Санкт-Петербургский государственный университет

Кафедра компьютерного моделирования и многопоточных систем

Лабораторная работа по дисциплине
Алгоритмы и структуры данных

«Обезличивание данных»

Выполнил студент:
Зайнуллин Мансур Альбертович
Группа 23.Б16-пу

Преподаватель:
Дик Александр Генадьевич
ассистент кафедры

07.11.2024
Санкт-Петербург

Оглавление

1 Цель работы

Цель данной работы — анонимизировать данные, сгенерированные в предыдущей лабораторной работе. Для этого применяются методы локального обобщения и маскирования, направленные на обеспечение конфиденциальности и достижение K-анонимности. Данные, созданные ранее, обеспечивают преемственность проекта. Ожидается, что в результате будет достигнут определённый уровень K-анонимности, что будет проверено в ходе работы.

2 Описание задачи

Задача, поставленная преподавателем, заключается в разработке алгоритма и программы для анонимизации данных, сгенерированных в предыдущей лабораторной работе. Основная цель — обеспечить конфиденциальность данных, применяя методы анонимизации, такие как локальное обобщение и маскирование, и достичь определённого уровня K-анонимности.

Программа должна обрабатывать входные данные, идентифицировать квази-идентификаторы и применять выбранные методы анонимизации. Важным аспектом является достижение K-анонимности, что требует тщательной настройки алгоритмов для обеспечения, что каждая запись в наборе данных не может быть однозначно идентифицирована.

Эта задача связана с предыдущей работой, где были сгенерированы данные, и требует их дальнейшей обработки для повышения уровня конфиденциальности. Ожидается, что в результате выполнения задачи будет достигнут баланс между анонимностью и полезностью данных.

3 Теоретическая часть

В данной работе реализуется анонимизация данных с целью обеспечения конфиденциальности и достижения K-анонимности.

3.1 Понятие K-анонимности

K-анонимность — это метод анонимизации, который гарантирует, что каждая запись в наборе данных неотличима от как минимум K-1 других записей по квази-идентификаторам. Квази-идентификаторы — это атрибуты, которые могут быть использованы для идентификации, но не являются уникальными сами по себе.

3.2 Расчёт K-анонимности

Чтобы рассчитать K-анонимность:

1. **Определите квази-идентификаторы**: Выберите атрибуты, которые будут использоваться для группировки данных.
2. **Группируйте строки**: Сгруппируйте записи по выбранным квази-идентификаторам.
3. **Подсчитайте записи**: Определите количество записей в каждой группе.
4. **Проверьте соответствие**: Убедитесь, что минимальное количество записей в любой группе не меньше K.

3.3 Методы анонимизации

1. **Локальное обобщение**:
    - **Дата и время**: Округляются до года и сезона.
    - **Стоимость**: Разбивается на диапазоны.
2. **Маскирование**:

- **Категория и бренд**: Полностью заменяются на символы '******'.

3. **Замена номеров карт**:

   - **Номер карты**: Заменяется на платёжную систему (например, VISA).

4. **Категоризация местоположения**:

   - **Долгота и широта**: Преобразуются в категории, такие как «До 4 км от центра».

3.4 Применение методов к данным

- **Название магазина**: Заменяется на категорию магазина.
- **Дата и время**: Округляются до года и сезона.
- **Долгота и широта**: Преобразуются в категории расстояния от центра.
- **Категория и бренд**: Полностью маскируются.
- **Номер карты**: Заменяется на платёжную систему.
- **Количество товаров и стоимость**: Обобщаются до определённых категорий.

3.5 Ожидаемая K-анонимность

Для набора данных в 2,000,000 строк ожидаемое значение K-анонимности составляет примерно 7.7. Это означает, что каждая запись в среднем неотличима от 7 других записей.

4 Блок-схемы

В этой главе представлены блок-схемы, иллюстрирующие логику работы программы `anonymize.py`.

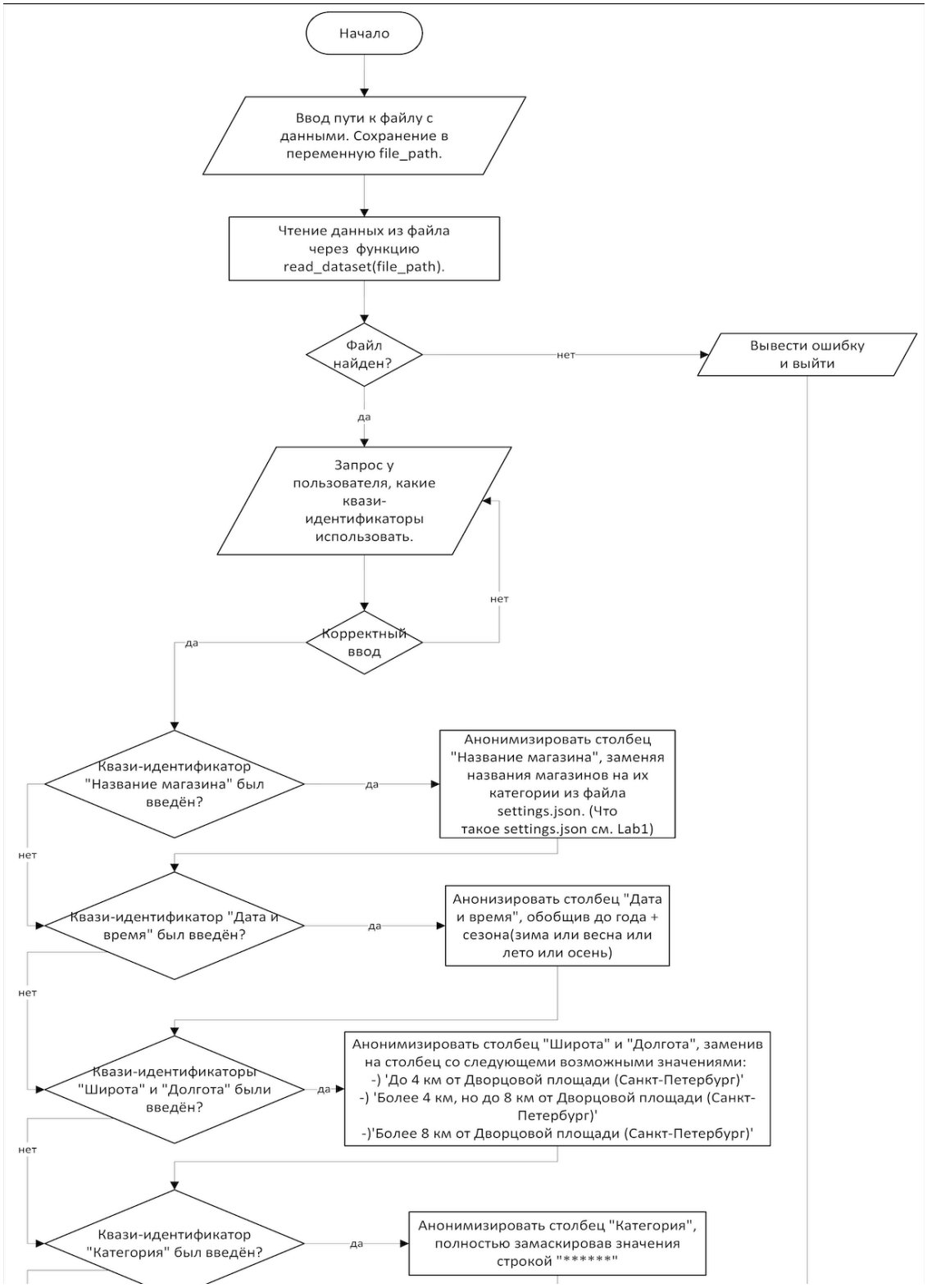

Figure 1: Блок-схема программы часть 1

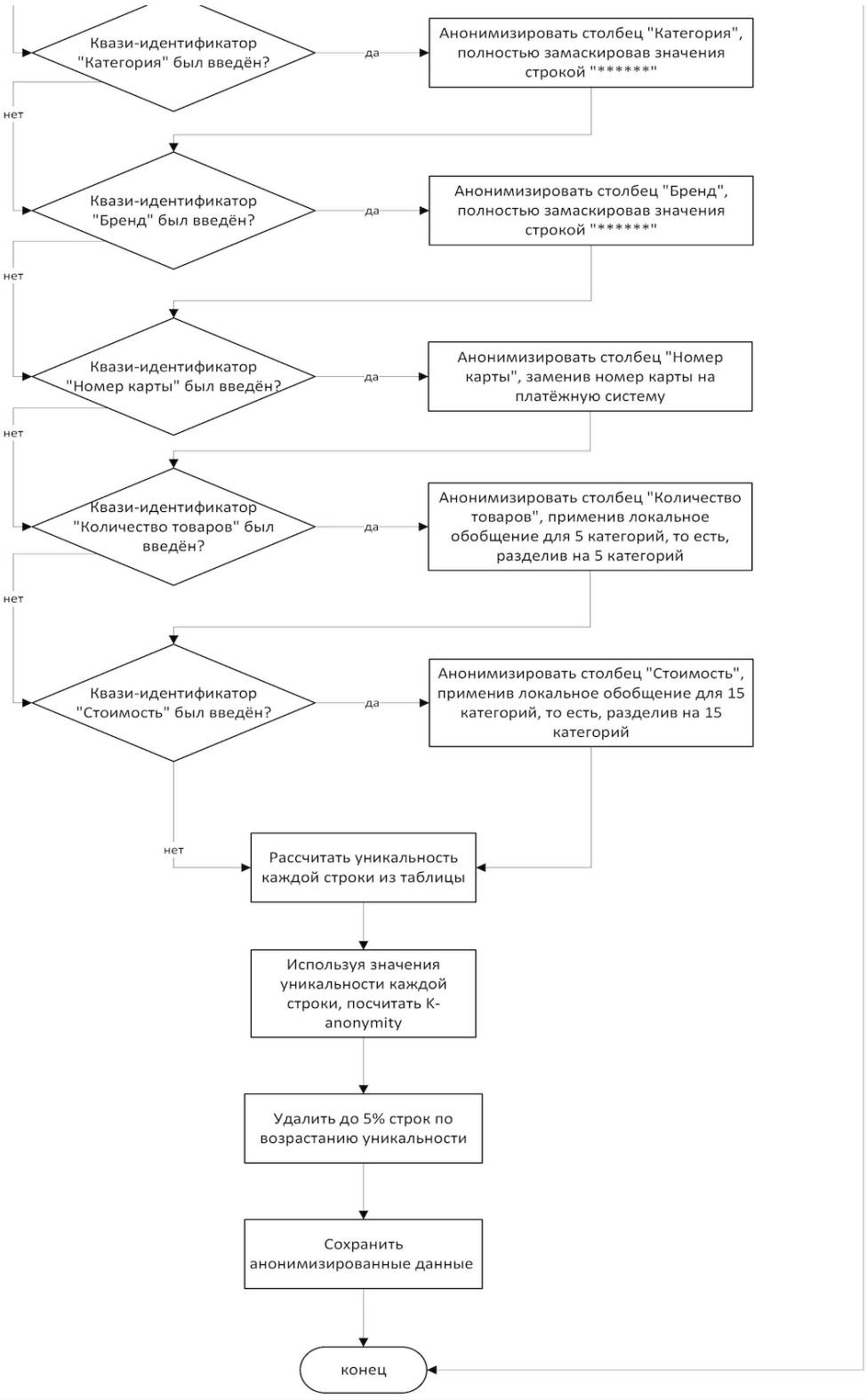

Figure 2: Блок-схема программы часть 2

5 Описание программы

В этой главе подробно рассматривается скрипт `anonymize.py`, предназначенный для анонимизации данных. Программа реализует различные методы анонимизации, чтобы обеспечить конфиденциальность данных и достичь K-анонимности.

5.1 Обзор программы

Программа состоит из нескольких функций, каждая из которых выполняет определённую задачу в процессе анонимизации. Эти функции работают совместно, чтобы обеспечить эффективную и безопасную обработку данных.

5.2 Описание функций

1. **Чтение и подготовка данных**:
    - `read_dataset(file_path)`: Загружает данные из указанного файла и преобразует их в DataFrame для дальнейшей обработки.
    - `get_quasi_identifiers(df)`: Запрашивает у пользователя выбор квази-идентификаторов, которые будут использованы для анонимизации.
2. **Анонимизация данных**:
    - `anonymize_shop_names(df)`: Заменяет названия магазинов на соответствующие категории.
    - `anonymize_location(df)`: Преобразует координаты в категории расстояний от центра.
    - `anonymize_datetime(df)`: Обобщает дату и время до года и сезона.
    - `mask(df, column_name)`: Полностью маскирует указанный столбец.

- `anonymize_card_number(df)`: Заменяет номер карты на название платёжной системы (например, VISA или MasterCard).
- `generalize_column(df, column_name, num_categories)`: Применяет локальное обобщение к выбранному столбцу данных.

3. **Оценка и улучшение K-анонимности**:

   - `calculate_row_uniqueness(df)`: Рассчитывает уникальность каждой строки данных.
   - `remove_rows_by_uniqueness(df, percentage=10)`: Удаляет строки с низкой уникальностью для повышения уровня анонимности.
   - `identify_bad_k_values(df, quasi_identifiers, max_rows=5)`: Идентифицирует группы с низким уровнем K-анонимности.

4. **Основная логика**:

   - `perform_anonymization(df, quasi_identifiers, remove_rows=True)`: Объединяет все этапы анонимизации.
   - `main()`: Координирует выполнение всех функций и обеспечивает вывод результатов.

6 Рекомендации пользователю

В этой главе вы найдёте простые и понятные инструкции по запуску и использованию программы для анонимизации данных. Следуйте шагам ниже, чтобы успешно выполнить процесс.

6.1 Подготовка среды

1. **Установка Python**:
    - Убедитесь, что у вас установлена версия Python 3.9. Если нет, скачайте её с официального сайта.
2. **Скачивание репозитория**:
    - Перейдите по ссылке и скачайте репозиторий. Если у вас установлен Git, используйте команду `git clone`.

6.2 Скачивание и подготовка данных

1. **Генерация или скачивание датасета**:
    - Сгенерируйте или скачайте синтетический датасет, используя код из первой лабораторной работы. Подробные инструкции можно найти в отчёте.
2. **Перенос файлов**:
    - Переместите сгенерированный датасет и файл `settings.json` в папку `Lab2` репозитория.

6.3 Установка зависимостей

1. **Установка библиотек**:
    - Откройте терминал и перейдите в папку `Lab2` с помощью команды `cd _ _` .

- Установите необходимые библиотеки, выполнив команду `pip install -r requirements.txt`.

6.4 Запуск программы

1. **Запуск программы**:

   - В терминале выполните команду `python3 anonymize.py`.
   - Следуйте указаниям, которые будут выводиться на экран.

6.5 Получение результатов

1. **Результаты**:

   - После завершения программы данные будут анонимизированы. Проверьте результаты в созданном файле.

6.6 Заключение

Поздравляем! Вы успешно анонимизировали данные. Если у вас возникли вопросы, обратитесь за помощью. Эти инструкции помогут вам легко и быстро использовать программу, даже если вы не знакомы с программированием.

7 Рекомендации программисту

7.1 Введение

Эта глава предоставляет краткие и чёткие инструкции для программистов по запуску и работе с программой `anonymize.py`. Предполагается, что вы знакомы с основами работы с Python и Git.

7.2 Подготовка среды

1. **Установка Python**:
   - Убедитесь, что у вас установлена версия Python 3.9. Если нет, скачайте её с официального сайта.
2. **Скачивание репозитория**:
   - Клонируйте репозиторий с помощью команды:

```
git clone
   https://github.com/MansurYa/labs-for-algorithms-and-data-structures
```

   - Или скачайте ZIP-файл и разархивируйте его.

7.3 Подготовка данных

1. **Генерация или скачивание датасета**:
   - Используйте код из первой лабораторной работы для генерации данных. Подробные инструкции доступны в отчёте.
2. **Перенос файлов**:
   - Переместите сгенерированный датасет и файл `settings.json` в папку `Lab2`.

7.4 Установка зависимостей

- Выполните команду в терминале:

```
pip install -r requirements.txt
```

7.5 Запуск программы

- Перейдите в папку Lab2:

```
cd labs-for-algorithms-and-data-structures/Lab2
```

- Запустите программу:

```
python3 anonymize.py
```

- Следуйте указаниям, выводимым на экран.

7.6 Заключение

После выполнения всех шагов данные будут успешно анонимизированы. Эти инструкции помогут вам быстро и эффективно настроить и запустить программу.

8 Контрольный пример

8.1 Введение

В этой главе представлен контрольный пример, демонстрирующий процесс анонимизации данных. Цель примера — показать, как исходные данные преобразуются в анонимизированный набор, обеспечивая конфиденциальность и достижение K-анонимности.

8.2 Исходные данные

На рисунке ниже представлена таблица с данными до анонимизации. Она включает такие столбцы, как Название магазина, Дата и время, Долгота, Широта, Категория, Бренд, Номер карты, Количество товаров и Стоимость.

	A	B	C	D	E	F	G	H	I
1	**Название магазина**	**Дата и время**	**Долгота**	**Широта**	**Категория**	**Бренд**	**Номер карты**	**Количество товаров**	**Стоимость**
2	Магнит	2017-08-23 06:12	59,97759	30,4368	Замороженные прод	MorningStar Farms	4023128611186297	37	500
3	OBI	2015-06-18 22:27	59,87128	30,34932	Напольные покрытия	Tarkett	4023128611186297	5	1500
4	Лента	2015-03-06 15:38	59,83165	30,36293	Молочные продукты	Campina	4023128611186297	18	85
5	Пятёрочка	2015-11-05 09:44	59,92358	30,3697	Орехи и сухофрукты	Семушка	4023128611186297	32	400
6	Афоня	2018-12-18 09:12	59,85589	30,32467	Электрика	IEK	5469663786773218	19	3200
7	Дизель спот	2020-06-22 10:12	59,96002	30,46879	Топливные системы	Bosch	5469663786773218	33	12000
8	Porivay Suit	2024-10-03 12:36	59,91518	30,35678	Одежда для йоги	Adidas	5469663786773218	13	6000
9	Петрович	2016-10-30 22:19	60,09559	30,30194	Освещение	TCP	6768051164126067	17	900
10	OBI	2019-10-10 11:36	60,08983	30,38236	Фасадные материалы	Mapei	6768051164126067	5	1300
11	DreamWhite	2015-03-02 17:12	59,9384	30,35918	Часы и украшения	Nokia	5484495377570242	5	20000
12	Азбука вкуса	2017-01-04 02:32	59,90529	30,31491	Специи и пряности	McCormick	5484495377570242	5	200

Figure 3: Исходные данные до анонимизации

8.3 Процесс анонимизации

Для анонимизации данных были использованы следующие методы:

- **Локальное обобщение**: Применено к дате и времени, количеству товаров и стоимости.
- **Маскирование**: Использовано для категорий и брендов.
- **Замена номеров карт**: Номера карт заменены на платёжные системы.
- **Категоризация местоположения**: Долгота и широта преобразованы в категории расстояния от центра.

8.4 Результаты анонимизации

На рисунке ниже представлена таблица с данными после анонимизации. Видно, что данные изменены для обеспечения конфиденциальности.

	A	B	C	D	E	F	G	H
1	**Название магазина**	**Дата и время**	**Категория**	**Бренд**	**Номер карты**	**чество тов**	**Стоимость**	**Расположение**
2	Продуктовые магазины	2017, лето	******	******	VISA	30-37	450-650	Более 8 км от Дворцовой площади (Санкт-Петербург)
3	Строительные магазины	2015, лето	******	******	VISA	5-15	1000-1500	Более 8 км от Дворцовой площади (Санкт-Петербург)
4	Продуктовые магазины	2015, весна	******	******	VISA	15-20	85-150	Более 8 км от Дворцовой площади (Санкт-Петербург)
5	Продуктовые магазины	2015, осень	******	******	VISA	30-37	250-400	Более 8 км от Дворцовой площади (Санкт-Петербург)
6	Строительные магазины	2018, зима	******	******	MASTERCARD	15-20	2500-4000	Более 8 км от Дворцовой площади (Санкт-Петербург)
7	Магазины автозапчастей и автотоваров	2020, лето	******	******	MASTERCARD	30-37	10000-14500	Более 8 км от Дворцовой площади (Санкт-Петербург)
8	Магазины одежды	2024, осень	******	******	MASTERCARD	5-15	4000-6000	Более 8 км от Дворцовой площади (Санкт-Петербург)
9	Строительные магазины	2016, осень	******	******	MAESTRO	15-20	650-900	Более 8 км от Дворцовой площади (Санкт-Петербург)
10	Строительные магазины	2019, осень	******	******	MAESTRO	5-15	1000-1500	Более 8 км от Дворцовой площади (Санкт-Петербург)
11	Магазины одежды	2015, весна	******	******	MASTERCARD	5-15	14500-20000	Более 8 км от Дворцовой площади (Санкт-Петербург)
12	Продуктовые магазины	2017, зима	******	******	MASTERCARD	5-15	150-250	Более 8 км от Дворцовой площади (Санкт-Петербург)

Figure 4: Анонимизированные данные

8.5 Анализ результатов

В результате анонимизации достигнута K-анонимность, равная 7 для таблицы в 1,000,000 строк. Это означает, что каждая запись неотличима от 6 других, что значительно повышает уровень конфиденциальности.

8.6 Заключение

Контрольный пример подтверждает эффективность применённых методов анонимизации. Данные успешно преобразованы, обеспечивая необходимый уровень конфиденциальности и K-анонимности.

9 Вывод

В ходе данной работы была успешно достигнута цель анонимизации данных, сгенерированных в предыдущей лабораторной работе. Применённые методы, такие как локальное обобщение и маскирование, доказали свою эффективность, обеспечив необходимый уровень K-анонимности. Скрипт `anonymize.py` продемонстрировал свою работоспособность и эффективность на контрольном примере, что подтверждает его пригодность для решения задач по защите конфиденциальной информации. Таким образом, работа показала важность использования данных методов для обеспечения конфиденциальности и безопасности данных.

10 Полезные ссылки

- Репозиторий всех лабораторных
- Ссылка на хранилище к этой лабораторной