

# Лабораторная работа №6. Коды Хаффмана.

## 1 Теоретические сведения

Сжатие данных – это процесс, обеспечивающий уменьшение объема данных путем сокращения их избыточности. Сжатие данных связано с компактным расположением порций данных стандартного размера.

Хаффмановское кодирование (сжатие) – это широко используемый метод сжатия, присваивающий символам алфавита коды переменной длины основываясь на вероятностях появления этих символов.

Идея алгоритма состоит в следующем: зная вероятности вхождения символов в исходный текст, можно описать процедуру построения кодов переменной длины, состоящих из целого количества битов. Символам с большей вероятностью присваиваются более короткие коды. Таким образом, в этом методе при сжатии данных каждому символу присваивается оптимальный префиксный код, основанный на вероятности его появления в тексте.

Префиксный код – это код, в котором никакое кодовое слово не является префиксом любого другого кодового слова. Эти коды имеют переменную длину.

Оптимальный префиксный код – это префиксный код, имеющий минимальную среднюю длину.

Алгоритм Хаффмана можно разделить на два этапа.

1. Определение вероятности появления символов в исходном тексте.  
Первоначально необходимо прочитать исходный текст полностью и подсчитать вероятности появления символов в нем (иногда подсчитывают, сколько раз встречается каждый символ). Если при этом учитываются все 256 символов, то не будет разницы в сжатии текстового или файла иного формата.
2. Нахождение оптимального префиксного кода. Далее находятся два символа  $a$  и  $b$  с наименьшими вероятностями появления и заменяются одним фиктивным символом  $x$ , который имеет вероятность появления, равную сумме вероятностей появления символов  $a$  и  $b$ . Затем, используя эту процедуру рекурсивно, находится оптимальный префиксный код для меньшего множества символов (где символы  $a$  и  $b$  заменены одним символом  $x$ ). Код для исходного множества символов получается из кодов замещающих символов путем добавления 0 или 1 перед кодом замещающего символа, и эти два новых кода принимаются как коды заменяемых символов. Например, код символа  $a$  будет соответствовать коду  $x$  с добавленным нулем перед этим кодом, а для символа  $b$  перед кодом символа  $x$  будет добавлена единица.

Коды Хаффмана имеют уникальный префикс, что и позволяет однозначно их декодировать, несмотря на их переменную длину.

## **2 Задание**

### **2.1 Прямая задача**

Необходимо написать программу, которая по заданному текстовому файлу строит код Хаффмана и осуществляет кодирование данного файла. Префиксный код и закодированный текст должны быть сохранены в разных текстовых файлах.

### **2.2 Обратная задача**

Необходимо написать программу, которая по двум текстовым файлам, в первом содержится префиксный код, во втором — закодированный текст, восстанавливает исходный текст и сохраняет его в текстовый файл.

### **Контрольные вопросы**

1. Что такое префиксный код?
2. Из-за чего сжатие методом Хаффмана может увеличить размер файла?
3. Каков алгоритм построения кода Хаффмана?