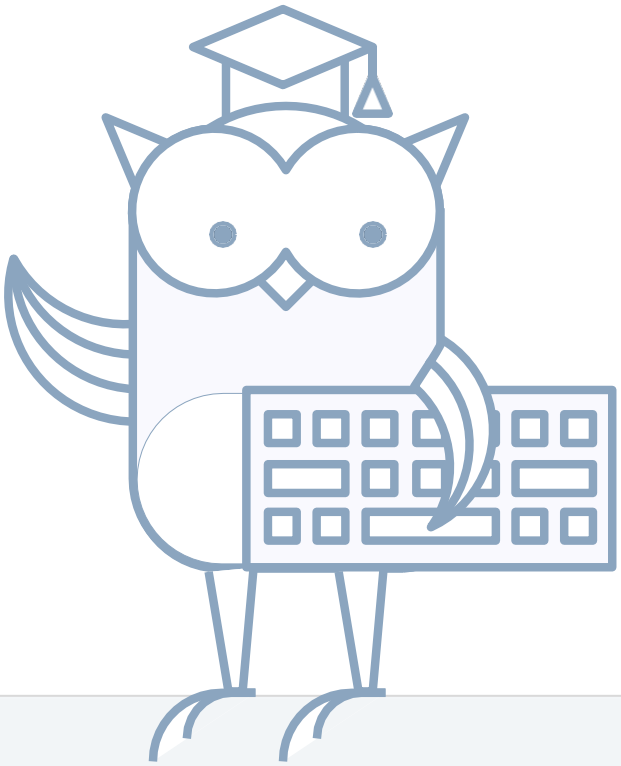
OTUS

Сбор данных: грязная работа своими руками

Учимся парсить сайты с Python



Мария Тихонова

Цель занятия

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

Содержание

- Зачем парсить?

Содержание

- Зачем парсить?
- Чем парсить?

Содержание

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

01

Зачем парсить?

Зачем парсить?

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

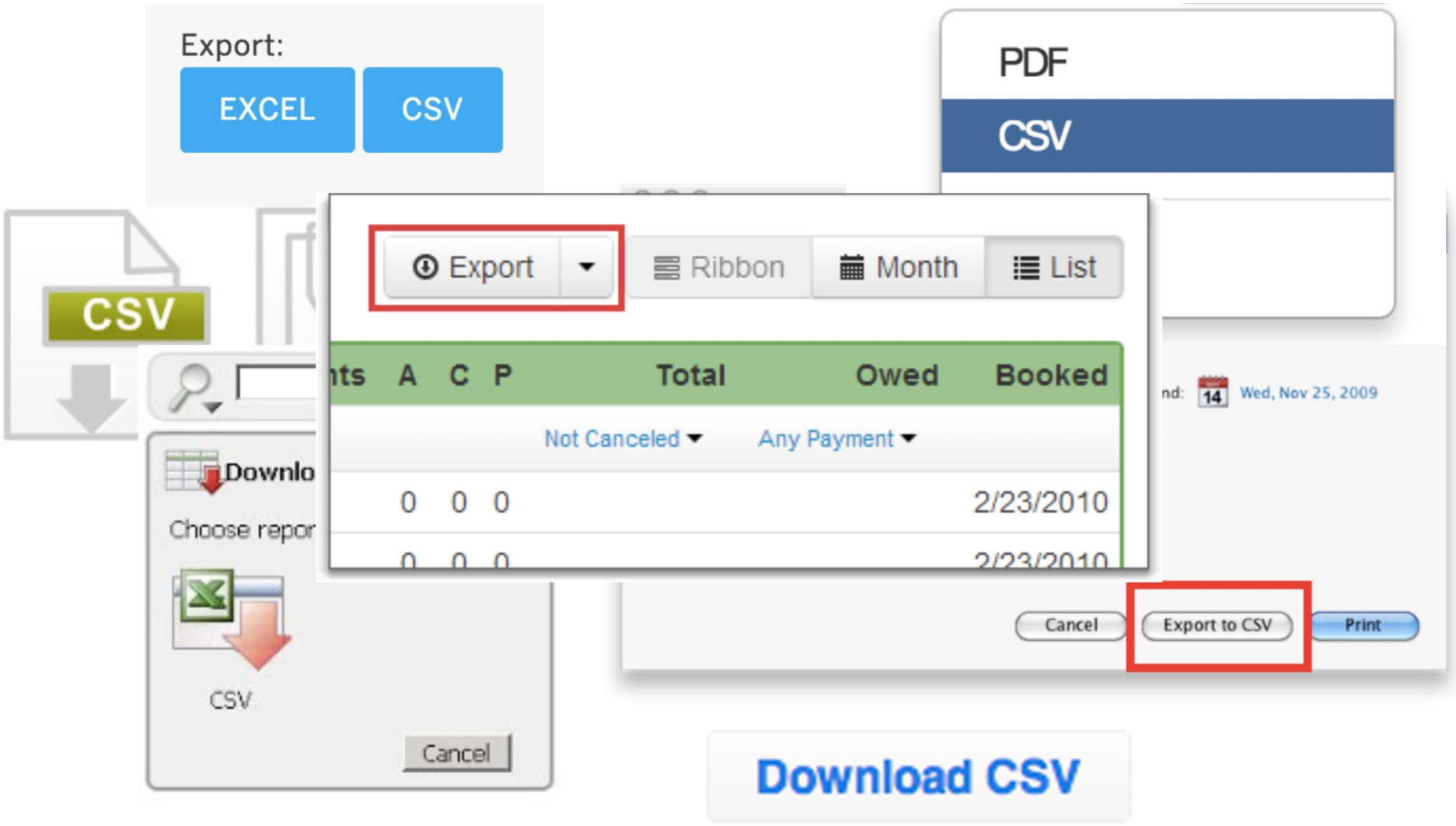
Зачем парсить?

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

Зачем парсить?

- 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?



02

Чем парсить?

Чем парсить?

- Конечно же питоном

Чем парсить?

- И еще парочкой библиотек

Requests

BeautifulSoup

03

Как парсить?

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 tag represents its children for the purposes of applying global attributes.
- The tag defines a division or a section in an HTML document
- And so on, and so on…

Как парсить?

- Открываем сайт с интересующими нас данными

Как парсить?

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

Как парсить?

- 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

Как парсить?

- 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

Практика

Переходим к практике!

Оценка результатов

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

Оценка результатов

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

- Есть ощущение, что это просто/сложно? Получилось бы повторить пройденные шаги?

Оценка результатов

- 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? :)

Мария Тихонова
Tg: @mashkka_ds

Спасибо
за внимание!