НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСИС»

ИНСТИТУТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

И КОМПЬЮТЕРНЫХ НАУК

КАФЕДРА ИНЖЕНЕРНОЙ КИБЕРНЕТИКИ

Описание алгоритма

нахождения минимального набора признаков для идентификации сущности

Выполнил учащийся:

Ромадова Ирина Олеговна

группа БПМ-22-4

Москва

2024 год

Данный проект реализует алгоритм поиска минимального набора признаков для однозначной идентификации сущности.

Алгоритм написан на языке Python и осуществляет поиск атрибутов в JSON-файле.

Структура проекта:

- app.py Оновной файл, где реализован алгоритм и вспомогательные функции.
- test.py Файл с тестами основных компонент.
- test.json пример входных данных.
- unit_test.json файл для тестирования.

Идеи для алгоритма:

Ниже описаны решения, которые были приняты в целях оптимизации

1. Уточним задачу следующим образом: требуется найти такой минимальный набор параметров, такой, что у любой пары записей, хотя бы один признак будет принимать разные значения
2. Было приято решение хранить информацию о записях и признаках для пары, для удобства
3. В реализации алгоритма был использован бинарный поиск для нахождения ответа, чтобы оптимизировать поиск.
4. В целях минимизирования перебора даже с использованием бинарного поиска было реализовано следующее: рассмотрим комбинации длины m/4 и 3m/4. (где m – число имеющихся у записи признаков). Если для комбинации длины m/4 найден ответ, то нет смысла рассматривать более длинные комбинации, потому что для них в данном случае заведомо будет иметься ответ. Аналогично с комбинациями длины 3m/4. Таким образом, появляется возможность

отсечь большое количество комбинаций, В худшем случае потребуется рассмотреть комбинации, которые лежат между ними.

Описание основых этапов работы алгоритма:

1. На вход алгоритму подается путь к файлу и имя для файла куда будет записан результат. В случае, если не указано имя для файла куда запишется ответ, то имя будет дано по умолчанию: output.csv. Далее происходит преобразование входных данных в формат, который будет удобен для обработки.
2. После этого определяется список признаков, среди которых нужно найти минимальный набор и заполняется "Массив сравнений", где хранится инфрмация о том, различен ли определенный признак у пары записей.
3. Проверка комбинаций длины m/4 и 3m/4, в целях оптимизации бинарного поиска
4. Производится поиск комбинаций минимальной длины с помощью бинарного поиска. Если наша комбинация является ответом, то достаточно, чтобы хотя бы одно значение признака было равно разным для каждой пары. Если мы нашли комбинацию, которая является ответом и при этом короче текущего ответа, мы обновляем результат.
5. Далее происходит запись в CSV файл по пути, который указывается при запуске.

Схема алгоритма:

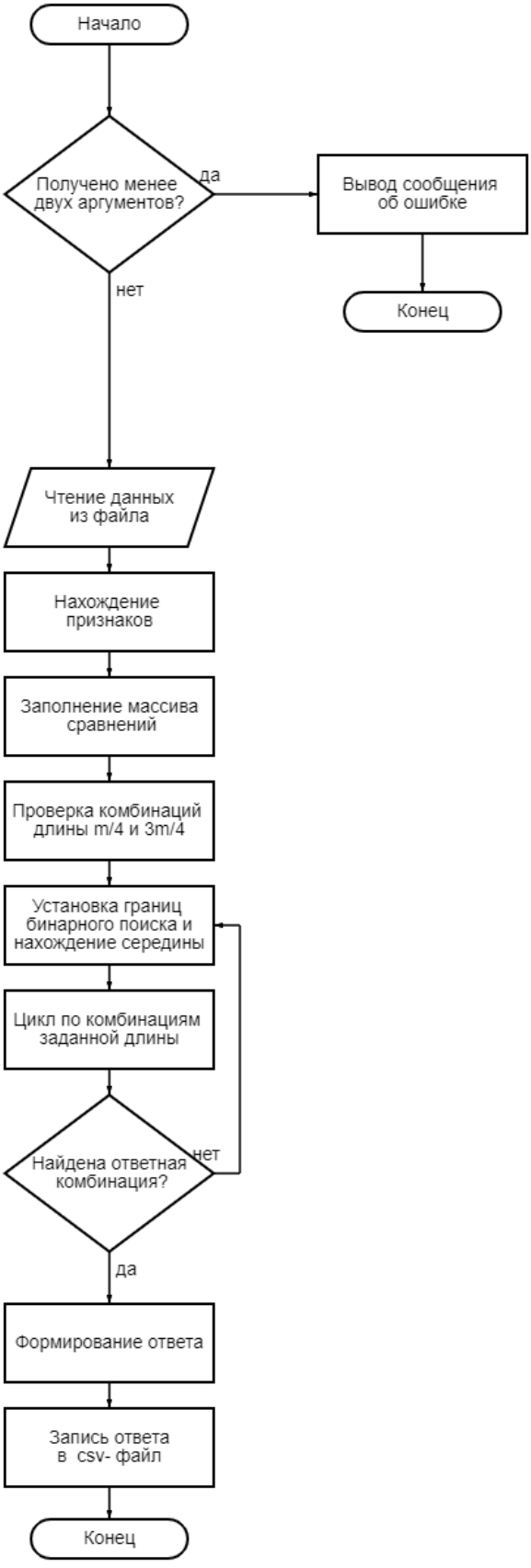

Рисунок 1 описание алгоритма