Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Кафедра вычислительных систем

ОТЧЕТ
по практической работе 3

по дисциплине «**Программирование**»

Выполнил:
студент гр. ИС-241
«25» апреля 2023 г. _____ /Олейников Д.О./

Проверил:
Старший преподаватель
«25» апреля 2018 г. _____ /Фульман Д.О./

Оценка «_____»

Новосибирск 2023

ОГЛАВЛЕНИЕ

ЗАДАНИЕ

Работа состоит из двух частей.

В первой части вам предлагается разобрать алгоритм кодирования целых чисел, называемый varint (**var**iable **int**eger). Такой способ кодирования позволяет использовать переменное количество байт для представления целых чисел и благодаря этому обеспечивает компактность данных. Вам нужно разработать приложение для записи и чтения чисел в сыром виде и в формате varint, сравнить эти два способа по эффективности и сделать выводы о применимости предложенного способа кодирования.

Во второй части необходимо самостоятельно реализовать алгоритм кодирования UTF-8. Этот алгоритм решает аналогичную задачу — позволяет кодировать целые числа переменным количеством байт, но используется для кодирования кодов символов и поддерживает обратную совместимость с кодировкой ASCII.

Задание 1

Разработайте приложение, которое генерирует 1000000 случайных чисел и записывает их в два бинарных файла. В файл `uncompressed.dat` запишите числа в несжатом формате, в файл `compressed.dat` — в формате varint. Сравните размеры файлов.

Реализуйте чтение чисел из двух файлов. Добавьте проверку: последовательности чисел из двух файлов должны совпадать.

Использование формата varint наиболее эффективно в случаях, когда подавляющая доля чисел имеет небольшие значения. Для выполнения работы используйте функцию генерации случайных чисел:

Задание 2

Разработать приложение для кодирования и декодирования чисел по описанному выше алгоритму.

Пример кодирования:

```
$ ./coder encode points.txt units.bin
```

ВЫПОЛНЕНИЕ РАБОТЫ

1 Задание.

В данном задание нам требуется 3 функции: для кодирования чисел, для декодирование, и для генерации случайного числа.

```
size_t encode_varint(uint32_t value, uint8_t *buf)
{
    assert(buf != NULL);
    uint8_t *cur = buf;
    while (value >= 0x80)
    {
        const uint8_t byte = (value & 0x7f) | 0x80;
        *cur = byte;
        value >>= 7;
        ++cur;
    }
    *cur = value;
    ++cur;
    return cur - buf;
}
```

Функция **encode_varint** кодирует целочисленное значение value в формате Base 128 Varint и сохраняет результат в буфер buf.

В самом начале идёт проверка, с помощью макроса assert, равен ли buf NULL.

Переменная cur указывает на текущую позицию в buf. В цикле while проверяется, является ли значение value больше или равным 0x80. Если это так, значит, должен записаться ещё один байт в буфер, т.к. каждый байт в формате Base 128 Varint содержит 7 бит данных и 1 бит продолжения. Создаётся byte, который содержит младшие 7 битов значения value и бит продолжения (0x80). Затем записывается этот в текущую позицию cur в буфере buf, сдвигается значение value на 7 бит вправо и увеличивается значение cur на 1. Когда значение value становится меньше, чем 0x80, записывается оставшееся значение value в текущую позицию в буфере buf и увеличивается значение cur на 1.

```
uint32_t decode_varint(const uint8_t **bufp)
{
    const uint8_t *cur = *bufp;
    uint8_t byte = *cur++;
    uint32_t value = byte & 0x7f;
    size_t shift = 7;
    while (byte >= 0x80)
    {
        byte = *cur++;
        value += (byte & 0x7f) << shift;
        shift += 7;
    }
    *bufp = cur;
    return value;
}
```

Функция **decode_varint** декодирует значение Base 128 Varint из буфера bufp и возвращает раскодированное значение.

Переменная cur указывает на текущую позицию в буфере bufp. Первый байт в буфере считывается в переменную byte, а младшие 7 битов этого байта сохраняются в переменной value. Затем цикл while, который продолжается до тех пор, пока значение байта byte имеет установленный старший бит (0x80). Внутри цикла считывается следующий байт из буфера bufp в переменную byte и добавляются младшие 7 битов этого байта к переменной value, сдвигая их на shift битов влево. Затем увеличивается значение shift на 7, чтобы готовиться к чтению следующего байта.

После того, как прочитались все байты, необходимые для кодирования значения value, устанавливается указатель bufp на следующую пустую позицию в буфере, где закончилось чтение. Возвращается раскодированное значение value.

```
uint32_t generate_number()
{
    const int r = rand();
    const int p = r % 100;
    if (p < 90)
    {
        return r % 128;
    }
    if (p < 95)
    {
        return r % 16384;
    }
    if (p < 99)
    {
        return r % 2097152;
    }
    return r % 268435455;
}
```

Функция generate_number генерирует случайное число.

```
size_t write_bin_file(FILE *file_unc, FILE *file_c)
{
    assert(file_unc != NULL);
    assert(file_c != NULL);

    uint8_t buf[4] = {};
    size_t size = 0;
    size_t full_size = 0;

    for (int i = 0; i < N; i++)
    {
        uint32_t numb = generate_number();
        fwrite(&numb, sizeof(numb), 1, file_unc);
        size = encode_varint(numb, buf);
        full_size += size;
        fwrite(buf, size, 1, file_c);
    }

    return full_size;
}
```

Функция **write_bin_file** записывает сгенерированные числа в два файла: в несжатом формате, в формате varint. Макрос assert проверяет не равны ли файлы NULL. Создаётся массив буфер, для передачи данных в функцию кодирования числа. Создаётся цикл для генерации случайных чисел и запись в файлы сначала без кодирования, а затем с кодированием. Возвращается полный размер закодированных чисел.

```c
// считывать из файла
int read_bin_file(FILE *file_unc, FILE *file_c, size_t size)
{
    assert(file_unc != NULL);
    assert(file_c != NULL);

    fseek(file_c, 0, SEEK_SET);
    fseek(file_unc, 0, SEEK_SET);

    uint32_t num_uncomp, num_comp;
    uint8_t *buf = malloc(sizeof(uint8_t) * size);
    const uint8_t *cur = buf;
    fread(buf, sizeof(uint8_t), size, file_c);

    for (int i = 0; i < N; i++)
    {
        num_comp = decode_varint(&cur);
        fread(&num_uncomp, sizeof(uint32_t), 1, file_unc);
        if (i < 5)
            printf("decode %d encode %d\n", num_comp, num_uncomp);
        if (num_comp != num_uncomp)
            return -1;
    }

    free(buf);
    return 0;
}
```

Функция, которая считывает цифры из файлов. Для начала, проверка с помощью макроса assert равны ли файлы NULL. С помощью fseek указатель перемещается на начало файла. Далее выделяется память для раскодирования чисел. Затем в buf считываются данные из кодированного файла. Числа декодируются и выводится 5 сообщений с кодированным значением и декодируемым. В результате buf очищается.

```c
int main()
{
    srand(time(NULL));

    FILE *file_c;
    if ((file_c = fopen("compressed.dat", "wb+")) == NULL)
    {
        fprintf(stderr, "Can't open compresed.dat\n");
        exit(EXIT_FAILURE);
    }

    FILE *file_unc;
    if ((file_unc = fopen("uncompressed.dat", "wb+")) == NULL)
    {
        fprintf(stderr, "Can't open uncompressed.dat\n");
        exit(EXIT_FAILURE);
    }

    size_t full_size = write_bin_file(file_unc, file_c);

    size_t size = ftell(file_c);
    printf("\nSize with encode: %ld\nSize with no encode: %d\n\n",
           full_size,
            N * 4);

    int test = read_bin_file(file_unc, file_c, size);
    if (test == -1)
        printf("- Последовательность чисел в файлах неверна!\n");
    else
        printf("+ Последовательность чисел в файлах верна!\n");

    printf("\n");

    fclose(file_c);
    fclose(file_unc);
    return 0;
}
```

В начале функции **main** устанавливается таймер, с помощью функции **srand** и внутри неё функции time, для того, чтобы когда функция выполнялась время было разное и генерировалось разное число. Далее создан указатель на файл, в котором будет храниться указатель на файл, в котором хранятся закодированные цифры. Если файл равен NULL, то выводится сообщение и программа завершается с ошибкой.

2 Задание

```c
int encode(uint32_t code_point, CodeUnits *code_unit)
{
    uint8_t count = 0;
    uint32_t i;

    for (i = code_point; i > 0; i >>= 1)
    {
        count++;
    }

    if (count <= 7)
    {
        code_unit->code[0] = code_point;
        code_unit->length = 1;
        return 0;
    }

    if (count <= 11)
    {
        code_unit->code[0] = 0xc0 | (code_point >> 6);
        code_unit->code[1] = 0x80 | (code_point & 0x3f);
        code_unit->length = 2;
        return 0;
    }

    if (count <= 16)
    {
        code_unit->code[0] = 0xe0 | (code_point >> 12);
        code_unit->code[1] = 0x80 | ((code_point & 0xfc0) >> 6);
        code_unit->code[2] = 0x80 | (code_point & 0x3f);
        code_unit->length = 3;
        return 0;
    }

    if (count <= 21)
    {
        code_unit->code[0] = 0xf0 | (code_point >> 18);
        code_unit->code[1] = 0x80 | ((code_point & 0x3f000) >> 12);
        code_unit->code[2] = 0x80 | ((code_point & 0xfc0) >> 6);
        code_unit->code[3] = 0x80 | (code_point & 0x3f);
        code_unit->length = 4;
        return 0;
    }

    return -1;
}
```

Функция принимает два параметра: число, которое необходимо закодировать, параметр, который будет являться результатом кодирования. С помощью цикла for считывается количество значащих бит кодируемого числа. После этого функция проверяет, сколько байтов потребуется для кодирования символа, и заполняет массив code в code_units соответствующими байтами. Если символ может быть закодирован в один байт, то функция сохраняет этот байт в code_units->code[0] и устанавливает длину code_units->length в 1. Если байт больше чем 1, то с помощью

масок и логических операций кодируются остальные байты, заполняется массив и устанавливается соответствующая длина.

```
uint32_t decode(const CodeUnits *code_unit)
{
    uint32_t code_point = 0;

    if ((code_unit->code[0] >> 7) == 0)
    {
        code_point = code_unit->code[0];
    }
    else if (code_unit->code[0] <= 0xdf)
    {
        code_point = ((code_unit->code[0] & 0x1f) << 6 | (code_unit->code[1] & 0x3f));
    }
    else if (code_unit->code[0] <= 0xef)
    {
        code_point = (((code_unit->code[0] & 0xf) << 12) | ((code_unit->code[1] & 0x3f) << 6) | (code_unit->code[2] & 0x3f));
    }
    else if (code_unit->code[0] <= 0xf7)
    {
        code_point = (((code_unit->code[0] & 0xf) << 18) | ((code_unit->code[1] & 0x3f) << 12) | ((code_unit->code[2] & 0x3f) << 6) | (code_unit->code[3] & 0x3f));
    }

    return code_point;
}
```

Функция принимает в качестве параметра закодированное представление числа. Есть 4 условия:

1. Если старший байт, сдвинутый вправо на 7 битов, равен нулю, то это число соответствует шаблону в первой строке таблицы, где количество значащих бит кодируемого числа равно 7. В данном случае переменной code_point присваивается значение старшего байта.
2. Если старший байт меньше, либо равен 0xdf (223), то закодированное число соответствует шаблону во второй строке таблицы, т.к. 0xdf – это 1101 1111 (т.е. максимальное число, которое может соответствовать второму шаблону). В данном случае, сначала с помощью побитового И и сдвига влево на 6 бит избавление от шаблонных битов в старшем байте, затем используется побитовое И для второго байта с числом 0x3f (63), чтобы снова избавиться от шаблонных битов. После этого используется операция побитовое ИЛИ, чтобы «склеить» байты.
3. Если старший байт меньше, либо равен 0xef (239), то закодированное число соответствует шаблону в третьей строке таблицы, т.к. 0xef – это 1110 1111 (т.е. максимальное число, которое соответствовать третьему шаблону). В данном случае сначала с помощью побитового И и сдвигов вправо избавление от шаблонных битов во всех байтах числа. После этого используется операция побитовое ИЛИ, чтобы «склеить» байты.
4. Если старший байт меньше, либо равен 0xf7 (247), то закодированное число соответствует шаблону в четвёртой строке таблицы, т.к. 0xf7 – это 11110111 (т.е. максимальное число, которое может соответствовать четвёртому шаблону).
   Функция возвращает результат декодирования в виде переменной code_point.

```c
int read_next_code_unit(FILE *in, CodeUnits *code_unit)
{
    code_unit->length = 0;
    uint8_t byte;

    while (code_unit->length == 0)
    {
        fread(&byte, sizeof(uint8_t), 1, in);
        if (feof(in))
        {
            return -1;
        }
        if (byte < 0x80)
        {
            code_unit->code[code_unit->length++] = byte;
            return 0;
        }
        if (byte >= 0xc0)
        {
            int i = 3;
            uint8_t buf;
            for (buf = byte >> 4; buf != 1; buf >>= 1, i--)
            {
                if ((buf == 0x3) || (buf == 0x7) || (buf == 0xf))
                {
                    code_unit->code[0] = byte;
                    int j = 0;
                    for (j; j != i; j++)
                    {
                        fread(&byte, sizeof(uint8_t), 1, in);
                        if ((byte >= 0x80) && (byte <= 0xbf))
                        {
                            code_unit->code[j + 1] = byte;
                        }
                        else if (!feof(in))
                        {
                            fseek(in, -1, SEEK_CUR);
                            return -1;
                        }
                        else
                        {
                            return -1;
                        }
                    }
                    code_unit->length = i + 1;
                    break;
                }
            }
        }
    }
    return 0;
}
```

Функция **read_next_code_unit** принимает несколько параметров: поток файла in и указатель на структуру code_unit. Сначала функция fread читает следующий байт из файла и сохраняет его в переменную byte, если достигнут конец файла, функция возвращает -1. if(byte < 0x80) – проверяет, является ли byte однобайтовой кодовой единицей. Если является, то он записывается в последовательность кодовых единиц «code_units», длина последовательности увеличивается на 1 и функция возвращает значение 0. If(byte => 0xc0) – проверяет, является ли byte началом многобайтовой кодовой единицы. Если да, то начинается цикл. Затем начинается цикл, который продолжается до тех пор, пока buf не станет равным единице. Внутри цикла переменная buf сдвигается на один бит вправо, а переменная i уменьшается на единицу.if((buf == 0x3) || (buf == 0x7) || (buf == 0xf)) – проверяет, является ли ‘byte’ началом двухбайтовой, трёхбайтовой или четырёхбайтовой кодовой единицей. Далее записывается первый байт и с помощью ещё одного цикла начинают записываться остальные байты, в 98 строчке проверяется, достигнут ли конец файла. Если нет, то происходит возврат указателя на предыдущий байт с помощью функции fseek,

и функция завершается с ошибкой. Если нет, то выполняются следующие условия. В завершение в 105 строчке устанавливается длина массива code на i+1 и выход из внутреннего цикла.

```
int write_code_unit(FILE *out, const CodeUnits *code_unit)
{
    fwrite(code_unit, sizeof(uint8_t), code_unit->length, out);
    return 0;
}
```

Функция **write_code_unit** записывает последовательность кодовых единиц из структуры CodeUnits в поток out.

```
int encode_file(const char *int_file_name, const char *out_file_name)
{
    uint32_t code_point;
    CodeUnits code_units;

    FILE *in_file = fopen(out_file_name, "wb");
    if (in_file == NULL)
    {
        printf("Failed to read file located at %s\n", int_file_name);
        return -1;
    }

    FILE *out_file = fopen(out_file_name, "wb");
    if (out_file == NULL)
    {
        printf("Failed to create file located at %s\n", out_file_name);
        return -1;
    }

    while (!feof(in_file))
    {
        fscanf(in_file, "%" SCNx32, &code_point);
        encode(code_point, &code_units);
        write_code_unit(out_file, &code_units);
    }

    fclose(in_file);
    fclose(out_file);
    return 0;
}
```

Функция **encode_file** принимает два параметра: путь к входному текстовому файлу и путь к выходному бинарному файлу. Сначала происходит попытка открыть для чтения входной текстовый файл, в случае неудачи, возвращается -1. Если всё успешно, то происходит попытка создать бинарный файл, в случае неудачи, возвращается -1. Если всё прошло успешно, то далее идёт цикл, который будет работать, пока не достигнет концы файла. В цикле из файла считываются числа и записываются в переменную code_point. Далее вызывается функция encode, чтобы закодировать считанные числа. После этого вызывается функция write_code_unit, чтобы записать в выходной бинарный файл закодированные числа. После завершения цикла, закрывается как текстовый фал, так и бинарный файл. В случае успеха, возвращается 0.

```c
int decode_file(const char *in_file_name, const char *out_file_name)
{
    CodeUnits code_units;
    FILE *in_file = fopen(in_file_name, "rb");
    if (in_file == NULL)
    {
        printf("Failed to read file located at %s\n", in_file_name);
        return -1;
    }

    FILE *out_file = fopen(in_file_name, "w");
    if (out_file == NULL)
    {
        printf("Failed to create file located at %s\n", out_file_name);
        return -1;
    }

    while (!feof(in_file))
    {
        if (!read_next_code_unit(in_file, &code_units))
        {
            fprintf(out_file, "%" PRIx32 "\n", decode(&code_units));
        }
    }

    fclose(in_file);
    fclose(out_file);

    return 0;
}
```

Функция **decode_file** принимает два параметра: путь к входному бинарному и путь к выходному текстовому файлу. Сначала попытка открыть для чтения входной бинарный файл, в случае неудачи, возвращается -1. Если всё успешно, то далее попытка создать или открыть уже имеющийся текстовый файл, в случае неудачи возвращается -1. Если всё прошло успешно, то далее идёт цикл, который будет работать, пока не достигнет конца файла. В цикле наблюдается условие, если возвращаемое значение функции read_next_code_unit равно нулю, то записывается в текстовый файл декодированное число. После завершения цикла, бинарный и текстовый файл закрываются. В случае успеха возвращается 0.

```c
#include "coder.h"
#include "command.h"
#include <string.h>

int main(int argc, char *argv[])
{
    if (argc == 4)
    {
        if (strcmp(argv[1], "encode") == 0)
        {
            encode_file(argv[2], argv[3]);
        }
        else if (strcmp(argv[1], "decode") == 0)
        {
            decode_file(argv[2], argv[3]);
        }
        else
        {
            printf("Usage:\n");
            printf("coder encode <in-file-name> <out-file-name>\n");
            printf("coder decode <in-file-name> <out-file-name>\n");
            return -1;
        }
    }
    else
    {
        printf("Usage:\n");
        printf("coder encode <in-file-name> <out-file-name>\n");
        printf("coder decode <in-file-name> <out-file-name>\n");
        return -1;
    }

    return 0;
}
```

В функции main проверяется количество введённых аргументов командной строки, затем с помощью функции strcmp проверяется, какое действие хочется совершить: закодировать или декодировать число. В случае неправильного ввода аргументов, main вернёт -1. В случае успеха 0.

ПРИЛОЖЕНИЕ

1 Задание

main.c

```
 1  #include <assert.h>
 2  #include <stddef.h>
 3  #include <stdint.h>
 4  #include <stdio.h>
 5  #include <stdlib.h>
 6  #include <string.h>
 7  #include <time.h>
 8
 9  #define N 1000000
10
11  uint32_t generate_number()
12  {
13      const int r = rand();
14      const int p = r % 100;
15      if (p < 90)
16      {
17          return r % 128;
18      }
19      if (p < 95)
20      {
21          return r % 16384;
22      }
23      if (p < 99)
24      {
25          return r % 2097152;
26      }
27      return r % 268435455;
28  }
29
30  size_t encode_varint(uint32_t value, uint8_t *buf)
31  {
32      assert(buf != NULL);
33      uint8_t *cur = buf;
34      while (value >= 0x80)
35      {
36          const uint8_t byte = (value & 0x7f) | 0x80;
37          *cur = byte;
38          value >>= 7;
39          ++cur;
40      }
41      *cur = value;
42      ++cur;
43      return cur - buf;
44  }
```

```
45
46  uint32_t decode_varint(const uint8_t **bufp)
47  {
48      const uint8_t *cur = *bufp;
49      uint8_t byte = *cur++;
50      uint32_t value = byte & 0x7f;
51      size_t shift = 7;
52      while (byte >= 0x80)
53      {
54          byte = *cur++;
55          value += (byte & 0x7f) << shift;
56          shift += 7;
57      }
58      *bufp = cur;
59      return value;
60  }
61
62  // записывать в файл
63  size_t write_bin_file(FILE *file_unc, FILE *file_c)
64  {
65      assert(file_unc != NULL);
66      assert(file_c != NULL);
67
68      uint8_t buf[4] = {};
69      size_t size = 0;
70      size_t full_size = 0;
71
72      for (int i = 0; i < N; i++)
73      {
74          uint32_t numb = generate_number();
75          fwrite(&numb, sizeof(numb), 1, file_unc);
76          size = encode_varint(numb, buf);
77          full_size += size;
78          fwrite(buf, size, 1, file_c);
79      }
80
81      return full_size;
82  }
83
84  // считывать из файла
85  int read_bin_file(FILE *file_unc, FILE *file_c, size_t size)
86  {
87      assert(file_unc != NULL);
88      assert(file_c != NULL);
89
90      fseek(file_c, 0, SEEK_SET);
91      fseek(file_unc, 0, SEEK_SET);
92
93      uint32_t num_uncomp, num_comp;
94      uint8_t *buf = malloc(sizeof(uint8_t) * size);
```

```
 95      const uint8_t *cur = buf;
 96      fread(buf, sizeof(uint8_t), size, file_c);
 97
 98      for (int i = 0; i < N; i++)
 99      {
100          num_comp = decode_varint(&cur);
101          fread(&num_uncomp, sizeof(uint32_t), 1, file_unc);
102          if (i < 5)
103              printf("encode: %d decode: %d\n", num_uncomp, num_comp);
104          if (num_comp != num_uncomp)
105              return -1;
106      }
107      printf("\n");
108
109      free(buf);
110      return 0;
111  }
112
113  int main()
114  {
115      srand(time(NULL));
116
117      FILE *file_c;
118      if ((file_c = fopen("compressed.dat", "wb+")) == NULL)
119      {
120          fprintf(stderr, "Can't open compresed.dat\n");
121          exit(EXIT_FAILURE);
122      }
123
124      FILE *file_unc;
125      if ((file_unc = fopen("uncompressed.dat", "wb+")) == NULL)
126      {
127          fprintf(stderr, "Can't open uncompressed.dat\n");
128          exit(EXIT_FAILURE);
129      }
130
131      size_t full_size = write_bin_file(file_unc, file_c);
132
133      size_t size = ftell(file_c);
134      printf("\nSize with encode: %ld\nSize with no encode: %d\n\n",
135             full_size,
136              N * 4);
137
138      int test = read_bin_file(file_unc, file_c, size);
139      if (test == -1)
140          printf("- Последовательность чисел в файлах неверна!\n");
141      else
142          printf("+ Последовательность чисел в файлах верна!\n");
143
144      printf("\n");
```

```
145
146     fclose(file_c);
147     fclose(file_unc);
148     return 0;
149 }
```

2 Задание

coder.c

```
 1 #include <stdio.h>
 2 #include "coder.h"
 3
 4 uint32_t decode(const CodeUnits *code_unit)
 5 {
 6     uint32_t code_point = 0;
 7
 8     if ((code_unit->code[0] >> 7) == 0)
 9     {
10         code_point = code_unit->code[0];
11     }
12     else if (code_unit->code[0] <= 0xdf)
13     {
14         code_point = ((code_unit->code[0] & 0x1f) << 6 | (code_unit-
15 >code[1] & 0x3f));
16     }
17     else if (code_unit->code[0] <= 0xef)
18     {
19         code_point = (((code_unit->code[0] & 0xf) << 12) | ((code_unit-
20 >code[1] & 0x3f) << 6) | (code_unit->code[2] & 0x3f));
21     }
22     else if (code_unit->code[0] <= 0xf7)
23     {
24         code_point = (((code_unit->code[0] & 0xf) << 18) | ((code_unit-
25 >code[1] & 0x3f) << 12) | ((code_unit->code[2] & 0x3f) << 6) |
26 (code_unit->code[3] & 0x3f));
27     }
28
29     return code_point;
30 }
31
32 int encode(uint32_t code_point, CodeUnits *code_unit)
33 {
34     uint8_t count = 0;
35     uint32_t i;
36
37     for (i = code_point; i > 0; i >>= 1)
```

```
38      {
39          count++;
40      }
41
42      if (count <= 7)
43      {
44          code_unit->code[0] = code_point;
45          code_unit->length = 1;
46          return 0;
47      }
48
49      if (count <= 11)
50      {
51          code_unit->code[0] = 0xc0 | (code_point >> 6);
52          code_unit->code[1] = 0x80 | (code_point & 0x3f);
53          code_unit->length = 2;
54          return 0;
55      }
56
57      if (count <= 16)
58      {
59          code_unit->code[0] = 0xe0 | (code_point >> 12);
60          code_unit->code[1] = 0x80 | ((code_point & 0xfc0) >> 6);
61          code_unit->code[2] = 0x80 | (code_point & 0x3f);
62          code_unit->length = 3;
63          return 0;
64      }
65
66      if (count <= 21)
67      {
68          code_unit->code[0] = 0xf0 | (code_point >> 18);
69          code_unit->code[1] = 0x80 | ((code_point & 0x3f000) >> 12);
70          code_unit->code[2] = 0x80 | ((code_point & 0xfc0) >> 6);
71          code_unit->code[3] = 0x80 | (code_point & 0x3f);
72          code_unit->length = 4;
73          return 0;
74      }
75
76      return -1;
77  }
78
79  int write_code_unit(FILE *out, const CodeUnits *code_unit)
80  {
81      fwrite(code_unit, sizeof(uint8_t), code_unit->length, out);
82      return 0;
83  }
84
85  int read_next_code_unit(FILE *in, CodeUnits *code_unit)
86  {
87      code_unit->length = 0;
```

```
 88      uint8_t byte;
 89
 90      while (code_unit->length == 0)
 91      {
 92          fread(&byte, sizeof(uint8_t), 1, in);
 93          if (feof(in))
 94          {
 95              return -1;
 96          }
 97          if (byte < 0x80)
 98          {
 99              code_unit->code[code_unit->length++] = byte;
100              return 0;
101          }
102          if (byte >= 0xc0)
103          {
104              int i = 3;
105              uint8_t buf;
106              for (buf = byte >> 4; buf != 1; buf >>= 1, i--)
107              {
108                  if ((buf == 0x3) || (buf == 0x7) || (buf == 0xf))
109                  {
110                      code_unit->code[0] = byte;
111                      int j = 0;
112                      for (j; j != i; j++)
113                      {
114                          fread(&byte, sizeof(uint8_t), 1, in);
115                          if ((byte >= 0x80) && (byte <= 0xbf))
116                          {
117                              code_unit->code[j + 1] = byte;
118                          }
119                          else if (!feof(in))
120                          {
121                              fseek(in, -1, SEEK_CUR);
122                              return -1;
123                          }
124                          else
125                          {
126                              return -1;
127                          }
128                      }
129                      code_unit->length = i + 1;
130                      break;
131                  }
132              }
          }
      }
      return 0;
  }
```


coder.h

```
#pragma once
#include <stdio.h>
#include <stdint.h>

enum
{
    MaxCodeLength = 4
};

typedef struct
{
    uint8_t code[MaxCodeLength];
    size_t length;
} CodeUnits;

int encode(uint32_t code_point, CodeUnits *code_units);
uint32_t decode(const CodeUnits *code_unit);
int read_next_code_unit(FILE *in, CodeUnits *code_units);
int write_code_unit(FILE *out, const CodeUnits *code_unit);
```

command.c

```
#include "coder.h"
#include "command.h"
#include <inttypes.h>
#include <stdio.h>

int encode_file(const char *int_file_name, const char *out_file_name)
{
    uint32_t code_point;
    CodeUnits code_units;

    FILE *in_file = fopen(out_file_name, "wb");
    if (in_file == NULL)
    {
        printf("Failed to read file located at %s\n", int_file_name);
        return -1;
    }

    FILE *out_file = fopen(out_file_name, "wb");
    if (out_file == NULL)
    {
```

```
21          printf("Failed to create file located at %s\n", out_file_name);
22          return -1;
23      }
24
25      while (!feof(in_file))
26      {
27          fscanf(in_file, "%" SCNx32, &code_point);
28          encode(code_point, &code_units);
29          write_code_unit(out_file, &code_units);
30      }
31
32      fclose(in_file);
33      fclose(out_file);
34      return 0;
35  }
36
37  int decode_file(const char *in_file_name, const char *out_file_name)
38  {
39      CodeUnits code_units;
40      FILE *in_file = fopen(in_file_name, "rb");
41      if (in_file == NULL)
42      {
43          printf("Failed to read file located at %s\n", in_file_name);
44          return -1;
45      }
46
47      FILE *out_file = fopen(in_file_name, "w");
48      if (out_file == NULL)
49      {
50          printf("Failed to create file located at %s\n", out_file_name);
51          return -1;
52      }
53
54      while (!feof(in_file))
55      {
56          if (!read_next_code_unit(in_file, &code_units))
57          {
58              fprintf(out_file, "%" PRIx32 "\n", decode(&code_units));
59          }
60      }
61
62      fclose(in_file);
63      fclose(out_file);
64
65      return 0;
66  }
```

command.h

```
#pragma once

int encode_file(const char *int_file_name, const char *out_file_name);
int decode_file(const char *in_file_name, const char *out_file_name);
```

main.c

```
#include "coder.h"
#include "command.h"
#include <string.h>

int main(int argc, char *argv[])
{
    if (argc == 4)
    {
        if (strcmp(argv[1], "encode") == 0)
        {
            encode_file(argv[2], argv[3]);
        }
        else if (strcmp(argv[1], "decode") == 0)
        {
            decode_file(argv[2], argv[3]);
        }
        else
        {
            return -1;
        }
    }
    else
    {
        return -1;
    }

    return 0;
}
```