

Parsing web Data with Python: Grab what you can

Dmitry Sergeyev

Why?

Why?

The screenshot displays a software application window with several export-related components:

- Top Right:** A dark blue button labeled "Export ▾".
- Left Sidebar:** A light gray panel titled "Export:" containing two blue buttons: "EXCEL" and "CSV".
- Middle Panel:** A modal dialog box with a red border. It features a toolbar with "Export" (highlighted with a red box), "Ribbon", "Month", and "List" buttons. Below the toolbar is a table with columns: "Total", "Owed", and "Booked". The table contains two rows of data:

Total	Owed	Booked
Not Canceled ▾	Any Payment ▾	2/23/2010
0 0 0		2/23/2010
- Bottom Right:** A large blue button labeled "Download CSV".
- Bottom Left:** A "Cancel" button.
- Bottom Center:** Three buttons: "Cancel", "Export to CSV" (highlighted with a red box), and "Print".
- Background:** A sidebar on the left with icons for "CSV" (with a green arrow pointing down), "Download" (with a red arrow pointing down), "Choose report" (with a red arrow pointing down), and "CSV" (with a red arrow pointing down). A date "Wed, Nov 25, 2009" is also visible.

Why?

Cause we can

How?

How?

Of course, Python



How?

And some libraries



Requests



BeautifulSoup

Where?

HTML

```
<section class="showcase news">
<div class="container">
    <h2>
        НОВОСТИ
        <a class="rss-button" href="/export/fast.xml" target="_blank">
            <i class="icon-svg7"></i>
        </a>
    </h2>
</div>
<section class="showcase-container">
    <section class="showcase-lines">
        <div class="date-col">
            <div class="date">
                <span class="date-content-day">
                    15
                </span>
                <span class="date-content-month">
                    октября
                </span>
            </div>
        </div>
        <div class="news-col">
            <article class="news-line">
                <a href="/posts/74934">
                    <div class="line-content">
                        <div class="line-post-label">
                            <h6 class="line-post-headline ">
                                Россия иnbsp;Индия подписали контракт наnbsp;поставку С-400
                            </h6>
                        </div>
                    </a>
                </article>
            <section class="showcase-row">

```

11:37

</div>

</h6>

HTML

HyperText Markup Language (*HTML*)

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

HTML

It's all about tags

TOP TAGS ON INSTAGRAM



HTML

It's all about tags

HTML5 Tags

The following tags are supported in [HTML5](#) (and/or the WHATWG HTML Living Standard).

<!--...-->

<!DOCTYPE>

<a>

<abbr>

<address>

<area>

<article> (NEW)

<aside> (NEW)

<form>

<h1>

<h2>

<h3>

<h4>

<h5>

<h6>

<head>

<pre>

<progress> (NEW)

<q>

<rb> (NEW)

<rp> (NEW)

<rt> (NEW)

<rtc> (NEW)

<ruby> (NEW)

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 `<div>` tag defines a division or a section in an HTML document
- And so on, and so on...

HTML

How can we use it

HTML

How can we use it

- Open the site with the data you need

HTML

How can we use it

- Open the site with the data you need
- Point on the element and go to “inspect element code” mode in your browser

HTML

How can we use it

- Open the site with the data you need
- Point on the element and go to “inspect element code” mode in your browser
- Find the corresponding tags surrounding your data

HTML

How can we use it

- Open the site with the data you need
- Point on the element and go to “inspect element code” mode in your browser
- Find the corresponding tags surrounding your data
- Let the Soup do the job

Lets get down to practice

Lets get down to practice

