**Тема 3. Кодирование**

1. **Кодирование**

# Алфавитный подход к измерению информации

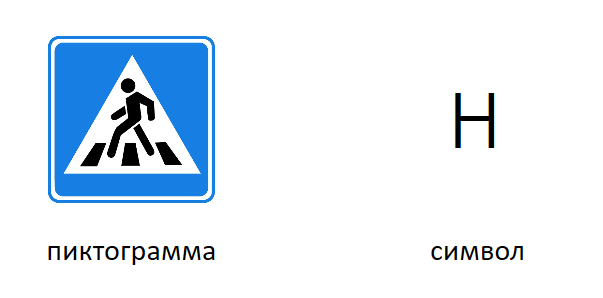
# Вероятностный подход к измерению информации

1. **Единицы измерения информации**

**1. Кодирование**

**Кодированием**  называют запись информации с использованием определенной знаковой системы. Также можно сказать, что кодирование – это представление информации в такой форме, которая будет удобна для хранения, передачи и обработки. В некоторых случаях кодирование – это замена одних знаков на другие, например, транслитерация символов.

**Знак** – это заменитель какого-либо объекта, предмета, явления, свойства. Каждый знак соответствует определённому образу. Знаками можно считать всевозможные пиктограммы и символы. Однако, если значения пиктограмм как правило интуитивно понятны, то значения символов могут зависеть от той знаковой системы к которой они принадлежат.



Например, на картинке мы видим знак пешеходного перехода. О его значении легко догадаться и его ни с чем не перепутаешь. А что это за символ? Может, это русская буква «н» или латинская буква «h», а может это символ водорода из периодической таблицы элементов или обозначение единицы измерения силы – Ньютона? Не зная, какая знаковая система используется, мы не сможем ответить на этот вопрос.

**Знаковая система** представляет собой множество входящих в неё знаков, то есть алфавит, и правила использования этих знаков. С этой точки зрения язык также является знаковой системой. Знаки в языке могут быть как символьными, так и звуковыми, поскольку мы используем письменную и устную речь. Если говорить о естественных языках, то есть тех языках, которые люди используют для общения, то они, как правило, гибкие и динамичные, в них много правил и есть исключения, чего не скажешь о формальных языках, где правила строгие и любые сочетания знаков всегда имеют одинаковый смысл.

**Алфавитом** называется набор знаков, используемых в языке. Мощностью алфавита называется количество знаков в нём. Например, мощность латинского алфавита равна 26, а русского – 33. В дальнейшем мощность алфавита будем обозначать буквой М.

Возьмем некоторый алфавит, скажем, состоящий из трёх букв {A, B, C}, т.е. имеющий мощность 3, и составим из него сообщения различной длины. Сообщением является любая последовательность символов алфавита, символы могут повторяться или не повторяться, но все сообщения должны быть разными. Для начала составим сообщения, состоящие из одного символа. Их получится только 3, столько же, сколько символов в алфавите. Теперь возьмем сообщения из двух символов. Получим следующие комбинации: АА, АВ, АС, ВА, ВВ, ВС, СА, СВ, СС. Всего 9 различных сообщений.

**Количество различных сообщений** N вычисляется по формуле:

**N = ML**

где М - мощность алфавита,

L - длина сообщения.

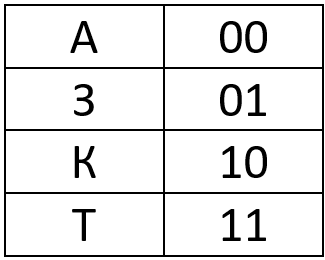
Чаще всего мы будем использовать для составления кодов алфавит, состоящий из нулей и единиц. Это обусловлено тем, что любая информация в памяти компьютера представлена в двоичном виде. Мощность такого алфавита равна двум. Значит количество различных сообщений или кодов будет вычисляться по формуле:

**N = 2L**

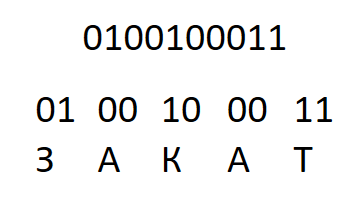
**Равномерное и неравномерное кодирование**

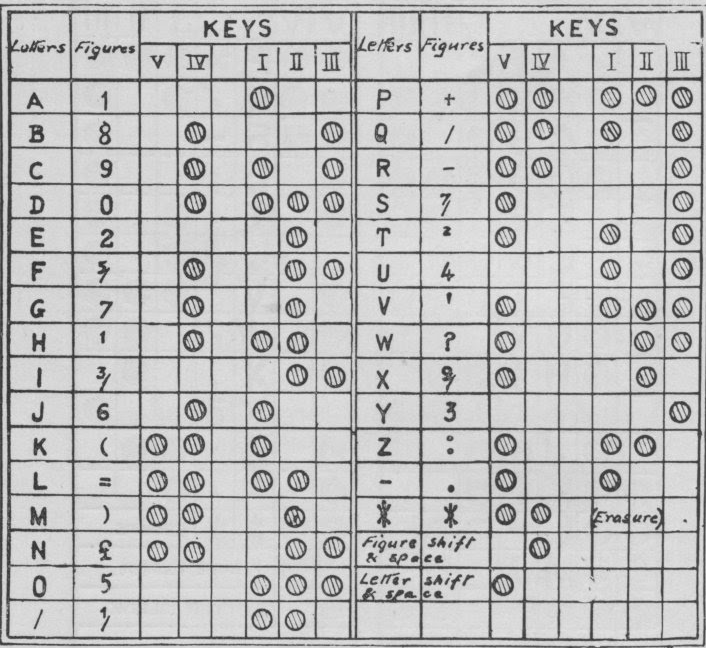
Составляя сообщения фиксированной длины, мы можем каждое из таких сообщений сделать кодом для определенного символа исходного алфавита. При этом каждый код или кодовое слово будет иметь равную длину. Такой код называется **равномерным**. Его удобство в том, что для декодирования сообщения достаточно разделить его на равные по длине коды и затем заменить их на символы исходного алфавита по таблице.

Например, 4 буквы некоторого алфавита закодированы следующим образом:



Декодируем сообщение. Для этого разделим его на группы по 2 двоичных разряда и выполним замену по таблице.





Код Бодо

Если обратиться к истории, то можно вспомнить, что одним из примеров использования равномерного двоичного кода был телеграфный код Бодо, разработанный французским инженером в начале 70-ых годов 19 века. Каждому символу в его коде отводилось 5 двоичных знаков. Вводился этот код с помощью 5 клавиш. Если клавиша была нажата, то это соответствовало единице в коде (на рисунке обозначено кружочком), а если не нажата – то нулю (пустая клетка на рисунке). Можно вычислить, что количество различных кодов максимально могло быть 2 в 5 степени – это 32.

Однако среди телеграфных кодов гораздо большее распространение получила всем известная азбука Морзе, созданная американским изобретателем Самюэлем Морзе в 1838 году.

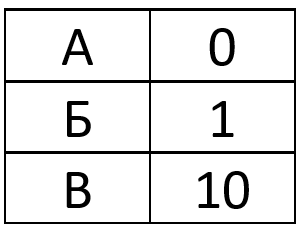


Азбука Морзе

Азбука Морзе служит примером **неравномерного двоичного кода**. В этом случае коды символов могут иметь различную длину. Как правило это обусловлено тем, что некоторые буквы встречаются в словах гораздо чаще, чем другие. И тогда целесообразно сделать их коды короче, тогда и суммарная длина сообщения получится меньше. Но возникает вопрос: как правильно декодировать полученное сообщение? Предположим мы получили сообщение "точка тире". Согласно таблице, так закодирована буква А. Но в то же время можно предположить, что в этом сообщении содержатся две буквы: Е и Т. Возникает неоднозначность при декодировании. Один и тот же код можно расшифровать несколькими способами. Чтобы такой проблемы не возникало, при передаче кода Морзе между символами необходимо делать паузы, тогда каждый символ можно будет декодировать однозначно. Но в каком-то смысле этот код становится не двоичным, а троичным, так как пауза – это тоже часть сообщения.

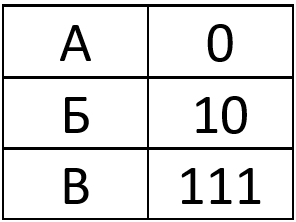
**Однозначное и неоднозначное кодирование**

Рассмотрим кодовую таблицу:



Декодируем сообщение 0100110. Первый символ в коде – это 0, это может быть только буква А, так как другие коды не начинаются с нуля. Но уже на втором символе возникает неоднозначность. То ли надо воспринимать 1 как букву Б, то ли 10 вместе как букву В. Оба варианта возможны. В итоге приходим к выводу, что декодировать данное сообщение можно несколькими способами: АБААББА, АВАБВ, АБААБВ. Такой код декодируется **неоднозначно**.

Рассмотрим другую кодовую таблицу:



Декодируем сообщение 100111. Здесь сразу видно, что первая буква кода – это Б, т.к. просто 1 не является кодом ни одной из букв алфавита, 100 также отсутствует в таблице. Остаётся только вариант 10. Затем 0 – это буква А и три единицы – это буква В. Получаем сообщение БАВ. Оно декодировано **однозначно**.

**Условие Фано**

От чего же зависит возможность однозначного декодирования сообщения при использовании неравномерного кода? В первом примере мы заметили, что если код буквы А – это 0, а код буквы В - это 01, то возникает неоднозначность, так как если 0 стоит рядом с 1, то мы можем декодировать их двумя способами. Чтобы такого не было, при использовании неравномерного кода мы должны соблюдать **условие Фано**: **ни одно кодовое слово не должно совпадать с началом другого кодового слова, то есть быть его частью.**

Очень удобно осуществлять проверку условия Фано с помощью дерева кодов. Будем откладывать влево ветки со значением 0, а вправо – со значением 1. Каждая ветка может разделяться ещё на 2. Когда ветка для того или иного кода получена, будем подписывать внизу букву, соответствующую данному коду.

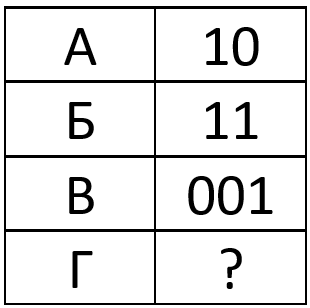


Видно, что ни одна ветка, которая уже занята буквой, не имеет продолжения. Это и означает, что условие Фано выполнено.

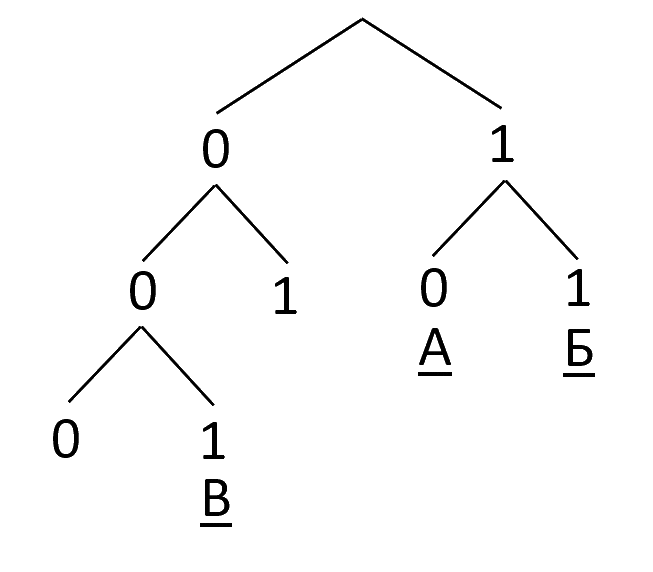
Рассмотрим задачу:

Символы алфавита закодированы неравномерным кодом. Буква А имеет код 10, буква Б - 11, буква В - 001. Какое кратчайшее кодовое слово может иметь буква Г?

Для начала запишем коды символов, которые нам даны.



 По этим кодам построим дерево.



Внимательно посмотрим, есть ли в этом дереве свободные ветви, то есть такие ветви, которые не заняты другими буквами. Видим, что это ветви, соответствующие кодам 000 и 01. Поскольку в условии задачи требуется найти **кратчайший** код, удовлетворяющий условию Фано, нам следует выбрать код **01**. Он и будет соответствовать букве Г.

Итак, в мы разобрали основные виды кодирования, научились подбирать длину кода и вычислять количество различных сообщений, работать с равномерными и неравномерными кодами.

# Алфавитный подход к измерению информации

**Алфавитом** называется конечное множество символов (это могут быть буквы, цифры и любые другие знаки) из которых состоит некоторая символьная последовательность, которую мы хотим закодировать и затем измерить её информационный объём.

Поскольку любой алфавит конечен, мы всегда можем сосчитать количество символов в нём. Это называется **мощностью алфавита**. Например, мощность алфавита, состоящего только из символов 0 и 1 равна двум. Мощность алфавита, состоящего из букв А, Б, В, Г, равна 4. Все десятичные цифры образуют алфавит с мощностью 10.

Для того, чтобы унифицировать принцип подсчёта количества информации, мы используем двоичное кодирование, то есть заменяем каждый символ нашего алфавита на определённую уникальную для каждого символа последовательность единиц и нулей. Например, для алфавита, состоящего только из символов «плюс» и «минус» можно использовать коды 0 и 1. В четырёхбуквенном алфавите коды будут длиннее: 00, 01, 10 и 11. А если мы будем кодировать десятичные цифры, то для каждой потребуется не менее четырёх двоичных знаков.

**-  0  
+ 1**

**А  00  
Б  01  
В  10  
Г  11**

**0  0000  
1  0001  
2  0010  
3  0011  
4  0100  
5  0101  
6  0110  
7  0111  
8  1000  
9  1001**

Именно количество двоичных разрядов в коде символа и определяет его информационный вес или количество информации. Один двоичный разряд принято называть словом **бит**, которое является сокращением от фразы «***bi***nary digi***t***», что переводится на русский как «двоичная цифра». Для того, чтобы измерить количество информации, необходимо сначала закодировать символ двоичным кодом, а затем подсчитать количество полученных нулей и единиц. 

### Формула Хартли

**2i=N**

N – мощность алфавита (количество различных символов),  
 i – количество информации (информационный вес одного символа из данного алфавита)

### Степени числа 2



**Пример 1.**Сколько бит необходимо для кодирования одного символа латинского алфавита?

{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y. Z}

N = 26

2i=26

Т.к. число 26 не является целой степенью числа 2, а количество бит, приходящееся на 1 символ, может быть только целым числом, мы подбираем ближайшую большую степень числа 2 к нашему N. Число 26 наиболее близко к числу 32, которое является степенью двойки.

i = 5 бит

**Пример 2.** Для кодирования одного символа некоторого алфавита выделено 3 бита информации. Сколько символов может быть в данном алфавите?

i = 3 бита

Т.к. мощность алфавита не обязательно является целой степенью числа 2, она должна быть строго больше, чем 2*i*−1, но меньше или равна 2*i*.

22 < N ≤23

4 < N ≤8

В нашем случае 3 битами может быть закодирован один символ алфавита мощностью от 5 до 8 символов.

Итак, мы рассмотрели алфавитный подход к измерению количества информации и научились определять мощность алфавита с помощью формулы Хартли.

# 3. Вероятностный подход к измерению информации

**Вероятность** – это мера возможности наступления некоторого события. Значение вероятности изменяется в диапазоне от нуля до единицы и является вещественным числом. Также можно выражать вероятность в процентах. Например, вероятность того, что после подбрасывания монеты выпадет решка – одна вторая или ноль целых пять десятых или 50%.

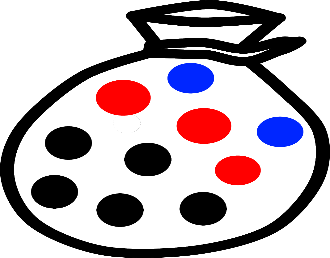
Для нахождения вероятности используется формула:

**p = m / n**

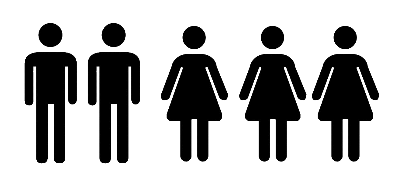
m - количество благоприятствующих исходов  
n - общее количество возможных исходов (m ≤*n*)

Вероятность не может быть больше единицы. Если вероятность равна 0, то событие никогда не наступает. Если вероятность равна 1, значит событие наступает всегда. При таких значениях вероятности говорят, что событие определено. Если же значение вероятности больше 0, но меньше 1, то в этом случае есть неопределенность.

**Пример 1.** Предположим, что в мешке лежат шарики разных цветов: 5 чёрных, 3 красных и 2 синих. Всего 10 шариков. Мы не глядя достаём один шарик. С какой вероятностью этот шарик окажется красного цвета?

 *p*кр​=3/10

**Пример 2.** В некоторой группе людей двое мужчин и три женщины. Мы случайным образом выбираем одного человека. Вероятность того, что выбранным человеком окажется женщина, выше, т.к. женщин в данной группе больше, чем мужчин.

 *p*м​=0,4  *p*ж​=0,6

События выбора женщины или мужчины **не равновероятны**, поэтому для нахождения количества информации нам потребуется другая формула.

Если нам известна вероятность наступления некоторого события, то количество информации, которое в нём содержится, может быть вычислено по **формуле Шеннона:**

**I = log21/р​**

Вернемся к примеру с мешком шариков. Найдём количество информации, содержащееся в одном из событий, например, что из мешка достали синий шарик. Для этого сначала вычислим вероятность наступления события, как отношение количества благоприятных исходов к общему числу исходов. Всего у нас 10 шариков, 2 из которых синих, значит вероятность равна 0,2. Теперь найдём обратную величину:  1/0,2 = 5. Наконец, найдём значение  **I = log25 = 2,32 бит**.

Когда мы имеем дело с полной системой событий, в которой одно из событий гарантированно наступит, но мы не знаем, какое именно, можно вычислить количество информации для всей системы событий в целом. Тогда формула Шеннона представляет собой сумму произведений вероятностей на двоичные логарифмы этих вероятностей:

**I = −∑*pi*​⋅*log*2​*pi*​**

Поскольку значения вероятностей всегда меньше единицы, двоичным логарифм будет отрицательным, а количество информации – это натуральное число, поэтому после вычисления суммы необходимо поменять её знак, тогда результат будет гарантированно положительным.

Вычислим количество информации для всей системы событий.

pчёрн = 0,5; pкр = 0,3; pсин = 0,2

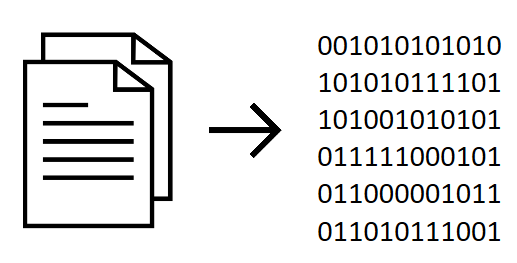
Подставим эти три значения в формулу:

I = −(0,5⋅*log*2​0,5+0,3⋅*log*2​0,3+0,2⋅*log*2​0,2)=1,49 бит

Итак, мы рассмотрели принцип вычисления количества информации в системе не равновероятных событий с помощью формулы Шеннона, научились находить вероятность наступления того или иного события.

1. **Единицы измерения информации**

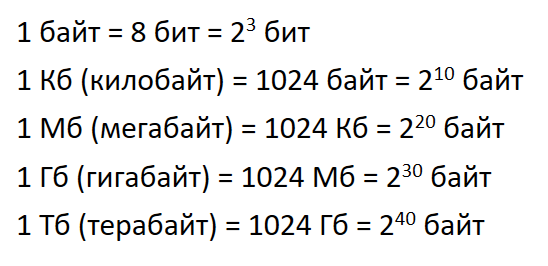
Чтобы измерить информацию, представленную в различных формах, её необходимо закодировать. Как мы уже знаем, вся информация в памяти компьютера представлена в двоичном виде и измерять мы будем количество двоичных знаков.



Мы уже познакомились с базовой единицей измерения количества информации – это **бит**(англ. bit - binary digit), то есть один двоичный разряд. Один бит – это то количество информации, которые мы получаем в результате осуществления одного из двух равновероятных событий. Бит может принимать только одно из двух устойчивых состояний, как правило они обозначаются цифрами 0 и 1.

С увеличением объёма информации нам понадобятся более крупные единицы измерения. Следующая по величине – **байт** (англ. byte - кусок). 1 байт вмещает в себя 8 бит. Исторически сложилось так, что в большинстве ЭВМ для адресации данных в памяти биты объединялись в октеты, то есть в группы из 8 разрядов. Также 8 – это два в третьей степени. Тем не менее, существовали компьютеры, в которых размер байта отличался и мог составлять 6, 7 или 9 бит, но с 70-ых годов прошлого века все приняли 8-битный размер байта в качестве стандарта.

С ростом объёма памяти компьютеров, емкости носителей информации, потребовались ещё более крупные единицы измерения: килобайты, мегабайты, гигабайты и другие.



Для перевода единиц измерения будем придерживаться правила: если в качестве исходных данных у нас меньшие единицы, скажем, биты или байты, и мы хотим перевести их в большие (мегабайты, гигабайты), то необходимо последовательно **разделить** на вес каждой из промежуточных единиц, а если наоборот из больших в меньшие – то последовательно **умножить** на вес каждой из промежуточных единиц.

**Например,**

4 Мбайта = 4×1024×1024×8=4×223бит=33554432 бита

122880 бит = 122880/1024×8​=15 Кбайт

Итак, мы рассмотрели основные единицы измерения информации и научились выражать количество информации в различных единицах.

**Вопросы для самоконтроля:**

1. Что такое знак? Приведите примеры знаков, используемых в общении людей.
2. Что такое знаковая система? Опишите десятичную систему счисления как знаковую систему.
3. Приведите примеры равномерных и неравномерных кодов, которые использовались ещё до изобретения компьютеров.
4. Почему нельзя применять неоднозначно декодируемые коды на практике?
5. Сформулируйте условие Фано.
6. В чём суть алфавитного подхода к измерению информации?
7. Что принято за минимальную единицу измерения информации?
8. Как определить информационный объём сообщения, представленного символами некоторого естественного или формального языка?
9. Что такое вероятность? Какие значения она может принимать?
10. Если в некоторой системе событий вероятность каждого из событий одинакова, то какой формулой лучше воспользоваться: Хартли или Шеннона?
11. Расшифруйте понятие "бит".
12. Отличается ли значение приставки "кило" применительно к метрам и к байтам? Почему?
13. Назовите объём памяти жесткого диска своего компьютера, объём памяти телефона, флеш-карты.