

[微知兮 \(http://www.weizhixi.com/\)](http://www.weizhixi.com/)[登录 \(http://www.weizhixi.com/user/login/index.html\)](http://www.weizhixi.com/user/login/index.html) [注册 \(http://www.weizhixi.com/user/reg/index.html\)](http://www.weizhixi.com/user/reg/index.html)

知兮、青

不积跬步，无以至千里；不积小流，无以成江海。

[首页 \(http://www.weizhixi.com/user/index/index/id/1.html\)](http://www.weizhixi.com/user/index/index/id/1.html) [关于我 \(http://www.weizhixi.com/user/index/about/id/1.html\)](http://www.weizhixi.com/user/index/about/id/1.html)

[加关注](#)
[发私信](#)

访问：14706次

积分：932

等级：LV.3

文章：78篇

评论：12条

[Go!](#)

java初探Tess4j识别图片文字

标签：[图文识别](#) (<http://www.weizhixi.com/user/index/articletags/uid/1/name/%E5%9B%BE%E6%96%87%E8%AF%86%E5%88%AB.html>)

□ 2018-01-13 □ 阅读(840) □ 评论(0)

想学习下识别图片中的文字，找到了Tess4j图文识别的方式，于是就初步探究下，玩下识别验证码。

第一步，下载

- 1、以3.4.2版本为例，下载Tess4j-3.4.2-src.zip。
- 2、下载中文字库，chi_sim.traineddata。

下载Tess4j参考：

<http://sourceforge.net/projects/tess4j/>

字库下载参考：

<https://github.com/tesseract-ocr/tessdata/tree/3.04.00>

api文档参考：

<http://tess4j.sourceforge.net/docs/docs-3.4/>

第二步，准备工作

- 1、解压Tess4J-3.4.2-src.zip。
- 2、把根目录的lib和dist相关jar拷贝到你的项目lib中。
- 3、再把tessdata目录拷贝到你的项目根目录中。
- 4、再把中文字库放入tessdata目录。
- 5、dll不用理，Tess4j.jar已经包含。
- 6、如果遇到异常，Error: Invalid memory access, Error opening data file ./tessdata/eng.traineddata说明tessdata路径不对。

压缩包目录:

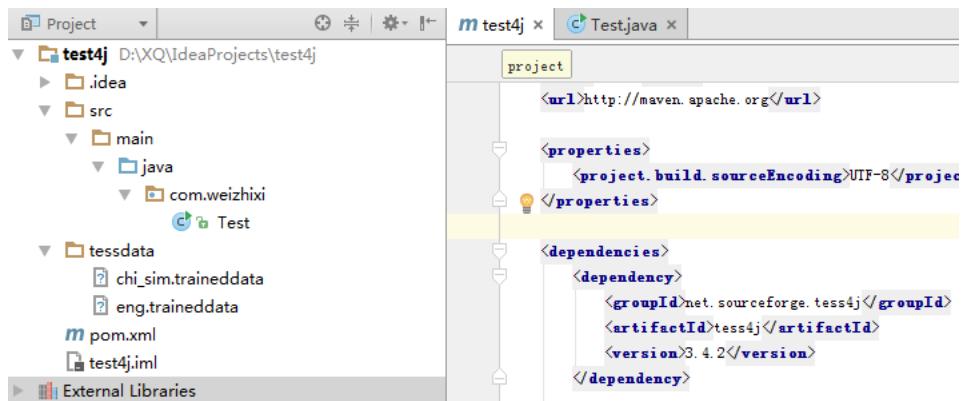
名称	压缩前	压缩后	类型	修改日期
.. (上级目录)			文件夹	
dist			文件夹	2017-11-14 21:20
lib			文件夹	2017-10-31 11:55
nbproject			文件夹	2017-10-12 08:00
src			文件夹	2017-01-26 19:38
tessdata			文件夹	2017-01-26 19:38
test			文件夹	2017-02-17 15:08
build.xml	5.3 KB	1.6 KB	XML 文档	2017-11-14 21:09
readme.html	5.4 KB	2.2 KB	360 Chrome HT...	2017-11-14 21:08
versionchanges.txt	3.8 KB	1.6 KB	文本文档	2017-11-14 21:05

如果是使用maven:

就在pom.xml加入即可。

```

1 <dependency>
2   <groupId>net.sourceforge.tess4j</groupId>
3   <artifactId>tess4j</artifactId>
4   <version>3.4.2</version>
5 </dependency>
```

我的项目test for java结构:**Tess4j依赖jar参考:**

```

1 commons-beanutils-1.9.2.jar
2 commons-io-2.6.jar
3 commons-logging-1.2.jar
4 ghostr4j-1.0.1.jar
5 hamcrest-core-1.3.jar
6 itext-2.1.7.jar
7 jai-imageio-core-1.3.1.jar
8 jboss-vfs-3.2.12.Final.jar
9 jcl-over-slf4j-1.7.25.jar
10 jna-4.1.0.jar
11 jul-to-slf4j-1.7.25.jar
12 junit-4.12.jar
13 lept4j-1.6.2.jar
14 log4j-1.2.17.jar
15 log4j-over-slf4j-1.7.25.jar
16 logback-classic-1.2.3.jar
17 logback-core-1.2.3.jar
18 slf4j-api-1.7.25.jar
19 xmlgraphics-commons-1.5.jar
```

第三步，开发测试**官方简单例子:**

```

1 package net.sourceforge.tess4j.example;
2
3 import java.io.File;
4 import net.sourceforge.tess4j.*;
5
6 public class TesseractExample {
7     public static void main(String[] args) {
8         File imageFile = new File("eurotext.tif");
9         ITesseract instance = new Tesseract(); // JNA Interface Mapping
10        // ITesseract instance = new Tesseract1(); // JNA Direct Mapping
11
12        try {
13            String result = instance.doOCR(imageFile);
14            System.out.println(result);
15        } catch (TesseractException e) {
16            System.err.println(e.getMessage());
17        }
18    }
19}

```

我的初探例子：

```

1 package com.weizhixi;
2
3 import net.sourceforge.tess4j.ITesseract;
4 import net.sourceforge.tess4j.Tesseract;
5 import net.sourceforge.tess4j.util.ImageHelper;
6 import javax.imageio.ImageIO;
7 import java.awt.image.BufferedImage;
8 import java.io.File;
9
10 public class Test {
11
12     public static void main(String[] args) throws Exception{
13         testEn();
14         //testZh();
15     }
16
17     //使用英文字库 - 识别图片
18     public static void testEn() throws Exception {
19         File imageFile = new File("C:/Users/XQ/Desktop/en.png");
20         BufferedImage image = ImageIO.read(imageFile);
21         //对图片进行处理
22         image = convertImage(image);
23         ITesseract instance = new Tesseract(); //JNA Interface Mapping
24         String result = instance.doOCR(image); //识别
25         System.out.println(result);
26     }
27
28     //使用中文字库 - 识别图片
29     public static void testZh() throws Exception {
30         File imageFile = new File("C:/Users/XQ/Desktop/zh.png");
31         BufferedImage image = ImageIO.read(imageFile);
32         //对图片进行处理
33         //image = convertImage(image);
34         ITesseract instance = new Tesseract(); //JNA Interface Mapping
35         instance.setLanguage("chi_sim"); //使用中文字库
36         String result = instance.doOCR(image); //识别
37         System.out.println(result);
38     }
39
40     //对图片进行处理 - 提高识别度
41     public static BufferedImage convertImage(BufferedImage image) throws Exception {
42         //按指定宽高创建一个图像副本
43         //image = ImageHelper.getSubImage(image, 0, 0, image.getWidth(), image.getHeight());
44         //图像转换成灰度的简单方法 - 黑白处理
45         image = ImageHelper.convertImageToGrayscale(image);
46         //图像缩放 - 放大n倍图像
47         image = ImageHelper.getScaledInstance(image, image.getWidth() * 3, image.getHeight() * 3);
48         return image;
49     }
50
51 }

```

处理倾斜图片：

如果图片字体倾斜的，可以用下面代码纠正

```

1  BufferedImage bi = ImageIO.read(imageFile);
2  ImageDeskew id = new ImageDeskew(bi);
3  double imageSkewAngle = id.getSkewAngle(); //获取倾斜角度
4  if ((imageSkewAngle > 0.05d || imageSkewAngle < -(0.05d))) {
5      bi = ImageHelper.rotateImage(bi, -imageSkewAngle); //纠偏图像
6  }

```

测试1:

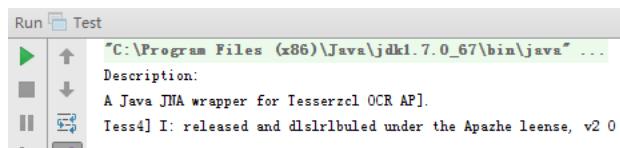
测试一张英文截图en.png。

Description:

A Java JNA wrapper for Tesseract OCR API.

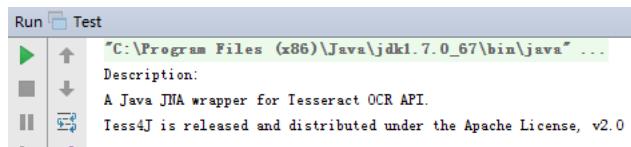
Tess4J is released and distributed under the Apache License, v2.0

未使用图像简单处理，运行读取图片文字：



发现有几次无法准确识别。

使用convertImage方法对图像简单处理，运行读取图片文字：



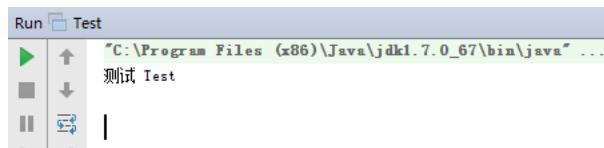
发现已经完全识别了。

测试2:

测试一张中文图片zh.png



用不用convertImage，测试结果都正常：

**测试3:**

来点复杂的图片：

cdefghijklmnopqrstuvwxyz ABCDEFGHIJKLMNOPQRSTUVWXYZ
34567890.:; ' " (!?) +-*=/

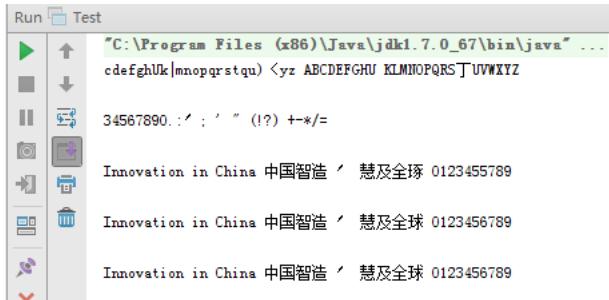
Innovation in China 中国智造，慧及全球 0123456789

Innovation in China 中国智造，慧及全球 0123456789

Innovation in China 中国智造，慧及全球 0123456789

来看看识别输出：

1、未使用图像处理



2、使用图像处理



发现识别度提高了很多，但部分还是未能够识别。

测试4：

识别干扰度比较低的简单验证码



识别结果：已经正确识别了。



经测试多张各种验证码，干扰度比较大的，扭曲字体的验证码不能识别。

关于训字库

训字库能提高中文字库的识别度。

需要下载中文字库：chi_sim.traindata

需要下载tesseract-ocr安装：tesseract-ocr-setup.exe

需要下载iTessBoxEditor用于修改box文件

至于怎么训字库，这里不展开说了。

初探总结

初探了一天，发现初级简单应用Tess4j：

- 1、只能识别几乎没有干扰，比较清晰的图片。
- 2、对图片灰度处理和放大处理，能提高识别度，但不是一定能起作用。
- 3、如果不准确的识别，可能要去训字库了，如测试识别图中的逗号，已经变成偏上的点了。
- 4、识别度受字体颜色、大小、清晰度、干扰度、扭曲、倾斜等度影响。
- 5、官方还提供了一些test例子，还有很多操作和应用。

初级应用只是简单的识别，能识别复杂度很大的图片文字，那是要很多牛B技术和逻辑的大神级操作。

如果想识别度很高很高几乎所有都能识别，又要快速集成、建议还是调用第三方识别API了，有些要收费的有些不用收费但有调用频次限制。

Demo下载

由于资源太大，我就不上传到我网站了。

请到我的网盘下载：

链接: <https://pan.baidu.com/s/1dHje9pR>

密码: z0bi

内含:

1、项目：基于maven_test4j例子项目.zip

2、官方Tess4j：Tess4J-3.4.2-src.zip

3、中文训字库：chi_sim.traineddata

原创文章，转载请注明出处：<http://www.weizhixi.com/user/index/article/id/59.html>

有用
0

无用
0

(<http://www.weizhixi.com/user/index/article/id/59.html#>) (<http://www.weizhixi.com/user/index/article/id/59.html#>) (<http://www.weizhixi.com/user/index/article/id/59.html#>)

(<http://www.weizhixi.com/user/index/article/id/59.html#>) (<http://www.weizhixi.com/user/index/article/id/59.html#>) (<http://www.weizhixi.com/user/index/article/id/59.html#>)

(<http://www.weizhixi.com/user/index/article/id/59.html#>)

« 上一篇：api接口签名加密请求（二） (<http://www.weizhixi.com/user/index/article/id/58.html>)

» 下一篇：springmvc3+hibernate4入门配置 (<http://www.weizhixi.com/user/index/article/id/60.html>)

分类： 编程语言 (<http://www.weizhixi.com/user/index/articlecat/uid/1/id/1.html>) / java (<http://www.weizhixi.com/user/index/articlecat/uid/1/id/9.html>)

点击(1017) 阅读(840) 评论(0) 收藏 举报

我的相关

java UTC时间格式化 时间带T Z

2018-02-01

(<http://www.weizhixi.com/user/index/article/id/70.html>)

api接口签名加密请求，从springmvc4项目搭建开始

2017-12-24

(<http://www.weizhixi.com/user/index/article/id/40.html>)

java生成m到n的随机数

2017-12-17

(<http://www.weizhixi.com/user/index/article/id/34.html>)

springmvc接收参数异常application/json not supported

2017-11-25

(<http://www.weizhixi.com/user/index/article/id/5.html>)

java分页工具类

2017-11-22

(<http://www.weizhixi.com/user/index/article/id/3.html>)

相关推荐

暂无相关

用户评论

您还没有登录，请 [[登录](http://www.weizhixi.com/user/login/index.html)] 或 [[注册](http://www.weizhixi.com/user/reg/index.html)]

暂无评论

分类

其他分类 (http://www.weizhixi.com/user/index/articlecat/uid/1/id/8.html)	1
java (http://www.weizhixi.com/user/index/articlecat/uid/1/id/9.html)	42
php (http://www.weizhixi.com/user/index/articlecat/uid/1/id/10.html)	16
js (http://www.weizhixi.com/user/index/articlecat/uid/1/id/11.html)	4
css (http://www.weizhixi.com/user/index/articlecat/uid/1/id/12.html)	1
html (http://www.weizhixi.com/user/index/articlecat/uid/1/id/13.html)	2
mysql (http://www.weizhixi.com/user/index/articlecat/uid/1/id/14.html)	6
nodejs (http://www.weizhixi.com/user/index/articlecat/uid/1/id/16.html)	2
说说 (http://www.weizhixi.com/user/index/articlecat/uid/1/id/17.html)	3
oracle (http://www.weizhixi.com/user/index/articlecat/uid/1/id/20.html)	1

阅读排行

java UTC时间格式化 时间带T Z (3323) (http://www.weizhixi.com/user/index/article/id/70.html)
wamp2.5安装imagine (1138) (http://www.weizhixi.com/user/index/article/id/2.html)
api接口签名加密请求，从springmvc4项目搭建开始 (611) (http://www.weizhixi.com/user/index/article/id/40.html)
thinkphp5项目如何在阿里云云虚拟主机部署 (543) (http://www.weizhixi.com/user/index/article/id/27.html)

最新评论

抖音出其不意的最火经典搞笑音乐
忧伤的猫咪: 新年好 (http://www.weizhixi.com/user/index/article/id/89.html#comment)
junit使用@Test报错
忧伤的猫咪: 博主好久没见你更新了，还有尽量写多点PHP，毕竟php是世界... (http://www.weizhixi.com/user/index/article/id/82.html#comment)
junit使用@Test报错
猫の物语: @忧伤的猫咪: 我也是 (http://www.weizhixi.com/user/index/article/id/82.html#comment)
junit使用@Test报错
忧伤的猫咪: 差点把密码忘记了 (http://www.weizhixi.com/user/index/article/id/82.html#comment)
tp5微信sdk之小改麦当劳的tp3.2微信sdk
看着星星数月亮: @知兮、青: 已经用其他办法配好了 谢谢 (http://www.weizhixi.com/user/index/article/id/61.html#comment)

阅读推荐

抖音出其不意的最火经典搞笑音乐 (52) (http://www.weizhixi.com/user/index/article/id/89.html)
js或php获取字符串长度中文1个字符，英文0.5个字符 (42) (http://www.weizhixi.com/user/index/article/id/88.html)
PHP过滤Emoji表情 (34) (http://www.weizhixi.com/user/index/article/id/87.html)
微信打开网页键盘收起后页面没下滑 (54) (http://www.weizhixi.com/user/index/article/id/86.html)
springmvc在jsp获取完整url地址含参数 (63) (http://www.weizhixi.com/user/index/article/id/85.html)