Двоичное кодирование текстовой информации



Презентацию подготовила Машкина Татьяна Анатольевна, учитель информатики

МБОУ «СОШ №92»

Вопросы для повторения

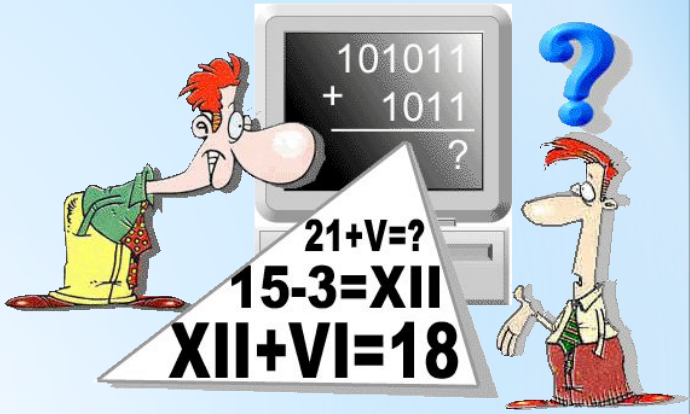
* Какие виды информации по способу представления вы знаете?
* Кодирование каких видов информации вы изучили на прошлых уроках?
* Что принято за единицу количества информации?
* Назовите производные единицы измерения количества информации.

В *40*-е годы прошлого столетия было положено начало созданию вычислительной машины.

Начиная с *60*-х годов, компьютеры все больше стали использовать для **обработки текстовой информации** и в настоящее время большая часть ПК в мире занято обработкой именно текстовой информации.

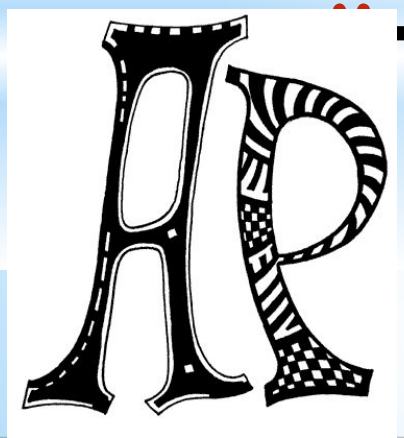


*Двоичное кодирование текстовой



+|| ++|||| +++|||

12 24 33

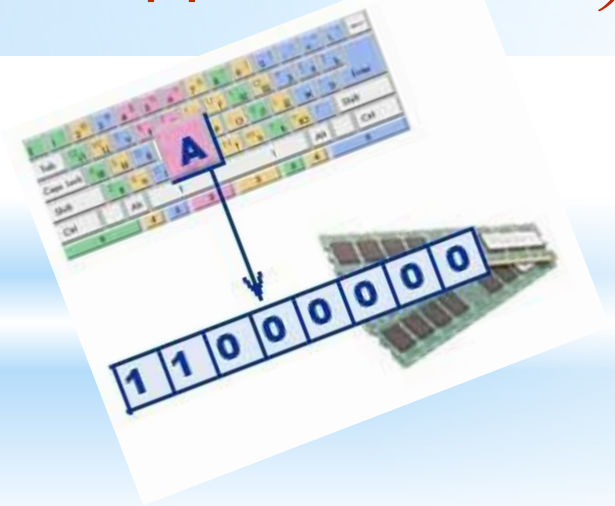


Таинственные знаки

* В памяти компьютера любой текст представляется последовательностью кодов символов, т. е. вместо самой буквы хранится ее номер в кодовой таблице. Изображение же букв и символов сформируется только в момент их вывода на экран или бумагу.

*Двоичное кодирование текстовой

Кодирование заключается в том, что каждому символу ставиться в соответствие уникальный двоичный код от *00000000* до *11111111* (или десятичный код от *0* до *255*).



*Двоичное кодирование текстовой

Важно, что присвоение символу конкретного кода – это вопрос соглашения, которое фиксируется кодовой таблицей.

Для разных типов ЭВМ используются различные кодировки

*Кодовые таблицы

Кодовые таблицы

* По началу применялось *7*-битная кодировка, которая могла представить *128* символов. С распространением *IBM PC* международным стандартом стала таблица кодировки *ASCII (American Standart Code for Information Interchange)* – Американский стандартный код для информационного обмена.

Позже она была расширена до *8* бит *(256* символов*)* и дошла в таком виде практически до сегодняшнего дня. При этом первая половина *(*символы *0-127)* были всегда одни и те же, соответствующие стандарту *ASCII*, а вторая половина таблицы *(*символы *128-255)* менялась в зависимости от страны, где она использовалась.

***Таблица кодировки**

ASCII

символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код
	32	00100000	8	56	00111000	P	80	01010000	h	104	01101000
!	33	00100001	9	57	00111001	Q	81	01010001	i	105	01101001
"	34	00100010	:	58	00111010	R	82	01010010	j	106	01101010
#	35	00100011	;	59	00111011	S	83	01010011	k	107	01101011
$	36	00100100	<	60	00111100	T	84	01010100	l	108	01101100
%	37	00100101	=	61	00111101	U	85	01010101	m	109	01101101
&	38	00100110	>	62	00111110	V	86	01010110	n	110	01101110
'	39	00100111	?	63	00111111	W	87	01010111	o	111	01101111
(	40	00101000	@	64	01000000	X	88	01011000	p	112	01110000
)	41	00101001	A	65	01000001	Y	89	01011001	q	113	01110001
*	42	00101010	B	66	01000010	Z	90	01011010	r	114	01110010
+	43	00101011	C	67	01000011	[	91	01011011	s	115	01110011
,	44	00101100	D	68	01000100	\	92	01011100	t	116	01110100
-	45	00101101	E	69	01000101	]	93	01011101	u	117	01110101
.	46	00101110	F	70	01000110	^	94	01011110	v	118	01110110
/	47	00101111	G	71	01000111	_	95	01011111	w	119	01110111
0	48	00110000	H	72	01001000	`	96	01100000	x	120	01111000
1	49	00110001	I	73	01001001	a	97	01100001	y	121	01111001
2	50	00110010	J	74	01001010	b	98	01100010	z	122	01111010
3	51	00110011	K	75	01001011	c	99	01100011	{	123	01111011
4	52	00110100	L	76	01001100	d	100	01100100	|	124	01111100
5	53	00110101	M	77	01001101	e	101	01100101	}	125	01111101
6	54	00110110	N	78	01001110	f	102	01100110	~	126	01111110
7	55	00110111	O	79	01001111	g	103	01100111	□	127	01111111

символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код
Ђ	128	10000000		160	10100000	А	192	11000000	а	224	11100000
Ѓ	129	10000001	Ў	161	10100001	Б	193	11000001	б	225	11100001
‚	130	10000010	ў	162	10100010	В	194	11000010	в	226	11100010
ѓ	131	10000011	J	163	10100011	Г	195	11000011	г	227	11100011
„	132	10000100	¤	164	10100100	Д	196	11000100	д	228	11100100
…	133	10000101	Ґ	165	10100101	Е	197	11000101	е	229	11100101
†	134	10000110	¦	166	10100110	Ж	198	11000110	ж	230	11100110
‡	135	10000111	§	167	10100111	З	199	11000111	з	231	11100111
€	136	10001000	Ё	168	10101000	И	200	11001000	и	232	11101000
‰	137	10001001	©	169	10101001	Й	201	11001001	й	233	11101001
Љ	138	10001010	Є	170	10101010	К	202	11001010	к	234	11101010
‹	139	10001011	«	171	10101011	Л	203	11001011	л	235	11101011
Њ	140	10001100	¬	172	10101100	М	204	11001100	м	236	11101100
Ќ	141	10001101	-	173	10101101	Н	205	11001101	н	237	11101101
Ћ	142	10001110	®	174	10101110	О	206	11001110	о	238	11101110
Џ	143	10001111	Ї	175	10101111	П	207	11001111	п	239	11101111
ђ	144	10010000	°	176	10110000	Р	208	11010000	р	240	11110000
‘	145	10010001	±	177	10110001	С	209	11010001	с	241	11110001
’	146	10010010	I	178	10110010	Т	210	11010010	т	242	11110010
“	147	10010011	і	179	10110011	У	211	11010011	у	243	11110011
”	148	10010100	ґ	180	10110100	Ф	212	11010100	ф	244	11110100
•	149	10010101	µ	181	10110101	Х	213	11010101	х	245	11110101
–	150	10010110	¶	182	10110110	Ц	214	11010110	ц	246	11110110
—	151	10010111	·	183	10110111	Ч	215	11010111	ч	247	11110111
□	152	10011000	ё	184	10111000	Ш	216	11011000	ш	248	11111000
™	153	10011001	№	185	10111001	Щ	217	11011001	щ	249	11111001
љ	154	10011010	є	186	10111010	Ъ	218	11011010	ъ	250	11111010
›	155	10011011	»	187	10111011	Ы	219	11011011	ы	251	11111011
њ	156	10011100	j	188	10111100	Ь	220	11011100	ь	252	11111100
ќ	157	10011101	S	189	10111101	Э	221	11011101	э	253	11111101
ћ	158	10011110	s	190	10111110	Ю	222	11011110	ю	254	11111110
џ	159	10011111	ї	191	10111111	Я	223	11011111	я	255	11111111

* Таблица расширенного кода *ASCII* Кодировка *Windows-1251 (CP1251)*

Двоичное кодирование текстовой информации в компьютере

Для представления текстовой информации достаточно 256 различных символов.

$N = 2^I$, $256 = 2^I$, $2^8 = 2^I$,I = 8 битов

Для кодирования каждого знака требуется количество информации, равное 8 битам.

Для представления текста в памяти компьютера необходимо представить его в двоичной знаковой системе.

Каждому знаку необходимо поставить в соответствие уникальный двоичный код в интервале от 00000000 до 11111111 (в десятичном коде от 0 до 255)

Для представления символов и соответствующих им кодов используется *кодовая таблица*.

В качестве стандарта во всем мире принята таблица ASCII (American Standard Code for Information Interchange – Американский стандартный код для обмена информацией). Условно таблица разделена на части:

* от 0 до 32 коды соответствуют операциям;
* с 33 по 127 соответствуют символам латинского алфавита, цифрам, знакам арифметических операций и знакам препинания;
* со 128 по 255 являются национальными.

Кодовая таблица

32		00100000	56	8	00111000	80	P	01010000	104	h	01101000
33	!	00100001	57	9	00111001	81	Q	01010001	105	i	01101001
34	“	00100010	58	:	00111010	82	R	01010010	106	j	01101010
35	#	00100011	59	;	00111011	83	S	01010011	107	k	01101011
36	$	00100100	60	<	00111100	84	T	01010100	108	l	01101100
37	%	00100101	61	=	00111101	85	U	01010101	109	m	01101101
38	&	00100110	62	>	00111110	86	V	01010110	110	n	01101110
39	‘	00100111	63	?	00111111	87	W	01010111	111	o	01101111
40	(	00101000	64	@	01000000	88	X	01011000	112	p	01110000
41	)	00101001	65	A	01000001	89	Y	01011001	113	q	01110001
42	*	00101010	66	B	01000010	90	Z	01011010	114	r	01110010
43	+	00101011	67	C	01000011	91	[	01011011	115	s	01110011
44	,	00101100	68	D	01000100	92	\	01011100	116	t	01110100
45	-	00101101	69	E	01000101	93	]	01011101	117	u	01110101
46	.	00101110	70	F	01000110	94	^	01011110	118	v	01110110
47	/	00101111	71	G	01000111	95	_	01011111	119	w	01110111
48	0	00110000	72	H	01001000	96	‘	01100000	120	x	01111000
49	1	00110001	73	I	01001001	97	a	01100001	121	y	01111001
50	2	00110010	74	J	01001010	98	b	01100010	122	z	01111010
51	3	00110011	75	K	01001011	99	c	01100011	123	{	01111011
52	4	00110100	76	L	01001100	100	d	01100100	124	|	01111100
53	5	00110101	77	M	01001101	101	e	01100101	125	}	01111101
54	6	00110110	78	N	01001110	102	f	01100110	126	~	01111110
55	7	00110111	79	O	01001111	103	g	01100111	127	□	11111111

В кодовой таблице ASCII латинские буквы (прописные и строчные) располагаются в алфавитном порядке. Расположение цифр также упорядочено по возрастанию значений.

Данное правило соблюдается и в других таблицах кодировки. Благодаря этому и в машинном представлении для символьной информации сохраняется понятие «алфавитный порядок».

*Принцип последовательного кодирования алфавита:

*Различные кодировки знаков

В настоящее время существуют пять различных кодировок для русских букв

(Windows, MS-DOS, КОИ-8, Mac, ISO), поэтому тексты созданные в одной кодировке не будут правильно отображаться в другой.

В последние годы широкое распространение получил новый международный стандарт кодирования текстовых символов Unicode, который отводит на каждый символ 2 байта (16 битов), что позволило закодировать многие алфавиты в одной таблице.

$N = 2^I$, $N = 2^{16}$, $N = 65536$

*Пример 1

Буква «i» в таблице кодировки символов имеет десятичный код 105. Что зашифровано последовательностью десятичных кодов:

108 105 110 107?

При расшифровке данной последовательности кодов не нужно обращаться к таблице кодировки символов.

Необходимо учесть принцип последовательного кодирования алфавитов и вспомнить порядок букв в латинском алфавите (…i, j, k, l, m, n, o, …)

Закодировано: «link»

*Пример 2

С помощью последовательности десятичных кодов:

99 111 109 112 117 116 101 114 закодировано слово «computer». Какая последовательность десятичных кодов будет соответствовать этому же слову, записанному заглавными буквами?

При шифровке слова не пользоваться таблицей кодировки символов. Необходимо учесть, что разница между десятичным кодом строчной буквы латинского алфавита и десятичным кодом соответствующей заглавной буквы равна 32.

Если «c» имеет код 99, то «C» имеет код 99-32=67.

COMPUTER – 67 79 77 80 85 84 69 82

* Какое количество информации необходимо для кодирования одного символа компьютерного алфавита?
* Слова «компьютер»?
* Слова «алфавит»?

Пример 3

*Тестовая проверочная работа

Вариант 1

1. Полный набор символов, используемый для кодирования, называют:
   1) шифром 2) алфавитом 3) синтаксисом 4) семантикой
2. Минимальным объектом, используемым для кодирования текста, является:
   1)бит 2) пиксель 3) символ 4) растр
3. Количество битов, необходимое для кодирования одного символа алфавита в коде Unicode, равно:
   1) 8 2) 16 3) 32 4) 256
4. Сообщение «урок» содержит следующее количество информации:
   1) 4 бита 2) 32 бита 3) 8 байт 4) 32 байта
5. Какие символы могут быть зашифрованы кодами таблицы ASCII 119 и 251?
   1) «д» и «ш» 2) «j» и «s» 3) «d» и «D» 4) «w» и «ы»

Вариант 2

1. Таблица кодировки символов устанавливает соответствие между:
   1) символами, их десятичными номерами и двоичными кодами
   2) символами разных алфавитов
   3) символами и количеством байтов, которые они занимают
   4) символами и номерами ячеек памяти, которые они занимают
2. Минимальным объектом, используемым для кодирования текста, является:
   1) бит 2) пиксель 3) символ 4) растр
3. Количество битов, необходимое для кодирования одного символа алфавита в коде ASCII, равно:
   1) 1 2) 8 3) 32 4) 256
4. В сообщение «класс!» содержится следующее количество информации:
   1) 5 битов 2) 6 битов 3) 5 байтов 4) 6 байтов
5. Код буквы «О» в некоторой таблице кодировки символов равен 142. Что зашифровано с помощью последовательности кодов: 145 143 142 144 146?
   1) CPORT 2) SPORT 3) СПОРТ 4) SNORT

Ответы для проверки

* Вариант 1

1	2	3	4	5
2	3	2	2	4

* Вариант 2

1	2	3	4	5
1	2	2	4	3

1. Повторить изученный материал, воспользовавшись конспектом лекции и учебником стр. 49- 52, ответить на контрольные вопросы.
2. Выполнить задания 2.1, 2.2 по учебнику
3. В таблице ниже представлена часть кодовой таблицы ASCII:

Символ	1	5	A	B	Q	a	b
Десятичный код	49	53	65	66	81	97	98
16-ричный код	31	35	41	42	51	61	62

Определить 16-ричный код символа “q”.

* Домашнее задание

* В Советском Союзе различные организации и сети, имевшие большое влияние на компьютерный и программный рынок тех времен, создавали свои кодировки (т.е. вторые половины таблицы), содержащие русские символы.

* Кодовые таблицы

В настоящее время существует 5 разных кодовых таблиц для русских букв (КОИ8, **СР**1251, СР866, Mac, ISO).

Широкое распространение получил новый международный стандарт Unicode, который отводит на каждый символ два байта. С его помощью можно закодировать 65536 ($2^{16} = 65536$) различных символов.



Кодовые таблицы для русских букв

* Проблемы с кодировками делятся на несколько типов. Первый тип - это отсутствие информации о кодировке.

*Проблемы с кодировками

* Проблемы второго типа - это когда кодировка в файле указана, но конечная программа такой кодировки не знает.

* Проблемы с кодировками

* Третий тип проблем, наоборот, связан с избытком информации о кодировках. Это актуальная в настоящее время проблема (например, для веб-страниц).

*Проблемы с кодировками

Цифры кодируются по стандарту *ASCII* в двух случаях – при вводе-выводе и когда они встречаются в тексте. Если цифры участвуют в вычислениях, то осуществляется их преобразование в другой двоичных код.



***Обратите внимание!**

Возьмем число *57.*

При использовании в тексте каждая цифра будет представлена своим кодом в соответствии с таблицей *ASCII.* В двоичной системе это – *00110101 00110111.*

При использовании в вычислениях, код этого числа будет получен по правилам перевода в двоичную систему и получим – *00111001.*

***Обратите внимание!**

Информационный объем текста

Сегодня очень многие люди для подготовки писем, документов, статей, книг и пр. используют **компьютерные текстовые редакторы**. Компьютерные редакторы, в основном, работают **с алфавитом размером** *256* **символов**.

В этом случае легко подсчитать объем информации в тексте. Если *1* **символ алфавита несет** *1* **байт информации**, то надо просто сосчитать количество символов; полученное число даст информационный объем текста в байтах.

Формулы для расчета информационного объема текста

$I = K \times i$, где

I-информационный объем сообщения

K- количество символов в тексте

i- информационный вес одного символа

$$2^i = N$$

*N- мощность алфавита

Задание *1*

* Мощность алфавита равна *256*. Сколько Кбайт памяти потребуется для сохранения *160* страниц текста, содержащего в среднем *192* символа на каждой странице*?*

* Мощность алфавита равна *64*. Сколько Кбайт памяти потребуется, чтобы сохранить *128* страниц текста, содержащего в среднем *256* символов на каждой странице*?*

Задание *1’

* Объем сообщения – *7,5* Кбайт. Известно, что данное сообщение содержит *7680* символов. Какова мощность алфавита*?*

*Задание *2*

* Объем сообщения равен *11* Кбайт. Сообщение содержит *11264* символа. Какова мощность алфавита*?*

***Задание** *2’*

Племя Мумбу-Юмбу использует алфавит из букв: α β γ δ ε ζ η θ λ μ ξ σ φ ψ, точки и для разделения слов используется пробел.

Сколько информации несет свод законов племени, если в нем *12* строк и в каждой строке по *20* символов*?*

*Задание *3*

* Для кодирования секретного сообщения используются *12* специальных значков-символов. При этом символы кодируются одним и тем же минимально возможным количеством бит. Чему равен информационный объем сообщения длиной в *256* символов?

Задание *3’*

Вопросы и задания:

* В чем заключается кодирование текстовой информации в компьютере?
* Закодируйте с помощью *ASCII*-кода свою фамилию, имя, номер класса.
* Считая, что каждый символ кодируется одним байтом, оцените информационный объем следующего предложения из пушкинского четверостишия:

Певец-Давид был ростом мал, Но повалил же Голиафа!

*Используемая литература

- Учебник Угринович Н.Д. Информатика и ИКТбазовый курс *9* класс*;*
- Яндекс-картинка
- Изображение 2
- Изображение 3
- Изображение 4
- Изображение 5
- Изображение 6
- http://inn.h1.ru/topic.shtml?h1=16&h2=7
- http://www.galaktionoff.ru/unpub/TTF.htm
- http://www.infospir.ru/articles/chto_takoe_kodirovka_2.php
- http://gym1.pupils.ru/img_school/gym1/Ekzamen10/Variant1.pdf