

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.....	3
Задачи.....	3
Описание методов .....	5
Блок-схема общего алгоритма.....	7
Примечания к блок-схеме общего алгоритма.....	8
Рекомендации программиста.....	9
Рекомендации пользователя .....	9
Контрольный пример .....	10
Вывод.....	12
Список литературы .....	13

**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 уникальных строк в дата-сети, согласно квази-идентификаторам. Вывести на экран уникальные строки, если переменная  $K=1$ .
- 8) Значение приемлемого  $K$ -анонимити для набора данных:
- а) до 51000 записей -  $K \geq 10$
  - б) до 105000 записей -  $K \geq 7$
  - с) до 260000 записей -  $K \geq 5$
- 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 отличие 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.

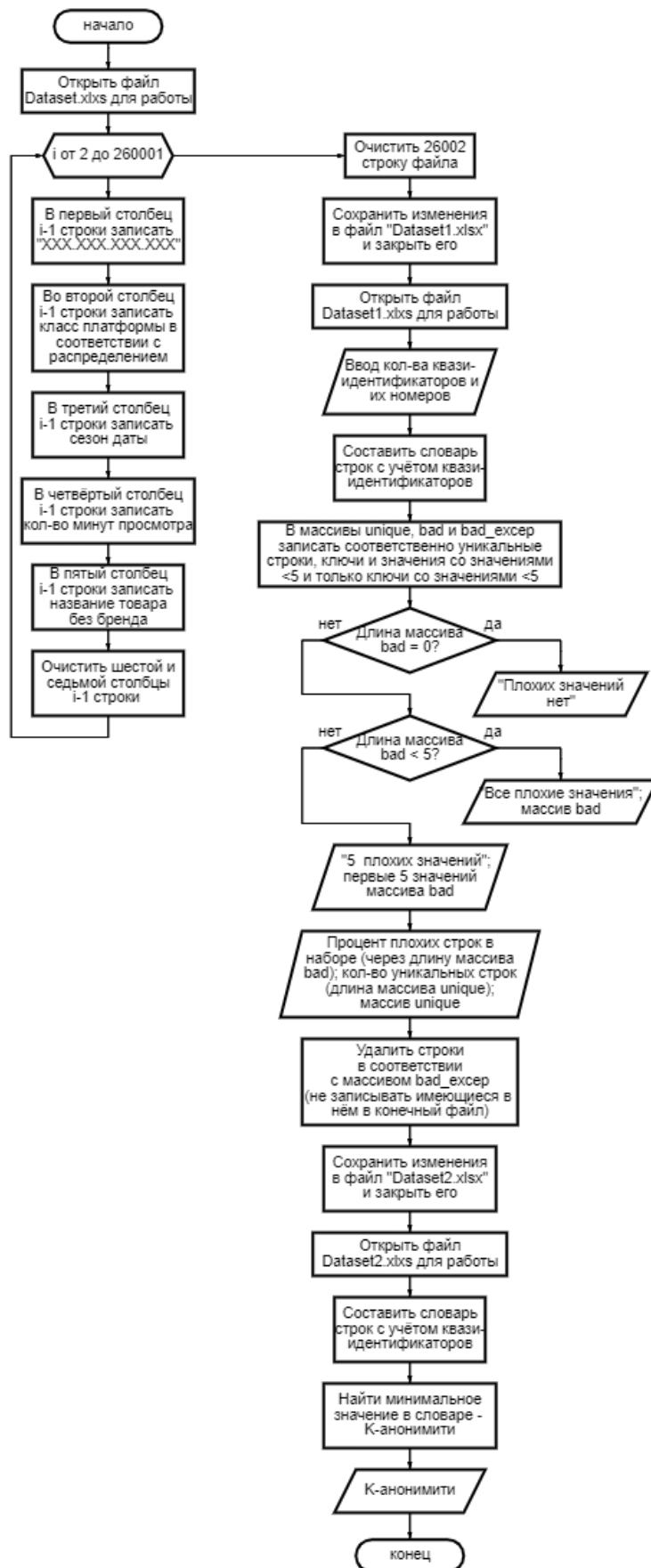


Рис. 1. Блок-схема общего алгоритма

## **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»;
- маскировка: 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 названием товара);
- локальное подавление для строк с низким значением К-анонимити.

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

К-анонимити рассчитывается как минимальное значение в построенном словаре.



### **Рекомендации программиста.**

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

### **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)

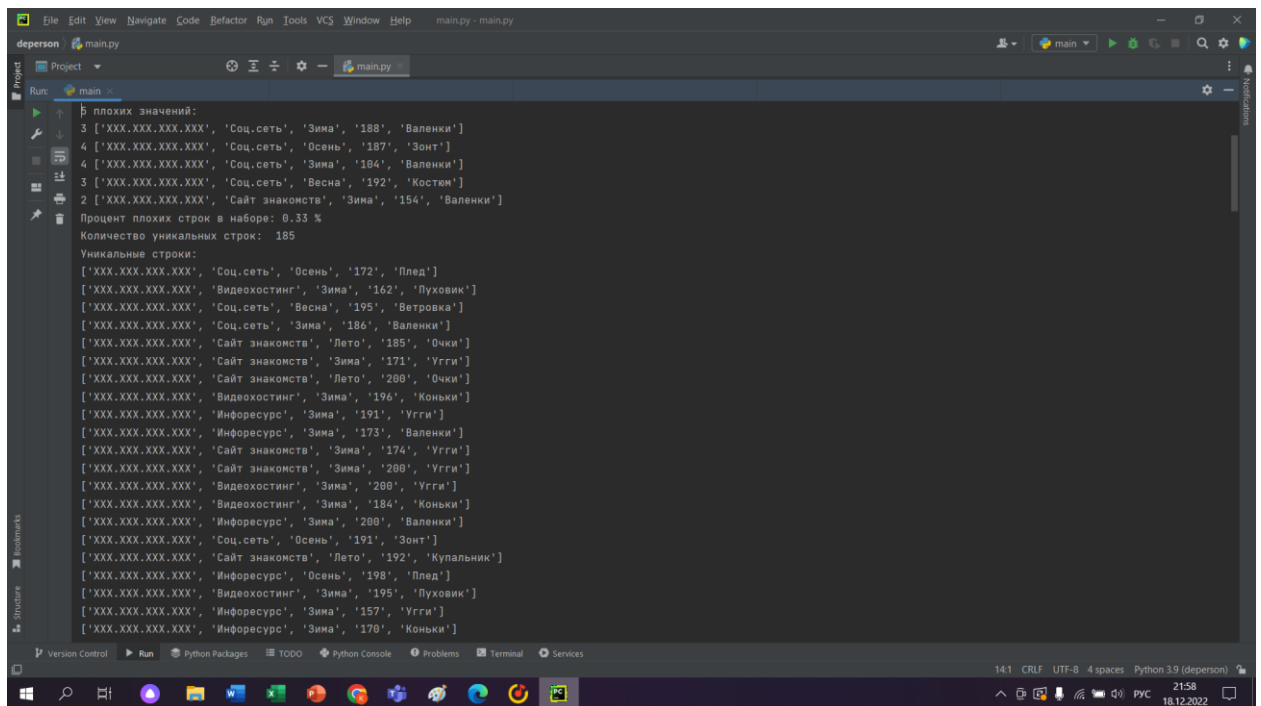
## Контрольный пример.

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.



```
5 плохих значений:
3 ['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '188', 'Валенки']
4 ['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Осень', '187', 'Зонт']
4 ['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '184', 'Валенки']
3 ['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Весна', '192', 'Костюм']
2 ['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '154', 'Валенки']

Процент плохих строк в наборе: 0.33 %
Количество уникальных строк: 185
Уникальные строки:
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Осень', '172', 'Плед']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '162', 'Пуховик']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Весна', '195', 'Ветровка']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '186', 'Валенки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Лето', '185', 'Очки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '171', 'Угги']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Лето', '200', 'Очки']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '196', 'Коньки']
['XXX.XXX.XXX.XXX', 'Инфоресурс', 'Зима', '191', 'Угги']
['XXX.XXX.XXX.XXX', 'Инфоресурс', 'Зима', '173', 'Валенки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '174', 'Угги']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '200', 'Угги']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '200', 'Угги']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '184', 'Коньки']
['XXX.XXX.XXX.XXX', 'Инфоресурс', 'Зима', '200', 'Валенки']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Осень', '191', 'Зонт']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Лето', '192', 'Купальник']
['XXX.XXX.XXX.XXX', 'Инфоресурс', 'Осень', '198', 'Плед']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '195', 'Пуховик']
['XXX.XXX.XXX.XXX', 'Инфоресурс', 'Зима', '157', 'Угги']
['XXX.XXX.XXX.XXX', 'Инфоресурс', 'Зима', '170', 'Коньки']
```

Рис. 2. Вывод результата работы программы

```

['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Весна', '198', 'Ветровка']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '193', 'Пуховик']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '189', 'Угги']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '198', 'Коньки']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '193', 'Валенки']
['XXX.XXX.XXX.XXX', 'Инфорекурс', 'Зима', '184', 'Валенки']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '162', 'Угги']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '183', 'Угги']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '185', 'Пуховик']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Зима', '196', 'Коньки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '186', 'Коньки']
['XXX.XXX.XXX.XXX', 'Инфорекурс', 'Зима', '195', 'Пуховик']
['XXX.XXX.XXX.XXX', 'Инфорекурс', 'Осень', '189', 'Зонт']
['XXX.XXX.XXX.XXX', 'Инфорекурс', 'Зима', '189', 'Коньки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '157', 'Коньки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Весна', '183', 'Костюм']
['XXX.XXX.XXX.XXX', 'Инфорекурс', 'Зима', '189', 'Угги']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '183', 'Коньки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '163', 'Валенки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '193', 'Валенки']
['XXX.XXX.XXX.XXX', 'Сайт знакомств', 'Зима', '183', 'Пуховик']
['XXX.XXX.XXX.XXX', 'Соц.сеть', 'Осень', '193', 'Зонт']
['XXX.XXX.XXX.XXX', 'Инфорекурс', 'Зима', '186', 'Угги']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Осень', '186', 'Плед']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Лето', '183', 'Купальник']
['XXX.XXX.XXX.XXX', 'Видеохостинг', 'Зима', '185', 'Коньки']
К-анонимити 5
--- 159.72336316108704 seconds ---
Process finished with exit code 0

```

Рис. 3. Вывод результата работы программы

```

108 ind = []
109 for i in range(n):
110     print("Введите номер квази-идентификатора: ")
111     ind.append(int(input()))
112
113
114 st = []
115 rows = 1
116 for i in range(1, 260001):
117     st.clear()
118     e = ""
119     elif a == "Обезличить данные"

```

```

C:\Users\user\PycharmProjects\deperson\venv\Scripts\python.exe C:/Users/user/PycharmProjects/deperson/main.py
Какую функцию сделать (Рассчитать К-анонимити/Обезличить данные)?
Введите количество квази-идентификаторов:
Введите номер квази-идентификатора:
Введите номер квази-идентификатора:
Введите номер квази-идентификатора:
Введите номер квази-идентификатора:
К-анонимити 1
Process finished with exit code 0

```

Рис. 4. Вывод результата работы программы

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

Обезличивание происходит с помощью удаления атрибутов, локального обобщения, маскирования, возмущения и локального подавления. Этих методов достаточно, так как при их использовании К-анонимити соответствует требуемому ( $\geq 5$  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>