## Tesseract样本训练

1. 下载jTessBoxEditor

<https://sourceforge.net/projects/vietocr/files/jTessBoxEditor/>

<https://ncu.dl.sourceforge.net/project/vietocr/vietocr.net/4.5/VietOCR.NET-4.5_64.zip>

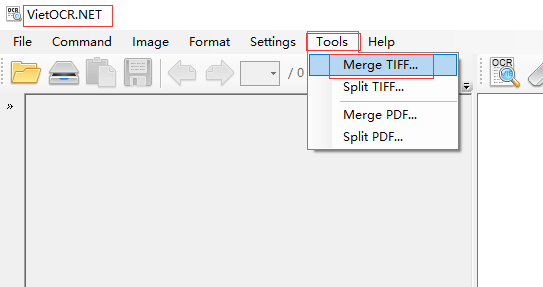
<https://nchc.dl.sourceforge.net/project/vietocr/jTessBoxEditor/jTessBoxEditorFX-2.0-Beta.zip>

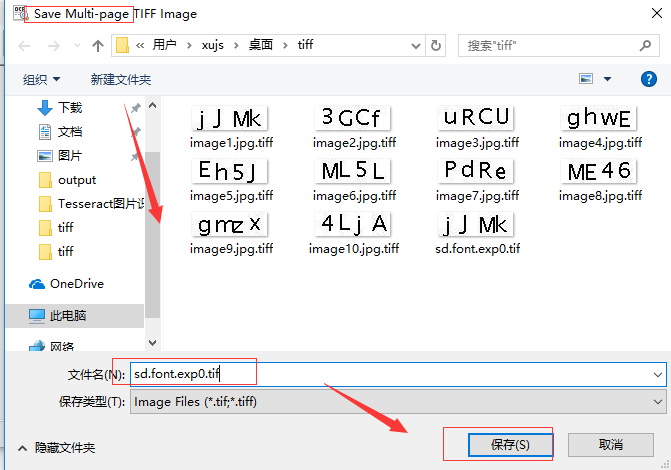
注意：TessBoxEditorFX-2.0-Beta.zip这里面的merge tiff文件的功能不可用，可以用VietOCR.NET-4.5\_64.zip这个里面的。

1. 下载样本数据
2. 把jpg的样本转化成tiff格式

<https://cn.office-converter.com/PNG-to-TIF>

1. 合并tiff文件





注意：保存到同一个目录下面，merge以后的tiff文件名是：sd.font.exp0.tif.

文件名的命名格式是有要求的：[lang].[fontname].exp[num].tif

lang：是自定义的lang名字

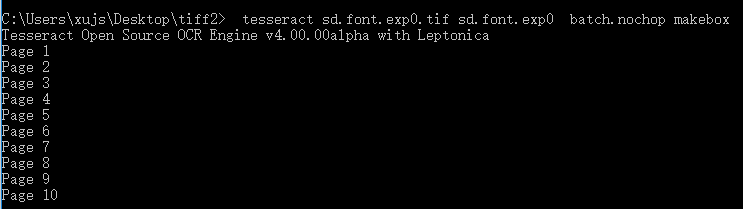
fontname：自定义的font的名字

exp：固定写法

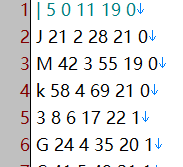
num：数字

1. 根据merge tiff文件“sd.font.exp0.tif”生成box文件“sd.font.exp0.box”

命令行：tesseract sd.font.exp0.tif sd.font.exp0 batch.nochop makebox



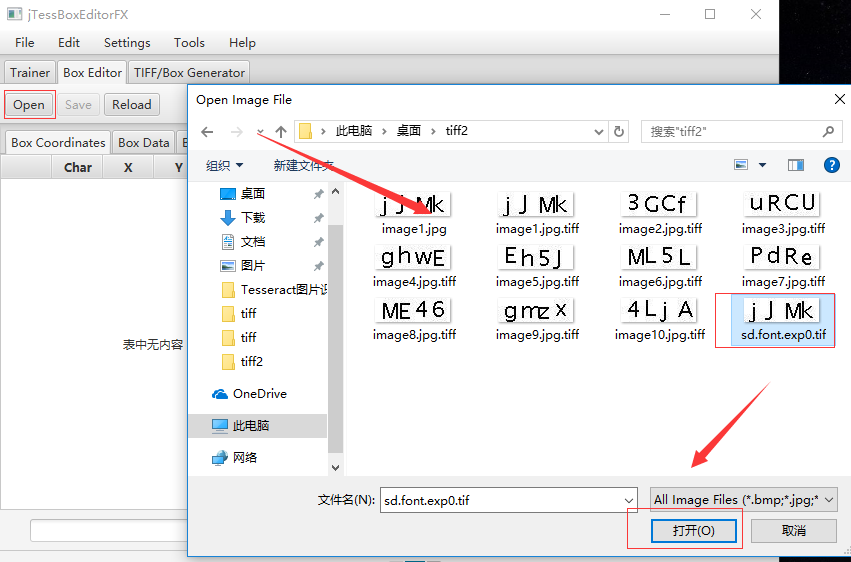
文件内容如下：

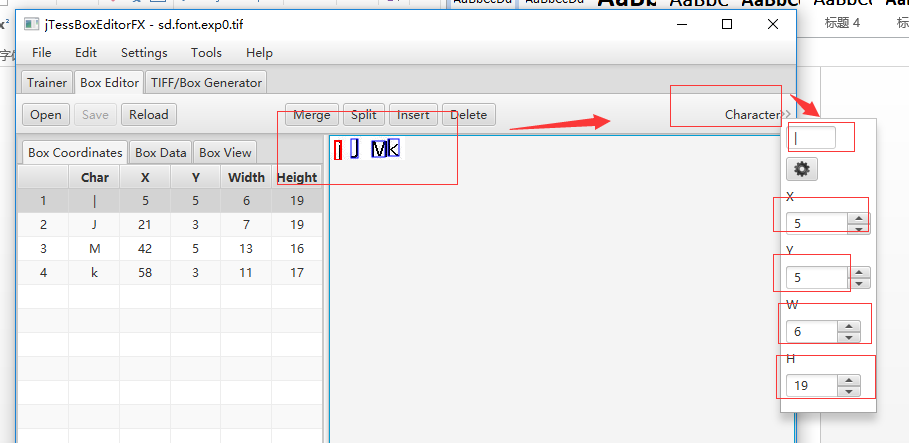


实际上就是识别出来的文字的坐标信息。

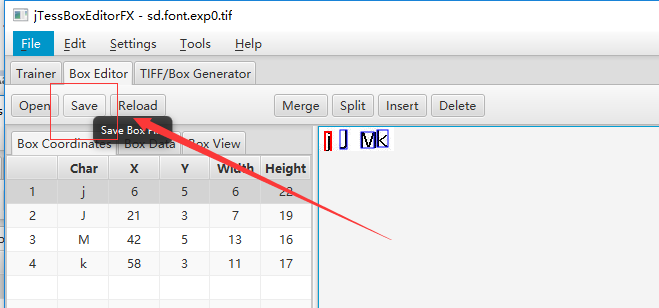
1. 文字校正

打开merge生成的sd.font.exp0.tif：





矫正完毕，点击保存：



1. 创建字体特征文件font\_properties

font\_properties不含有BOM头，文件内容格式如下：

<fontname> <italic> <bold> <fixed> <serif> <fraktur>

其中fontname为字体名称，必须与[lang].[fontname].exp[num].box中的名称保持一致。<italic> 、<bold> 、<fixed> 、<serif>、 <fraktur>的取值为1或0，表示字体是否具有这些属性。

这里在样本图片所在目录下创建一个名称为font\_properties的文件，用记事本打开，输入以下下内容：

font 0 0 0 0 0

这里全取值为0，表示字体不是粗体、斜体等等。

1. 生成语言文件。在样本图片所在目录下创建一个批处理文件sd.bat，输入如下内容:

rem 执行改批处理前先要目录下创建font\_properties文件

echo Run Tesseract for Training..

tesseract.exe sd.font.exp0.tif sd.font.exp0 nobatch box.train

echo Compute the Character Set..

unicharset\_extractor.exe sd.font.exp0.box

mftraining -F font\_properties -U unicharset -O sd.unicharset sd.font.exp0.tr

echo Clustering..

cntraining.exe sd.font.exp0.tr

echo Rename Files..

rename normproto sd.normproto

rename inttemp sd.inttemp

rename pffmtable sd.pffmtable

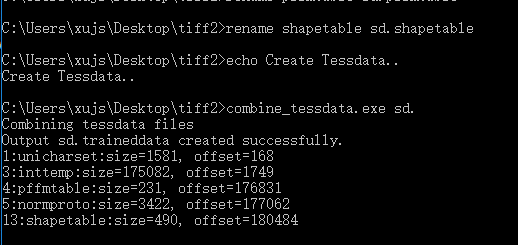
rename shapetable sd.shapetable

echo Create Tessdata..

combine\_tessdata.exe sd.

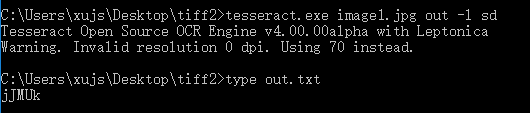
pause

运行这个批处理文件，需确认打印结果中的Offset 1、3、4、5、13这些项不是-1：



这样，一个新的语言文件“sd.traineddata”就生成了。

把sd.traineddata放到D:\Program Files (x86)\Tesseract-OCR\tessdata下面，就可以使用了：tesseract.exe number.jpg result -l eng



参考：https://www.cnblogs.com/samlin/p/Tesseract-OCR.html