Опис програми для «проріджения» даних і обчислення статистики параметра повторюваності

Постановка проблеми

Нехай досліджуваний текст складає за обсягом, наприклад, 10 МБ. За умови кодування UTF-8, це означає, що він містить 1 млн. символів. Такі тексти аж ніяк не можна вважати рекордно довгими, адже для дослідження параметра повторюваності нам потрібні великі тексти для якомога повнішої статистики.

Тоді за найпростішої умови запису в текстовий формат, файл звіту для параметра повторюваності міститиме номери позицій всіх символів і числа (наприклад, з точністю до 6–12 знаків після коми), які відповідають значенням параметра повторюваності на кожній позиції. Це особливість алгоритму та програми розрахунку параметра повторюваності – адже повторюваність неможливо розрахувати, пропустивши хоча би одну позицію в тексті.

Отже, розмір вихідного файлу складатиме порядку 100–300 МБ. Збереження даних до менш економних форматів означатиме ще більші розміри вихідного файлу. Іншими словами, постають проблеми збереження та подальшої обробки або графічного представлення даних v(t), якщо планується використати стандартні програмні засоби для цього (Excel, Origin тощо).

Для запобігання цим проблемам слід «прорідити» дані для параметра повторюваності. Найпростіша схема така: залишаємо в файлі звіту перші N точок, а далі беремо лише одну точку із P точок.

Для прикладу, можна обрати такі величини згаданих параметрів:

$$N = 1000, P = 1000.$$

Так можна на кілька порядків зменшити розміри файлів звіту, не надто втративши на інформативності графіка v(t).

Інша дрібна технічна проблема полягає в тому, що згадана програма формує csv файл, у якому різні точки розміщені не в різних стовпцях таблиці, а в різних рядках. Це не завжди зручно на практиці; до того ж, максимальна кількість рядків, наприклад, в Excel, є набагато меншою за максимальну кількість стовпців.

Вирішення проблеми

Програма для вирішення згаданих проблем була розроблена мовою C#, в середовищі Visual Studio. Запуск програми відбувається через .exe файл. Після цього відбудеться запуск самої програми. Далі необхідно відкрити csv файл.

Перша частина функціональності програми – це зменшення та транспонування заданого файлу за відповідними параметрами. Програма зберігає мінімізований файл зі значеннями введених параметрів у назві. Наприклад *M10_L10000_space0_bar5_minified_N100_P1000.csv.*

Друга – знаходження середнього значення та середнього квадратичного відхилення для параметра повторюваності на відрізку від заданого значення і до кінця файлу.

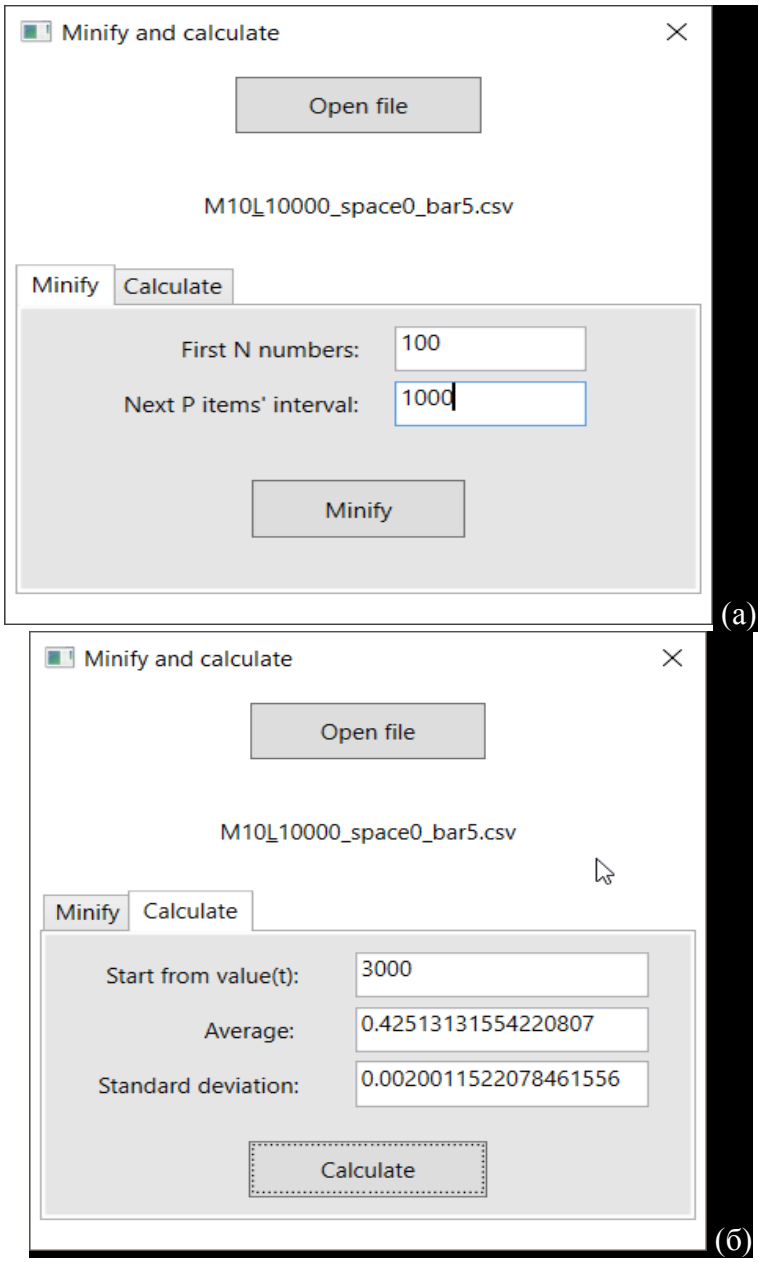

Рис 3.6. а) Режим проріджения б) Режим знаходження статистики, реалізовані в нашій програмі