Информатика

КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В КОМПЬЮТЕРЕ

10 класс

ИЗДАТЕЛЬСТВО БИНОМ

Ключевые слова

- текстовая информация
- кодирование
- кодовые таблицы

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.

Расширение кодировки ASCII

Расширение кодировки ASCII

	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	{	|	}	~	

Windows-1251 **КОИ-8**

	Ъ	Ѓ	,	ѓ	„	…	†	‡	€	‰	Љ	‹	Њ	Ќ	Ћ	Џ
9	ђ	‘	’	“	”	•	–	—		™	љ	›	њ	ќ	ћ	џ
A		Ў	ў	Ј	¤	Ґ	¦	§	Ё	©	Є	«	¬		®	Ї
B	°	±	І	і	ґ	µ	¶	·	ё	№	є	»	ј	Ѕ	ѕ	ї
C	А ю	Б а	В б	Г ц	Д д	Е е	Ж ф	З г	И х	Й и	К й	Л к	М л	Н м	О н	П о
D	Р п	С я	Т р	У с	Ф т	Х у	Ц ж	Ч в	Ш ь	Щ ы	Ъ з	Ы ш	Ь э	Э щ	Ю ч	Я ъ
E	а Ю	б А	в Б	г Ц	д Д	е Е	ж Ф	з Г	и Х	й И	к Й	л К	м Л	н М	о Н	п О
F	р П	с Я	т Р	у С	ф Т	х У	ц Ж	ч В	ш Ь	щ Ы	ъ З	ы Ш	ь Э	э Щ	ю Ч	я Ъ

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

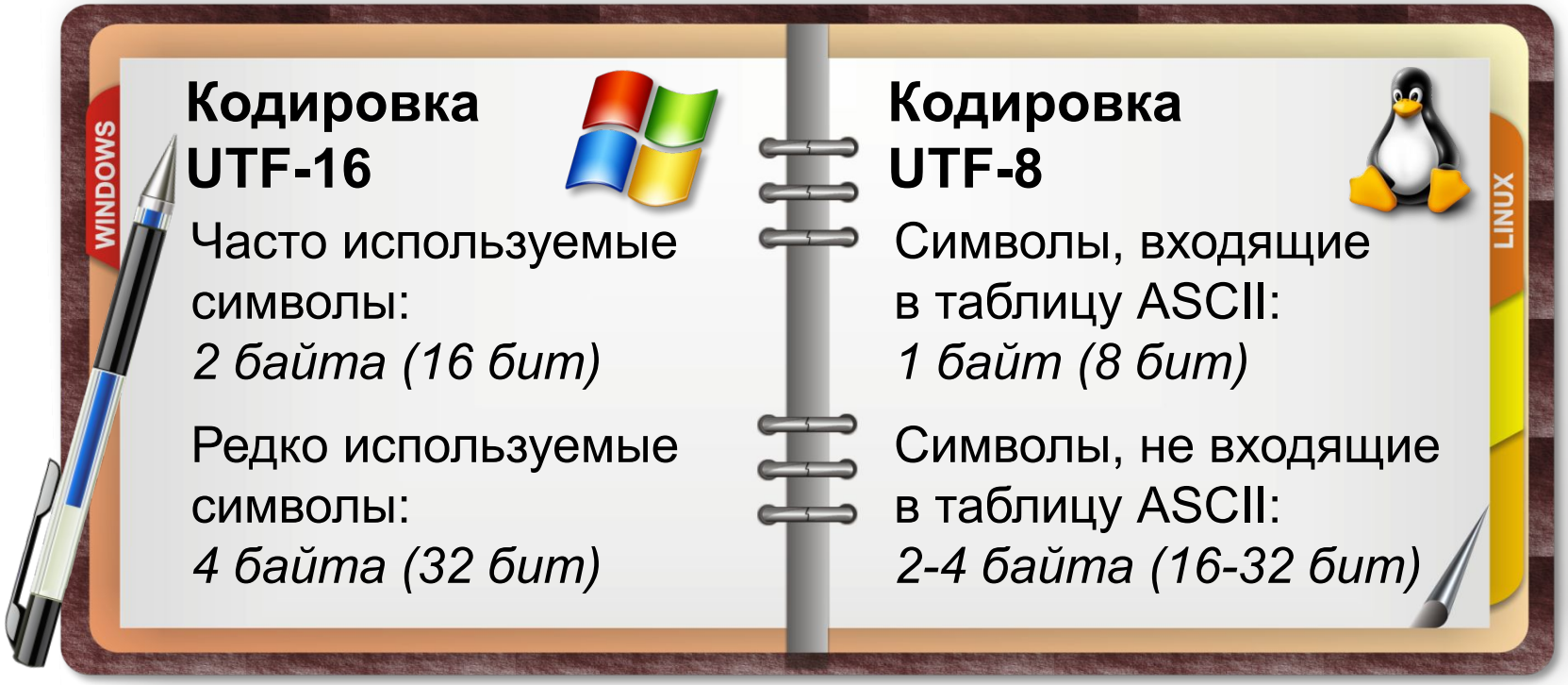
АРАБСКАЯ

АРМЯНСКАЯ

ЯПОНСКАЯ

Кодировки стандарта Unicode

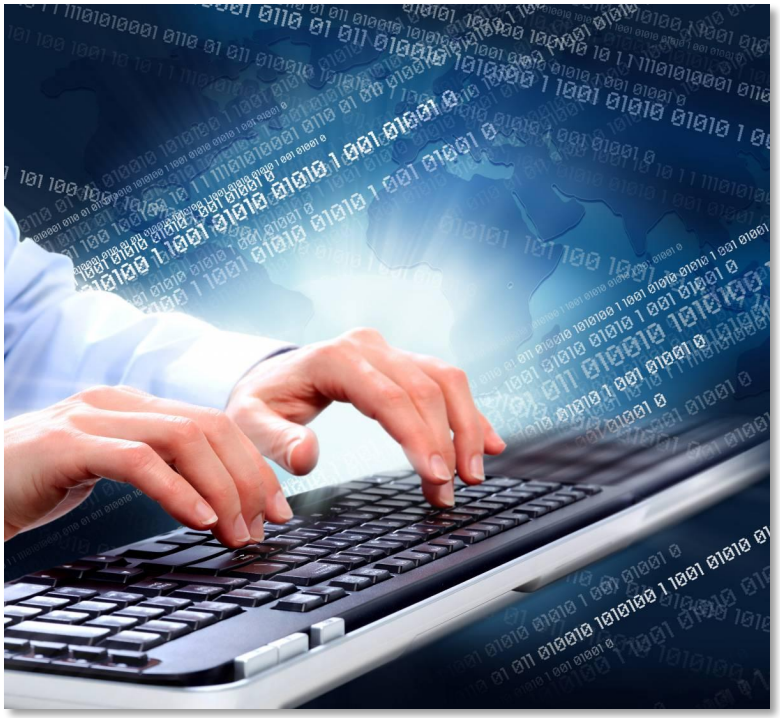
Для представления символов в памяти компьютера в стандарте Unicode имеется несколько кодировок.



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.



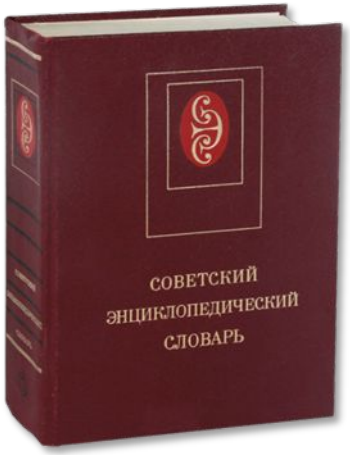
Количество символов в сообщении

$$I = K \cdot i$$

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

Самое главное

Текстовая информация по своей природе дискретна, так как представляется последовательностью отдельных символов.

В памяти компьютера хранятся специальные кодовые таблицы, в которых для каждого символа указан его двоичный код. Все кодовые таблицы, используемые в любых компьютерах и любых операционных системах, подчиняются международным стандартам кодирования символов.

Основой для компьютерных стандартов кодирования символов послужил код ASCII, рассчитанный на передачу только английского текста. Расширения ASCII-кодировки, в которых первые 128 символов кодовой таблицы совпадают с кодировкой ASCII, а остальные (с 128-го по 255-й) используются для кодирования букв национального алфавита, символов национальной валюты и т. п.

Самое главное

В 1991 году был разработан новый стандарт кодирования символов, получивший название Unicode (Юникод), позволяющий использовать в текстах любые символы любых языков мира. Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

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

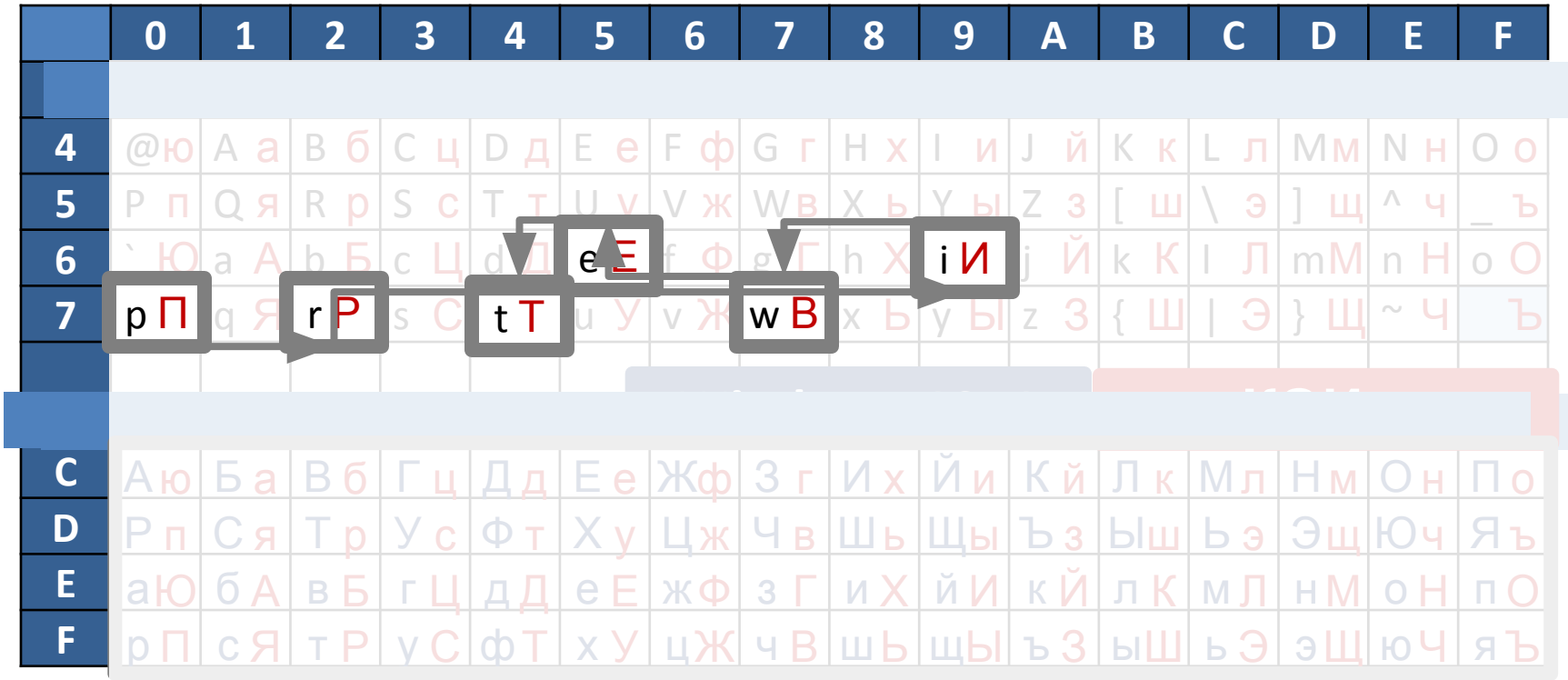
72 97 112 112 121 32 78 101 119 32 89 101 97 114 33

	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

ОТВЕТ

Подходы к расположению русских букв в различных кодировках

Задание 2. Сравните подходы к расположению русских букв в кодировках Windows-1251 и КОИ-8.



ПОДСКАЗКА - 1

ПОДСКАЗКА - 2

Вопросы и задания

Задание 3. 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:

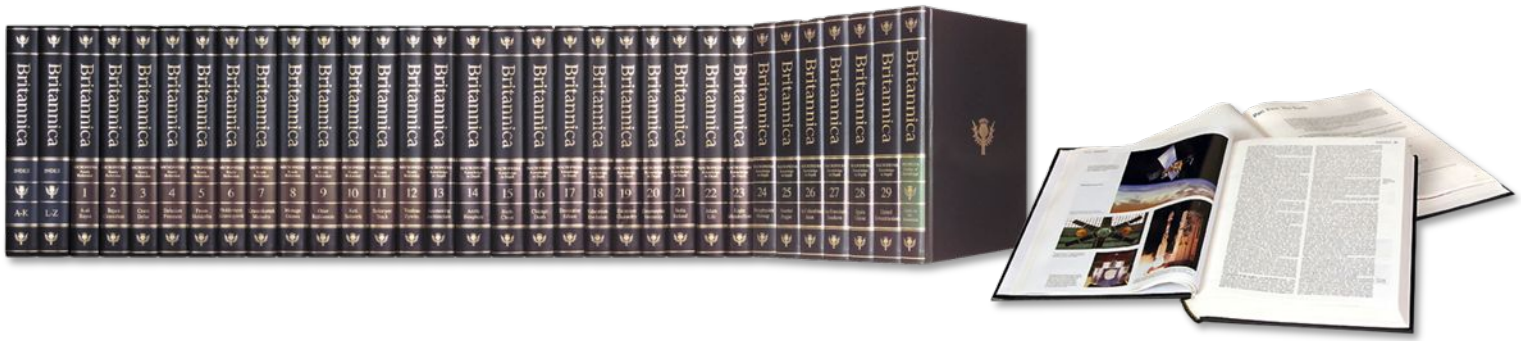
i = 2 байта

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



Информационные источники

- http://dev.bowdenweb.com/a/i/cons/utilities/unicode/unicode-2000px.png
- https://openclipart.org/image/2400px/svg_to_png/177279/Blank-Generic-Keyboard-Remix-by-Merlin2525.png
- http://arstyle.org/uploads/posts/2010-07/1278744192_1274782943_dreamstime_9113949-converted.jpg
- http://www.businesstoday.net.my/wp-content/uploads/2015/04/Computer-Programmer-Coding-Camp-shutterstock.jpg
- http://static.ozone.ru/multimedia/1005976053.jpg
- http://gimnnik.narod.ru/open-office/TextProcessor/p5aa1.html
- http://media.washtimes.com.s3.amazonaws.com/media/image/2012/03/14/encyclopaedia-britann_lea.jpg
- http://www.novilist.hr/var/novilist/storage/images/sci-tech/tehnologija/encyclopaedia-britannica-prekida-tiskanje-postaje-digitalna/1306075-1-cro-HR/Encyclopaedia-Britannica-prekida-tiskanje-postaje-digitalna.jpg