

Хабр



КАК СТАТЬ АВТОРОМ



Войти



empenoso

6 янв в 03:23

Как я научился оценивать популярность статей через парсинг показателей сайтов

Простой

7 мин

2.1K

Контент и копирайтинг*, Управление медиа*, Node.JS*, Open source*

Кейс

Технотекст 7

Сезон Open source

Я уже давно пишу статьи про различные аспекты IT-технологий, инвестиции, автоматизацию и умные дома на разных площадках: Хабр, Т—Ж, СмартЛаб, Пикабу, VC.ru и других.

За всё время накопилось примерно 250 статей, [которые по итогу свёл в таблицу](#). Но вот задумываться о популярности статей и их реальном эффекте стал относительно недавно.

LynxReport: учёт публикаций [Node.js]									
Файл Правка Вид Вставка Формат Данные Инструменты Расширения Справка									
Настройки Доступа									
Название									
1	A	B	C	D	E	F	G	H	I
	Название	Тип (задается вручную)	Дата выхода (задается вручную)	Выходные данные (задается вручную)	Автор(ы) (задается вручную)	Тема (задается вручную)	Тираж для печатных изданий	Просмотры	Ко
4	Таблицы и скрипты Гугл для бизнеса и анализа	Веб	2024-12-19	https://habr.com/ru/articles/871234/	М.В. Шардин	скрипты	-	950	
5	Топ-тренды Смартлаба 2024: что читали больше всего? Смартлаб	Веб	2024-12-19	https://smart-lab.ru/mobile/topic/1096274/	М.В. Шардин	Джва и гугл скрипты	-	4 687	
6	Что читали на Хабре в 2024 году: анализ статей с Node.js, Google Sheets и канпел ChatGPT Хабр	Веб	2024-12-18	https://habr.com/ru/articles/867068/	М.В. Шардин	Джва и гугл скрипты	-	7 109	
7	С бумажки на цифровую карту: генерация файла из таблицы для импорта на карту и геокодирование адресов с помощью Python Хабр	Веб	2024-12-12	https://habr.com/ru/articles/866558/	М.В. Шардин	Картография	-	1 974	
8	С бумажки на цифровую карту: генерация файла из таблицы для импорта на карту и геокодирование адресов с помощью Python Пикабу	Веб	2024-12-12	https://pikabu.ru/story/slaboumie_i_otvaga_kak_nayti_likvidnyie_obligatsii_s_dokhodnostyu_do_40_i_eksmeschnymi_vypplatami_na_moskovskoy_birzhe_12126449	М.В. Шардин	Картография	-	4 271	
9	Слабоумие и отвага: как найти ликвидные облигации с доходностью до 40% и ежемесячными выплатами Смартлаб	Веб	2024-12-06	https://smart-lab.ru/mobile/topic/1090879/	М.В. Шардин	Инвестиции	-	5 823	
10	Слабоумие и отвага: как найти ликвидные облигации с доходностью до 40% и ежемесячными фиксированными выплатами Хабр	Веб	2024-12-06	https://habr.com/ru/articles/863762/	М.В. Шардин	Инвестиции	-	7 609	
11	Слабоумие и отвага: как найти ликвидные облигации с доходностью до 40% годовых и ежемесячными фиксированными выплатами на Московской бирже Пикабу	Веб	2024-12-06	https://pikabu.ru/story/slaboumie_i_otvaga_kak_nayti_likvidnyie_obligatsii_s_dokhodnostyu_do_40_godovykh_i_eksmeschnymi_vypplatami_na_moskovskoy_birzhe_12098419	М.В. Шардин	Инвестиции	-	3 444	
12	Слабоумие и отвага: как найти ликвидные облигации с доходностью до 40% годовых и ежемесячными	Веб	2024-12-06	https://nationalanswer.club/post/1087/	М.В. Шардин	Инвестиции	-	нд	

Почему я решил собирать статистику публикаций?

РЕКЛАМА: ЧТОБЫ ПОСМОТРЕТЬ ЭТИХ ПУБЛИКАЦИЙ БЫЛО ПОДОБИТО СВОИМ ОПЫТОМ И ПРИ ЭТОМ



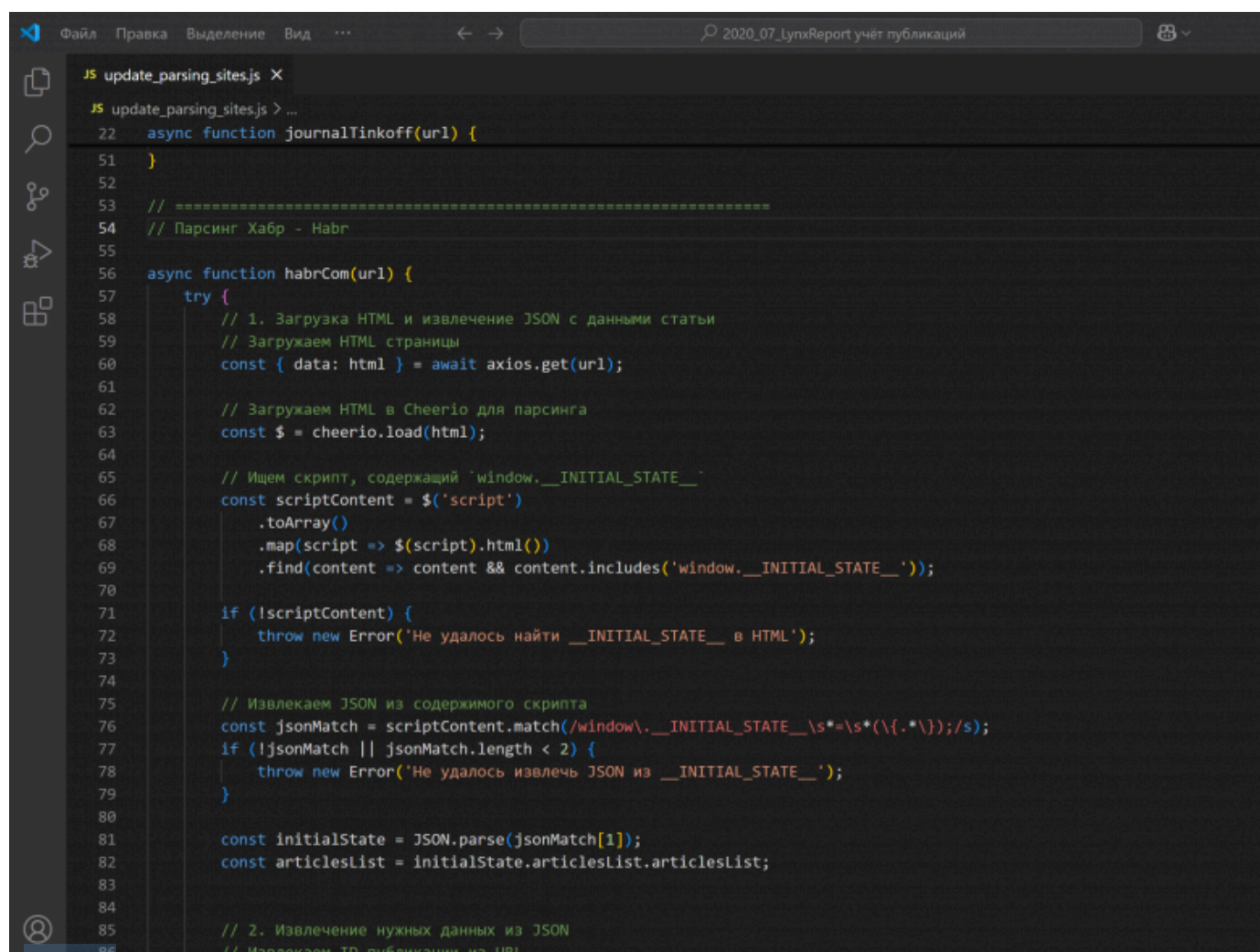
Горячая скидка 49%
на VPS в Малайзии

на Хабре статьи набирали большой рейтинг при малом количестве комментариев (но это редко).

Однако очевидно одно - статистика заставляет посмотреть на материалы со стороны. Например статья про [то, как я при помощи двух скриптов смог автоматически сгенерировать опись документов для 700 страниц](#) на непрофильном Пикабу собрала три месяца назад почти 75 тысяч просмотров и 80 комментариев, а на Хабре эта же тема была не особо популярна.

Или [статья про то, что читали на Хабре в 2024 году: анализ статей с Node.js, Google Sheets и каплей ChatGPT](#) - собрала на Хабре три недели назад рейтинг +68, но «всего» 7 тысяч просмотров.

Зачем я написал Open Source скрипт?



```
JS update_parsing_sites.js X
JS update_parsing_sites.js > ...
22  async function journalTinkoff(url) {
51  }
52
53  // =====
54  // Парсинг Хабр - Habr
55
56  async function habrCom(url) {
57    try {
58      // 1. Загрузка HTML и извлечение JSON с данными статьи
59      // Загружаем HTML страницы
60      const { data: html } = await axios.get(url);
61
62      // Загружаем HTML в Cheerio для парсинга
63      const $ = cheerio.load(html);
64
65      // Ищем скрипт, содержащий 'window.__INITIAL_STATE__'
66      const scriptContent = $('script')
67        .toArray()
68        .map(script => $(script).html())
69        .find(content => content && content.includes('window.__INITIAL_STATE__'));
70
71      if (!scriptContent) {
72        throw new Error('Не удалось найти __INITIAL_STATE__ в HTML');
73      }
74
75      // Извлекаем JSON из содержимого скрипта
76      const jsonMatch = scriptContent.match(/window\.__INITIAL_STATE__\s*=\s*(\{.*\});/s);
77      if (!jsonMatch || jsonMatch.length < 2) {
78        throw new Error('Не удалось извлечь JSON из __INITIAL_STATE__');
79      }
80
81      const initialState = JSON.parse(jsonMatch[1]);
82      const articlesList = initialState.articlesList.articlesList;
83
84
85      // 2. Извлечение нужных данных из JSON
86      // Извлекаем ID публикации из URL
```



Горячая скидка 49%
на VPS в Малайзии

Извлечение просмотров, комментариев, закладок и рейтинга из каждой статьи вручную занимало бы много времени, поэтому я решил пойти путём автоматизации. Написал скрипт, который скачивает эти данные по статьям и помещает сразу в одну удобную таблицу, где я вижу, какие темы стоит развивать дальше и на каких ресурсах.

Это [Open Source скрипт](#), размещенный на [Гитхабе](#), который состоит из Google Apps Script и Node.js частей и обе эти части работают с итоговой сводной Google Таблицей.

Если вы автор или агентство и хотите видеть полную картину популярности ваших материалов, выявлять, какие темы интересуют аудиторию, а какие требуют доработки, автоматизированный инструмент сбора данных - это то, что вам нужно. Я уже испытал это на себе и расскажу подробнее об этом в статье.

Особенности механизма парсинга сайтов, где размещены публикации

Изначально я использовал Google Apps Script внутри Гугл Таблиц для автоматизации извлечения просмотров, комментариев, закладок, рейтинга. Однако по мере того, как я расширялся на новые платформы, то обнаружил, что некоторые сайты ограничивают доступ к определенным данным, что потребовало от меня перехода на серверный язык JavaScript - Node.js.

Конкретный пример: сайт инвесторов и трейдеров Смартлаб имеет скрытое API для получения просмотров статей.

Если обращение идёт через Google Apps Script внутри Гугл Таблиц, то сервер просто не отдаёт просмотры, возвращая пустой ответ, потому что заголовки headers и User-Agent не поддерживаются Google Apps Script:

```
function fetchApiResponse() {  
  const url = 'https://smart-lab.ru/cgi-bin/gcn.fcgi?list=1083556&func=func8422&_=17319  
  
  try {  
    // Выполняем запрос к API с детализированными заголовками  
    const response = UrlFetchApp.fetch(url, {  
      method: 'get', // HTTP метод GET  
      muteHttpExceptions: true, // Не выбрасывать ошибки для HTTP ответов с кодами, отл
```



Горячая скидка 49%
на VPS в Малайзии

```
'Accept-Language': 'ru-RU,ru;q=0.9,en-US;q=0.8,en;q=0.7',
'Cache-Control': 'max-age=0',
'Sec-CH-UA': '"Chromium";v="130", "Google Chrome";v="130", "Not?A_Brand";v="99"',
'Sec-CH-UA-Mobile': '?0',
'Sec-CH-UA-Platform': '"Windows"',
'Sec-Fetch-Dest': 'document',
'Sec-Fetch-Mode': 'navigate',
'Sec-Fetch-Site': 'none',
'Sec-Fetch-User': '?1',
'Upgrade-Insecure-Requests': '1'
}
});

// Проверяем HTTP код ответа
const responseCode = response.getResponseCode();
if (responseCode !== 200) {
  console.error(`Ошибка: Получен HTTP код ответа ${responseCode}`);
  return;
}

// Получаем содержимое в виде обычного текста
const content = response.getContentText();

// Логируем необработанный ответ для отладки
console.log('Необработанный ответ API:', content);

} catch (error) {
  // Обрабатываем ошибки во время выполнения запроса
  console.error('Ошибка при выполнении запроса к API:', error.message);
}
}
```



Горячая скидка 49%
на VPS в Малайзии

Google Apps Script

Если же запрос идёт через Node.js и использование User-Agent, то в ответ возвращается число просмотров даже без использования эмуляции браузера:

```
const axios = require('axios');

async function fetchApiResponse() {
  const url = 'https://smart-lab.ru/cgi-bin/gcn.fcgi?list=1083556&func=func8422&_=173';

  try {
    // Установка заголовка User-Agent для имитации браузера Chrome
    const options = {
      headers: {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/53'
      }
    };

    // Выполнение запроса к API для получения просмотров
    const apiResponse = await axios.get(url, options);
    const responseText = apiResponse.data;

    // Логируем необработанный ответ для отладки
    console.log('Необработанный ответ API:', responseText);

  } catch (error) {
```



Горячая скидка 49%
на VPS в Малайзии

```
}  
}  
  
fetchApiResponse();
```

Node.js

Проблема оценки популярности статей

Обратная связь от аудитории может сильно варьироваться в зависимости от платформы, формата подачи материала и времени публикации.

Например, статья, выложенная одновременно на разных сайтах, может получить совершенно разные результаты. Уже упоминаемая выше статья "Как я при помощи двух скриптов смог автоматически сгенерировать опись документов для 700 страниц" собрала:

vc.ru:

225 / 3 / 2 / 2

Просмотры / Комментарии / Закладки / Рейтинг

Пикабу:

74 615 / 76 / 0 / 903

Просмотры / Комментарии / Закладки / Рейтинг



Горячая скидка 49%
на VPS в Малайзии

Просмотры / Комментарии / Закладки / Рейтинг

На Пикабу статья вызвала бурное обсуждение, в то время как на vc.ru аудитория осталась почти равнодушной.

Зачем собирать статистику публикаций?

Анализ метрик (просмотры, комментарии, закладки, рейтинг) помогает понять, что действительно интересует вашу аудиторию и какие темы популярны на различных ресурсах.

- **Просмотры:** показывают, сколько людей заинтересовалось темой.
- **Комментарии:** демонстрируют вовлечённость, обсуждения.
- **Закладки:** отражают ценность материала для долгосрочного использования.
- **Рейтинг:** индикатор общего одобрения.

Анализ этих данных помогают адаптировать стиль, темы и платформы для повышения эффективности публикаций.

Создание Open Source скрипта для сбора данных

За несколько лет написал [Google Apps Script](#), а затем и [Node.js](#), которые предназначены для автоматического сбора показателей публикаций с различных платформ.

Важно, что скрипты приспособлены к конкретной структуре каждой платформы, используя API (Тинькофф Журнал, Пикабу, Смартлаб) или парсинг HTML-кода (Хабр, vc.ru).



Горячая скидка 49%
на VPS в Малайзии

Различие Google Apps Script и Node.js версий

Если очень кратко, то версия на Google Apps Script проста в использовании, запускается прямо в браузере, идеальна для быстрой интеграции с таблицами Google, а Node.js версия мощнее, работает на сервере, подходит для более сложных задач и интеграций.

Google Apps Script – это облачное решение, подходящее для задач, не требующих больших вычислительных ресурсов. Он интегрирован в Google Workspace и позволяет легко работать с веб-страницами, API Google, используя встроенные функции. Однако его функционал ограничен, и он не может работать с ресурсоёмкими задачами или сложным парсингом.

Node.js – это серверная среда, использующая Javascript. Это мощное решение для сложного парсинга, работы с API, обработки больших объёмов данных и реализации сложных логик. Node.js позволяет использовать различные библиотеки (например, Axios, Cheerio, Puppeteer), расширяя возможности скрипта. Однако Node.js требует установки и настройки на сервере или локально.



Горячая скидка 49%
на VPS в Малайзии

Публикация скриптов на GitHub в рамках философии Open Source имеет несколько преимуществ. Во-первых, это позволяет другим пользователям использовать и адаптировать скрипт под свои нужды. Во-вторых, открытый исходный код способствует прозрачности и аудиту кода, что помогает обнаруживать и исправлять ошибки. В-третьих, сообщество разработчиков может помочь в улучшении и расширении функциональности скрипта, если захочет :)

Пример анализа собранной статистики для выбора будущих тем

Мне кажется немного некорректно проводить анализ собственных статей здесь и поэтому я оттолкнусь от анализа двух популярных платформ: [Хабра](#) и [Смартлаба](#) по всем их публикациям за 2024 год.



Горячая скидка 49%
на VPS в Малайзии

Исходные данные были получены мной путем парсинга двух ресурсов. Использовался автоматизированный сбор данных с использованием Node.js для парсинга HTML-страниц и последующего извлечения информации о каждой статье. Данные о просмотрах, комментариях, рейтинге и закладках позволяли оценить популярность и вовлеченность аудитории.

Фильтр по минимальному количеству просмотров (5 000 на Смартлабе, 30 000 на Хабре) и рейтингу (более 30 на Хабре) позволил отфильтровать статьи, которые не соответствовали критериям популярности.

Определение трендов и предпочтений аудитории для разных платформ: Хабр и Смартлаб.



Горячая скидка 49%
на VPS в Малайзии

Хабр более ориентирован на технические темы, такие как настройка программного обеспечения, обходы блокировок, разработка и советы по программированию. Наиболее популярные темы за 2024 год на Хабре - проблемы блокировки YouTube, решение проблем с GoodbyeDPI, уязвимости безопасности.

На Смартлабе, наоборот, доминируют темы, связанные с инвестициями, экономикой и геополитикой. Трендовые темы Смартлаба - санкции, рубль, ключевая ставка, дедолларизация, анализ рынка акций и облигаций.

Оптимизация публикаций под разные платформы: про что писать и на чём делать акценты.

Для эффективной публикации на каждой платформе важно учитывать интересы ее аудитории. На Хабре следует фокусироваться на конкретных технических проблемах и их решениях, практических советах и "человеческом" стиле. Статьи с интересными заголовками, посвященные узким, конкретным темам, вероятнее всего, получат большую вовлеченность.

На Смартлабе, ключевым является анализ геополитических и экономических трендов, связанных с финансами. Оптимально создавать качественный аналитический контент, предлагающий рекомендации по инвестированию, с подкреплением фактами и экспертизой.

Выводы

Мой Open Source скрипт для автоматического сбора статистики (просмотры, комментарии, закладки, рейтинг) позволяет анализировать популярность статей на разных платформах, выявлять тренды и понять что разные платформы могут требовать разного контента.

Работа с Open Source инструментом делает этот процесс масштабируемым и полезным для других авторов и агентств.

Автор: Михаил Шардин

 [Моя онлайн-визитка](#)

 [Telegram «Умный Дом Инвестора»](#)



Горячая скидка 49%
на VPS в Малайзии

Хабы: [Контент и копирайтинг](#), [Управление медиа](#), [Node.JS](#), [Open source](#)

Редакторский дайджест



Присылаем лучшие статьи раз в месяц

**189****17.5**

Карма

Рейтинг

Михаил Шардин [@empenoso](#)

[Автоматизация](#) / [Данные](#) / [Финансы](#) / [Умные дома](#)

[Подписаться](#)

[Хабр Карьера](#) [Сайт](#) [Сайт](#) [Github](#)



Горячая скидка 49%
на VPS в Малайзии

 Комментарии 11

Публикации

ЛУЧШИЕ ЗА СУТКИ

ПОХОЖИЕ

**Exosphere**

15 часов назад

Ещё 10 ошибок авторов Хабра

 11 мин 3.2K +93 35 65**vital_pavlenko**

22 часа назад

Больше нет входа в IT. Только выход

 2 мин 77K +49 78 324**duran-duran**

21 час назад

Трамплин в интернет: как мы ускорили запуск Яндекс Браузера

 6 мин 3.3K +42 9 29**Nickmob**

16 часов назад

**Горячая скидка 49%**

на VPS в Малайзии

Ретроспектива

 +37 29 8

Myskat_90

21 час назад

Распределённый инференс и шардирование LLM. Часть 1: настройка GPU, проброс в Proxmox и настройка Kubernetes



Сложный



14 мин



1.8K

Тutorial

 +33 46 0

slava_rumin

15 часов назад

Мое производство электроцитов приносит 40 млн в год. Спасибо нейросетям и СССР за конструкторскую школу



Простой



14 мин



22K

Интервью

 +28 48 53

AlexeyNadezhin

23 часа назад

Важное обновление BatteryTest 2



Простой



3 мин



2.5K

 +27 37 9

ru_vds

16 часов назад

Как создавались вокальные эффекты Daft Punk



Средний



13 мин



837



Горячая скидка 49%

на VPS в Малайзии

 +23 5 1**GordienkoAnd**

19 часов назад

Как настраивать сети: готовые решения Selectel для максимальной отказоустойчивости



Средний



20 мин



2.5K

Обзор

 +21 28 0**alizar**

20 часов назад

Почему из технологий делают культы



Простой



8 мин



1.3K

Обзор

 +21 7 6

Превращаем бездушные тикеты в эпическое приключение

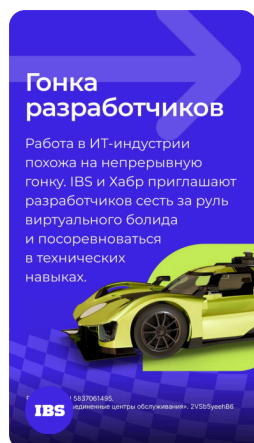
Промо

Показать еще

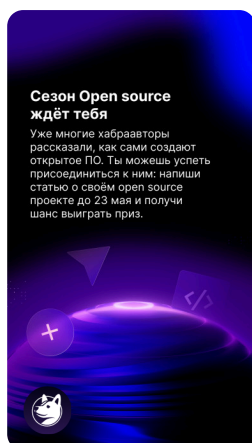
ИСТОРИИ

**Горячая скидка 49%**

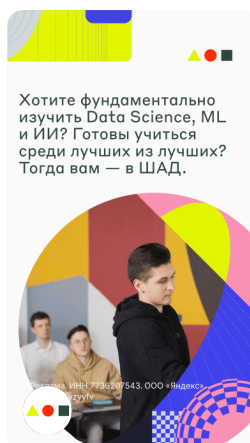
на VPS в Малайзии



Старт гонки разработчиков



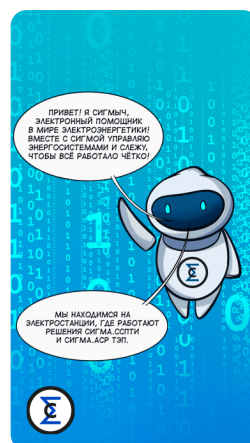
Торопись в сезон Open source



Открыт приём в Школу анализа данных



С Днём радио!



HELLO, WORLD — теперь на 110 киловольтах



Буду лучи

ВОПРОСЫ И ОТВЕТЫ

Как получить дерево категорий в PHP Simple HTML DOM Parser?

PHP · Средний · 0 ответов

Как парсить динамическую таблицу с помощью selenium?

Python · Простой · 0 ответов

Как сделать, чтобы при парсинге bs4 как то надо нажать на кнопку "показать еще", чтобы подгрузило еще 20 постов?

Python · Простой · 2 ответа

У меня не получается спарсить ссылку на фото с сайта, help?

Парсинг · Простой · 2 ответа

Как просмотреть скрытый sitemap.xml?

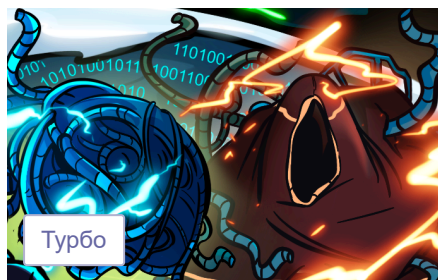
HTML · Простой · 3 ответа

[Больше вопросов на Хабр Q&A](#)

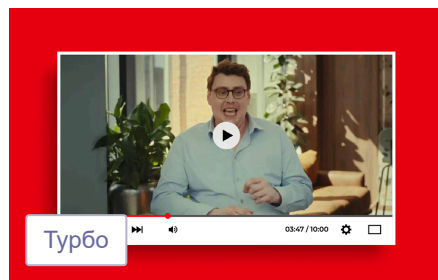
МИНУТОЧКУ ВНИМАНИЯ



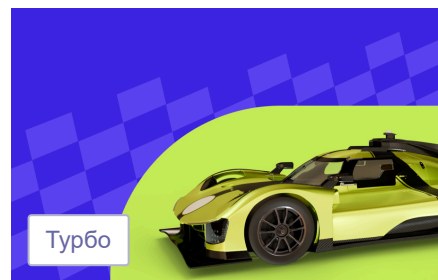
Горячая скидка 49%
на VPS в Малайзии



Как упорядочить хаос тикетов на спринте



Синтезируем речь без потери тембра спикера



Гонка разработчиков backend, frontend, mobile уже началась! Займи место на старте

РАБОТА

JavaScript разработчик

109 вакансий

Node.js разработчик

41 вакансия

Контент менеджер

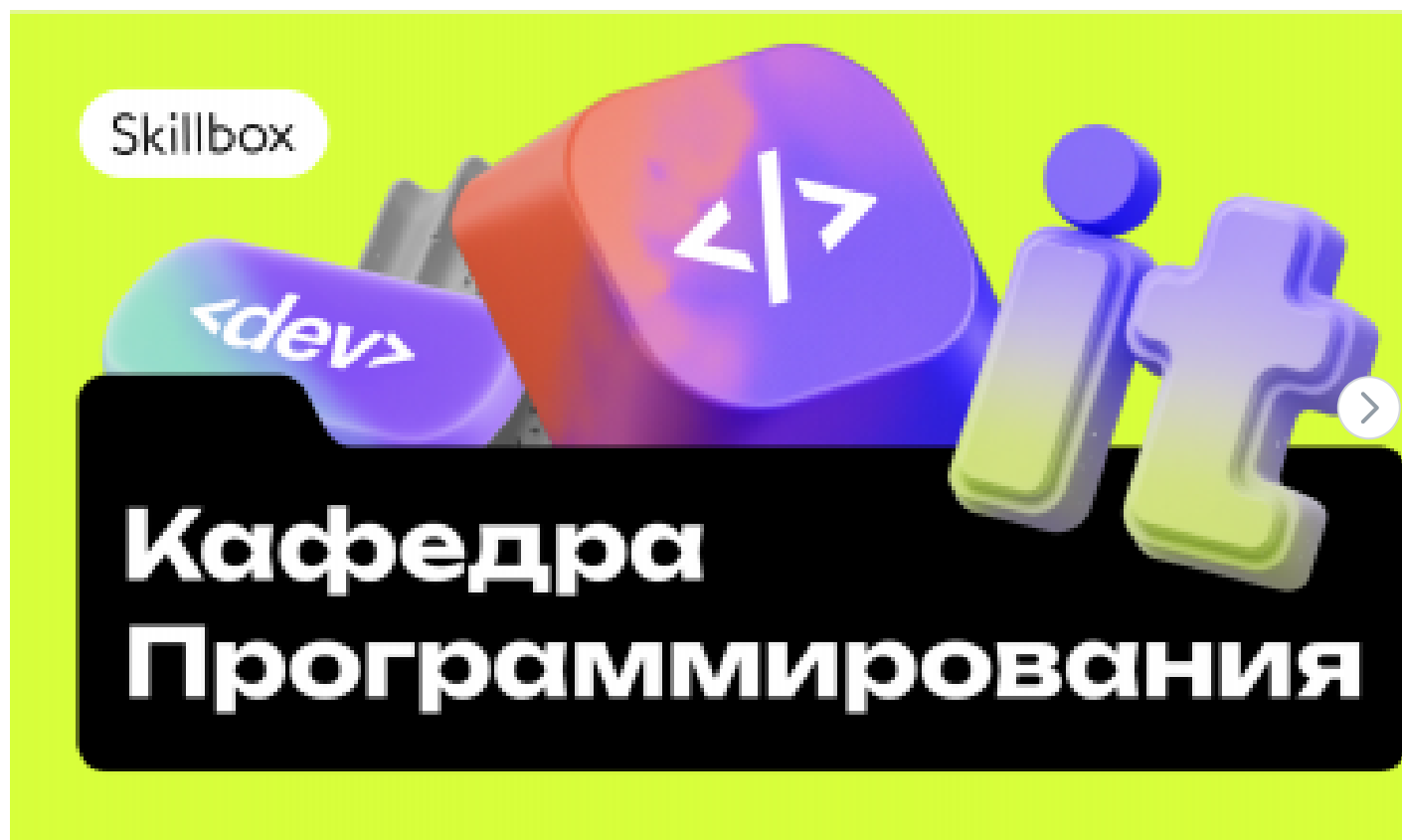
14 вакансий

[Все вакансии](#)

БЛИЖАЙШИЕ СОБЫТИЯ



Горячая скидка 49%
на VPS в Малайзии



17 апреля – 29 мая

Серия бесплатных офлайн-конференций «Кафедра Программирования» от Skillbox

Москва

Разработка

Больше событий в календаре

Хабр



Горячая скидка 49%
на VPS в Малайзии

техническая поддержка

© 2006–2025, Habr



Горячая скидка 49%
на VPS в Малайзии