|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | МИНОБРНАУКИ РОССИИ | | | | Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | | |   Институт Информационных технологий | |
|  | |
| Кафедра Математического обеспечения и стандартизации информационных технологий | |
|  | |
|  | |

|  |  |
| --- | --- |
| **ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ № 7** | |
| **по дисциплине** | |
| «Структуры и алгоритмы обработки данных**»**  **Тема: «**Кодирование и сжатие данных методами без потерь**»** | |
|  | |
| Выполнил студент группы ИКБО-10-21 | Смольников А.Б. |
| Принял преподаватель | Филатов А.С. |

|  |  |  |
| --- | --- | --- |
| Лабораторная работа выполнена | «\_\_»\_\_\_\_\_\_\_202\_\_ г. | *(подпись студента)* |
|  |  |  |
| «Зачтено» | «\_\_»\_\_\_\_\_\_\_202\_\_ г. | *(подпись руководителя)* |

Москва 2022

# **Цель работы**

Получение практических навыков и знаний по выполнению сжатия данных рассматриваемыми методами.

# **Постановка задачи**

Вариант 22. Вид дерева: АВЛ

1. Составить программу сжатия текста, используя метод RLE (run length encoding/кодирование длин серий/групповое кодирование). Рассчитать коэффициент сжатия в тестах.
   1. Провести тестирование на длинной серии повторяющихся символов.
   2. Провести тестирование на длинной серии неповторяющихся символов.
2. Составить программу сжатия текста алгоритмом Лемпеля-Зива (LZ77), LZ78. Рассчитать коэффициент сжатия в тестах.
   1. Провести тестирование реализации LZ77 в соответствии с индивидуальным вариантом задания, используя двухсимвольный алфавит (0, 1). Описать процесс восстановления сжатого текста.
   2. Провести тестирование реализации LZ78 в соответствии с индивидуальным вариантом задания. Описать процесс восстановления сжатого текста.
3. Составить программу сжатия и восстановления текста алгоритмами Шеннона-Фано и Хаффмана.
   1. Провести тестирование метода Шеннона-Фано в соответствии с индивидуальным вариантом задания. Представить в отчете таблицу формирования кода, изобразить префиксное дерево, рассчитать коэффициент сжатия.
   2. Провести тестирование метода Хаффмана на строке, содержащей ваше ФИО. Построить таблицу частот встречаемости символов, сформировать алфавит исходной строки и посчитать частоту вхождений и вероятность появления символов. Изобразить префиксное дерево Хаффмана. Показать кодирование каждого символа в строке. Рассчитать коэффициент сжатия, среднюю длину кода и дисперсию.
   3. Применить алгоритм к большому текстовому файлу. Архивировать тот же файл любым архиватором. Выполнить сравнительный анализ сжатия этими способами.
4. Составить отчет, отобразив в нем описание выполнения всех этапов разработки, тестирования и код всей программы со скриншотами результатов тестирования.

Таблица 1. Задания варианта №22

|  |  |  |  |
| --- | --- | --- | --- |
| Вариант | Сжатие данных по методу Лемпеля-Зива LZ77 | Закодировать следующую фразу, используя код LZ78 | Закодировать фразу методами Шеннона-Фано |
| 2 | 0100100010010000101 | упупапекапекаупуп  (upupapekapekaupup) | One, two, Freddy's coming for you  Three, four, better lock  your door Five, six, grab a crucifix  Seven, eight, gonna stay up late. |

# **Решение**

* 1. **Теоретическое введение**

Кодирование длин серий (RLE) — алгоритм сжатия данных, заменяющий повторяющиеся символы (серии) на один символ и число его повторов. Серией называется последовательность, состоящая из нескольких одинаковых символов. При кодировании строка одинаковых символов, составляющих серию, заменяется строкой, содержащей сам повторяющийся символ и количество его повторов.

Алгоритмы LZ77 обеспечивают сжатие путем замены повторяющихся вхождений данных ссылками на единственную копию этих данных, существовавшую ранее в несжатом потоке данных. Совпадение кодируется парой чисел, называемых парой длина-расстояние, что эквивалентно утверждению "каждый из следующих символов длины равен символам, расположенным точно на расстоянии символов за ним в несжатом потоке".

Чтобы определить совпадения, кодировщик должен отслеживать некоторый объем самых последних данных, таких как последние 2 КБ, 4 КБ или 32 КБ. Структура, в которой хранятся эти данные, называется скользящим окном, поэтому LZ77 иногда называют сжатием со скользящим окном. Кодеру необходимо сохранить эти данные для поиска совпадений, а декодеру необходимо сохранить эти данные для интерпретации совпадений, на которые ссылается кодер. Чем больше скользящее окно, тем дольше кодировщик может выполнять поиск ссылок на создание.

Алгоритмы LZ78 сжимают последовательные данные путем построения словаря последовательностей токенов на основе входных данных, а затем заменяют второе и последующее вхождение последовательности в потоке данных ссылкой на словарную запись. Наблюдение состоит в том, что количество повторяющихся последовательностей является хорошей мерой неслучайного характера последовательности. Алгоритмы представляют словарь в виде n-арного дерева, где n - количество токенов, используемых для формирования последовательностей токенов. Каждая словарная запись имеет вид dictionary[...] = {index, token}, где index - это индекс словарной записи, представляющей ранее просмотренную последовательность, а token - это следующий токен из входных данных, который делает эту запись уникальной в словаре.

В области сжатия данных кодирование Шеннона–Фано, названное в честь Клода Шеннона и Роберта Фано - это название, данное двум различным, но родственным методам построения префиксного кода на основе набора символов и их вероятностей (оцененных или измеренных).

Метод Шеннона выбирает префиксный код, в котором исходному символу присваивается длина кодового. Один из распространенных способов выбора кодовых слов использует двоичное разложение кумулятивных вероятностей. Этот метод был предложен в "Математической теории коммуникации" Шеннона (1948), его статье, вводящей в область теории информации.

Метод Фано делит исходные символы на два набора ("0" и "1") с вероятностями, максимально близкими к 1/2. Затем эти наборы сами делятся на два, и так далее, пока каждый набор не будет содержать только один символ. Кодовым словом для этого символа является строка из "0" и "1", которая записывает, на какую половину делений он пришелся. Этот метод был предложен в более позднем техническом отчете Фано (1949).

Коды Шеннона–Фано неоптимальны в том смысле, что они не всегда достигают минимально возможной ожидаемой длины кодового слова, как это делает кодирование Хаффмана. Однако коды Шеннона–Фано имеют ожидаемую длину кодового слова в пределах 1 бита от оптимальной. Метод Фано обычно производит кодирование с более короткими ожидаемыми длинами, чем метод Шеннона. Однако метод Шеннона легче поддается теоретическому анализу.

Код Хаффмана — это особый тип оптимального префиксного кода, который обычно используется для сжатия данных без потерь. Процесс поиска или использования такого кода осуществляется с помощью кодирования Хаффмана.

Выходные данные алгоритма Хаффмана можно рассматривать как кодовую таблицу переменной длины для кодирования исходного символа (например, символа в файле). Алгоритм выводит эту таблицу из расчетной вероятности или частоты появления (веса) для каждого возможного значения исходного символа. Как и в других методах энтропийного кодирования, более распространенные символы обычно представляются с использованием меньшего количества битов, чем менее распространенные символы. Метод Хаффмана может быть эффективно реализован, находя код за время, линейное количеству входных весов, если эти веса отсортированы.

* 1. **Функции задания №1**

Для кодирования методом RLE используется функция encodeRLE, принимающая на вход строку и итеративно проходя по каждому символу определяет серии и формирует закодированную строку.

|  |
| --- |
| string encodeRLE(string data){  string result;  const unsigned long long len = data.length();  if(len==0) return "";  unsigned long long cnt=1;  for(unsigned long long i=0;i<len;i++){  if(i==len-1 || data[i]!=data[i+1]){  result+=to\_string(cnt);  result+=data[i];  cnt=1;  } else cnt++;  }  return result;  } |
|  |

* 1. **Функции задания №2**

Для реализации алгоритма сжатия LZ77 используем класс с идентичным названием, а также структуру единичной записи. В классе созданы все необходимые методы и поля. Метод fillNode является «конструктором» записи, исходный код опущен в связи с простотой.

|  |
| --- |
| struct Node{  unsigned char offset;  unsigned char length;  char next;  };  class LZ77{  int windowLength=8;  Node fillNode(int off,int l, char nex);  public:  vector<Node> encoded;  vector<Node> encode(string text);  string decode(vector<Node> encoded);  int getEncodedLength();  void printEncoded();  }; |

Метод encode() принимает на вход сжимаемую строку, и итеративно рассматривая символы, заполняет вектор с результатом. В ходе сжатия выполняется поиск повторяющихся вхождений в окне и запись их в соответствующую структуру данных.

|  |
| --- |
| vector<Node> LZ77::encode(std::string text) {  vector<Node> result;  windowLength=text.length();  result.push\_back(fillNode(0,0,text[0]));  for(int i=1;i<text.length();i++) {  string window;  if(true){  window = text.substr(0,i);  }else{  window= text.substr(i-windowLength,windowLength);  }  //finding match in buffer  int matchLength=0;  int matchOffset=0;  for(int j=0;j<window.length();j++){  int k=0;  int pos=j+k;  int startPos=j+k;  while(window[pos]==text[i+k]){  k++;  pos++;  if(pos>=window.length()){  pos-=(int)window.length();  }  }  if(k>matchLength){  matchLength=k;  matchOffset=i-startPos;  }  }  if(matchLength==0){  result.push\_back(fillNode(0,0,text[i]));  }else{  result.push\_back(fillNode(matchOffset,matchLength,text[i+matchLength]));  i+=matchLength;  }  }  encoded=result;  return result;  } |

Метод decode() принимает на вход вектор структур записей и итеративно декодирует строку, используя сдвиги в соответствии с записями.

|  |
| --- |
| string LZ77::decode(vector<Node> encoded){  string result;  for(int i=0;i<encoded.size();i++){  if(encoded[i].offset==0){  result+=encoded[i].next;  }else{  int pos=result.length()-encoded[i].offset;  int resultLen=result.length();  for(int j=0;j<encoded[i].length;j++){  result+=result[pos];  pos++;  if(pos>=resultLen){  pos-=resultLen;  }  }  result+=encoded[i].next;  }  // cout<<result<<endl;  }  return result;  } |

Методы printEncoded() и getEncodedLength() являются служебными, первый выводит архивированную строку в консоль, второй считает объем памяти, занимаемый закодированной информацией.

|  |
| --- |
| void LZ77::printEncoded() {  for(int i=0;i<encoded.size();i++){  if(encoded[i].next=='\0'){  cout<<"("<<(int)encoded[i].offset<<","<<(int)encoded[i].length<<", NTERM)"<<endl;  }else{  cout<<"("<<(int)encoded[i].offset<<","<<(int)encoded[i].length<<", "<<encoded[i].next<<")"<<endl;  }  }  }  int LZ77::getEncodedLength() {  return encoded.size()\*(1+1+1);  } |

Для реализации LZ78 аналогично существует соответствующий класс, а также структура записи. В классе реализованы необходимые методы и поля.

|  |
| --- |
| struct LZ78node{  unsigned char index;  char next;  };  class LZ78{  vector<LZ78node> encoded;  map<string,int> dictionary;  public:  int getEncodedLength();  void printEncoded();  vector<LZ78node> encode(string text);  string decode();  }; |

Метод encode() принимает на вход исходную строку и возвращает вектор записей, закодированную строку. В процессе итеративно рассматривает текст и выполняет поиск по словарю максимального совпадения, также заполняет сам словарь, рассматривая случай, когда после рассмотрения строки в буфере остались символы.

|  |
| --- |
| vector<LZ78node> LZ78::encode(string text) {  int index=0;  string buffer="";  vector<LZ78node> answer;  for(int i=0;i<text.length();i++){  if(dictionary.find(buffer+text[i]) != dictionary.end()){  buffer+=text[i];  }else{  answer.push\_back(LZ78node{static\_cast<unsigned char>(dictionary[buffer]), text[i]});  dictionary[buffer+text[i]]=dictionary.size()+1;  buffer="";  }  }  if(!buffer.empty()){  char last\_ch = buffer[buffer.length()-1];  buffer.pop\_back();  answer.push\_back(LZ78node{static\_cast<unsigned char>(dictionary[buffer]),last\_ch});  }  encoded=answer;  return answer;  } |

Методы getEncodedLength() и printEncoded() являются служебными и отвечают за возврат занимаемой памяти и читаемый вывод закодированной последовательности соответсвенно.

|  |
| --- |
| int LZ78::getEncodedLength() {  return 2\*encoded.size();  }  void LZ78::printEncoded() {  for(int i=0;i<encoded.size();i++){  cout<<(int)encoded[i].index<<" "<<encoded[i].next<<endl;  }  } |

* 1. **Функции задания №3**

Для архивации текста методом Шеннона-Фано используется специальный класс ShennonFano, который включает в себя поля:

* Целочисленный тип - длина строки;
* Вектор логических значений – ответ;
* Ассоциативный массив char, int – счетчик;
* Вектор кортежей char, double – вероятность;
* Ассоциативный массив кодов символов;
* Дерево закодированных символов.

В реализации задачи будем считать, что логическое значение в памяти занимает 1 бит, а не 1 байт. Такой подход связан с невозможностью динамического использования bitset в STL (возможно только в библиотеке boost). Для простоты и читаемости используем логические значения, а не контейнер char.

|  |
| --- |
| class ShennonFano{  long long s\_length=0;  vector<bool> answer;  map<char,int>count;  vector<tuple<char,double>>dataProbability;  map<char,string>codeTable;  int tempCnt=0;  struct treeNode{  char c;  string code;  treeNode \*left;  treeNode \*right;  };  treeNode \*rootTree = new treeNode;  public:  void scanText(string text);  vector<bool> encode(string text);  void buildCodeTree(vector<tuple<char,double>>data, string curValue, treeNode \*root);  string decode();  int getByteSize();  ~ShennonFano();  }; |

Метод scanText() принимает на вход изначальную строку и выполняет первоначальное сканирование строки, заполняет ассоциативный массив количеством встречаемых символов, инициализирует поле s\_length, а также заполняет вектор вероятностей и сортирует его по второму полю корежа.

|  |
| --- |
| void ShennonFano::scanText(string text) {  for(auto c: text){  count.find(c) == count.end() ? count[c]=1 : count[c]++;  }  s\_length=text.length();  for(auto it : count){  dataProbability.emplace\_back(it.first,(double)it.second/ (double)s\_length);  }  sort(dataProbability.begin(),dataProbability.end(),[](tuple<char,double> const &a, tuple<char,double> const &b){  return get<1>(a) >= get<1>(b);  });  } |
|  |
|  |

Метод buildCodeTree() принимает на вход контейнер вероятностей, текущий код, а также родительский элемент и рекурсивно строит дерево получившихся кодов. В процессе выполнения вектор разделяется на две части так, чтобы различие в сумме вероятностей было минимально. Далее создаются два дополнительных вектора, хранящие в себе левую и правую часть, метод вызывается снова с обновленными параметрами. Условием выхода из рекурсии является единичная длина вектора, тогда создается лист и выполнение заканчивается.

|  |
| --- |
|  |
| VoidShennonFano::buildCodeTree(vector<tuple<char,double>>data,string curValue, treeNode \*root) {  if(data.size()==1){  root->c = get<0>(data[0]);  root->code = curValue;  codeTable[get<0>(data[0])] = curValue;  tempCnt++;  return;  }else if(data.size()==0){  return;  }  int minPos=1;  double minDelta=1.0;  for(int i=1;i<data.size();i++){  double accumulateLeft=0.0;  double accumulateRight=0.0;  for(int j=0;j<data.size();j++){  if(j<i){  accumulateLeft+=get<1>(data[j]);  }else{  accumulateRight+=get<1>(data[j]);  }  }  double delta = abs(accumulateLeft-accumulateRight);  if(delta<minDelta){  minDelta=delta;  minPos=i;  }  }  vector<tuple<char,double>>leftV;  vector<tuple<char,double>>rightV;  leftV = vector<tuple<char,double>>(data.begin(),data.begin()+minPos);  rightV = vector<tuple<char,double>>(data.begin()+minPos,data.end());  treeNode \*leftNode = new treeNode;  treeNode \*rightNode = new treeNode;  root->left = leftNode;  root->right = rightNode;  buildCodeTree(leftV, curValue+"0",root->left);  buildCodeTree(rightV, curValue+"1",root->right);  } |

Метод encode() принимает на вход исходный текст, вызывает два ранее описанных подготовительных метода. В процессе выполнения по построенному дереву кодов при помощи ассоциативного массива итеративно кодирует строку.

|  |
| --- |
|  |
| vector<bool> ShennonFano::encode(string text) {  scanText(text);  buildCodeTree(dataProbability,"",rootTree);  // for(auto c: codeTable){  // cout << c.first << " " << c.second << endl;  // }  for(auto c: text){  for(auto code: codeTable[c]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  return {};  } |

Метод decode() выполняет декодирование строки по данным из памяти. В процессе выполнения итеративно рассматривает двоичный код и восстанавливает информацию по словарю кодов, возвращает строку.

|  |
| --- |
| string ShennonFano::decode() {  string result="";  string temp\_code;  while(!answer.empty()){  temp\_code+=answer.front() ? "1" : "0";  answer.erase(answer.begin());  for(auto it: codeTable){  if(it.second==temp\_code){  result+=it.first;  temp\_code.clear();  break;  }  }  }  return result;  } |
|  |

Метод getByteSize() рассчитывает размер закодированной строки и возвращает целое число.

|  |
| --- |
| int ShennonFano::getByteSize() {  if(answer.size()%8==0){  return answer.size()/8;  }else{  return answer.size()/8+1;  }  } |
|  |

* 1. **Функции задания №4**

Для реализации сжатия по Хаффману реализуем соответствующий класс, который включает в себя поля:

* Структура записи, состоящая из строки и веса;
* Структура узла дерева, хранящая запись и указатели;
* Вектор записей data;
* Вектор логических значений (результат);
* Ассоциативный массив кодов символов;
* Ассоциативный массив для декодирования;
* Указатель на корень дерева.

|  |
| --- |
| class Huffman{  struct Node{  string s="";  int weight=0;  };  struct treeNode{  Node data;  treeNode \*left;  treeNode \*right;  treeNode \*parent;  };  vector<Node> data;  vector<bool> answer;  map<string,string>codeTable;  map<string,string>decodeTable;  treeNode \*rootTree;  public:  void scanText(string dataText);  int find\_el(char c);  void createCodes(treeNode \*pNode, string code);  void printEncoded();  int getEncodedLength();  string decode();  vector<bool> encode(string dataText, string fileName="", bool isFile=false);  long long s\_length=0;  }; |
|  |

Метод scanText() принимает на вход исходную строку и выполняет первоначальное сканирование текста, заполняя поле длины текста и вектор записей символом и количеством его вхождений в текст.

|  |
| --- |
| void Huffman::scanText(string dataText) {  for(char c:dataText){  s\_length++;  int index = find\_el(c);  if(index!=-1){  this->data[index].weight++;  }else{  this->data.push\_back(Node{string(1,c),1});  }  }  } |
|  |

Метод encode() принимает на вход строку – исходный текст, имя файла и логическую переменную isFille. Если пользователь хочет обработать строку, вызывается метод сканирования и вектор записей сортируется, иначе сканируются все строки файла по имени. Далее заполняется вектор узлов дерева (для удобного пользования). Алгоритм построения использует приоритетную очередь, в которой узлу с наименьшей вероятностью присваивается наивысший приоритет. Пока в очереди находится более одного узла, удалить два узла с наивысшим приоритетом (наименьшей вероятностью) из очереди, создать новый внутренний узел с этими двумя узлами в качестве дочерних и с вероятностью, равной сумме вероятностей двух узлов. Добавить новый узел в очередь. Оставшийся узел является корневым, дерево завершено. Далее создаются и выводятся коды для каждого узла. В самом конце происходит непосредственно кодировка текста по полученному ассоциативному массиву. Если кодируется файл, все строки читаются итеративно.

|  |
| --- |
| vector<bool> Huffman::encode(string dataText, string fileName, bool isFile) {  if(!isFile) {  scanText(dataText);  sort(this->data.begin(), this->data.end(), [&](Node a, Node b) { return a.weight > b.weight; });  }else{  ifstream file(fileName);  string str;  while (getline(file, str)) {  scanText(str+"\n");  }  sort(this->data.begin(), this->data.end(), [&](Node a, Node b) { return a.weight > b.weight; });  }  vector<treeNode\*> nodes;  for(int i=0;i<this->data.size();i++){  nodes.push\_back(new treeNode{this->data[i], nullptr, nullptr, nullptr});  }  while(nodes.size()>1){  treeNode \*left = nodes[nodes.size()-1];  treeNode \*right = nodes[nodes.size()-2];  treeNode \*parent = new treeNode{Node{"",left->data.weight+right->data.weight},left,right,nullptr};  left->parent=parent;  right->parent=parent;  nodes.pop\_back();  nodes.pop\_back();  nodes.push\_back(parent);  sort(nodes.begin(), nodes.end(), [&](treeNode \*a, treeNode \*b) { return a->data.weight > b->data.weight; });  }  this->rootTree = nodes[0];  //create codes for all tree  createCodes(rootTree,"");  //cout map  for(auto i:codeTable){  cout<<i.first<<" "<<i.second<<endl;  }  //encode data  if(!isFile){  for(char c : dataText){  for(char code : codeTable[string(1,c)]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  }else{  ifstream file(fileName);  string str;  while (getline(file, str)) {  for(char c : str){  for(char code : codeTable[string(1,c)]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  for(char code : codeTable[string(1,'\n')]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  }  return answer;  } |
|  |

Метод createCodes() принимает на вход узел и код, рекурсивно обходит дерево в глубину и инициализирует коды символов в соответствии с частотой их встречаемости.

|  |
| --- |
| void Huffman::createCodes(Huffman::treeNode \*pNode, string code) {  if(pNode->left== nullptr && pNode->right== nullptr){  codeTable[pNode->data.s]=code;  decodeTable[code]=pNode->data.s;  return;  }  createCodes(pNode->left,code+"0");  createCodes(pNode->right,code+"1");  } |
|  |

Метод decode() выполняет расшифровку строки по известному набору префиксных кодов алфавита, итеративно рассматривая двоичный код и декодируя подстроки.

|  |
| --- |
| string Huffman::decode(){  string decoded;  string code;  for(bool i:answer){  if(i) code+="1";  else code+="0";  if(decodeTable.find(code)!=decodeTable.end()){  decoded+=decodeTable[code];  code="";  }  }  return decoded;  } |
|  |

Методы printEncoded() и getEncodedLength() являются вспомогательными, выводят закодированную строку и считают занимаемую память соответственно.

|  |
| --- |
| void Huffman::printEncoded(){  for(bool i: answer){  cout<<i;  }  cout<<endl;  }  int Huffman::getEncodedLength() {  return answer.size()/8;  } |
|  |

* 1. **Интерфейс**

При запуске программы пользователю предоставляется текстовое меню

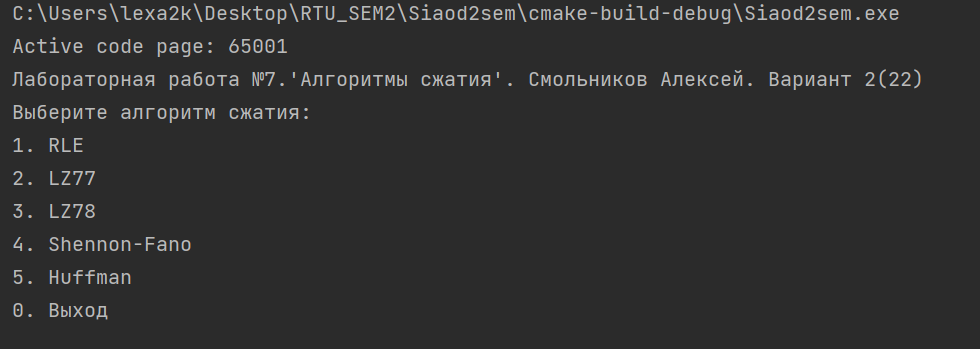


Рисунок 1. Интерфейс программы

# **Тестирование**

Тестирование первого метода проведем на двух строках, повторяющейся и неповторяющейся.

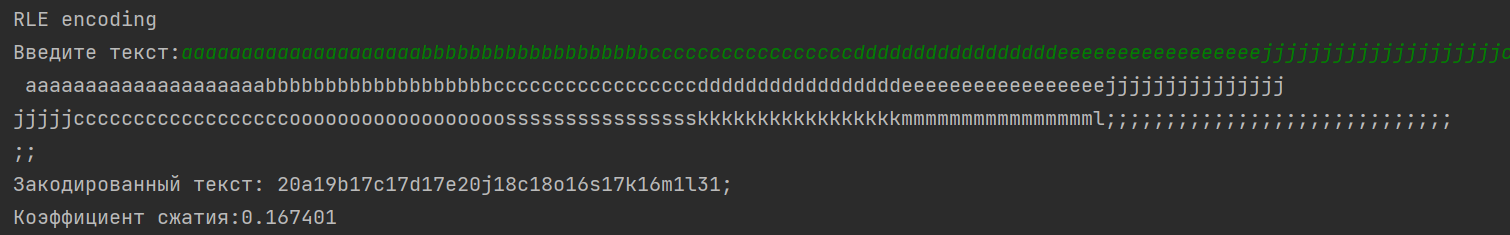


Рисунок 2. Повторяющаяся последовательность

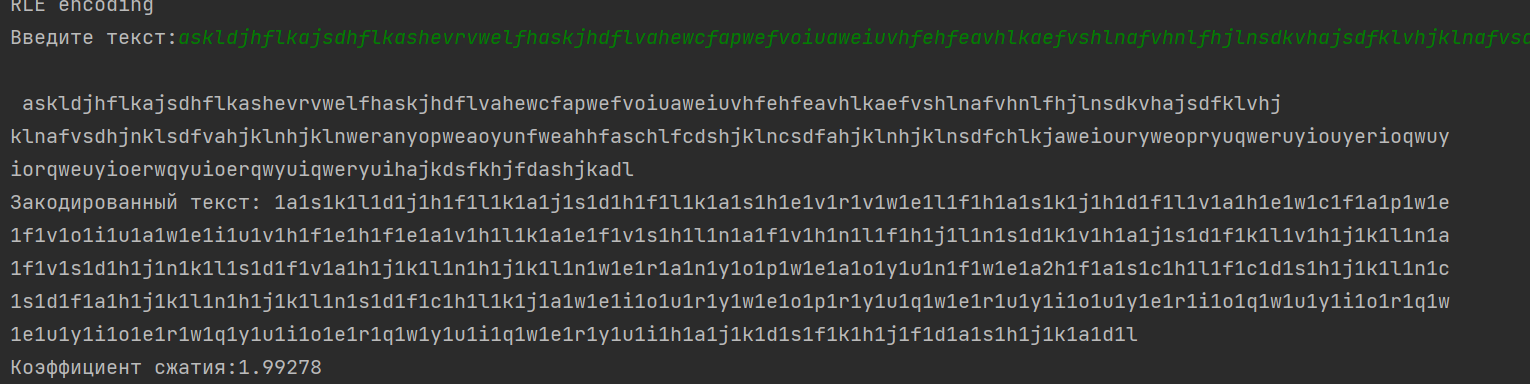


Рисунок 3. Неповторяющаяся последовательность

Использование RLE, очевидно, эффективно только при повторяющихся серий символов, при этом если каждые 2 рядом стоящих символа различны, то коэффициент – 2, если длина последовательности из одного символа стремится к бесконечности, то коэффициент сжатия стремится к нулю.

Тестирование алгоритма сжатия LZ77 проведем на строке из персонального варианта: «0100100010010000101».

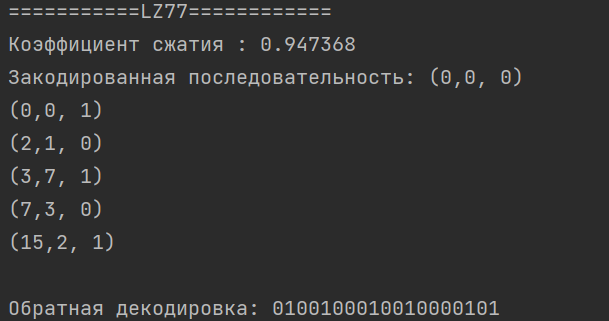


Рисунок 4. LZ77

Вручную декодируем полученную последовательность. Пробелов в строке нет, они используются для улучшения визуального восприятия

Таблица 2. Описание LZ77

|  |  |  |
| --- | --- | --- |
| Закодированная структура в памяти | Результат | Описание действия |
| (0,0, 0) | 0 | Ничего не берется из окна, ставим последний символ |
| (0,0, 1) | 0 1 | Ничего не берется из окна, ставим последний символ |
| (2,1, 0) | 01 0 0 | Сдвигаем на 2, ставим один символ и последний «0». |
| (3,7, 1) | 0100 1000100 1 | Сдвигаем на 3, берем 7 символов (если длина больше сдвига, начинаем сначала), ставим последний «1» |
| (7,3, 0) | 010010001001 000 0 | Сдвигаем на 7, берем 3 символов, ставим последний «0» |
| (15,2, 1) | 0100100010010000 10 1 | Сдвигаем на 15, берем 2 символа и ставим «1» |

Из таблицы 2 видно, что результат совпадает с изначальной строкой, значит кодировка верна.

Протестируем алгоритм LZ78 на строке из персонального варианта. Для удобства преобразуем текст «упупапекапекаупуп» в эквивалентный «upupapekapekaupup». Из рисунка 5 и таблицы 3 видно, что алгоритм работает корректно.

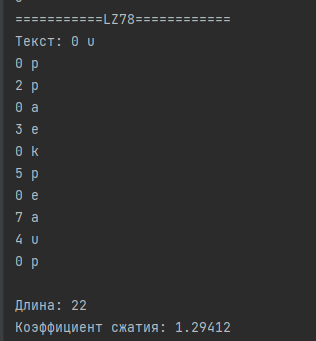


Рисунок 5. Результат LZ78

Рассмотрим процесс восстановления сжатого текста.

Таблица 3. Декодирование LZ78

|  |  |
| --- | --- |
| Структура в памяти | Результат |
| 0 u | u |
| 0 p | up |
| 2 p | up u p |
| 0 a | upup a |
| 3 e | upupa p e |
| 0 k | upupape k |
| 5 p | upupape k ap |
| 0 e | upupapekap e |
| 7 a | upupapekape k a |
| 4 u | upupapekapeka up |
| 0 p | upupapekapekaup up |

Проведем тестирование для алгоритма Шеннона-Фано в соответствии с персональным вариантом:

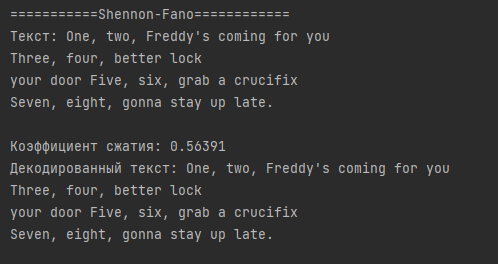


Рисунок 6. Результат работы кодировки Шеннона-Фано

Рассмотрим таблицу вероятностей:

Таблица 4. Вероятности в кодировке Шеннона

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| space | 0.165414 | n | 0.037594 | d | 0.0225564 | w | 0.0075188 |
| e | 0.0827068 | u | 0.037594 | x | 0.0150376 | ' | 0.0075188 |
| o | 0.075188 | \n | 0.0300752 | v | 0.0150376 | . | 0.0075188 |
| r | 0.0676692 | g | 0.0300752 | F | 0.0150376 | k | 0.0075188 |
| , | 0.0601504 | c | 0.0300752 | l | 0.0150376 | p | 0.0075188 |
| i | 0.0451128 | y | 0.0300752 | h | 0.0150376 | O | 0.0075188 |
| t | 0.0451128 | s | 0.0225564 | b | 0.0150376 | S | 0.0075188 |
| a | 0.037594 | f | 0.0225564 | T | 0.0075188 | m | 0.0075188 |

Следуя данному алгоритму, будем разделять полученные данные так, чтобы уравнить сумму весов слева и справа. В итоге, когда размер массива сократится до единицы, получим следующее дерево:

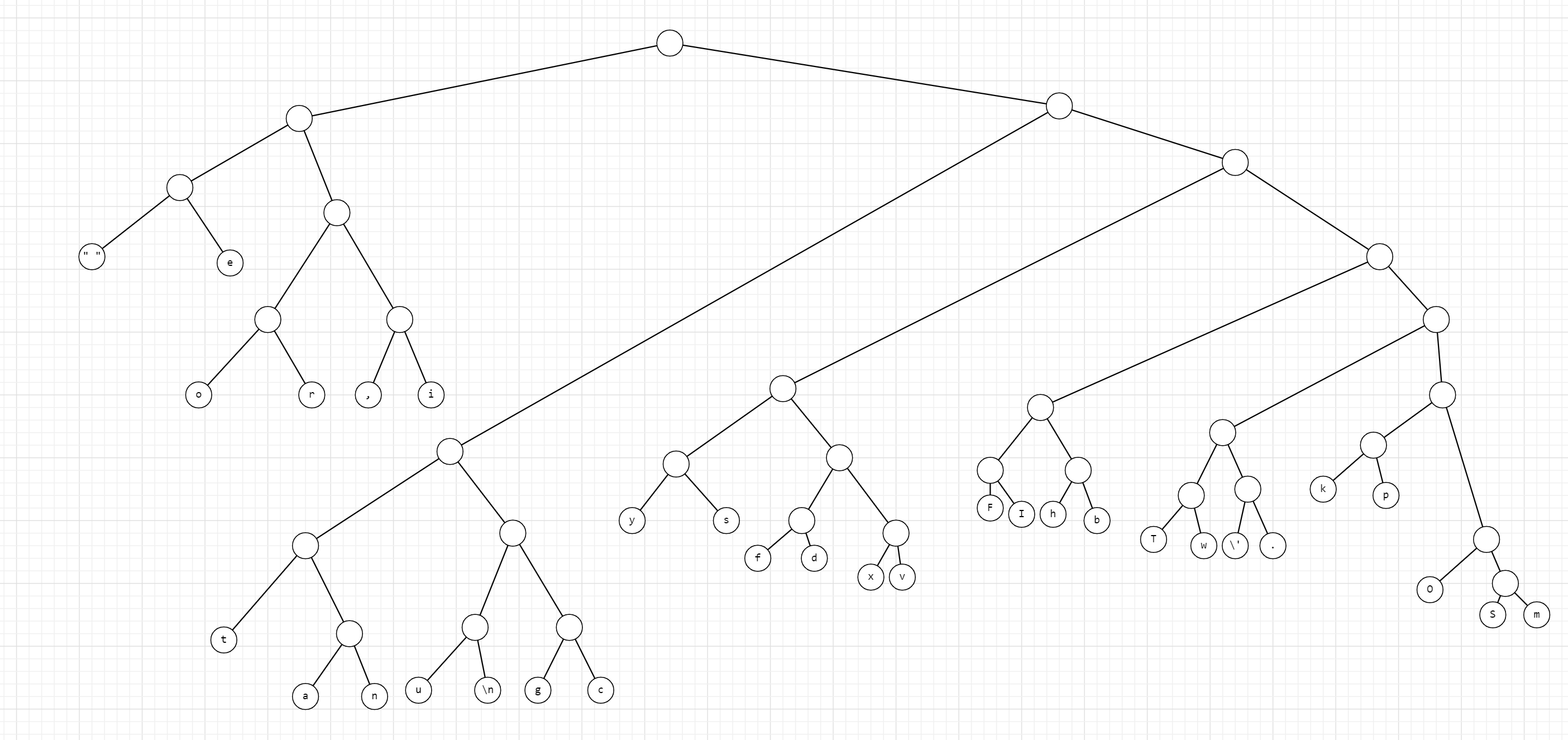


Рисунок 7. Полученное дерево

Таблица 5. Коды в шифровке по Шеннону

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| “ “ | 000 | a | 10010 | i | 0111 | s | 11001 |
| ' | 1111010 | b | 111011 | k | 1111100 | t | 1000 |
| , | 0110 | c | 10111 | l | 111001 | u | 10100 |
| . | 1111011 | d | 110101 | m | 11111111 | v | 110111 |
| F | 111000 | e | 001 | n | 10011 | w | 1111001 |
| O | 1111110 | f | 110100 | o | 0100 | x | 110110 |
| S | 11111110 | g | 10110 | p | 1111101 | y | 11000 |
| T | 1111000 | h | 111010 | r | 0101 | \n | 10101 |

На данной строке коэффициент сжатия: 0.56391

Рассмотрим результат кодировки методом Хаффмана в соответствии с персональным вариантом, строка “Smolnikov Alexey Borisovich”:

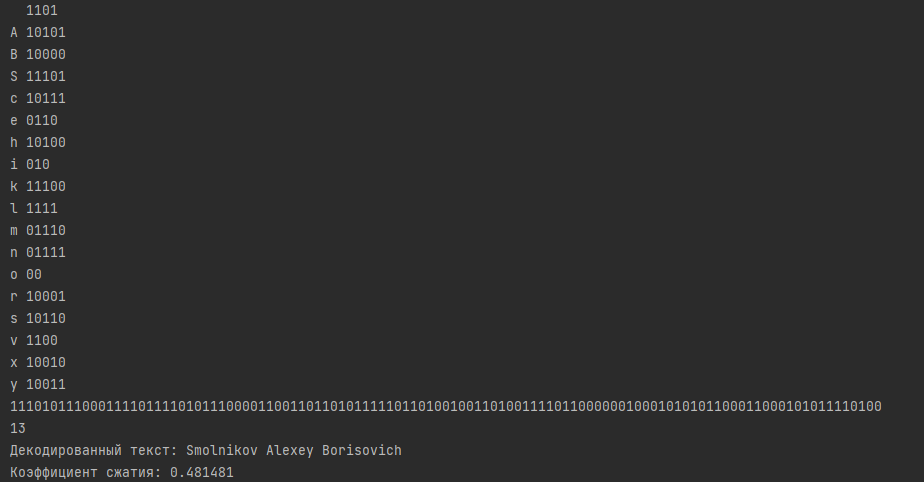


Рисунок 8. Кодировка по Хаффману

Таблица 6. Алфавит ФИО

|  |  |  |
| --- | --- | --- |
| Символ | Частота | Вероятность |
| i | 3 | 0,11111111111 |
| l, v, “ ”, e | 2 | 0,07407407407 |
| A, h, c, s, r, B, y, x, S, k, n, m | 1 | 0,03703703703 |

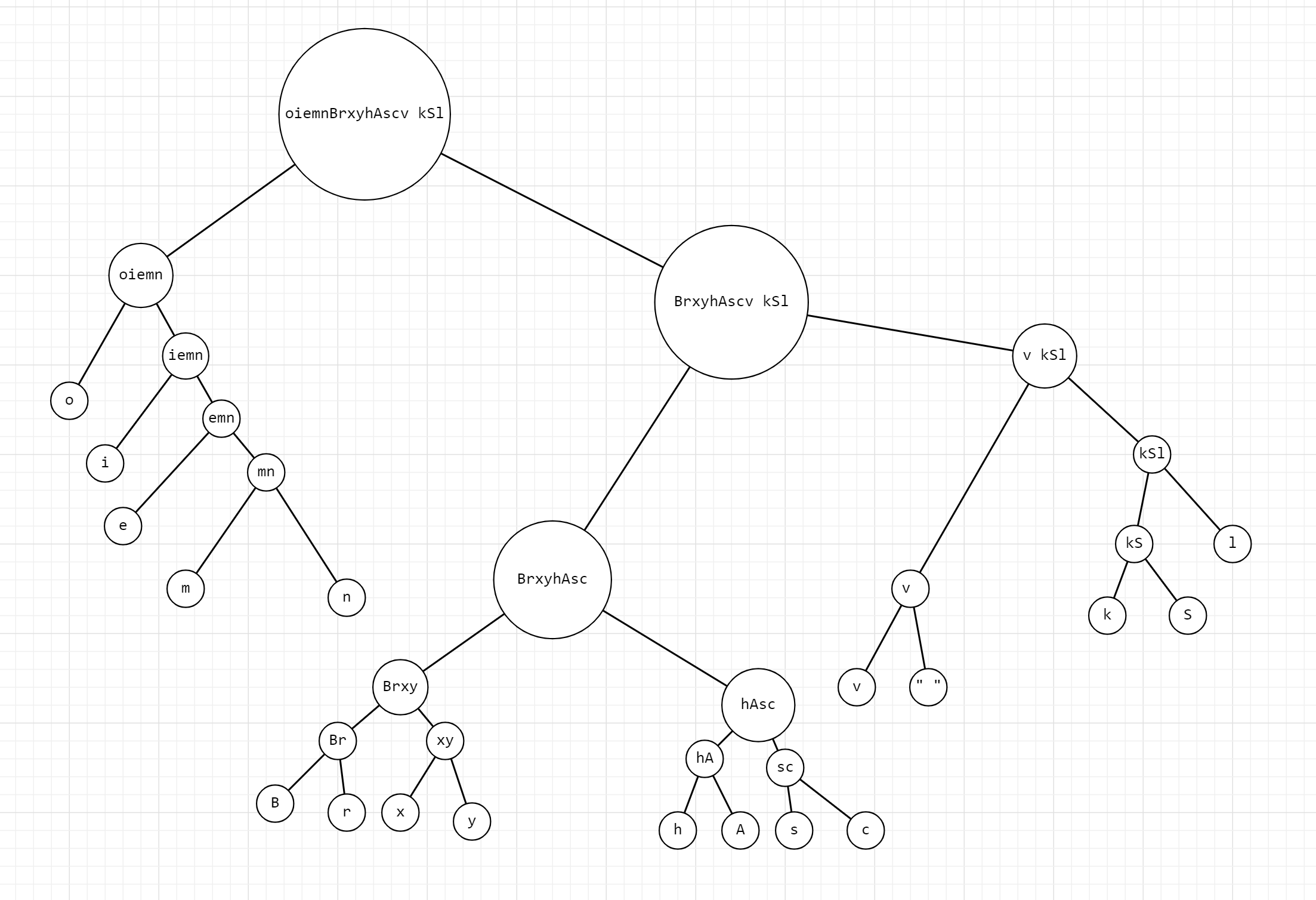


Рисунок 9. Дерево кодов по Хаффману

Коэффициент сжатия для данной строки – 0,481; средняя длина кода символа – 4,5 бита.

Рассчитаем дисперсию:

Применим алгоритм сжатия к файлу world192.txt и сравним результаты:

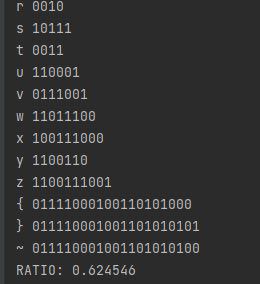


Рисунок 10. Коэффициент сжатия

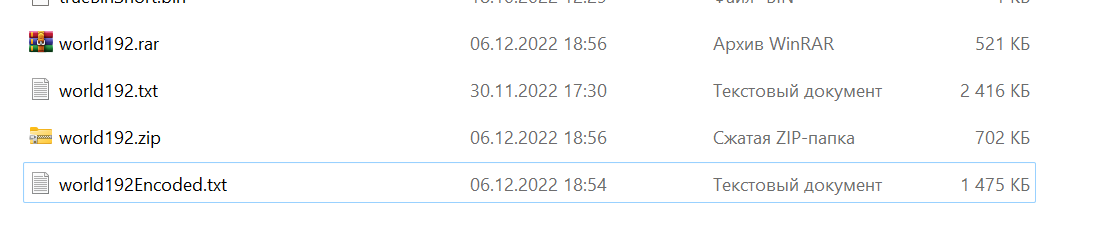


Рисунок 11. Сжатые файлы

Из рисунков 10 и 11 видно, что алгоритм Хаффмана в данной имплементации на выходе предоставляет коэффициент сжатия 0.62, в то время как более сложные алгоритмы сжатия намного эффективнее, коэффициент сжатия от 0.21 до 0.29.

# **Выводы**

В результате выполнения работы были:

* получены навыки по выполнению сжатия данных;
* были реализованы некоторые алгоритмы на языке С++.

# **Исходный код программы**

main.cpp

|  |
| --- |
| //  // Created by lexa2k on 27.11.2022.  //  #include <iostream>  #include <string>  #include "LZ77.h"  #include "LZ78.h"  #include "ShennonFano.h"  #include "az100000.h"  #include "Huffman.h"  using namespace std;  double calcucateCompressionRatio(const string& text, const string& compressedText){ return static\_cast<double>(compressedText.length())/static\_cast<double>(text.length()); }  double calcucateCompressionRatio(const string& text, const int compressedSize){ return static\_cast<double>(compressedSize)/static\_cast<double>(text.length()); }  string encodeRLE(string data){  string result;  const unsigned long long len = data.length();  if(len==0) return "";  unsigned long long cnt=1;  for(unsigned long long i=0;i<len;i++){  if(i==len-1 || data[i]!=data[i+1]){  result+=to\_string(cnt);  result+=data[i];  cnt=1;  } else cnt++;  }  return result;  }  string decodeRLE(string dataRLE){  string result;  const unsigned long long len = dataRLE.length();  if(len==0) return "";  unsigned long long cnt=0;  for(unsigned long long i=0;i<len;i++){  if(dataRLE[i]>='0' && dataRLE[i]<='9'){  cnt\*=10;  cnt+=dataRLE[i]-'0';  } else {  for(unsigned long long j=0;j<cnt;j++) result+=dataRLE[i];  cnt=0;  }  }  return result;  }  int main(){  system("chcp 65001");  int choise=-1;  cout<<"Лабораторная работа №7.'Алгоритмы сжатия'. Смольников Алексей. Вариант 2(22)"<<endl;  while(choise!=0){  cout<<"Выберите алгоритм сжатия:"<<endl;  cout<<"1. RLE"<<endl;  cout<<"2. LZ77"<<endl;  cout<<"3. LZ78"<<endl;  cout<<"4. Shennon-Fano"<<endl;  cout<<"5. Huffman"<<endl;  cout<<"0. Выход"<<endl;  cin>>choise;  switch (choise) {  case 1:{  cout<<"RLE encoding"<<endl;  string RLEtext;  cout<<"Введите текст: ";  cin>>RLEtext;  string textRLE = encodeRLE(RLEtext);  cout<<"Закодированный текст: "<<textRLE<<endl;  cout<<"Коэффициент сжатия:"<<calcucateCompressionRatio(RLEtext,textRLE)<<endl;  cout<<"Encoding long random a-z string (100 000). Ratio: "<<calcucateCompressionRatio(encodeRLE(az100000),az100000)<<endl;  break;  }  case 2:{  cout<<"===========LZ77============"<<endl;  LZ77 lz77;  string text;  text = "0100100010010000101";  lz77.encode(text);  cout<<"Коэффициент сжатия : "<<calcucateCompressionRatio(text,lz77.getEncodedLength())<<endl;  cout<<"Закодированная последовательность: ";  lz77.printEncoded();  cout<<endl;  cout<<"Обратная декодировка: "<<lz77.decode(lz77.encoded)<<endl;  break;  }  case 3:{  cout<<"===========LZ78============"<<endl;  LZ78 lz78;  string lz78text;  lz78text = "upupapekapekaupup";  lz78.encode(lz78text);  cout<<"Текст: "<< lz78text<<endl;  lz78.printEncoded();  cout<<endl;  cout<<"Длина: "<<lz78.getEncodedLength()<<endl;  cout<<"Коэффициент сжатия: "<<calcucateCompressionRatio(lz78text,lz78.getEncodedLength())<<endl;  break;  }  case 4:{  cout<<"===========Shennon-Fano============"<<endl;  string shenonText;  cout<<"Текст: ";  shenonText = "One, two, Freddy's coming for you \n"  "Three, four, better lock \n"  "your door Five, six, grab a crucifix \n"  "Seven, eight, gonna stay up late.\n";  cout<<shenonText<<endl;  ShennonFano ShFano;  ShFano.encode(shenonText);  cout<<"Коэффициент сжатия: "<<calcucateCompressionRatio(shenonText,ShFano.getByteSize())<<endl;  cout<<"Декодированный текст: "<<ShFano.decode()<<endl;  ShFano.printCodeTable();  break;  }  case 5:{  cout<<"===========Huffman============"<<endl;  Huffman huffman;  string huffmanText;  cout<<"Текст: ";  huffmanText = "Smolnikov Alexey Borisovich";  huffman.encode(huffmanText);  huffman.printEncoded();  cout<<huffman.getEncodedLength()<<endl;  cout<<"Декодированный текст: "<<huffman.decode()<<endl;  cout<<"Коэффициент сжатия: "<<calcucateCompressionRatio(huffmanText,huffman.getEncodedLength())<<endl;  cout<<"===========Huffman (encoding large)============"<<endl;  Huffman huffmanLarge;  huffmanLarge.encode("","world192.txt",true);  cout<<"RATIO: "<<(double)huffmanLarge.getEncodedLength()/(double)huffmanLarge.s\_length<<endl;  break;  }  case 0:{  cout<<"Выход"<<endl;  break;  }  }  }  return 0;  } |

LZ77.h

|  |
| --- |
| //  // Created by lexa2k on 27.11.2022.  //  #ifndef SIAOD2SEM\_LZ77\_H  #define SIAOD2SEM\_LZ77\_H  #include <vector>  #include <string>  #include <iostream>  using namespace std;  //for LZ77 use alphabet [0,1] and encode string "0100100010010000101"  struct Node{  unsigned char offset;  unsigned char length;  char next;  };  class LZ77{  int windowLength=8;  Node fillNode(int off,int l, char nex);  public:  vector<Node> encoded;  vector<Node> encode(string text);  string decode(vector<Node> encoded);  int getEncodedLength();  void printEncoded();  };  #endif //SIAOD2SEM\_LZ77\_H |

LZ77.cpp

|  |
| --- |
| //  // Created by lexa2k on 27.11.2022.  //  //for LZ77 use alphabet [0,1] and encode string "0100100010010000101"  #include "LZ77.h"  Node LZ77::fillNode(int off,int l, char nex){  return Node{off,l,nex};  }  vector<Node> LZ77::encode(std::string text) {  vector<Node> result;  windowLength=text.length();  result.push\_back(fillNode(0,0,text[0]));  for(int i=1;i<text.length();i++) {  string window;  if(true){  window = text.substr(0,i);  }else{  window= text.substr(i-windowLength,windowLength);  }  //finding match in buffer  int matchLength=0;  int matchOffset=0;  for(int j=0;j<window.length();j++){  int k=0;  int pos=j+k;  int startPos=j+k;  while(window[pos]==text[i+k]){  k++;  pos++;  if(pos>=window.length()){  pos-=(int)window.length();  }  }  if(k>matchLength){  matchLength=k;  matchOffset=i-startPos;  }  }  if(matchLength==0){  result.push\_back(fillNode(0,0,text[i]));  }else{  result.push\_back(fillNode(matchOffset,matchLength,text[i+matchLength]));  i+=matchLength;  }  }  encoded=result;  return result;  }  string LZ77::decode(vector<Node> encoded){  string result;  for(int i=0;i<encoded.size();i++){  if(encoded[i].offset==0){  result+=encoded[i].next;  }else{  int pos=result.length()-encoded[i].offset;  int resultLen=result.length();  for(int j=0;j<encoded[i].length;j++){  result+=result[pos];  pos++;  if(pos>=resultLen){  pos-=resultLen;  }  }  result+=encoded[i].next;  }  // cout<<result<<endl;  }  return result;  }  void LZ77::printEncoded() {  for(int i=0;i<encoded.size();i++){  if(encoded[i].next=='\0'){  cout<<"("<<(int)encoded[i].offset<<","<<(int)encoded[i].length<<", NTERM)"<<endl;  }else{  cout<<"("<<(int)encoded[i].offset<<","<<(int)encoded[i].length<<", "<<encoded[i].next<<")"<<endl;  }  }  }  int LZ77::getEncodedLength() {  return encoded.size()\*(1+1+1);  } |

LZ78.h

|  |
| --- |
| //  // Created by lexa2k on 29.11.2022.  //  #ifndef SIAOD2SEM\_LZ78\_H  #define SIAOD2SEM\_LZ78\_H  #include <string>  #include <vector>  #include <iostream>  #include <map>  using namespace std;  struct LZ78node{  unsigned char index;  char next;  };  class LZ78{  vector<LZ78node> encoded;  map<string,int> dictionary;  public:  int getEncodedLength();  void printEncoded();  vector<LZ78node> encode(string text);  string decode();  };  #endif //SIAOD2SEM\_LZ78\_H |

LZ78.cpp

|  |
| --- |
| //  // Created by lexa2k on 29.11.2022.  //  #include "LZ78.h"  int LZ78::getEncodedLength() {  return 2\*encoded.size();  }  void LZ78::printEncoded() {  for(int i=0;i<encoded.size();i++){  cout<<(int)encoded[i].index<<" "<<encoded[i].next<<endl;  }  }  vector<LZ78node> LZ78::encode(string text) {  int index=0;  string buffer="";  vector<LZ78node> answer;  for(int i=0;i<text.length();i++){  if(dictionary.find(buffer+text[i]) != dictionary.end()){  buffer+=text[i];  }else{  answer.push\_back(LZ78node{static\_cast<unsigned char>(dictionary[buffer]), text[i]});  dictionary[buffer+text[i]]=dictionary.size()+1;  buffer="";  }  }  if(!buffer.empty()){  char last\_ch = buffer[buffer.length()-1];  buffer.pop\_back();  answer.push\_back(LZ78node{static\_cast<unsigned char>(dictionary[buffer]),last\_ch});  }  encoded=answer;  return answer;  }  string LZ78::decode(){  string res="";  for(int i=0;i<encoded.size();i++){  }  } |

ShennonFano.h

|  |
| --- |
| //  // Created by lexa2k on 29.11.2022.  //  #ifndef SIAOD2SEM\_SHENNONFANO\_H  #define SIAOD2SEM\_SHENNONFANO\_H  #include <string>  #include <vector>  #include <iostream>  #include <map>  #include <tuple>  using namespace std;  class ShennonFano{  long long s\_length=0;  vector<bool> answer;  map<char,int>count;  vector<tuple<char,double>>dataProbability;  map<char,string>codeTable;  int tempCnt=0;  struct treeNode{  char c = '\r';  string code;  treeNode \*left;  treeNode \*right;  };  public:  treeNode \*rootTree = new treeNode;  void scanText(string text);  vector<bool> encode(string text);  void buildCodeTree(vector<tuple<char,double>>data, string curValue, treeNode \*root);  string decode();  int getByteSize();  void printCodeTable();  ~ShennonFano();  };  #endif //SIAOD2SEM\_SHENNONFANO\_H |

ShennonFano.cpp

|  |
| --- |
| //  // Created by lexa2k on 29.11.2022.  //  #include <fstream>  #include "ShennonFano.h"  #include <fstream>  void ShennonFano::scanText(string text) {  for(auto c: text){  count.find(c) == count.end() ? count[c]=1 : count[c]++;  }  s\_length=text.length();  for(auto it : count){  dataProbability.emplace\_back(it.first,(double)it.second/ (double)s\_length);  }  sort(dataProbability.begin(),dataProbability.end(),[](tuple<char,double> const &a, tuple<char,double> const &b){  return get<1>(a) >= get<1>(b);  });  }  vector<bool> ShennonFano::encode(string text) {  scanText(text);  buildCodeTree(dataProbability,"",rootTree);  // for(auto c: codeTable){  // cout << c.first << " " << c.second << endl;  // }  for(auto c: text){  for(auto code: codeTable[c]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  return {};  }  ShennonFano::~ShennonFano() {  answer.clear();  count.clear();  dataProbability.clear();  codeTable.clear();  // treeNode \*temp = rootTree;  // while(temp!= nullptr){  // if(temp->left!= nullptr){  // temp = temp->left;  // delete temp->right;  // } else {  // delete temp;  // temp = nullptr;  // }  // }  }  void ShennonFano::buildCodeTree(vector<tuple<char,double>>data,string curValue, treeNode \*root) {  if(data.size()==1){  root->c = get<0>(data[0]);  root->code = curValue;  codeTable[get<0>(data[0])] = curValue;  tempCnt++;  return;  }else if(data.size()==0){  return;  }  int minPos=1;  double minDelta=1.0;  for(int i=1;i<data.size();i++){  double accumulateLeft=0.0;  double accumulateRight=0.0;  for(int j=0;j<data.size();j++){  if(j<i){  accumulateLeft+=get<1>(data[j]);  }else{  accumulateRight+=get<1>(data[j]);  }  }  double delta = abs(accumulateLeft-accumulateRight);  if(delta<minDelta){  minDelta=delta;  minPos=i;  }  }  vector<tuple<char,double>>leftV;  vector<tuple<char,double>>rightV;  leftV = vector<tuple<char,double>>(data.begin(),data.begin()+minPos);  rightV = vector<tuple<char,double>>(data.begin()+minPos,data.end());  treeNode \*leftNode = new treeNode;  treeNode \*rightNode = new treeNode;  root->left = leftNode;  root->right = rightNode;  buildCodeTree(leftV, curValue+"0",root->left);  buildCodeTree(rightV, curValue+"1",root->right);  }  int ShennonFano::getByteSize() {  if(answer.size()%8==0){  return answer.size()/8;  }else{  return answer.size()/8+1;  }  }  string ShennonFano::decode() {  string result="";  string temp\_code;  while(!answer.empty()){  temp\_code+=answer.front() ? "1" : "0";  answer.erase(answer.begin());  for(auto it: codeTable){  if(it.second==temp\_code){  result+=it.first;  temp\_code.clear();  break;  }  }  }  return result;  }  void ShennonFano::printCodeTable() {  for(auto it: codeTable){  cout << it.first<< endl;  }  for(auto it: codeTable){  cout << it.second<< endl;  }  } |

Huffman.h

|  |
| --- |
| //  // Created by lexa2hk on 30.11.2022.  //  #ifndef SIAOD2SEM\_HUFFMAN\_H  #define SIAOD2SEM\_HUFFMAN\_H  #include <string>  #include <vector>  #include <tuple>  #include <iostream>  #include <map>  #include <numeric>  #include "fstream"  using namespace std;  class Huffman{  struct Node{  string s="";  int weight=0;  };  struct treeNode{  Node data;  treeNode \*left;  treeNode \*right;  treeNode \*parent;  };  vector<Node> data;  vector<bool> answer;  map<string,string>codeTable;  map<string,string>decodeTable;  treeNode \*rootTree;  public:  void scanText(string dataText);  int find\_el(char c);  void createCodes(treeNode \*pNode, string code);  void printEncoded();  int getEncodedLength();  string decode();  vector<bool> encode(string dataText, string fileName="", bool isFile=false);  long long s\_length=0;  };  #endif //SIAOD2SEM\_HUFFMAN\_H |

Huffman.cpp

|  |
| --- |
| //  // Created by lexa2hk on 30.11.2022.  //  #include "Huffman.h"  void Huffman::scanText(string dataText) {  for(char c:dataText){  s\_length++;  int index = find\_el(c);  if(index!=-1){  this->data[index].weight++;  }else{  this->data.push\_back(Node{string(1,c),1});  }  }  }  vector<bool> Huffman::encode(string dataText, string fileName, bool isFile) {  if(!isFile) {  scanText(dataText);  sort(this->data.begin(), this->data.end(), [&](Node a, Node b) { return a.weight > b.weight; });  }else{  ifstream file(fileName);  string str;  while (getline(file, str)) {  scanText(str+"\n");  }  sort(this->data.begin(), this->data.end(), [&](Node a, Node b) { return a.weight > b.weight; });  }  for(auto el:this->data){  cout << el.s << " " << el.weight << endl;  }  vector<treeNode\*> nodes;  for(int i=0;i<this->data.size();i++){  nodes.push\_back(new treeNode{this->data[i], nullptr, nullptr, nullptr});  }  while(nodes.size()>1){  treeNode \*left = nodes[nodes.size()-1];  treeNode \*right = nodes[nodes.size()-2];  treeNode \*parent = new treeNode{Node{"",left->data.weight+right->data.weight},left,right,nullptr};  left->parent=parent;  right->parent=parent;  nodes.pop\_back();  nodes.pop\_back();  nodes.push\_back(parent);  sort(nodes.begin(), nodes.end(), [&](treeNode \*a, treeNode \*b) { return a->data.weight > b->data.weight; });  }  this->rootTree = nodes[0];  //create codes for all tree  createCodes(rootTree,"");  //cout map  for(auto i:codeTable){  cout<<i.first<<" "<<i.second<<endl;  }  //encode data  if(!isFile){  for(char c : dataText){  for(char code : codeTable[string(1,c)]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  }else{  ifstream file(fileName);  string str;  while (getline(file, str)) {  for(char c : str){  for(char code : codeTable[string(1,c)]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  for(char code : codeTable[string(1,'\n')]){  if(code=='0') answer.push\_back(false);  else answer.push\_back(true);  }  }  }  //write answer to file (every 8 bits to char)  ofstream file("encoded.txt");  string str;  for(int i=0;i<answer.size();i++){  if(answer[i]) str+="1";  else str+="0";  if(str.length()==8){  file<<char(stoi(str, nullptr, 2));  str="";  }  }  if(str.length()!=0){  while(str.length()!=8){  str+="0";  }  file<<char(stoi(str, nullptr, 2));  }  file.close();  return answer;  }  int Huffman::find\_el(char c) {  for(int i=0;i<data.size();i++){  if (string(1,c) == data[i].s)  return i;  }  return -1;  }  void Huffman::createCodes(Huffman::treeNode \*pNode, string code) {  if(pNode->left== nullptr && pNode->right== nullptr){  codeTable[pNode->data.s]=code;  decodeTable[code]=pNode->data.s;  return;  }  createCodes(pNode->left,code+"0");  createCodes(pNode->right,code+"1");  }  void Huffman::printEncoded(){  for(bool i: answer){  cout<<i;  }  cout<<endl;  }  int Huffman::getEncodedLength() {  return answer.size()/8;  }  string Huffman::decode(){  string decoded;  string code;  for(bool i:answer){  if(i) code+="1";  else code+="0";  if(decodeTable.find(code)!=decodeTable.end()){  decoded+=decodeTable[code];  code="";  }  }  return decoded;  } |