

SEOWORK

Парсинг

Горобий Сергей

Руководитель отдела аналитики в SEOWORK

- Более 6 лет в digital
- Ранее был ведущим SEO специалистом в WebIT и МВидео
- Опыт SEO и аналитики крупнейших Ecom России – Ozon, AliExpress Россия, МВидео, Лента, Эльдorado и др.



Содержание

1. Что это такое и зачем? Парсинг vs Скрапинг vs Краулинг
2. Где используем в платформе?
3. Как парсим?
4. Синтаксический анализ HTML. Селекторы.
5. XPath. Синтаксис.
6. Парсим с Screaming Frog. Краткие основы + фишки.
7. Получаем «болванку» или неполный контент. Что делать дальше?
8. Что делать если заблокировали?
9. Что делать если все перепробовали, а результата все равно нет?
10. BONUS. Пишем бота через BAS

Что это такое и зачем? Парсинг vs Скрапинг vs Краулинг



Где используем в платформе?

- WS
- Поисковая выдача
- Слепки HTML и их разбор в «Аналитике»
- Краулинг в «Тех мониторе»
- Большинство «инструментов»
- Данные с внешних источников (системы аналитики, вебмастера и т.д.)
- Скрапинг для создания сегментов

Как парсим?

- Вручную в браузере ([Scraper](#), [Data Scraper](#))
- Desktopные приложения (Screaming Frog, A-Parser, Content Downloader)
- Настройка в Excel ([ParserOK](#))
- Прямой парсинг в Google Таблицы (функции IMPORTXML и IMPORTHTML)

И еще много других способов и парсеров ([см. статью](#)).

Синтаксический анализ HTML. Селекторы.

- CSS selector (CSSPath)

CSSPath = 'html > body > span:nth-of-type(1) a'

- Xpath selector ([удобное расширение для браузера](#))

Xpath = 'html/body/span[1]//a'

- jQuery selector

XPath. Выбираем ноду ([ссылка](#))

Selecting Nodes

XPath uses path expressions to select nodes in an XML document. The node is selected by following a path or steps. The most useful path expressions are listed below:

Expression	Description
<i>nodename</i>	Selects all nodes with the name " <i>nodename</i> "
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

In the table below we have listed some path expressions and the result of the expressions:

Path Expression	Result
bookstore	Selects all nodes with the name "bookstore"
/bookstore	Selects the root element bookstore Note: If the path starts with a slash (/) it always represents an absolute path to an element!
bookstore/book	Selects all book elements that are children of bookstore
//book	Selects all book elements no matter where they are in the document
bookstore//book	Selects all book elements that are descendant of the bookstore element, no matter where they are under the bookstore element
//@lang	Selects all attributes that are named lang

XPath. Предикаты ([ссылка](#))

Predicates

Predicates are used to find a specific node or a node that contains a specific value.

Predicates are always embedded in square brackets.

In the table below we have listed some path expressions with predicates and the result of the expressions:

Path Expression	Result
/bookstore/book[1]	Selects the first book element that is the child of the bookstore element. Note: In IE 5,6,7,8,9 first node is [0], but according to W3C, it is [1]. To solve this problem in IE, set the SelectionLanguage to XPath: <i>In JavaScript: <code>xml.setProperty("SelectionLanguage","XPath");</code></i>
/bookstore/book[last()]	Selects the last book element that is the child of the bookstore element
/bookstore/book[last()-1]	Selects the last but one book element that is the child of the bookstore element
/bookstore/book[position()<3]	Selects the first two book elements that are children of the bookstore element
//title[@lang]	Selects all the title elements that have an attribute named lang
//title[@lang='en']	Selects all the title elements that have a "lang" attribute with a value of "en"
/bookstore/book[price>35.00]	Selects all the book elements of the bookstore element that have a price element with a value greater than 35.00
/bookstore/book[price>35.00]/title	Selects all the title elements of the book elements of the bookstore element that have a price element with a value greater than 35.00

XPath. Оси. ([ссылка](#))

XPath Axes

An axis represents a relationship to the context (current) node, and is used to locate nodes relative to that node on the tree.

AxisName	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself
attribute	Selects all attributes of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself
following	Selects everything in the document after the closing tag of the current node
following-sibling	Selects all siblings after the current node
namespace	Selects all namespace nodes of the current node
parent	Selects the parent of the current node
preceding	Selects all nodes that appear before the current node in the document, except ancestors, attribute nodes and namespace nodes
preceding-sibling	Selects all siblings before the current node
self	Selects the current node

XPath. Неизвестные ноды, несколько путей. ([ссылка](#))

Selecting Unknown Nodes

XPath wildcards can be used to select unknown XML nodes.

Wildcard	Description
*	Matches any element node
@*	Matches any attribute node
node()	Matches any node of any kind

In the table below we have listed some path expressions and the result of the expressions:

Path Expression	Result
/bookstore/*	Selects all the child element nodes of the bookstore element
//*	Selects all elements in the document
//title[@*]	Selects all title elements which have at least one attribute of any kind

Selecting Several Paths

By using the | operator in an XPath expression you can select several paths.

In the table below we have listed some path expressions and the result of the expressions:

Path Expression	Result
//book/title //book/price	Selects all the title AND price elements of all book elements
//title //price	Selects all the title AND price elements in the document
/bookstore/book/title //price	Selects all the title elements of the book element of the bookstore element AND all the price elements in the document

XPath. Операторы. ([ссылка](#))

XPath Operators

Below is a list of the operators that can be used in XPath expressions:

Operator	Description	Example
	Computes two node-sets	//book //cd
+	Addition	6 + 4
-	Subtraction	6 - 4
*	Multiplication	6 * 4
div	Division	8 div 4
=	Equal	price=9.80
!=	Not equal	price!=9.80
<	Less than	price<9.80
<=	Less than or equal to	price<=9.80
>	Greater than	price>9.80
>=	Greater than or equal to	price>=9.80
or	or	price=9.80 or price=9.70
and	and	price>9.00 and price<9.90
mod	Modulus (division remainder)	5 mod 2

XPath. Функции.

Основные полезные функции:

- `count()`
- `last()` / `position()`
- `string-length()`
- `contains()`
- `starts-with()` или `ends-with()`
- `boolean()`
- `substring-before()` или `substring-after()`
- `normalize-space()`

Больше функций ([источник 1](#), [источник 2](#), [источник 3](#))

Парсим с Screaming Frog. Краткие основы + фишки.



Screamingfrog

Парсим с Screaming Frog. Краткие основы + фишки.

- File и Mode
- Configuration -> Spider
- Configuration -> Custom, HTTP Header, User-Agent, Speed
- Configuration -> Authentication
- Configuration -> API Access
- Configuration -> Кратко по остальным
- Bulk Export и Reports
- Crawl Analysis
- Sitemaps
- Visualisations

Получаем «болванку» или неполный контент. Что делать дальше?



Получаем «болванку» или неполный контент. Что делать дальше?

- Поменять user-agent на поискового бота (перед изменением можно посмотреть view source под другим ботом в браузере)
- Включить и настроить рендеринг (настраиваем вкладки в Configuration)
- Отправлять с запросом уникальные HTTP Headers (User-agent, Cookie, Асцепт, Асцепт-Encoding)
- Тянем сохраненные копии из поисковых систем (или SW)
- Создаем и настраиваем бота

Что делать если заблокировали?



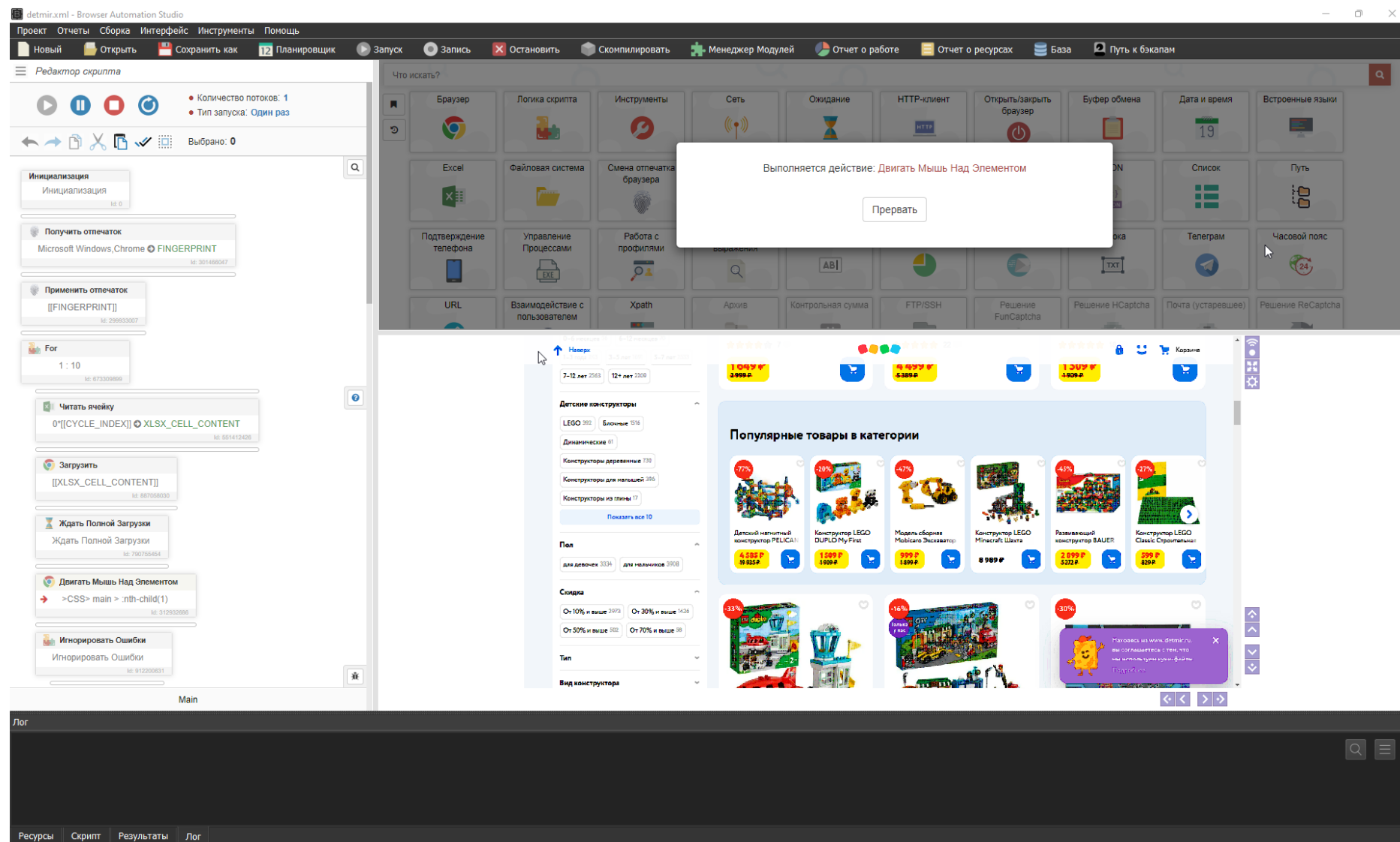
Что делать если заблокировали?

- Отправлять с запросом уникальные HTTP Headers (берем из реального браузера)
- Включить рендеринг (если JS не включен, то может послужить триггером для блокировки)
- Уменьшить кол-во потоков и URL/сек
- Соблюдаем правила robots.txt (триггер - много запросов на заблокированные в robots URL)
- Следуем редиректам
- Прокси-ротатор

Что делать если заблокировали?

- Таймауты (каждый раз паузы разной длинны)
- Имитация скrolла и кликов (каждый раз разные в разном месте)
- Подключить VPN
- Ротация HTTP Headers
- Антикапча
- Пауза при блокировке на первом же URL и попытка возобновления парсинга через время
- Фингерпринты

BONUS. Пишем бота через BAS



Спасибо! Вопросы?



 <https://seowork.ru>

<https://dataforce.io>



 Группа [facebook.com](https://www.facebook.com)



 Канал на [youtube.com](https://www.youtube.com)



 Канал в телеграм

https://t.me/seowork_hub

