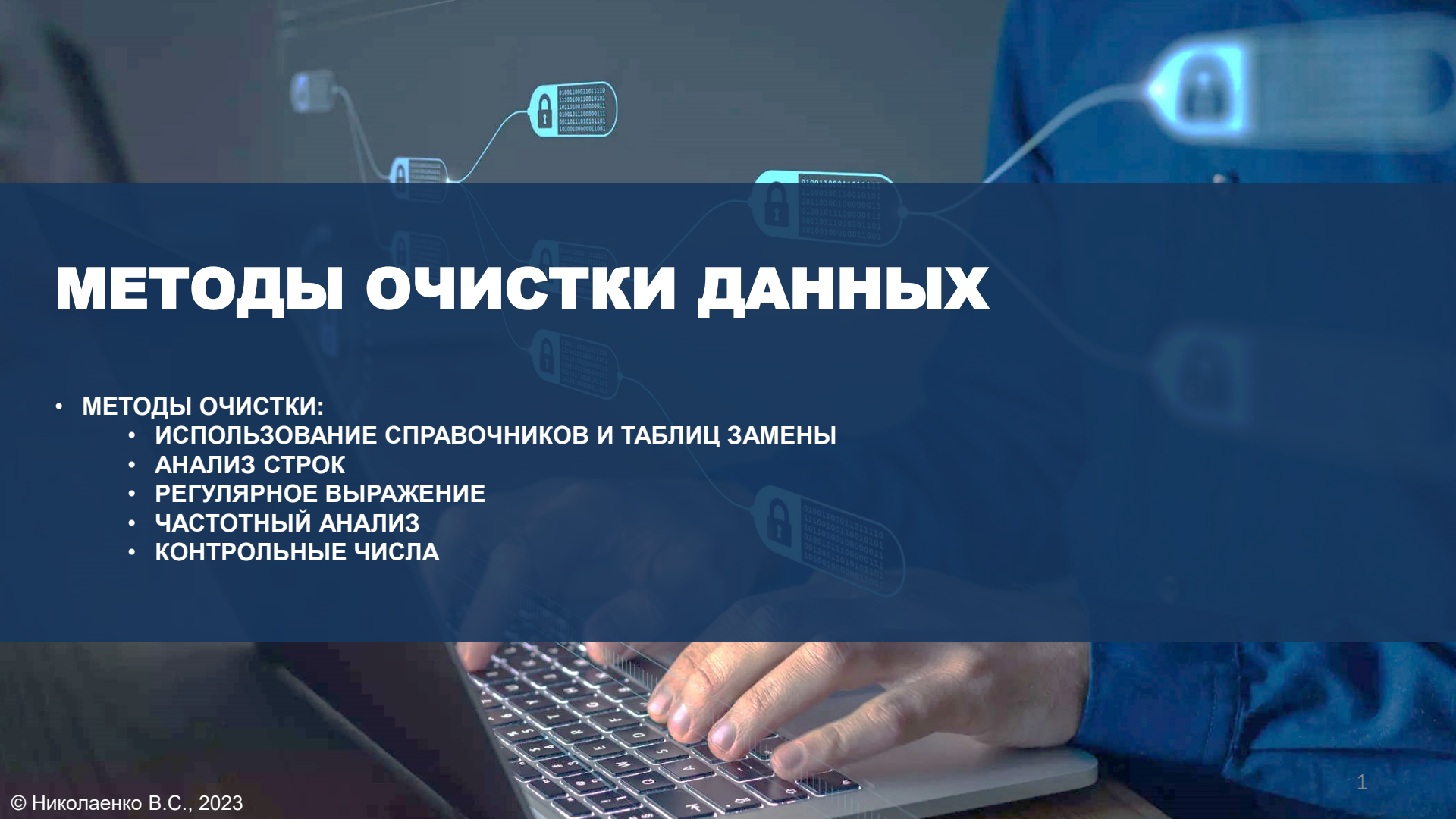


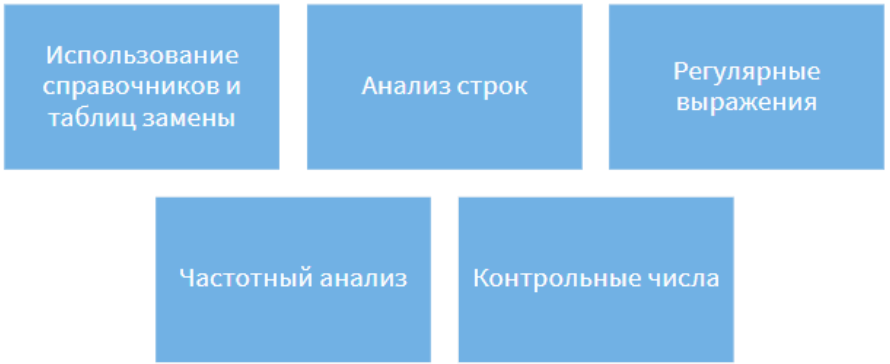
МЕТОДЫ ОЧИСТКИ ДАННЫХ

- МЕТОДЫ ОЧИСТКИ:
  - ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ
  - АНАЛИЗ СТРОК
  - РЕГУЛЯРНОЕ ВЫРАЖЕНИЕ
  - ЧАСТОТНЫЙ АНАЛИЗ
  - КОНТРОЛЬНЫЕ ЧИСЛА

МЕТОДЫ ОЧИСТКИ

Существуют различные методы очитки данных и повышения их качества. Использование того или иного метода во многом зависит от типов ошибок, которые встречаются в данных. Рассмотрим данные методы подробнее:

- Использование справочников и таблиц замены.
- Анализ строк.
- Регулярные выражения.
- Частотный анализ.
- Контрольные числа.



ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ

Данный метод можно использовать двумя способами: как для проверки данных на предмет загрязнения, так и непосредственно для очитки.

ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ: ПЕРВЫЙ СПОСОБ

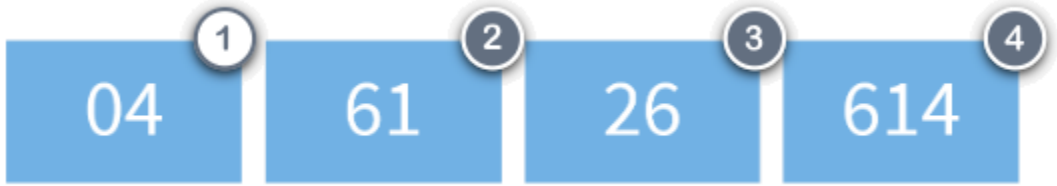
Прежде, чем приступить к обработке данных, необходимо подготовить нужные справочники: адресные справочники (например, в России есть официальные справочник ГАР – государственный адресный реестр), справочники имен и фамилий, телефонов и любых других значений. После этого значения полей исходных данных сравниваются с соответствующим справочником. Если значение найдено в справочнике, оно считается корректным, если нет – исходные данные загрязнены и нуждаются в очитке.

ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ: ВТОРОЙ СПОСОБ

Существуют различные классификаторы, которые могут помочь в заполнении пропусков и обогащения данных.

Классификатор – это справочник, состоящий из наименований объектов, классификационных группировок, на которые они разбиты по степени сходства, и идентифицирующих их кодов. Рассмотрим использование классификатора на примере классификатора **Банковских Идентификационных кодов – БИК**.

ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ: ВТОРОЙ СПОСОБ



Номер БИК служит для однозначной идентификации банка при проведении платежей. Это **девятизначное число, которое начинается с цифр 04** – кода РФ.

Следующие два символа указывают на территорию РФ по ОКАТО (Общероссийский классификатор объектов административно-территориального деления). В приведенном примере 61 – код Рязанской области по ОКАТО.

Далее идет условные номер подразделения по региону, 26 – условный номер подразделения расчетной сети Банка России по Рязанской области.

Последнее три цифры БИК номера должны совпадать с последними цифрами в корреспондентском счете банка. 614 – последние три цифры корреспондентского счета Рязанского отделения №8606 ПАО Сбербанк.

Таким образом, если исходные данные содержат информацию о банках, и в некоторых записях пропущен регион, но имеется БИК банка, пропущенные значения можно восстановить с помощью третьего и четвертого символов БИК.

ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ: ВТОРОЙ СПОСОБ

Общероссийский классификатор – справочник, который представляет собой систематизированный перечень объектов с указанием наименований и кодов объектов технико-экономической и социальной информации. Является официальным документом (рисунок).

Полное название: **Общероссийские классификаторы технико-экономической и социальной информации**. Названия таких классификаторов имеют аббревиатуры и обычно начинаются с ОК.

Требования к таким справочникам приведены в **Единой системе классификации и кодирования технико-экономической и социальной информации РФ** (ЕСКК ТЭСИ).

Полный их перечень содержится в **Общероссийском классификаторе информации об общероссийских классификаторах**.

ИСПОЛЬЗОВАНИЕ СПРАВОЧНИКОВ И ТАБЛИЦ ЗАМЕНЫ: ВТОРОЙ СПОСОБ

В России существует несколько справочников адресов, оператором официальных справочников является **Федеральная налоговая служба** (ФНС). Официальным форматом с2021 года является ГАР — Государственный адресный реестр, утвержден приказом ФНС России от 13.05.2020 № ЕД-7-6/329@.

Адреса в формате **ФИАС** (Федеральная информационная адресная система), который все еще повсеместно используется, ФНС перестает предоставлять с августа 2021. Некоторые также используют **КЛАДР** (классификатор адресов Российской Федерации), но он давно устарел.

Разница в том, что в ГАР адреса предоставляются в структуре муниципального. деления, тогда как ранее они предоставлялись в структуре административно-территориального деления.

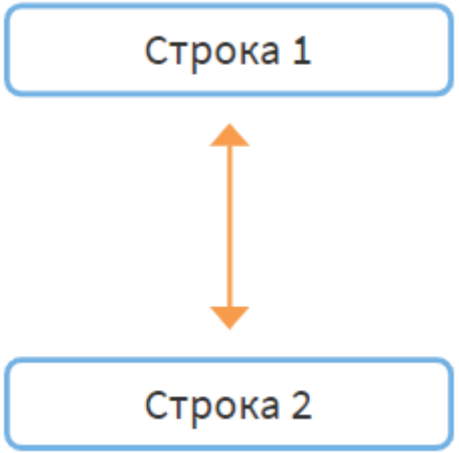
Для ознакомления и подготовки информационных систем пользователей ФИАС к работе с новым форматом на портале (https://fias.nalog.ru/) еженедельно размещаются файлы выгрузки адресной информации в формате ГАР.

АНАЛИЗ СТРОК

Суть данного метода заключается в подборе значения, максимально похожего на введенное некорректное значение.

Существуют различные алгоритмы анализа строк, которые также называют алгоритмами нечеткого поиска. Мы поговорим о достаточно распространенном методе – **расчете расстояния между строками**.

Для расчета расстояния между двумя строками широко применяется **расстояние Левенштейна**, или **редакционное расстояние**. Это мера разницы двух строк символов, которая определяется как минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа другим для превращения одной строки в другую. Расстояние названо по имени советского математика **Владимира Иосифовича Левенштейна**, который впервые упомянул подобную задачу в 1965 году при исследовании схожести последовательностей 0-1.



АНАЛИЗ СТРОК

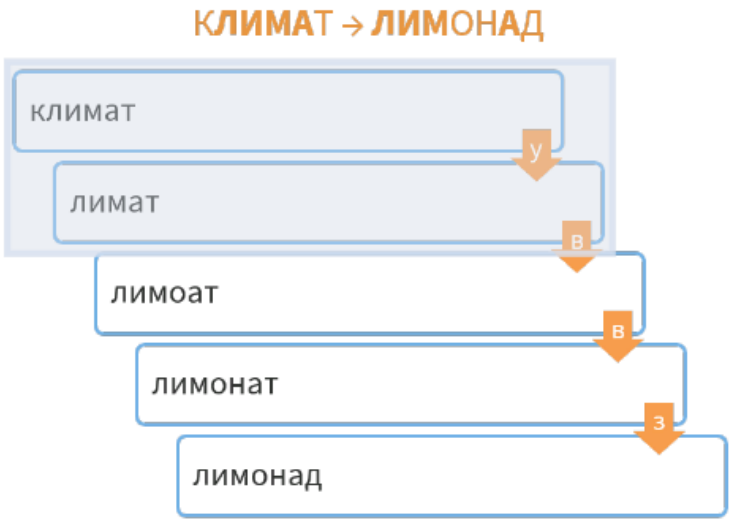
Рассмотрим пример. Нам необходимо получить из слова **климат** слово **лимонад**.

Частично слова совпадают, но позиции совпадающих символов различны. Уравняем наибольшую совпадающую часть: проведем операцию удалению(у) над исходным словом (рисунок).

Буква **а** нам тоже пригодиться, но она находится не на своей позиции. Проведем две операции ставки (в): добавим после **лим** на буквы **о** и **н**.

Легко заметить, что теперь искомое слово отличается от полученного нами всего одной буквой. Производим операцию замены(з) буквы **т** на **д** и получаем **лимонад**.

Таким образом, нам понадобиться четыре операции (одно удаление, две вставки и одна замена), чтобы превратить слово **климат** в слово **лимонад**. Значит, расстояние между этими словами равно четырем.



АНАЛИЗ СТРОК

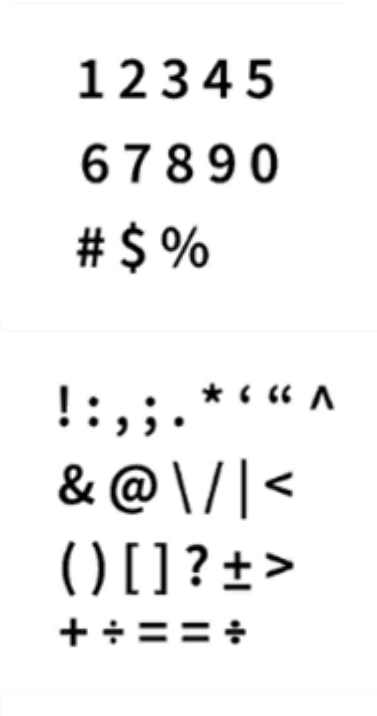
Существует модификация данного расстояния – **расстояние Дамерау-Левенштейна**. Фредерик Дамерау был видным исследователем по компьютерной обработке естественного языка. Он выяснил, что при наборе текста вручную 80% ошибок появляется из-за случайной перестановки двух соседних символов, или **транспозиции**. Например, слово **лимонад** можно набрать как **лиомнад**. Поэтому при вычислении расстояния Дамерау-Левенштейна к расчету количества операций вставки, замены и удаления добавляется количество операций **транспозиции**.

РЕГУЛЯРНОЕ ВЫРАЖЕНИЕ

Регулярное выражение, по сути, представляет собой шаблон строки. Такие шаблоны можно использовать во многих современных приложениях и языках программирования. С их помощью мы можем находить строки, которые соответствуют или не соответствуют нужному шаблону.

Регулярные выражения могут помочь решить целый ряд задач. Можно проверить данные на соответствие форматов, извлечь нужную часть строки, произвести замену части строки или всей строки, например, для исправления опечаток, привести информацию к единому виду.

Это универсальное средство, которое позволяет обрабатывать различные виды ошибок в данных. Поэтому на нем мы более подробно остановимся в отдельном блоке.



ЧАСТОТНЫЙ АНАЛИЗ

Метод основан на анализе **частоты появления** значений или их комбинаций. На основе частоты можно определять корректные значения: чем чаще одно и то же слово или название появляется среди значений поля, тем больше вероятность, что оно корректное.

Например, есть ряд компаний, которые приобретают у компании-поставщика продукцию. Внутри компании-поставщика нет списка контрагентов в едином виде, он разрознен, его части находятся у нескольких менеджеров. При этом разные сотрудники могут указывать название одной и той же компании по-разному: с организационно-правовой формой и без, полностью или сокращенно и т.д. Если собрать такой список воедино, одна и та же компания может встретиться несколько раз. Для стандартизации можно подсчитать частоту появления одинаковых значений. Наиболее часто встречающиеся значения будут считаться корректными.

Под **организационно-правовой формой** (ОПФ) понимается способ закрепления (формирования) и использования организацией имущества и вытекающие из этого ее правовое положение и цели предпринимательской деятельности.

ЧАСТОТНЫЙ АНАЛИЗ

0К028-2012. Общероссийский классификатор организационно-правовых форм (утв. Приказом Росстандарта от 16.10.2012 № 505-ст).

Для дальнейшей обработки можно составить справочник из корректных значений в едином формате и сравнить с ним исходный список на схожесть названий. Таким образом из списка будут исключены дубликаты, а корректные названия будут стандартизированы.

Можно использовать комбинации значений. Возьмем поля **Имя** и **Пол**. В исходных данных имя **Алексей** 9 раз. встречается с полом **Женский** и 48 раз с полом **Мужской**. На основе найденной частоты в девяти некорректных случаях мы можем исправить ошибку, заменив значение на **Мужской**. Аналогичным образом можно комбинировать и другие поля: ОПФ предприятия с названием и адресом, паспортные данные с ФИО и т.д.

Имя	Количество совпадений	
	Мужской	Женский
Алексей	**48**	9
Евгения	20	**50**
Оксана	7	**84**
Андрей	**99**	3
Георгий	**78**	5

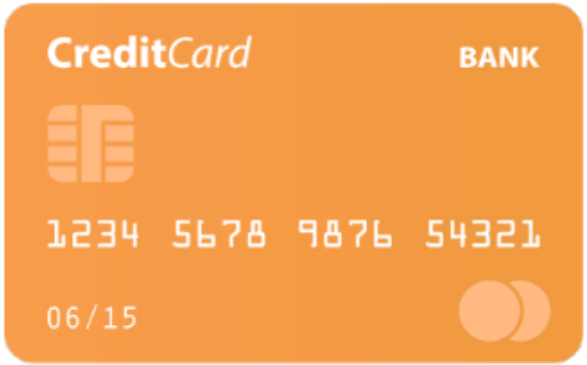
Имя	Пол
Алексей	Мужской
Евгения	Женский
Оксана	Женский
Андрей	Мужской
Георгий	Мужской

КОНТРОЛЬНЫЕ ЧИСЛА

Данный метод подходит для числовых данных, в Росси это: **номера банковских карт и счетов, ИНН, ОГРН, страховые номера индивидуального лицевого счета (СНИЛС)** и т.д. Внутри указанных номеров (чаще всего в конце) содержится число (цифра), с помощью которого можно проверить их правильность: **контрольное число**.

Кроме проверки контрольное число, как правило, позволяет восстановить одну потерянную цифру номера, при условии, что известна позиция этой цифры, и все остальные цифры в номере правильные. При неизвестной позиции один из выходов — перебрать все варианты и передать их на оценку эксперту, который сможет выбрать наиболее правдоподобное значение.

Для разных номеров применяются различные алгоритмы расчета контрольных чисел. Как правило, числа формируются с помощью математических операций над остальными цифрами, входящими в номер, например, это может быть последняя цифра суммы остальных.



КОНТРОЛЬНЫЕ ЧИСЛА

Для примера рассмотрим алгоритм формирования контрольного числа для СНИЛС. Страховой номер индивидуального лицевого счета имеет вид: **XXX-XXX-XXXYY**. Контрольным числом здесь являются две последние цифры. Проверка контрольного числа возможна только для номеров больше, чем **001-001-998**.

Пусть у нас есть номер **121-388-227 38**. Контрольным числом в нем является **38**.

Каждая цифра номера (за исключением самой контрольной части) умножается на номер ее позиции, рассчитанный с конца. Тогда для нашего номера **121-388-227 38**:

1×9=9
2×8=16
1×7=7
3×6=18
8×5=40
8×4=32
2×3=6
2×2=4
7×1=7

Полученные произведения складываются: **9+16+7+18+ 40 + 32 +6+4+7= 139**.

Если полученная сумма меньше **100**, то сама сумма и будет контрольным числом. Мы получили сумму **139**. Продолжим расчеты.

Если сумма равна **100** или **101**, контрольным числом будет **00**. В нашем случае данный вариант также не подходит.

Если сумма превышает **101**, то она делится по остатку на **101**, и контрольное число будет равно остатку от деления по аналогии с пунктами 3 и 4. Это как раз наш случай. Проводим операцию деления на **101** и получаем остаток **38**. Контрольная сумма совпадает с исходным номером, значит, номер можно считать корректным.