**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

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

**ОТЧЕТ**

**по лабораторной работе №2**

**по дисциплине «Алгоритмы и структуры данных»**

**на тему «Обезличивание данных.»**

**Студент гр. 23Б15-пу**

**Беляева А.П.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2024 г.**

**Оглавление**

Цель работы...................................................................................................................3

Описание задачи............................................................................................................3

Теоретическая часть......................................................................................................4

Основные шаги программы...........................................................................................5

Блок-схема программы..................................................................................................6

Описание программы.....................................................................................................8

Рекомендации пользователя........................................................................................9

Рекомендации для программиста................................................................................9

Исходный код программы..............................................................................................9

Контрольный пример...................................................................................................10

Вывод.............................................................................................................................11

Источники......................................................................................................................11

**Цель работы**

Цель данной лабораторной работы заключается в изучении и применении методов обезличивания данных для обеспечения конфиденциальности личной информации, а также в проверке К-анонимности обработанного набора данных. В процессе выполнения работы требуется разработать программу для выполнения обезличивания и анализа результатов с точки зрения конфиденциальности.

**Описание задачи (формализация задачи)**

Основной задачей лабораторной работы является разработка программы, которая будет:

* Считывать входной файл (результат работы первой лабораторной);
* Осуществлять обезличивание данных с использованием одного или нескольких методов (на выбор);
* Оценивать уровень К-анонимности для каждого обезличенного набора данных;
* Выводить значения К-анонимности для самых уязвимых строк;
* Сравнивать полезность исходного и обезличенного набора данных.

**Теоретическая часть**

**Обезличивание данных** (анонимизация) представляет собой процесс обработки данных с целью удаления или маскировки персональной информации для защиты конфиденциальности. Существуют различные методы обезличивания, которые могут быть использованы для разных типов данных.

**К-анонимность** — это параметр, который определяет минимальное количество записей в наборе данных, которые неотличимы друг от друга по значениям квази-идентификаторов. К-анонимность обеспечивает защиту от рекордационной атаки, при которой злоумышленник может идентифицировать человека, сопоставив информацию в базе данных с внешними источниками.

Методы обезличивания, которые можно использовать:

1. **Локальное обобщение** — замена конкретных значений на более общие категории;
2. **Агрегация** — группировка данных по определенным категориям;
3. **Возмущение** — добавление шума к данным;
4. **Микро-агрегация** — объединение данных в группы с последующим усреднением значений;
5. **Перемешивание** — случайное переставление значений;
6. **Создание псевдонимов** — замена идентификаторов вымышленными именами;
7. **Маскеризация** — скрытие части данных;
8. **Локальное подавление** — удаление отдельных данных для повышения конфиденциальности;
9. **Удаление атрибутов** — полное удаление столбцов с квази-идентификаторами;
10. **Метод декомпозиции** — разбиение данных на отдельные таблицы.

**Основные шаги программы**

* **Загрузка входного файла.** Программа начинает с чтения CSV-файла, который содержит набор данных.
* **Выбор квази-идентификаторов.** Пользователь указывает атрибуты, которые должны быть защищены.
* **Применение обезличивания.** Программа обрабатывает набор данных выбранным методом.
* **Рассчет К-анонимности.** Программа оценивает уровень К-анонимности для обезличенного набора.
* **Повышение** **К-анонимности. Программа удаляет строки, число повторений которых меньше допустимой К-анонимности.**
* **Вывод результатов.** Программа выводит пять "плохих" значений К-анонимности и процент записей с этими значениями.
* **Оценка полезности.** Программа сравнивает полезность данных до и после обезличивания, оценивая количество уникальных записей и степень сохранения исходных характеристик.

# Блок схема программы

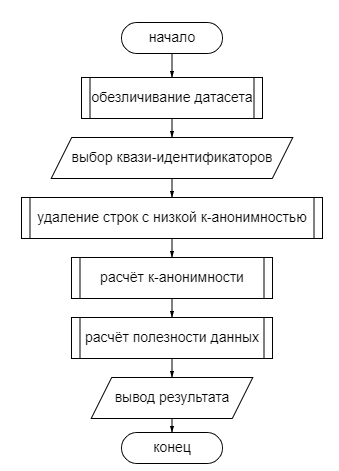
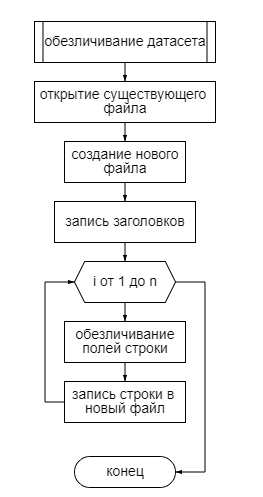
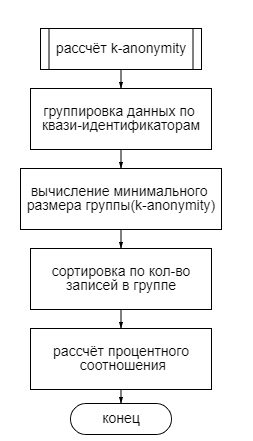


Рис 1. Блок-схема программы



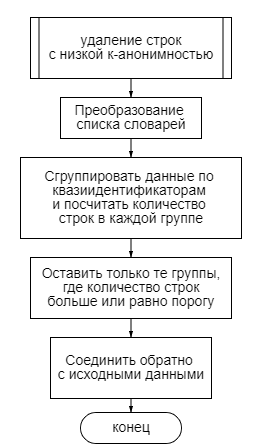
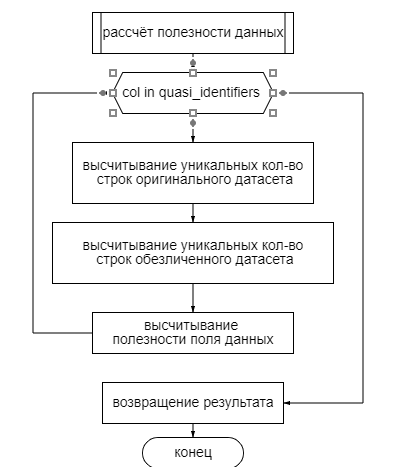


Рис 2. Блок-схема подпрограмм

# Описание программы

Программа разработана на языке Python с использованием библиотек pandas, csv и io для работы с данными. Пользователю предоставляется возможность выбора квази-идентификаторов, после чего программа автоматически применяет выбранные действия и высчитываются результаты.

Таблица 1. main.py

|  |  |  |
| --- | --- | --- |
| Функция | Описание | Возвращаемое значение |
| de\_name | Обезличивание поля ФИО | None |
| de\_passport | Обезличивание поля Паспортные данные | None |
| de\_marsh | Обезличивание полей Откуда, Куда | None |
| de\_train\_type | Обезличивание поля Рейс | None |
| de\_wagon | Обезличивание поля Вагон и место | None |
| de\_price | Обезличивание поля Стоимость | None |
| de\_card | Обезличивание поля Карта оплаты | None |
| de\_date | Обезличивание полей Дата отъезда и Дата приезда | None |
| evaluate\_data\_utility | Сравнение полезности | pandas.core.frame.DataFrame |
| calculate\_k\_anonymity\_from\_dict | Расчёт k-anonymity | tuple |
| remove\_low\_frequency\_rows | Удаление уникальных строк | list |

# Рекомендации пользователя

Для использования программы необходимо открыть файл main.py в среде разработки, нажать на кнопку запуска программы и в показавшемся окошке указать квази-идентификаторы. Рекомендуется использовать системные требования с достаточно большим объёмом памяти, так как датасет может достигать больших размеров (50 000 строк и более).

# Рекомендации программиста

Для запуска программы убедитесь, что у вас установлен Python и необходимые библиотеки, такие как pandas, csv и io. Следует поддерживать актуальные версии этих библиотек. Код можно запустить из среды разработки, либо через консоль.

# Исходный код программы

**https://github.com/hysterria/Depersonalization.git**

# Контрольный пример

# 1. Запуск программы

# Для запуска программы используйте файл **main.py**.

2. Ввод квази-идентификаторов.

После запуска программы пользователю предложено ввести квази-идентификаторы(Рис. 2). Они определяют по каким полям высчитывается k-anonymity

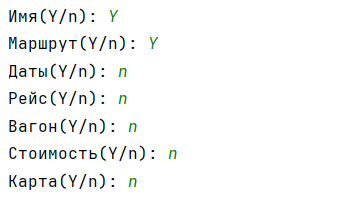


Рис 2. пример ввода

3. Обезличивание датасета

Далее программа приступает к формированию обезличенного датасета, используя введенные параметры, затем сгенерированный датасет сохраняется в new\_new\_train\_tickets\_dataset.cvs.

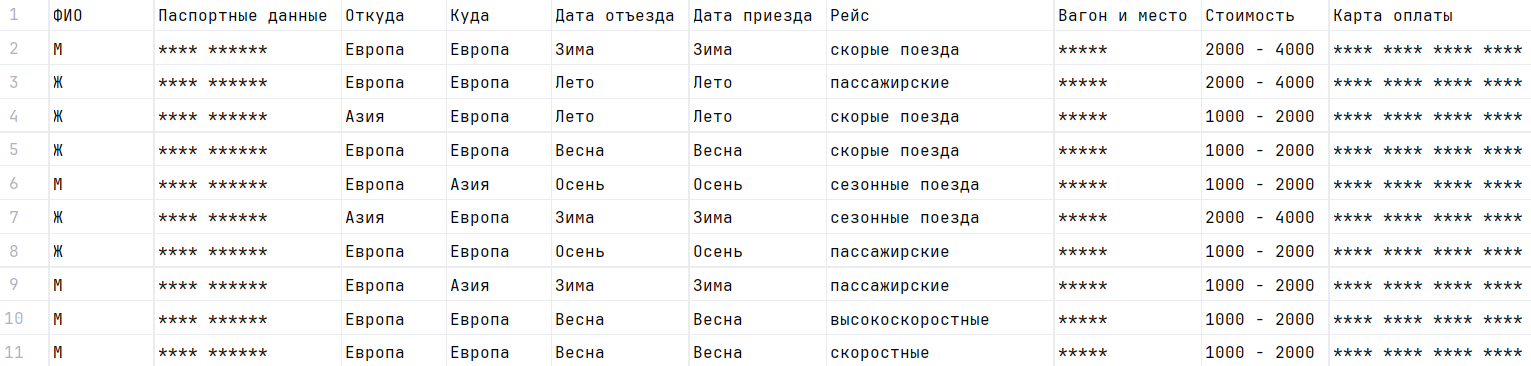


Рис 7. пример обезличенного датасета

# Вывод

В ходе выполнения лабораторной работы была разработана программа для обезличивания данных с использованием оценки уровня К-анонимности и нескольких методов обезличивания. Программа позволяет пользователю настраивать параметры обезличивания и оценивать качество защиты данных.

# Источники

1. Pandas – pandas documentation // Pandas URL:  
    <https://pandas.pydata.org/>.
2. CSV – csv documentation // CSV URL: <https://docs.python.org/3/library/csv.html>
3. io - io documentation // IO URL:

<https://docs.python.org/3/library/io.html>

1. k-anonymity information - <https://en.wikipedia.org/wiki/K-anonymity>