Санкт-Петербургский государственный университет

Факультет прикладной математики – процессов управления

Лабораторная работа №1

**Отчет**

по дисциплине «Алгоритмы и структуры данных»

Генерация синтетических данных

Вариант 4

Автор работы: Добренкова Л.С.

Группа: 22.Б15-пу

Преподаватель: Дик А.Г.

Санкт-Петербург, 2023

Оглавление

[1. Цель работы 3](#__RefHeading___Toc4888_848729285)

[2. Задачи 3](#__RefHeading___Toc4890_848729285)

[3. Теоретическая часть 3](#__RefHeading___Toc4892_848729285)

[4. Алгоритм генерации 4](#__RefHeading___Toc4894_848729285)

[5. Описание программы 10](#__RefHeading___Toc4896_848729285)

[6. Рекомендации пользователя 11](#__RefHeading___Toc4898_848729285)

[7. Рекомендации программиста 12](#__RefHeading___Toc12522_3894689155)

[8. Контрольный пример 13](#__RefHeading___Toc12524_3894689155)

[9. Вывод 14](#__RefHeading___Toc4900_848729285)

# **Цель работы**

Целью настоящей лабораторной работы является разработка программы с графическим интерфейсом для генерации датасета, представляющего собой информацию о пациентах, обратившихся в частную поликлинику с возможностью установки вероятностей для генерации различных полей.

# **Задачи**

1. Изучить методы генерации данных.
2. Выбрать методы генерации, подходящие под ограничения.
3. Создать словари для генерации.
4. Добавить графический интерфейс.
5. Добавить возможность менять вероятности генерации.
6. Протестировать программу.

# **Теоретическая часть**

Формат паспорта состоит из серии - 4 цифры и номера - 6 цифр. В серии первые 2 цифры – номер региона, в котором выдан паспорт, вторые 2 – год выдачи. Номер является просто счетчиком при совпадении серий (паспорта должны быть уникальны).

СНИЛС представляет собой 11 цифр, 2 последние являются хэш-суммой остальных. Для упрощения это будут случайные цифры. Остальная часть – счетчик (СНИЛСы должны быть уникальны).

Даты обращения пациента и результата анализов должны быть рабочими днями в рабочее время.

Банковская карта состоит из 16 цифр. Первая цифра – номер платежной системы (например 2 – мир). Следующие 5 цифр – номер банка в этой платежной системе. Остальные – счетчик.

Количество возможных комбинаций симптомов не менее 5000. Анализов - 250.

При разработке алгоритма программы требуется учесть все описанные условия.

# **Алгоритм генерации**

С помощью интерфейса пользователь указывает, в каком процентном соотношении встречаются банковские карты и платежные системы. Программа начинает свою работу только при условии, что общая вероятность всех событий составляет ровно сто процентов.

Генерируется пол пациента, затем его ФИО (случайно берутся из словарей).

Генерируется серия паспорта, по ней берется номер (счетчик) и увеличивается на 1.

Генерируется СНИЛС – первые 9 цифр – счетчик, увеличивающийся на 1. 2 последние цифры генерируются случайно.

Генерируются симптомы пациента – набор размером 7 из 15 возможных вариантов. Количество возможных симптомов – сочетания из 15 по 7 = 6435

Генерируется проводящий прием врач.

Генерируются назначаемые анализы. Аналогично симптомам, но из 10 по 5 = 252 возможных вариантов.

Генерируется дата и время приема, а также дата и время посещения врача (в течение 24-72 часов с момента приема). Для генерации случайной даты выбирается случайное количество дней до текущего момента. Если этот день выходной, вычитаем из даты 2 дня (чтобы не попасть снова в выходной). Время выбирается случайно из диапазона 8-17ч.

Дата получения результатов анализов получается с помощью сдвига даты обращения на день вперед. Если попали в выходной (это может быть только суббота), сдвигаем еще на 2 вперед, попадая в понедельник.

Генерируется стоимость анализов (шаг цены 100).

Все полученные данные записываются в файл.

На рисунках 1-7 представлена блок-схема алгоритма.

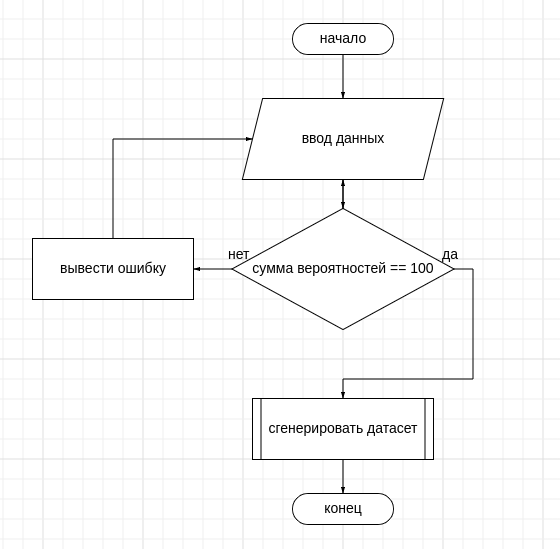


Рис. 4.1 Блок-схема основной программы.



Рис. 4.2 Блок-схема генерации датасета.

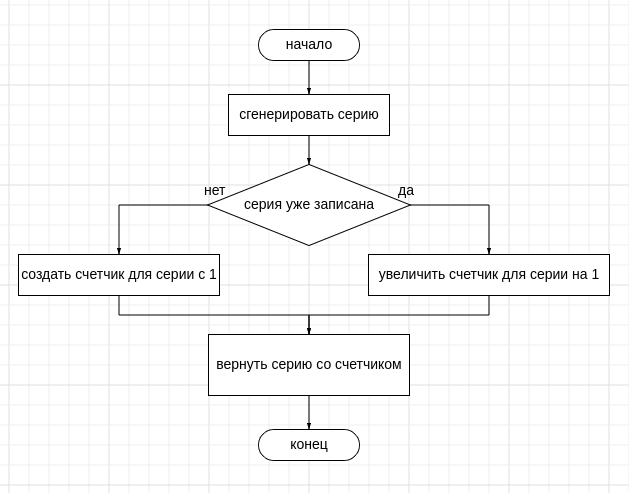


Рис. 4.3 Блок-схема метода generate класса PassportGenerator



Рис. 4.4 Блок-схема метода generate класса SnilsGenerato

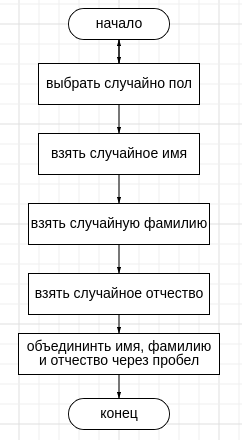


Рис. 4.5 Блок-схема метода generate класса NamesGenerator

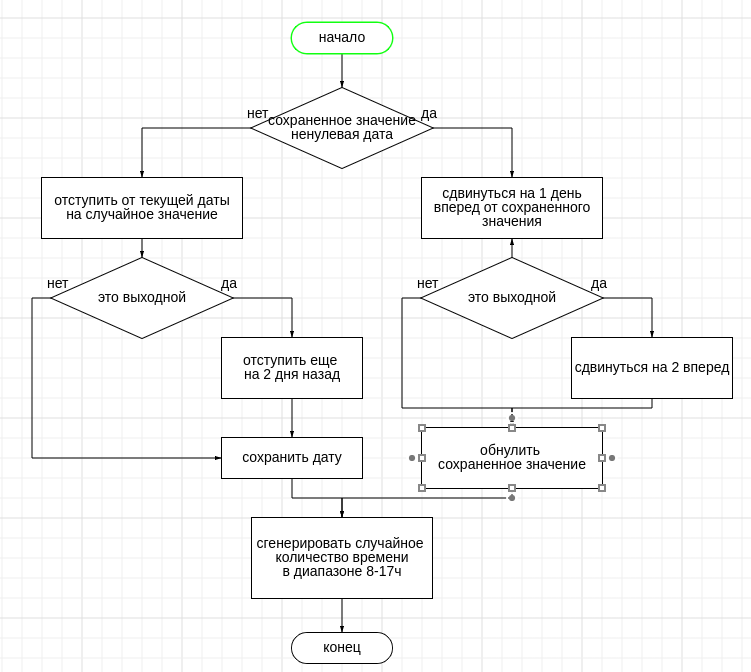


Рис. 4.6 Блок-схема метода generate класса DatetimeGenerator

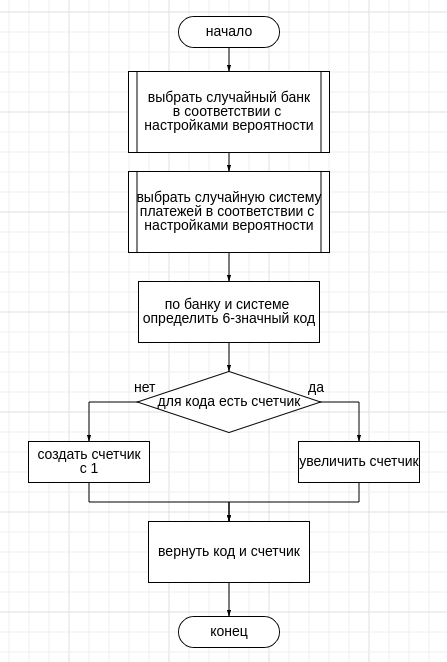


Рис. 4.7 Блок-схема метода generate класса CardGenerator

# **Описание программы**

Программа реализована на языке python 3.10 с использованием зависимостей: datetime, tkinter.

В программе используются 6 классов и 1 функция для генерации данных.

Таблица 5.1 Описание классов

|  |  |
| --- | --- |
| Имя класса | Назначение |
| PassportGenerator | Генерирует случайную серию, от нее берется номер, являющийся счетчиком. Хранит счетчики для каждой серии. |
| SnilsGenerator | Генерирует СНИЛС - первые 9 цифр – счетчик, 2 последние случайные. Хранит счетчик СНИЛСов. |
| NamesGenerator | Генерирует ФИО. Каждый элемент берется случайно из хранящейся соответствующей базы. |
| DatetimeGenerator | Генерирует дату и время. Дата всегда – будний день. При первом запуске - случайная дата. При втором запуске – следующий день от предыдущей генерации. Случайное рабочее время. |
| SamplesGenerator | Генерирует набор симптомов или анлизов. Хранит соответствующие базы. |
| CardGenerator | Учитывая вероятности генерирует банк, систему, берет 6-значный ключ. Остальное – счетчик. |

Таблица 5.2 Описание функций

|  |  |  |
| --- | --- | --- |
| Имя функции | Параметры функции | Назначение |
| compute | n, out\_file, banks\_p, systems\_p | Генерирует n записей, сохраняет в файл out\_file. |

# **Рекомендации пользователя**

Для запуска программы необходимо установить python версии не ниже 3.10, а так же зависимости tkinker, datetime.

В каталоге «dictionary» с программой необходимо иметь специальные файлы словари. В директории должно быть 10 текстовых документов:

1. analyzis.txt - файл, содержащий словарь назначаемых анализов.
2. card\_keys.txt - файл, содержащий словарь кодов банков в каждой платежной системе.
3. doctors.txt - файл, содержащий словарь врачей, работающих в поликлинике и проводящих приемы.
4. symptoms.txt – файл, содержащий словарь симптомов, с которыми обращается пациент
5. names\_1.txt – файл, содержащий словарь славянских мужских имен.
6. names\_2.txt – файл, содержащий словарь славянских женских имен.
7. lastnames\_1.txt – файл, содержащий словарь славянских мужских фамилий.
8. lastnames\_2.txt – файл, содержащий словарь славянских женских фамилий.
9. patronymics\_1.txt – файл, содержащий словарь славянских мужских отчеств.
10. patronymics\_2.txt – файл, содержащий словарь славянских женских отчеств.

После запуска программы требуется:

1. Выбрать вероятности для генерации карт оплаты: платежных систем и банков. Если в одной из характеристик суммарная вероятность не дотянет или превысит 100, появится уведомление об ошибке.
2. В поле «Count» необходимо ввести одно целое число – количество строк генерации.
3. Для генерации необходимо заполнить все поля, нажать «Compute» дождаться уведомления об окончании. Файл с датасетом будет сохранен как значение поля «Output filename».

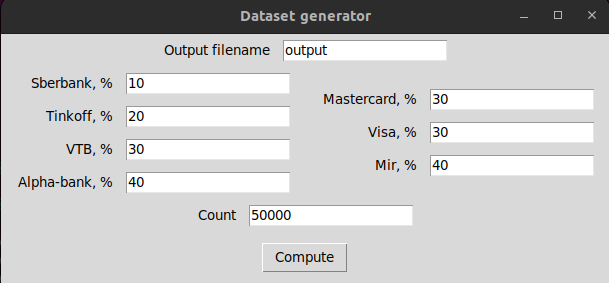
# **Рекомендации программиста**

Для запуска программы необходима 64-битная операционная система Windows, Linux или macOS. Для работы с кодом необходима среда разработки, совместимая с python 3.1 и библиотеки datetime и tkinter.

Исходный код программы и необходимые текстовые файлы доступны по ссылке: <https://github.com/v131v/algo_labs/tree/main/lab1>

# **Контрольный пример**

В данном разделе представлен пример, демонстрирующий работу программы.

Рис. 8.1 Ввод данных в интерфейс программы

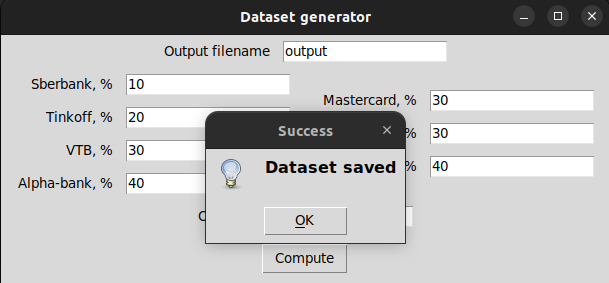


Рис. 8.2 Уведомление об успешной генерации

# **Вывод**

В результате выполнения данной работы разработана программа генерации датасета с синтетическими данными платной поликлиники.