跟我一起学编程系列课程：

第一部分 预备知识

第四章 常用编码规则

第二节

字符编码规则

- 字符编码规则
- 变形国标码

本节内容：字符编码规则

■字符编码规则：ASCII码、ANSI字符集、Unicode字符集。

■变形国标码：国标码是16位编码，高8位表示汉字符的区号，低8位表示汉字符的位号。

一、字符编码规则

计算机只能存储二进制数0和1，那么该如何表示字符呢？

■ASCII码字符

ASCII是美国国家标准信息交换码的英文缩写。每个字符由一个唯一的7位整数表示。只使用了每个字节的低7位，共128个字符，对应标准美国键盘上的字母和符号。剩下的最高位被各种计算机用来创建私有字符集，比如IBM PC的扩展ASCII码。

■ANSI字符集

如图4-1所示。美国国家标准委员会（ANSI）定义了一个8位的字符集，用于表示256个字符，前128个对应标准美国键盘上的字母和符号。后128个字符用于表示特殊字符，如其他语言字母表中的字母、重音符号、货币符号和分数等。MS-Windows Me/95/98使用ANSI字符集。

一、字符编码规则

ASCII 字符代码表 一

高四位 / 低四位		ASCII非打印控制字符										ASCII 打印字符																
		0000					0001					0010		0011		0100		0101		0110		0111						
		0					1					2		3		4		5		6		7						
		+进制	字符	ctrl	代码	字符解释	+进制	字符	ctrl	代码	字符解释	+进制	字符	+进制	字符	+进制	字符	+进制	字符	+进制	字符	+进制	字符	ctrl				
0000	0	0	BLANK NULL	^@	NUL	空	16	►	^P	DLE	数据链路转意	32		48	0	64	@	80	P	96	`	112	p					
0001	1	1	☺	^A	SOH	头标开始	17	◄	^Q	DC1	设备控制 1	33	!	49	1	65	A	81	Q	97	a	113	q					
0010	2	2	☻	^B	STX	正文开始	18	↕	^R	DC2	设备控制 2	34	"	50	2	66	B	82	R	98	b	114	r					
0011	3	3	♥	^C	ETX	正文结束	19	!!	^S	DC3	设备控制 3	35	#	51	3	67	C	83	S	99	c	115	s					
0100	4	4	♦	^D	EOT	传输结束	20	¶	^T	DC4	设备控制 4	36	$	52	4	68	D	84	T	100	d	116	t					
0101	5	5	♣	^E	ENQ	查询	21	§	^U	NAK	反确认	37	%	53	5	69	E	85	U	101	e	117	u					
0110	6	6	♠	^F	ACK	确认	22	■	^V	SYN	同步空闲	38	&	54	6	70	F	86	V	102	f	118	v					
0111	7	7	•	^G	BEL	震铃	23	↨	^W	ETB	传输块结束	39	'	55	7	71	G	87	W	103	g	119	w					
1000	8	8	◘	^H	BS	退格	24	↑	^X	CAN	取消	40	(	56	8	72	H	88	X	104	h	120	x					
1001	9	9	○	^I	TAB	水平制表符	25	↓	^Y	EM	媒体结束	41	)	57	9	73	I	89	Y	105	i	121	y					
1010	A	10	◙	^J	LF	换行/新行	26	→	^Z	SUB	替换	42	*	58	:	74	J	90	Z	106	j	122	z					
1011	B	11	♂	^K	VT	竖直制表符	27	←	^[	ESC	转意	43	+	59	;	75	K	91	[	107	k	123	{					
1100	C	12	♀	^L	FF	换页/新页	28	∟	^\	FS	文件分隔符	44	,	60	<	76	L	92	\	108	l	124	|					
1101	D	13	♪	^M	CR	回车	29	↔	^]	GS	组分隔符	45	-	61	=	77	M	93	]	109	m	125	}					
1110	E	14	♫	^N	SO	移出	30	▲	^6	RS	记录分隔符	46	.	62	>	78	N	94	^	110	n	126	~					
1111	F	15	☼	^O	SI	移入	31	▼	^-	US	单元分隔符	47	/	63	?	79	O	95	_	111	o	127	Δ	^Back space				

注：表中的ASCII字符可以用:ALT + "小键盘上的数字键" 输入

图4-1 128字符ASCII表

一、字符编码规则

提示

1.注意观察ASCII表，常用的ASCII字符对应的16进制数值需要熟记于心。如响铃字符的ASCII值07H，退格字符08H，TAB制表符09H，换行符0AH，回车符0DH，空格符20H，CTRL+B 的ASCII值02H，CTRL+C的ASCII值03H。

2.大写字母“A~Z”的ASCII值为41H~5AH，小写字母“a~z”的ASCII值为61H~7AH，数字符“0~9”的ASCII值为30H~39H。

3.可见字符的ASCII值从20H开始，到7EH结束。

一、字符编码规则

■Unicode标准

计算机软件中表示各种不同国家的语言有上百种编码方案，比较混乱。由此创建Unicode标准作为定义字符和符号的统一方法。Unicode标准定义了所有主要语言中使用的字母、符号及标点。Unicode有三种编码形式：

●**UTF-8**：ASCII码在UTF-8编码中占用一个字节，其字节值和ASCII码值相同。所有Unicode字符都可以用一种变长的编码系统表示。

●**UTF-16**：用于访问效率和存储空间并重的环境中。例如：Windows NT/2000/XP使用UTF-16编码，每个字符用16个二进制数据位编码。

●**UTF-32**：用于不太关心存储空间的环境。每个字符都使用32个二进制数据位编码，宽度固定。

二、变形国标码

有了ASCII码，计算机可以处理数字、字母等字符，但是并不能处理汉字符。

我们国家1981年5月对六千多个常用汉字制定了交换码的国家标准，即GB2312-80《信息交换用汉字编码字符集—基本集》。该标准规定了汉字信息交换的基本汉字符和一般图形字符，共计7445个，其中汉字分成两个等级共计6763个。该标准同时也给定了它们的二进制编码，即国标码。后来的字符集GBK收录20912个汉字，最新的字符集GB18030收录27533个汉字。

国标码是16位编码，高8位表示汉字符的区号，低8位表示汉字符的位号。实际上，为了给汉字符编码，该标准把代码表分成94个区，每个区94个位。区号和位号都从21H开始。一级汉字安排在30H区至57区，二级汉字安排在58H至77区。

机内码是汉字在计算机内部使用的编码。汉字的机内码采用变形国标码，其变换方法为：变形国标码=国标码+8080H，即将两个字节的最高位由0改1，其余7位不变。

区位码转换为国标码的方式：国标码是由区位码稍作转换得到。先将十进制区码和位码转换为十六进制的区码和位码，再将这个代码加上2020H，就得到国标码。

二、变形国标码

有了ASCII码，计算机可以处理数字、字母等字符，但是并不能处理汉字符。

我们国家1981年5月对六千多个常用汉字制定了交换码的国家标准，即GB2312-80《信息交换用汉字编码字符集—基本集》。该标准规定了汉字信息交换的基本汉字符和一般图形字符，共计7445个，其中汉字分成两个等级共计6763个。该标准同时也给定了它们的二进制编码，即国标码。后来的字符集GBK收录20912个汉字，最新的字符集GB18030收录27533个汉字。

国标码是16位编码，高8位表示汉字符的区号，低8位表示汉字符的位号。实际上，为了给汉字符编码，该标准把代码表分成94个区，每个区94个位。区号和位号都从21H开始。一级汉字安排在30H区至57区，二级汉字安排在58H至77区。

机内码是汉字在计算机内部使用的编码。汉字的机内码采用变形国标码，其变换方法为：变形国标码=国标码+8080H，即将两个字节的最高位由0改1，其余7位不变。

区位码转换为国标码的方式：国标码是由区位码稍作转换得到。先将十进制区码和位码转换为十六进制的区码和位码，再将这个代码加上2020H，就得到国标码。

二、变形国标码

举例

某汉字区号为34，位号为56。区位码：3456。
34 = 0010 0010B = 22H
56 = 0011 1000B = 38H
国标码：2238H + 2020H = 4258H
变形国标码：4258H + 8080H = C2D8H

练习

熟悉常用的各种编码规则，仔细观察编码规则中的规律。

昆山爱达人信息技术有限公司

视频提供

视频录制：编程达人

联系电话：
0512-57882866

官网地址：
www.bcdaren.com

联系公众号：
昆山爱达人

联系QQ：
1250121864

编程达人APP：