

通过读写文本文件小结“关于python处理中文编码的问题”

2019年1月14日 星期一 上午 0:37

一、引言

无论学习什么程序语言，字符串这种数据类型总是有着非常重要。然而最近在学习python这门语言，想要显示中文，总是出现各种乱码。于是在网上查了很多资料，各说纷纭，我也尝试了许多的方法，有时候可以正常显示，有时候确实乱码，让我摸不着头脑。于是自己利用python读写中文的文本文件来尝试去摸索python中的中文编码问题。比较幸运的是，最后能够正常的读取出文本里面的中文数据并且显示，而且还能将中文的结果数据写入文本文件中。但是本文仅仅只是总结处理中文乱码问题的小结，并没有将其编码的原理弄透。那么，下面就让我们开始吧。

二、准备工作

1、首先得建立一个文本文件（**编码方式是ascii**），文本文件的内容如下：

编号，雨量，站点位置

1，10.2，南京

2，45，北京

3，78，上海

2、给这个文件文件的每一行建立一个数据层也就是建立储存记录的类

```
class Rain:
    def __init__(self,id,acc,site):
        self.id=id
        self.acc=acc
        self.site=site
```

三、错误及其改正

1、错误一

首先需要建立一个py文件去着手写我们的代码。创建我的py文件之后我还什么代码都没写，仅仅是写了两行注释之后，保存一下，就发现下面的Console结果框就出来了错误。

代码截图如下：

```
1 '''
2 Created on 2015年10月18日
3 用于读取txt雨量文本数据
4 @author: tjm
5 '''
6
7
8
9
```

Console PyUnit

<terminated> D:\program\Java\PythonOne\src\Test\FileHandle.py

File "D:\program\Java\PythonOne\src\Test\FileHandle.py", line 2
SyntaxError: Non-ASCII character '\xe5' in file D:\program\Java\PythonOne\

原因分析：

原来Python的源代码默认的编码是ascii编码，而我的源代码文件中的注释含有中文，只是ascii编码所不能表达的字符，固然被python解释器解释的时候会出现如下错误。

解决方案：

只需要在文件的第一或者第二行，也只能是第一，第二行加上如下代码：

```
#coding:utf-8
```

这行代码的意思是，让解释器用utf-8的方式去解释源代码文件。

2、错误二

建立好一个py文件之后就要读取文本里面的文件，代码如下：

```
f=open("raindata.txt","r")
f.readline()#第一行是列，可以将文件移到第二行开始处
for line in f:
    print line
```

结果出现的结果中文都是乱码的，如下：

```
1 10.2°C
2 45
3 78M
```

原因：

因为txt文本文件中的中文都不是ascii编码，所以在读取出来的时候需要读取出来的字符串

经过解码才能正常显示。

解决方案：

只需要在读取的文件后面进行解码就好了。代码如下：

```
f=open("raindata.txt","r")
f.readline()#第一行是列，可以将文件移到第二行开始处
for line in f:
    print line.decode("gb2312")
```

结果如下：

1, 10.2, 南京

2, 45, 北京

3, 78, 上海

3、错误三

将数据正确读取出来之后就需要把每一行的数据存储在对象中。代码如下：

```
f=open("raindata.txt","r")
f.readline()#第一行是列，可以将文件移到第二行开始处
for line in f:
    lines=line.decode("gb2312").split(" ")
    obj=Rain(lines[0],lines[1],lines[2])
    data.append(obj)
```

结果出现了`split(" ",)`这个方法错误，提示说这个方法里面的参数不是中文编码。如下：

```
lines=line.decode("gb2312").split(" ")
```


```
UnicodeDecodeError: 'ascii' codec can't decode byte 0xef in position 0: ordinal not in range(128)
```

原因：

在网上看了这种错误的解决方案，说是因为我们的解决方案一种把py的源代码改为了utf-8的编码，所以可以解决该文件中所有的中文问题，但是调用的方法如何使其他模块中的方法，而方法还出现中文的话，就会提示错误。

解决方案：

既然知道了原因，那么解决的办法是，把整个环境的编码默认编码方式都改成utf-8就好了。更改的代码如下：

```

import sys
default_encoding="utf-8"
if (default_encoding!=sys.getdefaultencoding()):
    reload(sys)
    sys.setdefaultencoding(default_encoding)
data=[]
f=open("raindata.txt","r")
f.readline()#第一行是列，可以将文件移到第二行开始处
for line in f:
    lines=line.decode("gb2312").split(" ")
    obj=Rain(lines[0],lines[1],lines[2])
    data.append(obj)
f.close()
print len(data)
```



这样就解决了。

4、错误四

当把文本文件的编码方式换成了utf-8之后，上面的代码就出错了，错误如下：

Traceback (most recent call last):

File "D:\program\Java\PythonOne\src\Test\FileHandle.py", line 24, in <module>

lines=line.decode("gb2312").split(" ")

UnicodeDecodeError: 'gb2312' codec can't decode bytes in position 3-4: illegal multibyte sequence

原因：

这是因为本来文件的编码是utf-8，所以用gb2312的编码方式去解码，无疑，那肯定是错误。然而前面已经设置了本系统默认的编码方式就是utf-8，所以只需要将读出来的文本去掉gb2312的编码方式就好了。代码如下：

```
import sys
default_encoding="utf-8"
if (default_encoding!=sys.getdefaultencoding()):
    reload(sys)
    sys.setdefaultencoding(default_encoding)
data=[]
f=open("raindata.txt","r")
f.readline()#第一行是列，可以将文件移到第二行开始处
for line in f:
    lines=line.split(" ")
    obj=Rain(lines[0],lines[1],lines[2])
    data.append(obj)
f.close()
```

这就解决了文本文件为utf-8的编码方式了。

5、错误五：

完成了读出文本数据的工作之后，接下就是将读入的数据写入文本文件的了。代码如下：

代码如下：

```
f1=open('result.txt','w')
for vs in data:
    f1.write(vs.id+", "+vs.acc+", "+vs.site)
    f1.write("\n")
```

这段写入数据的代码的分两种情况。

1、解释器的默认编码是utf-8的，而文本文件的编码也是utf-8的，直接写入，写入到txt的结果不会乱码。

2、而文本文件的编码也是ascii的，写入的时候需要编码之后再写入，更改的代码如下：

```
f1=open('result.txt','w')
for vs in data:
    f1.write((vs.id+", "+vs.acc+", "+vs.site).encode("gb2312"))
    f1.write("\n")
f1.close()
```

到此为止，利用python进行读写文件的已经能够成功的运行的。然而，还补充一点，这是在打印列表中的中文时所照成的问题，并不是乱码的问题。

6、错误6

我们在打印含有中文的列表的时候中文得不到有效的输出，而是以utf-8的编码输出。代码

如下：

```
strs=['你好','hello']  
print strs
```

产生的结果如下：

```
['\xe4\xbd\xa0\xe5\xa5\xbd', 'hello']
```

解决方案：

把该列表一项一项的输出就没有问题。代码如下：

```
strs=['你好','hello']  
print strs[0],strs[1]
```

结果为：

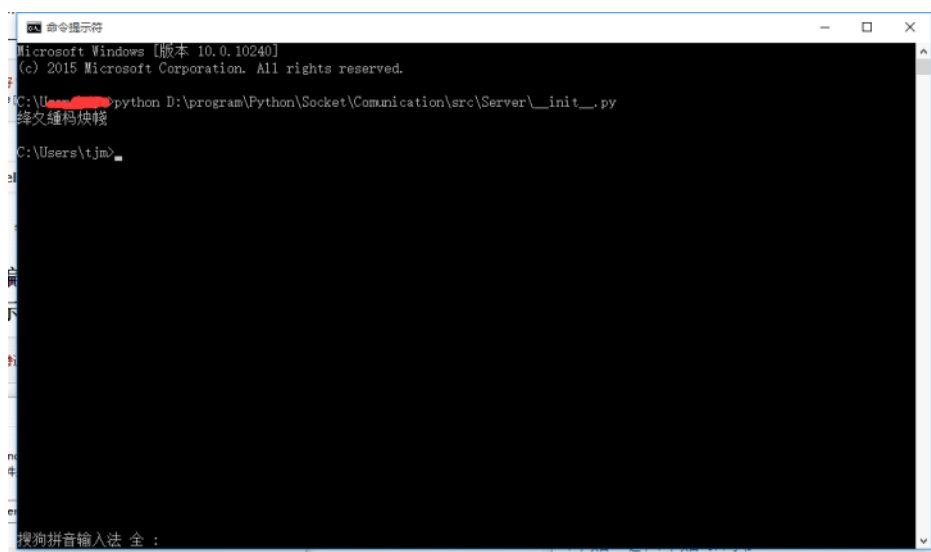
```
你好 hello
```

7、错误7

我编写python的脚本的编辑器为eclipse，在eclipse中的控制台中的输出结果（中文）是正确编码的，但是在cmd命令窗口中执行py脚本显示的确实乱码。打印代码如下：

```
print '等待连接'
```

然后我在cmd窗口中执行脚本的时输出的情况如下图：



出现了乱码。

解决方案：

要让这个字符串以utf-8的方式去实现，更改的代码如下图：

```
print u'等待连接'
```

这下不管是在eclipse的控制台中输出的，还是cmd命令窗口中输出都是正常显示。

四、总结

尝试过这么多错误之后终于正确的将文本的数据读取，也成功的数据写入到文本文件中。总结一下就那么几点。

- 源码中的编码方式
- 环境中的默认编码方式

- 结果文件中的编码方式

毕竟是初学python，对于上面的解决方案的解释的原理可能是错误的，希望各位大牛在看到错误的之后能及时指出来，在此感激不尽。

来自 <<https://www.cnblogs.com/mingjatang/p/4890420.html>>