

2. 数据的表示与存储

2.7 文本信息处理



2. 数据的表示与存储

2.7 典型信息表示

01 文本信息与ASCII编码

02 典型国际编码

03 文本乱码的根源

2



2. 数据的表示与存储

2.7.1、二进制数据的解读

00	1	1	0	0	0	0	00	1	1	0	0	0	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
----	---	---	---	---	---	---	----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

e f .

... 1024

32 16

21

32位的无符号数：

32位的有符号数：

32位的单精度浮点数：

4个单字节数：

0x30, 0x31, 0x32, 0x33
'0', '1', '2', '3'



2. 数据的表示与存储

2.7.1、文本信息

ASCII编码

Unicode编码

Utf-8编码

GB2312编码

GBK编码

GB18030编码

•Characters, not glyphs: 字符，而不是字形。

位

样子

字体

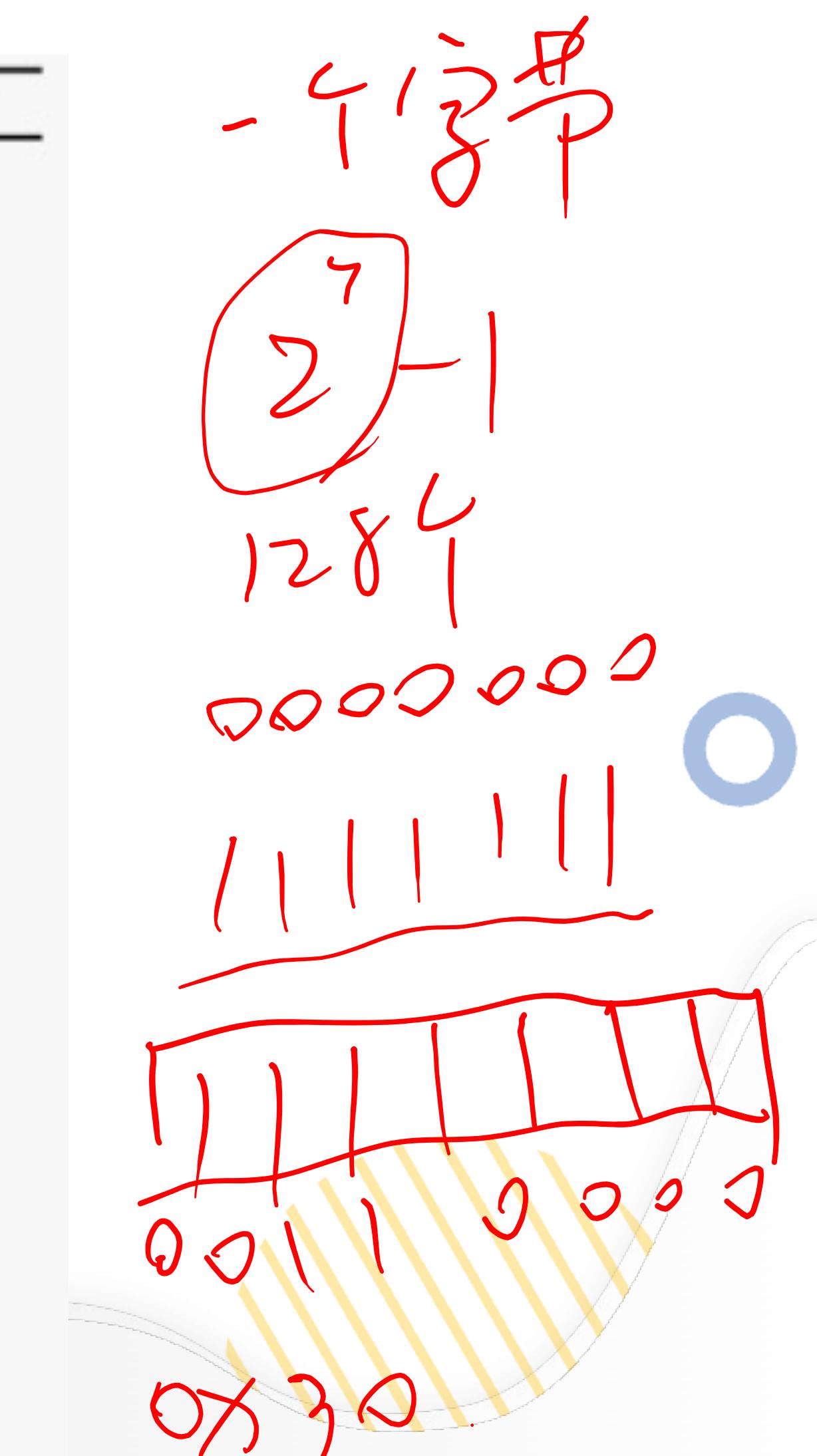


2. 数据的表示与存储

2.7.1、ASCII编码表<https://zh.wikipedia.org/wiki/ASCII>

Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char
0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	'
1	1	1		33	21	41	!	65	41	101	A	97	61	141	a
2	2	2		34	22	42	"	66	42	102	B	98	62	142	b
3	3	3		35	23	43	#	67	43	103	C	99	63	143	c
4	4	4		36	24	44	\$	68	44	104	D	100	64	144	d
5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
7	7	7		39	27	47	'	71	47	107	G	103	67	147	g
8	8	10		40	28	50	(72	48	110	H	104	68	150	h
9	9	11		41	29	51)	73	49	111	I	105	69	151	i
10	A	12		42	2A	52	*	74	4A	112	J	106	6A	152	j
11	B	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	l
13	D	15		45	2D	55	-	77	4D	115	M	109	6D	155	m
14	E	16		46	2E	56	.	78	4E	116	N	110	6E	156	n
15	F	17		47	2F	57	/	79	4F	117	O	111	6F	157	o
16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
19	13	23		51	33	63	3	83	53	123	S	115	73	163	s
20	14	24		52	34	64	4	84	54	124	T	116	74	164	t
21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
23	17	27		55	37	67	7	87	57	127	W	119	77	167	w
24	18	30		56	38	70	8	88	58	130	X	120	78	170	x
25	19	31		57	39	71	9	89	59	131	Y	121	79	171	y
26	1A	32		58	3A	72	:	90	5A	132	Z	122	7A	172	z
27	1B	33		59	3B	73	;	91	5B	133	[123	7B	173	{
28	1C	34		60	3C	74	<	92	5C	134	\	124	7C	174	
29	1D	35		61	3D	75	=	93	5D	135]	125	7D	175	}
30	1E	36		62	3E	76	>	94	5E	136	^	126	7E	176	-
31	1F	37		63	3F	77	?	95	5F	137	-	127	7F	177	

7bit 海贼宝藏
专注IT教育在线学习平台





2. 数据的表示与存储

2.7.1、ASCII编码特点

字符集：ASCII编码定义了128个字符，包括大小写英文字母、数字0-9、标点符号、控制字符（如换行、回车）以及一些其他符号。

编码方式：每个ASCII字符被分配了一个从0到127的数字编码。例如，大写字母“A”的ASCII编码是65，小写字母“a”的编码是97。

二进制表示：ASCII字符使用7位二进制数来表示，这意味着每个字符可以用一个字节（通常是8位）中的7位来表示。例如，字母“A”的二进制表示是01000001。- 0x41

控制字符：ASCII中的控制字符用于文本控制。例如，字符编码为10的“换行”（LF）用于表示文本行的结束。

兼容性：由于ASCII仅使用7位，它可以在8位字节的系统中使用而不会引起兼容性问题。这使得ASCII在早期计算机系统中非常流行。

局限性：ASCII仅能表示英语字母和基本符号，它无法表示其他语言的字符。

0xa

8位机



2. 数据的表示与存储

2.7.2、Unicode编码

1. 全球一致性：Unicode旨在为全球范围内所有的字符提供一个唯一的编码，解决了之前不同编码系统不兼容的问题。

2. 广泛的字符集：Unicode涵盖了几乎所有现存和历史上的字符系统，包括拉丁文、希腊文、阿拉伯文、汉字、日文假名、韩文、印度文字等。

3. 不同的编码方案：Unicode有几种实现方式，最常见的是UTF-8、UTF-16和UTF-32。这些编码方案以不同的方式存储Unicode字符。

1. **UTF-8**：使用1到4个字节来表示每个字符，兼容ASCII，是最常用的Web编码。

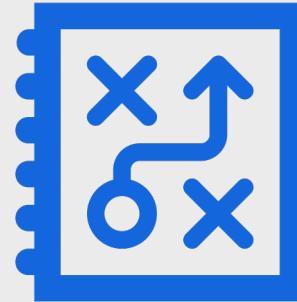
2. **UTF-16**：使用2个或4个字节表示每个字符，常用于现代操作系统和环境中。

3. **UTF-32**：每个字符固定使用4个字节，简化了字符的处理，但占用空间较多。

4. **代码点**：在Unicode中，每个字符分配一个唯一的编号，称为“代码点”。这些代码点通常表示为“U+”后跟一个十六进制数，例如，英文字符“A”的Unicode代码点是U+0041。

5. 扩展性：Unicode定期更新，以包含新的字符和符号。

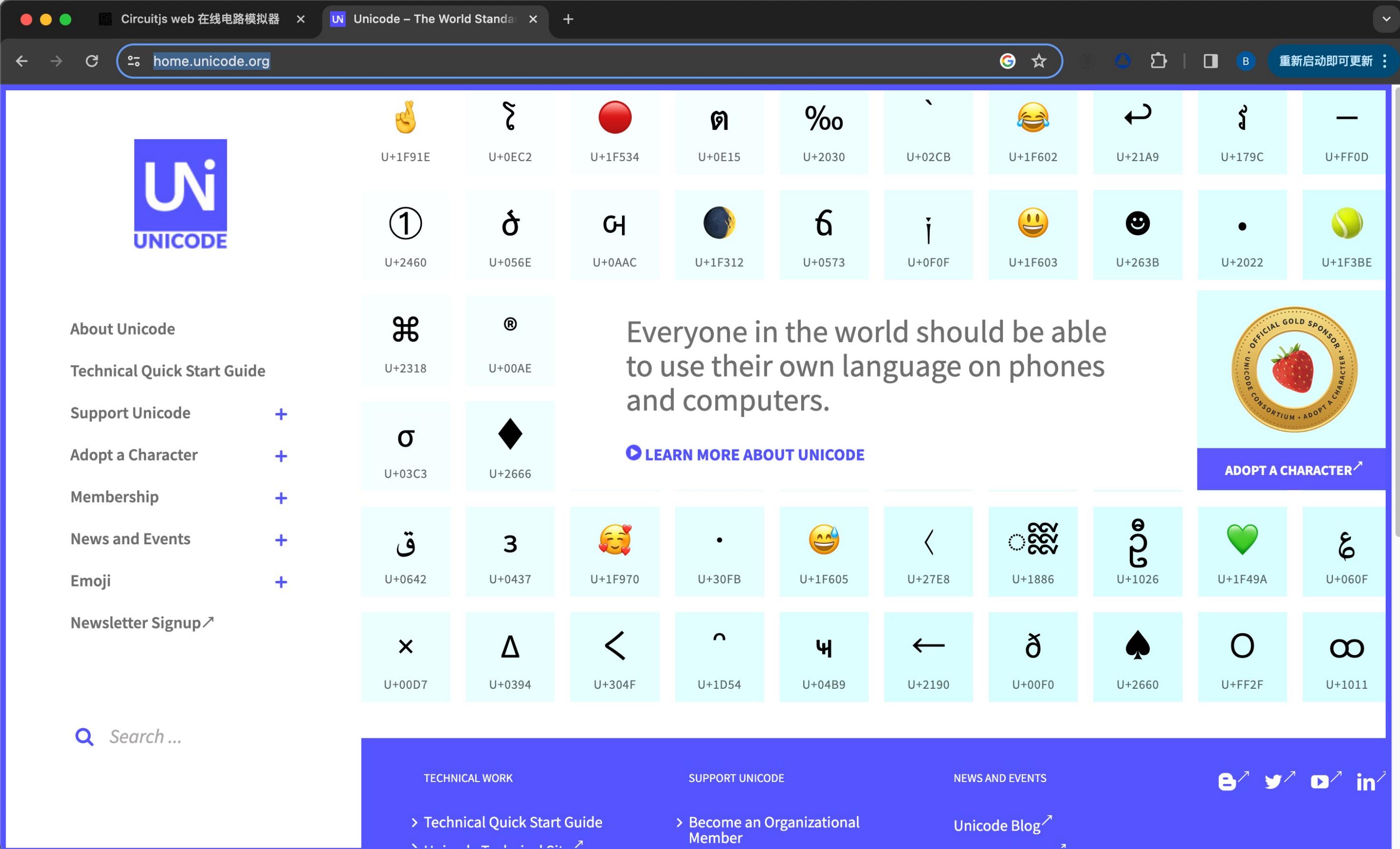
9~12
回向单刀
A



2. 数据的表示与存储

2.7.2、Unicode编码

<https://home.unicode.org/>



The screenshot shows the homepage of the Unicode website (<https://home.unicode.org/>). The page features a large grid of various Unicode characters at the top, including symbols like a yellow hand, a red ball, a smiley face, and a strawberry. Below this, there's a sidebar with links to "About Unicode", "Technical Quick Start Guide", "Support Unicode", "Adopt a Character", "Membership", "News and Events", and "Emoji". A search bar is located at the bottom left. The main content area includes a "Learn More About Unicode" button and a "Strawberry" badge for being an "OFFICIAL GOLD SPONSOR" of the "UNICODE CONSORTIUM - ADOPT A CHARACTER". At the bottom, there are links for "Technical Quick Start Guide", "Become an Organizational Member", and "Unicode Blog".



2. 数据的表示与存储

2.7.2、Unicode编码

<https://zh.wikipedia.org/wiki/Unicode>

zh.wikipedia.org/wiki/Unicode字符平面映射

维基百科
自由的百科全书

搜索维基百科 搜索

创建账号 登录 ...

中文维基百科Facebook粉丝专页正式上线，邀请大家一同关注。 [关闭]

Unicode字符平面映射 [\[编辑\]](#)

文 13种语言 ▾

目录 [\[隐藏\]](#)

条目 讨论 汉 漢 大陆简体 ▾

阅读 编辑 查看历史 工具 ▾

维基百科，自由的百科全书

目前的Unicode字符分为17组编排，每组称为平面（Plane），而每平面拥有65536（即 2^{16} ）个代码点。然而目前只用了少数平面。

平面	始末字符值	中文名称	英文名称
0号平面	U+0000 – U+FFFF	基本多文种平面	Basic Multilingual Plane, 简称BMP
1号平面	U+10000 – U+1FFFF	多文种补充平面	Supplementary Multilingual Plane, 简称SMP
2号平面	U+20000 – U+2FFFF	表意文字补充平面	Supplementary Ideographic Plane, 简称SIP
3号平面	U+30000 – U+3FFFF	表意文字第三平面	Tertiary Ideographic Plane, 简称TIP
4号平面 至 13号平面	U+40000 – U+DFFFF	(尚未使用)	
14号平面	U+E0000 – U+EFFFF	特别用途补充平面	Supplementary Special-purpose Plane, 简称SSP
15号平面	U+F0000 – U+FFFFF	保留作为私人使用区（A区） [1]	Private Use Area-A, 简称PUA-A
16号平面	U+100000 – U+10FFFF	保留作为私人使用区（B区） [1]	Private Use Area-B, 简称PUA-B

要有更详细的描述，请参阅：[基本多文种平面与辅助平面](#)。

[基本多文种平面 \[\\[编辑\\]\]\(#\)](#)

填答简短问卷，帮助我们改善维基 ×



2. 数据的表示与存储

2.7.2、Unicode编码

<http://www.chi2ko.com/tool/CJK.htm>

字体编辑用中日韩汉字Unicode编码表

编著：资深中韩翻译金圣镇 金圣镇

Copyright © 2004 资深中韩翻译金圣镇 www.chi2ko.com All Rights Reserved. 版权所有 不得转载

一	丁	ㄅ	七	上	ㄒ	ㄈ	万	丈	三	上	下	ㄉ	不	与	ㄅ
4E00	4E01	4E02	4E03	4E04	4E05	4E06	4E07	4E08	4E09	4E0A	4E0B	4E0C	4E0D	4E0E	4E0F
ㄅ	丑	ㄋ	专	且	丕	世	丂	丘	丙	业	丛	东	丝	丞	丢
4E10	4E11	4E12	4E13	4E14	4E15	4E16	4E17	4E18	4E19	4E1A	4E1B	4E1C	4E1D	4E1E	4E1F
丂	丂	丢	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂
丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂
丰	卯	串	弗	临	擧	丶	丶	丸	丹	为	主	丂	丽	举	丂
4E30	4E31	4E32	4E33	4E34	4E35	4E36	4E37	4E38	4E39	4E3A	4E3B	4E3C	4E3D	4E3E	4E3F
丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂
丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂	丂
乐	乚	兵	兵	乔	𠂔	乖	乘	乘	乙	乚	乚	乚	九	乞	也
4E50	4E51	4E52	4E53	4E54	4E55	4E56	4E57	4E58	4E59	4E5A	4E5B	4E5C	4E5D	4E5E	4E5F
习	乡	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊
4E60	4E61	4E62	4E63	4E64	4E65	4E66	4E67	4E68	4E69	4E6A	4E6B	4E6C	4E6D	4E6E	4E6F
买	乱	姿	乳	哲	乳	惠	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊
4E70	4E71	4E72	4E73	4E74	4E75	4E76	4E77	4E78	4E79	4E7A	4E7B	4E7C	4E7D	4E7E	4E7F
龜	乾	亂	𠂊	亂	丂	了	丂	予	争	爭	事	事	二	亍	亏
4E80	4E81	4E82	4E83	4E84	4E85	4E86	4E87	4E88	4E89	4E8A	4E8B	4E8C	4E8D	4E8E	4E8F
亏	云	互	元	五	井	三	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊	𠂊
4E90	4E91	4E92	4E93	4E94	4E95	4E96	4E97	4E98	4E99	4E9A	4E9B	4E9C	4E9D	4E9E	4E9F
丂	亡	亢	宀	交	亥	亦	产	亨	亩	𠂊	享	京	亭	亮	富
4EA0	4EA1	4EA2	4EA3	4EA4	4EA5	4EA6	4EA7	4EA8	4EA9	4EAA	4EAB	4EAC	4EAD	4EAE	4EAF
京	匱	亲	毫	亮	襄	亶	廉	諱	亹	人	亼	𠂊	𠂊	𠂊	𠂊
4FB0	4FB1	4FB2	4FB3	4FB4	4FB5	4FB6	4FB7	4FB8	4FB9	4FBA	4FBB	4FBC	4FBD	4FBE	4FBF



2. 数据的表示与存储

2.7.2、UTF-8编码特点

1到4个字节，包含控制码和字符码。

unicode和utf-8的映射关系

110xxxxx 10xxxxxx

0x00 89
0000 0001 0001 00 |
| 0000 0001 0001 00 |

Unicode中的字符集范围	Utf-8编码方式	说明
0000 0000 ~ 0000 007F	0xxxxxxxx	完全兼容ASCII
0000 0080 ~ 0000 07FF	110xxxxx 10xxxxxx	110表示需要两个字节
0000 0800 ~ 0000 FFFF	<u>1110xxxx</u> 10xxxxxx 10xxxxxx	1110表示需要三个字节
0001 0000 ~ 0010 FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx	表示需要四个字节



2. 数据的表示与存储

2.7.2、UTF-8编码特点

1. 兼容ASCII: UTF-8的设计使得所有的ASCII字符都可以使用一个字节表示，且与ASCII编码相同。这意味着ASCII文本无需修改即可作为UTF-8文本使用。

2. 变长编码: UTF-8使用1到4个字节来表示每个Unicode字符。字节的数量取决于字符的Unicode码点：

- 1.U+0000 至 U+007F (基本的拉丁字母、数字和符号) 用一个字节表示。
- 2.U+0080 至 U+07FF (包括拉丁字母补充、希腊字母、西里尔字母等) 用两个字节表示。
- 3.U+0800 至 U+FFFF (包括大多数其它语言的字符和符号) 用三个字节表示。
- 4.U+10000 至 U+10FFFF (包括少数语言和特殊符号) 用四个字节表示。

3. 前导字节和连续字节: UTF-8中的每个字节都可以通过其最高位 (bit) 来识别。单字节字符的最高位是0，而多字节字符的首字节从110到1110开始，随后字节都以10开始。

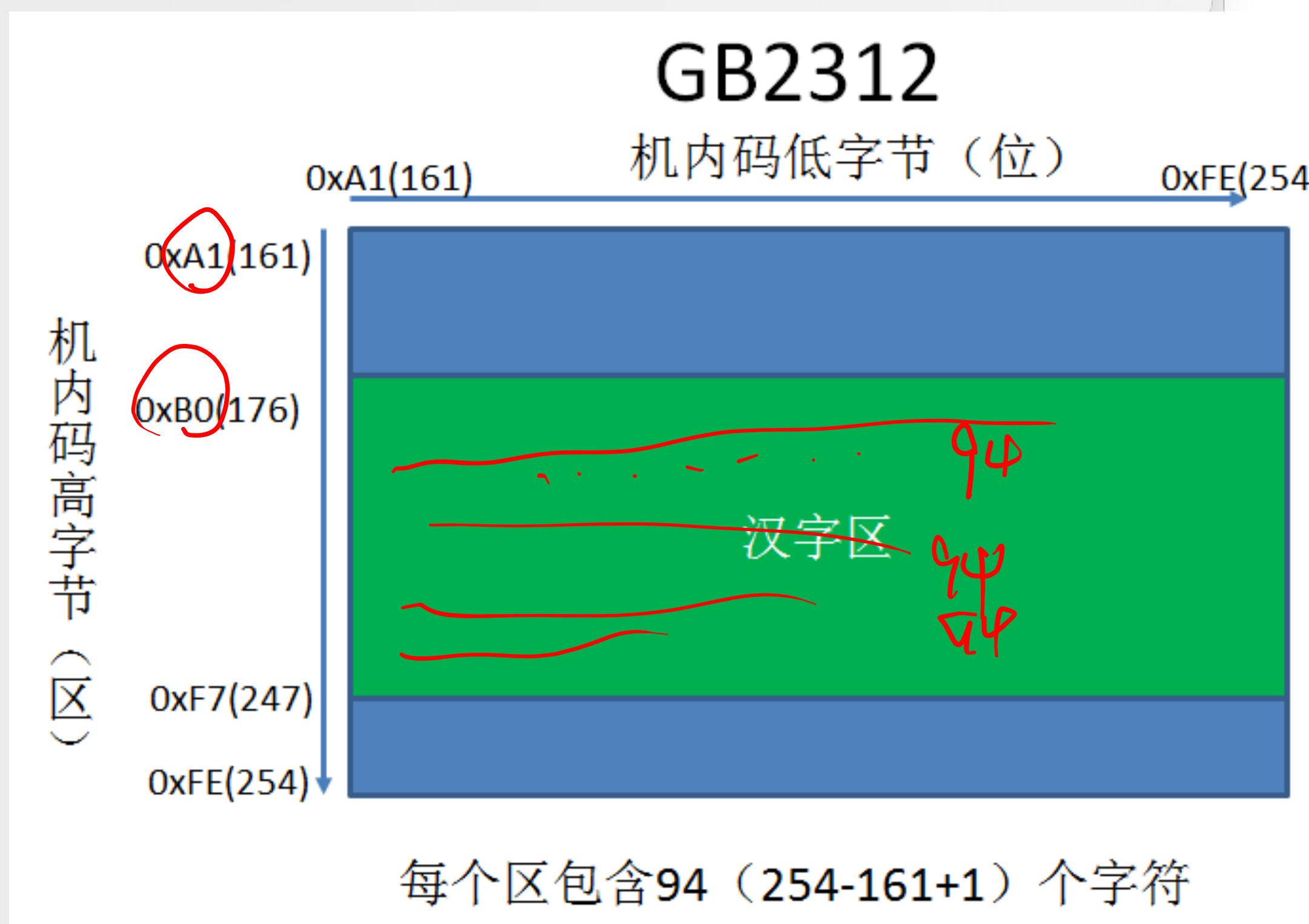
4. 自同步: UTF-8编码的一个重要特性是自同步性，这意味着从任何字节开始，都可以确定它是一个字符的开始还是一个字符中间的字节。这有助于文本搜索和恢复。



2. 数据的表示与存储

2.7.2、GB2312中国最早使用的汉字编码

GB2312 (~~区位码~~)：1980年发布，收录简化汉字及符号、字母、日文假名等共7445个图形字符，其中汉字占6763个。

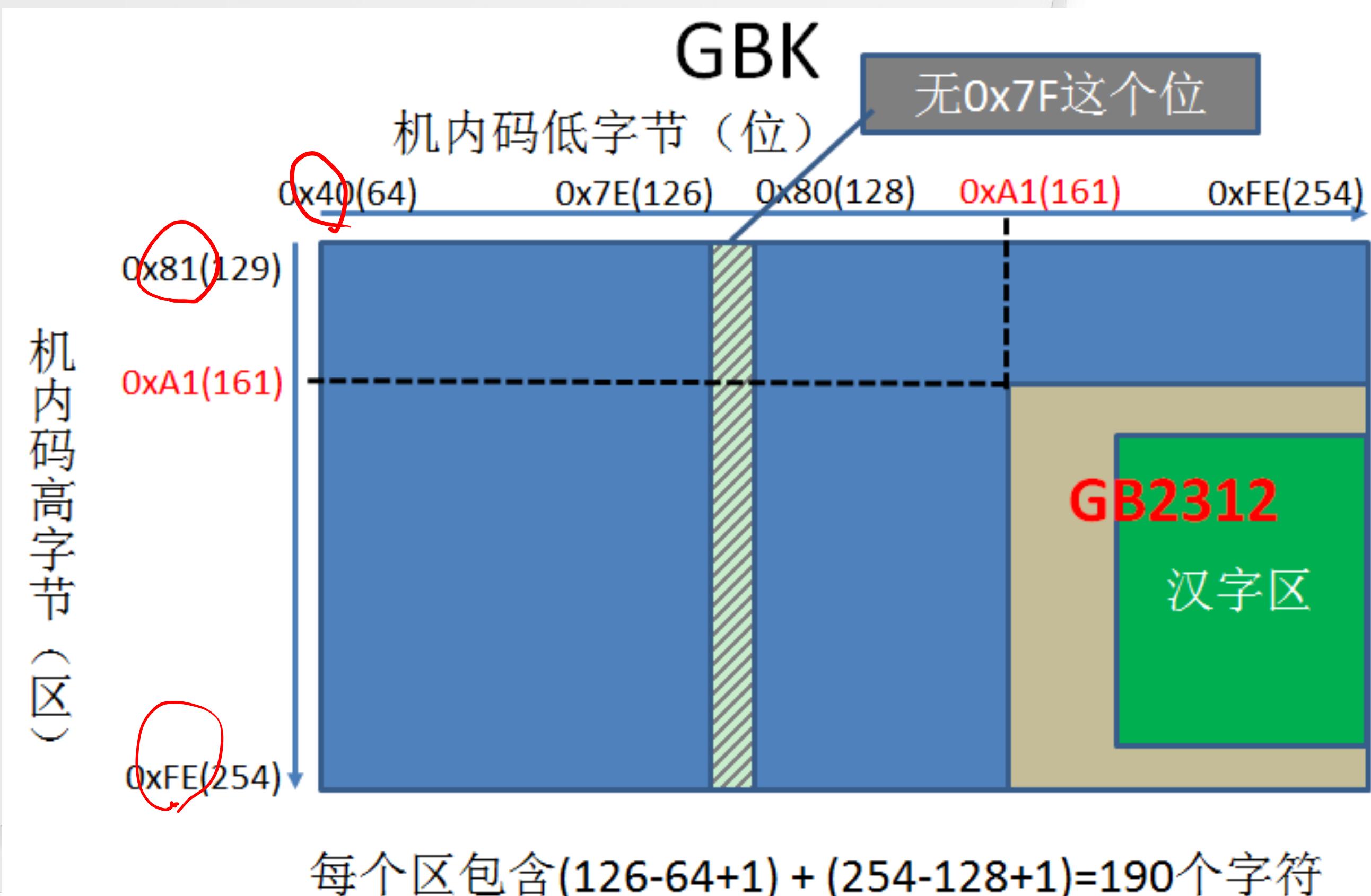




2. 数据的表示与存储

2.7.2、GBK编码

1.1995年发布，和GB2312兼容，汉字数有2万多个





2. 数据的表示与存储

2.7.2、GB18030编码

1.2005年发布，和Unicode2.0兼容，汉字数有2万多个

UTF-8

1-byte codes: {0x00-0x7F} Same as ASCII codes

2-byte codes: {0x81-0xFE}{0x40-0x7E} and {0x81-0xFE}{0x80-0xFE} Same as GBK codes

4-byte codes: {81-FE}{30-39}{81-FE}{30-39}

Maps linearly to Unicode codes as:

GB+81308130 ... = U+0080 ... U+FFFF

GB+90308130 ... = U+10000 ... U+10FFFF



2. 数据的表示与存储

2.7.3、为什么显示乱码？

汉字	Unicode编码结果	UTF-8编码结果
一	4E00	E4 B8 80
二	4E8C	E4 BA 8C
三	4E09	E4 B8 89

汉字	GBK编码结果	说明
涓	E4B8	GBK是按照2个字节解码的，取E4B8
€	80	80E4这个编码在GBK中没有，只取80
浜	E4BA	接着再取2个字节
嵐	8CE4	接着再取2个字节



2. 数据的表示与存储

2.2. 本节总结

1. 对二进制数据的不同解读会得到不同的信息
2. 文字信息识别必须靠统一的编码
3. Unicode国际编码和Utf-8编码的区别和联系

欢迎参与学习

WELCOME FOR YOUR JOINING

船说：计算机基础