Злобин Роман Юрьевич, ФИТ-2 Статистическое сжатие.

Белорусская транскрипция: злобiн раман юр’евiч

Всего 18 символов, из них: з – 1, л – 1, о – 1, б – 1, i – 2, н – 2, р – 2, а – 2, ю – 1, ‘ – 1, e – 1, в – 1, ч – 1, м – 1.

Получаем следующую статистику: 14 – различных символов, следовательно 4 бита на каждый.

P(i) = 0,111

P(н) = 0,111

P(р) = 0,111

P(а) = 0,111

P(з) = 0,0556

P(л) = 0,0556

P(о) = 0,0556

P(б) = 0,0556

P(ю) = 0,0556

P(‘) = 0,0556

P(е) = 0,0556

P(в) = 0,0556

P(ч) = 0,0556

P(м) = 0,0556

Проверка: 0,111 \* 4 + 0,0556 \* 14 = 1

Метод Шеннона – Фано:

Берём отсортированные вероятности появления символов и разделяем их на две равные (близкие) по вероятности группы, и элементам верхней группы приписываем 1, а нижней – 0. Получившиеся группы делим таким же образом, до тех пор, пока не образуются группы по одному элементу.

P(i) = 0,111

P(н) = 0,111

\_.\_.\_.\_.\_.\_.\_

P(р) = 0,111

P(а) = 0,111

P(з) = 0,0556

\_\_\_\_\_\_\_\_\_\_\_

P(л) = 0,0556

P(о) = 0,0556

\_..\_..\_..\_..\_..\_

P(б) = 0,0556

P(ю) = 0,0556

\_.\_.\_.\_.\_.\_.\_

P(‘) = 0,0556

P(е) = 0,0556

\_..\_..\_..\_..\_..\_

P(в) = 0,0556

\_\_.\_\_.\_\_.\_\_

P(ч) = 0,0556

P(м) = 0,0556

111

110

\_.\_.\_.\_.\_.\_.\_

101

1001

1000

\_\_\_\_\_\_\_\_\_\_\_

0111

0110

\_..\_..\_..\_..\_..\_

0101

0100

\_.\_.\_.\_.\_.\_.\_

0011

0010

\_..\_..\_..\_..\_..\_

0001

\_\_.\_\_.\_\_.\_\_

00001

00000

злобiн – 1000 0111 0110 0101 111 110

Исходная кодировка – 24 бит, сжатая – 22 бита.

Таким образом сжали на 2 бит.

Метод Хаффмана.

Отличается от метода Шеннона – Фано лишь в части кодирования символов исходного алфавита. В данном случае бинарные коды создаются на основе дерева, ветви которого обозначаются бинарными символами. Бинарным кодом символа исходного алфавита будет последовательность обозначений ветвей дерева от корня до листа, соответствующего этому символу.

Построение дерева начинается с сортирования символов исходного алфавита в порядке убывания. Далее выбираются два символа (ai, aj) с наименьшими вероятностями (р(ai), р(aj)) и объединяются в узел. Ветви этого узла обозначаются «1» и «0». Данный узел рассматривается далее как новый, виртуальный символ (aij), которому будет соответствовать вероятность р(aij) = р(ai) + р(aj). Такой виртуальный символ будет рассматриваться далее наравне с остальными символами исходного алфавита. Два его потомка из дальнейшего рассмотрения исключаются. Создаются новые узлы дерева по тому же принципу. Корень дерева образуют два символа с наибольшими вероятностями.

В данном построении красные ветви обозначают «1», синие – «0».

P(i) = 0,111

P(а) = 0,111

P(н) = 0,111

P(р) = 0,111

P(з) = 0,0556

P(л) = 0,0556

P(о) = 0,0556

P(б) = 0,0556

P(е) = 0,0556

P(ю) = 0,0556

P(‘) = 0,0556

P(в) = 0,0556

P(ч) = 0,0556

P(м) = 0,0556

P(..) = 0,1112

P(..) = 0,1112

P(..) = 0,1112

P(..) = 0,1112

P(..) = 0,1112

P(..) = 0,222

P(..) = 0,222

P(..) = 0,2224

P(..) = 0,2224

P(..) = 0,3332

P(..) = 1

P(..) = 0,4444

P(..) = 0,5556

В соответствии с графом изображённом выше, имеем следующие бинарные коды:

i 001

н 001

р 1111

а 1110

з 1101

л 1100

о 1011

б 1010

ю 1001

‘ 1000

е 0111

в 0110

ч 0101

м 0100

злобiн – 1101 1100 1011 1010 001 001

Исходная кодировка – 24 бит, сжатая – 22 бита.

Таким образом сжали на 2 бит.