关于字符编码

> 关于c++中的坑，当使用string的时候，往里面填写中文，则填入的是gbk编码（显示出来的数由于是负数，要注意补码）。如果使用wstring填入中文，则使用的是unicode编码。

1.Ascii编码(8位表示一个字符)：American Standard Code for Information Interchange，美国信息互换标准代码。把这些0×20以下的字节状态称为”控制码”。他们又把所有的空格、标点符号、数字、大小写字母分别用连续的字节状态表示，一直编到了第127号，这样计算机就可以用不同字节来存储英语的文字了。**共{0x00~0xFF}(256个状态)，使用了{0x00~0x7F}(128个状态)，未使用{0x80~0xFF}(128个状态)。**

2.扩展字符集(8位表示一个字符)：8位的字节一共可以组合出256(2的8次方)种不同的状态。扩展字符集就是使用Ascii编码剩下的状态(256-127)。**共{0x00~0xFF}(256个状态)，使用了{0x00~0xFF}(256个状态)。**

3.GB2312编码(对Ascii编码的中文扩展)：{0x00~0x7F}(128个状态)与Ascii编码相同(半角符号)，都是8位表示一个字符。{0x80~0xFF}(128个状态)，两个状态连在一起表示一个汉字，其中高字节(前一个字节)从{0xA1~0xF7}(87个状态)，低字节(后一个字节)从{0xA1~0xFE}(94个状态)，可组合出7000多简体汉字，同时把数字标点字母也重编成两个字节长(全角符号)。

4.GBK 标准(对GB2312编码的扩展)：由于GB2312编码的许多人名无法给出，所以GBK标准规定高字节满足大于0x7F(127)，低位无要求的条件。因此增加了20000多个汉字(包括繁体字)和符号。

5.GB18030编码(对GBK编码的扩展)：由于少数民族使用计算机，又加了几千个新的少数民族的字。

6.unicode编码(16位表示一个字符)：Universal Multiple-Octet Coded Character Set”，简称 UCS。对于Ascii编码原有的半角符号，保持低位不变，扩展成16位。因此，使用unicode编码保存英文文本，会浪费一倍空间。unicode编码，让中文字符和英文字符都占用两个字节，因此strlen函数不再把中文字符当成两个英文字符处理了。**unicode编码的问题在于：区分unicode编码和ascii编码，英文文本浪费空间问题。**

7.UTF标准(UCS Transfer Format)：UTF是一种面向传输的标准，是解决UCS(unicode编码)网络传输问题而出的标准。**UTF-8就是每次8个位传输数据，而UTF-16就是每次16个位。**目前使用最广的是UTF-8标准。UTF-8最大的一个特点，就是它是一种变长的编码方式。**它可以使用1~4个字节表示一个字符。**根据不同的符号而变化字节长度，**当字符在ASCII码的范围时，就用一个字节表示**，保留了ASCII字符一个字节的编码作为它的一部分。**注意的是unicode一个中文字符占2个字节，而UTF-8一个中文字符占3个字节。**从unicode到uft-8并不是直接的对应，而是要过一些算法和规则来转换。

在UTF-8传输标准下的unicode实际上是可变字节数的编码，它可以使用1~4个字节表示一个unicode。转换规则如下：

unicode编码字符范围(十六进制表示)	UTF-8编码方式(二进制表示)
00 00 00 00 ~ 00 00 00 7F	0XXXXXXX
00 00 00 80 ~ 00 00 07 FF	110XXXXX 10XXXXXX
00 00 08 00 ~ 00 00 FF FF	1110XXXX 10XXXXXX 10XXXXXX
00 01 00 00 ~ 00 10 FF FF	11110XXX 10XXXXXX 10XXXXXX 10XXXXXX

例子：新建文本，输入“联通”二字，保存后再打开乱码的原因。
新建文本后，默认以Ascii编码方式输入文本，当在文本中输入中文字符时，实际上使用的是Ascii的扩展GB系列的编码。所以在GB系列的编码下，“联通”二字的内部编码为：
联：**110**00001 **10**101010
通：**110**01101 **10**101000
通过观察可以看到，该编码方式与UTF-8的二字节规范一致，当再次打开文本时，文本误以为编码方式为UTF-8，因为UTF-8需要处理掉头信息"110"和"10"转换为unicode码，所以编码变为：
0000 0000 0110 1010(unicode编码 006A)
0000 0011 0110 1000(unicode编码 0368)