**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

**“Большие данные и распределенная цифровая платформа”**

**ОТЧЁТ**

**По лабораторной работе №2**

**По дисциплине “Алгоритмы и структуры данных”**

**На тему “Обезличивание датасета”**

**Вариант №1**

**Студент гр. 23Б16-пу**

**Пушкарев Н.П.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2024 г.**

[**Цель работы 3**](#_gjdgxs)

[**Описание задачи 3**](#_30j0zll)

[**Теоретическая часть 4**](#)

[**Основные шаги программы 7**](#_1fob9te)

[**Описание программы 7**](#_3znysh7)

[**Блок-схема программы 9**](#_tyjcwt)

[**Рекомендации пользователя 10**](#_1t3h5sf)

[**Рекомендации программиста 10**](#_4d34og8)

[**Контрольный пример 10**](#_2s8eyo1)

[**Исследование 12**](#)

[**Вывод 12**](#_17dp8vu)

# Цель работы

Целью лабораторной работы является написание програмного обеспечения, для обезличивания датасета, состоящего из данных о магазинах в Санкт-Петербурге: название магазина, координаты, время покупки, категории товаров, бренды, количество товаров, стоимость и данные о платежных картах; и расчёта К-анонимити.

# Описание задачи

⦁ Программа должна считывать входной файл (Итоговый файл 1-ой лабораторной работы)

⦁ Программа делиться по функционалу

⦁ Обезличивание входного датасета.

⦁ Высчитывание К-анонимити входного датасета.

⦁ У пользователя есть возможность указывать Квази-идентификаторы в программе.

⦁ На выбор, производим обезличивание дата сета по данным методам:

1. Локальное обобщение
2. Агрегация
3. Возмущение
4. Микро-агрегация
5. Перемешивание
6. Создание псевдонимов
7. Маскеризация
8. Локальное подавление
9. Удаление атрибутов
10. Метод декомпозиции

⦁ Используя метод К-анонимити необходимо рассчитать К для обезличенного набора.

⦁ Вывести 5 "плохих" значений К-анонимити (если их меньше, то все возможные). Данные переменной К вывести в процентах из всего набора.идентификаторам. Вывести уникальные строки, если переменная К=1.

⦁ Вывести приемлемого значения К-анонимити для набора данных:

1. до 51000 записей - К>=10
2. до 105000 записей - К>=7
3. до 260000 записей - К>=5

⦁ Оценить полезность данных, путем сравнения обезличенного набора с исходным набором.

# Теоретическая часть

К-анонимити — это характеристика датасета, при которой каждая строка данных имеет не менее чем К - 1 других строк с идентичными значениями. Это гарантирует, что идентификация конкретных субъектов данных становится практически невозможной, в то время как полезность данных сохраняется.

Методы обезличивания:

1. Локальное обобщение

Данная техника уменьшает специфичность атрибутов за счет подмены точного значения на обобщенное. Применяется для геоданных, временных интервалов и финансовых параметров.

- Варианты применения:

- перекодирование сверху-вниз или снизу-вверх;

- округление значений.

- Характеристики:

- Увеличивает размеры классов, снижая риск повторной идентификации.

- Эффективен для профилей злоумышленников типа "прокурор", "журналист".

- Потеря информации рассчитывается на основе метрик энтропии.

2. Агрегация

Построение обобщений на основе групповых операций для выделенных атрибутов. Используется для статистических целей.

- Характеристики:

- Снижает риск повторной идентификации путем использования агрегирующих функций.

- Эффективен для всех профилей злоумышленников.

- Потеря информации рассчитывается на основе метрик энтропии.

3. Возмущение

Внесение шума в данные для сохранения основных статистических закономерностей при снижении точности. Применимо для бинарных данных, цифровых объектов и статистических наборов.

- Характеристики:

- Снижает риск повторной идентификации и исключает квази-идентификаторы.

- Эффективен для профилей злоумышленников типа "прокурор".

- Потеря информации оценивается по мерам взаимной полезности.

4. Микро-агрегация

Формирование групп записей с указанием диапазона значений вместо точных данных. Применяется для возраста, геолокаций и финансовых данных.

- Характеристики:

- Увеличивает классы эквивалентности, снижая риск повторной идентификации.

- Неэффективен для малых наборов данных.

- Эффективен для профиля "маркетолог".

5. Перемешивание

Перемешивание аналогичных записей при сохранении значимости всего набора. Используется для временных интервалов в совокупности с другими методами.

- Характеристики:

- Снижает риск повторной идентификации.

- Эффективен для профиля "журналист".

6. Создание псевдонимов

Замена идентификаторов на условные обозначения или цифровые ключи. Применяется к частным данным.

-Характеристики:

- Эффективен для атак типа "связывание".

- Потери информации незначительны.

7. Маскеризация

Замена части записи заполнителями по шаблону. Применимо для номеров паспортов, телефонов и карт.

-Характеристики:

- Эффективен для всех видов атак.

- Потери информации минимальны.

8. Локальное подавление

Удаление или перекодирование редких записей. Применяется для медицинской и биометрической информации.

- Характеристики:

- Снижает риск повторной идентификации, но ведет к существенной потере данных.

9.Удаление аргументов

Удаление чувствительного контента без добавления замен. Применяется к прямым идентификаторам.

-Характеристики:

- Эффективен для всех видов атак.

- Потеря информации максимальна.

10. Метод декомпозиции

Разделение атрибутов на несколько подмножеств с раздельным хранением. Не является собственно методом обезличивания, но поддерживает его.

- Характеристики:

- Потеря данных рассчитывается на основе классификационных метрик.

Виды идентификаторов:

1. Прямые идентификаторы — данные, которые однозначно идентифицируют конкретное лицо и подлежат обязательной анонимизации.

2. Квази-идентификаторы — данные, которые в комбинации с другими могут позволить установить личность.

3.Чувствительные идентификаторы — данные, утечка которых может причинить значительный вред.

# Основные шаги программы

1. Запуск программы anonimization.py
2. Пользователь выбирает квази-идентификаторы
3. Когда пользователь выбирает обезличивание, датасет обезличивается
4. Когда пользователь выбирает подсчет К-анонимити, сохраняются все строки с низким К-анонимити
5. удаляются строки, относящиеся к маленьким классам эквивалентности, но не более 5% от всего датасета
6. выводиться топ 5 худших К с процентами от оставшегося датасета

# Описание программы

файл anonimization.py

функции:

Таблица 1. Описание функций

| Название функции | Описание функции |
| --- | --- |
| calculate\_k\_anonymity | Считает повторения строк. Возвращает минимальное значение (К-анонимность), максимальное и количество повторений каждой уникальной строки. |
| mask\_brand | Бренд заменятеся на один из разделов товаров. |
| mask\_card | Оставляет только платежную систему. |
| mask\_categories | Заменяет товары на разделы. |
| mask\_coordinates | От координат оставляет только точную широту и долготу (округляет до целого). Удаляет время покупки. |
| mask\_num | Количество товаров заменяется на диапазоны. |
| mask\_price | Цена заменяется на диапазоны. |
| mask\_shop | Магазины заменяются на разделы товаров. |
| class AnonymizationApp | |
| \_\_init\_\_ | Инициирующая функция приложения на Qt |
| initUI | Графическая часть приложения. |
| executeAction | Выбор действия в зависимости от выполняемой приложением функции. При обезличивании вызываются функции mask и создается anonimization.xlsx, при подсчете k-anonimity считается начальное k-anonimity, удаляются строки (не более 5% датасета), сохраняются уникальные строки и снова считается k-anonimity. |

# 

# Блок-схема программы

# 

Рис1. Блок схема основной программы и подпрограммы

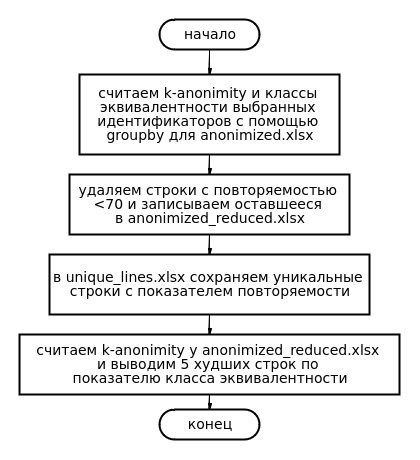


Рис2. Блок-схема второй подпрограммы

# Рекомендации пользователя

* Имейте в папке laba\_2/ файл output.xlsx с датасетом
* Для запуска программы, откройте файл anonimization.py в pycharm и нажмите Shift+F10
* в окне приложения выберите необходимые квази-идентификаторы
* дождитесь окончания работы программы. в этот момент будет выведено топ 5 худших К поле обезличивания всего датасета

# Рекомендации программиста

* Поддерживайте актуальность библиотек pandas и python (не ниже 3.11)
* Имейте библиотеку PyQt5.Widgets.

# Контрольный пример

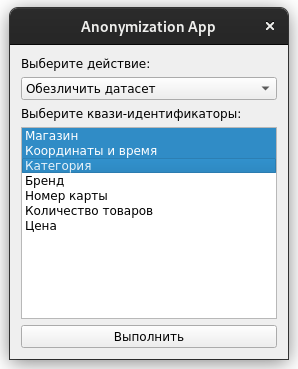
Запускаем файл anonimization.py; выбираем квази-идентификаторы, по которым хотим обезличить датасет.

Рис3. Пример ввода данных для обезличивания

Ждём окончания работы программы, по итогу будет выведено done и получен файл anonimized.xlsx.

Запускаем вычисление k-anonimity, также выбрав квази-идентификаторы.

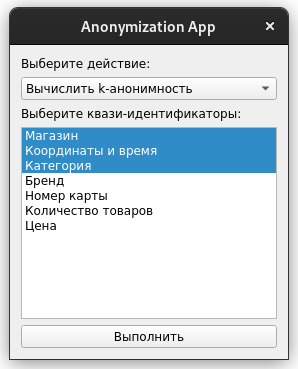


Рис4. Пример ввода данных для обезличивания

Далее будет посчитано начальное К-анонимити, будут удалены строки с низким показателем и записаны в unique\_lines.xlsx уникальные строки. Обновленным датасетом будет anonimized\_reduced.xlsx. У него снова пересчитается k-anonimity и будет выведено 5 худших значений.

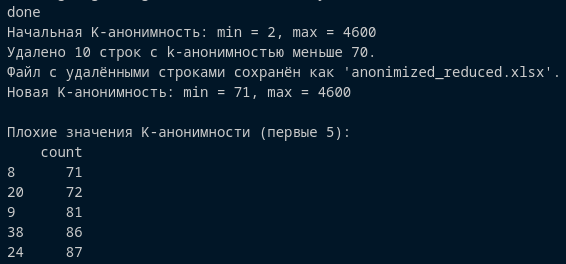
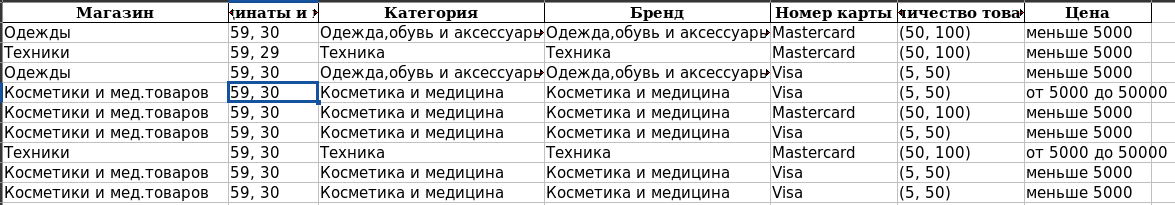


Рис5. Пример вывода после выполнения вычисления k-anonimity

Рис6. Пример обезличенного датасета

# Исследование

Из всех разрешённых для использования методов обезличивания квази-идентификаторов были использованы:

1. Локальное подавления, так как этот метод позволяет избавиться от редких строк и сильно поднять К-анонимность.
2. Возмущение данных для округления координат.
3. Создание псевдонимов для карт, названий магазинов, брендов и товаров.
4. Локальное обобщение, для создания диапазонов данных. Таким образом были обезличены цены и количество товаров.

После обезличивания всего датасета К-анонимити получается в районе 70, что позволяет утверждать, что датасет был обезличен хорошо, при этом была сохранена понятность и полезность датасета для “конечного пользователя”. Без проблем можно проводить аналитику и подсчитывать статистику по покупкам различных товаров в магазинах разной направленности в Санкт-Петербурге.

# Вывод

В рамках работы было написано программное обеспечение, которое обезличивает датасет и считает k-anonimity в разные моменты обезличивания. При датасете в 50000 строк k-anonimity получается около 70, что обеспечивает высокую защищённость датасета, при этом этот датасет остается достаточно полезным, целым и понятным.