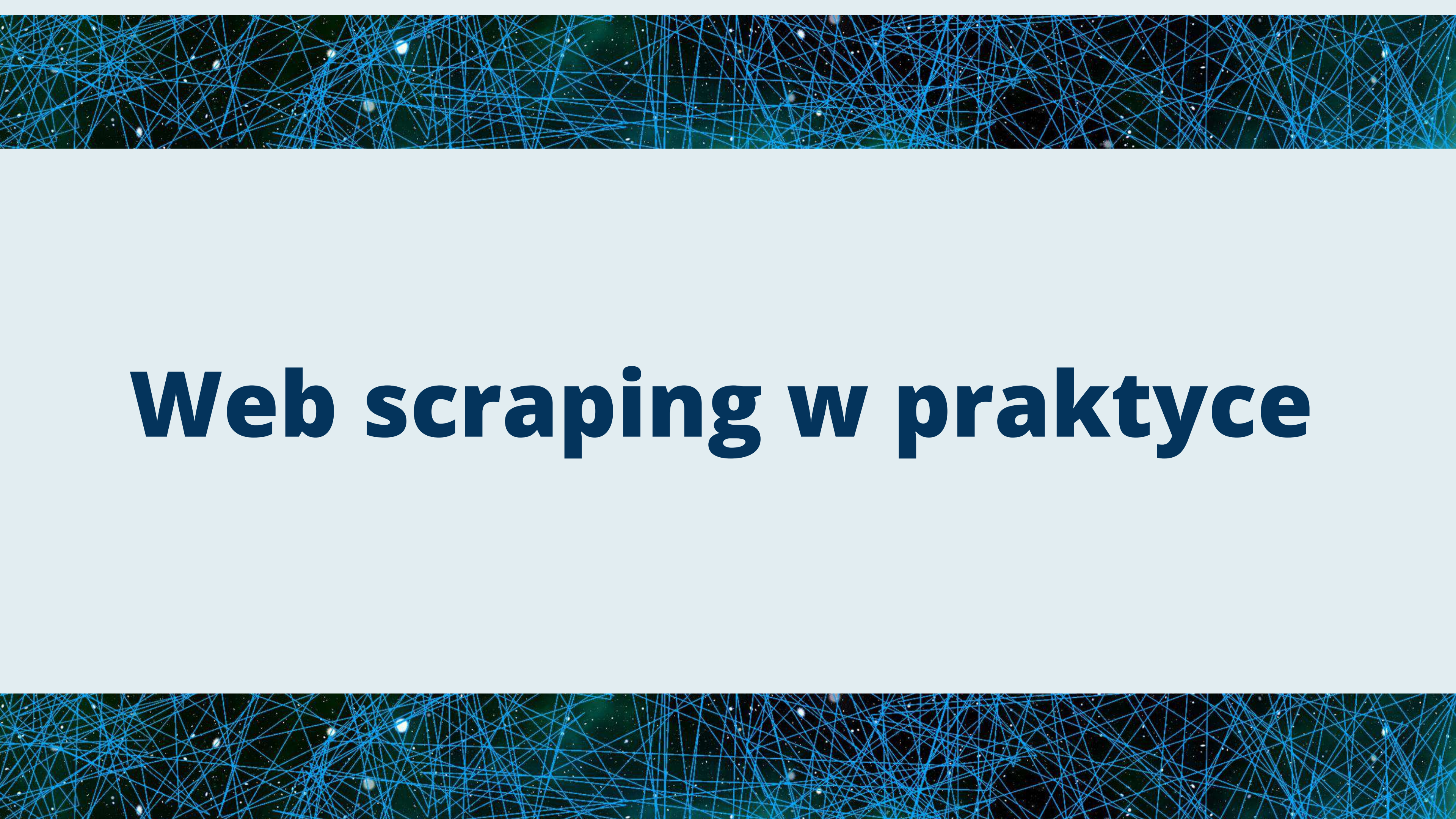
Biblioteki

requests

BeautifulSoup

requests_html

Scrapy



Web scraping w praktyce


```
from requests_html import HTMLSession
```

```
headers = {  
    'User-Agent': "Kamil.Kwapisz.pl"  
}
```

```
url = "http://pywaw.org/spotkania/"
```

```
session = HTMLSession()
```

```
r = session.get(url, headers=headers)
```

```
print(r.status_code) # 200
```


[Home](#)[Spotkania](#)[Prelegenci](#)[Sponsorzy](#)[FAQ](#)

Spotkania

PyWaw #94

18.05.2020**SZCZEGÓŁY**

Kamil Kwapisz: Web scraping i web crawling

Piotr Tynecki: Fagoterapia 2.0, czyli Python w służbie walki z infekcjami bakteryjnymi

PyWaw #93

30.03.2020**SZCZEGÓŁY**

Szymon Przedwojski: Testowanie jednostkowe – jak robić to dobrze?

Zbigniew Siciarz: Gdyby Sokrates programował, tudzież o obserwowalności słów kilka

PyWaw #92

17.02.2020**SZCZEGÓŁY**

Jarek Potiuk, Tomek Urbaszek: Introduction to Airflow

Grzegorz Kocjan: Od legacy do czystej architektury

PyWaw #91

20.01.2020**SZCZEGÓŁY**

Aleksander Kawala: Django DB Optimizations


```
▼ <div class="row">
  ::before
  ▼ <div class="span10 list-caption">
    ► <div class="pull-right">...</div>
```

```
▼ <h2>
  ▼ <a href="/94/" itemprop="sameAs">
```

```
    <span itemprop="name">PyWaw #94</span> == $0
```

```
  </a>
```

```
</h2>
```

```
</div>
```

```
::after
```

```
</div>
```

```
► <ul class="unstyled">...</ul>
```

html body #wrapper div div div div div div **h2 a** span


```
from urllib.parse import urljoin

relative_links = r.html.find('h2 > a')
speech_links = list()

for link in relative_links:
    href = link.attrs.get('href')
    speech_link = urljoin(link.url, href)
    speech_links.append(speech_link)

print(speech_links[:2])
```

```
['http://pywaw.org/94/', 'http://pywaw.org/93/']
```



```
for speech_url in speech_link:  
    data = scrape_details(speech_url)
```

```
def scrape_details(url):  
    session = HTMLSession()  
    r = session.get(url, headers=headers)  
    data = extract_data(r.html)
```

```
def extract_data(html):  
    speakers = html.xpath('.//h5//a//text()')  
    talks = html.xpath(".//h3//span[@itemprop='name']//text()")[:2]  
    date = r.html.find('h4', first=True).text  
    descriptions = r.html.xpath(".//p[@itemprop='description']//text()")[:2]  
    # [...]
```

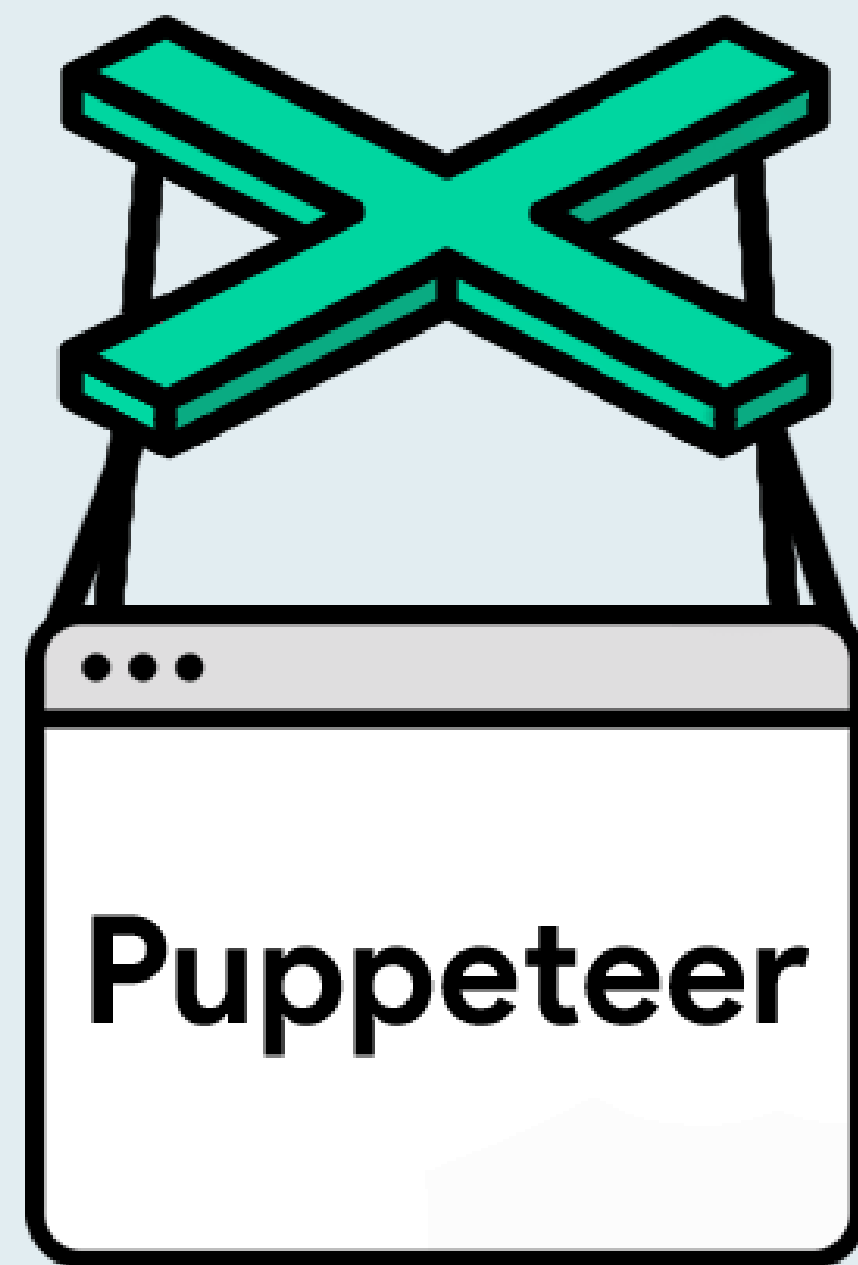



Problemy w web scrapingu

The background of the image is a dark blue field with a complex, abstract pattern of thin, light blue lines that crisscross and intersect, creating a web-like or neural network appearance. Scattered throughout this pattern are small, bright white and light blue dots, resembling stars or data points in a digital space.

{.js}

JavaScript




```
r = session.get('https://zawodprogramista.pl/')  
print(r.status_code)
```

200

```
r.html.find('h1')
```

[]

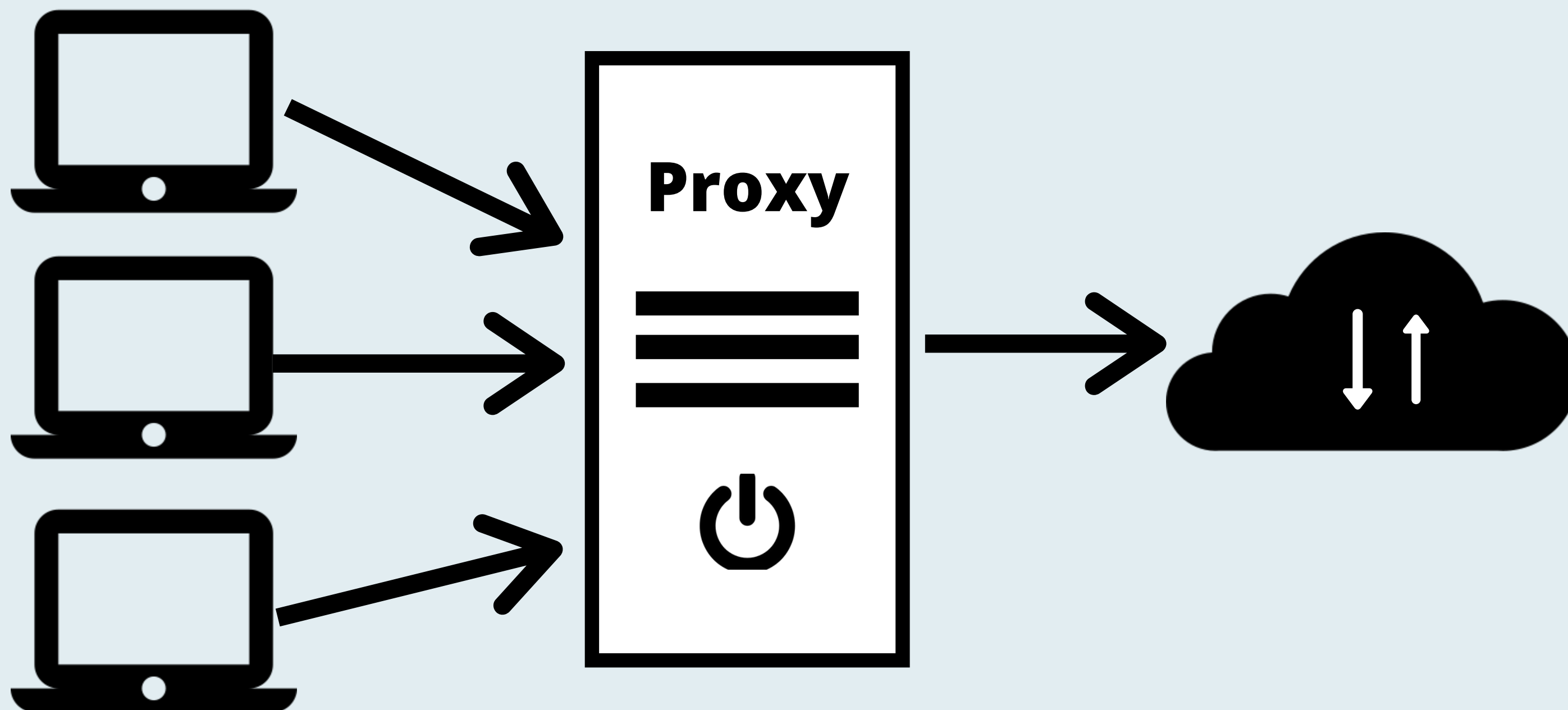
```
r.html.render()
```

```
r.html.find('h1')
```

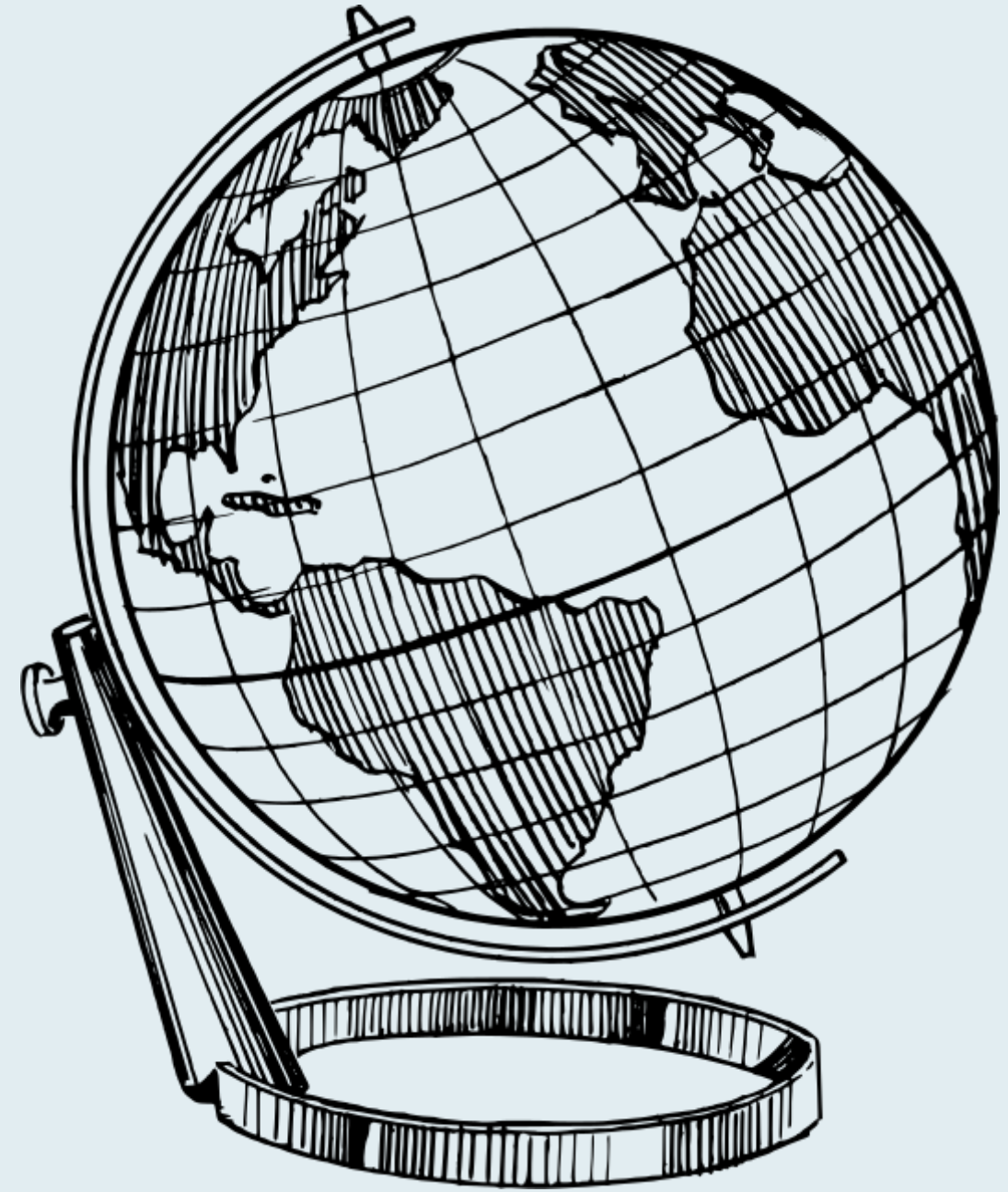
```
[<Element 'h1' class=('line-height-scale-5', 'font-scale-6', 'text-align-left',  
e', 'headline')>,
```




**ACCESS
DENIED**



Geolokalizacja



Problem captcha

XAEA-12



I'm not a robot



reCAPTCHA

**Brak danych na
stronie i
niestandardowe
sytuacje**



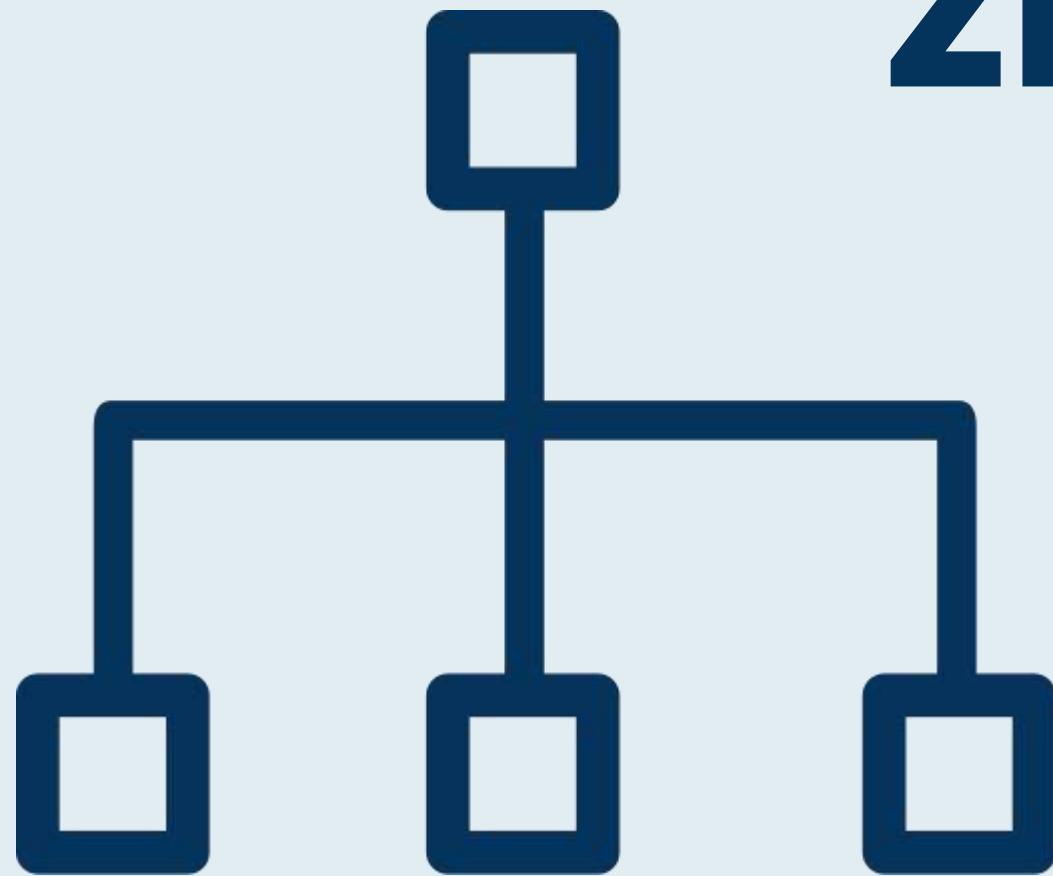

```
import re
```

```
text = "$525.55 sale price, $5.00 shipping"  
regex = re.compile(r'\$(\d[\d.,]*)')  
re.findall(regex, text)
```

```
['525.55', '5.00']
```


A co jeśli...

**zmieni się struktura
strony?**



ZASTOSOWANIA WEB SCRAPINGU



Machine Learning



Raporty i research



Pobieranie newsów/ogłoszeń



Data Driven Development

Dziękuję za uwagę! :)



KAMIL.KWAPISZ.PL



KONTAKT@KAMIL.KWAPISZ.PL



KAMIL.KWAPISZ.PL/FANPAGE



KAMIL.KWAPISZ.PL/GRUPA