**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**КАФЕДРА КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ И МНОГОПРОЦЕССОРНЫХ СИСТЕМ**

**ОТЧЕТ**

**по лабораторной работе**

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

**на тему “Обезличивание данных”**

|  |  |
| --- | --- |
| Студент | Михайлов И. С. |
| Преподаватель | Щеголева Н. Л. |

Санкт-Петербург

2022 г.

[Цель работы 2](#_Toc2053149552)

[Задача 3](#_Toc783382027)

[Теоретическая часть 3](#_Toc2135864452)

[Алгоритм метода 4](#_Toc879095325)

[Диаграмма классов 4](#_Toc1155500797)

[Карта кода 4](#_Toc2060043948)

[Описание классов 5](#_Toc205159044)

[Рекомендации пользователя 5](#_Toc1126372176)

[Рекомендации программиста 5](#_Toc1344947604)

[Контрольный пример 5](#_Toc1750627196)

[Вывод 6](#_Toc1098356228)

[Ссылки 7](#_Toc1987858249)

# Цель работы

Обезличить входной датасет.

# Задача

Обезличить базу данных из 2-й лабораторной работы, применяя различные способы обезличивания. Вычислить k-anonymity получившегося датасета.

# Теоретическая часть

К-анонимити — свойство таблицы с обезличенными данными, которое является минимальным количеством повторяющихся строк. Если к-анонимити равно 1 — это значит, что в таблице есть уникальная строка, по которой можно вычислить человека, описываемого этой строкой. Чем больше этот параметр — тем качественнее обезличены данные.

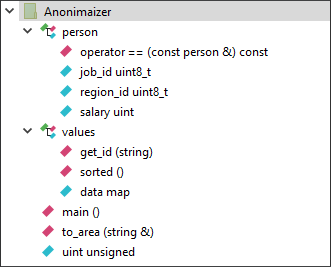
Для обезличивания были использованы следующие методы:

* **Удаление атрибутов.** Под данной техникой понимается удаление чувствительного контента без добавления замен. Метод используется для удаления прямых идентификаторов, а также удаления избыточных квази-идентификаторов. (Был применен для ФИО, номера телефона).
* **Локальное обобщение.** В рамках данной техники предполагается уменьшение специфичности атрибута за счет подмены точного значения атрибута его общим значением. Применяется для гео-данных, временных интервалов, финансовых параметров. (Был применен для адреса).
* **Микро-агрегация.** Техника предполагает формирование групп записей, для которых вместо точных значений указывается диапазон. Метод рекомендуется к применению для параметров возраста, географических локаций, финансовых транзакций, биометрических данных. Один из самых широко применяемых методов. (Был применен для зарплаты).
* **Локальное подавление.** Техника предполагает удаление или перекодирование относительно редких записей данных. В отличии от метода удаления атрибутов нацелен на удаление строк или значений для заданных атрибутов в выделенных строках. Наибольшее применение имеет для медицинской и биометрической информации, а также больших наборов статистических данных с аномалиями. (Был применен для строк со слишком маленьким к-анонимити).

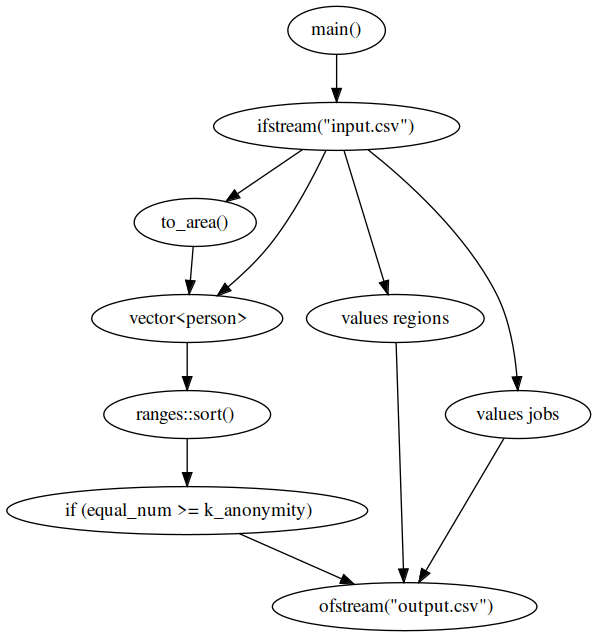
# Алгоритм метода

* В массив заносятся строки исходной таблицы, обезличенные следующим образом:
  + ФИО и телефон заменяются символом *'\*'*
  + Полный адрес заменяется названием района
  + Должность остается без изменений
  + Точное значение зарплаты заменяется диапазоном в 20000
* Массив сортируется по всем полям — это нужно, чтобы эквивалентные строки стояли друг за другом
* Выполняется проход по отсортированному массиву, для каждой группы эквивалентных строк подсчитывается их количество. Если оно:
  + Меньше требуемого к-анонимити, то эти строки игнорируются
  + Больше или равно требуемому к-анонимити, то строки записываются в выходной файл

# Диаграмма классов

Рис. 1. Диаграмма классов

# Карта кода

Рис. 2. Карта кода

# Описание классов

В программе для работы с данными используется структура person и класс values. Ниже представлено их описание.

|  |  |  |
| --- | --- | --- |
| **Имя** | **Наследование** | **Описание** |
| person | - | Представление обезличенной строки таблицы |
| values | - | Заменяет строку на ид. Используется с полями адрес и должность для ускорения и упрощения их обработки. Метод sorted() возвращает массив, из которого по ид можно получить исходную строку. |

Табл. 1

# Рекомендации пользователя

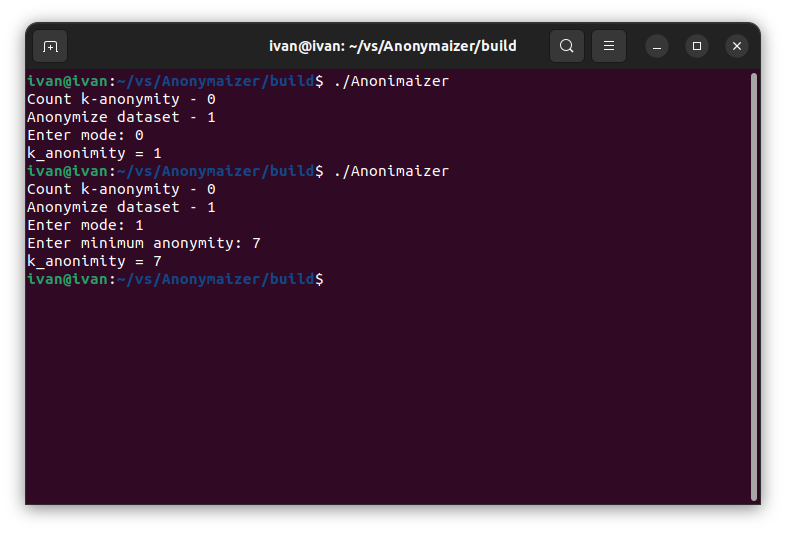
В папке с программой должен быть входной файл - input.csv. После запуска программа спросит желаемый к-анонимити. Чем больше к-анонимити - тем больше строк будет исключено.

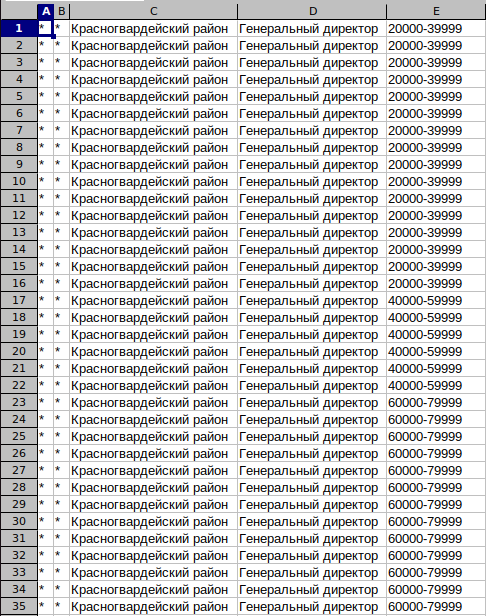
# Рекомендации программиста

Программа собирается с помощью cmake версии не ниже 3.8 и использует с++ 20-й версии. Никакие сторонние библиотеки не нужны. Сборка проверялась на компиляторе gcc 11-й версии.

# Контрольный пример

В первом случае программа подсчитывает k-anonymity, а во втором анонимизирует таблицу.

Рис. 3. Пример выполнения программы

Рис. 4. Полученный датасет

# Вывод

Были исследованы и реализованы методы обезличивания данных. С помощью этой реализации было произведено обезличивание датасета из прошлого задания и определение к-анонимити получившегося датасета.

Обезличенные данные содержат район, должность и диапазон зарплаты. По этим данным можно составить следующую статистику:

* Средняя зарплата по профессиям
* Средняя зарплата по районам
* Количество фирм по районам
* Среднее количество в фирме охранников, секретарей и т. д.

# Ссылки

* [futuris-tech/Anonymaizer (github.com)](https://github.com/futuris-tech/Anonymaizer)