# Кодирование. Префиксные коды. Код Фано, код Хаффмана.

Рассматривается двоичное кодирование символов из произвольного алфавита $A$.

Кодом для алфавита $A$ называется соответствие элементов (символов) алфавита двоичным последовательностям. Для каждого слова $W$ в алфавите $A$, т.е. для любой последовательности букв алфавита $A$, рассмотрим последовательность $\alpha(W)$ из 0 и 1, которая получается, если записать подряд коды всех букв из $W$ без разделителей.

Код называется *однозначным* (или *разделимым*), если $\alpha(P) \neq \alpha(Q)$ для любых $P \neq Q$.

Код называется *префиксным*, если ни один код буквы не являются началом кода другой буквы (слово $\alpha(P)$ не является началом слова $\alpha(Q)$, если буквы $P$ и $Q$ различны).

1. (*разминочное*) Доказать, что префиксный код является однозначным.
2. Привести пример однозначного кода, не являющегося префиксным.
3. Пусть кодом каждого из данных чисел является его двоичное представление наименьшей возможной длины (без лидирующих нулей), например: $1 \to 1$, $5 \to 101$ и т.д.

   Является ли кодирование однозначным:

   (a) $L = \{1, 2, 4, 9, 50\}$
   (b) $L = \{1, 2, 4, 17, 98\}$
4. Построить двоичный префиксный код с заданной последовательностью длин кодовых слов:

   (a) $S = \{1, 2, 3, 3\}$
   (b) $S = \{1, 2, 4, 4, 4, 4\}$
   (c) $S = \{1, 2, 3, 3, 4, 4, 4, 4\}$

Пусть для каждой буквы $a$ алфавита $A$ фиксирована её *частота* $p(a)$ положительное число, причём сумма частот всех букв равна 1. Тогда средняя длина кода это сумма:

$$E = \sum_{a \in A} p(a) \cdot |\alpha(a)|$$

где $|\alpha(a)|$ — длина кодового слова $\alpha(a)$ буквы $a$.

Если в слове длины $N$ буква $a$ встречается с частотой $p(a)$, то таких букв будет $Np(a)$ и на их кодирование уйдёт $Np(a)|\alpha(a)|$ битов.
Общая длина кода будет $\sum_{a \in A} Np(a)|\alpha(a)|$ и в среднем на кодирование каждой буквы уйдёт $E$ битов.

К сожалению, выбрать много коротких кодовых слова мешает одно препятствие. Длины кодовых слов $n_1, n_2, \ldots, n_k$ должны удовлетворять неравенству:

$$\sum_i 2^{-n_i} \leqslant 1$$

называемому *неравенством Крафта-Макмиллана.*

6. Доказать, что для всякого префиксного кода выполняется неравенство Крафта-Макмиллана.
7. Пусть даны $k$ целых положительных чисел $n_1, n_2, \ldots n_k$, удовлетворяющих неравенству Крафта-Макмиллана. Докажите, что можно построить префиксный код для $k$-буквенного алфавита с длинами кодовых слов $n_1, n_2, \ldots n_k$.

8* Докажите, что неравенство Крафта-Макмиллана выполняется не только для префиксного кода, но и для любого однозначного кода.

Задача о коде минимальной средней длины: для данных положительных $p_1, \ldots, p_k$, равных в сумме 1, найти положительные $n_1, \ldots, n_k$, для которых выполнено неравенство Крафата-Макмиллана, а сумма

$$\sum_{i=1}^{k} p_i n_i$$

является минимально возможной среди наборов чисел $n_1, \ldots, n_k$, удовлетворяющих неравенству.

9. Некоторые наблюдения:
   (a) Докажите, что для алфавита из двух букв оптимальный код состоит из двух слов длины 1, независимо от частот букв.
   (b) Пусть частоты расположены в убывающем порядке: $p_1 > p_2 > \cdots > p_k$. Докажите, что тогда длины слов оптимального кода идут в неубывающем порядке: $n_1 \leqslant n_2 \leqslant \cdots \leqslant n_k$.
   (c) Верно ли наблюдение 9(b), если частоты расположены в невозрастающем порядке (т.е. возможны равные)?

10. Пусть частоты расположены в невозрастающем порядке: $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_k$, а длины слов в *оптимальном* коде расположены в неубывающем порядке $n_1 \leqslant n_2 \leqslant \cdots \leqslant n_k$.

    Доказать, что $n_{k-1} = n_k$.

11. *Код и алгоритм Хаффмена*

    Докажите, что задачу поиска длин кодовых слов оптимального кода для $k$ упорядоченных частот $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_k$ можно свести к задаче поиска длин оптимального кода для $k-1$ частот $p_1, p_2, \ldots, p_{k-2}, p_{k-1} + p_k$ (частоты двух самых редких букв объединены).

12. Рассмотрим строку над алфавитом из четырёх символов A, C, G, T с частотами $31\%, 20\%, 9\%, 40\%$ соответственно. Каким будет код Хаффмана?

13. Пусть частоты вхождения символов $a, b, c, d, e$ в строку равны $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}$ соответственно.
    (a) Каким будет код Хаффмана?
    (b) Если использовать его для сжатия строки длины $10^6$ с указанными выше частотами вхождения пяти символов, какова будет длина закодированной строки (в битах)?

14. Может ли код Хаффмана для трёхбуквенного алфавита оказаться таким (для каких-то частот):
    (a) {0, 10, 11}
    (b) {0, 1, 00}
    (c) {10, 01, 00}

15. Докажите:
    (a) Если частота одного из символов превосходит $\frac{2}{5}$, то обязательно найдётся символ, код которого будет состоять из 1 бита.
    (b) Если частоты всех символов меньше $\frac{1}{3}$, то коды всех символов будут длиннее одного бита.

16. Какова максимальная длина кодового слова в коде Хаффмана для $n$ символов? Приведите пример частот, при которых достигается такая длина.

17. *Задание по программированию 1*

    Для файлов (https://prog.sch57.ru/books/data/) посчитайте частоты вхождения символов (обратите внимание, что некоторые файлы бинарные, для них алфавит состоит из 256 возможных значений байтов). Тут можно посмотреть, как работать с бинарными файлами.

    Для посчитанных частот вычислите объём сжатого файла (в битах). Придумайте, как хранить дерево, описывающее код Хаффмана.

18. *Задание по программированию 2*

    Напишите программу-архиватор, сжимающую Джойсовского Улисса больше, чем на 40%.