检测文本编码类型

<http://blog.csdn.net/love4Mario/article/details/52402290>

file = open(fileName, "rb")#要有"rb"，如果没有这个的话，默认使用gbk读文件。

buf = file.read()

result = chardet.detect(buf)

file = open(fileName,"r",encoding=result["encoding"])

content = file.readlines()

本文实例讲述了Python简单检测文本类型的方法。分享给大家供大家参考，具体如下：

1. 是否为带BOM头的UTF8文件

def IsUtf8BomFile(pathfile):

if b'\xef\xbb\xbf' == open(pathfile, mode='rb').read(3)):

return True

return False

2．用cchardet库。

>>> import cchardet

>>> cchardet.detect(open(pathfile, 'rb').read())

{'encoding': 'UTF-8', 'confidence': 0.9900000095367432}

<https://stackoverflow.com/questions/6707657/python-detect-charset-and-convert-to-utf-8>

import icu

def convert\_encoding(data, new\_coding='UTF-8'):

coding = icu.CharsetDetector(data).detect().getName()

if new\_coding.upper() != coding.upper():

data = unicode(data, coding).encode(new\_coding)

return data

import cchardet

def convert\_encoding(data, new\_coding = 'UTF-8'):

encoding = **cchardet**.detect(data)['encoding']

if new\_coding.upper() != encoding.upper():

data = data.decode(encoding, data).encode(new\_coding)

return data