Lecture2-4 国际化 Internationalization

1. 介绍

- 国际化是指制作可以根据不同的地点和语言进行输入和输出的程序
- 根据所编码字符的数量，字符编码可以采用不同的形式，编码的字符数与每个表示的长度有直接关系，通常用字节数来衡量，**要编码更多的字符本质上意味着需要更长的二进制表示**

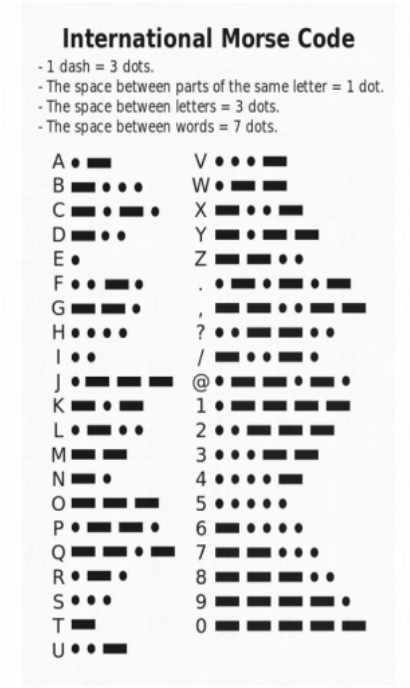
2. 常见的编码

Base64 encoding

Index	Binary	Char	Index	Binary	Char	Index	Binary	Char	Index	Binary	Char
0	000000	A	16	010000	Q	32	100000	g	48	110000	w
1	000001	B	17	010001	R	33	100001	h	49	110001	x
2	000010	C	18	010010	S	34	100010	i	50	110010	y
3	000011	D	19	010011	T	35	100011	j	51	110011	z
4	000100	E	20	010100	U	36	100100	k	52	110100	0
5	000101	F	21	010101	V	37	100101	l	53	110101	1
6	000110	G	22	010110	W	38	100110	m	54	110110	2
7	000111	H	23	010111	X	39	100111	n	55	110111	3
8	001000	I	24	011000	Y	40	101000	o	56	111000	4
9	001001	J	25	011001	Z	41	101001	p	57	111001	5
10	001010	K	26	011010	a	42	101010	q	58	111010	6
11	001011	L	27	011011	b	43	101011	r	59	111011	7
12	001100	M	28	011100	c	44	101100	s	60	111100	8
13	001101	N	29	011101	d	45	101101	t	61	111101	9
14	001110	O	30	011110	e	46	101110	u	62	111110	+
15	001111	P	31	011111	f	47	101111	v	63	111111	/

Character Encoding

- 国际摩斯码编码了 26 英语字母 A 到 Z，一些非英语字母，阿拉伯数字和一组小的标点符号和程序信号，大写字母和小写字母之间没有区别



ASCII 字符集

Value	Char	Value	Char	Value	Char	Value	Char	Value	Char
32	*space*	51	3	70	F	89	Y	108	l
33	!	52	4	71	G	90	Z	109	m
34	"	53	5	72	H	91	[	110	n
35	#	54	6	73	I	92	\	111	o
36	$	55	7	74	J	93	]	112	p
37	%	56	8	75	K	94	^	113	q
38	&	57	9	76	L	95	_	114	r
39	'	58	:	77	M	96	'	115	s
40	(	59	;	78	N	97	a	116	t
41	)	60	<	79	O	98	b	117	u
42	*	61	=	80	P	99	c	118	v
43	+	62	>	81	Q	100	d	119	w
44	'	63	?	82	R	101	e	120	x
45	–	64	@	83	S	102	f	121	y
46	.	65	A	84	T	103	g	122	z
47	/	66	B	85	U	104	h	123	{
48	0	67	C	86	V	105	i	124	|
49	1	68	D	87	W	106	j	125	}
50	2	69	E	88	X	107	k	126	~

Unicode

- 不同的编码方案单独开发并在当地的地理环境中实践开始变得具有挑战性
- 不难理解，虽然编码很重要，但解码对于理解表示也同样重要。只有在广泛使用一致或兼容的编码方案时，这才有可能实现
- 这一挑战催生了一种被称为 **Unicode** 的单一编码标准，它可以**容纳世界上所有可能的字符**，这包括那些正在使用的字符，甚至那些已经失效的字符
- 因此，如何将这些编码点编码成位取决于 Unicode 内部的特定编码方案

Value	Character	Source
1071	Я	Russian (Cyrillic)
3593	้	Thai
5098	Ꮩ	Cherokee
8478	℞	Letterlike Symbols
8652	⇌	Arrows
10287	⡯	Braille
13407	侦	Chinese/Japanese/Korean (Common)

UTF-32

UTF-32 是一种 Unicode 编码方案，使用四个字节来表示定义的每个代码点，显然，每个字符使用 4 个字节会降低空间效率

- 例如：“UTF-32” 中的 “T” 为 00000000 00000000 00000000” 01010100 （前 3 个字节是不必要的）

UTF-8

可变长度编码：UTF-8 是另一种使用可变长度字节进行编码的 Unicode 编码方案。虽然它通常使用单个字节来编码字符，但如果需要，它可以使用更多的字节，从而节省空间

- 例如：“T” 在 “UTF-8” 中是 01010100
- 但 “語” 在 “UTF-8”是中是 11101000 10101010 11101000

3. 常见的编码头

BOM

BOM (byte-order mark) 文件编码头，即字节顺序标记

它是插入到以 UTF-8, UTF16 或 UTF-32 编码文件开头的特殊标记

用来标记多字节编码文件的**编码类型和字节顺序 big-endian 或 little-endian**

一般用来识别文件的编码类型

- 根据字节序的不同
  - UTF-16 可以被实现为 UTF-16LE 或 UTF-1B6E
  - UTF-32 可以被实现为 UTF-32LE 或 UTF-32BE

Unicode编码	UTF-16LE	UTF-16BE	UTF32-LE	UTF32-BE
0x006C49	49 6C	6C 49	49 6C 00 00	00 00 6C 49
0x020C30	43 D8 30 DC	D8 43 DC 30	30 0C 02 00	00 02 0C 30