pyladies Bratislava #10

Web scraping with requests and BeautifulSoup

Components of a web page

- HTML — contain the main content of the page.
- CSS — add styling to make the page look nicer.
- JS — Javascript files add interactivity to web pages.
- Images — image formats, such as JPG and PNG allow web pages to show pictures.

HTML

tags, properties, …

```
<a href="https://www.python.org"  class="library-link">Python</a>
```

- **child** — a child is a tag inside another tag
- **parent** — a parent is the tag another tag is inside
- **sibiling** — a sibiling is a tag that is nested inside the same parent as another tag

Requests library

- pip install requests
- https://realpython.com/python-requests/

```
>>> response = requests.get('https://weather.com/weather/today/l/48.15,17.11')

    <Response [200]>
```

Requests library

- Status codes

```
if response.status_code == 200:
    print('Success!')
elif response.status_code == 404:
    print('Not Found.')
```

- Content

```
>>> response.content
>>> response.text
>>> response.json()
```

Requests library

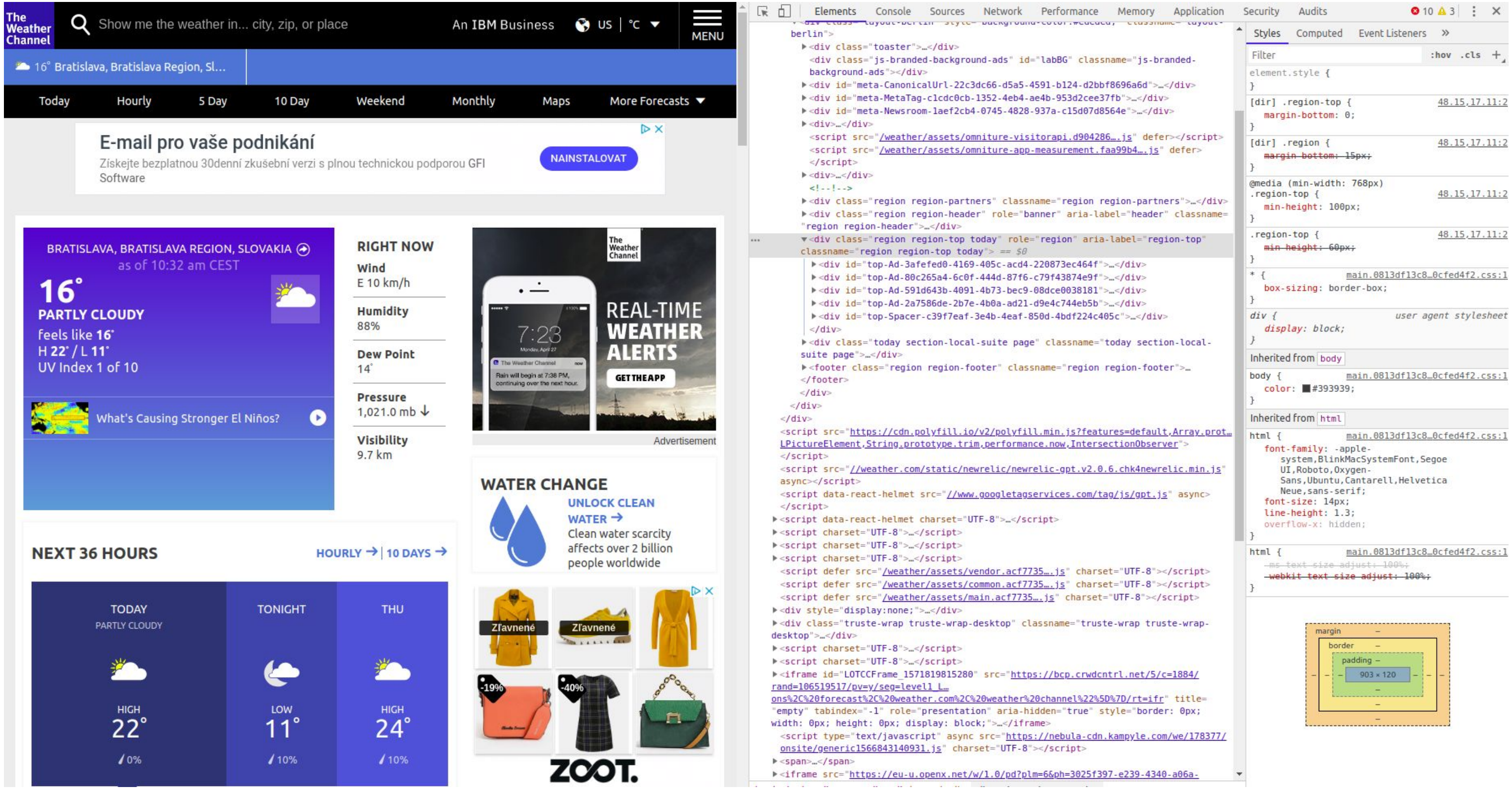
- Request headers and query parameters

```
response = requests.get(
    'https://api.github.com/search/repositories' ,
    params={'q': 'requests+language:python' },
    headers={'Accept': 'application/vnd.github.v3.text-match+json' },
 )
```

- Other methods

```
>>> requests.post('https://httpbin.org/post' , data={'key':'value'})
>>> requests.put('https://httpbin.org/put' , data={'key':'value'})
>>> requests.delete('https://httpbin.org/delete' )
>>> requests.head('https://httpbin.org/get' )
>>> requests.patch('https://httpbin.org/patch' , data={'key':'value'})
>>> requests.options('https://httpbin.org/get' )
```

Dev tools



E-mail pro vaše podnikání

Získejte bezplatnou 30denní zkušební verzi s plnou technickou podporou GFI Software

Clean water scarcity affects over 2 billion people worldwide

```
berlin">
  ▶<div class="toaster">…</div>
   <div class="js-branded-background-ads" id="labBG" classname="js-branded-
   background-ads"></div>
  ▶<div id="meta-CanonicalUrl-22c3dc66-d5a5-4591-b124-d2bbf8696a6d">…</div>
  ▶<div id="meta-MetaTag-c1cdc0cb-1352-4eb4-ae4b-953d2cee37fb">…</div>
  ▶<div id="meta-Newsroom-1aef2cb4-0745-4828-937a-c15d07d8564e">…</div>
  ▶<div>…</div>
   <script src="/weather/assets/omniture-visitorapi.d904286….js" defer></script>
   <script src="/weather/assets/omniture-app-measurement.faa99b4….js" defer>
   </script>
  ▶<div>…</div>
   <!--!-->
  ▶<div class="region region-partners" classname="region region-partners">…</div>
  ▶<div class="region region-header" role="banner" aria-label="header" classname=
  "region region-header">…</div>
… ▼<div class="region region-top today" role="region" aria-label="region-top"
  classname="region region-top today"> == $0
    ▶<div id="top-Ad-3afefed0-4169-405c-acd4-220873ec464f">…</div>
    ▶<div id="top-Ad-80c265a4-6c0f-444d-87f6-c79f43874e9f">…</div>
    ▶<div id="top-Ad-591d643b-4091-4b73-bec9-08dce0038181">…</div>
    ▶<div id="top-Ad-2a7586de-2b7e-4b0a-ad21-d9e4c744eb5b">…</div>
    ▶<div id="top-Spacer-c39f7eaf-3e4b-4eaf-850d-4bdf224c405c">…</div>
   </div>
  ▶<div class="today section-local-suite page" classname="today section-local-
  suite page">…</div>
  ▶<footer class="region region-footer" classname="region region-footer">…
  </footer>
  </div>
 </div>
</div>
<script src="https://cdn.polyfill.io/v2/polyfill.min.js?features=default,Array.prot…
LPictureElement,String.prototype.trim,performance.now,IntersectionObserver">
</script>
<script src="//weather.com/static/newrelic/newrelic-gpt.v2.0.6.chk4newrelic.min.js"
async></script>
<script data-react-helmet src="//www.googletagservices.com/tag/js/gpt.js" async>
</script>
▶<script data-react-helmet charset="UTF-8">…</script>
▶<script charset="UTF-8">…</script>
▶<script charset="UTF-8">…</script>
▶<script charset="UTF-8">…</script>
<script defer src="/weather/assets/vendor.acf7735….js" charset="UTF-8"></script>
<script defer src="/weather/assets/common.acf7735….js" charset="UTF-8"></script>
<script defer src="/weather/assets/main.acf7735….js" charset="UTF-8"></script>
▶<div style="display:none;">…</div>
▶<div class="truste-wrap truste-wrap-desktop" classname="truste-wrap truste-wrap-
desktop">…</div>
▶<script charset="UTF-8">…</script>
▶<script charset="UTF-8">…</script>
▶<iframe id="LOTCCFrame_1571819815280" src="https://bcp.crwdcntrl.net/5/c=1884/
rand=106519517/pv=y/seg=level1_L…
ons%2C%20forecast%2C%20weather.com%2C%20weather%20channel%22%5D%7D/rt=ifr" title=
"empty" tabindex="-1" role="presentation" aria-hidden="true" style="border: 0px;
width: 0px; height: 0px; display: block;">…</iframe>
 <script type="text/javascript" async src="https://nebula-cdn.kampyle.com/we/178377/
 onsite/generic1566843140931.js" charset="UTF-8"></script>
▶<span>…</span>
▶<iframe src="https://eu-u.openx.net/w/1.0/pd?plm=6&ph=3025f397-e239-4340-a06a-
```

```
element.style {
}
[dir] .region-top {                    48.15,17.11:2
  margin-bottom: 0;
}
[dir] .region {                        48.15,17.11:2
  ~~margin-bottom: 15px;~~
}
@media (min-width: 768px)
.region-top {                          48.15,17.11:2
  min-height: 100px;
}
.region-top {                          48.15,17.11:2
  ~~min-height: 60px;~~
}
* {                       main.0813df13c8…0cfed4f2.css:1
  box-sizing: border-box;
}
div {                           user agent stylesheet
  display: block;
}
Inherited from body
body {                    main.0813df13c8…0cfed4f2.css:1
  color: ■#393939;
}
Inherited from html
html {                    main.0813df13c8…0cfed4f2.css:1
  font-family: -apple-
    system,BlinkMacSystemFont,Segoe
    UI,Roboto,Oxygen-
    Sans,Ubuntu,Cantarell,Helvetica
    Neue,sans-serif;
  font-size: 14px;
  line-height: 1.3;
  overflow-x: hidden;
}
html {                    main.0813df13c8…0cfed4f2.css:1
  ~~-ms-text-size-adjust: 100%;~~
  ~~-webkit-text-size-adjust: 100%;~~
}
```

BeautifulSoup library

- pip install beautifulsoup4
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/

BeautifulSoup library

```
from bs4 import BeautifulSoup

soup = BeautifulSoup(response,
'html.parser')

print(soup.prettify())
```

BeautifulSoup library

```python
from bs4 import BeautifulSoup

soup = BeautifulSoup(response,
'html.parser')

print(soup.prettify())
```

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link2">
    Tillie
   </a>
   ; and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

BeautifulSoup library

```
soup.title
# <title>The Dormouse's story</title>
soup.title.name
# 'title'
soup.title.string
# 'The Dormouse's story'
soup.title.parent.name
# 'head'
soup.p
# <p class="title"><b>The Dormouse's story</b></p>
soup.p['class']
# 'title'
soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>
soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]
soup.find(id="link3")
 # <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
list(soup.p.children)
# [<b>The Dormouse's story</b>]
```

BeautifulSoup library

```
soup.find_all(name, attrs)
```

```
soup.find_all("a", class_="sister")
soup.find_all("a", attrs={"class": "sister"})
```

```
soup.find(id="link2")
```

BeautifulSoup library

CSS selectors

- **p a** — finds all a tags inside of a p tag
- **body p > a** — finds all a tags directly inside of a p tag inside of a body tag
- **p.outer-text** — finds all p tags with a class of outer-text
- **p#first** — finds all p tags with an id of first
- **body p.outer-text** — finds any p tags with a class of outer-text inside of a body tag

```
soup.select("div p")
```

BeautifulSoup library

extract text from the element

```
soup.select("div p")[0].get_text()
```

Send an email

https://www.afternerd.com/blog/how-to-send-an-email-using-python-and-smtplib/