**使用Tesseract 进行OCR 文字识别**

tesseract 下载<https://digi.bib.uni-mannheim.de/tesseract/>

Java JDK（运行训练工具需要java环境）

jTessBoxEditor

<http://down2.opdown.com:8019/opdown/jTessBoxEditor.zip>

将训练数据打包程一个tif文件：使用jTessBoxEditor 工具合成tif文件，命名为xxx.xxx.exp0.tif([语言名称].[字体名称].exp0.tif)

Ps. 需要提前准备默认训练数据库（eng.traineddate/chi\_sim.traineddata）

训练英文相关数据集

tesseract lang.font.exp0.tif lang.font.exp0 -l eng -- psm 7 batch.nochop makebox

用jTessBoxEditor打开tif文件，然后根据实际情况修改box文件。调整字符框对应的xyz参数。注意如果你有很多数据，它可能会分页，需要逐页调整。

图形用户界面, 文本, 应用程序

描述已自动生成

产生自符特征文件生成.tr 文件

tesseract lang.font.exp0.tif lang.font.exp0 -- psm 7 nobatch box.train

计算字符集生成unicharset 文件

unicharset\_extractor lang.font.exp0.box

定义字体特征文件，新建font\_properties文件，（无txt后缀）内容写入

font 0 0 0 0 表示默认普通字体。font要与tif文件内的font一致

聚集字符特征

shapeclustering -F font\_properties -U unicharset lang.font.exp0.tr

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

cntraining lang.font.exp0.tr

修改文件命名

mv inttemp font.inttemp  
mv normproto font.normproto  
mv pffmtable font.pffmtable  
mv shapetable font.shapetable  
mv unicharset font.unicharset

合并文件

Combine\_tessdata font.

生成font.traineddata 文件， 移入tessdata文件夹内

若出现empty page 时 可使用-- psm 7

**Pytesseract 使用**

from PIL import Image  
import pytesseract

读取网页url格式图片

res=requset.get(“url链接”)

image=Image.open(BytesIO(res.content)) – 读取为byte形式

print(pytesseracrt.image\_to\_string(image,lang=”font”)

--------------------------------------------------------------------------------

直接读取图片

image = Image.open(“xxx.png”)

print(pytesseracrt.image\_to\_string(image,lang=”font”)