在电工导C的第一个实验中，发现在utf8这些编码的转换上被折磨得很惨。下面是一点心得，更多内容可参见《python中文全攻略》。

在正常的使用中，发现python的默认中文编码方式应该是acsii，但是在大部分情况下，网站或其他一些需要我们解析的html源代码中的中文是采用utf8的编码方式的，因此，经过我长时间的百度，发现要在程序中加入下列三行代码，使得python对于中文的默认编码方式改为utf8：

import sys

reload(sys)

sys.setdefaultencoding('utf8')

（ps：据说这三行代码对python2.x的版本不适用）

经过上面三行的代码修改后，我在python的控制台中运行了一下代码，发现结果如下：

>>> s='你好'

>>> s

'\xe4\xbd\xa0\xe5\xa5\xbd'（此时为utf8的编码）

>>> print s

浣犲ソ

>>> s.decode('utf8')

u'\u4f60\u597d'

>>> print s

浣犲ソ

>>> s

'\xe4\xbd\xa0\xe5\xa5\xbd'

>>> print s.decode('utf8')（对s进行utf8的解码，发现可以被还原成汉字）

你好

>>> s.decode('utf8').encode('gbk')

'\xc4\xe3\xba\xc3' （gbk的编码）

>>> i=s.decode('utf8').encode('gbk')

>>> i

'\xc4\xe3\xba\xc3'

>>> print i

你好

>>> print i.decode('gbk')

你好 （这两段貌似说明gbk的编码不用解码也可以直接翻译）

>>>

>>> i=s.decode('utf8').encode('unicode')

Traceback (most recent call last):

File "<interactive input>", line 1, in <module>

LookupError: unknown encoding: Unicode （说明unicode的转码不能按照平常的方式来进行）

>>> i=unicode(s,'utf8')

>>> i

u'\u4f60\u597d'（“你好”的unicode编码）

>>> print i

你好

>>>

但是有时候在给一段文字进行编码时会显示ascii由于……导致……地方不能进行编码，此时应该将要编码的对象强制转化为string类型的再进行编码

例如：str(j.string).encode('utf8')

在得出上述方法的过程中，我也进行了其他的实验：

>>> import sys

>>> reload(sys)

>>> sys.setdefaultencoding('utf8')

>>> s='我们'

>>> s

'\xe6\x88\x91\xe4\xbb\xac'

>>> s.decode('gbk')

u'\u93b4\u621c\u6ed1' (这里是因为s的utf8编码凑巧也可以翻译成gbk的编码，具体见下文)

>>> s.encode('utf8')

'\xe6\x88\x91\xe4\xbb\xac' （这里发现在编译器中自动将中文转化成了utf8格式的编码，因此对其encode和直接输出好像是一样的）

>>> s.decode('utf8')

u'\u6211\u4eec' （这里是将utf8编码转为unicode编码）

>>> s='我们'

>>> s.decode('utf8')

u'\u6211\u4eec'

>>> s.encode('gbk') (看来可以直接转为gbk编码，和下边的效果一样)

'\xce\xd2\xc3\xc7'

>>> s.decode('utf8').encode('gbk')

'\xce\xd2\xc3\xc7'

>>> s='我们'

>>> s.encode('utf8')

'\xe6\x88\x91\xe4\xbb\xac'

>>> print s.decode('utf8')

我们

>>> s.encode('utf8').decode('utf8')

u'\u6211\u4eec'

>>> print s.encode('utf8').decode('utf8')

我们

>>> print u'\u6211\u4eec'

我们

>>> s.decode('utf8').encode('gbk')

'\xce\xd2\xc3\xc7'

>>> b=unicode(s,'utf8')

>>> b

u'\u6211\u4eec' (s的unicode编码)

>>> b=unicode(s,'gbk')

>>> b

u'\u93b4\u621c\u6ed1' (正好s的utf8编码也是例外一个词的gbk编码，因此可解)

>>> print s.decode('gbk')

鎴戜滑 （这个词的gbk编码和“我们”的utf8编码相同）