Java判断文件编码格式

转自：http://blog.csdn.net/zhangzh332/article/details/6719025

一般情况下我们遇到的文件编码格式为GBK或者UTF-8。由于中文Windows默认的编码是GBK，所以一般只要判定UTF-8编码格式。对于UTF-8编码格式的文本文件，其前3个字节的值就是-17、-69、-65，所以，判定是否是UTF-8编码格式的代码片段如下：

Java代码

```
java.io.File f=new java.io.File("待判定的文本文件名");
try{
  java.io.InputStream ios=new java.io.FileInputStream(f);
  byte[] b=new byte[3];
  ios.read(b);
  ios.close();
  if(b[0]==-17&&b[1]==-69&&b[2]==-65)
     System.out.println(f.getName()+"编码为UTF-8");
  else System.out.println(f.getName()+"可能是GBK");
}catch(Exception e){
   e.printStackTrace();
}
```

上述代码只是简单判定了是否是UTF-8格式编码的文本文件，如果项目对要判定的文本文件编码不可控（比如用户上传的一些HTML、XML等文本），可以采用一个现成的开源项目：cpdetector，它所在的网址是：http://cpdetector.sourceforge.net/。它的类库很小，只有500K左右，利用该类库判定文本文件的代码如下（由于cpdetector的算法使用概率统计，所以结果并不是100%准确的，但是是迄今为止我见过的最准确的....@_@）：

Java代码

```
/*--------------------------------------------------------------------------
  detector是探测器，它把探测任务交给具体的探测实现类的实例完成。
  cpDetector内置了一些常用的探测实现类，这些探测实现类的实例可以通过add方法
  加进来，如ParsingDetector、 JChardetFacade、ASCIIDetector、UnicodeDetector。
  detector按照“谁最先返回非空的探测结果，就以该结果为准”的原则返回探测到的
  字符集编码。
--------------------------------------------------------------------------*/
cpdetector.io.CodepageDetectorProxy detector =
cpdetector.io.CodepageDetectorProxy.getInstance();
/*-------------------------------------------------------------------------
  ParsingDetector可用于检查HTML、XML等文件或字符流的编码,构造方法中的参数用于
  指示是否显示探测过程的详细信息，为false不显示。
-----------------------------------------------------------------------*/
detector.add(new cpdetector.io.ParsingDetector(false));
/*-------------------------------------------------------------------------
  JChardetFacade封装了由Mozilla组织提供的JChardet，它可以完成大多数文件的编码
  测定。所以，一般有了这个探测器就可满足大多数项目的要求，如果你还不放心，可以
  再多加几个探测器，比如下面的ASCIIDetector、UnicodeDetector等。
 -------------------------------------------------------------------------*/
detector.add(cpdetector.io.JChardetFacade.getInstance());
//ASCIIDetector用于ASCII编码测定
detector.add(cpdetector.io.ASCIIDetector.getInstance());
//UnicodeDetector用于Unicode家族编码的测定
```

```
detector.add(cpdetector.io.UnicodeDetector.getInstance());
java.nio.charset.Charset charset = null;
File f=new File("待测的文本文件名");
try {
      charset = detector.detectCodepage(f.toURL());
} catch (Exception ex) {ex.printStackTrace();}
if(charset!=null){
     System.out.println(f.getName()+"编码是："+charset.name());
}else
    System.out.println(f.getName()+"未知");
```

上面代码中的detector不仅可以用于探测文件的编码，也可以探测任意输入的文本流的编码，方法是调用其重载形式：

Java代码

```
charset=detector.detectCodepage(待测的文本输入流,测量该流所需的读入字节数);
```

上面的字节数由程序员指定，字节数越多，判定越准确，当然时间也花得越长。要注意，字节数的指定不能超过文本流的最大长度。

判定文件编码的具体应用举例：
属性文件(.properties)是Java程序中的常用文本存储方式，象STRUTS框架就是利用属性文件存储程序中的字符串资源。它的内容如下所示：

Java代码

```
#注释语句
属性名=属性值
```

读入属性文件的一般方法是：

Java代码

```
FileInputStream ios=new FileInputStream("属性文件名");
Properties prop=new Properties();
prop.load(ios);
ios.close();
```

利用java.io.Properties的load方法读入属性文件虽然方便，但如果属性文件中有中文，在读入之后就会发现出现乱码现象。发生这个原因是load方法使用字节流读入文本，在读入后需要将字节流编码成为字符串，而它使用的编码是"iso-8859-1",这个字符集是ASCII码字符集，不支持中文编码，所以这时需要使用显式的转码：

Java代码

```
String value=prop.getProperty("属性名");
String encValue=new String(value.getBytes("iso-8859-1"),"属性文件的实际编码");

```

标签： web开发

明明是悟空
关注 - 4
粉丝 - 92
+加关注

« 上一篇：FileUpload之FileItem
» 下一篇：SAE Java相关问题小结

posted @ 2014-05-16 21:08 明明是悟空 阅读(5247) 评论(0) 编辑 收藏

相关博文：
· cpdetector获取文件编码
· Java如何获取文件编码格式
· 代码文件的编码不统一导致的坑
· xml、文件操作功能类
· Linux查看文件编码格式及文件编码转换<转>

最新新闻：
· 生物在量子进化？你可能正在利用量子现象生存
· 特斯拉大幅调整Model X和S价格和配置
· Drive.ai、Waymo率先商业化，智能驾驶加速进入冲击阶段
· 自由软件基金会认为商业性条款是非自由的
· 天猫双十一销售额破1682亿元：超过去年全天 提前8小时
» 更多新闻...

2014年10月 (13)
2014年9月 (15)
2014年8月 (37)
2014年7月 (2)
2014年6月 (3)
2014年5月 (26)
2014年4月 (17)
2014年3月 (11)
2014年2月 (7)
2014年1月 (6)
2013年11月 (14)
2013年10月 (8)
2013年9月 (12)
2013年8月 (20)

最新评论

1. Re:Jacob的使用出

***.ocx 是什么东西？到？？

2. Re:线程安全的单例

感谢大佬

3. Re:如何在java程序

或者shell脚本

大哥，如何在系统或者t只能执行指定的几个命令行mkdir ，不让执行rm

4. Re:jsp放在web-inf

最近正好在学Servlet和p访问的问题,弄了好久,

解决办法;
博主的文章给我带来了
浅!!
很有帮助,谢谢.

5. Re:二进制的计算（
补码存储数据）

@bambooo~1+（-1
原码运算=（0001+11
000=0...

阅读排行榜

1. 如何在java程序中调
hell脚本(39314)

2. linux mysql 操作命

3. 1080P、720P、4C
理论带宽(21157)

4. 修改Tomcat编码方
39)

5. 实例讲解教你读懂路

评论排行榜

1. JSP/Servlet的编码

2. struts2文件下载及
="inputName">input
m>的理解(2)

3. 二进制的计算（计算
存储数据）(2)

4.（三）WebRTC手记
(2)

5. 谷歌开源项目Chrom
与项目构建（Win7+vs

推荐排行榜

1. JAVA 的wait(), not
ized同步机制(3)

2. jsp放在web-inf下的

3. C语言中结构体的位
(2)

4. JAVA 正则表达式、
正则代码(2)

5. 表现层(jsp)、持久
务层 (逻辑层、service
abean) 、控制层 (act