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 – 1

Студент гр. 23Б15-пу
Антонян А. А.

Преподаватель
Дик А.Г.

Санкт-Петербург
2024 г.

Оглавление

Цель работы

Целью лабораторной работы является разработка механизма для обезличивания датасета, содержащего информацию о покупках в магазинах. При создании и генерации такого набора данных необходимо учитывать различные требования и условия, связанные с конфиденциальностью данных. Генератор данных должен уметь создавать большие наборы, которые включают:

- Название магазина
- Координаты покупки (включая дату, время и географические координаты)
- Категорию и бренд товара
- Номер платежной карты
- Количество товаров и стоимость покупки

Теоретическая часть

Для создания и обезличивания датасета, описывающего транзакции покупок в магазинах, используется несколько программных модулей, которые работают с различными данными. Основное внимание уделено обезличиванию данных с помощью таких методов, как локальное обобщение и удаление атрибутов, а также оценке с помощью метрики k-анонимности.

1. **Загрузка данных**: Данные загружаются из XML-файла.
2. **Модули анонимизации**:
    - **Обезличивание названий магазинов**: Применён метод локального обобщения. Конкретные названия магазинов заменяются на их тематику, что помогает скрыть прямые

идентификаторы торговых точек, сохраняя при этом связь с их типом.

- **Обезличивание координат**: Также обезличены методом локального обобщения. Все географические координаты заменяются на одно фиксированное значение — "Санкт-Петербург.

- **Обезличивание даты и времени**: Применён метод локального обобщения. Остается только год транзакции, удаляя точное время и дату покупки.

- **Обезличивание категорий и брендов**: Применен метод удаления атрибутов. Удаляются столбцы, содержащие информацию о бренде и категории товара.

- **Обезличивание номеров карт**: Применен метод локального обобщения. Номера карт заменяются на названия банков, к которым они относятся.

- **Обезличивание количества товаров и стоимости**: Применен метод локального обобщения. Количество товаров и их стоимость разделяются на диапазоны.

3. **Гибкость анонимизации**: Пользователю предоставляется возможность выбора, какие поля необходимо анонимизировать.

4. **K-анонимность**: Для оценки уровня анонимизации данных используется метрика k-анонимности. Подсчитывается количество уникальных записей, а также его процент по соотношению к длине датасета.

5. **Процентное соотношение K-анонимности**: После вычисления k-анонимности выводится процентное соотношение записей, соответствующих каждому значению k.

6. **Сохранение результатов**: После завершения процесса обезличивания данные сохраняются в формате

Описание задачи

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

1) Программа должна считывать входной файл (Итоговый файл 1-ой лабораторной работы).

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 (если их меньше, то все возможные). Данные переменной К вывести в процентах из всего набора.

Основные шаги программы

1. **Запуск программы**: Пользователь запускает программу, и загружается датасет из файла purchases.xml.
2. **Обезличивание названий магазинов**: В зависимости от выбора пользователя, происходит замена названий магазинов на тематику.
3. **Обезличивание координат**: Если пользователь выбирает эту опцию, все координаты заменяются на "Санкт-Петербург".
4. **Обезличивание даты и времени**: Дата и время покупок обрезаются до года, оставляя только год покупки.
5. **Обезличивание категорий**: Если выбрано, все категории товаров удаляются.
6. **Обезличивание брендов**: В случае выбора пользователем, все бренды удаляются.
7. **Обезличивание номеров карт**: Номера карт заменяются на названия банков в зависимости от первых четырех цифр номера карты.
8. **Обезличивание количества товаров**: Количество товаров заменяется на диапазон.
9. **Обезличивание стоимости**: Итоговая стоимость товаров заменяется на диапазон.
10. **Ввод пользователя**: Пользователь выбирает, какие данные он хочет обезличить, через последовательный ввод (да/нет).
11. **Расчет K-анонимности**: Происходит группировка и расчет K-анонимности, чтобы определить уровень обезличенности данных.

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

13. **Запись обезличенных данных**: Все обезличенные данные сохраняются в новый файл depersonalized_purchases.xml.

14. **Отчет о выполнении программы**: Выводится информация о выполнении программы, включая примерный результат и статистику.

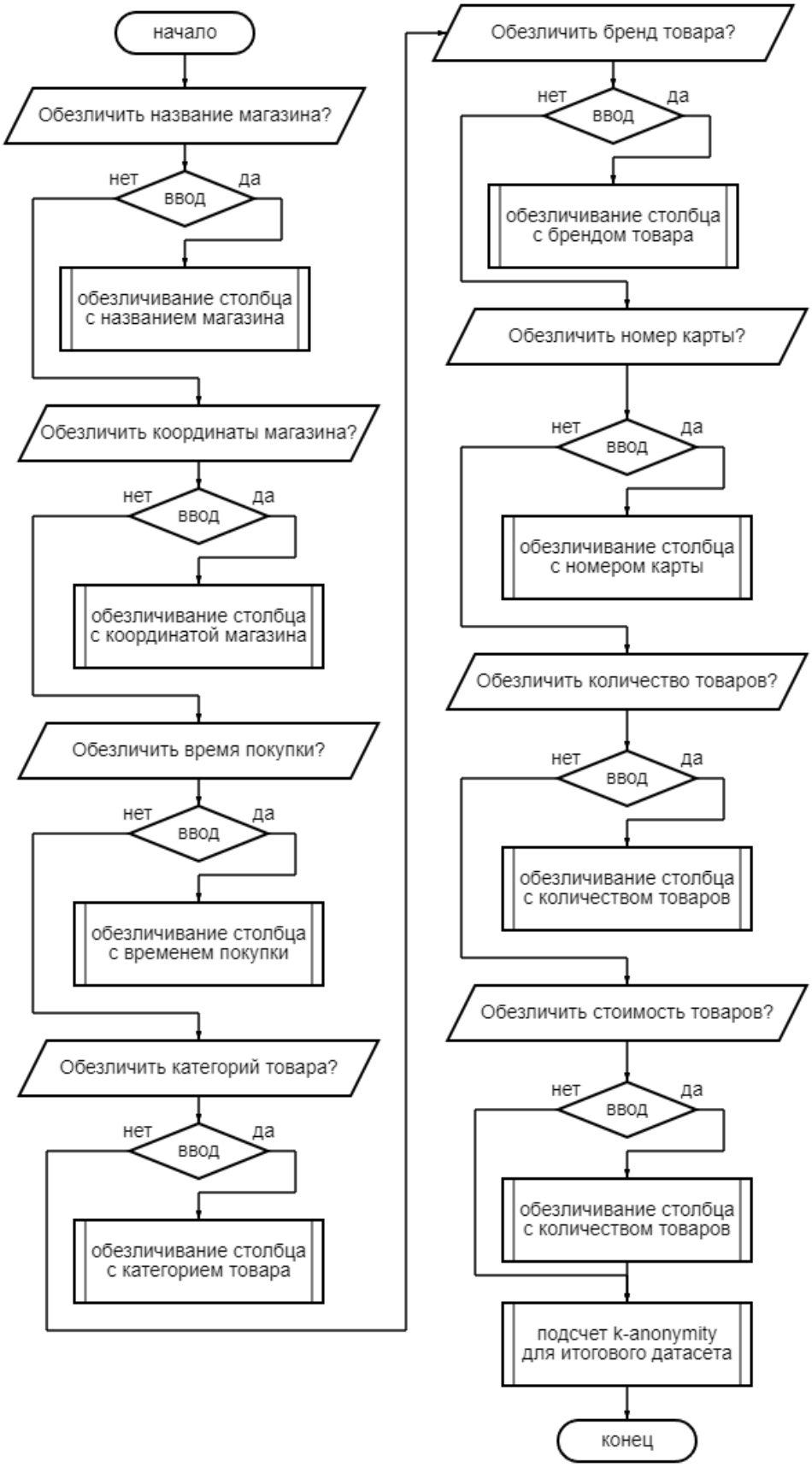

Рис 1. Блок-схема основной программы

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 покупках в магазинах. В процессе разработки программы использовался следующий модуль:

Таблица1 functions.py

Функция	Описание	Возвращаемое значение
calculate_k_anonymity()	Считает топ n значений k для датасета и высчитывает их процентное соотношение.	None
cost_depersonalization()	Обезличивает стоимость покупки, заменяя ее на диапозон.	pd.DataFrame
shop_depersonalization()	Обезличивает название магазина, заменяя его на тематику.	pd.DataFrame
quantity_depersonalization()	Обезличивает количество товаров, заменяя количество на диапазон	pd.DataFrame
brand_depersonalization()	Обезличивает бренд товара, удаляя столбец из датасета	pd.DataFrame
card_number_depersonalization()	Обезличивает номера карт, заменяя их на названия банков на основе первых 4 цифр номера.	pd.DataFrame
coords_depersonalization()	Обезличивает координаты, заменяя их на фиксированное значение "Санкт-Петербург".	pd.DataFrame
time_depersonalization()	Обезличивает дату и время, оставляя только год.	pd.DataFrame
category_depersonalization()	Обезличивает категорию товара, удаляя столбец из датасета.	pd.DataFrame

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 purchases.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 названия магазинов, координаты и даты. Вводите ответы в соответствии с предложенными вариантами (y - да, n - нет). После завершения работы программы обезличенные данные будут сохранены в depersonalized_purchases.xml.

Рекомендации программиста

Для поддержания актуальности и работоспособности программы используйте последние версии библиотек, особенно pandas[1]. Следите за правильной структурой данных в файле dataset.xml, чтобы избежать ошибок при загрузке и анонимизации данных. Применяйте практики надлежащего именования переменных и функций для улучшения читаемости кода.

Регулярно тестируйте программу на различных наборах данных, проверяя корректность сгенерированных данных, таких как номера карт, стоимости товаров и выбор магазинов.

Исходный код программы:

https://github.com/ArseniiAntonian/spbu-algorithms-and-data-structures

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 purchases.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:

   - Обезличивание названий магазинов
   - Обезличивание координат
   - Обезличивание даты и времени
   - Обезличивание категорий
   - Обезличивание брендов
   - Обезличивание номеров карт
   - Обезличивание количества товаров
   - Обезличивание стоимости

   Каждую из этих опций можно включить, введя “y”, или отключить, введя “n”.

```
Обезличить название магазина? [y/n]y
Обезличить координаты магазина? [y/n]y
Обезличить время покупки? [y/n]y
Обезличить категорий товара? [y/n]y
Обезличить бренд товара? [y/n]y
Обезличить номер карты? [y/n]y
Обезличить количество товаров? [y/n]y
Обезличить стоимость покупки? [y/n]y
```

Рис 2. Пример выбора квази-идентификаторов

3. **Обработка данных и вывод результатов**

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

	theme	state	time	quantity	cost	bank
0	Продукты	Санкт-Петербург	2021	5-20	10000-500000	ВТБ
1	Электроника	Санкт-Петербург	2022	5-20	10000-500000	Сбер
2	Животные	Санкт-Петербург	2022	5-20	100-10000	Альфа банк
3	Животные	Санкт-Петербург	2023	5-20	100-10000	ВТБ
4	Электроника	Санкт-Петербург	2024	5-20	500000-2000000	Сбер
5	Электроника	Санкт-Петербург	2023	5-20	10000-500000	ВТБ
6	Продукты	Санкт-Петербург	2021	5-20	10000-500000	ВТБ
7	Животные	Санкт-Петербург	2022	5-20	10000-500000	Т-банк
8	Электроника	Санкт-Петербург	2021	5-20	10000-500000	Альфа банк
9	Продукты	Санкт-Петербург	2021	5-20	10000-500000	Альфа банк

Рис 3. Пример датасета

```
Топ 5 уникальных значений k-анонимности:
K = 48, Процент = 0.0017%
K = 63, Процент = 0.0033%
K = 66, Процент = 0.0017%
K = 71, Процент = 0.0017%
K = 73, Процент = 0.0017%
```

Рис 4. Пример рассчитанных значений K

Вывод

В рамках данной работы был разработан алгоритм для обезличивания данных о покупках в магазинах. Программа анализирует существующий датасет, который включает в себя такие параметры, как названия магазинов, координаты, дата и время покупок, категории и бренды товаров, номера карт, количество и стоимость покупок. 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. Программа также рассчитывает значения K-анонимности, что позволяет пользователю оценить степень обезличенности данных.

Источники

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