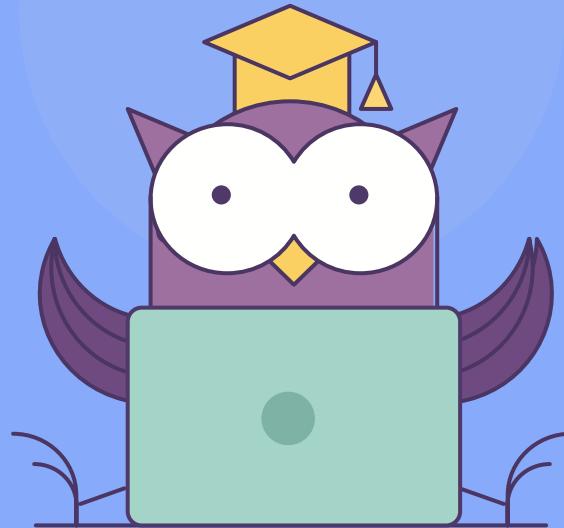


O T U S

ОНЛАЙН-ОБРАЗОВАНИЕ

# Меня хорошо слышно && видно?



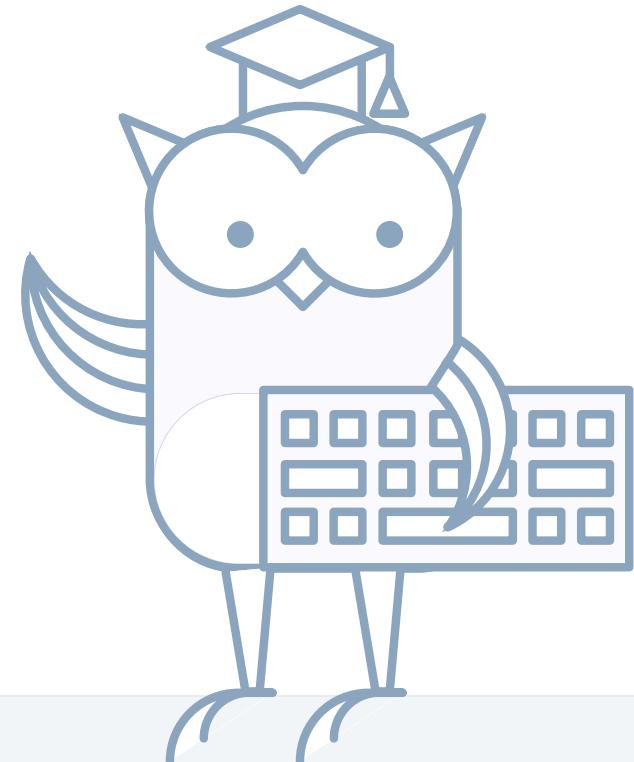
Напишите в чат, если есть проблемы!

Ставьте + если все хорошо



# Сбор данных: грязная работа своими руками

Учимся парсить сайты с Python



- Научитесь создавать собственные датасеты
- Освоите библиотеки, необходимые для парсинга сайтов
- Поймёте принцип работы парсеров

- Зачем парсить?

- Зачем парсить?
- Чем парсить?

- Зачем парсить?
- Чем парсить?
- Как парсить?

01

Зачем парсить?

- У вас есть идея и есть данные, но их мало?

- У вас есть идея и есть данные, но их мало?
- У вас есть идея, но нет данных?

- У вас есть идея и есть данные, но их мало?
- У вас есть идея, но нет данных?
- У вас нет идеи, и нет данных?

# Зачем парсить?

O T U S

The screenshot shows a software interface with several export options:

- Export:** EXCEL, CSV
- PDF**
- CSV**
- CSV** (highlighted with a red box)
- Download** (highlighted with a red arrow)
- Choose report**
- CSV** (highlighted with a red box)

A central modal window displays a report table:

Category	Total	Owed	Booked
Not Canceled	0 0 0	2/23/2010	2/23/2010
Any Payment	0 0 0	2/23/2010	2/23/2010

The "Export" button in the modal is also highlighted with a red box.

At the bottom right of the modal, there are buttons: **Cancel**, **Export to CSV** (highlighted with a red box), and **Print**.

A large blue button at the bottom center says **Download CSV**.

02

Чем парсить?

# Чем парсить?

O T U S

- Конечно же питоном



- И еще парочкой библиотек



Requests



BeautifulSoup

03

Как парсить?

```
▼<div id="main" class="main">
  ▶<div class="advert hidden-mobile hidden-compact hidden-narrow">...</div>
  <div class="advert advert--m_main_top hidden-compact hidden-narrow hidden-wide"></div>
  ▶<div id class="header header--normal js-header">...</div>
  ▼<div class="container">
    ▼<header class="post-header post-article_header">
      ▼<div class="post-meta post-header_meta">
        <a href="/mags/photo" class="post-meta_magazine magazine">Фотожурнал</a>
        <a href="/tags/310" class="post-meta_item post-meta_tag">
          Искусство фотографии
        </a>
      ▶<span class="post-meta_item post-meta_published-at">...</span>
      </div>
    ▼<h1 class="post-title post-header_title post-article_title article-feature-title">
      "
      Правда или ложь?
      <span class="thin">Насколько правдива документальная фотография?</span> == $0
    </h1>
    ▼<p class="post-excerpt post-header_excerpt post-article_excerpt">
      "
      Принято считать, что документальный снимок передает реальность достоверно, но на
      "
    </p>
  ▶<div class="post-authors post-header_authors post-article_authors">...</div>
</header>
```

# HTML

HyperText Markup Language (*HTML*)

*is the standard markup language for creating web pages and web applications*

# HTML

It's all about tags

- The `<a>` tag is used for creating an `a` element (also known as an "anchor" element). The `a` element represents a [hyperlink](#). This is usually a link to another document.
- The `<h1>...<h6>` tags represent a level 1...6 headings in an HTML document.
- The `<span>` tag represents its children for the purposes of applying global attributes.
- The `<div>` tag defines a division or a section in an HTML document
- And so on, and so on...

- Открываем сайт с интересующими нас данными

- Открываем сайт с интересующими нас данными
- Находим интересующий нас элемент и переходим в браузере в режим “inspect element code”

- Открываем сайт с интересующими нас данными
- Находим интересующий нас элемент и переходим в браузере в режим “inspect element code”
- Находим соответствующие тэги, окружающие наш элемент

- Открываем сайт с интересующими нас данными
- Находим интересующий нас элемент и переходим в браузере в режим “inspect element code”
- Находим соответствующие тэги, окружающие наш элемент
- Дальше суп всё сделает сам

Переходим к практике!

- Стало ли понятнее, зачем нужно парсить и что это вообще означает?

- Стало ли понятнее, зачем нужно парсить и что это вообще означает?
- Есть ощущение, что это просто/сложно? Получилось бы повторить пройденные шаги?

- Стало ли понятнее, зачем нужно парсить и что это вообще означает?
- Есть ощущение, что это просто/сложно? Получилось бы повторить пройденные шаги?
- Самое главное – захотелось ли что-то где-то собрать? :)



# Сергеев Дмитрий

Sergeyev.D.A@yandex.ru

+7 906 779 76 97

Спасибо  
за внимание!

