В. А. Стромский
(*ГрГУ им. Я. Купалы, Гродно*)

К ВОПРОСУ КОДИРОВАНИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

Современные цифровые носители информации позволяют хранить широкий спектр различных её видов при том, что в памяти большинства устройств могут непосредственно храниться только двоичные значения – 0 и 1. Это возможно благодаря разнообразным кодировкам – способам представления информации последовательностями бит. В случае текстовой информации, чем больше различных символов необходимо закодировать, тем длиннее должна быть последовательность битов. Сопоставляя каждому символу последовательность из 8 бит,

можно закодировать $2^8 = 256$ различных символов. Одна из наиболее распространённых в мире кодировок, UTF-8, поддерживает подавляющее большинство печатных символов языков мира, используя ряд правил и последовательности длинной от 1 до 4 байтов (символы кириллицы кодируются 2 байтами). Таким образом, нередки случаи, когда для кодирования каждого символа текста содержащего не более 2^n уникальных символов используется более n бит, что лишний раз расходует память компьютера. Следовательно, в большинстве случаев способ хранения текстовой информации можно оптимизировать.

В настоящей работе разработана программа, которая кодирует текстовую информацию по следующему алгоритму: составление списка всех n уникальных символов, используемых в конкретном тексте; запись в начало нового кодового представления текста 5 бит, хранящих длину x битовой последовательности, кодирующей один символ данного текста ($x = \lceil log_2 n \rceil$); запись последовательности всех n уникальных символов в изначальной кодировке, что необходимо для декодирования; запись самой текстовой информации, где каждый символ кодируется своим номером в записанном ранее списке уникальных символов (числом длинной x в двоичной системе счисления).

Закодированная таким образом моно язычная текстовая информация в большинстве случаев будет занимать намного меньше памяти (как минимум, на 12.5% меньше для объёма свыше 1 Кб при изначальной однобайтной кодировке (напр., ANSI) и на более 50% для объёма свыше 2 Кб при изначальной двухбайтной кодировке (например, кириллица в UTF-8)).