编码发展

在讲 python 编码之前，首先讲一讲计算机编码的发展过程，计算机保存信息是通过一串 01 二进制表示，计算机编码就是将字符保存成二进制的方式。

ASCII：一开始是国外发展的计算机，说的是英语，因此当时想的是将所有的英文字母、数字、标点符号进行编码，用计算机存储，但是这些字符加起来也就 128 个，用 8bit 来表示就足够了。

EASCII：后来又加入了一下希腊字母、拉丁符号、计算符号等，将 8bit 所能表示的字符全部占满，成为 EASCII。EASCII 是完全兼容 ASCII 的。

gbk：随着计算机不断的发展和普及，计算机不再仅仅在一些欧美国家使用，就比如说计算机普及到了中国，中国人民不仅仅只在计算机上显示英文，想把中文也存储在计算机中，但是中文博大精深，EASCII 已经没有剩余可以用来表示中文了，即使 EASCII 全用来表示中文也不够，因此中国人创造了自己国家的编码，gb2312。规定：一个小于 127 的字符的意义与原来相同，但两个大于 127 的字符连在一起时，就表示一个汉字，前面的一个字节（他称之为高字节）从 0xA1 用到 0xF7，后面一个字节（低字节）从 0xA1 到 0xFE。这样既能存储中文，也能兼容 ASCII。后来由于中国的汉字过多，又将 gb2312 进行了扩展，不再要求两个大于 127 的字符表示一个汉字，只要符合前一个字节大于 127，则其加上后一个字节一起表示汉字，即对低字节没有大于 127 的要求。

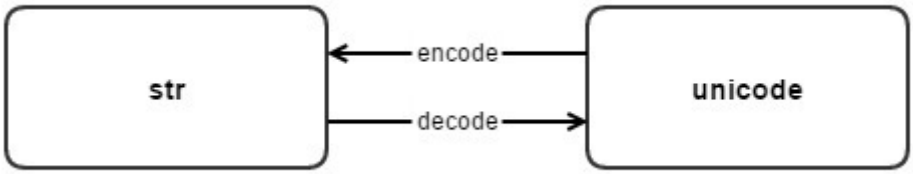
Unicode:由于各个国家都发展了自己的编码方式，为了显示别的国家的文字必须使用该国家相应的编码方式，于是为了统一，ISO（国际标谁化组织）发展了 unicode。用两个字节表示一个字符，对于 ASCII 中的编码保存原编码，将长度由 8 位变为 16 位。

Utf-8:通过 unicode 可以看到一个缺点，即原来的一个英文文件现在使用 unicode，空间增长了一倍，为了节约存储空间，这在网络传输上及其不利，UTF-8 的前 1~4 个字节表示一个符号，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做为它的一部分，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 utf-8 并不是直接的对应，而是要过一些算法和规则来转换。

Unicode 符号范围(十六进制)	UTF-8 编码方式（二进制）
0000 0000-0000 007F	0xxxxxxx
0000 0080-0000 07FF	110xxxxx 10xxxxxx
0000 0800-0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx
0001 0000-0010 FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

Python2x 编码

python2x 中有两种类型的字符串，str 和 unicode,unicode 是 unicode 编码的字符串，而 str 的编码则是可能是 UFT-8,gbk 等不同的编码方式，unicode 是应用程序保存的字符串类型，而 str 是输出时的字符串类型，str 和 unicode 是通过 decode 和 encode 转换的

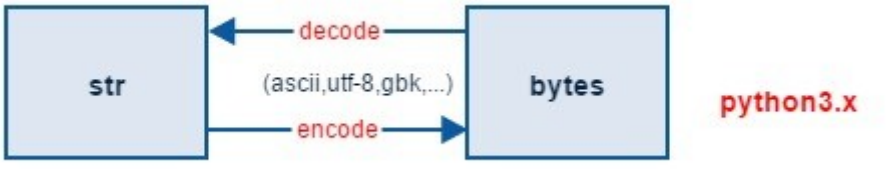


由于 python 的出现时间在 unicode 出现之前，所以系统默认的 str 编码方式是 ascii,将程序中的数据输出到文件时会将 unicode 的字符串转为 str 字符串，这就涉及到了 unicode 通过 encode 转为 str,在保存在文件中，当我们没有指定编码类型时，会采用系统默认的编码方式，即 ascii 编码。而 unicode 中如果存在中文等字符，进行 ascii 编码时就会报错。

'ascii' codec can't encode characters in position 0-14: ordinal not in range(128)

正确的方式应该在输出到文件之前将 unicode 字符串 encode 成正确编码的 str。可以使用 codecs 库，以我们指定的编码方式正确输出字符串到文件。

Python3x 编码



Python2 字符串设计上的一些缺陷：

使用 ASCII 码作为默认编码方式，对中文处理很不友好。

把字符串的牵强地分为 unicode 和 str 两种类型，误导开发者

Python3 把系统默认编码设置为 UTF-8，文本字符和二进制数据区分得更清晰，分别用 str 和 bytes 表示。文本字符全部用 str 类型表示，str 能表示

Unicode 字符集中所有字符，而二进制字节数据用一种全新的数据类型，用 bytes 来表示。

数据库存储格式

创建论文表，专家表，关联表，语句见附件 create_table.sql