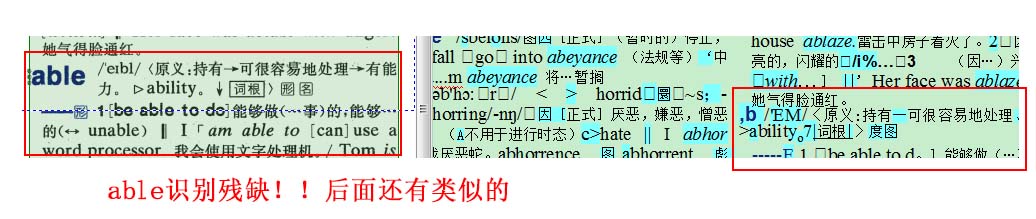
**从PDF开始，一种双色词典[词头-页码]的制备**

前些天，有人私信，说能否帮忙制作词典，随后便发了一个PDF文件给我。我看了一下，PDF质量挺高的。

《英汉多功能学习词典》小西友七等著，商务，2009，双色版。



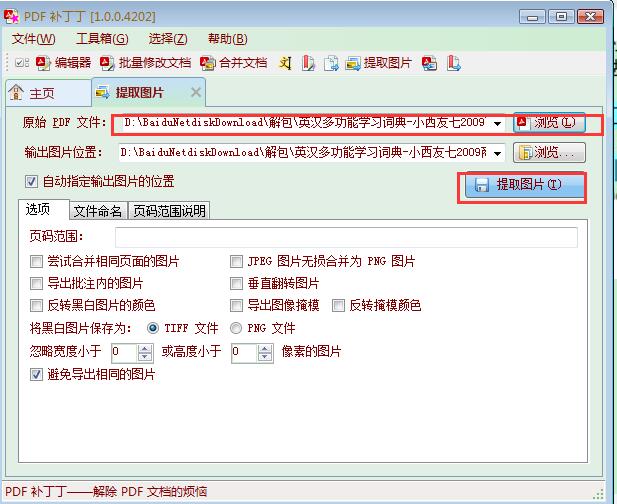
怎么提取词头呢?试着OCR了，发现真的很不理想，特别是会出现某些单词根本没有识别出来或者有残缺。



看了下PDF，这词典双色印刷，这蓝色的词头应该可以提取出来。

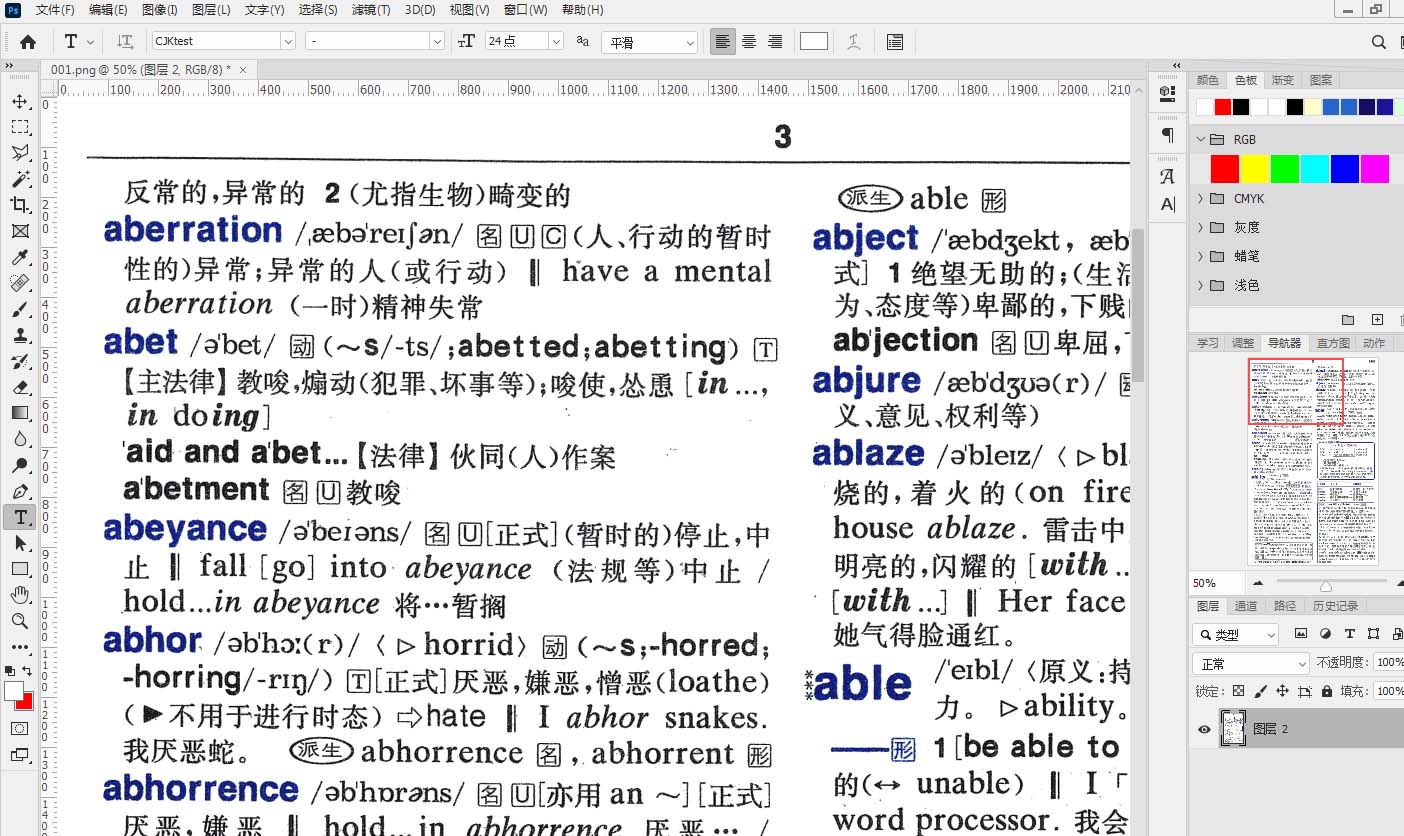
1. **从PDF中提取图片**

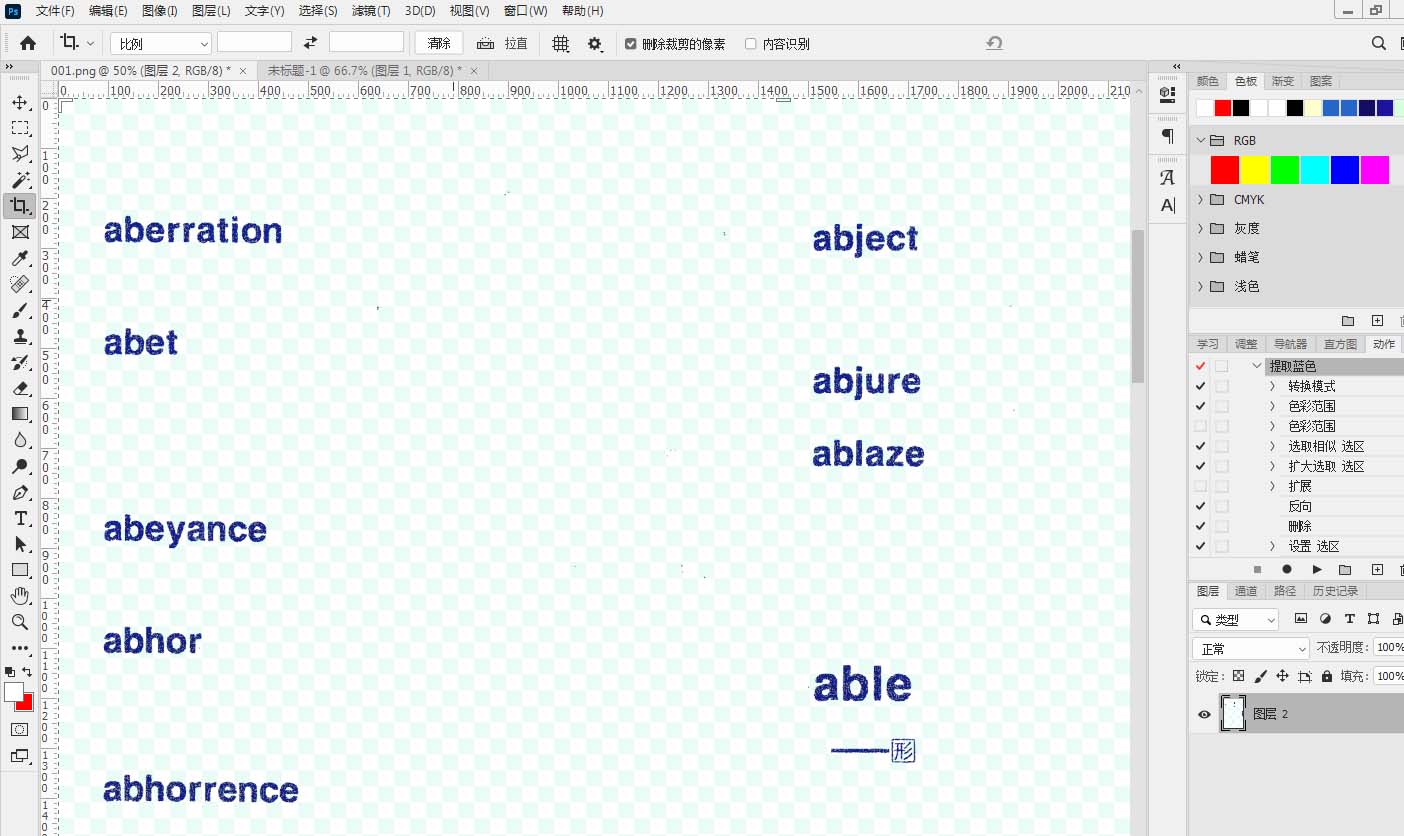
可以用PDF补丁丁等工具提取。



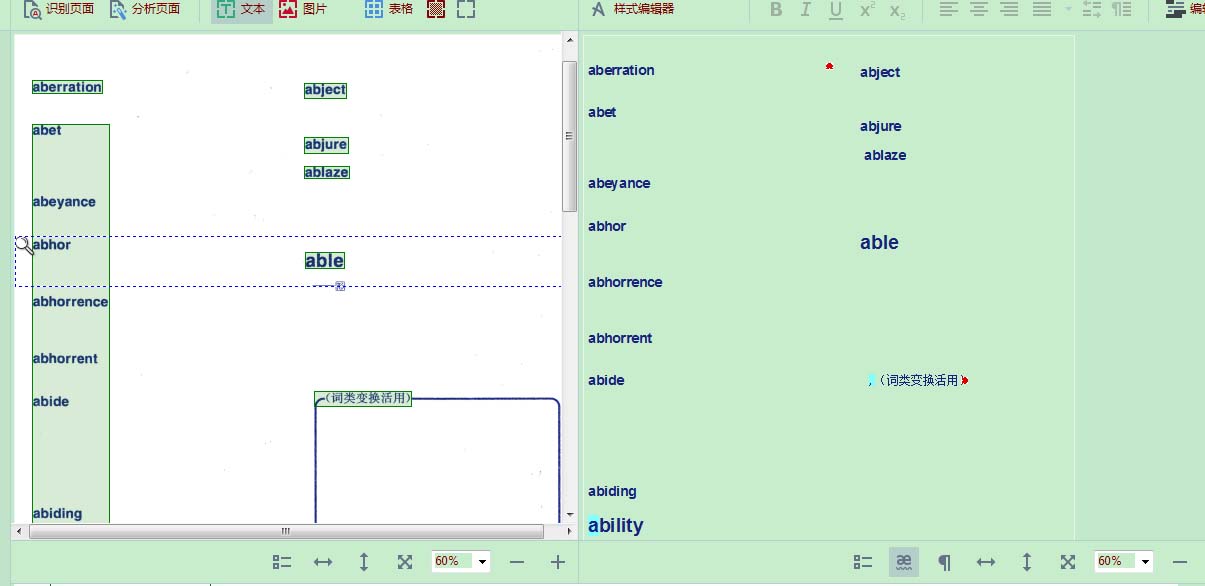
**二、提取蓝色词头**

打开PHOTOSHOP，提取这些蓝色词头。主要操作：色彩范围，选择相似，扩大选取，反向，删除。





导出，放到ABBYY FineReader Ocr一识别，这个真奈斯。



这么纯净的词头，完全可以一用。

效果不错啊！回到PHOTOSHOP，录个动作，跑PHOTOSHOP的自动处理了。1680张很快就跑完。

**三、OCR**

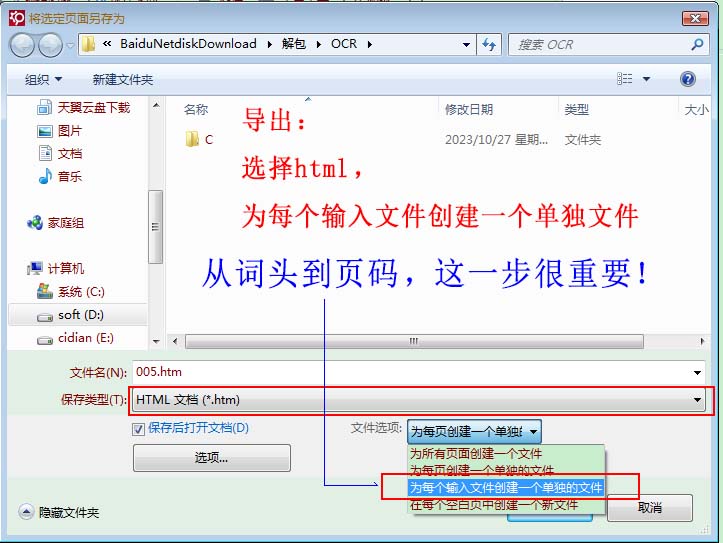
全部导入ABBYY FineReader Ocr，先分栏，然后OCR。

当然，说是纯净的词头，一提到纯净这词，就代表着不够纯净，百分之多少多少多少多少。OCR后自然存在一些问题：一些干扰的字符、不需要的文字、几个单词挤在一行，得手动把它们处理好。工作还是不小的。

此时用ABBYY FineReader Ocr的好处就展现出来，可以进行校对，不需要的直接删除。

导出，不错。导出的时候，注意一下。**文件格式选择：HTML，选项选择：为每个输入文件创建一个单独文件。这一步很重要。**

我建议是保存成HTML，不建议直接导出成txt文件。导出成txt，会发生几个单词挤在一行问题，增加处理难度。



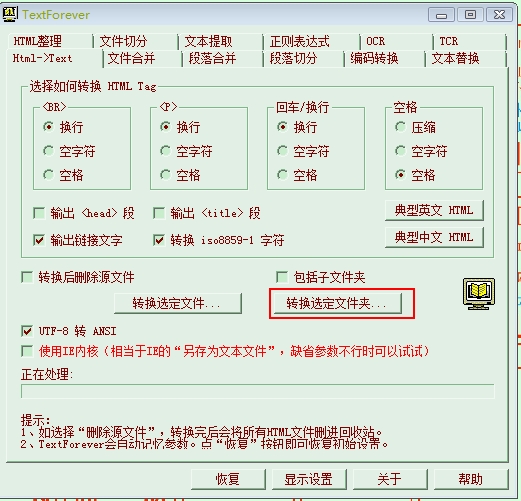
这样，得到了，一些以页码数字命名的文件，里面是该页的词头。

1. **合并文件及最后的处理**。

方法一：

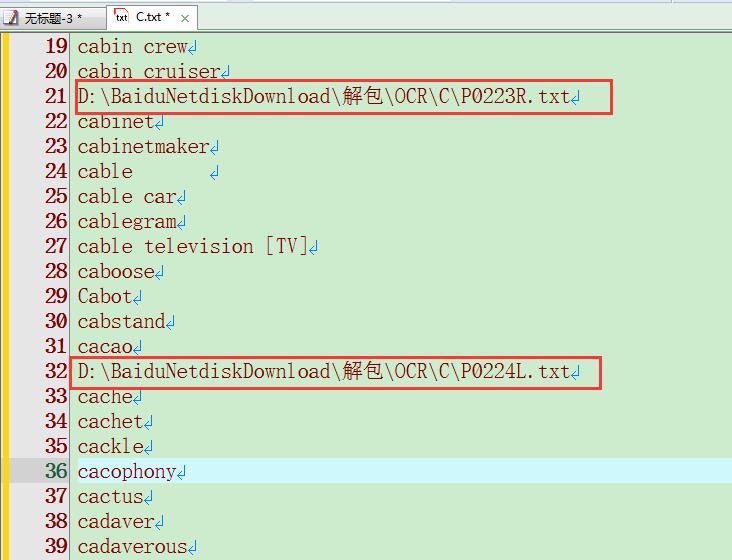
接下来，就是合并文件。但可以来一步，HTML转txt。也可以不用转txt，直接合并。（转不转TXT，大概就是多几个、少几个替换的差别。）

需要做的重要一步，就是插入文件名：在文件内容前面加注文件名。





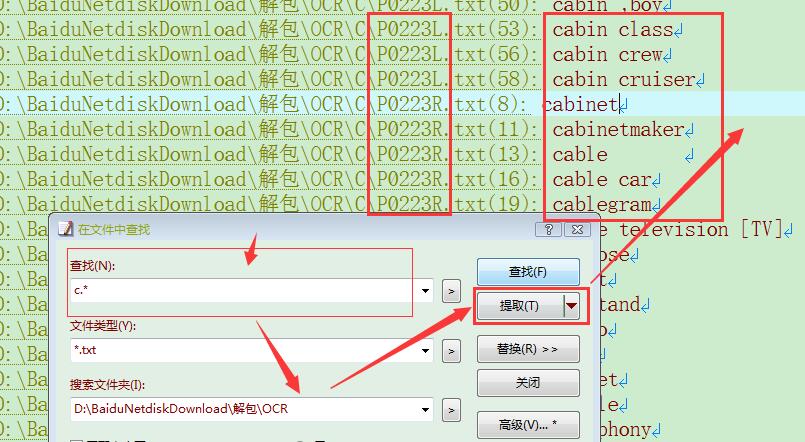
打开合并文件，随便处理一下，就得到对应的 词头-页码





**方法二，不合并文件，直接从文件提取**

同样可以先来一步HTML转TXT。（但也不是必要的，这个就看正则怎么写了。）



这个方法2,就更容易处理了。略。

**相关PDF及OCR工程见度盘：**

链接：https://pan.baidu.com/s/1BG3BA0xZEoms-8tLBfHFLQ?pwd=dvc8

提取码：dvc8