

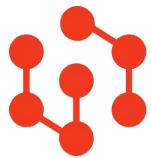
СИБИРСКИЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

SIBERIAN
FEDERAL
UNIVERSITY

Представление и кодирование текстовых данных в цифровой форме.

Корикова Арина, Морозова Ирина, Варлаганвоа Виктория

ГФ25-01Б



Введение

Для передачи информации между собой люди используют знаки и символы. Начав с простейших условных жестов, человек создал целый мир знаков, главным средством общения в котором стал язык. С появлением вычислительных машин возникла задача представления в цифровой форме нечисловых величин, и в первую очередь — символов, слов, предложений и текста.

1. **Кодирование** — это процесс представления информации в виде последовательности условных обозначений.
2. **Кодировка** — последовательность символов из некоторого алфавита, используемая для кодирования информации.
3. **Код** — уникальное двоичное число без знака, соответствующее определённому символу.



Стандарт ASCII

Широкое распространение персональных компьютеров фирмы IBM привело к тому, что стандарт ASCII приобрёл статус международного. В таблице ASCII содержится 256 символов и их кодов. Таблица состоит из двух частей: основной и расширенной. Основная часть (символы с кодами от 0 до 127 включительно) является базовой и в соответствии с принятым стандартом не может быть изменена. В неё входят управляемые символы (им соответствуют коды с 1 по 31), арабские цифры, буквы латинского алфавита, знаки препинания, специальные символы. При преобразовании в двоичную форму коды представляют собой семиразрядные целые двоичные числа в диапазоне от $000\ 0000_2 = 0016 = 0$ до $111\ 1111_2 = 7F_{16} = 127$.

Основная таблица ASCII

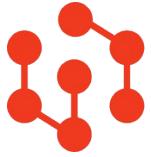
Диапазон кодов	Тип символов	Особенности
0-31 ($00\text{-}1F_{16}$)	Управляющие символы	Не отображаются на экране
32-127 ($20\text{-}7F_{16}$)	Графические символы	Цифры, буквы, знаки препинания

ASCII Code Chart																
0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI	
DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US	
	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	0	
P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	-	
~	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL	

Примеры:

$08_{16} = BS$ — стирание предыдущего символа

$42_{16} = "B"$ — отображение символа «B»



Расширенная таблица ASCII

В соответствии с утверждёнными стандартами эта часть таблицы изменяется в зависимости от национального алфавита страны, в которой она используется, и способа кодирования. Именно поэтому при 17 наименовании программ, документов и других объектов желательно использовать латинские буквы, содержащиеся в основной, неизменяемой части таблицы, так как русскоязычные имена при несоответствии таблиц кодирования будут отображаться неверно.

Кодовые страницы — это расширение кода ASCII.

Extended ASCII используется для:

- Национальных алфавитов
- Символы псевдографики
- Специальных символов



Кодовая страница CP1251

Пользуется довольно большой популярностью. CP1251 выгодно отличается от других 8-битных кириллических кодировок наличием практически всех символов, используемых в русской типографике для обычного текста, а также всех символов для языков, близких к русскому: украинского, белорусского, сербского и болгарского.

Кодовая таблица Windows-1251

0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F		
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
1	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
2	!	"	#	\$	%	&	()	*	+					/		
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	
4	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	
5	@	А	Б	С	Д	Е	Ғ	Г	Ҳ	Ӣ	Ҷ	Ҹ	Ҳ	Ҵ	ҵ	Ҷ	
6	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	
7	P	Q	R	S	T	U	V	W	X	Y	Z	[]	^	-		
8	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	
9	'	а	б	с	д	е	ғ	ҳ	ӣ	ҷ	қ	լ	м	п	օ		
10	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	
11	р	q	г	с	т	ү	в	w	ҳ	ӯ	з	{		}}	~	□	
12	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	
13	Ҧ	Ү	,	Ҫ	"	...	ҩ	Ҫ	ҩ	ҩ	Ҧ	Ү	Ұ	ұ	Ҳ	ҳ	
14	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	
15	Ү	ү	'	Ү	"	-	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
16	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	
17	Ҩ	ҩ	Ҫ	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
18	A	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
19	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
20	B	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
21	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
22	C	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
23	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
24	D	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
25	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
26	E	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
27	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	Ү	
28	F	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255



Кодовая страница КОИ8

Разработчики КОИ-8 расположили символы русского алфавита в таблице таким образом, что позиции кириллических символов соответствуют их фонетическим аналогам в английском алфавите в базовой таблице. Это означает, что если в тексте, написанном в КОИ-8, убрать восьмой бит каждого символа (вычесть 128), то получится читаемый текст, хотя он и написан латинскими символами. Но из-за такого решения символы кириллицы оказались расположены не в алфавитном порядке.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	-		г	л	л	т	т	т	+	■	■	■	■	■	■	■
9
A	=		ф	ё	п	г	г	т	т	т	т	т	т	т	т	т
B			т	т	т	т	т	т	т	т	т	т	т	т	т	т
C	ю	я	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н
D	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ
E	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
F	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ

Пример: Слова «Русский Текст»
превратились бы в «гUSSKIJ тEKST»



Стандарт Unicode

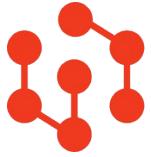
Стандарт символьной кодировки, в котором присутствует русский алфавит **Юникод**, был разработан в 1991 году и позволяет кодировать 2¹⁶ = 65 536 символов. Во многих странах Азии 256 кодов явно не хватало для кодирования национальных алфавитов. В 1991 году производители программных продуктов и организаций, утверждающие стандарты, пришли к соглашению о разработке единого стандарта. Этот стандарт построен на 16-битной схеме кодирования и получил название UNICODE. Способы кодирования символов таблицы Юникод, то есть преобразования номеров ячеек таблицы Юникод в двоичные коды, образуют кодовое пространство, состоящее из трёх кодов семейства UTF (формат преобразования Юникода): UTF-8, UTF-16 и UTF-32.

Характеристика	UTF-8	UTF-16	UTF-32
Тип кодирования	Переменная длина	Переменная длина	Фиксированная длина
Единица кода	От 1 до 4 байт	2 или 4 байта	Всегда 4 байта
Совместимость с ASCII	Да, полная	Нет	Нет
Применение	Интернет, Linux, macOS	Windows API, Java, .NET	Специализированные задачи, в которых важна скорость доступа к символу



Различия:

- **UTF-8:** UTF-8 очень эффективен для латинских алфавитов (например, английского), так как символы ASCII занимают всего 1 байт. Это делает его наиболее компактным для большинства западных языков.
- **UTF-16:** UTF-16 может быть эффективнее для языков с большим количеством иероглифов (например, китайского, японского), поскольку многие символы кодируются 2 байтами, в то время как в UTF-8 они заняли бы 3 байта.
- **UTF-32:** UTF-32 наименее эффективен по занимаемому пространству, так как каждый символ, независимо от его типа, занимает 4 байта. Файл с простым английским текстом в UTF-32 будет в четыре раза больше, чем тот же файл в UTF-8.



Практические рекомендации:

- Для имен файлов и программ рекомендуется использовать латинские буквы из основной таблицы ASCII.
- Unicode устраняет проблему несовместимости кодовых таблиц.
- Современные технические средства компенсируют увеличение размера файлов в Unicode.