Информатика

КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В КОМПЬЮТЕРЕ

10 класс

ИЗДАТЕЛЬСТВО БИНОМ

Компьютерное представление текстовой информации

Для компьютерного представления текстовой информации достаточно:

...	...
64	01000000
65	01000001
66	01000010
67	01000011
68	01000100

Определить алфавит (множество всех символов)

Присвоить каждому символу алфавита порядковый номер

Перевести номер символа в двоичную систему счисления

Кодировка ASCII

American Standard Code for Information Interchange – американский стандартный код для обмена информацией, разработанный в 1960-х годах в США.

Стандарт Unicode

Unicode — это «уникальный код для любого символа, независимо от платформы, независимо от программы, независимо от языка» (www.unicode.org).

Стандарт Unicode был разработан в 1991 году и описывает алфавиты всех известных, в том числе и «мертвых», языков. Для языков, имеющих несколько алфавитов или вариантов написания (японского и индийского), закодированы все варианты.
В кодировку Unicode внесены все математические и иные научные символьные обозначения и даже некоторые придуманные языки (язык эльфов из трилогии Дж. Р. Р. Толкина «Властелин колец»).

Клавиатуры некоторых стран мира

РУССКАЯ

АМЕРИКАНСКАЯ

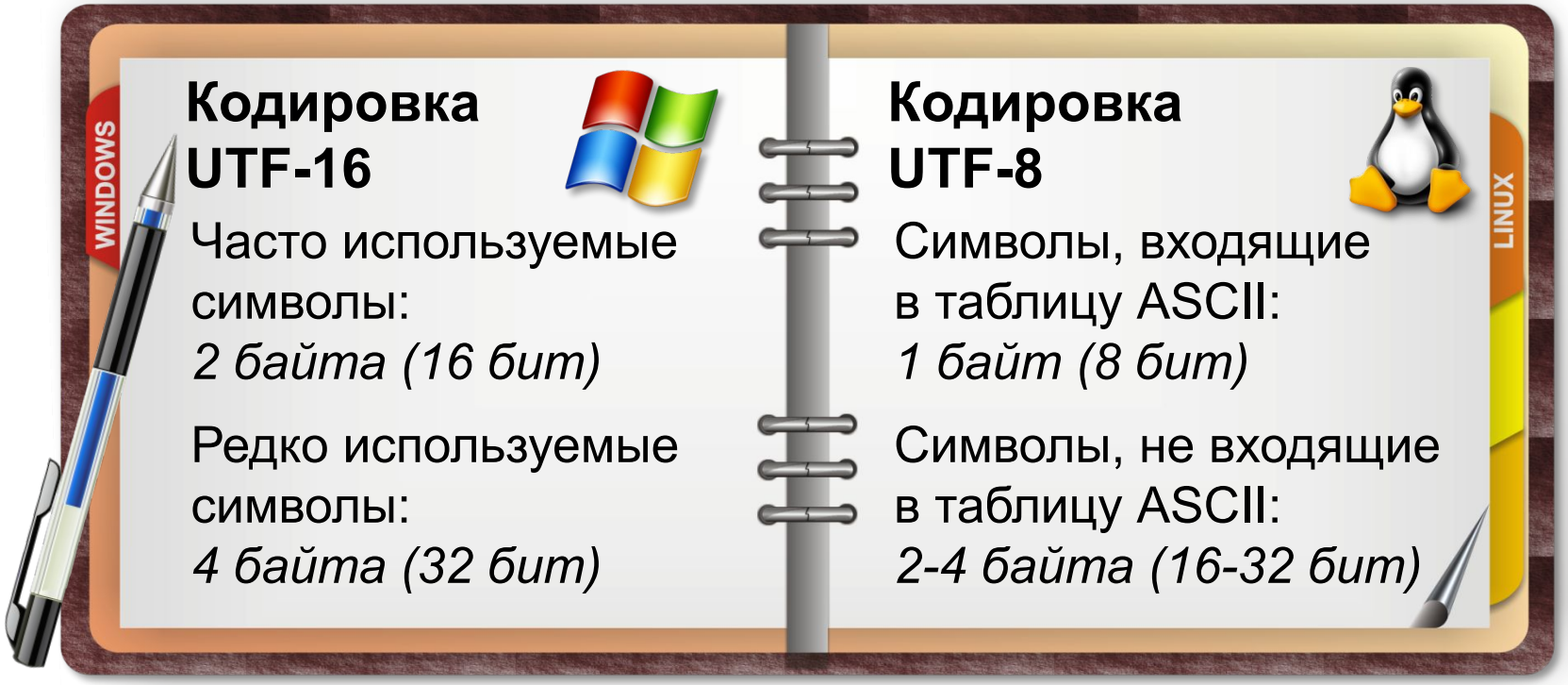
АРАБСКАЯ

АРМЯНСКАЯ

ЯПОНСКАЯ

Кодировки стандарта Unicode

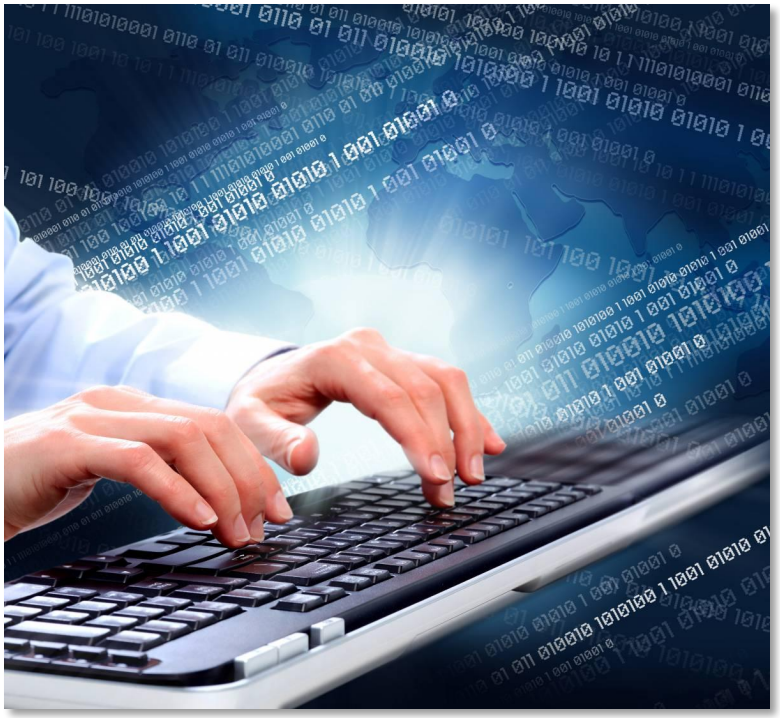
Для представления символов в памяти компьютера в стандарте Unicode имеется несколько кодировок.



Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

Информационный объем сообщения

Информационным объёмом текстового сообщения называется количество бит (байт, килобайт, мегабайт и т. д.), необходимых для записи этого сообщения путём заранее оговоренного способа двоичного кодирования.



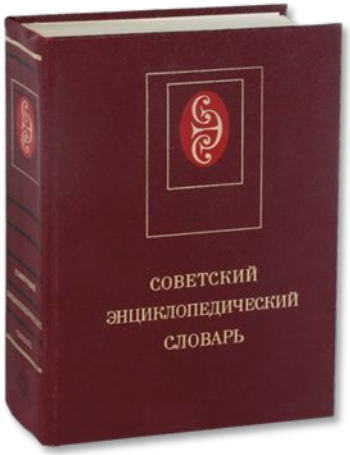
Количество символов в сообщении

$$I = K \cdot i$$

ASCII, КОИ-8, Windows-1251, …
1 символ = 1 байт

Unicode
1 символ = 2 байта

Вопросы и задания



В Советском энциклопедическом словаре (1983 года издания) 1600 страниц. На одной странице размещается в среднем 100 строк по 140 символов (включая пробелы) в каждой. Найдите объем (в Мбайтах) текстовой информации в словаре, если при записи используется кодировка «*один символ — один байт*».

Дано:

i = 1 байт

K = 1600·100·140

I - ?

I = K·i

$$I = \frac{1600 \cdot 100 \cdot 140}{1024 \cdot 1024} \text{ Мб} \approx 21{,}36 \text{ Мб}$$

Ответ: 21,36 Мбайта

Вопросы и задания

Задание 1. Представьте в кодировке ASCII текст

Happy New Year!

а) шестнадцатеричным кодом

48 61 70 70 79 20 4E 65 77 20 59 65 61 72 21

б) десятичным кодом

72 97 112 112 121 32 78 101 119 32 89 101 97 114 33

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Для представления в шестнадцатеричном коде необходимо записать адрес ячейки, где находится нужный символ (строка+столбец).Для представления в десятичном коде выполняем перевод из 16-ой с.с. В 10-ую с.с.

48 (16-ой с.с.) -> X (10-ой с.с)

Вопросы и задания

Задание 2. В 15-м издании энциклопедии Britannica 32 тома, в каждом из которых порядка 1000 страниц. На одной странице размещается в среднем 70 строк по 120 символов (включая пробелы) в каждой. Найдите объем текстовой информации в энциклопедии, если при записи используется кодировка Unicode («*один символ — два байта*»).

Дано:

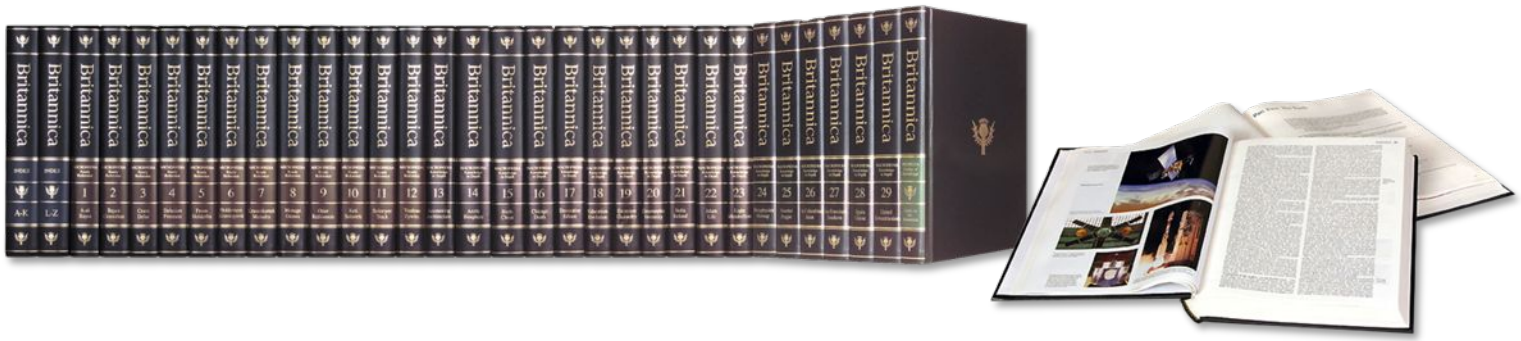
i = 2 байта

K = 32·1000·70·120

I = K·i

$$I = \frac{32 \cdot 1000 \cdot 70 \cdot 120 \cdot 2}{1024 \cdot 1024} \text{ Мб} \approx 513 \text{ Мб}$$

Ответ: 513 Мбайт



Задания для самостоятельного выполнения

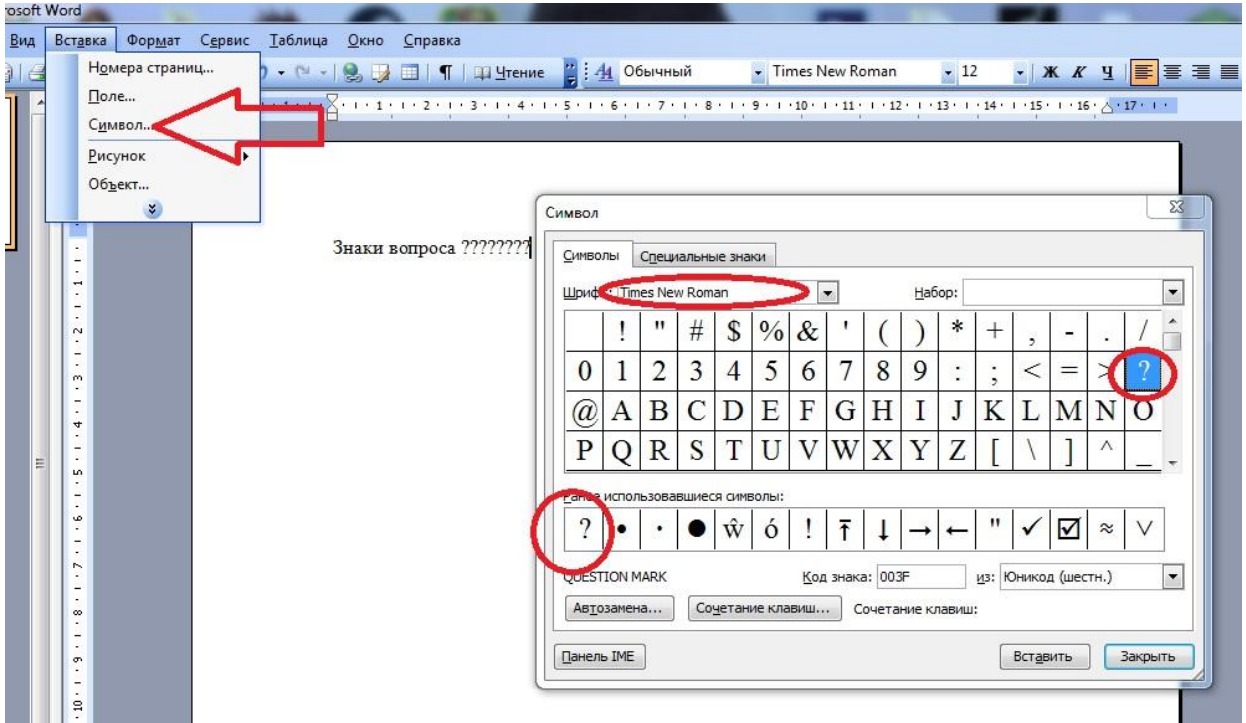
1. С помощью таблицы кодировки ASCII

А) декодируйте (расшифруйте) сообщение

64 65 73 6B 74 6F 70

Б) запишите в десятичном коде сообщение SCHOOL

2. В текстовом процессоре MS WORD откройте таблицу символов (вкладка ВСТАВКА-СИМВОЛ-ДРУГИЕ СИМВОЛЫ)



В поле ШРИФТ установите Times New Roman, в поле из-кириллица (дес.)

Вводя в поле Код знака десятичные коды символов, расшифруйте сообщение

196	238	240	238	227	243	32
238	241	232	235	232	242	32
232	228	243	249	232	233	46