Где искать данные?

Алексей Смагин | 2022

Для анализа нам подойдут любые объекты, которые имеют одинаковые свойства

Люди часто собирают информацию о похожих объектах

Автоматическим путём

- информация о звонках каждого абонента
- банковские транзакции
- штрафы о превышении скорости

Ручным путём

- количество пациентов в больницах
- список персонажей в Симпсонах
- список товаров в каталоге IKEA

Кто собирает информацию и выкладывает её в открытый доступ

Кто собирает информацию и выкладывает её в открытый доступ

Государство
раскрывает показатели
своей эффективности

Кто собирает информацию и выкладывает её в открытый доступ

Государство
раскрывает показатели своей эффективности

Компании
создают каталоги продуктов для своих клиентов

Кто собирает информацию и выкладывает её в открытый доступ

Государство
раскрывает показатели своей эффективности

Компании
создают каталоги продуктов для своих клиентов

Энтузиасты
собирают данные для общего блага

Кто собирает информацию и выкладывает её в открытый доступ

Государство
раскрывает показатели своей эффективности

Компании
создают каталоги продуктов для своих клиентов

Энтузиасты
собирают данные для общего блага

Исследователи
собирают данные для себя, но делятся ими со всеми

Для анализа нам нужны данные в табличной форме

(с другими мы не умеем работать)

Откуда взять табличные данные

- Найти данные уже в табличной форме
- Конвертировать из других форматов
- Собрать таблицу вручную
- Собрать данные при помощи автоматизированных средств
- Запросить данные

Источники данных

Государственная статистика

Росстат (rosstat.gov.ru)
• Статистика по темам (население, экономика, цены, окружающая среда…)
• Обследования, наблюдения, переписи (разные статистические показатели по одной теме)
• Публикации: готовые работы — буклеты, книги, аналитические отчёты, которые содержат статистику и инфографику

Базы данных Росстата с поиском:
ЕМИСС (fedstat.ru)
Витрина статистических данных (showdata.gks.ru)

Методология

На сайте Росстата есть раздел «методология», который рассказывает, как именно считался определённый статистический показатель.

Методология позволяет понять, как нам уместнее использовать данные, можно ли вообще доверять этим данным.

Порталы открытых данных

Свои порталы есть у почти каждого российского (и не российского) ведомства, например: открытые данные Минкульта, Минфина, Минздрава, ФНС, Роскомндазора, Реформы ЖКХ, NASA.
Также порталы бывают у компаний — OZON, ДомКлик; или у городов. регионов: Москвы, Екатеринбурга, ХМАО, Лондона.

Как искать: Через поисковик вбить «открытые данные + название ведомства» (или Open Data для англоязычных источников)

Что там искать: статистику, которой нет на Росстате; дезагрегированные данные; геоданные.

Паспорт открытых данных

Идентификатор набора данных:	1949
Идентификационный номер:	7710474791-CallfireAdmArea
Наименование набора данных:	Данные вызовов подразделений пожарно-спасательного гарнизона города Москвы по административным округам
Описание набора данных:	Количественные показатели вызовов подразделений пожарно-спасательного гарнизона города Москв административным округам
Владелец набора данных:	Департамент по делам гражданской обороны, чрезвычайным ситуациям и пожарной безопасности города Москвы
Ответственные за набор данных:	Ф.И.О.: Финогенова Татьяна Владимировна E-mail: FinogenovaTV@mos.ru Телефон: (495) 609-02-12
Гиперссылка (URL) на набор:	Ссылка на последнюю версию набора данных
Формат данных:	JSON
Описание структуры набора данных:	structure-20160706(vs1).json
Дата первой публикации:	19.10.2015
Дата последнего внесения изменений:	11.01.2021
Периодичность обновления:	По мере поступления
Содержание последнего изменения:	Выпуск релиза
Дата актуальности набора данных:	17.02.2021

Мировая статистика

World Bank (data.worldbank.org)
Собирает статистику по важнейшим показателям всех стран мира: ВВП, рождаемость, бедность, курящее население, экология…

Our World in Data (ourworldindata.org)
Делают инфографику и исследования, связанные со статистикой в разных странах, иногда делятся эксклюзивными наборами данных

Поисковые запросы

Google Trends
(trends.google.com)

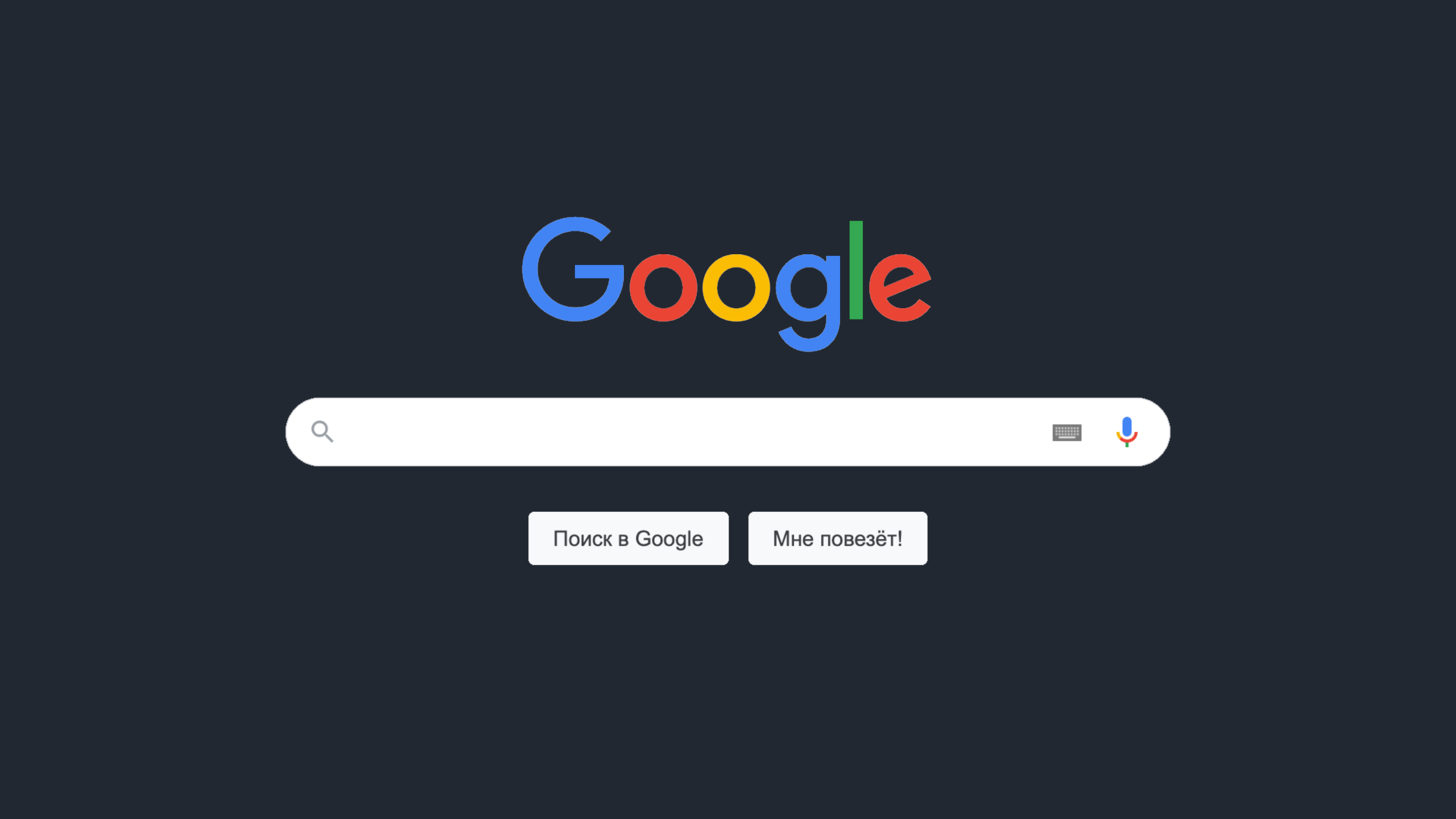
Яндекс Wordstat
(https://wordstat.yandex.ru/)

Госзакупки

Официальный портал — zakupki.gov.ru

Портал «Госрасходы» — spending.gov.ru
Удобный поиск и выгрузки данных

Где ещё найти данные?


Google
Поиск в Google
Мне повезёт!

Что ищем?

- Готовые наборы данных
- Аналитические отчёты и исследования
- Опросы

Важно:

- найти первоисточник
- понимать методологию расчётов

«Фишки» поиска

• Напишите запрос в кавычках, чтобы найти точное сочетание слов
• Напишите минус перед словом, чтобы исключить его из поиска
• При необходимости, ограничьте язык и время появления данных (выпадающее меню при поиске)

«Фишки» поиска

• Ограничьте зону поиска, ключевое слово **site**. Например: *население site:fedstat.ru*
• Ищите файлы определённого типа, ключевое слово **filetype**. Например: *пожары filetype:xlsx*
• Используйте оператор OR, чтобы связать ключевые слова. Например: *пожары filetype:xlsx OR filetype:csv*

https://datasetsearch.research.google.com/

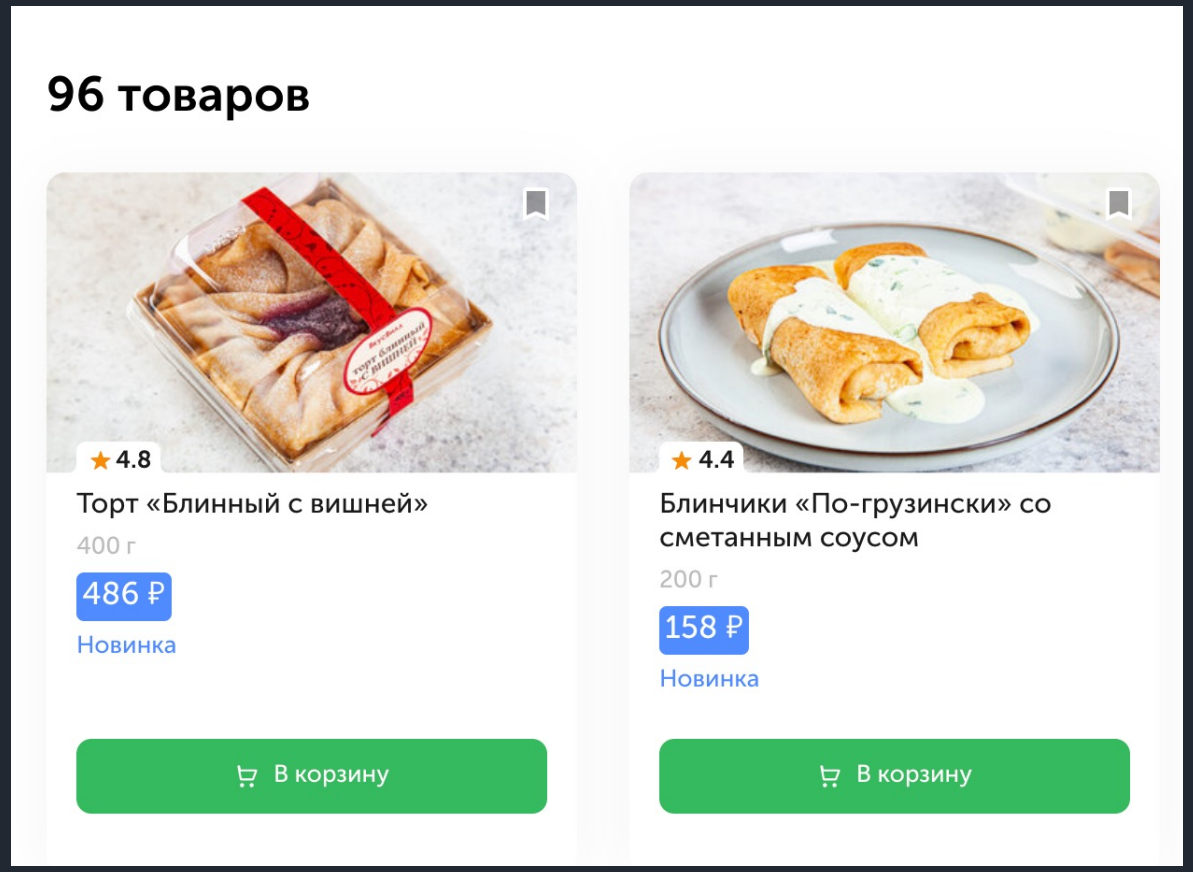
Собираем данные вручную

Из разных источников в интернете
• стоимость обучения в вузах
• количество новостей про протестные акции в разных изданиях

Из окружающего мира
• размеры карманов в джинсах
• список групп, выступавших на митингах
• стоимость блюд в столовой

Если нет возможности собрать данные автоматическим путём
• определить пол человека
• посчитать количество ресторанов, доступных в Delivery Club

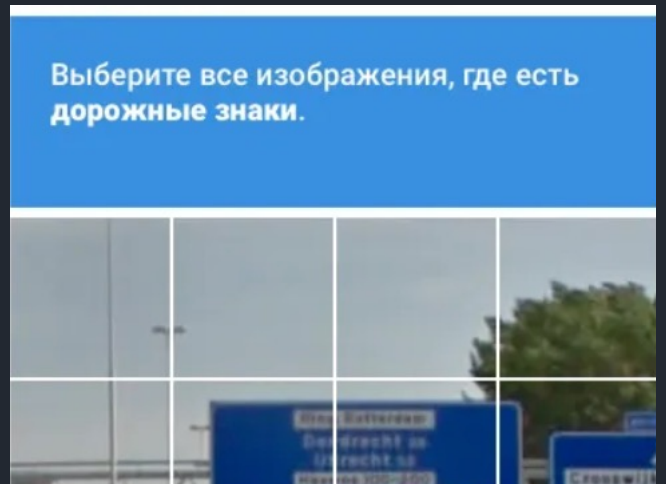
Собираем данные автоматическим образом



Если на сайте есть много страниц, устроенных похожим образом, его можно **скрейпить** — написать инструкцию, которую выгрузит оттуда нужные данные

Проблемные сайты для скрейпинга без программирования

• Рассчитывайте, что вы сможете написать скрипт, чтобы скачать сайт максимум в несколько тысяч страниц
• Ничего не выйдет, если для получения данных пользователю нужно делать что-то сложное (например, заполнять формы)
• Ничего не выйдет, если придётся вводить капчу
• Тяжело выгружать сайты с бесконечной лентой



Запрашиваем данные

Иногда интересные данные не выкладывают в открытый доступ, но их можно попросить у компаний или государства.

- заложите время на ответ — лучше несколько недель
- всегда нужно иметь план «Б» — данными могут не захотеть делиться

Дополнительные лайфхаки

Карты данных «Инфокультуры»

Чаты в телеграме:
Open Data Russia Chat — https://t.me/opendatarussiachat
Data Jounalism Russia Chat — https://t.me/ddjrus