码点(Code Point):每个字符对应一个唯一的编号,如汉字"你"的码点是 U+4F60。 范围:Unicode 目前定义的码点范围是 0x0000 ~ 0x10FFFF,可以表示超过一百万个字符。

本质: Unicode 只是一个"编号表",并没有规定具体如何存储这些编号。

一种字符集(Character Set),它的目标是为世界上所有的字符分配唯一的编号(码点,Code Point),以便在不同的系统、平台、语言之间实现字符的统一编码

Unicode(统一码、万国码、单一码、标准万国码)是 业界的一种标准,它可以使电脑得以体现世界上数十种 文字的系统

Unicode是字符集,UTF-32/ UTF-16/ UTF-8是三种字符 编码方案

每个字符都使用4字节。就空间而言,是效率非常低的。因此UTF-32使用并不广泛

.

最明显的优点是它在空间效率上比UTF-32高两倍,因为每个字符只需要2-4个字节来存储

UTF-16 也支持可变

一种针对[Unicode]的<mark>可变长度字符编码</mark>

UTF-8

| Unicode编码范围(十六进制) | UTF-8编码方式(二进制) |
|-------------------|-------------------------------------|
| 000000 - 00007F | 0xxxxxxx (ASCII编码) |
| 000080 - 0007FF | 110xxxxx 10xxxxxx |
| 000800 - 00FFFF | 1110xxxx 10xxxxxx 10xxxxxx |
| 010000 - 10FFFF | 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx |
| | |

1、Unicode 码点是: 20013

4、 utf-8 表示: 11100100 10111000

2、转成二进制->

10101101

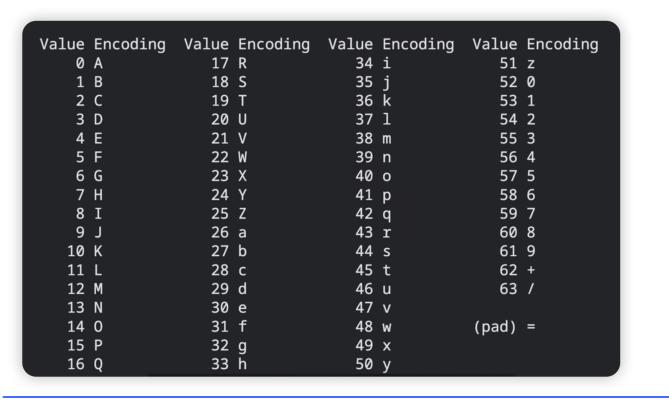
0100111000101101

1、128个US-ASCII字符只需一个字节编码(Unicode范围由U+0000至U+007F)。

2、带有附加符号的拉丁文、希腊文、西里尔字母、亚美尼亚语、希伯来文、阿拉伯文、叙利亚文及它拿字母则需要二个字节编码(Unicode范围由U+0080至U+07FF)。

- 3、类似汉字通常使用三个字节编码。
- 4、其他极少使用的Unicode辅助平面的字符使用四字节编码。

ASCII使用 7 个二进制位表示字符



base64 使用 6 个二进制位,对应能表示 64 个字符,所以叫 base64

- 1、对应的 ascii 编码是 72 101 108 110
- 3、按照6位分组010010000110010101 101100 011011 000110 111100 不足6位的用0补齐
- 4、转位十进制 18 6 21 44 27 6 60
- 5、对应的 base64 SGVsbG8 (编码后的长度要是 4 的倍数,不足用=补齐)SGVsbG8=

比如 Hello 如何转换

每5个二进制位编码成1个字

2-7 base32

base64

