MapReduce 程序处理中文时出现乱码

侯在钱

HADOOP 在使用的默认编码是 UTF-8，如果源文件编码格式是其它类型（如 GBK)，则会出现乱码。

解决方法：

在 Map 程序中增加如下的语句，申明使用 GBK 编码：

```
public void map(Object key, Text value,  Context  context)  throws
IOException,InterruptedException {

    String line = null;//编码后的每一行数据
    try {
        //解决中文乱码问题
        line = new String(value.getBytes(), "GBK");
    } catch (UnsupportedEncodingException e) {
        e.printStackTrace();
    }
    //业务处理
    String[] array = line.split(",");
    if (array!=null&&array.length==2) {
        Text outputKey = new Text(array[0].trim());
        int score = Integer.parseInt(array[1].trim());
        IntWritable outputValue = new IntWritable(score);
        context.write(outputKey, outputValue);
    }
}
```

另外，要保证 Linux 的客户端工具（像 SecureCRT）能正常显示中文。