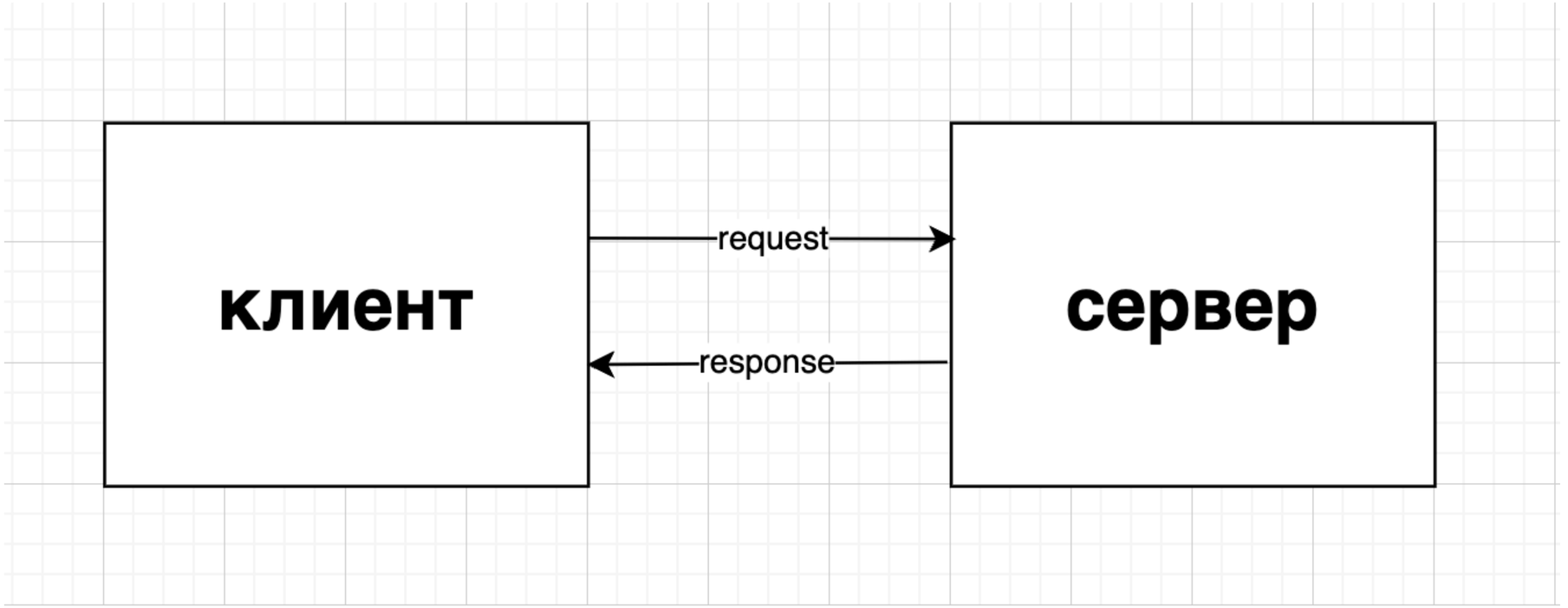
ЛЕКЦИЯ 5

Веб-скрейпинг. Библиотеки requests, BeautifulSoup. Парсинг html файлов

- HTTP запросы
- Типы ответов
- Библиотека requests
- Библиотека BeautifulSoup

HTTP – это протокол для получения информации. Он является основой любого обмена данными в Интернете и представляет собой клиент-серверный протокол, что означает, что запросы инициируются получателем, например веб-браузером.

HTTP ЗАПРОСЫ



GET

http://url/get-request/?text=get+request

POST

path : http://url/post-request

body : {«text» : «post request»}

ТИПЫ ОТВЕТА

- XML
- HTML
- JSON

https://cbr.ru/scripts/XML_daily.asp?date_req=02/03/2002

https://requests.readthedocs.io/en/latest/index.html

https://www.boredapi.com/api/activity

Модуль Requests позволяет отправлять HTTP/1.1 запросы очень легко. Вам не нужно вручную добавлять строки запросов к вашим URL-адресам или кодировать данные POST.

```
pip install requests
```

2XX (успешные ответы)

4XX (ошибка на стороне клиента)

```
import requests

response=requests.get('
https://ya.ru/')
print(response)
```

3XX (перенаправление)

5XX (ошибка на стороне сервера)

БИБЛИОТЕКА REQUESTS. ПРИМЕР ИСПОЛЬЗОВАНИЯ. JSON

```
import requests

response=requests.get('https://
www.boredapi.com/api/activity').json()
print(response)
```

```
import requests

response=requests.get('https://
requests.readthedocs.io/en/latest/index.html
').text
print(response)
```

BeautifulSoup – это библиотека на Python для извлечения данных из HTML и XML файлов. Она работает с вашим любимым парсером, предоставляя идиоматические способы навигации, поиска и модификации дерева разбора. Обычно она экономит программистам часы или дни работы.

```
pip install beautifulsoup4
```

БИБЛИОТЕКА BEAUTIFULSOUP. ПРИМЕР РАБОТЫ

```
import requests
from bs4 import BeautifulSoup

url="https://mfd.ru/currency/?currency=USD"
r = requests.get(url)


soup = BeautifulSoup(r.text, 'html.parser')
data = soup.find('table', {'class': 'mfd-table mfd-currency-
table'}).find_all('td')

print(data)
```

БИБЛИОТЕКА BEAUTIFULSOUP. ПРИМЕР РАБОТЫ. РЕЗУЛЬТАТ

```
[<td>с 19.01.2024</td>, <td>88.6610</td>, <td><span class="mfd-u">+0.307</span></td>,
<td>с 18.01.2024</td>, <td>88.3540</td>, <td><span class="mfd-u">+0.7083</span></td>,
<td>с 17.01.2024</td>, <td>87.6457</td>, <td><span class="mfd-d">-0.0315</span></td>,
<td>с 16.01.2024</td>, <td>87.6772</td>, <td><span class="mfd-d">-0.4552</span></td>,
<td>с 13.01.2024</td>, <td>88.1324</td>, <td><span class="mfd-d">-0.6494</span></td>,
<td>с 12.01.2024</td>, <td>88.7818</td>, <td><span class="mfd-d">-1.6222</span></td>,
<td>с 10.01.2024</td>, <td>90.4040</td>, <td><span class="mfd-u">+0.7157</span></td>,
<td>с 30.12.2023</td>, <td>89.6883</td>, <td><span class="mfd-d">-0.6158</span></td>,
<td>с 29.12.2023</td>, <td>90.3041</td>, <td><span class="mfd-d">-1.4028</span></td>,
<td>с 27.12.2023</td>, <td>91.7069</td>, <td><span class="mfd-d">-0.2621</span></td>,
<td>с 26.12.2023</td>, <td>91.9690</td>, <td><span class="mfd-u">+0.0301</span></td>,
<td>с 23.12.2023</td>, <td>91.9389</td>, <td><span class="mfd-u">+0.2327</span></td>,
<td>с 22.12.2023</td>, <td>91.7062</td>, <td><span class="mfd-u">+1.3006</span></td>,
<td>с 21.12.2023</td>, <td>90.4056</td>, <td><span class="mfd-u">+0.3186</span></td>,
```

Как «очистить» данные и сделать их пригодными для дальнейшего использования ?

КОНТРОЛЬНЫЕ ВОПРОСЫ