## Tesseract做图片识别

### Linux（CentOS7）下的安装

参考：<https://github.com/EisenVault/install-tesseract-redhat-centos/blob/master/install-tesseract.sh>

#### 准备环境

yum -y update

yum -y install libstdc++ autoconf automake libtool autoconf-archive pkg-config gcc gcc-c++ make libjpeg-devel libpng-devel libtiff-devel zlib-devel

安装leptonica <http://www.leptonica.org/download.html>

tar -zvxf leptonica-1.74.4.tar.gz

cd leptonica-1.74.4

./autobuild

./configure

make

make install

安装tesseract <https://github.com/tesseract-ocr/tesseract>

unzip tesseract-master.zip

cd tesseract-master

./autogen.sh

PKG\_CONFIG\_PATH=/usr/local/lib/pkgconfig LIBLEPT\_HEADERSDIR=/usr/local/include ./configure --with-extra-includes=/usr/local/include --with-extra-libraries=/usr/local/lib

LDFLAGS="-L/usr/local/lib" CFLAGS="-I/usr/local/include" make

make install

#### 下载数据 https://github.com/tesseract-ocr/tessdata

cd /usr/local/share/

上传eng.traineddata

到此就安装完成了。

### Windows下的安装

下载地址：<http://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-setup-4.00.00dev.exe>

windows上的安装就不介绍了，注意安装好了以后要添加环境变量：

path中添加：D:\Program Files (x86)\Tesseract-OCR\tessdata

新建TESSDATA\_PREFIX：D:\Program Files (x86)\Tesseract-OCR\tessdata

## 使用

命令行执行：tesseract image1.jpg out --psm 7 -l eng

image1.jpg：是输入文件名

out：识别出来的文字存放在out.txt这个文件中

-l：代表文字的类型，eng就是英文。

--psm：代表不同的切割方式

0 Orientation and script detection (OSD) only.

1 Automatic page segmentation with OSD.

2 Automatic page segmentation, but no OSD, or OCR.

3 Fully automatic page segmentation, but no OSD. (Default)

4 Assume a single column of text of variable sizes.

5 Assume a single uniform block of vertically aligned text.

6 Assume a single uniform block of text.

7 Treat the image as a single text line.

8 Treat the image as a single word.

9 Treat the image as a single word in a circle.

10 Treat the image as a single character.

11 Sparse text. Find as much text as possible in no particular order.

12 Sparse text with OSD.

13 Raw line. Treat the image as a single text line, bypassing hacks that are Tesseract-specific.

1. JAVA代码

