Как искать данные?

Школа прикладного анализа данных Data-Diving

Басина Полина

аналитик Центра прикладного анализа больших данных

О чем поговорим?

Как определять какие данные нам нужны?

Каким требованиям должны отвечать данные?

Где и как искать необходимые данные?

Для чего нужны данные?

Шаг 1

Понять заказчика

Шаг 2

Сформулировать цель анализа

Определить какие нужны данные и сформулировать требования к ним

Шаг 1. Понять заказчика

- Для чего необходимо это исследование?
- Какие есть готовые решения?
- Как будут использоваться результаты?
- В чем заключается цель аналитики?

Шаг 2. Цель анализа данных

1. **Что мы изучаем?** (объект анализа)
2. **Какие свойства и/или характеристики**

   объекта исследования мы хотим изучить?

   Что хотим найти? (предмет анализа)
3. **Где мы хотим найти?**
4. **Зачем нам это нужно?**

Данные

Требования к данным

- Контент
- Форматы
- Источники
- Временные ограничения

Данные — это набор фиксированных сведений

Данные ≠ Информация

Обработка

Типы данных

Структурированные данные

- Упорядоченные, стандартный формат
- Таблицы, базы данных со связанными строками и столбцами (реляционные базы данных)

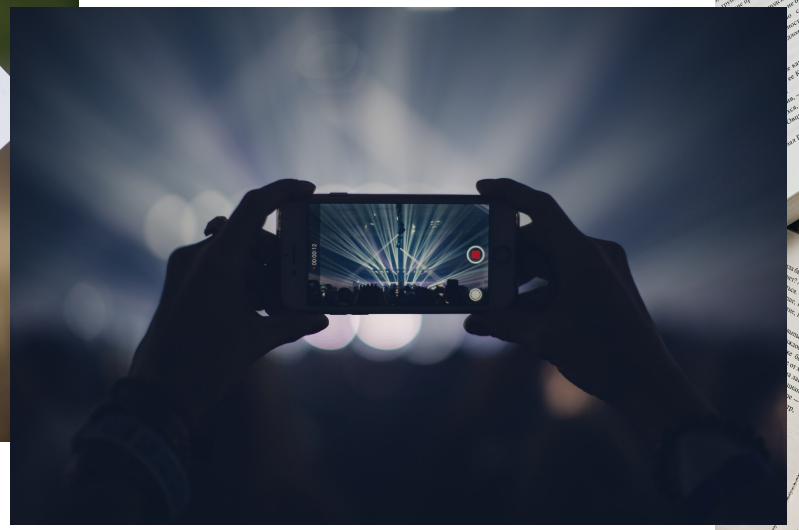
	A	B	C	D	E	F	G	H
1	*Заказчик*	*Товар*	*Категория*	*Дата*	*Менеджер продаж*	*Регион*	*Закупка*	*Продажа*
2	Рамстор	Ванильное небо	Печенья	01.01.2005	Петров	Восток	4032	10416
3	Рамстор	Попугай	Батончики	01.01.2005	Петров	Восток	1200	2436
4	Копейка	Сырные	Крекеры	02.01.2005	Григорьев	Центр	1449	3128
5	Копейка	Чесночные	Крекеры	03.01.2005	Григорьев	Центр	5916	6612
6	Метро	Картофельные чипсы	Крекеры	03.01.2005	Григорьев	Центр	363	517
7	Рамстор	Браво	Батончики	04.01.2005	Петров	Восток	920	2300
8	Ашан	Укроп	Крекеры	04.01.2005	Михайлов	Запад	1850	2500
9	Рамстор	Банановый Рай	Батончики	05.01.2005	Петров	Восток	9555	20839

https://www.planetaexcel.ru/upload/medialibrary/ca1/ca15044a17a56900a421068621de196c.gif

Типы данных

Неструктурированные данные

- Нет заранее заданной структуры, разнообразные формы (изображения, тексты, видео, аудио)



Типы данных

Полуструктурированные данные

- Структурированные + неструктурированные данные
- Нет строгой структуры, но содержат теги и различные маркеры, которые обеспечивают иерархическую структуру записей (xml, json)

JSON Example

```
{ "PONumber"             : 1600,
  "Reference"            : "ABULL-20140421",
  "Requestor"            : "Alexis Bull",
  "User"                 : "ABULL",
  "CostCenter"           : "A50",
  "ShippingInstructions" : { "name"   : "Alexis Bull",
                             "Address": { "street"  : "200 Sporting Green",
                                          "city"    : "South San Francisco",
                                          "state"   : "CA",
                                          "zipCode" : 99236,
                                          "country" : "United States of America" },
                             "Phone" : [ { "type" : "Office", "number" : "909-555-7307" },
                                         { "type" : "Mobile", "number" : "415-555-1234" } ] },
  "Special Instructions" : null,
  "AllowPartialShipment" : false,
  "LineItems"            : [ { "ItemNumber" : 1,
                               "Part"       : { "Description" : "One Magic Christmas",
                                                "UnitPrice"   : 19.95,
                                                "UPCCode"     : 13131092899 },
                               "Quantity"   : 9.0 },
                             { "ItemNumber" : 2,
                               "Part"       : { "Description" : "Lethal Weapon",
                                                "UnitPrice"   : 19.95,
                                                "UPCCode"     : 85391628927 },
                               "Quantity"   : 5.0 } ] }
```

Данные

Виды данных:

- собственные
- сторонние данные
- «потенциальные» данные

Способы получения данных:

- Сбор/получение первичной исходной информации
- Получение данных из вторичных источников
- Получение данных с использованием API
- Парсинг

Популярные источники данных:

- Социальные сети и мессенджеры
- Интернет-СМИ
- Общедоступные базы данных
- Контент сайтов
- Поисковые запросы
- Собственные данные организаций
- Открытые данные
- Большие пользовательские данные

Способы формирования маркерных слов

- Изучение предметной области
- Анализ литературы
- Экспертная оценка
- Словари
    - Словарь эмотивной лексики
      (беспокойство, страх, вера, радость, смелость и т.д.)
    - Словарь русского языка короновирусной эпохи
    - Прагматичные маркеры русской повседневной речи
    - Новое в русской лексике
    - Тематические словари
    - Словарь оценочных слов и выражений русского языка РуСентиЛекс
    - NLPub (корпусы, тексты, тезаурусы, словари)

- Использование специальных сервисов
    - Rusvectores (похожие слова и различные операции)
    - КартаСлов (синонимы, ассоциации и др.)
    - ЯндексВордстат
    - GoogleTrends

Всегда учитывайте источник и стиль речи

Способы формирования маркерных слов

Маркер	% валидных
{Ковивак}, {Ковивак,прививка}	100
{"Спутник лайт"}	100
{ЭпиВакКорона}	100
{Гам-Ковид-Вак}	100
{вакцинофилы}, {вакцинофобы}	100
{коронобесие}	100
{Querdenken 711}	100
{ковидоистерия}	100
{ковид-фейки}	100
{ковидофобы}, {ковидоскептики}, {ковид-агностики}	97
{антипрививочник}, {антиваксер,прививка}, {антиваксер,вакцина,прививка}, {антиваксер,антипрививочник}	95
{"большая фарма"}, {"большая фарма, вакцина"}	83
{пцр}	80
{ковид-диссиденты,прививка}, {ковид-диссиденты,пцр}, {ковид-диссиденты}, {вакцина,ковидодиссиденты},	80
{вакцина,прививка}	70
{sputnik,вакцина,прививка}, {sputnik,пцр}, {sputnik,вакцина}	66
{вакцина}	62
{вакцина,чипирование}, {вакцина,прививка,факцина}, {вакцина,пцр,факцина}, {вакцина,шмурдяк}	61
{лжепандемия}, {вакцина,лжепандемия}	43
{вакцина,прививка,пцр}	35
{прививка}	26
{чипирование}	15
{жижа},{жижа,прививка}, {вакцина,жижа}, {вакцина,жижа,шмурдяк}, {вакцина,жижа,ковидодиссиденты,прие	7
{кремлеботы}, {вакцина,кремлеботы}	5
{Большой брат}	5
{sputnik}	2
{шмурдяк}	1

Подведем итоги

- Для того, чтобы определить какие данные нам необходимо собрать нужно понять заказчика и перевести его запрос в цель исследования. Только после этого вы сможете сформировать релевантные требования к данным.
- Требования к данным заключаются в том, чтобы определить, что нам нужно, в каком формате, из какого источника и за какой период.
- В процессе анализа надо отталкиваться от решаемой задачи и подбирать под нее данные, а не брать имеющуюся информацию и придумывать, что из них можно «выжать».
- В качестве основных источников следует выделить – данные, которые предоставляет вам заказчик; данные, которые вы можете купить; данные, которые необходимо собрать.