Информатика

ОЦЕНКА КОЛИЧЕСТВЕННЫХ ПАРАМЕТРОВ ТЕКСТОВЫХ ДОКУМЕНТОВ

ОБРАБОТКА ТЕКСТОВОЙ ИНФОРМАЦИИ

7 класс

Издательство БИНОМ

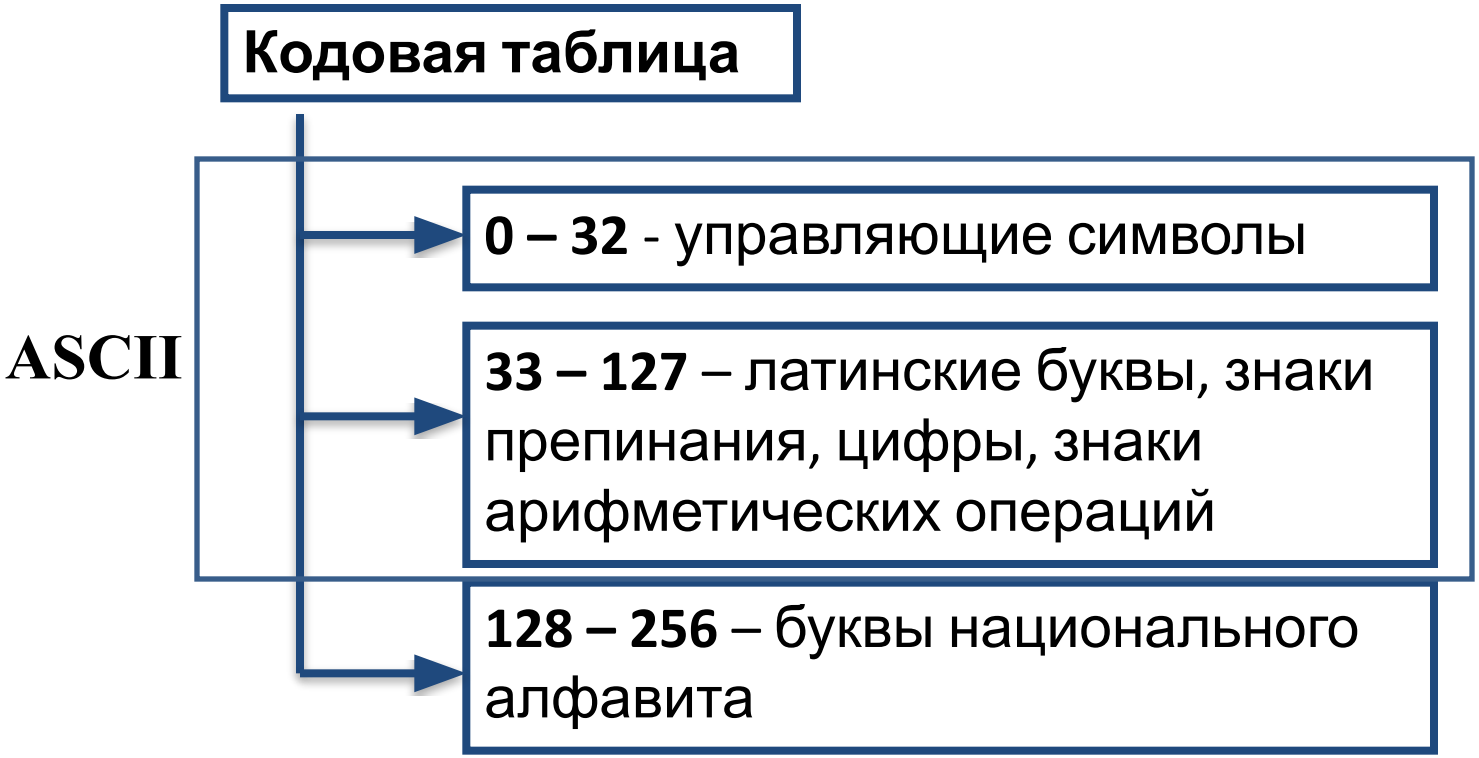
Ключевые слова

- кодовая таблица
- восьмиразрядный двоичный код
- информационный объём текста

Представление текстовой информации в памяти компьютера

Текст состоит из символов - букв, цифр, знаков препинания и т. д., которые компьютер различает по их **двоичному коду**.

Соответствие между изображениями символов и кодами символов устанавливается с помощью **кодовых таблиц**.



Представление текстовой информации в памяти компьютера

Фрагмент кодовой таблицы ASCII

Коды русских букв в разных кодировках

Символ	Кодировка			
	Windows		КОИ-8	
	десятичный код	двоичный код	десятичный код	двоичный код
А	192	11000000	225	11100001
Б	193	11000001	226	11100010
В	194	11000010	247	11110111
,	44	00101100	6	54	00110110
-	45	00101101	7	55	00110111
.	46	00101110	8	56	00111000
/	47	00101111	9	57	00111001
A	65	01000001	**N**	78	01001110
B	66	01000010	**O**	79	01001111
C	67	01000011	**P**	80	01010000

Стандарт кодирования символов Unicode позволяет пользоваться более чем двумя языками.

В Unicode каждый символ кодируется шестнадцатиразрядным двоичным кодом. Такое количество разрядов позволяет закодировать 65 536 различных символов: $2^{16} = 65\,536$.

Информационный объём фрагмента текста

I - информационный объём сообщения

K – количество символов

i – информационный вес символа

$$I = K \times i$$

В зависимости от разрядности используемой кодировки информационный вес символа текста, создаваемого на компьютере, может быть равен:

- 8 битов (1 байт) - **восьмиразрядная кодировка**;
- 6 битов (2 байта) - **шестнадцатиразрядная кодировка**.

Информационный объём фрагмента текста - это количество битов, байтов (килобайтов, мегабайтов), необходимых для записи фрагмента оговорённым способом кодирования.

Информационный объём фрагмента текста

Задача 1. Считая, что каждый символ кодируется одним байтом, определите, чему равен информационный объём следующего высказывания Жан-Жака Руссо:

Тысячи путей ведут к заблуждению, к истине - только один.

Решение

В данном тексте 57 символов (с учётом знаков препинания и пробелов). Каждый символ кодируется одним байтом. Следовательно, информационный объём всего текста - 57 байтов.

Ответ: 57 байтов.

Информационный объём фрагмента текста

Задача 2. В кодировке Unicode на каждый символ отводится два байта. Определите информационный объём слова из 24 символов в этой кодировке.

Решение.

$I = 24 \times 2 = 48$ (байтов).

Ответ: 48 байтов.

Информационный объём фрагмента текста

Задача 3. Автоматическое устройство осуществило перекодировку информационного сообщения на русском языке, первоначально записанного в 8-битовом коде, в 16-битовую кодировку **Unicode**. При этом информационное сообщение увеличилось на 2048 байтов. Каков был информационный объём сообщения до перекодировки?

Решение

Информационный вес каждого символа в 16-битовой кодировке в два раза больше информационного веса символа в 8-битовой кодировке. Поэтому при перекодировании исходного блока информации из 8-битовой кодировки в 16-битовую его информационный объём должен был увеличиться вдвое, другими словами, на величину, равную исходному информационному объёму. Следовательно, информационный объём сообщения до перекодировки составлял 2048 байтов = 2 Кб.

Информационный объём фрагмента текста

Задача 4. Выразите в мегабайтах объём текстовой информации в «Современном словаре иностранных слов» из 740 страниц, если на одной странице размещается в среднем 60 строк по 80 символов (включая пробелы). Считайте, что при записи использовался алфавит мощностью 256 символов.

Решение

$K = 740 \times 80 \times 60$

$N = 256$

I - ?

$I = K \times i$

$N = 2^i$

$256 = 2^i = 2^8$, $i = 8$

К = 740 × 80 × 60 × 8 = 28 416 000 бит = 3 552 000 байтов =

= 3 468,75 Кбайт $\approx$ 3,39 Мбайт.

Ответ: 3,39 Мбайт.

Самое главное

Текст состоит из символов - букв, цифр, знаков препинания и т. д., которые человек различает по начертанию. Компьютер различает вводимые символы по их двоичному коду. Соответствие между изображениями и кодами символов устанавливается с помощью ***кодовых таблиц***.

В зависимости от разрядности используемой кодировки информационный вес символа текста, создаваемого на компьютере, может быть равен:

- 8 битов (1 байт) - ***восьмиразрядная кодировка***;
- 6 битов (2 байта) - ***шестнадцатиразрядная кодировка***.

Информационный объём фрагмента текста - это количество битов, байтов (килобайтов, мегабайтов), необходимых для записи фрагмента оговорённым способом кодирования.

Вопросы и задания

1. Почему кодировки, в которых каждый символ кодируется цепочкой из восьми нулей и единиц, называются иначе однобайтовыми?

2. С какой целью была введена кодировка Unicode?

Вопросы и задания

3. Считая, что каждый символ кодируется одним байтом, определите, чему равен информационный объём следующего высказывания Алексея Толстого:

Не ошибается тот, кто ничего не делает, хотя это и есть его основная ошибка.

1) 512 битов
2) 608 битов
3) 8 Кбайт
4) 123 байта

Вопросы и задания

4.В кодировке ASCII каждый символ кодируется 8 битами. Определите информационный объём сообщения в этой кодировке:

Длина данного текста 32 символа.

1) 32 бита
2) 320 битов
3) 32 байта
4) 256 байтов

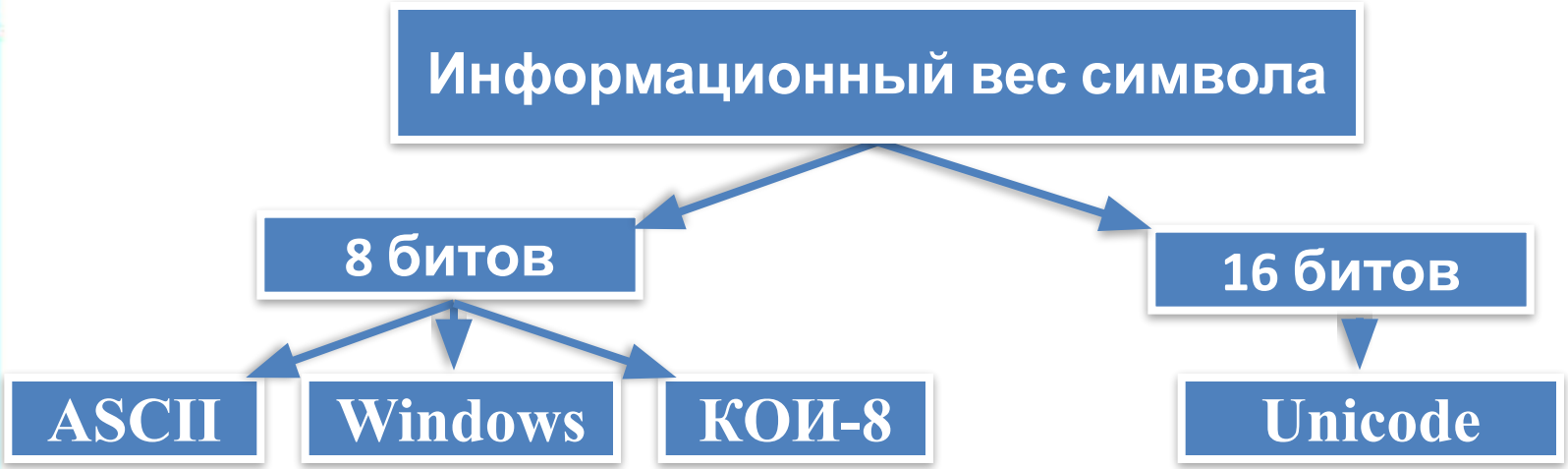
Вопросы и задания

5. В какой кодировочной таблице можно закодировать 65 536 различных символов?

1) ASCII
2) Windows
3) КОИ-8
4) Unicode

Опорный конспект

Компьютер различает вводимые символы по их двоичному коду. Соответствие между изображениями и кодами символов устанавливается с помощью **кодовых таблиц**.



$$I = K \times i$$

I - информационный объём сообщения

K – количество символов

i – информационный вес символа