

# WEBARCHIVE MASTER



БЕЗШАБЛОННЫЙ ПАРСИНГ  
ТЕКСТА НА ВСЕХ ЯЗЫКАХ



ПОЛНОСТЬЮ ОТКРЫТЫЙ  
ШАБЛОН



ПРОВЕРКА ОТВЕТА СЕРВЕРА  
НА ОТВЕТ 200



ПОДГОТОВКА К ПРОВЕРКЕ  
НА УНИКАЛЬНОСТЬ



ФИЛЬТРАЦИЯ МУСОРА - CSS,  
КАРТИНКИ И Т.Д. - ЧИСТЫЙ ТЕКСТ!



...И МНОГОЕ ДРУГОЕ...

**WebArchiveMaster** - программа парсинга контента из ВебАрхива. Программа полностью автоматизирована и позволяет разгрузить своё время на 90%. Программа работает в связке с PHP скриптом, который можно поставить на любой хостинг или использовать **Open Server** - <https://ospanel.io> (рекомендуется).

## Принцип работы

Принцип работы очень прост - нужно только вставить домены в текстовый файл и запустить программу - все остальное она сделает сама. Никаких настроек нет, так-как все настроено на максимальную производительность. Разберем на примере:

Domens.txt

```
1 kuritenazdorovie.ru
2 ufazdorovie.ru
3 kosmetikazdorovo.ru
4 nazdorovie-spb.ru
5 seksualnoe-zdorovie.ru
6 goodzдорovie.ru
7 sportzдорovie.ru
8 retseptzдорovya.ru
9 mir-zdorove.ru
10 seksualnoe-zdorovie.ru
11 albon-zdorovya.ru
12 x-zdorovo.ru
13 apteka-zdorovya.ru
14 voda-istochnik-zdorovya.ru
15 krepkozdorove.ru
16 magazinzdorovie.ru
17 vipzdorov.ru
18 vestnikzдорov.ru
19 clubzдорove.ru
20 semjzдорova.ru
21 arm-zdorovo.ru
22 centr-zdorovie.ru
23 liniadzдорovia.ru
24 dushvnoezdorovie.ru
25 chelovek-zdorovee.ru
26 dlja-zdorovya.ru
27 clubzдорov.ru
28 natur-zdorovie.ru
29 bitzдорovim.ru
30 detskoe-zdorovie.ru
```

Скопируйте домены в файл: Domens.txt, запустите программу и можете отдыхать.

Директория должна находиться по адресу:  
**c:\bot\WebАрхив**

# Установка скрипта

Разберем установку бесшаблонного парсинга - скачиваем **Open Server**.

Встречайте: Open Server!



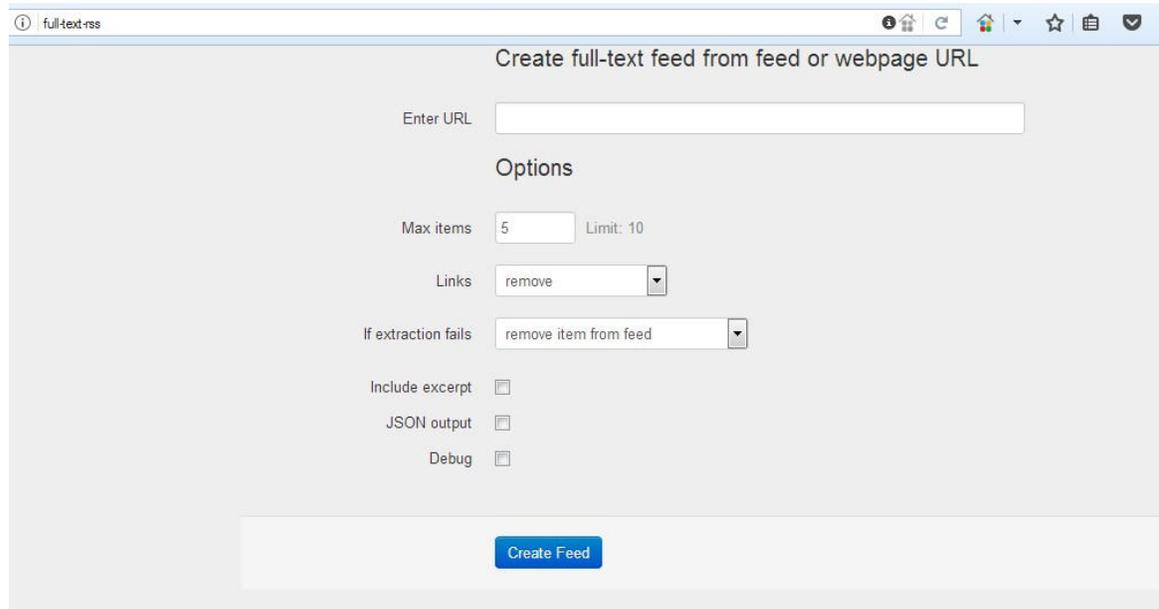
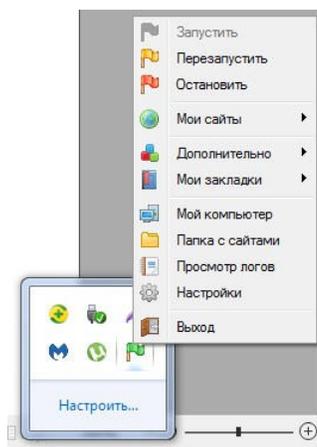
Open Server Panel — это портативная серверная платформа и программная среда, созданная специально для веб-разработчиков с учётом их рекомендаций и пожеланий.

Программный комплекс имеет богатый набор серверного программного обеспечения, удобный, многофункциональный продуманный интерфейс, обладает мощными возможностями по администрированию и настройке компонентов. Платформа широко используется с целью разработки, отладки и тестирования веб-проектов, а так же для предоставления веб-сервисов в локальных сетях.

Хотя изначально программные продукты, входящие в состав комплекса, не разрабатывались специально для работы друг с другом, такая связка стала весьма популярной среди пользователей Windows, в первую очередь из-за того, что они получали бесплатный комплекс программ с надёжностью на уровне Linux серверов.

Удобство и простота управления безусловно не оставят вас равнодушными, за время своего существования Open Server зарекомендовал себя как первоклассный и надёжный инструмент необходимый каждому веб-мастеру.

## Запускаем локальный сервер



Запускаем сервер, после запуска вставляем в браузер название скрипта (база данных не требуется) и все, программа готова к работе. Точно также можно установить на хостинг - просто копируете скрипт на домен или поддомен и все готово к работе.

В scraper.txt настраивается путь к скрипту. Если вы установили на поддомене <http://feed.cheerfulness.ru>, то так и пишете, если на локальном сервере, то пишете: <http://full-text-rss/>

# Принцип работы

Как работает программа - берет выборочно домен и проверяет его на ответ 200 (сайт работает). Если сайт работает, домен удаляется и берется следующий. После получения нужного домена, программа подключается к Вебархиву и запрашивает количество файлов за все годы (не по снелшотам). Если файлов нет, возвращается к выбору другого домена. Если файлы есть, программа забирает ссылки и включает фильтрацию (css, png, jpg, reply и т.д.).

После этого чистит ссылки и включает скрипт скрапинга, забирает текст и начинает его чистить от всего мусора, тегов и т.д.

Программа пишет все статьи в один файл без заголовков, и этому есть причины - все это отработана неделями тестирования и выбран лучший вариант среди тысяч - разберем некоторые:

Парсить приходится самые различные системы управления контентом, как cms, так и обычные сайты и фреймворки и самоделки, в которых просто нет зацепок для программы, типа Title или H1 - их просто может не быть. Поэтому программа работает так - берет текст, фильтрует его, пишет в один файл, затем удаляет дубли (здесь ещё один из тысяч подводных камней - сайты имеют неявные дубли, и одна страница может открываться как по адресам: /page&?p233, так и по /ozdorovlenie/ и по /ozdorovlenie.html, к тому же стоят редиректы и другие всевозможные перенаправления.

Это одна и та же страница, и это создает очень большие проблемы не только для поисковых систем.

Поэтому все пишется в один файл и после того, как все страницы скачены, программа удаляет все дубли и затем каждую страницу сохраняет в отдельный файл. Это нужно для массовой проверки через антиплагиат - я использую **eTXT Антиплагиат**, она позволяет использовать пакетную проверку хоть тысячи файлов. Для капчи я использую **XEvil**.

Вот как это выглядит (готовый сайт):

Готовая статья19	txt	5 750	21.08.2017 07:37
Готовая статья18	txt	11 401	21.08.2017 07:37
Готовая статья17	txt	12 018	21.08.2017 07:37
Готовая статья16	txt	10 086	21.08.2017 07:37
Готовая статья15	txt	2 916	21.08.2017 07:37
Готовая статья14	txt	19 867	21.08.2017 07:37
Готовая статья13	txt	2 992	21.08.2017 07:37
Готовая статья12	txt	2 188	21.08.2017 07:37
Готовая статья11	txt	4 512	21.08.2017 07:37
Готовая статья10	txt	3 804	21.08.2017 07:37
Готовая статья9	txt	10 309	21.08.2017 07:37
Готовая статья8	txt	12 723	21.08.2017 07:37
Готовая статья7	txt	4 243	21.08.2017 07:37
Готовая статья6	txt	4 195	21.08.2017 07:37
Готовая статья5	txt	19 689	21.08.2017 07:37
Готовая статья4	txt	11 733	21.08.2017 07:37
Готовая статья3	txt	3 837	21.08.2017 07:37
Готовая статья2	txt	19 647	21.08.2017 07:37
Готовая статья1	txt	2 570	21.08.2017 07:37
Статьи в одном файле	txt	266 643	21.08.2017 07:37
Все текстовые данные	txt	267 542	21.08.2017 07:37

Все статьи сохраняются в папку с названием домена, с которого они были скачены. Это сделано для того, чтобы, если статьи понравятся, можно попытаться восстановить дроп.

Готовые сайты <Папка>

nahezdorovie.ru	<Папка>
www.nahezdorovie.ru	<Папка>
zdorovoru.ru	<Папка>
www.zdorovoru.ru	<Папка>
korzinkazdorovya.ru	<Папка>
budtezdorovimi.ru	<Папка>
kladezzdorovya.ru	<Папка>
zdorovemoe.ru	Дата создания: 21.08.2017 7:26
zdorovie-glaza.ru	<Папка>
zdorove-pitanie-pohudenie.ru	<Папка>
praktikzdorovie.ru	<Папка>
www.praktikzdorovie.ru	<Папка>
zdorovaja.ru	<Папка>
zdorovie-live.ru	<Папка>
klinica-zdorovie.ru	<Папка>
www.klinica-zdorovie.ru	<Папка>
osnovazdorov.ru	<Папка>
zdorovesad.ru	<Папка>
hobby-zdorove.ru	<Папка>
obninsk-zdorovie.ru	<Папка>
...	...

## Что делать со статьями

Расскажу про свою методику. После того, как скачено около тысячи текстов, я выбираю 500-600 статей по размеру (3 - 15 тысяч символов одна статья, остальные сбрасываю в резервную папку для саттелитов или дорвеев), пакетно загружаю в **eTXT** и запускаю проверку на уникальность. Я ставлю настройки 80% уникальности и антиплагиат сам раскидывает их в разные папки - прошедшие уникальность и не прошедшие.

Затем я за копейки покупаю на Телдери старый трастовый сайт 2-3 лет, который давно не обновлялся и работает в убыток и публикую на нем статьи. Статьи очень хорошо заходят и сидят в выдаче, многие мои сайты были приняты во все биржи, некоторые в РСЯ. На молодом сайте так делать опасно, так как статьи иницированы, и скорее всего яндекс про них знает- уникальность позволяет только определить, что этих статей нет на других сайтах.

Продавал статьи на бирже и не имел ни одного отрицательного отзыва, но жадность сгубила и на объемах биржа спалила, из-за того, что кто-то еще продавал эти же статьи. Но деньги успел вывести, неплохую сумму. Так что поакуратнее, даже если статья показывает 100% уникальности на всех сервисах антиплагиата, не факт, что вас не забанят при загрузке статьи на биржу, т.к. у них своя база и каждую загруженную статью они сравнивают, не было ли такой ранее.

# Где взять брошенные домены

Здесь я покажу, где взять брошенные домены по тематикам

Интересные домены

Дата регистрации от 01.08.2010 до 01.08.2017 от  посетителей в сутки  освобожденные домены

ТИЦ от   есть в YаCa PR от   есть в DMOZ Маска  Подобрать домены ↓

auto\*[7], \*ivanov\*, \*[4-6]- ВИДЫ МАСОК

Доменное имя	ТИЦ	страниц в Яндексе	PR	страниц в Google	посетители	просмотры	Возраст домена, лет	Оплачен до
<a href="#">kuritenazdorovie.ru</a>	0	193	0	380	..	..	5	свободен
<a href="#">ufazdorovie.ru</a>	0	0	2	0	..	..	1	свободен
<a href="#">youzdorov.ru</a>	0	0	0	0	..	..	1	06.06.2017 <sup>-77</sup>
<a href="#">kosmetikazdorovo.ru</a>	0	0	0	0	..	..	4	05.01.2018
<a href="#">nazdorovie-spb.ru</a>	20	0	0	0	..	..	3	04.10.2017
<a href="#">seksualnoe-zdorovi....ru</a>	0	0	0	0	..	..	6	свободен
<a href="#">goodzdorovie.ru</a>	0	0	1	0	..	..	6 мес.	21.02.2018
<a href="#">sportzdorovie.ru</a>	0	0	0	0	..	..	2	17.09.2017 <sup>-26</sup>
<a href="#">retseptzdorovya.ru</a>	0	0	..	..	..	..	1	01.03.2018
<a href="#">mir-zdorove.ru</a>	0	0	0	0	..	..	7 мес.	06.01.2018

## 1. Выбираем дату дропов за все года (2010-2017)

Дата регистрации от 01.08.2010 до 01.08.2017 от  посетителей в сутки  освобожденные домены

ТИЦ от   есть в YаCa PR от   есть в DMOZ Маска  Подобрать домены ↓

auto\*[7], \*ivanov\*, \*[4-6]- ВИДЫ МАСОК

2. Выбираем тематику - для этого используем маску: так-как мы выбрали тематику о здоровье, используем маску `*zdoro*`, если хотите что-то про женщин, можно выбрать маску `*wom*` или `*kraso*` и т.д.

Маска  Подобрать домены ↓

auto\*[7], \*ivanov\*, \*[4-6]- ВИДЫ МАСОК

www.nomina.ru/domains?date\_type=created&date\_from=01.08.2010&date\_to=01.08.2017&views\_type=vis&name=\*zdoro\*&

3. Нажимаем "Подобрать домены" и в командной строке браузера получаем ссылку, которую вставляем в файл "Парсинг доменов.txt" - программа соберет все домены с данной фильтрацией.