САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ – ПРОЦЕССОВ УПРАВЛЕНИЯ

ОТЧЕТ

по лабораторной работе №3

по дисциплине «Алгоритмы и структуры данных»

на тему «Обезличивание данных»

Вариант 3 (просмотр рекламы)

Выполнила
студентка 2 курса
группы 21-Б15.ПУ
Павлова Ксения Андреевна

Преподаватель
Дик Александр Геннадьевич

Санкт-Петербург
2022

СОДЕРЖАНИЕ

Цель: написать программу обезличивания набора данных и высчитывания К-анонимити.

Задачи:

Часть пользовательская:

1) Программа должна считывать входной файл
2) Программа делиться по функционалу
   a) Обезличивание входного датасета.
   b) Высчитывание К-анонимити.
3) У пользователя есть выбор, какую функцию сделать. Возможность указывать Квази-идентификаторы в К-анонимити.

Часть программиста:

4) На ваше усмотрение, производим обезличивание дата сета по данным методам (комбинирование приветствуются, но главное - доказать эффективность данной комбинации методов):
   a) Локальное обобщение
   b) Агрегация
   c) Возмущение
   d) Микро-агрегация
   e) Перемешивание
   f) Создание псевдонимов
   g) Маскеризация
   h) Локальное подавление
   i) Удаление атрибутов
   j) Метод декомпозиции
5) Используя метод К-анонимити рассчитайте К для обезличенного набора.
6) Вывести плохих 5 значений К-анонимити (если их меньше, то все возможные). Данные переменные К вывести в процентах из всего набора.

7) Количество уникальный строк в дата-сете, согласно квази-идентификаторам. Вывести на экран уникальные строки, если переменная К=1.
8) Значение приемлемого К-анонимити для набора данных:
   a) до 51000 записей - К>=10
   b) до 105000 записей - К>=7
   c) до 260000 записей - К>=5
9) Оценивание полезности данных, если сравнивать обезличенный набор с исходным.

Описание методов.

I. Возмущение.

Настоящая техника предполагает внесение шума в данные, которые перестают быть точными или правдивыми, но сохраняют основные статистические закономерности. Метод применим для бинарных данных (да/нет, например, пола), цифровых BLOBS (фотографии), статистически частых наборов данных. Позволяет сохранить статистическую ценность набора данных при незначительных потерях информации.

II. 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.

III. Удаление атрибутов.

Под данной техникой понимается удаление чувствительного контента без добавления замен. Метод используется для удаления прямых идентификаторов, а также удаления избыточных квази-идентификаторов.

IV. 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.

V. Маскеризация.

В рамках данной техники проводится замена части записи заполнителями по определенному шаблону. Хотя в результате исходной строке сопоставляется строка определенного шаблона, что позволяет

отнести этот метод к псевдонимизации, по ряду характеристик этот метод может быть сопоставлен микро-агрегации: в результате применения шаблонов возникают обобщенные группы. Применим для идентификаторов (номера паспортов, телефоны, номера карт).

Блок-схема общего алгоритма.

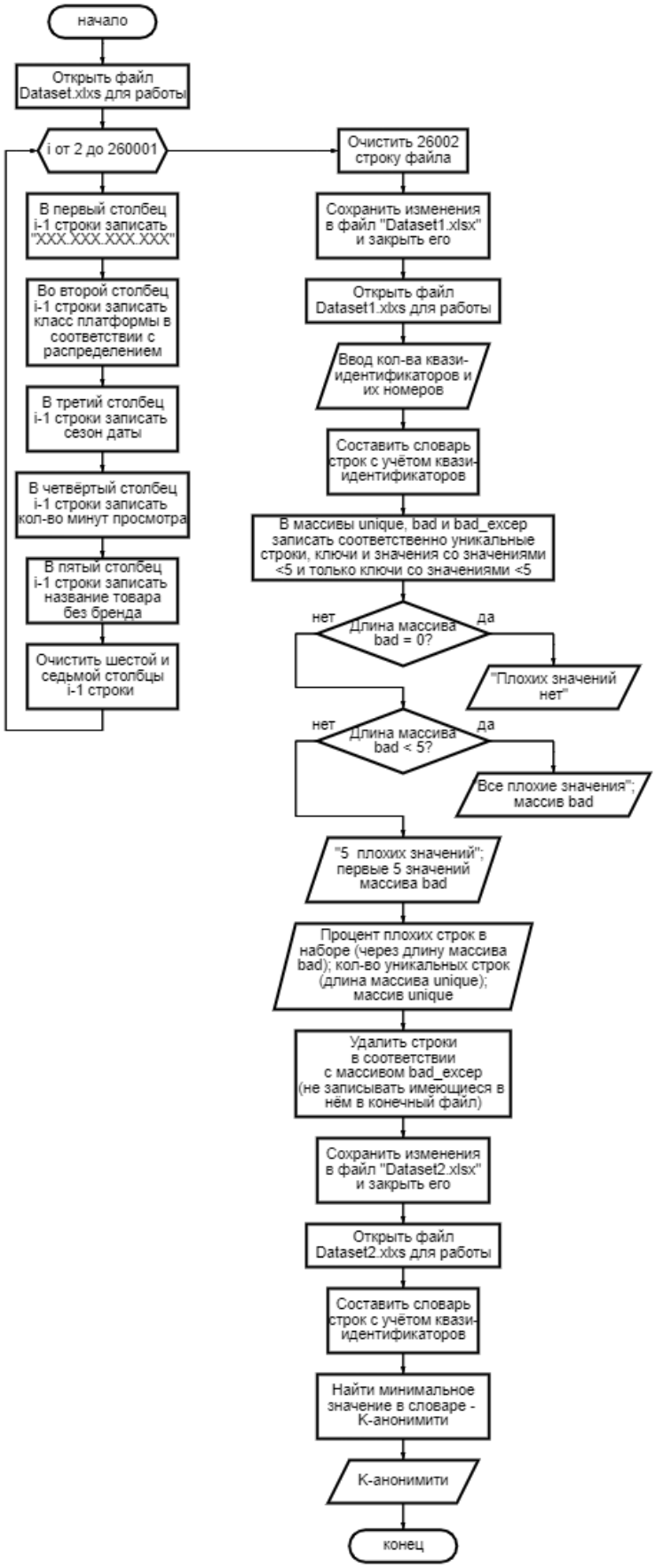

Рис. 1. Блок-схема общего алгоритма

Примечания к блок-схеме общего алгоритма (рис. 1).

Работа с файлами формата xlsx происходит с помощью библиотеки xlsxwriter.

Обезличивание происходит с помощью методов, описанных в разделе «Описание методов»:

- удаление атрибутов: столбец 1 «Пользователь» и столбец 5 «Кол-во рекламы»;
- маскеризация: столбец 2 «IP адрес» в формат «XXX.XXX.XXX.XXX»;
- локальное обобщение: столбец 3 «Платформа» по классам «Соц.сеть», «Видеохостинг», «Сайт знакомств», «Инфоресурс» и столбец 4 «Дата» по сезонам года;
- возмущение: столбец 6 «Время просмотра рекламы» (точное время заменяется только минутами) и столбец 7 «Вид рекламы» (товар с брендом заменяется только название товара);
- локальное подавление для строк с низким значением K-анонимити.

В создающихся словарях ключ – строка, состоящая из квази-идентификаторов, значение – количество таких строк.

K-анонимити рассчитывается как минимальное значение в построенном словаре.

Рекомендации программиста.

Для работы с кодом необходимо приложение PyCharm и установленные библиотеки time, xslxswriter.

Рекомендации пользователя.

Перед запуском программы убедитесь, что в папке с файлом программы находится таблица «Dataset.xlsx» с исходными данными.

Во время работы алгоритма необходимо вводить количество квази-идентификаторов и их номера. Максимальное возможное количество – 5. Номера квази-идентификаторов соответствуют номеру столбца в обезличенной таблице:

- 1 – IP адресс
- 2 – Платформа
- 3 – Дата (сезон)
- 4 – Время просмотра
- 5 – Вид рекламы (товар)

Контрольный пример.

В данном разделе представлен пример работы данной программы.

На рис. 2, рис. 3 и рис. 4 представлен результат работы программы.

Количество квази-идентификаторов – 5.

Номера квази-идентификаторов от 1 до 5.

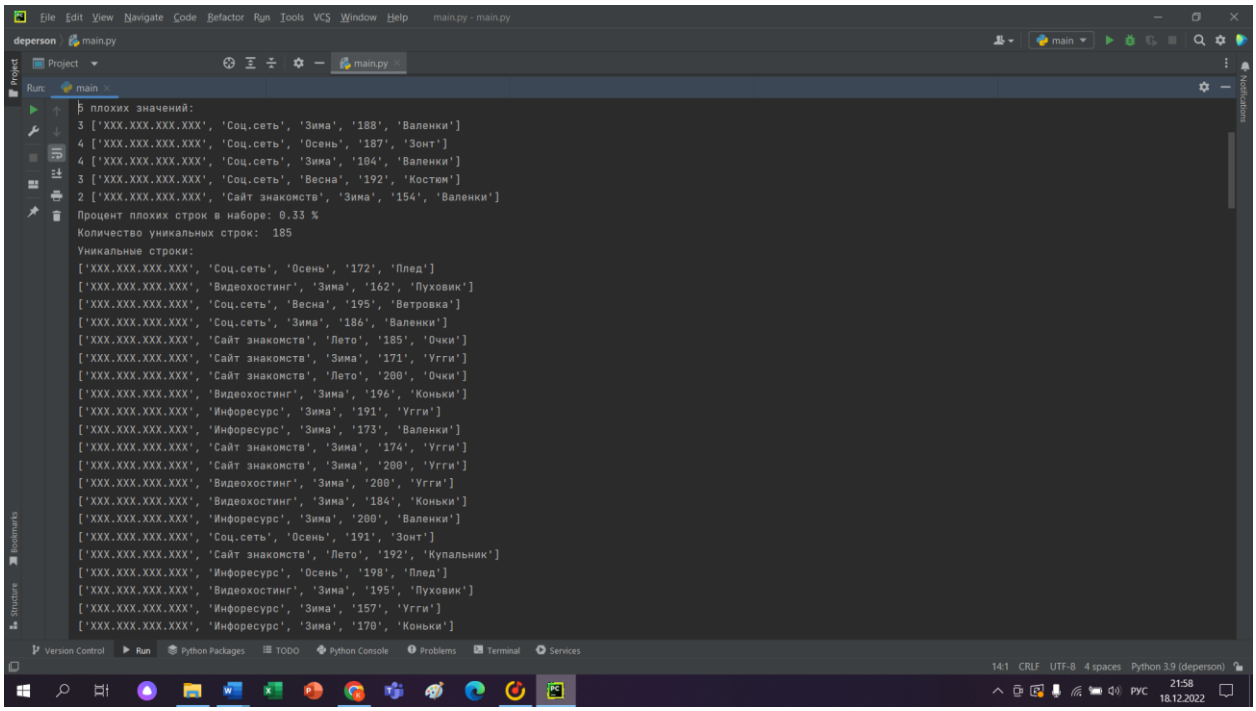

Рис. 2. Вывод результата работы программы

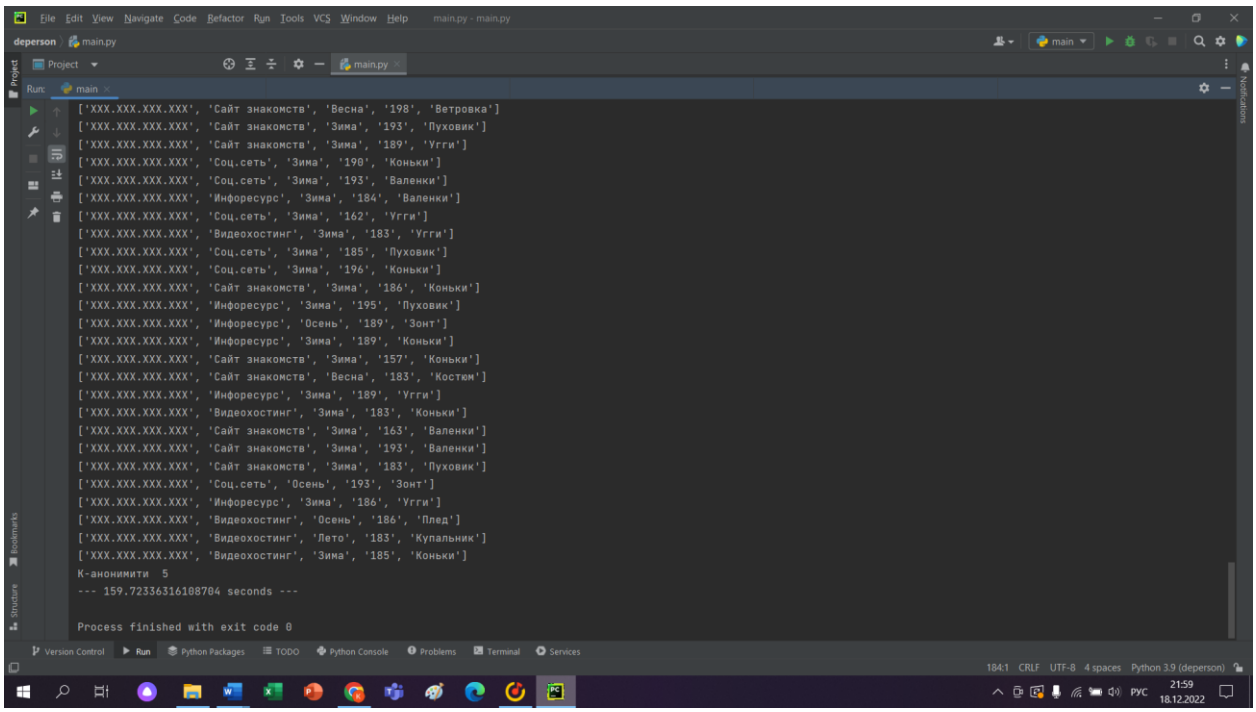

Рис. 3. Вывод результата работы программы

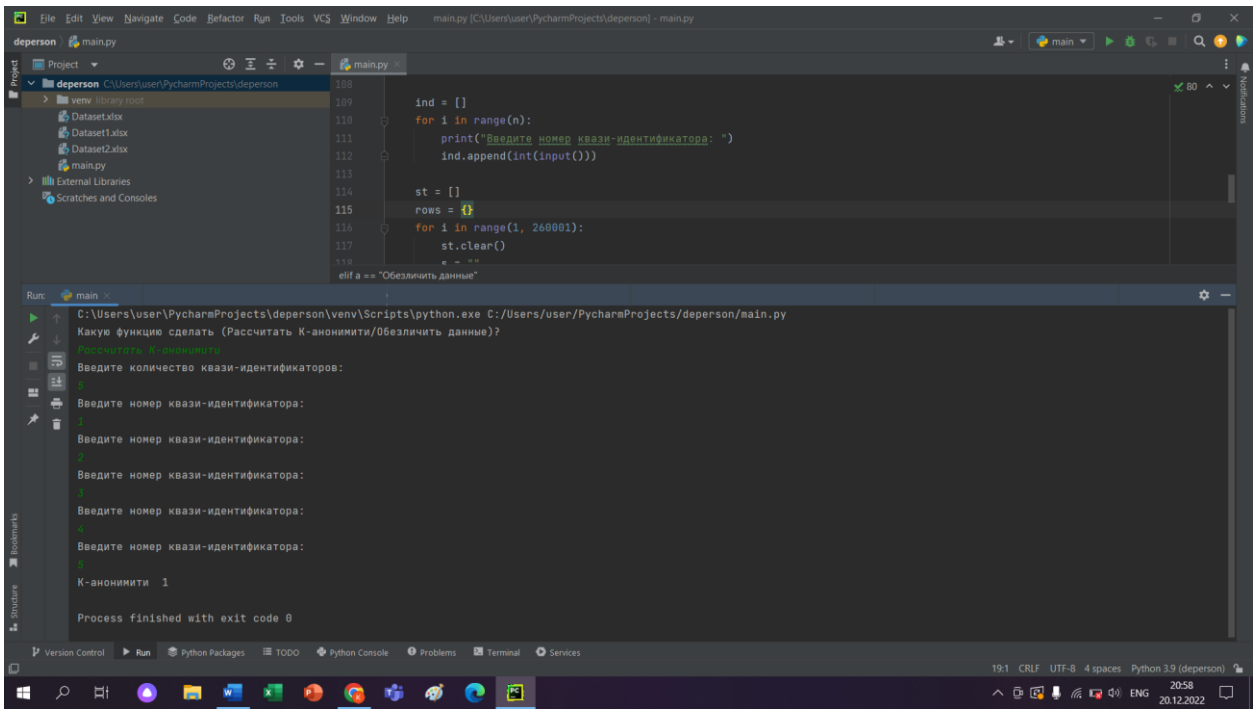

Рис. 4. Вывод результата работы программы

Вывод: в ходе выполнения данной работы была написана программа обезличивания набора данных и высчитывания K-анонимити.

Обезличивание происходит с помощью удаления атрибутов, локального обобщения, маскеризации, возмущения и локального подавления. Этих методов достаточно, так как при их использовании K-анонимити соответствует требуемому (>= 5 для 260000 записей). При этом конечные данные являются полезными, так как по ним можно собрать статистику, например, в каком классе платформ, какой товар и в какое время года рекламировался больше.

Список литературы.

1. Introducing Python. Modern computing in simple packages / Bill Lobanovic.
2. Презентация «Методы обезличивания» / аспирант Дик А.Г.
3. Ссылка на код, необходимый и полученные файлы: https://github.com/ksenkap/depersonalization