**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**Факультет прикладной математики – процессов управления**

**отчет**

**по лабораторной работе**

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

**на тему «Обезличивание данных»**

**1 вариант**

|  |  |  |
| --- | --- | --- |
| Студент гр. 22.Б15 |  | Агишев А.Б. |
| Преподаватель |  | Дик А.Г. |

**Санкт-Петербург**

**2023 г.**

**Содержание**

[1. Цель работы 2](#_Toc151413474)

[2. Задача 2](#_Toc151413475)

[3. Теоретическая часть 2](#_Toc151413476)

[4. Описание алгоритма 4](#_Toc151413477)

[5. Описание программы 7](#_Toc151413478)

[5.1 Описание классов 7](#_Toc151413479)

[5.2. Описание функций 7](#_Toc151413480)

[5.3. Описание переменных 7](#_Toc151413481)

[6. Рекомендации пользователю 9](#_Toc151413482)

[7. Рекомендации программисту 9](#_Toc151413483)

[8. Контрольный пример 10](#_Toc151413484)

[9. Заключение 12](#_Toc151413485)

# **Цель работы**

Разработать алгоритм и программу, которая обезличивает набор синтетических данных.

# **Задача**

Приобрести и закрепить знания в области обезличивания данных. Обезличить набор синтетических данных, которые обладают следующими наборами свойств: ФИО, номер телефона, адрес работы, должность, З/П. Рассчитать k-anonymity для обезличенного набора. Достичь приемлемого значения k-anonymity.

# **Теоретическая часть**

*Обезличивание данных* — это процесс обработки персональных данных с целью удаления или замены всех прямых или косвенных идентификационных характеристик, которые могут связать данные с конкретным человеком.

*Методы обезличивания* — техники, применяемый к персональным данным с целью снижения риска идентификации субъекта персональных данных. В результате применения соответствующего метода из набора данных удаляется идентифицирующая личность человека информация. Оценка риска обезличенных данных производится для каждого набора данных и каждого метода. Все методы являются параметризуемыми, то есть содержат настраиваемые характеристики (такие как набор квази-идентификаторов), определяющие конкретный алгоритм в отношении конкретного набора данных в заданных условиях (сценариях и контексте).

*K-анонимность (k-anonymity)* — концепция в области обезличивания данных, представляющая собой методологию, при которой каждая запись в наборе данных неотличима от по крайней мере k-1 других записей по определенному набору атрибутов. Этот подход направлен на устранение уникальных идентификационных признаков и сведение к минимуму риска нежелательного раскрытия личной информации при анализе данных.

Применение комплекса методов, таких как "удаление атрибутов", "маскеризация" и "локальное обобщение", предоставляет эффективные стратегии для достижения максимальных значений k-anonymity в датасете. Каждый из этих методов активно воздействует на обезличивание данных, соблюдая баланс между сохранением структуры данных и снижением риска идентификации.

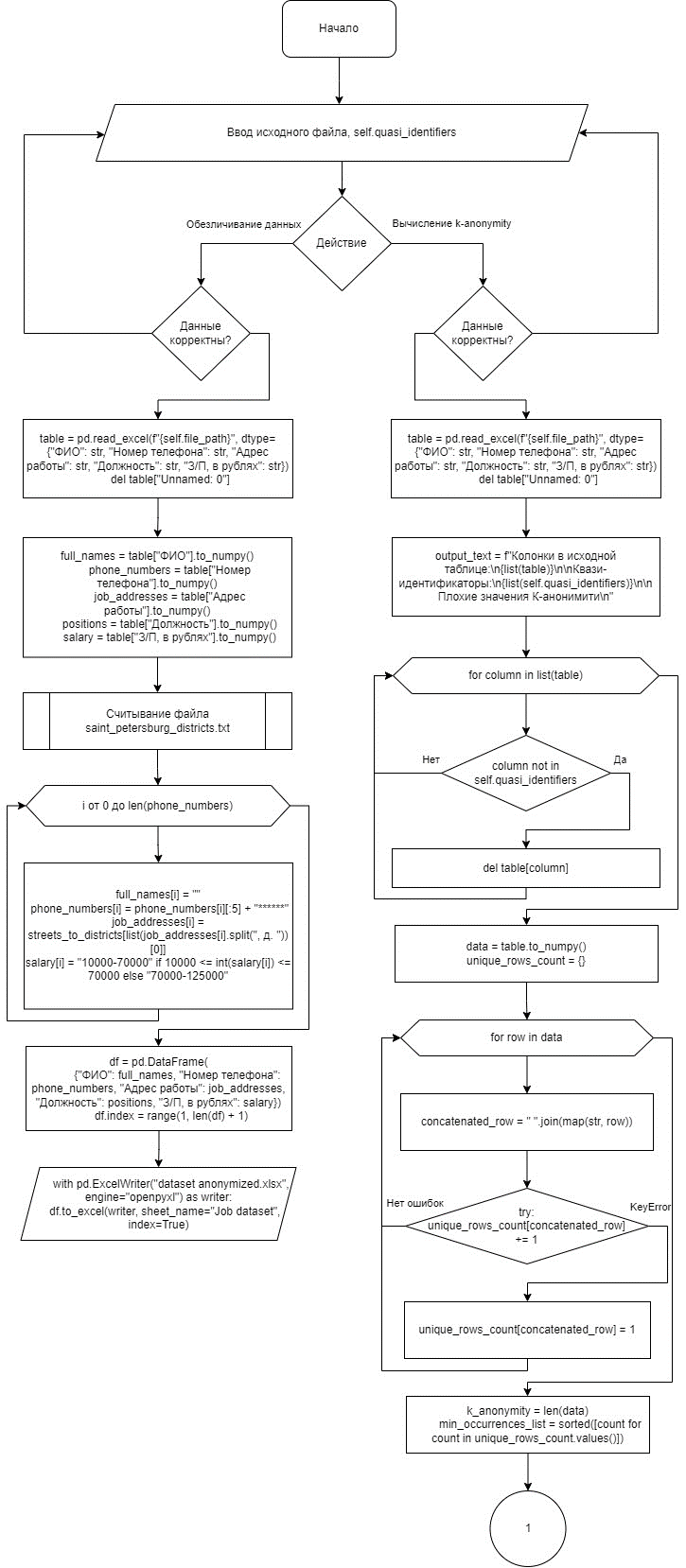
Применение метода "удаление атрибутов" к ФИО эффективно исключает чувствительные личные данные, такие как полные имена, что существенно снижает уникальность данных и, следовательно, улучшает k-anonymity. Метод "маскеризации" номеров телефонов, в свою очередь, обеспечивает сохранение общей структуры данных, предотвращая идентификацию по номерам телефонов.

Применение метода "локальное обобщение" для адресов работ и зарплат позволяет уменьшить детализацию данных путем агрегации близких значений. Например, агрегация адресов работ в более общие районы и локальное обобщение зарплат в более широкие интервалы снижают точность данных, сохраняя при этом общие тенденции.

Этот комплексный подход акцентирует баланс между обезличиванием и сохранением полезности данных, что способствует повышению уровня k-anonymity.

# **Описание алгоритма**

1. Считывание файла для обработки и квази-идентификаторов в k-anonymity из поля ввода интерфейса.
2. При нажатии кнопки *«Обезличивание данных»*:
   1. Проверка на корректность выбора файла. Если файл некорректен или отсутствует, то появляется текст ошибки.
   2. Считывание исходного файла, а также файла с соответствием улиц Санкт-Петербурга к районам *«saint\_petersburg\_districts.txt»*.
   3. Применение метода “удаление атрибутов” для ФИО, метода “маскеризация” для номеров телефонов, метода “локальное обобщение” для адресов работ и З/П.
   4. Сохранение обезличенного набора данных в таблицу Excel.
3. При нажатии кнопки *«Вычисление k-anonymity»*:
   1. Проверка на корректность выбора файла. Если файл некорректен или отсутствует, то появляется текст ошибки.
   2. Считывание исходного файла.
   3. Удаление идентификаторов в исходном файле, не являющихся квази-идентификаторами.
   4. Поиск количества вхождений каждой строки в обработанном файле.
   5. Поиск k-anonymity, вывод 5 строк данных, которые имеют худшие значения k-anonymity.
   6. Вывод количества уникальных строк в обработанном файле.



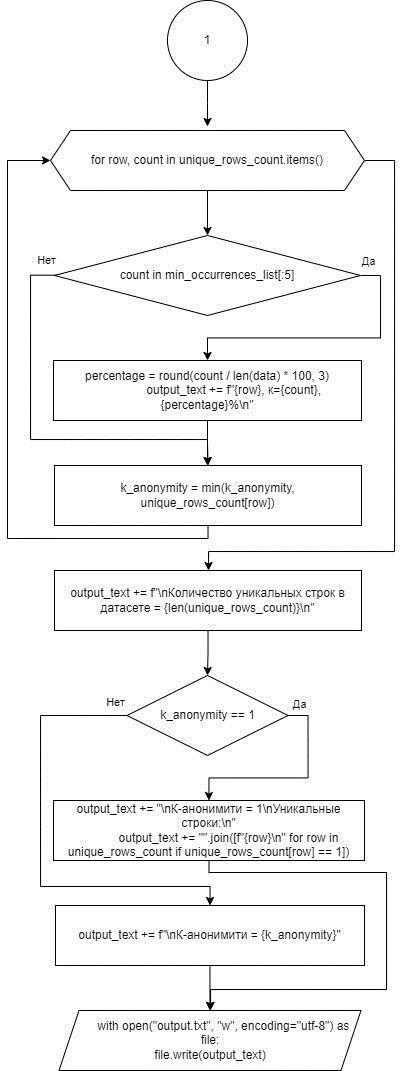


Рисунок 4.1–4.2. Блок-схема алгоритма.

# **Описание программы**

## **Описание классов**

В программе используется 2 класса: 1 связанный с интерфейсом программы, 1 связанный с обработкой набора данных. В таблице 5.1 представлено описание классов.

*Таблица 5.1. Описание классов*

|  |  |  |
| --- | --- | --- |
| Имя класса | Наследование | Описание |
| InterfaceApp | — | Создание интерфейса |
| DatasetDepersonalization | — | Обезличивание набора данных, вычисление k-anonymity |

## **5.2. Описание функций**

Описание функций класса *«DatasetDepersonalization»* представлено в таблице 5.2.

*Таблица 5.2. Описание функций класса «DatasetDepersonalization»*

|  |  |  |
| --- | --- | --- |
| Входные данные | Имя | Описание |
| Путь до файла, состояние интерфейса, квази-идентификаторы | \_\_init\_\_ | Инициализация класса |
| — | depersonalization | Обезличивание набора данных |
| — | calc\_k\_anonymity | Вычисление k-anonymity |

## **5.3. Описание переменных**

*Таблица 5.3. Описание переменных функции «depersonalization»*

|  |  |  |
| --- | --- | --- |
| Тип | Имя | Описание |
| pd.DataFrame | table | Исходная таблица |
| np.ndarray | full\_names | Набор ФИО |
| np.ndarray | phone\_numbers | Номера телефонов |
| np.ndarray | job\_addresses | Адреса работ |
| np.ndarray | positions | Должности |
| np.ndarray | salary | Зарплаты |
| Dict[str, str] | streets\_to\_districts | Словарь с соответствием улиц к району |
| pd.DataFrame | df | Итоговая таблица обезличенных данных |
| Dict[str, int] | column\_widths | Название колонок в итоговой таблице и их ширина |
| worksheet | worksheet | Название листа в Excel |

*Таблица 5.4. Описание переменных функции «calc\_k\_anonymit»*

|  |  |  |
| --- | --- | --- |
| Тип | Имя | Описание |
| pd.DataFrame | table | Исходная таблица |
| str | output\_text | Итоговый текст для вывода, содержащий данные об обработке |
| np.ndarray | data | Данные таблицы |
| Dict[str, int] | unique\_rows\_count | Количество вхождений каждой строки данных среди всего набора данных |
| int | k\_anonymity | Значение k-anonymity |
| List[int] | min\_occurrences\_list | Отсортированный массив значений k-anonymity для строк |

# **Рекомендации пользователю**

Поле *«Входной файл»* устанавливает исходный файл для последующей обработки. Если файл не выбран, то программа выдаст ошибку. Поля с указанием квази-идентификаторов указывает на то, какие идентификаторы будут использоваться при вычислении k-anonymity.

Кнопка *«Обезличивание данных»* отвечает за обезличивание входного набора данных. Данные сохраняются в файл *«dataset anonymized.xlsx»* в директории, хранящей в себе программу.

Кнопка *«Вычисление k-anonymity»* отвечает за вычисления k-anonymity. Программа выполнит генерацию синтетических данных на основе заданных настроек.

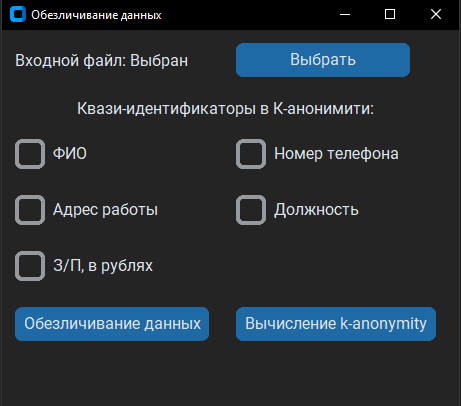
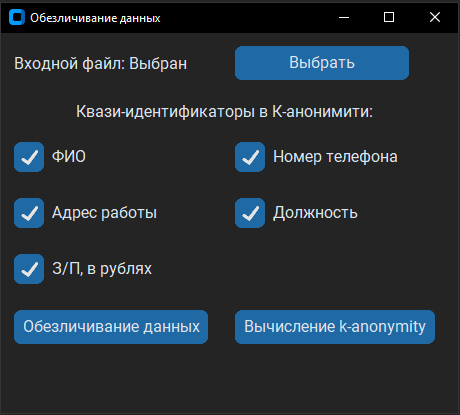
Для завершения работы нажмите на крестик в левом верхнем углу.

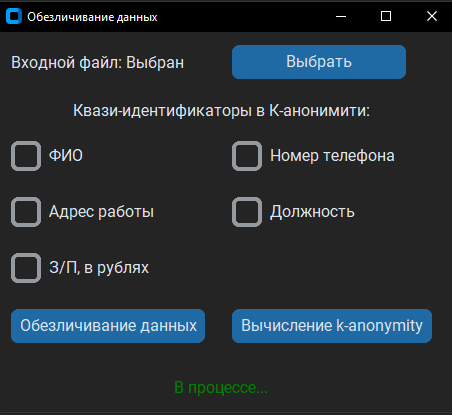
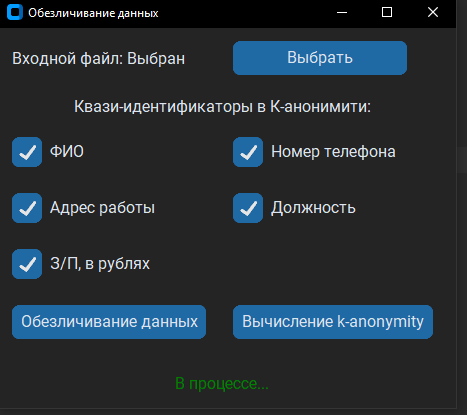
# **Рекомендации программисту**

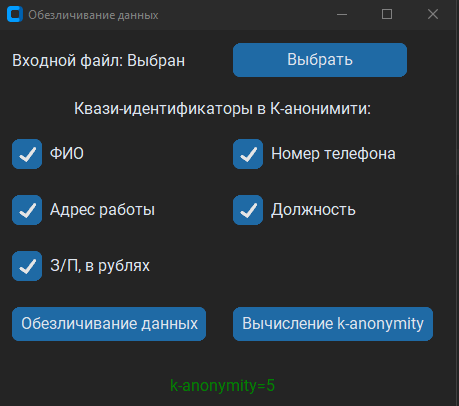
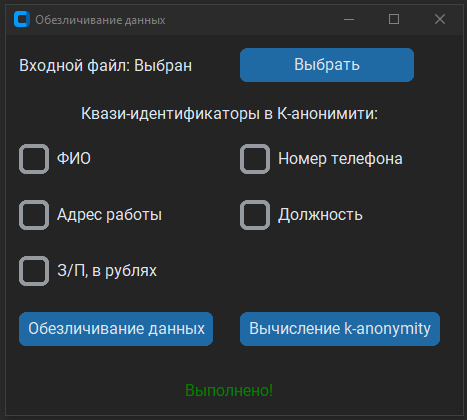
Для запуска программы необходим Python версии не ниже 3.10.6, а также 64-битная операционная система Windows/Linux/macOS. Предварительно необходимо установить библиотеки: customtkinter версии не ниже 5.2.0, pandas версии не ниже 2.1.0. Для работы с кодом необходим PyCharm версии не ниже 2022.2.1.

Минимальное необходимое место на диске: 0.5 МБ. Минимальное необходимое количество оперативной памяти: 500 МБ.

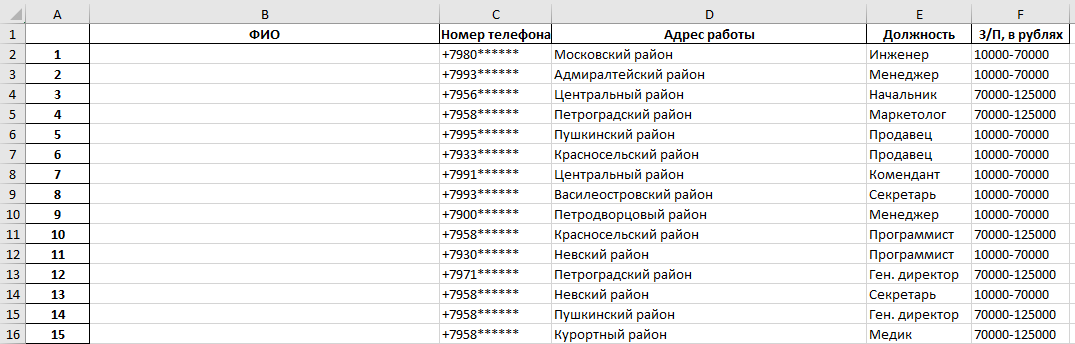
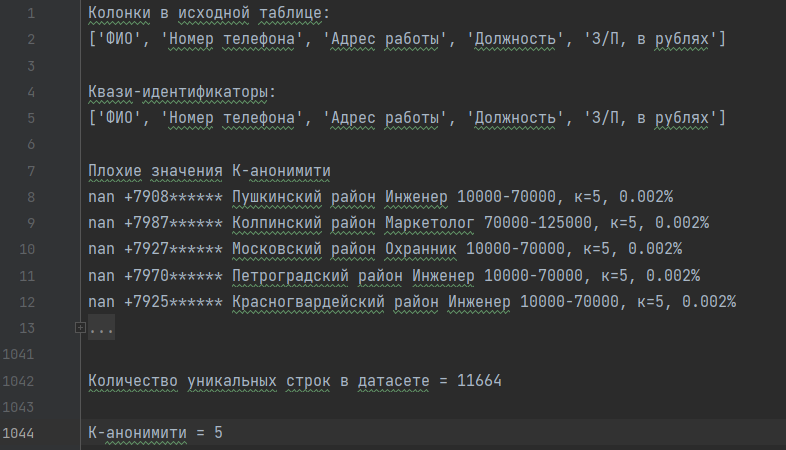
# **Контрольный пример**

 В данном разделе представлены два контрольных примера (описываются они параллельно), демонстрирующие способность программы обезличивать данные и вычислять k-anonymity.

*Рисунок 8.1–8.2. Интерфейс программы*

*Рисунок 8.3–8.4. Окно программы в процессе выполнения функции*

*Рисунок 8.5–8.6. Окно программы после окончания выполнения функции*



*Рисунок 8.7–8.8. Результат работы программы*

# **Заключение**

В рамках данного отчета была успешно достигнута поставленная цель - разработка алгоритма и программы для обезличивания набора синтетических данных. Лабораторная работа позволила не только приобрести, но и закрепить знания в области обезличивания данных. Был создан набор обезличенных данных, соответствующих определенным требованиям и характеристикам, что имеет важное значение для последующего использования в различных учебных целях. Также в рамках обезличивания данных было проведено вычисление k-anonymity, обеспечивая уровень анонимности и безопасности данных.