

# Парсинг HTML. BeautifulSoup.

Урок 2



# План курса

1

Основы клиент-серверного взаимодействия. Парсинг API.

2

Парсинг HTML. BeautifulSoup.

3

СУБД MongoDB и ClickHouse в Python

4

Парсинг HTML. XPath.

5

Scrapy.

6

Scrapy. Парсинг фото и файлов.

7

Selenium в Python.

8

Работа с данными.

9

Инструменты разметки наборов данных.



## Что будет на уроке сегодня

- 📌 Как работает веб-скрейпинг?
- 📌 Законен ли веб-скрейпинг?
- 📌 Введение в BeautifulSoup.
- 📌 Парсинг HTML с помощью BeautifulSoup
- 📌 Скрейпинг веб-страницы





Вопрос

Какую информацию для анализа  
можно получить с веб-страниц?



# Пример сайта, содержащего информацию

**apteka.de**

На латинице: название, PZN или дейст. вещество

Войти

Все категории ▾ Марки ▾ Косметика ▾ Семья ▾ Фитнес ▾ Пищевые добавки ▾ Оздоровление ▾

### Лучшие категории


**Фитнес**  
Похудение и диета  
спорт

**Почки, мочевой пузырь и простата**  
Здоровый мочевой пузырь  
Мочекаменная болезнь

**Укрепление здоровья и большая концентрация**  
Энергия и душевные силы






## Сильная помощь при заболеваниях желудка и кишечника.

Сохранить сейчас »



The image shows two products: a white bag of Bekunis (Biotin and Folic Acid) and a yellow box of FAKTU (Folic Acid). The Bekunis bag is labeled 'Bekunis' and 'FÜR DIE DARMGESUNDHEIT'. The FAKTU box is labeled 'FAKTU lind' and 'Folsäure'. There is a blue arrow pointing from the left sidebar towards the products.

## Пример сайта, содержащего информацию

Сортировка: Актуальность ▾		 	
	-39 % <sup>1</sup>		-52 % <sup>1</sup>
ALLERGO-MOMELIND von DoppelherzPharma		MometaxHEXAL Heuschnupfenspray	
18 g		18 g	
<del>13,45 €</del>		<del>22,24 €</del>	
450,56 € / 1 kg		588,33 € / 1 kg	
<b>8,11 €</b>		<b>10,59 €</b>	
			
		-50 % <sup>1</sup>	
		Lorano akut Tabletten	
		20 St	
		<del>10,02 €</del>	
		<b>4,99 €</b>	

# Пример сайта, содержащего информацию

1 аллергический ринит > ALLERGO-MOMELIND von DoppelherzPharma



-39 %<sup>1</sup>

## ALLERGO-MOMELIND von DoppelherzPharma 18 g

Zur symptomatischen  
Behandlung der Beschwerden  
eines Heuschnupfens. Für  
Erwachsene.

Форма выпуска: Nasendosierspray

Содержание: 18 g

Базовый тариф: 450,56 € / 1 kg

Фармацевтический номер:


16665753

Производитель: Queisser Pharma  
GmbH & Co. KG

**8,11 €** вместо 13,45 € UVP

Вкл. НДС" при необходимости включая  
[стоимость доставки](#)

● ● ● в наличии

 [1008248087](#)

★ [Оценки](#)



Как работает веб-скрейпинг?







## Последовательность операций при веб-скрейпинге

- 💡 Запросить содержимое (HTML-код) определенного URL с сервера
- 💡 Загрузить полученное содержимое
- 💡 Определить элементы страницы, содержащие нужную нам информацию
- 💡 Извлечь и при необходимости выполнить парсинг элементов страницы в датасет



## Правовые нормы, касающиеся веб-скрейпинга

1. Правила о гражданско-правовой ответственности и о причинении вреда имуществу.
2. Уголовная ответственность за преступления в сфере компьютерной информации.
3. Нормы договорного права.
4. Право интеллектуальной собственности.
5. Правила о персональных данных.



## Выдержка из Пользовательского соглашения Авито

Без согласия Компании запрещено использовать технические средства для взаимодействия с сервисом в обход обычного порядка использования баз данных и программ для ЭВМ. В том числе запрещено использовать автоматизированные скрипты для сбора информации на Авито, а также для автоматической регистрации профилей.



## Пример файла `www.example.com/robots.txt`

```
User-agent: Googlebot  
Disallow: /nogooglebot/
```

```
User-agent: *  
Allow: /
```

```
Sitemap: http://www.example.com/sitemap.xml
```



## Правила скрейпинга

- 💡 никогда не делайте скрейп чаще, чем это необходимо.
- 💡 рассмотрите возможность кэширования содержимого, которое вы скрейпируете, чтобы оно загружалось только один раз.
- 💡 встраивайте в код паузы (для этого есть функция `time.sleep()`), чтобы не перегружать сервер слишком большим количеством запросов.



# Введение в Beautiful Soup





## Особенности BeautifulSoup

- 💡 Парсинг документов HTML и XML.
- 💡 Навигация по парсинг-дереву (parse tree).
- 💡 Доступ к тегам и атрибутам.
- 💡 Изменение HTML-кода.
- 💡 Поддержка Unicode.



## Parsing tree

— это структура данных, которая представляет синтаксическую структуру строки текста в соответствии с правилами формальной грамматики.







## Parsing tree

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example</title>
  </head>
  <body>
    <h1>Welcome to my website!</h1>
    <p>This is a paragraph of text.</p>
  </body>
</html>
```



## Parsing tree

```
from bs4 import BeautifulSoup
import requests

url = "https://example.com"
response = requests.get(url)
html = response.content

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

print(soup.title.string)
```



# Методы BeautifulSoup



## **find()**

```
a = soup.find('a')
```



## **find\_all()**



## **Доступ к атрибутам.**

```
link = soup.find('a')
```

```
href = link['href']
```



## **get()**

```
link = soup.find('a')
```

```
href = link.get('href', None)
```



## **string и text**

```
link = soup.find('a')
```

```
text = link.text
```



# Методы BeautifulSoup



## **contents**

```
html = soup.find('html')
```

```
children = html.contents
```



## **descendants**

```
html = soup.find('html')
```

```
descendants = html.descendants
```



## **parent и parents**

```
link = soup.find('a')
```

```
href = link['href']
```



## **next\_sibling и previous\_sibling**



## **find\_all() и find()**

```
div = soup.find('div')
```

```
links = div.find_all('a')
```



```
<div>
  <p>This is a <b>bold</b> statement.</p>
  <ul>
    <li>Item 1</li>
    <li>Item 2</li>
  </ul>
</div>
```