Технологии обработки текстовой информации. Кодовые таблицы

Технологии обработки текстовой информации. Кодовые таблицы

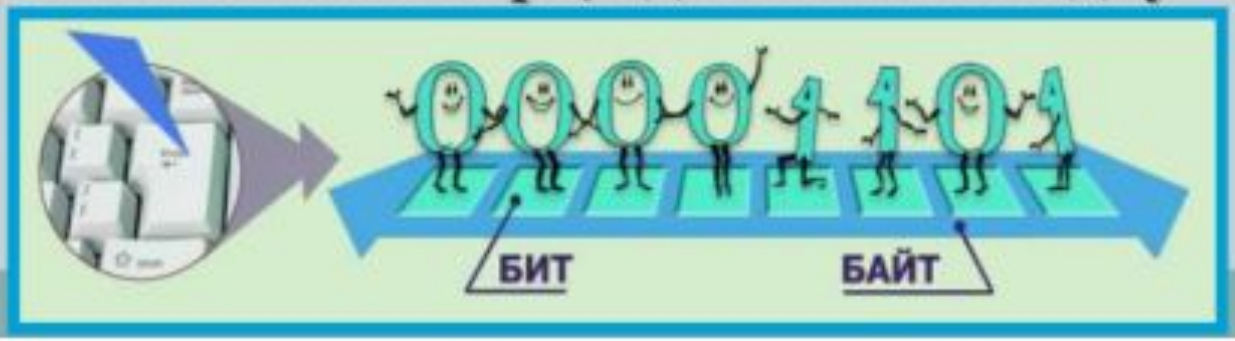
- *Текстовая информация* – это информация, представленная в виде букв, знаков препинания и специальных символов некоторой знаковой системы. Буквы и другие знаки принято называть символами. Набор их конечен. Текстовую информацию иногда называют символьной. Знаковая система содержит еще и правила выполнения операций над знаками (грамматика, синтаксис).
- *Кодирование* – это процесс представления каждого символа в виде кода.
- *Код* – набор условных обозначений для представления информации.
- Количество знаков в коде называется *длиной кода*.

- *Естественные языки* – это знаковые системы с исключениями из правил. Поэтому их нельзя использовать для кодирования информации с последующей обработкой на компьютере. Знаковые системы со строгими правилами называются *формальными*.
- Для компьютерного кодирования информации используется формальная двоичная знаковая система. Физическая природа знаков двоичного компьютерного кода – это электрические импульсы (наличие импульса или его отсутствие). При кодировании используется *кодировочная таблица*. Таблица устанавливает взаимно однозначное соответствие между знаками и их кодами.

Кодирование текстовой информации

Человек различает знаки по их начертанию, а компьютер - по их двоичным кодам. При вводе в компьютер текстовой информации происходит ее двоичное кодирование.

Пользователь нажимает на клавиатуре клавишу со знаком, и в компьютер поступает определенная последовательность из восьми электрических импульсов (двоичный код знака). Код знака хранится в оперативной памяти компьютера, где занимает одну ячейку.



Кодирование текстовой информации

Для представления текстовой информации достаточно 256 различных знаков.

По формуле $\mathbf{N=2^I}$ можно вычислить, какое количество информации необходимо, чтобы закодировать каждый знак:

$$\mathbf{N = 2^I \Rightarrow 256 = 2^I \Rightarrow 2^8 = 2^I \Rightarrow}$$

$$\mathbf{I = 8 \text{ бит} = 1 \text{ байт}}$$

Для кодирования **одного символа** требуется **один байт** информации

Кодирование текстовой информации

- ***Декодирование*** – процесс обратный кодированию, т.е. код символа преобразуется в его изображение. Процесс декодирования информации осуществляется при выводе информации из оперативной памяти компьютера на экран монитора, например, или на листинг с помощью принтера.

Таблица кодировки

При кодировании каждому символу алфавита ставиться в соответствие уникальный двоичный код.

Таблица кодировки – это таблица, в которой всем символам компьютерного алфавита поставлены в соответствие порядковые номера (коды).

Кодировки знаков

Двоичный код	Десятичный код	КОИ-8	Windows	MS-DOS	Mac	ISO
00000000	0					
...						
00001000	8	удаление последнего символа (клавиша {Backspace})				
...						
00001101	13	перевод строки (клавиша {Enter})				
...						
00100000	32	клавиша {Пробел}				
00100001	33	!				
...						
01011010	90	Z				
...						
01111111	127	□				
10000000	128	─	Ъ	А	А	к
...						
11000010	194	б	В	─	─	Т
...						
11001100	204	л	М	╣	╣	Ъ
...						
11011101	221	щ	Э	_	Ё	и
...						
11111111	255	ъ	я	нераз. пробел	нераз. пробел	п

В существующих кодовых таблицах десятичные коды :

- **от 0 до 32** соответствуют операциям (перевод строки, ввод пробела и т.д.);
- **от 33 по127** соответствуют знакам латинского алфавита, цифрам, знакам арифметических операций и знакам препинания;
- **от 128 по 255** в различных национальных кодировках одному и тому же коду соответствуют разные знаки.

Десятичные коды некоторых символов в различных кодировках

В настоящее время существуют пять различных кодовых таблиц для русских букв (Windows, MS-DOS, КОИ-8, Mac, ISO) поэтому тексты, созданные в одной кодировке, не будут правильно отображаться в другой.

Символ	Windows	MS-DOS	КОИ-8	Mac	ISO	Unicode
А	192	128	225	128	176	1040
В	194	130	247	130	178	1042
М	204	140	237	140	188	1052
Э	221	157	252	157	205	1069
я	255	239	241	223	239	1103

Таблицы кодировки русскоязычных символов

КОИ-8

Таблицы кодировки русскоязычных символов

CP1251(Windows)

Ѓ 128	ѓ 129	‚ 130	ѐ 131	„ 132	… 133	† 134	‡ 135	€ 136	‰ 137	Љ 138	‹ 139	њ 140	Ќ 141	ћ 142	Џ 143
ђ 144	‘ 145	’ 146	“ 147	” 148	• 149	– 150	— 151	˜ 152	™ 153	љ 154	› 155	ќ 156	ѝ 157	џ 158	Ў 159
nbsp 160	ў 161	Ј 162	¤ 163	Ґ 164	¦ 165	§ 166	Ё 167	© 168	Є 169	« 170	¬ 171	shy 172	® 173	Ї 174	° 175
± 176	І 177	і 178	ґ 179	µ 180	¶ 181	· 182	ё 183	№ 184	є 185	» 186	ј 187	Ѕ 188	ѕ 189	ї 190	А 191
Б 192	В 193	Г 194	Д 195	Е 196	Ж 197	З 198	И 199	Й 200	К 201	Л 202	М 203	Н 204	О 205	П 206	Р 207
С 208	Т 209	У 210	Ф 211	Х 212	Ц 213	Ч 214	Ш 215	Щ 216	Ъ 217	Ы 218	Ь 219	Э 220	Ю 221	Я 222	а 223
б 224	в 225	г 226	д 227	е 228	ж 229	з 230	и 231	й 232	к 233	л 234	м 235	н 236	о 237	п 238	р 239
с 240	т 241	у 242	ф 243	х 244	ц 245	ч 246	ш 247	щ 248	ъ 249	ы 250	ь 251	э 252	ю 253	я 254	255

Таблицы кодировки русскоязычных символов

CP866 (MS-DOS)

А 128	Б 129	В 130	Г 131	Д 132	Е 133	Ж 134	З 135	И 136	Й 137	К 138	Л 139	М 140	Н 141	О 142	П 143
Р 144	С 145	Т 146	У 147	Ф 148	Х 149	Ц 150	Ч 151	Ш 152	Щ 153	Ъ 154	Ы 155	Ь 156	Э 157	Ю 158	Я 159
а 160	б 161	в 162	г 163	д 164	е 165	ж 166	з 167	и 168	й 169	к 170	л 171	м 172	н 173	о 174	п 175
▒ 176	▒ 177	▓ 178	│ 179	┤ 180	╡ 181	╢ 182	╖ 183	╕ 184	╣ 185	║ 186	╗ 187	╝ 188	╜ 189	╛ 190	┐ 191
└ 192	┴ 193	┬ 194	├ 195	─ 196	┼ 197	╞ 198	╟ 199	╚ 200	╔ 201	╩ 202	╦ 203	╠ 204	═ 205	╬ 206	╧ 207
╨ 208	╤ 209	╥ 210	╙ 211	╘ 212	╒ 213	╓ 214	╫ 215	╪ 216	┘ 217	┌ 218	▉ 219	▄ 220	▌ 221	▐ 222	▀ 223
р 224	с 225	т 226	у 227	ф 228	х 229	ц 230	ч 231	ш 232	щ 233	ъ 234	ы 235	ь 236	э 237	ю 238	я 239
Ё 240	ё 241	Є 242	є 243	Ї 244	ї 245	Ў 246	ў 247	° 248	∙ 249	· 250	√ 251	№ 252	¤ 253	■ 254	nbsp 255

Таблицы кодировки русскоязычных символов

ISO

128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
nbsp 160	Ё 161	Ђ 162	Ѓ 163	Є 164	S 165	I 166	Ї 167	J 168	Љ 169	Њ 170	Ћ 171	Ќ 172	shy 173	Ў 174	Џ 175
А 176	Б 177	В 178	Г 179	Д 180	Е 181	Ж 182	З 183	И 184	Й 185	К 186	Л 187	М 188	Н 189	О 190	П 191
Р 192	С 193	Т 194	У 195	Ф 196	Х 197	Ц 198	Ч 199	Ш 200	Щ 201	Ъ 202	Ы 203	Ь 204	Э 205	Ю 206	Я 207
а 208	б 209	в 210	г 211	д 212	е 213	ж 214	з 215	и 216	й 217	к 218	л 219	м 220	н 221	о 222	п 223
р 224	с 225	т 226	у 227	ф 228	х 229	ц 230	ч 231	ш 232	щ 233	ъ 234	ы 235	ь 236	э 237	ю 238	я 239
№ 240	ё 241	ђ 242	ѓ 243	є 244	s 245	і 246	ї 247	j 248	љ 249	њ 250	ћ 251	ќ 252	§ 253	ў 254	џ 255

Mac

А 128	Б 129	В 130	Г 131	Д 132	Е 133	Ж 134	З 135	И 136	Й 137	К 138	Л 139	М 140	Н 141	О 142	П 143
Р 144	С 145	Т 146	У 147	Ф 148	Х 149	Ц 150	Ч 151	Ш 152	Щ 153	Ъ 154	Ы 155	Ь 156	Э 157	Ю 158	Я 159
† 160	° 161	ѝ 162	£ 163	§ 164	• 165	¶ 166	ы́ 167	® 168	© 169	™ 170	Ѐ 171	ѐ 172	ё 173	ѐ 174	ѐ 175
Ё 176	± 177	≤ 178	≥ 179	ѕ́ 180	µ 181	ґ 182	Ѕ́ 183	Ю́ 184	ю́ 185	Я́ 186	я́ 187	Е́ 188	е́ 189	и́ 190	о̀ 191
э̀ 192	ю̀ 193	¬ 194	√ 195	ƒ 196	≈ 197	Δ 198	« 199	» 200	… 201	nbsp 202	О́ 203	о́ 204	И́ 205	и́ 206	я̀ 207
– 208	— 209	“ 210	” 211	‘ 212	’ 213	÷ 214	„ 215	у̀ 216	Ы́ 217	у́ 218	у́ 219	№ 220	Ё 221	ё 222	я 223
а 224	б 225	в 226	г 227	д 228	е 229	ж 230	з 231	и 232	й 233	к 234	л 235	м 236	н 237	о 238	п 239
р 240	с 241	т 242	у 243	ф 244	х 245	ц 246	ч 247	ш 248	щ 249	ъ 250	ы 251	ь 252	э 253	ю 254	¤ 255

Кодовая таблица ASCII

- Для разных типов ЭВМ используются различные таблицы кодировки.
- С распространением персональных компьютеров типа IBM PC международным стандартом стала таблица кодировки под названием ASCII (American Standard Code for Information Interchange) – американский стандартный код для информационного обмена

sp 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	

- Этот формат оперирует с 256 численными кодами, имеющими значения от 0 до 255. В соответствие каждому коду ставится определенный символ (буква, цифра, знак препинания, математический символ или символ псевдографики). Это соответствие задается с помощью стандартных кодовых таблиц с различными номерами (например, таблица 866 предназначена для русскоязычных пользователей). Не содержит форматирования текста, поэтому является переносимым между различными операционными системами и программами.

- ANSI (American National Standard Interface) кодировка в среде Windows. У этих кодировок совпадают те части, которые относятся к латинскому алфавиту, специальным символам, цифрам, знакам препинания и математическим операциям, а различаются относящиеся к другим алфавитам и псевдографике

Понятие кодировки Unicode(UCS - 2)

В последние годы широкое распространение получил новый международный стандарт кодирования текстовых символов Unicode, который отводит на каждый символ 2 байта (16 битов). По формуле можно определить количество символов, которые можно закодировать согласно этому стандарту: $N = 2^I = 2^{16} = 65\,536$.

Такого количества символов достаточно, чтобы закодировать не только русский и латинский алфавиты, цифры, знаки и математические символы, но и греческий, арабский, иврит и другие алфавиты.

Текст - последовательность символов компьютерного алфавита.

Текстовая информация - это информация, выраженная с помощью естественных и формальных языков в письменной форме (прописные и строчные буквы русского и латинского алфавитов, цифры, знаки и математические символы).