Unicode的学名是"Universal Multiple-Octet Coded Character Set"，简称为UCS。1990年开始研发，1994年正式发布1.0版本，2020年发布13.0版本。Unicode是国际组织制定的可以容纳世界上所有文字和符号的国际标准编码。Unicode用数字0-0x10FFFF来映射这些字符，最多可以容纳1114112个字符，或者说有1114112个码位。UTF-8、UTF-16、UTF-32都是将数字转换到程序数据的编码方案。

UTF 是英文 Unicode Transformation Format 的缩写，意为把 Unicode 字符转换为某种格式。UTF 系列编码方案（UTF-8、UTF-16、UTF-32）均是由 Unicode 编码方案衍变而来，以适应不同的数据存储或传递，它们都可以完全表示 Unicode 标准中的所有字符。目前，这些衍变方案中 UTF-8 被广泛使用，而 UTF-16 和 UTF-32 则很少被使用。

1. UTF-8 使用一至四个字节为每个字符编码，其中大部分汉字采用三个字节编码，少量不常用汉字采用四个字节编码。因为 UTF-8 是可变长度的编码方式，相对于 Unicode 编码可以减少存储占用的空间，所以被广泛使用。

* 表示ASCII字符的UNICODE字符，被编码成1个字节，并且UTF-8表示与ASCII字符表示是一样的。所有其他的UNICODE字符转化成UTF-8将需要至少2个字节。每个字节由一个换码序列开始。第一个字节由唯一的换码序列，由n位连续的1加一位0组成, 首字节连续的1的个数表示字符编码所需的字节数。
* Unicode转换为UTF-8时，可以将Unicode二进制从低位往高位取出二进制数字，每次取6位，如上述的二进制就可以分别取出为如下示例所示的格式，前面按格式填补，不足8位用0填补。
* 注：Unicode转换为UTF-8需要的字节数可以根据这个规则计算：如果Unicode小于0X80（Ascii字符），则转换后为1个字节。否则转换后的字节数为Unicode二进制位数+3再除以5。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * **表一 UTF-8转换表** | | | | |
| **Unicode/UCS-4** | bit数 | **UTF-8** | byte数 | 备注 |
| 0000 ~  007F | 0~7 | **0**XXX XXXX | 1 |  |
| 0080 ~  07FF | 8~11 | **110**X XXXX  **10**XX XXXX | 2 |  |
| 0800 ~  FFFF | 12~16 | **1110**XXXX  **10**XX XXXX  **10**XX XXXX | 3 | 基本定义范围：0~FFFF |
| 1 0000 ~  1F FFFF | 17~21 | **1111 0**XXX  **10**XX XXXX  **10**XX XXXX  **10**XX XXXX | 4 | Unicode6.1定义范围：0~10 FFFF |
| 20 0000 ~  3FF FFFF | 22~26 | **1111 10**XX  **10**XX XXXX  **10**XX XXXX  **10**XX XXXX  **10**XX XXXX | 5 | 说明：此非unicode编码范围，属于UCS-4 编码  早期的规范UTF-8可以到达6字节序列，可以覆盖到31位元（通用字符集原来的极限）。尽管如此，2003年11月UTF-8 被 RFC 3629 重新规范，只能使用原来Unicode定义的区域， U+0000到U+10FFFF。根据规范，这些字节值将无法出现在合法 UTF-8序列中 |
| 400 0000 ~  7FFF FFFF | 27~31 | **1111 110**X  **10**XX XXXX  **10**XX XXXX  **10**XX XXXX  **10**XX XXXX  **10**XX XXXX | 6 |

实际表示ASCII[字符](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6)的UNICODE字符，将会编码成1个[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82)，并且UTF-8表示与ASCII字符表示是一样的。

示例

UNICODE 0xCA(1100 1010) 编码成UTF-8将需要2个[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82)：C3 8A， 过程如下：

0xCA(1100 1010)处于0080 ~07FF之间，从上文中的转换表可知对其编码需要2bytes，即两个字节，其对 应 UTF-8格式为： 110X XXXX, 10XX XXXX。从此格式中可以看到，对其编码还需要11位，而0xCA(1100 1010)仅有8位，这时需要在其二进制数前补0凑成11位: 000 1100 1010, 依次填入110X XXXX 10XX XXXX的空位中， 即得 1100 0011 1000 1010（C38A）。

同理，UNICODE 0xF03F (1111 0000 0011 1111) 编码成UTF-8将需要3个字节::0xEF 80 BF，对应格式为：1110XXXX, 10XX XXXX, 10XX XXXX，编码还需要16位，将1111 0000 0011 1111(F03F)依次填入，可得 1110 1111 1000 0000 1011 1111（EF 80 BF）。

1. UTF-16 使用二或四个字节为每个字符编码，其中大部分汉字采用两个字节编码，少量不常用汉字采用四个字节编码。UTF-16 编码有大尾序和小尾序之别，即 UTF-16BE 和 UTF-16LE。
2. UTF-32 使用四个字节为每个字符编码，使得 UTF-32 占用空间通常会是其它编码的二到四倍。UTF-32 与 UTF-16 一样有大尾序和小尾序之别，编码前会放置 U+0000FEFF 或 U+FFFE0000 以区分。