Coding Systems (2.1.2)

Character Representation using ASCII and UNICODE. Collating Sequence

Introduction

Representation of Text Using an 8-Bit Coding System

A bit is the smallest unit in a computer system and it can only store the binary value of 0 or 1. So, with a single bit we are able to represent only 2 different values.

Data such as text, pictures and videos stored in a computer system is stored as a sequence of binary digits. Using a single bit is not enough to store these types of files. So if we're going to store a large quantity of data in a computer system, more bits are required. Every time a bit is combined with another bit, more different combinations can be stored. So if we have two bits which can be used to store data, there are a total of 2^2=4 different combinations.

Combination	Value of Bit 1	Value of Bit 2
1	0	0
2	0	1
3	1	0
4	1	1

The following tables shows all the different combinations possible using from 1 to 8 bits.

Number of Bits	Calculation	Number of Combinations
1	2^1	2
2	2^2	4
3	2^3	8
4	2^4	16
5	2^5	32
6	2^6	64
7	2^7	128
8	2^8	256

ASCII

In the early days of computing, computers were mainly used for calculations. Then scientists realised that computers can also be used to store and manipulate text. But there was a problem; a computer can only understand binary signals so a plan had to be created to be able to represent characters as binary signals.

For this purpose ASCII (American Standard Code for Information Interchange) was developed. In ASCII every single character is assigned a binary code, so whenever a character is stored on the computer system, it is stored as a group of binary numbers. This system, also allowed computers to communicate with each other, hence files containing text can be copied from one computer to another.

The first ASCII system used a series of 7 Bits. This meant that 2^7=128 different characters can be used. These 128 characters are represented in the following table (Note that a decimal number is used instead of a binary number for easier reading).

Code	Char
0	NUL
1	SOH
2	STX
3	ETX
4	EOT
5	ENQ
6	ACK
7	BEL
8	BS
9	TAB
10	LF
11	VT
12	FF
13	CR
14	SO
15	SI
16	DLE
17	DC1
18	DC2
19	DC3
20	DC4
21	NAK
22	SYN
23	ETB
24	CAN
25	EM
26	SUB
27	ESC
28	FS
29	GS
30	RS
31	US

Code	Char
32	Space
33	!
34	“
35	#
36	$
37	%
38	&
39	‘
40	(
41	)
42	*
43	+
44	,
45	-
46	.
47	/
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	?

Code	Char
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
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	_

Code	Char
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
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	DEL

The first 32 characters were only used in the beginning for transmission purposes. Nowadays these are not used anymore. After this was developed it was noted that it consisted only of English alphabet, so an extra bit was added to accommodate more languages, graphics and mathematical symbols. By adding an extra bit 128 more characters can be stored since 2^8=256.

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

Try It: *You can check the symbol on your computer by typing in its decimal code. Open Notepad (Start > Programs > Accessories > Notepad). First make sure that NUM LOCK is switched on, then press and hold ALT (on the left hand side of the keyboard) and from the Numpad enter the 3 digits that you want (example 234 for Ω). Release the ALT, and the symbol should appear.*

It is very important to use a standard system because otherwise the same text would look different on another computer. Let's assume that one computer uses ASCII representation and another computer uses XYZ system (it's invented!) to represent characters. If the codes do not match, the text will look different on another computer. Let's assume that the word HELLO is typed on a computer using ASCII code. Using the ASCII table we obtain the following:

Letter	H	E	L	L	O
Decimal Code	72	69	76	76	79
Binary Code	1001000	1000101	1001100	1001100	1001111

Now this is an example of the XYZ coding system:

Code	Char
1	A
2	B
3	C
4	D
5	E
6	F
...	

Code	Char
69	!
70	“
71	£
72	$
73	%
74	^
75	&

Code	Char
76	*
77	(
78	)
79	-
80	_
81	+
...	

So if we look up the same binary codes in this coding system we end up with a very different text message!

Binary Code	1001000	1000101	1001100	1001100	1001111
Decimal	72	69	76	76	79
Letter	$	!	*	*	-

So, when the second computer with the XYZ representation system receives the text file from the computer with the ASCII representation system instead of HELLO, one will see $!**-, which is completely wrong! That’s why it’s important to have a standard system so that the text remains consistent on all computers.

Since nowadays computers are used across the world a new system is being used which is UNICODE. UNICODE still uses the same 256 characters in order to be compatible with ASCII but then it uses a different system and in certain cases with more bits, to represent all the different characters of all the languages in the world! A list of UNICODE characters can be accessed on *http://en.wikipedia.org/wiki/List_of_Unicode_characters*.

Collating Sequence

The character codes assigned to the characters in ASCII are very important since they can be used to sort letters and distinguishing from upper to lower case letters. For instance if the letters C, V and B have to be sorted, they can be easily sorted by looking at their ASCII code and use it to sort text in ascending (A→Z) or descending order (Z→A).

Character	Code
C	67
V	86
B	66

If we want to sort C,V and B in ascending order, they can be easily sorted by first sorting the code, therefore 66, 67 and 86, hence we will obtain the sorted letters B,C and V!

***