



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!



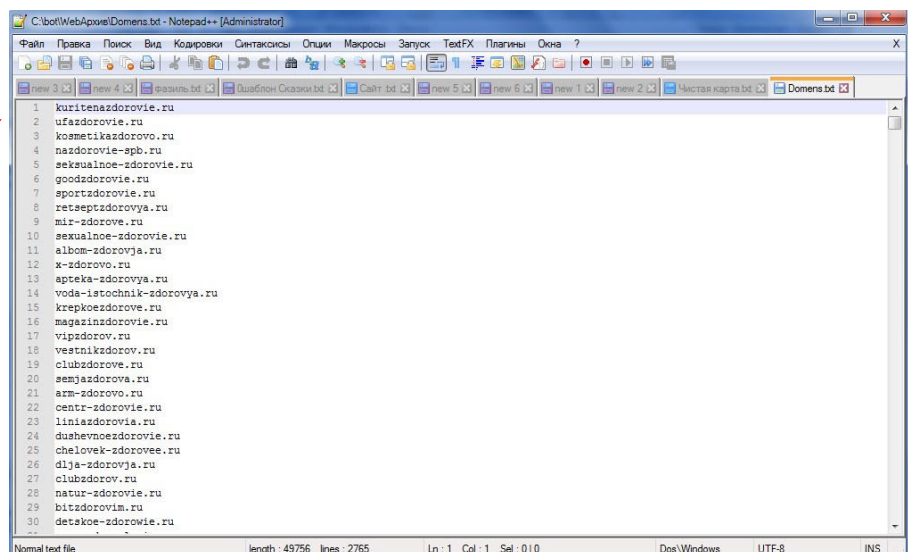
...И МНОГОЕ ДРУГОЕ...

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



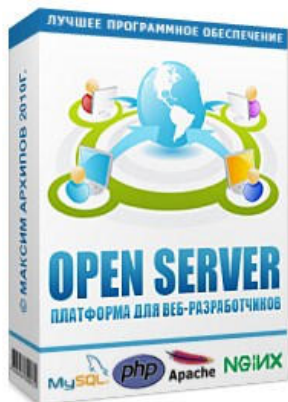
Скопируйте домены в файл: Domens.txt, запустите программу и можете отдыхать.

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

Установка скрипта

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

Встречайте: Open Server!



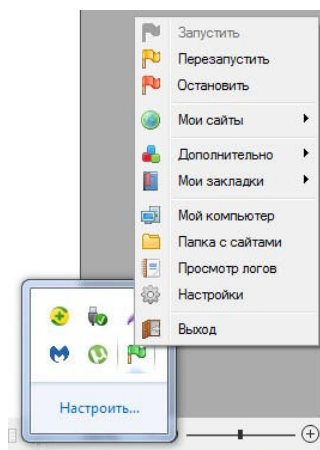
Open Server Panel — это портативная серверная платформа и программная среда, созданная специально для веб-разработчиков с учётом их рекомендаций и пожеланий.

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



full-text-rss

Create full-text feed from feed or webpage URL

Enter URL

Options

Max items Limit: 10

Links

If extraction fails

Include excerpt ☐

JSON output ☐

Debug ☐

Create Feed

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/>. Скрипт **full-text-rss**, находящийся в папке, нужно перенести на локальный сервер или хостинг. Программа будет к нему обращаться для парсинга.

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>
zdrovemoie.ru	Дата создания: 21.08.2017 7:26
zdorovie-glaza.ru	<Папка>
zdrove-pitanie-pohudenie.ru	<Папка>
praktikzdorovie.ru	<Папка>
www.praktikzdorovie.ru	<Папка>
zdrovaja.ru	<Папка>
zdorovie-live.ru	<Папка>
klinica-zdorovie.ru	<Папка>
www.klinica-zdorovie.ru	<Папка>
osnovazdorov.ru	<Папка>
zdrovesad.ru	<Папка>
hobby-zdrove.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 инициированы, 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 поаккуратнее, 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

Брошенные домены можно взять на expireddomains.net. Регистрируетесь, вводите ключевое слово в поиск, например, если нужны домены о здоровье, то пишете: zdo или zdr и скачиваете домены списком.

Вот пример работы программы за два часа - скачено около 300 текстов, выборочная проверка текста показала, что уникального контента очень много.

ПРОВЕРКА ТЕКСТА НА УНИКАЛЬНОСТЬ

[+ Новый текст](#)

Время проверки уникальности: 11.09.2017 13:53 (UTC +03:00) [Архив текстов](#) [API проверки](#)

Проверка уникальности

Уникальность: **100.00%**

[Получить ссылку на проверку](#)
[Зафиксировать уникальность](#)
[Получить кнопку уникальности](#)

[Подробнее](#)

Проверка орфографии

В тексте найдена 1 ошибка:

• я являются

[Подробнее](#)

SEO-анализ текста

Всего символов: **2321** Заспамленность: **49%**
Без пробелов: **2022** Вода: **10%**
Количество слов: **300**

[Подробнее](#)

Подсвечено: Неуникальные фрагменты

Диета и питание, являются самыми важными факторами поддержания здоровья. С точки зрения науки, питание является неотъемлемой частью здоровья организма, способствует его росту и развитию. Самыми необходимыми продуктами питания являются продукты, содержащие белки, жиры, углеводы, витамины и минералы. Хорошее питание способствует развитию хорошего здоровья. Во все времена, питание являлось самым важным фактором при здоровом и больном организме. Человечество постоянно боролось, чтобы получить продукты питания. В девятнадцатом веке, белки, жиры и углеводы были признаны самыми важными составляющими в продуктах питания человека. Ученые уделяли им особое внимание, т. к. они имели энергетическую ценность. Открытие витаминов в двадцатом веке, стало значительным достижением в области науки о питании. Примерно в году, все известные витамины и аминокислоты были уже известны. Питание получило признание в области научной медицины и с корнями уходило в физиологию и биохимию. Большой прогресс был достигнут в течение последних нескольких лет в области питания и его практического применения для удовлетворения ежедневных потребностей в питании человека. Для поддержания хорошего здоровья мы требуем предоставления питательных веществ, энергии и воды в определенных объемах. Конкретные питательные вещества должны включать девять основных аминокислот, несколько жирных кислот, четыре витаминов растворимых жиром, десять водорастворимых витаминов. Ряд неорганических веществ, в том числе четыре минерала, три электролита и микроэлементы должны быть в рационе питания. Необходимое количество основных питательных веществ зависит от состояния здоровья, возраста и уровня активности. Диета играет важную роль в поддержании и развитии вашего здоровья. Каждый должен строго соблюдать

Версии текста:

🔒 Минуту назад (UTC +03:00)

Уникальность	100%	Орфография	1	
Всего символов	2321	Заспамленность	49%	
Без пробелов	2022	Вода	10%	
Количество слов	300			

🔒 2 минуты назад (UTC +03:00)

Уникальность	100%	Орфография	1	
Всего символов	2344	Заспамленность	50%	
Без пробелов	2041	Вода	10%	
Количество слов	304			

1. Устранена проблема с Config.cfg - теперь задаваемое количество минимальных символов обрабатывает корректно.
2. Добавлен новый файл конфигурации - Kategory.cfg, который отвечает за исключение анонсов статей. Как правило, работает только со стандартными CMS и отсекает около 60% ненужных коротких анонсов. Но за счет чистки ненужных данных снижена скорость. По умолчанию в Kategory.cfg стоит цифра "1", что означает включение, вы можете отключить чистку категорий и ускорить парсинг, поставив "0". Можно при работающем проекте менять значения.
3. Теперь все данные сохраняются в одну папку, без "www"
4. Отрегулирован PHP скрипт, но мусор все равно будет цеплять - если текст небольшой, а данных на странице много (комментарии, рекламные слоганы, которые бывают больше текста), то неизбежно бесшаблонный парсер захватит их. Если текст чистый более-менее, то всё ненужное отсекается.
5. Новый файл конфигурации - Zapros.cfg. Он отвечает за проверку доменов на работоспособность - если домен работает, то он не проверяется, а берется следующий из списка. По умолчанию стоит цифра "1", если вы считаете, что теряется слишком много доменов, которые отвечают на запрос, что работают, а в действительности отключены и припаркованы (сервер не отдаёт ошибки), то можете поставить "0" и домены проверяться не будут. Но возрастает вероятность парсинга заведомо не уникальных текстов.