# 计算机编码

## Unicode字符集

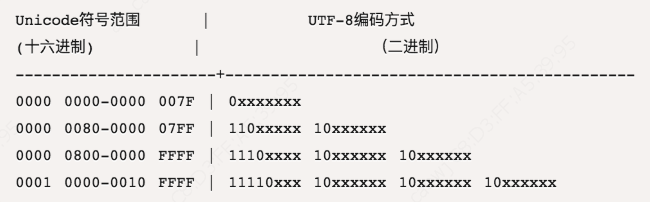
Unicode是一种字符集， 并未规定其二进制如何存储。

UTF(Unicode Transformation Format)Unicode转换格式，是规定Unicode如何存储的实现方式。

UTF-8表示使用1~6个字节不定长度大小来存储。

UTF-16表示使用2个或4个字节长度来存储。

UTF-32表示使用4个字节定长来存储。



以汉字严为例，演示如何实现 UTF-8 编码。

严的 Unicode 是4E25（100111000100101），根据上表，可以发现4E25处在第三行的范围内（0000 0800 - 0000 FFFF），因此严的 UTF-8 编码需要三个字节，即格式是1110xxxx 10xxxxxx 10xxxxxx。然后，从严的最后一个二进制位开始，依次从后向前填入格式中的x，多出的位补0。这样就得到了，严的 UTF-8 编码是11100100 10111000 10100101，转换成十六进制就是E4B8A5

|  |  |  |  |
| --- | --- | --- | --- |
| 例子 | 字符集 | 编码方式 | 存储方式 |
| 严(E4B8A5) | Unicode | UTF-8,根据Unicode字符集进行转换。 | EF BB BF E4 B8 A5;(大端存储)EF BB BF表示使用UTF-8编码, E4 B8 A5为真正编码。 |
| 严(E4B8A5) | Unicode | UTF-8无BOM编码 | E4 B8 A5;(大端存储)E4 B8 A5为真正编码。无需EF BB BF表示编码方式。 |
| 严(4E25) | Unicode | Unicode  (默认USC-2 little endian) 使用2个字节存储,无需转换。 | FF FE 25 4E; FF FE表示小端存储类型，25 4E为真正编码。 |
| 严(4E25) | Unicode | Unicode big endian  (USC-2 big endian ) 使用2个字节存储，无需转换。 | FE FF 4E 25 ; FE FF表示大端存储类型，4E 25为真正编码。 |

ANSI是默认的编码方式。对于英文文件是ASCII编码，对于简体中文文件是GB2312编码（只针对 Windows 简体中文版，如果是繁体中文版会采用 Big5 码。