**Санкт-Петербургский государственный университет Факультет прикладной математики и процессов управления**

**Лабораторная работа №4**

**Деобезличивание дата-сета**

**Вариант- 19**

|  |  |  |
| --- | --- | --- |
| Студент гр. Б15-ПУ |  | Шайдуров В.Д. |
| Преподаватель |  | Щеголева Н.Л. |

Санкт-Петербург

2022 г.

**Оглавление**

[**1. Цель работы 3**](#_heading=h.gjdgxs)

[**2. Задача 3**](#_heading=h.30j0zll)

[**3. Описание программы 3**](#_heading=h.1fob9te)

**4**[**. Вывод 5**](#_heading=h.4d34og8)

[**5. Ссылки: 6**](#_heading=h.2s8eyo1)

# **Цель работы**

Деобезличить исходный дата-сет.

# **Задача**

Получить информацию из представленного дата-сета student\_19.csv, применяя различные способы взлома.

# **Описание программы**

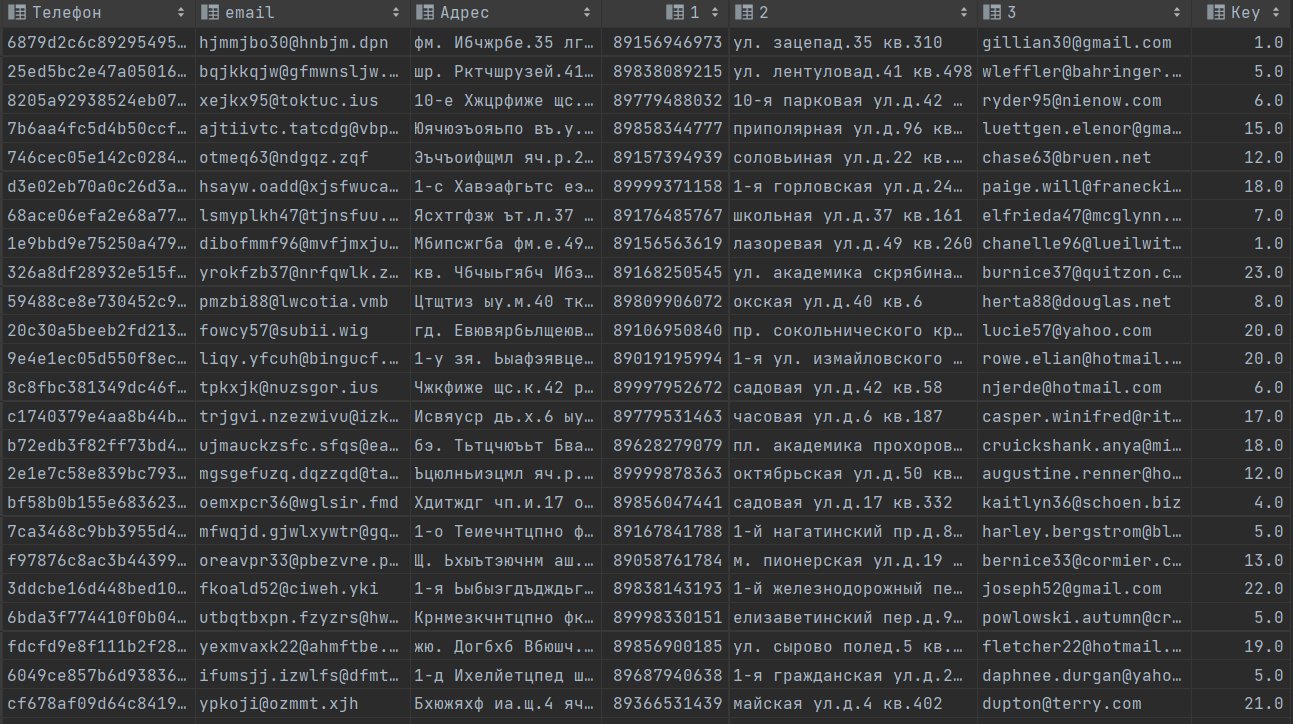
Исходный дата-сет имеет три столбца с указанием категории: номер телефона, электронная почта и адрес.

В строчках вместо номеров телефона был представлен результат хеш функции, которая была визуально определена как хеш-функция sha1. При помощи программного обеспечения hashcat данный столбец был взломан по маске состоящей из 11 цифр. Таким образом, работая через консоль, данная команда выглядела следующим образом (hashcat -m 100 -a 3 ./Downloads/student.txt -o ./Downloads/hucked.txt ?d?d?d?d?d?d?d?d?d?d?d - D 1) был получен результат взлома, который был записан в текстовый файл hucked.txt. Student.txt это файл в которой был скопирован столбец с номерами телефонов. Из-за специфики работы hashcat в сохранённом файле hucked.txt была напутана исходная последовательность, поэтому первым действием программы является распределения номеров телефонов в правильном порядке и их запись в соседнюю колонку под именем “1” в исходный дата-сет student\_19.csv.

Следующим шагом было, так же, визуальное определение шифра цезаря в столбцах: электронная почта и адрес (легче всего это было заметить на местах в которых должен быть записан провайдер почты @gmail и так далее или на местах адреса, где должно стоять обозначение квартиры [кв.]). Исходя из данной информации, путём перебора были подобраны необходимые сдвиги для каждой строки, таким образом используя сдвиги удалось расшифровать второй и третий столбцы исходной таблицы. Получившийся результаты были записаны в исходный файл student\_19.csv под колонками “2” и “3”. А также была добавлена колонка “Key”, которая показывает на какой сдвиг была преобразована строка. Как оказалось в дальнейшем, что значение почты и адреса имеют одинаковые сдвиги, поэтому было решено оставить общий столбец “Key” для почты и адреса.

На рисунке 3.1 представлен получившийся результат деанонимизации.

*Рисунок 3.1*



# **Вывод**

В ходе выполнения данной работы были определены методы, которыми пытались скрыть информацию в дата-сете student\_19.csv и было успешно реализовано её раскрытие.

# **Ссылки:**

С программной реализацию можно познакомиться, перейдя по ссылке [github](https://github.com/Mooncake911/Generation-of-medical-data).