Кодирование текстовой информации

Как вы, наверное, уже знаете, современные компьютерные системы уже давно не работают только с числовой информацией, как это было на начальном этапе их развития. Они наряду с числами могут обрабатывать и буквы, и слова, причем на различных языках, и так называемые **специальные символы**, к которым относят следующие знаки: . , ! ? " # $ % ^ & * ( ) _ - = + \ > < и др. Очевидно, что любое математическое выражение или строка букв состоит из этих отдельных элементов — букв, цифр, знаков и т. д. Будем в дальнейшем называть их **символами**. Для представления такой информации в компьютерах создают специальные коды, которые так и называют — **символьные коды.**

Это очень существенный момент, не зря еще в 1964 г. в нашей стране был выпущен посвященный этим вопросам государственный общесоюзный стандарт — ГОСТ 10859—64. В последние 15—20 лет эти вопросы определяются международными стандартами. Например, большие вычислительные машины типа IBM 360 и IBM 370 и их аналоги, которые доминировали в мире в 1970—1980-е гг., использовали **специальный код EBCDIC** (Extended Binary Coded Decimal Interchange Code — расширенный двоично-десятичный код обмена информацией), у нас его аналогом был **код ДКОИ-8** (двоичный код для обмена информацией, 8 бит). **Код ASCII** применяется в ПК, совместимых с IBM, работающих под управлением операционной системы MS DOS (Microsoft Disk Operating System — дисковая ОС).

Сейчас, как вы наверняка знаете, **в основном используется ОС Microsoft Windows**. Она применяет так называемую **ANSI-ко-**

дировку. Но эта кодировка ориентирована на английский язык и не содержит символов кириллицы (русских букв), поэтому американская компания **Microsoft** — разработчик Windows — **создала русскую версию ANSI-кодировки**, которая будет приведена далее. **Наша задача** — не заучивать эти многочисленные символы, а **понять механизм формирования кодов**, который, кстати, практически одинаков во всех символьных компьютерных кодах.

Вы уже знаете, что в компьютере каждый символ представлен в виде байта, состоящего из восьми двоичных разрядов, называемых битами. Вы помните также, что содержимым бита может быть либо «0», либо «1», а также то, что 8 бит могут дать 256 комбинаций из «0» и «1». Вы также знаете, что 4 бит, представляющие тетраду байта, «свернуты» в 16-ричную цифру. Если вы внимательно посмотрите на приведенную таблицу русской версии ANSI-кодировки (рис. 2.11), то увидите, что **каждая строка** и **столбец начинаются** с **16-ричной цифры** и **соответствующей битовой комбинации**, а на пересечении строк и столбцов в соответствующих клетках находятся **кодируемые символы**. Возьмем, например, **знак «+»**. В строке стоит 16-ричная цифра 2 и соответствующая ей кодовая комбинация — тетрада 0010, а в столбце — 16-ричная цифра B и соответствующая ей кодовая комбинация 1011. Следовательно, **знак «+» в компьютере представлен кодом 00101011** (в строках указана первая половина восьмибитового кода, а в столбцах — вторая), в 16-ричном виде этот код можно записать как 2B. Но в каждой клеточке указано еще какое-то десятичное число, в данном случае 43. Это не что иное, как переведенные в десятичную систему или двоичное число 00101011, или 16-ричное число 2B. Действительно,

$$00101011 = 1 + 2 + 8 + 32 = 43;$$

$$2B = 11 + 2 \cdot 16 = 11 + 32 = 43.$$

		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
		0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0	0000	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0001	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
2	0010	32	33 !	34 “	35 #	36 $	37 %	38 &	39 ‘	40 (	41)	42 *	43 +	44 ,	45 –	46 .	47 /
3	0011	48 0	49 1	50 2	51 3	52 4	53 5	54 6	55 7	56 8	57 9	58 :	59 ;	60 <	61 =	62 >	63 ?
4	0100	64 @	65 A	66 B	67 C	68 D	69 E	70 F	71 G	72 H	73 I	74 J	75 K	76 L	77 M	78 N	79 O
5	0101	80 P	81 Q	82 R	83 S	84 T	85 U	86 V	87 W	88 X	89 Y	90 Z	91 [	92 \	93]	94 ^	95 _
6	0110	96 `	97 a	98 b	99 c	100 d	101 e	102 f	103 g	104 h	105 i	106 j	107 k	108 l	109 m	110 n	111 o
7	0111	112 p	113 q	114 r	115 s	116 t	117 u	118 v	119 w	120 x	121 y	122 z	123 {	124 |	125 }	126 ~	127 □
8	1000	128 Ђ	129 Ѓ	130 ‚	131 ѓ	132 „	133 …	134 †	135 ‡	136 □	137 ‰	138 Љ	139 ‹	140 Њ	141 Ќ	142 Ћ	143 Џ
9	1001	144 ђ	145 ‘	146 ’	147 “	148 ”	149 •	150 –	151 —	152	153 ™	154 љ	155 ›	156 њ	157 ќ	158 ћ	159 џ
A	1010	160	161 Ў	162 ў	163 Ј	164 ¤	165 Ґ	166 ¦	167 §	168 Ё	169 ©	170 Є	171 «	172 ¬	173	174 ®	175 Ї
B	1011	176 °	177 ±	178 І	179 і	180 ґ	181 µ	182 ¶	183 ·	184 ё	185 №	186 є	187 »	188 j	189 S	190 s	191 ї
C	1100	192 А	193 Б	194 В	195 Г	196 Д	197 Е	198 Ж	199 З	200 И	201 Й	202 К	203 Л	204 М	205 Н	206 О	207 П
D	1101	208 Р	209 С	210 Т	211 У	212 Ф	213 Х	214 Ц	215 Ч	216 Ш	217 Щ	218 Ъ	219 Ы	220 Ь	221 Э	222 Ю	223 Я
E	1110	224 а	225 б	226 в	227 г	228 д	229 е	230 ж	231 з	232 и	233 й	234 к	235 л	236 м	237 н	238 о	239 п
F	1101	240 р	241 с	242 т	243 у	244 ф	245 х	246 ц	247 ч	248 ш	249 щ	250 ъ	251 ы	252 ь	253 э	254 ю	255 я

Рис. 2.11.Структура кода ANSI

Если мы спросим вас, какой комбинацией кодируется заглавная буква, то, посмотрев в таблицу, вы легко определите, что русская буква «Ю» имеет код 11011110, или в 16-ричном виде — DE, а в десятичном — 222.

Итак, этот **вариант ANSI-кодировки применяется в Windows для всех текстовых шрифтов, содержащих русские буквы**. При работе с компьютером многие пользователи используют как MS DOS, так и Windows, отличающиеся друг от друга, помимо прочего, системами кодирования символьной информации. Windows

содержит стандартные средства для перехода от одной кодировки к другой, которые часто выполняются автоматически и не доставляют особых забот пользователю. Тем не менее об этом необходимо помнить, чтобы осознанно анализировать те внештатные ситуации, которые могут возникнуть при использовании компьютера.

Ответить на вопросы и решить задания:

1. Как определить сколько байт содержит текст последнего абзаца параграфа в кодировке ANSI?
2. Для кодирования букв Р, С, Н, О, Г решили использовать двоичное представление чисел 0, 1, 2, 3 и 4 соответственно (с сохранением одного незначащего нуля в случае одноразрядного представления). Закодируйте последовательность букв НОСОРОГ таким способом и результат запишите восьмеричным кодом.
3. Измеряется температура воздуха, которая может быть целым числом от -30 до 34 градусов. Какое наименьшее количество бит необходимо, чтобы закодировать одно измеренное значение?
4. Определить максимальное количество страниц текста, содержащего по 80 символов в каждой строке и 64 строки на странице, которое может содержать файл, сохраненный на гибком магнитном диске объемом 10Кбайт. (кодировка ASCII)
5. Автоматическое устройство осуществило перекодировку информационного сообщения на русском языке, первоначально записанного в коде Windows-1251, в кодировку Unicode. При этом информационное сообщение увеличилось на 400 бит. Какова длина сообщения в символах?