PCS5　Python 脚本文件

概述

脚本一般都是以文本形式存在的文件。它不无须像 C 语言那样编译成二进制文件执行，而是由特定解释器对脚本解释执行，所以只要系统上有相应的解释器就可以做到跨平台运行。Python 脚本比较特殊的是，在直接执行 Python 脚本时是解释执行，但若在该脚本中导入了另一个模块，这个模块会产生.pyc 字节码文件。

应用

以下是一个普通的 Python 脚本文件，实现合并多个 html 为可用文本。

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

""" Html To Text
@author: lizzie
@contract: shengyan1985@gmail.com
@see: ...
@version:0.1
"""
import os
from html2text import *
import chardet
import sys
```

```
reload(sys)
sys.setdefaultencoding('utf8')

YAHOO_DIR = 'J:\\yahoo_data\\'
YAHOO_TXT = YAHOO_DIR+ 'txt\\all.txt'

def html_to_txt():
    """将多个 html 文件合并为一个 txt 文件，统一编码为 utf-8 or ascii
    """
    ft = open(YAHOO_TXT, 'w')
    start = 1
    while 1:
        filename = YAHOO_DIR+ str(start) + '.html'
        if os.path.isfile(filename):
            fp = open(filename, 'r')
            htmltxt = ''.join(fp.readlines())
            if not htmltxt or not len(htmltxt):
                continue
            fp.close()

            codedetect = chardet.detect(htmltxt)["encoding"]
            #检测得到修改之前的编码方式
            print codedetect
     if not codedetect in ['utf-8', 'ascii']:
            htmltxt = unicode(htmltxt, codedetect).encode('utf-8')
            codedetect = chardet.detect(htmltxt)["encoding"]
            #检测得到修改之后的编码方式
                print 'change', codedetect

            ft.write(html2txt(htmltxt))
            print 'Success change html to txt %s' % start
            start += 1
        else:
            break
    ft.close()

if __name__ == '__main__':
    html_to_txt()
```

接下来依次介绍 Python 脚本的各个部分：

#!/usr/bin/python 这句话表示该脚本文件用哪个解释器来执行，这里是 Python 解释器。另

一种写法是#!/usr/bin/env python，两种写法是有区别的。区别详述请见网址：

http://groups.google.ro/group/python-cn/msg/843a13f6d8be7eab

精巧地址：*http://bit.ly/3TFHPC*

-*- coding: utf-8 -*-指定字符编码方式。有关字符编码方式和字符集的相关知识参见 PCS6 和以下这个链接。

字符编码和字符集介绍：

http://www.ruanyifeng.com/blog/2007/10/ascii_unicode_and_utf-8.html

精巧地址：*http://bit.ly/2BEWnC*

import os 表示导入 os 模块。from html2text import *、import chardet、import sys 同样是导入相应模块。这里的 html2txt 实现的是将某个 html 文件去除 html 标签，提取可用信息，转变成纯文本。

定义全局变量。全局变量通常用大写字母来标识。

定义函式。其中:表示函式开始，函式内利用缩进体现块与块之间的不同。#后表示注释。

最后调用函式执行。

执行脚本。在命令行中输入 python pcs-5-1.py 就可以执行。

小结

Python 脚本文件的结构是非常清晰的，很容易掌握。