

彻底搞懂 python 中文乱码问题

转：<https://blog.csdn.net/joyfixing/article/details/79971667>

总结 windows cmd 窗口下不支持utf-8，想要显示中文必须转换为gbk或者unicode，而Python idle 中这三种编码都支持。中文乱码的出现都是由于编码不一致导致的，存储的是用utf-8，打印的时候用gbk就会乱码了，所有要保证不乱码尽量保持统一，建议全部使用unicode

直接以gbk编码保存数据，作者不推荐此方法

```
# encoding:gbk
s = "中文"
print s
print repr(s)
```

cmd命令行可以支持直接打印unicode, 中文字符前加u表示用unicode存储

```
# encoding:utf-8
## 表示字符用unicode存储，而不是默认的utf-8进行存储
s = u"中文"
print s
print repr(s)
```

把中文强制转换为unicode编码

通过前面的encoding:utf-8声明表示里面的中文是以utf-8进行编码的，通过decode可从utf-8解码成unicode

```
# encoding:utf-8
s = "中文"

### utf-8 decode成unicode
u = s.decode('utf-8')
print u
print type(u)
print repr(u)
```

强制转为gbk, 不同编码之间的转换，都需要先转换成unicode，再转换到其它的编码

```
# encoding:utf-8
s = "中文"

### utf-8 decode成unicode
u = s.decode('utf-8')

### 然后unicode再编码成gbk
g = u.encode('gbk')
```

```
print g
print type(g)
print repr(g)
```

总结一口气说了这么多，不知道你们看懂没？想要不乱码，记住以下5点法则就好。

文件存储为utf-8格式，编码声明为utf-8 # encoding:utf-8 出现汉字的地方前面加 u 不同编码之间不能直接转换，要经过unicode中间跳转 cmd 下不支持utf-8编码 raw_input提示字符串只能为gbk编码

————— 版权声明：本文为CSDN博主「joyfixing」的原创文章，遵循CC 4.0 BY-SA版权协议，转载请附上原文出处链接及本声明。原文链接：<https://blog.csdn.net/joyfixing/article/details/79971667>