есть

число.

Пифаго

Кодирование и р декодирование информации

Теория

Кодирование и декодирование информации

Кодирование — это преобразование информации из одной ее формы представления в другую, наиболее удобную для её хранения, передачи или обработки.

Декодирование — процесс восстановления изначальной формы представления информации, т. е. обратный процесс кодирования, при котором закодированное сообщение переводится на язык, понятный получателю. В более широком плане это:

а) процесс придания определенного смысла полученным сигналам;

б) процесс выявления первоначального замысла, исходной идеи отправителя, понимания смысла его сообщения.

Алфавит

В основе каждого текста лежит **алфавит** – конечное множество символов. В основе русского языка лежит алфавит, называемый кириллицей, состоящий из 33 строчных и 33 заглавных букв. В основе английского языка лежит латиница – алфавит, состоящий из 26 строчных и 26 заглавных букв. Пусть задан алфавит T, содержащий m символов: $\mathbf{T=\{t_1, t_2, \dots t_m\}}$

Словом S в алфавите T называют любую последовательность символов алфавита:

$S = s_1 s_2 \dots s_k$,

где s_i- это символы алфавита. Число символов в слове – k называют **длиной слова**.

Мощность алфавита – это количество символов в нем.

Алфавит

При нажатии на клавиатурную клавишу компьютер получает сигнал в виде двоичного числа, р˙асшифровку которого можно найти в кодовой таблице – внутреннем представлении знаков в ПК. Стандартом во всем мире считают таблицу ASCII.

Для хранения одного символа двоичного кода электронно-вычислительная машина выделяет 1 байт, то есть 8 бит. Эта ячейка может принимать только два значения: 0 и 1. Получается, что один байт позволяет зашифровать 256 разных символов, ведь именно такое количество комбинаций можно составить. Эти сочетания и являются ключевой частью таблицы ASCII.

Алфавит

ASCII

UNICODE

Долгое время при работе с текстами, сохраняемыми в компьютере, используется код ASCII. Такой алфавит, содержащий 256 различных символов, мог включать латиницу и кириллицу, цифры, знаки операций, знаки препинания, скобки и другие символы. Но все-таки этого алфавита недостаточно, чтобы можно было хранить в памяти компьютера тексты на любых естественных языках.

Сегодня для хранения текстов используется кодировка из 2 байтов, называемая UNICODE кодировкой, позволяющая словами из 16 битов кодировать содержащий $2^{16}=65536$ символов. алфавит,

Неоднозначное кодирование

Пример.
Пусть у нас есть алфавит из 3-х символов – А, М, П.
Введем следующую кодировку: А-0, М-1, П-10.
Рассмотрим закодированный текст: **1010**.
Этому тексту соответствует два слова – МАМА и ПП.
Как видите, введенная кодировка не обеспечивает однозначное кодирование.

Если при кодирование выполняется условие Фано, то декодирование однозначно.

Условие Фано

Условие Фано: никакое кодовое слово не совпадает с началом другого кодового слова.

Коды, длякоторых выполняется условие Фано, называют **префиксными** (префикс слова — это его начальный фрагмент).

Все сообщения, закодированные с помощью префиксных кодов, декодируются однозначно.

Префиксные коды имеют важное практическое значение — они позволяют декодировать символы полученного сообщение по мере его получения, не дожидаясь, пока всё сообщение будет доставлено получателю.

Нужно знать

Прямое условие Фано

Неравномерный код может быть однозначно декодирован, если никакой из кодов не совпадает с началом (префиксом) какого-либо другого, более длинного кода.

A	B	C
10	11	001

D: 00

недопустимо:

C - 001

D – 00

Код D совпадает с началом кода C

A	B	C
10	11	00

D: 11

недопустимо:

B - 11

D – 11

Код D совпадает с кода B

A	B	C
100	110	010

D: 00

допустимо:

Прямое условие Фано выполнено.

Обратное условие Фано

Неравномерный код может быть однозначно декодирован, если никакой из кодов не совпадает с окончанием (постфиксом) какого-либо другого, более длинного кода.

A	B	C
10	11	001

D: 01
недопустимо:
C - 001
D – 01
Код D совпадает с концом кода C

A	B	C
10	11	00

D: 11
недопустимо:
B - 11
D – 11
Код D совпадает с кода B

A	B	C
100	110	010

D: 01
допустимо:
Обратное условие Фано выполнено.

Условия Фано

Для однозначности последовательности кодов достаточно хотя бы одного из двух выполнения следующих условий Фано:

- при выполнении прямого условия Фано последовательность кодов однозначно декодируется с начала;
- при выполнении обратного условия Фано последовательность кодов однозначно декодируется с конца.

Правило Фано – это достаточное, но необходимое условие однозначного декодирования.

Задача 3

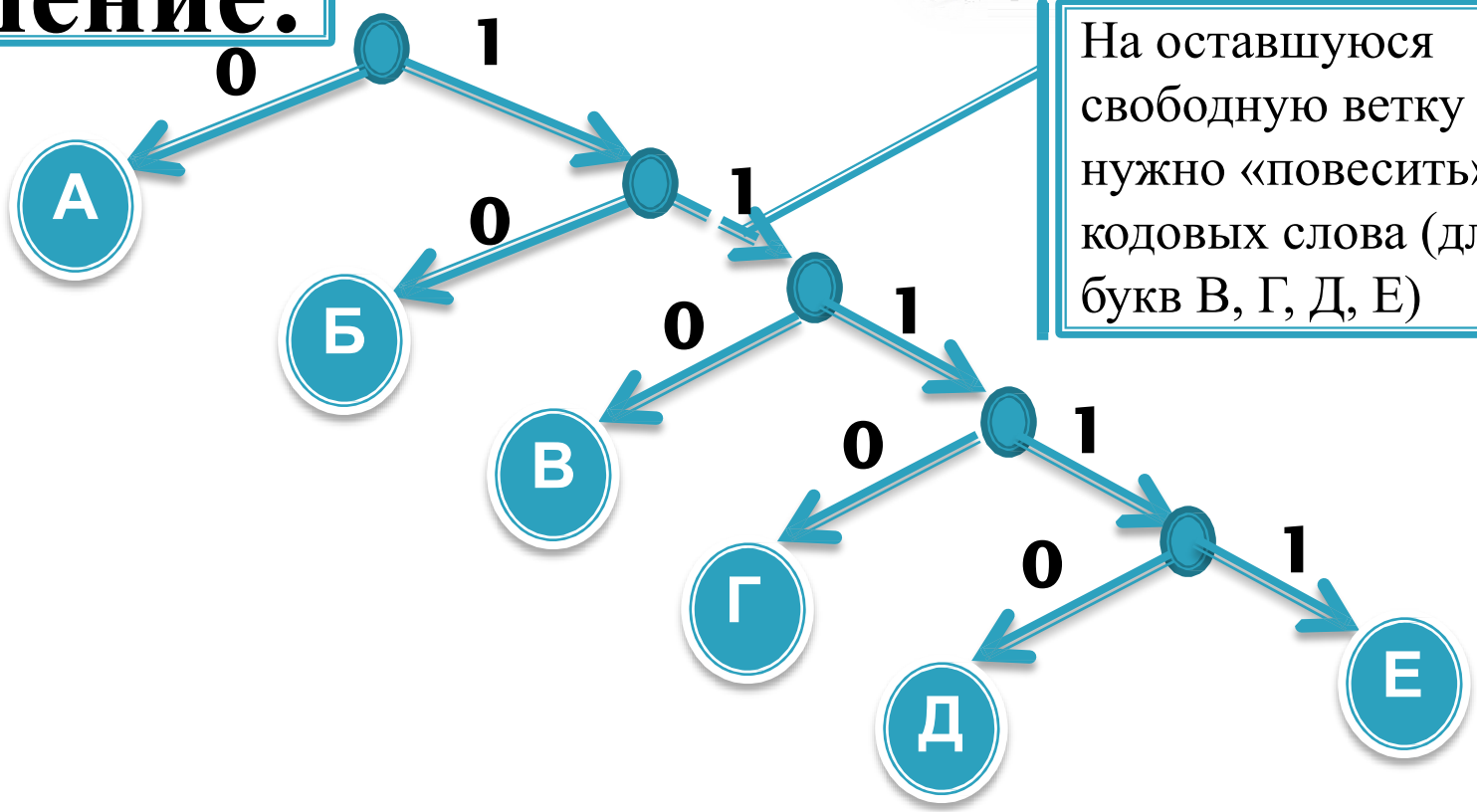
Для кодирования некоторой последовательности, состоящей из букв А, Б, В, Г, Д, Е, решили использовать неравномерный двоичный код, удовлетворяющий условию Фано. Для буквы А использовали кодовое слово 0; для буквы Б – кодовое слово 10. Какова наименьшая возможная сумма длин всех шести кодовых слов?

Это задание удобнее решать с помощью дерева: условие Фано выполняется тогда, когда все выбранные кодовые слова заканчиваются в листьях дерева.

Подсказка

Задача 3

Решение:

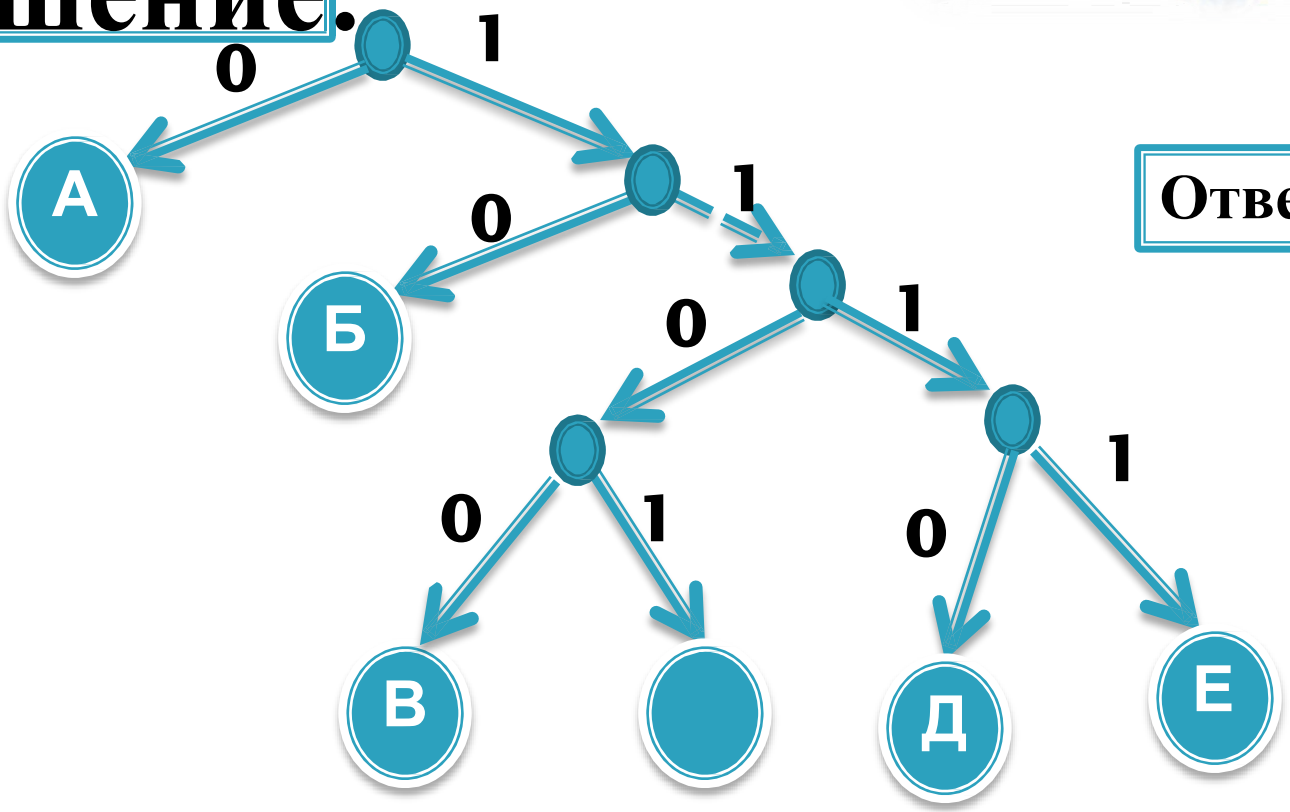


На оставшуюся свободную ветку нужно «повесить» 4 кодовых слова (для букв В, Г, Д, Е)

суммарная длина кодовых слов будет в этом случае равна **1 + 2 + 3 + 4 + 2·5 = 20**
(А-0, Б-10, В-110, Г-1110, Д-11110, Е-11111)

Задача 3

Решение:

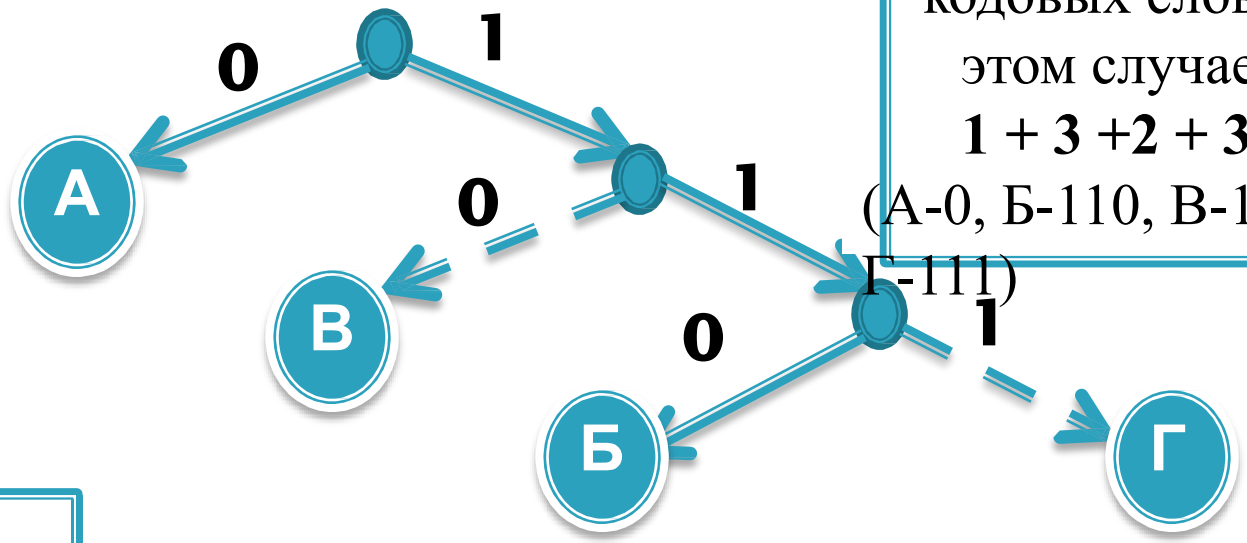


Ответ: 19

суммарная длина кодовых слов будет в этом случае равна **1 + 2 + 4·4 = 19**
(А-0, Б-10, В-1100, Г-1101, Д-1110, Е-1111)

Задача 4

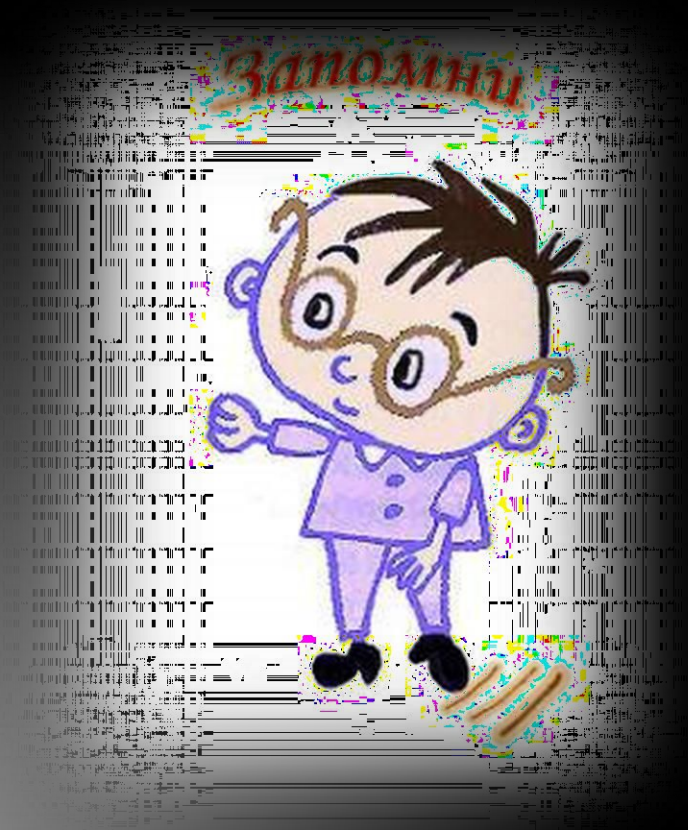
Для кодированиянекоторой последовательности, состоящей
•из букв А, Б, В, Г, решили использовать неравномерный двоичный код, удовлетворяющий условию Фано. Для буквы А использовали кодовое слово 0, для буквы Б – кодовое слово 110. Какова наименьшая возможная суммарная длина всех четырёх кодовых слов?



суммарная длина кодовых слов будет в этом случае равна
1 + 3 +2 + 3 = 9
(А-0, Б-110, В-10, Г-111)

Ответ: 9

Нужно помнить



Кодирование – это перевод информации с одного языка на другой (запись в другой системе символов, в другом алфавите).

Обычно кодированием называют перевод информации с «человеческого» языка на формальный, например, в

двоичный код, а декодированием – обратный переход.

Один символ исходного сообщения может заменяться одним символом нового кода или несколькими символами, а может быть и наоборот – несколько символов исходного сообщения заменяются одним символом в новом коде (китайские иероглифы обозначают целые слова и понятия).

Кодирование может быть *равномерное* и *неравномерное*

При равномерном кодировании все символы кодируются кодами равной длины.

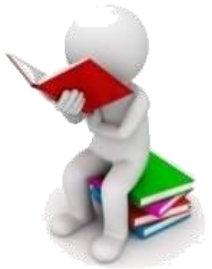
При неравномерном кодировании разные символы могут кодироваться кодами разной длины, это затрудняет декодирование.

Закодированное сообщение можно однозначно декодировать с начала, если выполняется *условие Фано*: никакое кодовое слово не является началом другого кодового слова;

закодированное сообщение можно однозначно декодировать с конца, если выполняется *обратное условие Фано*: никакое кодовое слово не является окончанием другого кодового слова.

Условие Фано – это достаточное, но не необходимое условие однозначного декодирования.

Задача 5

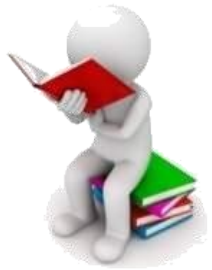


Для трехбуквенного алфавита {А,М, П} используется кодировка А-01, М-10, П-001. Какой код минимальной длины следует задать длякодировки буквы Т, добавляемой в алфавит?

Решение:

Для нового символа, добавляемого в алфавит, нельзя использовать код, состоящий из одного символа, так как будет нарушено условие Фано. Для кода, состоящего из двух символов, возможен только один вариант, удовлетворяющий условию Фано, Т-11.

Ответ: 11



Для четырехбуквенного алфавита {А, М, П, Т} используется кодировка А-01, М-10, П-001, Т-11. Можно ли уменьшить длину кода одного из символов, сохраняя однозначность декодирования?

Ответ: П-00

Задача 7

По каналу связи передаются сообщения, содержащие только 4 буквы: A, B, C, D. Для передачи используется двоичный код, допускающий однозначное декодирование. Для букв используются такие кодовые слова: A-111, B-0, D-110.

Укажите кратчайшее кодовое слово для буквы C, при котором код будет допускать однозначное декодирование. Если таких кодов несколько, укажите код с наименьшим числовым значением.

Решение:

Коды 1 и 0 являются началом кода данных букв.

Коды 00 и 01 нельзя использовать, так как код буквы B является их началом. Следовательно, минимальный код для буквы C будет 10.

Ответ: 10

Задача 8

Для передачи по каналу связи сообщения, состоящего только из символов А, Б, В и Г, используется неравномерный (по длине) код: А-100, Б-111, В-110, Г-0. Через канал связи передаётся сообщение: ВАБГАВ. Закодируйте сообщение данным кодом. Полученную двоичную последовательность переведите в шестнадцатеричный вид.

Решение:

Закодируем сообщение ВАБГАВ – 1101001110100110.

Полученную двоичную последовательность переведем в шестнадцатеричный вид.

1101|0011|1010|0110

D 3 A 6

Ответ: D3A6

Задача 9

По каналу связи передаются сообщения, содержащие только 3 буквы: A, B, C. Для передачи используется двоичный код, допускающий однозначное декодирование. Для букв A и B используются такие кодовые слова: A: 11, B: 0.

Укажите кратчайшее кодовое слово для буквы C, при котором код будет допускать однозначное декодирование. Если таких кодов несколько, укажите код с наименьшим числовым значением.

Решение:

Коды 1 и 0 являются началом кода данных букв.

Коды 00 и 01 нельзя использовать, так как код буквы B является их началом. Следовательно, минимальный код для буквы C будет 10.

Ответ: 10

Выполни самостоятельно

Задание 1. По каналу связи передаются сообщения, содержащие только 4 буквы: A, B, C, D; для передачи используется двоичный код, допускающий однозначное декодирование. Для букв A, B, D используются такие кодовые слова: A: 0, B: 10, D: 110. Укажите кратчайшее кодовое слово для буквы C, при котором код будет допускать однозначное декодирование. Если таких кодов несколько, укажите код с наименьшим числовым значением.

Ответ: 111

Задание 2. Для передачи по каналу связи сообщения, состоящего только из символов А, Б, В и Г, используется неравномерный (по длине) код: А-00, Б-11, В-100, Г-011. Через канал связи передаётся данное сообщение: ГГВАГВ. Закодируйте сообщение. Полученную двоичную последовательность переведите в шестнадцатеричный вид.

Ответ: 7C1C

Задание 3. Для передачи по каналу связи сообщения, состоящего только из символов А, Б, В и Г, используется неравномерный (по длине) код: А-00, Б-11, В-010, Г-011. Через канал связи передаётся сообщение: ГБВАВГ. Закодируйте сообщение данным кодом. Полученную двоичную последовательность запишите в восьмеричной системе счисления.

Ответ: 75023

Задание 4. Для передачи по каналу связи сообщения, состоящего только из символов А, Б, В и Г, используется неравномерный (по длине) код: А-111, Б-110, В-10, Г-0. Через канал связи передаётся сообщение: ВАБГАВ. Закодируйте сообщение данным кодом. Полученную двоичную последовательность запишите в восьмеричной системе счисления.

Ответ: 27636

Задание 5. По каналу связи передаются сообщения, содержащие только 3 буквы: A, B, C; для передачи используется двоичный код, допускающий однозначное декодирование. Для букв A и B используются такие кодовые слова: A: 10, B: 0. Укажите кратчайшее кодовое слово для буквы C, при котором код будет допускать однозначное декодирование. Если таких кодов несколько, укажите код с наименьшим числовым значением.

Ответ: 11

Задание 6. По каналу связи передаются сообщения, содержащие только 4 буквы: A, B, C, D; для передачи используется двоичный код, допускающий однозначное декодирование. Для букв A, B, D используются такие кодовые слова: A: 111, B: 0, D: 100. Укажите кратчайшее кодовое слово для буквы C, при котором код будет допускать однозначное декодирование. Если таких кодов несколько,