tesseract-ocr tips

【基本用法】

1、官方：https://github.com/tesseract-ocr

2、基本语法：

```
Usage:tesseract.exe imagename outputbase [-l lang] [-
psm pagesegmode] [configfile...]

pagesegmode values are:
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.
-l lang and/or -psm pagesegmode must occur before anyconfigfile.

Single options:
-v --version: version info
--list-langs: list available languages for tesseract engine

tesseract imagename outputbase [-l lang] [-psm pagesegmode] [configfile...]
tesseract    图片名  输出文件名 -l 字库文件 -psm pagesegmode 配置文件
```

3、示例：

（1）、

```
tesseract code.jpg code -l chi_sim -psm 7 digits
code 生成code.txt的结果文件
-l chi_sim 表示用简体中文字库
-psm 7 表示告诉tesseract code.jpg图片是一行文本，默认为 3
configfile 参数值为tessdata\configs 和 tessdata\tessconfigs 目录下的文件名
digits 内容为 tessedit_char_whitelist 0123456789-. 表示数字
```

（2）、白名单

```
tesseract code.jpg code -l eng -psm 7 -
c tessedit_char_whitelist="ABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789"
```

（2）、黑名单

```
tesseract code.jpg code -l eng -psm 7 -
c tessedit_char_blacklist="abcdefghijklmnopqrstuvwxy"
```

【训练】（以训练arial字体为例）

1、准备一张字体图片如下。

2、用 jTessBoxEditor 将图片转为tif文件，将tif文件命名为 eng.arial.font.exp0.tif。注意这里其实可以选多张图片。

```
Tools -> Merge TIFF...
```

3、生成坐标文件（.box）。

```
tesseract.exe eng.arial.font.exp0.tif eng.arial.font.exp0 batch.nochop makebox
```

【语法】：tesseract [lang].[fontname].exp[num].tif [lang].[fontname].exp[num] batch.nochop makebox

lang为语言名称，fontname为字体名称，num为序号；在tesseract中，一定要注意格式。

4、在当前目录创建 font_properties 文件，内容如下。

```
arial.font.exp0.box 1 1 1 0 0
```

【语法】：<fontname> <italic> <bold> <fixed> <serif> <fraktur>

fontname为字体名称，italic为斜体，bold为黑体字，fixed为默认字体，serif为衬线字体，fraktur德文黑字体，1和0代表有和无，精细区分时可使用。

5、字符校正。

打开jTessBoxEditor，BOX Editor -> Open，打开 eng.arial.font.exp0.tif，注意多页时页面切换。

6、执行批处理文件（arial.bat），生成.traineddata文件。

```
echo Run Tesseract for Training..
tesseract eng.arial.font.exp0.tif eng.arial.font.exp0 nobatch box.train

echo Compute the Character Set..
unicharset_extractor eng.arial.font.exp0.box
mftraining -F font_properties -U unicharset -
O arial.unicharset eng.arial.font.exp0.tr


echo Clustering..
cntraining eng.arial.font.exp0.tr

echo Rename Files..
rename normproto arial.normproto
rename inttemp arial.inttemp
rename pffmtable arial.pffmtable
rename shapetable arial.shapetable

echo Create Tessdata..
combine_tessdata arial.

echo. & pause
```

7、将生成文件中的arial.traineddata 文件拷贝到相应tessdata目录就可以使用啦！

```
tesseract code.jpg code -l arial
```

【参考文献】

1、啥都不懂也能识别验证码

2、Tesseract-OCR的简单使用与训练

3、Adding New Fonts to Tesseract 3 OCR Engine

4、Python做简单的验证码识别(ocr)

*** walker ***