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 датасета»

Вариант – 2

Студент гр. 23Б15-пу
Трофимов И. А.

Преподаватель
Дик А.Г.

Санкт-Петербург
2024 г.

Оглавление

1. Цель работы: 4
2. Теоретическая часть 4
3. Описание задачи 7
4. Основные шаги программы 7
5. Описание программы 10
6. Рекомендации пользователя 11
7. Рекомендации программиста 11
8. Исходный код программы: 11
9. Контрольный пример 12
10. Вывод 14
11. Источники 14

Цель работы:

Целью лабораторной работы является разработка механизма для обезличивания медицинского датасета. При работе с такими данными необходимо учитывать требования по защите персональной информации и возможности восстановления исходных данных. Датасет должен содержать такие данные, как:

- Врач
- ФИО пациента
- Паспортные данные
- СНИЛС
- Информация о банковской карте
- История визитов к врачам (симптомы, анализы, дата посещения врача и получения анализов, стоимость приема).

Теоретическая часть

Для обезличивания медицинских данных используются различные методы и алгоритмы, направленные на защиту конфиденциальной информации, при этом стараясь сохранить полезность набора данных для дальнейшего анализа. В данной работе основной акцент делается на обезличивание полей, которые могут раскрыть личность пациента или его медицинскую историю. К таким полям относятся: ФИО, СНИЛС, номер карты, симптомы, анализы и другие идентификаторы, которые могут использоваться для реидентификации.

Основные методы обезличивания данных

Для обеспечения конфиденциальности медицинских данных в данной работе используются следующие методы обезличивания:

1. **Локальное обобщение**

   - Локальное обобщение заключается в замене детализированных значений данных на более общие категории или диапазоны. Например, вместо точного возраста пациента можно указать возрастной диапазон (например, 20-30 лет). Этот метод часто используется для таких полей, как даты посещений врача и стоимости услуг.
   - Пример: Точные даты посещений врача могут быть заменены только годом посещения (например, вместо "2023-06-03" используется "2023").

2. **Маскирование данных**

   - Маскирование данных — это метод, при котором часть информации скрывается, оставляя только обобщенные или частичные значения. Этот метод подходит для полей, таких как СНИЛС или номер карты, где важно сохранить структуру данных, но не позволить восстановить точные значения.
   - Пример: Паспорт "4019 738384" может быть замаскирован как "37** ******".

3. **Удаление атрибутов**

   - В некоторых случаях для предотвращения утечки данных необходимо полностью удалить определенные атрибуты из набора данных. Этот метод применяется, если атрибут несет слишком высокий риск идентификации или не является критически важным для анализа.
   - Пример: Поля со СНИЛС могут быть удалены из датасета для полного исключения возможности восстановления личности.

4. **Локальное подавление**

   - Локальное подавление применяется к отдельным строкам данных, где существует риск утечки информации. В этом случае, данные в этих строках могут быть полностью скрыты или удалены, если они представляют угрозу для конфиденциальности.
   - Пример: Если определенная запись пациента является уникальной по своим характеристикам, она может быть скрыта полностью для предотвращения раскрытия личности.

Оценка уровня анонимности данных

Для оценки уровня анонимности данных используется метрика K-анонимити. K-анонимность позволяет измерить, насколько защищены данные от реидентификации. В основе метода лежит идея, что каждый набор данных должен иметь хотя бы K записей, которые невозможно отличить друг от друга по выбранным квази-идентификаторам. Чем выше значение K, тем выше уровень анонимности данных.

- Квази-идентификаторы: Это поля, которые сами по себе не являются уникальными, но в комбинации могут привести к идентификации личности. Примером таких данных могут служить сочетания полей "Паспортные данные", "Анализы" и "Дата посещения врача".

- Расчет K-анонимити: Для расчета K-анонимити данные группируются по квази-идентификаторам, и для каждой группы подсчитывается количество записей. Если для определенных комбинаций квази-идентификаторов количество записей меньше допустимого значения K, эти записи считаются "плохими" с точки зрения анонимности.

Описание задачи

Задача состоит в обезличивании датасета, содержащего данные о покупке со следующими требованиями:

1) Программа должна считывать входной датасет.
2) Программа делится по функционалу
    a. Обезличивание входного датасета.
    b. Высчитывание К-анонимити входного датасета.
3) У пользователя есть возможность указывать Квази-идентификаторы в программе.
4) Используя метод К-анонимити рассчитать К для обезличенного набора.
5) Вывести 5 "плохих" значений К-анонимити. Для моего размера датасета в 50к хорошее значение К-анонимити считается > 9. Данные переменной К вывести в процентах из всего набора.

Основные шаги программы

1. **Запуск программы**: Пользователь запускает программу, и загружается датасет из файла data_set.xml.
2. **Обезличивание ФИО**: ФИО заменяется на пол человека.
3. **Обезличивание Паспортных данных**: Остается только страна.
4. **Обезличивание Снилс**: Остается только первая цифра, все остальное заменяется на *****.

5. **Обезличивание симптомов, врачей, анализов**: Если выбрано, все затирается.

6. **Обезличивание цены**: Задается примерный диапазон, вместо точного значения.

7. **Обезличивание банковских карт**: Остается информация только о банке.

8. **Ввод пользователя**: Пользователь выбирает, какие данные он хочет обезличить, через последовательный ввод (y/n).

9. **Расчет K-анонимности**: Происходит группировка и расчет K-анонимности, чтобы определить уровень обезличенности данных.

10. **Вывод результата**: Выводятся первые 5 значений K-анонимности и процентное соотношение этих значений относительно общего количества записей.

11. **Запись обезличенных данных**: Все обезличенные данные сохраняются в новый файл data_set_new.xml.

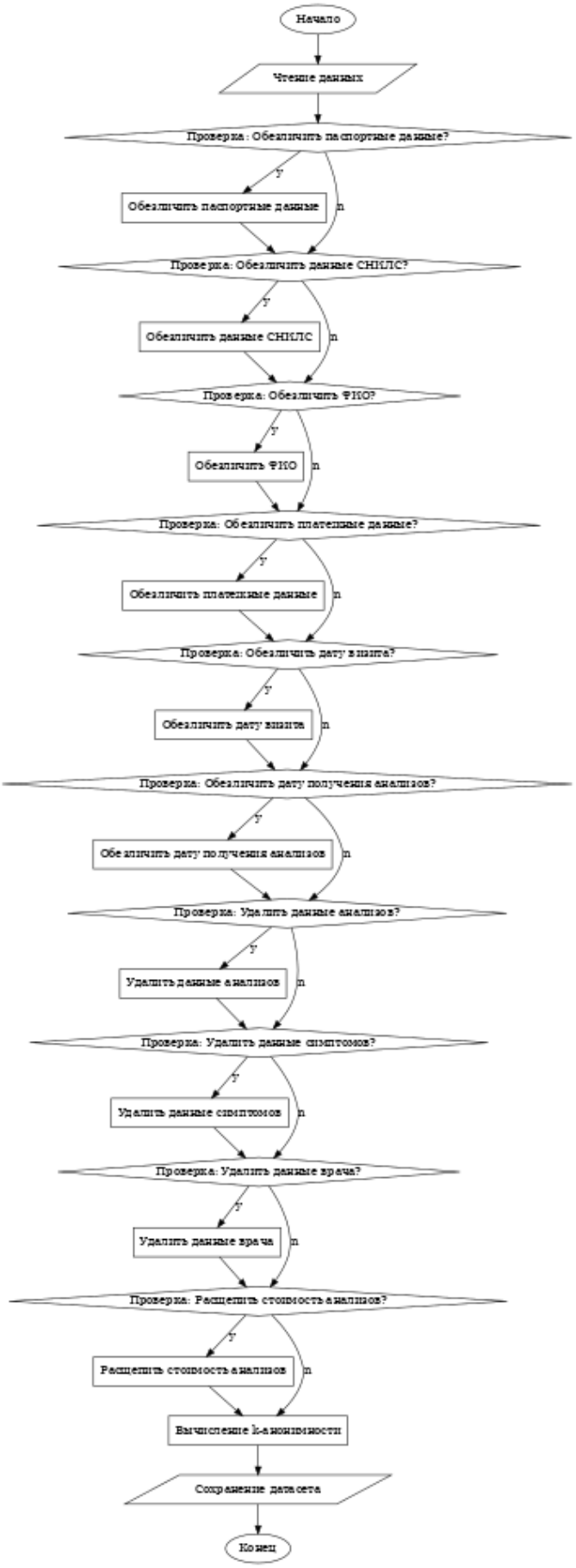

Рис 1. Блок-схема основной программы

Описание программы

Программная реализация написана на языке Python 3.12.2 с использованием библиотеки pandas[1]. Программа организована через модульную структуру, направлена на обезличивание данных о визитах людей ко врачам. В процессе разработки программы использовались следующие методы:

Таблица1 depers.py

Функция	Описание	Возвращаемое значение
xml_to_dataframe	Считывает датасет из xml файла	pd.DataFrame
mask_passport_data	Обезличивает данные, оставляя только их часть.	pd.DataFrame
mask_snils	Обезличивает данные, оставляя только их часть.	pd.DataFrame
anonymize_name_fields	Обезличивает данные, оставляя только их часть.	pd.DataFrame
anonymize_payment_info	Обезличивает данные, оставляя только их часть.	pd.DataFrame
anonymize_date_in_dataframe	Обезличивает данные, оставляя только их часть.	pd.DataFrame
remove_column	Обезличивает данные, затирая столбец из датасета	pd.DataFrame
anonymize_total_analysis_cost	Обезличивает данные, заменяя их на диапозон.	pd.DataFrame
save_to_xml	Сохраняет датасет в xml	None
find_worst_k_anonymity	Подсчитывает значения K-анонимити	list

Рекомендации пользователя

Для запуска программы убедитесь, что у вас установлен Python и необходимые библиотеки, такие как pandas [1]. Код можно запустить в среде разработки или через командную строку, используя консоль для настройки параметров и генерации данных. Также убедитесь, что все модули программы находятся в одной директории для корректного выполнения. Запуск программы производится через файл main.py. Перед запуском убедитесь, что ваш файл data_set.xml правильно отформатирован и содержит минимум 50000 строк для корректной работы с данными.

При запуске программы вам будет предложено обезличить различные аспекты данных, такие как симптомы, врачи и даты. Для того чтобы выбрать желаемые аспекты, нужно ввести их номера, согласно тексту в консоли. После завершения работы программы обезличенные данные будут сохранены в data_set_new.xml.

Рекомендации программиста

Для поддержания актуальности и работоспособности программы используйте последние версии библиотек, особенно pandas. Следите за правильной структурой данных в файле dataset.xml, чтобы избежать ошибок при загрузке и анонимизации данных.

Регулярно тестируйте программу на различных наборах данных, проверяя корректность обезличенных данных, таких как номера карт, стоимости визитов и т.д.

Исходный код программы:

https://github.com/hanglider/Labs/tree/af4b6dec52fd41196d7a8361b1cdd7e55907d9fd/2

Контрольный пример

1. **Запуск программы**

   Для запуска программы используйте файл main.py. Этот скрипт отвечает за обезличивание данных о покупках на основе информации из файла data_set.xml. Программа загружает данные и позволяет пользователю выбрать квази идентификаторы для обезличивания.

2. **Выбор параметров анонимизации**

   После запуска программы пользователю будет предложено выбрать, какие квази идентификаторы он хочет обезличить (Рис. 2). Пользователь может выбрать следующие опции:

   - Паспортные данные
   - Данные снилса
   - ФИО
   - Платежные данные
   - Дату визита
   - Дату получения анализов
   - Данные анализов
   - Данные симптомов
   - Данные врача
   - Стоимость анализов

   Каждую из этих опций можно выбрать, [y/n].

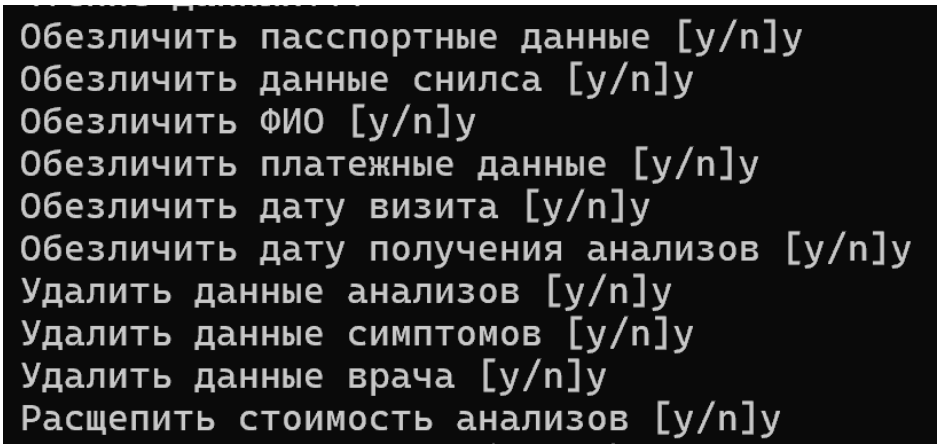

Рис 2. Пример выбора квази-идентификаторов

3. **Обработка данных и вывод результатов**

   После выбора параметров программа обрабатывает данные, а затем сохраняет обезличенные данные в файл data_set_new.xml (Рис. 3). Программа также рассчитывает значения K-анонимности и выводит их на экран, чтобы пользователь мог оценить уровень анонимизации данных (Рис. 4).

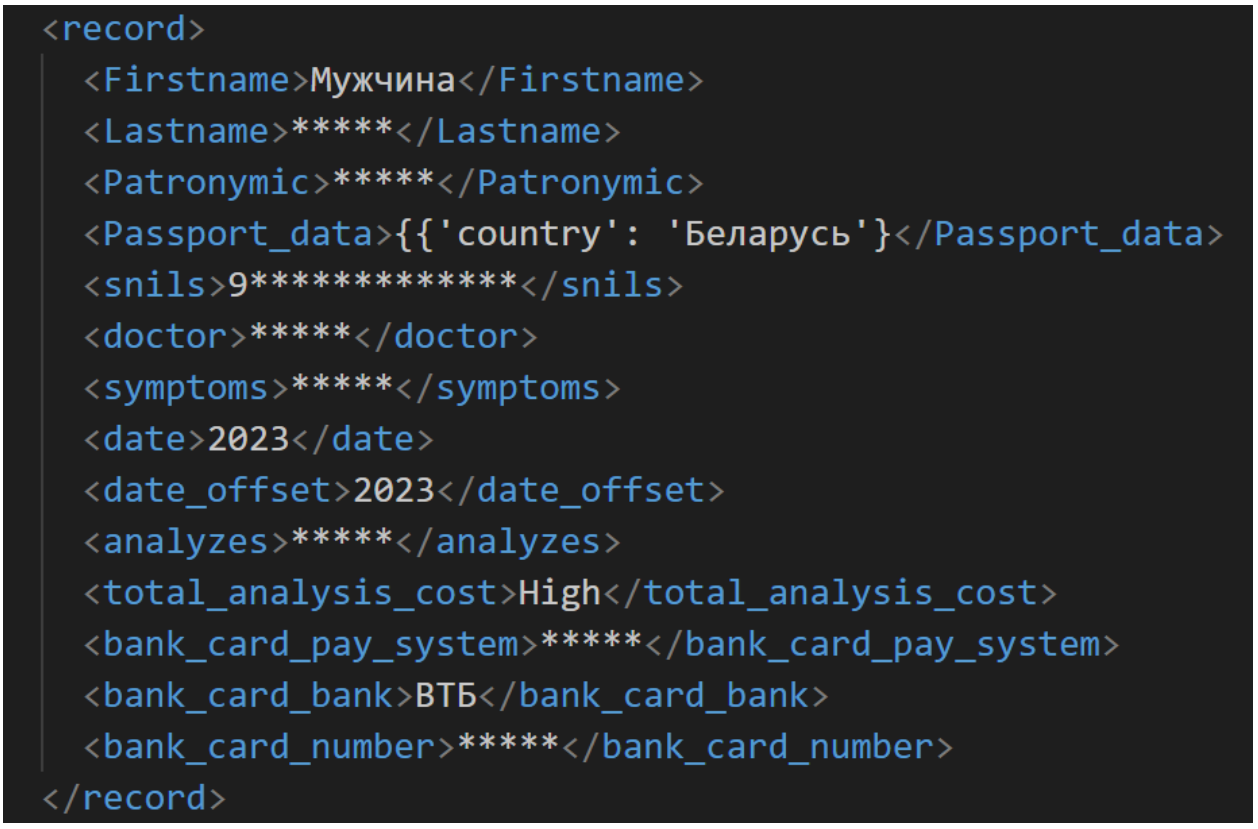

Рис 3. Пример датасета

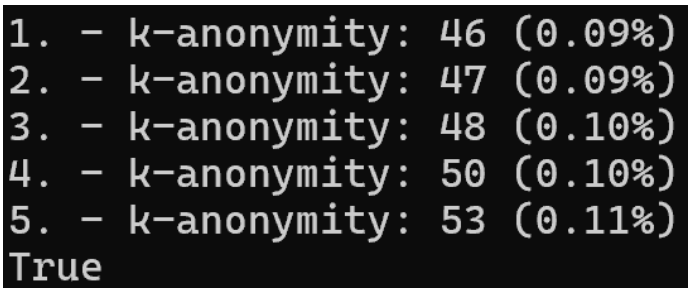

Рис 4. Пример рассчитанных значений K

Вывод

В рамках данной работы был разработан алгоритм для обезличивания данных о визитах людей ко врачу. Программа анализирует существующий датасет, который включает в себя такие параметры, как симптомы, врачи, дата посещения и дата получения анализов, номера карт, стоимость визитов и т.д. Реализованный алгоритм обеспечивает возможность обезличивания различных аспектов данных, что повышает уровень конфиденциальности информации.

В процессе работы программа позволяет пользователю настраивать квази-идентификаторы, выбирая, какие именно данные необходимо скрыть. Это обеспечивает гибкость в обработке данных и позволяет адаптировать программу под специфические требования к анонимности. Обезличенные данные сохраняются в формате XML, что упрощает их последующую обработку и анализ. Программа также рассчитывает значения K-анонимности, что позволяет пользователю оценить степень защиты анонимизируемых данных.

Источники

1. Pandas’ documentation // Pandas URL: https://pandas.pydata.org/docs/ (дата обращения: 2.10.2024).