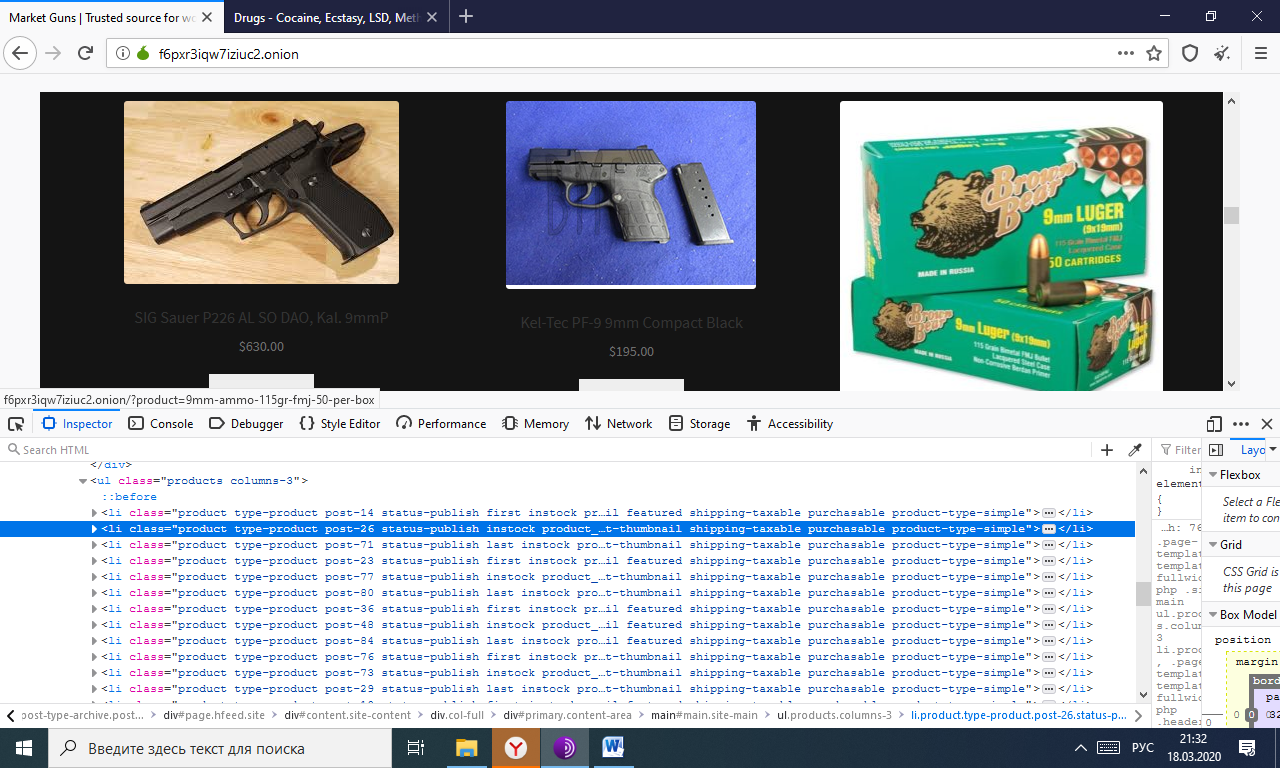
1 сайт(http://f6pxr3iqw7iziuc2.onion/):

Каталог в виде списка (тэг <ul>, класс “products columns-3”)

Товары под тэгом <li>, классы данных тэгов практически одинаковые, но тем не менее различаются:



Основная информация о товаре содержится под тэгом <a>(класс: “woocommerce-LoopProduct-link woocommerce-loop-product\_\_link”):

Наименование – тэг <h2>, класс “woocommerce-loop-product\_\_title”

Цена – тэг <span>, класс “price”

URL – атрибут тэга <a> href, т е прикреплён к первому объединяющему тэгу <a>

*<a href="http://f6pxr3iqw7iziuc2.onion/?product=kel-tec-pf-9-9mm-compact-black" class="woocommerce-LoopProduct-link woocommerce-loop-product\_\_link">*

*<img src="http://f6pxr3iqw7iziuc2.onion/wp-content/uploads/2018/08/Kel-Tec-PF-9-9mm-Compact-Black.jpg" class="attachment-woocommerce\_thumbnail size-woocommerce\_thumbnail" alt="Kel-Tec PF-9 9mm Compact Black" width="250" height="188">*

*<h2 class="woocommerce-loop-product\_\_title">Kel-Tec PF-9 9mm Compact Black</h2>*

*<span class="price">*

*<span class="woocommerce-Price-amount amount">*

*<span class="woocommerce-Price-currencySymbol">$</span>195.00</span></span>*

*</a>*

2 сайт(http://wzvsz3g6dodj2fyh.onion/):

Каталог под тэгом <div>, класс “features\_items”

Товары под тэгом <div>, класс “col-sm-4 ”

Основная информация о товаре содержится под тэгом под тэгом <div> (класс “productinfo text-center”):

Наименование – тэг <p> (внутри второго <a>)

Цена – тэг <h2> (внутри <div>)

URL – атрибут тэга <a> href, т е фактически он прикреплён на название товара

<div class="productinfo text-center">

<a href="/product/88"><img src="/upload/images/products/88.jpg" alt=""></a>

<h2>$ 39.00</h2>

<a href="/product/88">

<p>Pure Love - Cannabis 20g</p>

</a>

<a href="/cart/add/88" class="btn btn-default add-to-cart"><i class="fa fa-shopping-cart"></i>Add to cart</a>

</div>

3 сайт(http://elite6c3wh756biv7v2fyhnoitizvl2gmoisq7xgmp2b2c5ryicottyd.onion/index.php?cid=2):

Каталог находится под тэгом <td> (тэг, предназначенный для создания одной ячейки таблицы, что странно).

Товары представлены в виде таблиц и соответственно находятся под тэгами <table> (класс “offer-vertical”)

Основная информация о товаре находится под тэгом <td> (класс “offer-vertical-td-right”). Внутри этого тэга мы можем снова увидеть, конечно же, парочку таблиц (:

Наименование – тэг <a> (внутри <table> с классом “offer-vertical-title-table”)

Цена – общий тэг <span> (без класса), внутри два тэга <font>: с классами <green>(USD) и <red> (криптовалюта)

URL – атрибут тэга <a> (внутри <table> с классом “offer-vertical-title-table”) href, т е фактически он прикреплён на название товара.

4,5,6 сайт: (<http://bitpharplh2t5fxm.onion/>, http://tuu66yxvrnn3of7l.onion/, http://ll6lardicrvrljvq.onion/)

Сайты, состоящие из одной страницы с одной или несколькими табличками товаров.

Итак каталог объединен в таблицу (тэг <table>, “ table1”), где товары содержатся внутри тэга <tbody>.

Основная информация о товаре находится под тэгом <tr> (строка таблицы).

Наименование –первый тэг <td>

Цена – второй тэг <td>

Порядок тэгов td важен. Ссылка на товар отсутствует.

7 сайт (<http://luckp47s6xhz26rn.onion/>):

Каталог находится под тэгом <table> (с атрибутом id)

Основная информация о товаре находится под тэгом <td> (класс “figure”):

Наименование и цена находятся внутри тэга <p> в виде текста

<td class="figure" style="width:300px;height:200px">

<p style="color:white;font-size:1.1em;text-decoration:none">

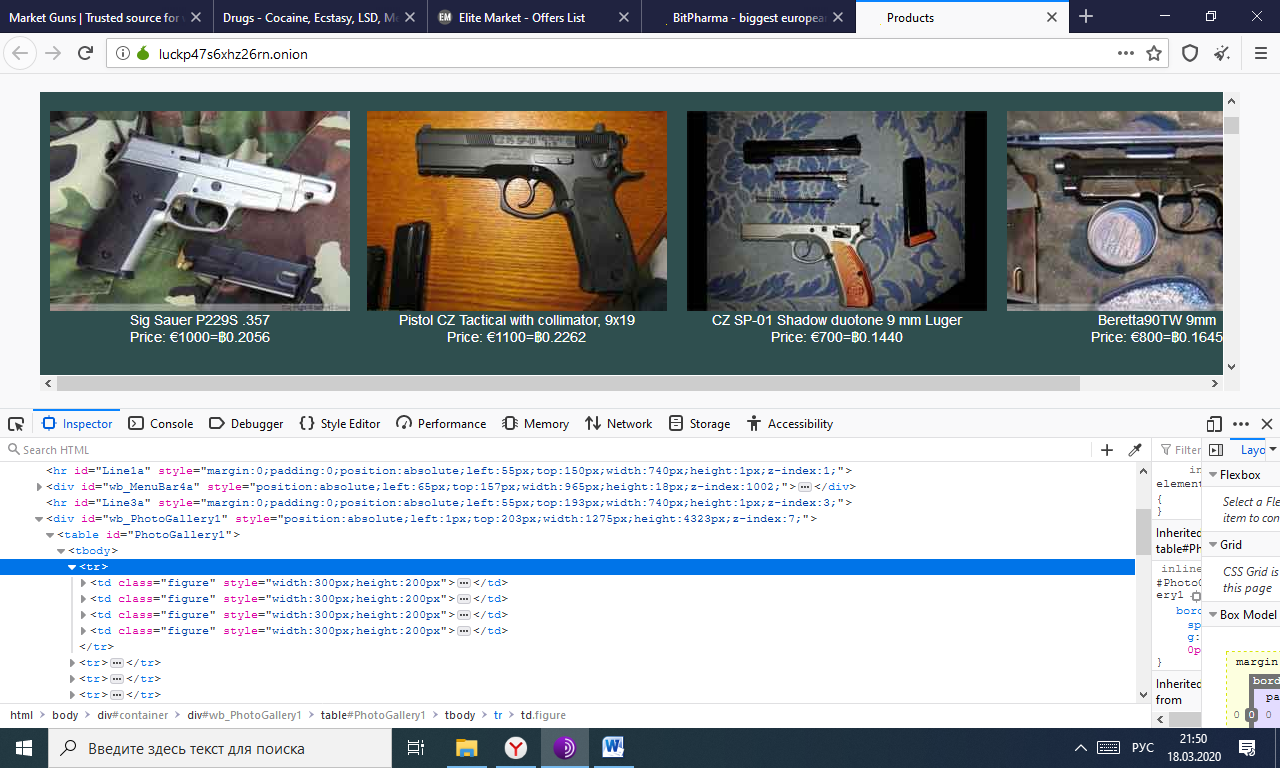
<img alt="" src="images/1.jpg" style="width:300px;" align="bottom">

Sig Sauer P229S .357

<br>

Price: €1000=฿0.2062</p>

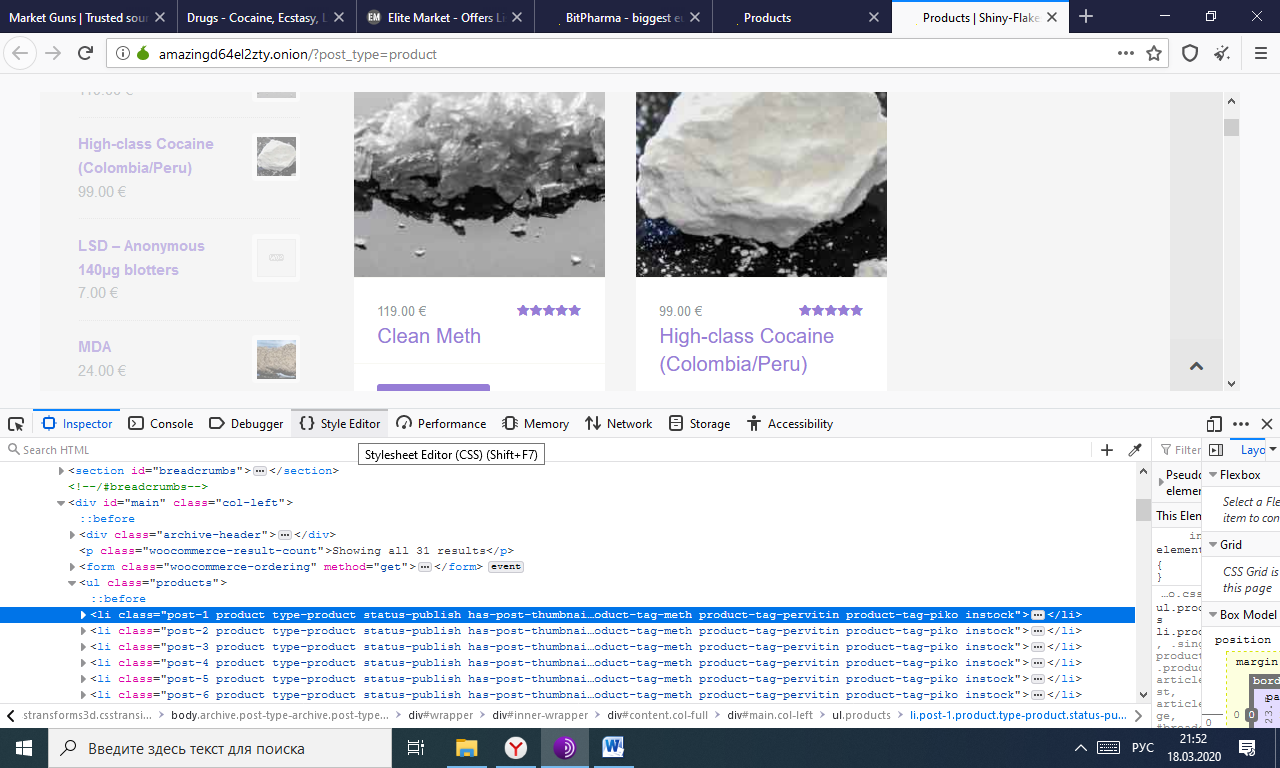
</td>



8 сайт (<http://amazingd64el2zty.onion/?post_type=product>):

Каталог находится под тэгом <ul>, класс “products” (представляет собой список)

Информация о каждом товаре находится под тэгами <li>, внутри которого содержится два тэга <a>

(классы тэгов незначительно различаются)

Наименование – тэг <h3>

Цена – тэг <span>, класс “price”

URL – атрибут тэга <a> href (тэг <a> родительский для цены и названия)

<li class="post-1 product type-product status-publish has-post-thumbnail first shipping-taxable purchasable product-type-simple product-tag-crystal product-tag-glass product-tag-ice product-tag-meth product-tag-pervitin product-tag-piko instock">

<a href="http://amazingd64el2zty.onion/?product=shiny-flakes-sauberes-meth">

<img src="http://amazingd64el2zty.onion/wp-content/uploads/2013/12/11- 250x250.jpg" class="attachment-shop\_catalog wp-post-image" alt="sauberes\_meth" width="250" height="250">

<div class="star-rating"><span style="width:100%"><strong class="rating">5.00</strong> out of 5</span></div>

<span class="price">

<span class="amount">

119.00&nbsp;€

</span>

</span>

<h3>Clean Meth</h3>

</a>

<a href="/?post\_type=product&amp;add-to-cart=1" rel="nofollow" data-product\_id="1" data-product\_sku="" data-quantity="1" class="button add\_to\_cart\_button product\_type\_simple">Add to cart</a>

</li>

9 сайт(<http://gunstry2lzpvf47i.onion/>)

Каталог товаров содержится под тэгом <div> (класс “container”)

Каждый отдельный товар внутри под тэгом <div>, класс “col-md-3 product”

Далее в виде сплошного текста под тэгами <p> можно найти наименование, цену.

<div class="col-md-3 product">

<img src="images/uzi.jpg" alt="">

<br>

<hr>

<br>

<p><b>Uzi </b></p>

<p>Caliber : 9mm</p>

<p>Weight : 33OZ</p>

<p>Capacity : 15 rounds</p>

<br>

<p>Price : 800$ (0.1538952 BTC)</p>

<form action="payment.php" method="post">

<input type="number" name="quantity" class="input" placeholder="Quantity" required="">

<input type="hidden" value="800" name="price">

<input type="hidden" value="Uzi" name="name">

<input type="submit" name="order" value="Order" class="input">

</form>

</div>

10 сайт (<http://wacky2yx73r2bjys.onion/>)

Каталог представлен списком. Тэг <ul>, класс “wc-block-grid\_\_products”

Товары находятся под тэгами <li> (элементы списка), класс “ wc-block-grid\_\_products”

Название находится в картинке):

URL – тэг <a>(класс “wc-block-grid\_\_product-link”) атрибут href

Цена – внутри одного из тэгов <span> (класс “woocommerce-Price-currencySymbol”)

Закономерности

Итак основной отличительной чертой списка товаров является содержание однотипных элементов (совпадающие классы, тэги).

Несмотря на некоторые исключения, пожалуй, самый эффективный способ найти информацию о товаре – это найти тэг с его ценой (знаки USD, $, EUR, €, ฿ в помощь).

Далее проще подниматься вверх по структуре сайта от найденного тэга и пытаться выйти в каталог товаров.

Для этого мы должны проверять дочерние элементы каждого родителя нашего тэга (являются ли они однотипными).

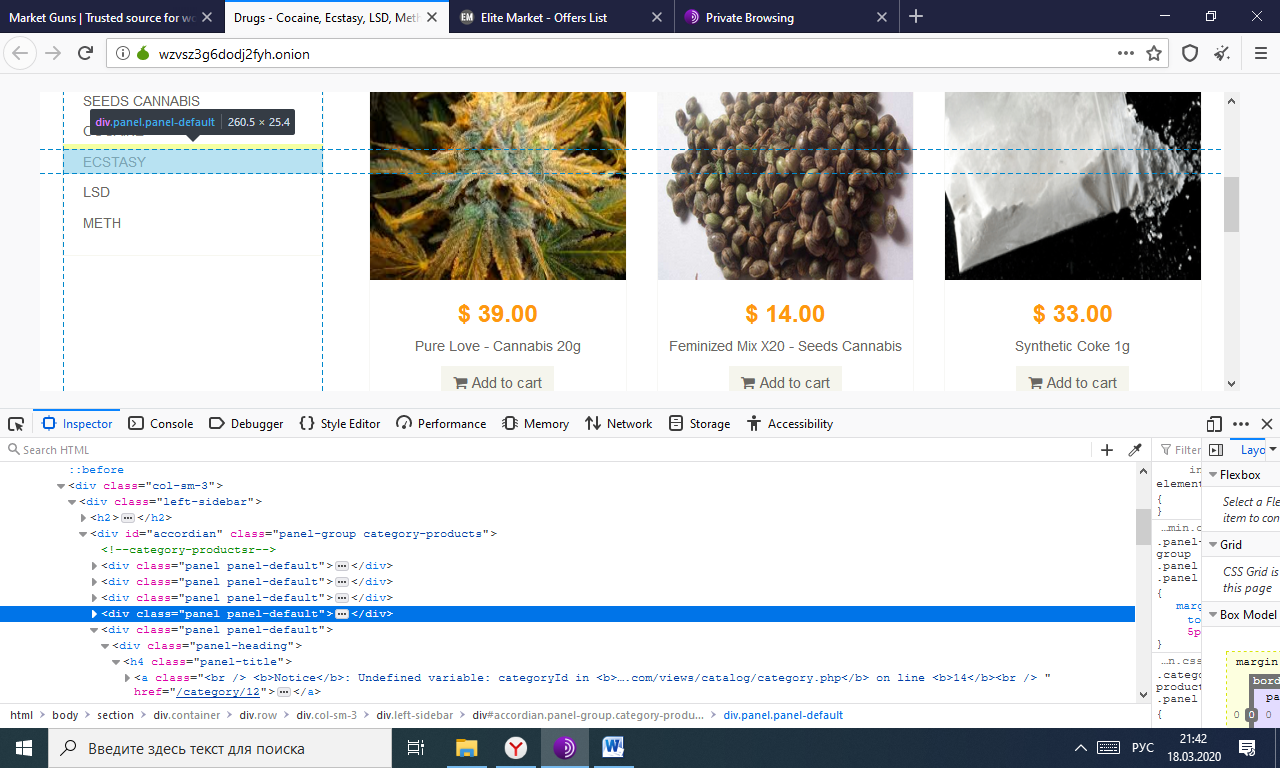
Если да, то далее заходим в каждый и ищем:

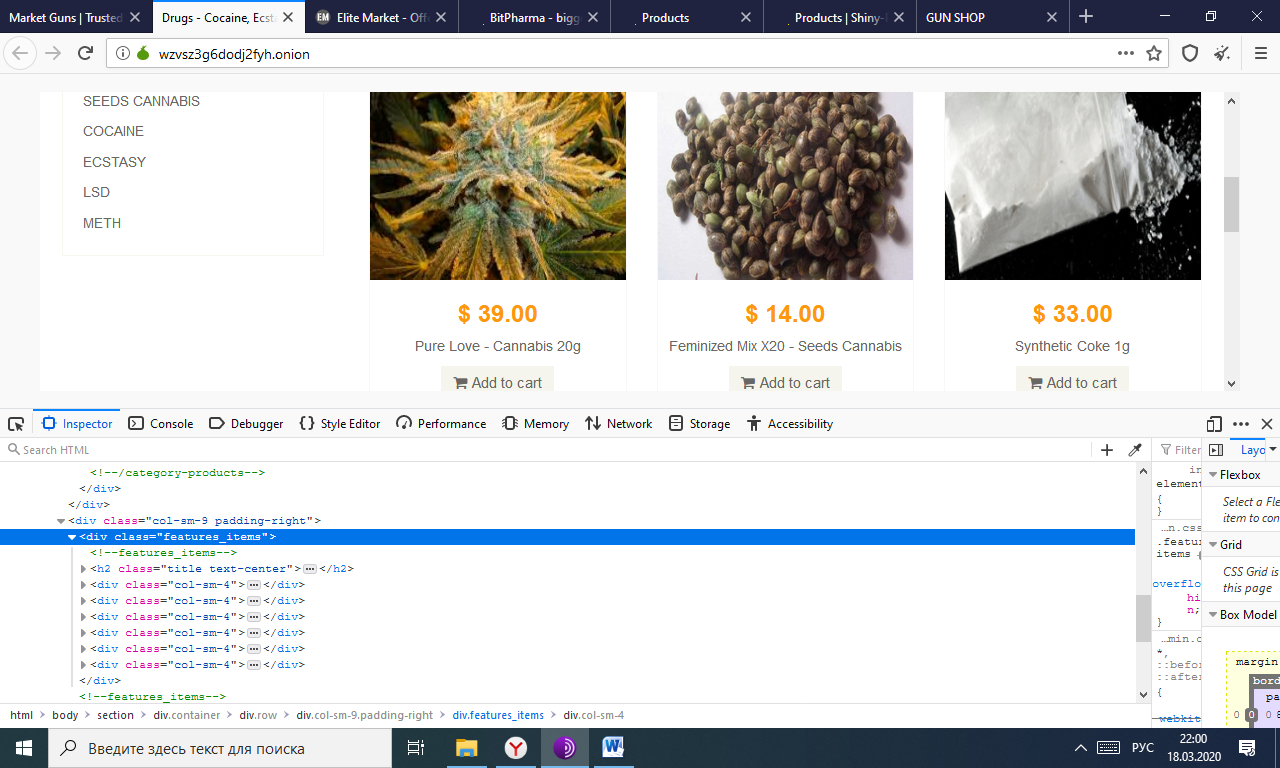
* Текст, включающий какие то плохие слова про наркотики, оружие и так далее. Возможно это будет являться наименованием товара.
* Снова значки (USD, $, EUR, €, ฿). Скорее всего это его цена.
* И тэг <a> и его атрибут href (не src, т к он чаще всего содержит ссылку на картинку) – это обычно ссылка. Тут нужно внимательнее смотреть на ссылку (полная или нет).

На многих сайтах нельзя найти все три искомые составляющие данных о товаре. Иногда не хватает URL, а иногда даже и названия.

Поиск названий тэгов, где чаще всего может содержаться информация о товаре или цене, как мне кажется, идея безрезультатная. Товар может быть как в <div>, так и в <li>, <tr>, <td>, <table>. Разве что только цена чаще всего находится в тэге <span>, но не всегда так.

Проблемы:

А это пример списка подразделов, который очень легко бует спутать с каталогом. (Как отличить каталог от такого списка?)

Наряду с элементами, обозначающими товары в каталоге есть лишние элементы, как например заголовок <h2> (Как тогда определить каталог, если мы будем находить все таки отличающиеся ото всех в нём элементы?)

Ниже представлена архитектура одного из крупных интернет магазинов.

Это просто исключение из всех возможных правил.

