ДВОИЧНОЕ КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

Что такое текст?

Первые компьютеры были созданы для обработки числовой информации, но начиная с 60-х годов XX века, появилась возможность обработки и текстовой (символьной) информации.

Компьютерный текст – любая последовательность символов из компьютерного алфавита. Текст может быть на естественном языке (например, русском или английском), может содержать химические или математические формулы, таблицы и пр. Главное, чтобы все символы, используемые в тексте, входили в **компьютерный алфавит.**

Алфавит должен включать латинские и русские прописные и строчные буквы, цифры, знаки препинания и арифметических операций, специальные знаки.

Для представления текста в компьютерном алфавите используется 256 символов

В с п о м н и м …

Любая информация представляется в памяти компьютера в двоичном виде.

Для компьютерной обработки текста необходимо **кодирование** – преобразование входной информации (каждого символа текста) в форму, воспринимаемую компьютером, т.е. двоичный код.

Чтобы вывести текст из памяти на экран или печать, нужно **декодирование** - преобразование двоичных кодов в символы.

Алфавит – множество символов для записи текста.

Мощность алфавита (N) – количество символов в алфавите.

Определить **информационный вес символа (**i**)**, т. е. **количество бит для представления одного символа** из алфавита указанной мощности, можно по формуле Хартли:

N = 2^{I} ($256 = 2^{8}$ i = 8 бит = 1 байт).

Для кодирования одного символа в алфавите мощностью 256 символов требуется 8 бит или 1 байт информации.

Что такое т а б л и ц а к о д и р о в к и ?

Чтобы поставить в соответствие каждому символу числовой код нужна **таблица кодировки – стандарт, в котором всем символам компьютерного алфавита поставлены в соответствие порядковые номера в двоичной системе счисления.**

Международным стандартом является таблица кодировки ASCII

Для кодировки русских букев существует пять различных кодовых таблиц : КОИ-8, С31251, СР866, Mac, ISO

В последнее время появился новый международный стандарт UNICODE, в котором для представления символа отводится 2 байта.

Стандарт ASCII

В 1967 году в США был введен код **ASCII (American Standart Code for Information Interchange - Американский Стандартный Код для Обмена Информацией).** В нем каждому символу ставился в соответствие 7-битный двоичный код, всего 128 символов (кодов), из них:

Управляющие (коды от 0 до 31, а также 127), не отображаются на экране;

Отображаемые (коды от 32 до 126). Код 32 – пробел, отображает пропуск на экране. Далее следуют знаки препинания, скобки, арабские цифры (0 - 9), некоторые знаки, латинские прописные, затем строчные буквы, знаки.

После модификации в 1977 году стандартом был принят 1 байт, и каждому символу в ASCII поставлено число от 00000000 до 11111111 (0 - 255 в десятичной системе счисления).

Коды 0 -127 - являются международным стандартом,

128 – 255 используются для национальных алфавитов и специальных знаков (расширенная таблица).

Кодировка кириллицы

Стандарт CP866, альтернативная кодировка — кодовая страница, где все специфические европейские символы в верхней половине кодовой таблицы были заменены на кириллицу. Популярен в среде MS-DOS и OS/2. Разработана в ВЦ АН СССР, для которого впервые в СССР была закуплена партия IBM PC.

Стандарт CP1251. Кодовая страница Microsoft CP1251 создана Microsoft как стандарт для кодировки кириллицы в Windows.

Стандарт КОИ8. В нем символы русской кириллицы поместили так, что позиции символов кириллицы соответствуют их фонетическим аналогам в английском алфавите. Это означает, что если в тексте, написанном в КОИ8, убрать восьмой бит каждого символа, то мы имеем "читабельный" текст, хотя он и написан английскими символами. KOI8-R быстро стал фактически стандартом для русской кириллицы в Internet

Кодировка ASCII (коды 0 – 127)

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

К о д и р о в к а Windows-1251 (CP1251)

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

Стандарт KOI8-R

─ 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	nbsp 154	⌡ 155	° 156	² 157	· 158	÷ 159
═ 160	║ 161	╒ 162	ё 163	╓ 164	╔ 165	╕ 166	╖ 167	╗ 168	╘ 169	╙ 170	╚ 171	╛ 172	╜ 173	╝ 174	╞ 175
╟ 176	╠ 177	╡ 178	Ё 179	╢ 180	╣ 181	╤ 182	╥ 183	╦ 184	╧ 185	╨ 186	╩ 187	╪ 188	╫ 189	╬ 190	© 191
ю 192	а 193	б 194	ц 195	д 196	е 197	ф 198	г 199	х 200	и 201	й 202	к 203	л 204	м 205	н 206	о 207
п 208	я 209	р 210	с 211	т 212	у 213	ж 214	в 215	ь 216	ы 217	з 218	ш 219	э 220	щ 221	ч 222	ъ 223
Ю 224	А 225	Б 226	Ц 227	Д 228	Е 229	Ф 230	Г 231	Х 232	И 233	Й 234	К 235	Л 236	М 237	Н 238	О 239
П 240	Я 241	Р 242	С 243	Т 244	У 245	Ж 246	В 247	Ь 248	Ы 249	З 250	Ш 251	Э 252	Щ 253	Ч 254	Ъ 255

Стандарт CP866

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	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	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
Ё	ё	Є	є	Ї	ї	Ў	ў	°	∙	·	√	№	¤	■	nbsp
240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255

Сортировка текста по алфавиту

В таблицах кодировки соблюдается **принцип последовательного кодирования (лексикографический)**: в начале упорядочены цифры (от 0 до 9), затем приводится латинский алфавит: прописные (большие), затем - строчные (маленькие) буквы, во второй части таблицы дается кириллица (русский алфавит), также сначала прописные, затем строчные буквы. Этот принцип позволяет **сортировать символьную информацию.**

П р и м е р

Изучите в приведенных кодировках размещение символов**.**

Попробуйте определить: в каком порядке будут идти фрагменты текста «excel», «байт», «8в», «10г», «9а», «10а», если упорядочить их по возрастанию?

Ответ: «10а», «10г», «8в», «9а», «excel», «байт»

Почему? Сначала будут упорядочены по возрастанию коды первых символов, затем, затем среди одинаковых первых символов, будут упорядочены вторые и т. д.

Информационный объем текста

Статьи, рефераты, дипломы и прочие документы, подготавливают на компьютере в текстовых редакторах. Обычно известно, какая кодировка используется программой. Все это позволяет определить **информационный объем** документа.

Пример

Пусть реферат содержит 32 страницы; на каждой странице — 32 строки, в каждой строке — 64 символа. Определить информационный объем реферата в кодировке ASCII .

Одна страница содержит $32 \times 64 = 2^5 \times 2^6 = 2^{11}$ символов. Тогда в всем реферате: $32 \times 2^{11} = 2^5 \times 2^{11} = 2^{16}$ символов. В кодировке ASCII для хранения символа требуется один байт. Объем реферата: $2^{16} \times 1 = 2^{16}$ байт $= 2^{16} / 2^{10} = 2^6$ Кбайт = 64 Кбайта

Решаем задачи сами…

1. Оцените информационный объем сообщения в битах и байтах, представленного в кодировке ASCII:
2. **В одном килограмме 1000 граммов**
3. Какое сообщение закодировано в кодировке Windows-1251:
   00110101001000001110000111100000111010111110101111101110111000 10
4. Считая, что каждый символ кодируется двумя байтами, оцените информационный объем следующего предложения из пушкинского четверостишия:
   Певец-Давид был ростом мал, Но повалил же Голиафа!
5. Выбрать слово, имеющее наибольшую сумму кодов символов в таблице кодировки ASCII.
6. A. окно; B. кино; C. ника; D. конь; E. ночь.
7. Выбрать слово, имеющее наибольшую сумму кодов символов в таблице кодировки ASCII.
8. A. 2b2d; B. файл; C. file; D. 1999; E. 2001.
11. Декодируйте следующее сообщение, записанное восьмибитовой кодировке:
    01010101 01110000 00100000 00100110 00100000
    01000100 01101111 01110111 01101110
11. Определите вид кодировки и декодируйте следующие сообщения:
    а) 235 207 212 197 204 216 206 201 203 207 215
    б) 213 224 244 244 236 224 237

Ответьте на вопросы…

1. Что такое компьютерный текст?
2. Что такое кодирование?
3. Что такое алфавит? мощность алфавита?
4. Что такое таблица кодировки?
5. В чем суть принципа последовательного кодирования?
6. Какие вам известны таблицы кодировки?
7. Какие Вам известны таблицы кодировок?
8. Как определить информационный объем текста.?