字符编码

二进制

山东肥城

山东肥城

字节

字符

a123212 哈 😄

ASCII

Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	C
000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	
001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	
002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	
003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	
004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	&#100	
005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	&#101	
006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	&#102	
007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	&#103	
010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	&#104	
011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	&#105	
012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	&#106	
013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	&#107	
014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	&#108	
015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	&#109	
016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	&#110	
017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	&#111	
020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	&#112	
021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	&#113	
022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	&#114	
023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	&#115	
024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	&#116	
025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	&#117	
026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	&#118	
027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	&#119	
030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	&#120	
031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	&#121	
032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	&#122	
033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	&#123	
034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174	&#124	
035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	&#125	
036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	&#126	
037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127	

144	É	160	á	176	▒	192	└	208	╨	224	α	240	≡
145	æ	161	í	177	▒	193	┴	209	╤	225	ß	241	±
146	Æ	162	ó	178	▓	194	┬	210	╥	226	Γ	242	≥
147	ô	163	ú	179	│	195	├	211	╙	227	π	243	≤
148	ö	164	ñ	180	┤	196	─	212	╘	228	Σ	244	⌠
149	ò	165	Ñ	181	╡	197	┼	213	╒	229	σ	245	⌡
150	û	166	ª	182	╢	198	╞	214	╓	230	µ	246	÷
151	ù	167	º	183	╖	199	╟	215	╫	231	τ	247	≈
152	ÿ	168	¿	184	╕	200	╚	216	╪	232	Φ	248	°
153	Ö	169	⌐	185	╣	201	╔	217	┘	233	Θ	249	∙
154	Ü	170	¬	186	║	202	╩	218	┌	234	Ω	250	·
155	¢	171	½	187	╗	203	╦	219	█	235	δ	251	√
156	£	172	¼	188	╝	204	╠	220	▄	236	∞	252	ⁿ
157	¥	173	¡	189	╜	205	═	221	▌	237	φ	253	²
158	₧	174	«	190	╛	206	╬	222	▐	238	ε	254	■
159	ƒ	175	»	191	┐	207	╧	223	▀	239	∩	255	

混乱时代

你会如何设计?

- 字符表（Character repertoire）

- 给字符表里的抽象字符编上一个数字，也就是字符集合到一个整数集合的映射。这种映射称为编码字符集（CCS:Coded Character Set）

- 将CCS里字符对应的整数转换成有限长度的比特值，便于以后计算机使用一定长度的二进制形式表示该整数。这个对应关系被称为字符编码表（CEF:Character Encoding Form）

Unicode

UTF-8 UTF-16 GBK

Unicode Transformation Format

```
Unicode符号范围 | UTF-8编码方式
(十六进制) |  (二进制)
---------------------+-----------------------------------------
0000 0000-0000 007F | 0xxxxxxx
0000 0080-0000 07FF | 110xxxxx 10xxxxxx
0000 0800-0000 FFFF | 1110xxxx 10xxxxxx 10xxxxxx
0001 0000-0010 FFFF | 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
```

“严” unicode

4E25 (100111000100101)

从“严”的最后一个二进制位开始，依次从后向前填入格式中的x，多出的位补0。这样就得到了，“严”的UTF-8编码>是“11100100 10111000 10100101”，转换成十六进制就是0xE4B8A5。

Byte order mark

0010 0000 1010 1100

字体

1. 字体是用来显示字符的，将不可见的内存中的字符显示为可见的图形。

2. 日常使用的字符是由特定字符集（比如 Unicode）定义的。

3. 并非所有字体都遵循你用的字符集。

iconv -f UTF8 -t

GB18030 a.csv >b.csv

QA