

写在前面	2
第一步 分割 PDF 到图片（以 PDFPATCHER 为例）	3
第二步 处理图片（以 COMICENHANCERPRO 为例）	4
第三步 识别（ABBYY 及汉王）	6
第四步 文本常见错误处理（EMEDITOR 和宏）	9
第五步 对比文本（BC）	10
第六步 合并段落（TEXTFOREVER）	13

写在前面

bc 是 beyond compare 4 软件的简称，作用在于比对文本差异处纠错，经实践证明比（大部分人的）肉眼一校靠谱（且速度更快）。此方法非自创，出处已不可考，在此感谢各位前辈。

该文档只有从一份 PDF 到 txt 文本的过程，后续的排版工作不用在教程里面找了，没有（sad）。

Beyond compare 从 PDF 到 txt 所需软件（群文件内查找）及大概步骤是这样的：

- 1 PDFPatcher（或 PDF Image Extraction Wizard 等）——拆分 PDF 到图片、图片组合为 PDF。
- 2 ComicEnhancerPro——处理图片（如切掉页眉页脚边侧多余页码等）。
- 3 汉王 ocr 以及 ABBYY——识别文字（abbyy 识别之后保留双层 PDF 及 TXT 各一份）。
- 4 Emeditor 和宏——运行宏进行文字处理。
- 5 Beyond Compare 4——对比文本并纠错，称之为“一校”。
6. TextForever（或排版助手）——合并文本段落。
- ~~7 自行粗排一次（该过程省略——因为我不会），即可进行读校。~~

所以首先，你需要一份不错的 **PDF**，如果找不到也买不到并且不具备自行扫描的条件，可以不看了。（如果已有好的文本仅需排版，也请无视此文档。）

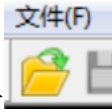
第一步 分割 PDF 到图片（以 PDFPatcher 为例）

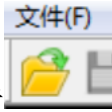
打开软件选择左边提取图片



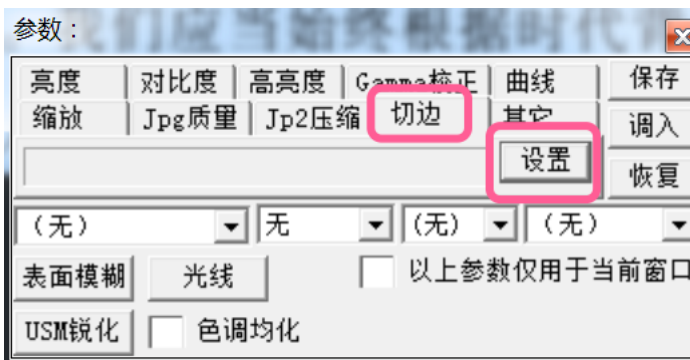
PDF 拖入后自行选择图片输出位置，文件命名和页码范围默认即可。点击“提取图片”后等待完成。

第二步 处理图片（以 ComicEnhancerPro 为例）



打开软件 Ctrl+O 或点击 ，选择刚才分割 PDF 后的图片导出位置，随机挑选一张或多张图（不要全选，因为并没有用，而且软件处理时有可能会崩溃）；或直接打开刚才的文件夹，随机挑选几张图拖入软件。

参数调节有很多，重点是切边（点击切边，再点击下面的设置）：



① 纠偏区域，如果觉得不需要就默认为无。

② 页面大小选择“内容框大小”或“保持不变，内容框外填白”（推荐），此

时⑤区域预览界面会有图中的绿色线（也就是处理后留下来的实际文本区域），红色外框是指所需文本的最大边框，如果有页眉页脚需要去掉，将红色外框调整至合适位置然后自行在④区域浏览页面检测是否合适。

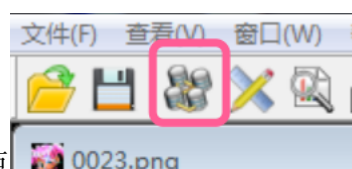
③ 打勾，尤其遇到示例图片里面这种有页眉页脚（甚至边侧页码）的 PDF。“忽略斑点直径”的程度自行调整。（比如示例图中这张图，如果此处选 8 就会把页眉也框在绿色范围内，选 20 就自动无视页眉了。需要手动将红色边框线拉到合适的位置，自行在软件内调节。非常好的功能，不过如果不熟悉的话……那还是慎用吧。）

④ 预览任意页，此处是指整个 PDF 拆分后的图片文件夹里面按顺序的任意页面。可使用四个按钮快速预览边缘是否多切或者少切。

⑤ 预览区域，红色是手动框选的范围，绿色是实际切边后的界面。

确保无问题后点④界面右下侧的“确定”按钮。

（其他“柔化”“锐化”“色彩”等各取所需自行调节。不同的 PDF 甚至同一本 PDF 的不同页面所需参数不同。只需调整单页时，勾选参数窗口右下角的“以上参数仅用于当前窗口”。）



参数选完以后点“批量处理”按钮，会将该文件夹内所有图片按照所选参数处理。



自行指定文件夹，扩展名 jpg 即可，其他默认。（此窗口也可补充设置参数。）

第三步 识别 (abbyy 及 汉王)

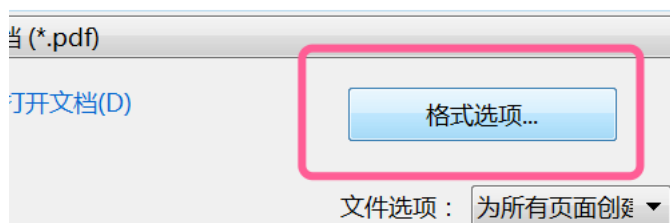
1. Abbyy (此处指 ABBYY FineReader 12 版本——因为没钱买 14 版) 打开后直接把刚才第二步处理过的图片全选丢入, 点击上方“读取”或 Ctrl+shift+R, 等待完成即可。

注意事项: (此处以@捕鱼的方法为例)

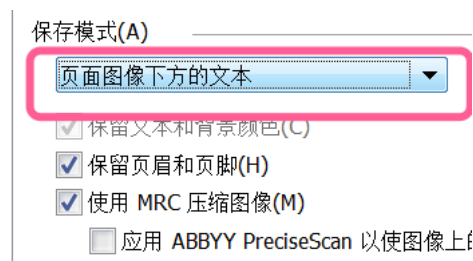
①为了提高识别率, 语言一定要选对 (比如: 简体中文+英语), 但默认的简体中文貌似是 CJK 字典, 会识别出一些异体字, 可自行参考以下链接进行替换。

<https://www.hi-pda.com/forum/viewthread.php?tid=1247258>

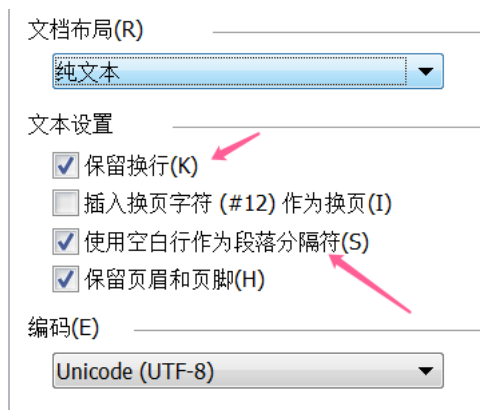
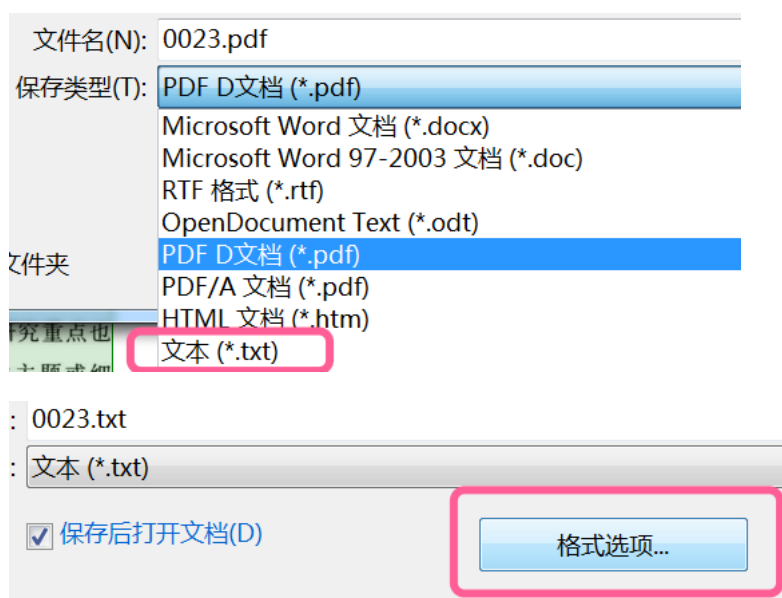
②识别完成后, 点击保存右边的↓箭头, 选择 PDF:



点击“格式选项”, “保存模式”设置为页面下方的文本 (也就是双层可搜索 PDF, 便于读校时查找文本。若该 PDF 无法搜索文本, 建议换一个 PDF 阅读器, 如 Sumatra PDF 阅读器, 见群文件。或福昕一类)。



回到主页再点一次保存右边的↓箭头, 选保存为其他格式——TXT——格式选项:



第一项和第三项勾上，后面再说为什么。

保存（一般我命名为“书名 ab.txt”）之后 abbyy 这一步就结束了。

2. 汉王。好像没什么设置的（默认就是简体中文+英文）……文件夹内所有图片丢入之后**全选左边图片列表**，点 F8 开始识别，识别完之后再次**全选左边图**

片列表点上方菜单栏“输出”，选择“到指定格式文件”存为txt即可。（一般我命名为“书名 hw.txt”。）

第四步 文本常见错误处理（Emeditor 和宏）

下载“排版专用宏 20170525”，打开 Emeditor，点击上方菜单栏“宏”——自定义——右上角“添加”，选择刚才下的宏，然后“确定”（可在此界面调整宏的位置）。

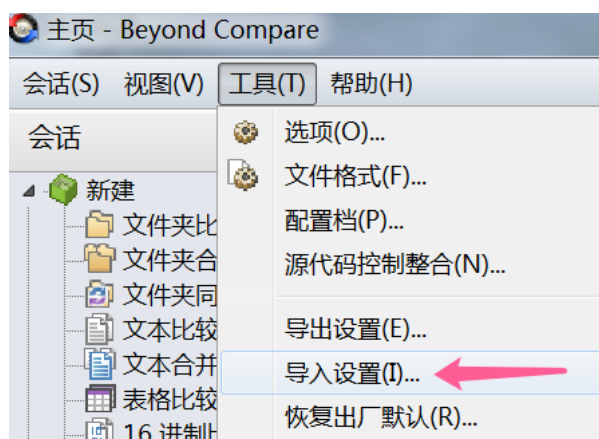
将刚才的 txt 拖入 Emeditor，选择运行宏——排版专用宏 20170525（如果提示缺少对象……可能是提示你找个男朋友和女朋友……好吧其实重启一下 em 软件就好了），第 2 项格式重排的 123 都可以点（不要点“4 清除空行”），每运行一项需要再次运行宏才能运行下一项，如下图：



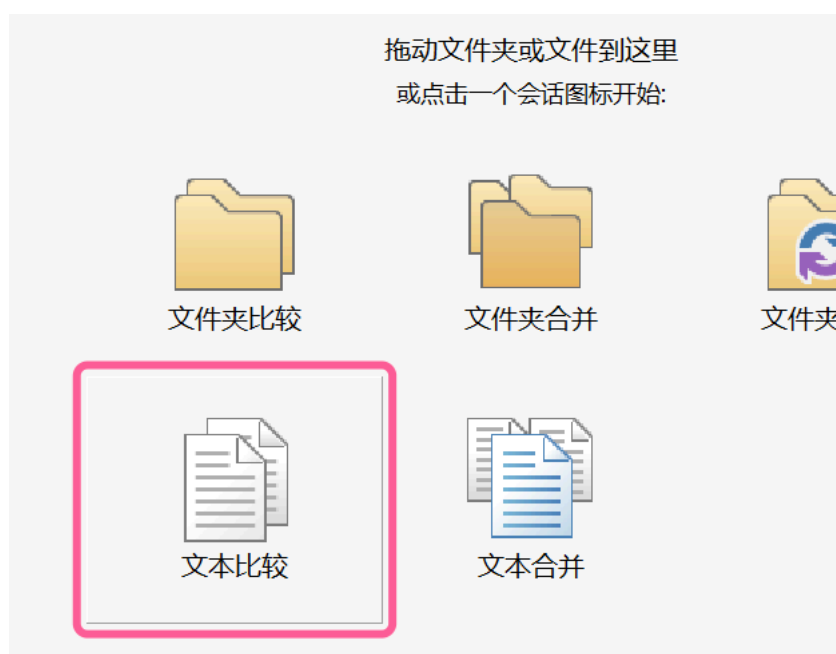
整理完毕后即可进行 bc。（记得运行完之后 Ctrl+S 保存。）

第五步 对比文本 (bc)

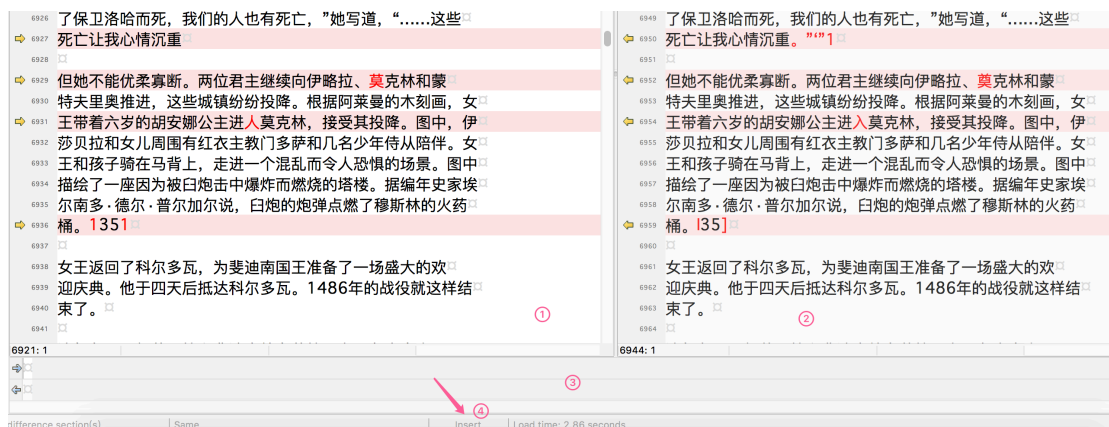
下载 bc 和 bc 配置文件之后，导入一下配置文件（需要选择的都可以全选，开心就好）：



将刚才的两个文档拖入或者点“文本比较”然后选择刚才的两个文档：



界面是下面这样的（也可能是上下视图，在上方“视图”里面可以自行选择左右或者上下）：



①区域我习惯放 ab 识别的文本。红色部分表示差异（如果肉眼看觉得是一样的，但显示为红色，那可能是……你看错了）。

②区域我习惯放汉王识别的文本。

③区域是每一行的对比区域，光标放在这里再按快捷键可以只复制该行到另一侧而不用特别的快捷键。（看不懂这句没关系，去实战一下就好了。）

④区域表示输入文字是插入或覆盖方式（按键盘上的 insert 键切换）。

导入配置后更改文本到另一侧的默认快捷键是 Ctrl+U（右边/下面的文本对，复制到左边/上面）和 Ctrl+D（反之）。查找下一个差异是 Ctrl+N，上一个 Ctrl+P。（所有快捷键都可以自行设置，比如我就是用 control+J/K/N/P 四个组合键，这样左手手指一直接住 control 就好了，右手手指也不需要滑动太远——JKNP 隔得近，懒人大法 1。）

如果文本上某个字词两个软件都识别错误，自行在①/②/③区域修改就好了。另外该软件只能复制一行到另一边，所以如果一行出现两个错误，一个 ab 对，一个汉王对，那只能手动改一个了再复制到另一边了。

有时一个字（尤其一些生僻字）会被两个软件都识别错误，并且识别为同一个字，BC 只能看差异，所以是显示不出来的，只能靠读校的时候依照 PDF 改正。

小提示：如果在①②区域无法对文本进行编辑，将鼠标放到①或者②区域，并点击右键，勾选“完整编辑”即可，如下图（示范图中是英文）：



下图这种情况，左边的符号代表此处文本内有个空白行，右边在此处没有空白行。此时需要搜索双层 PDF 查询该处是否需要分段再进行修改（比如原书这里分段了，就留下换行符，以便后期合并的时候处理）。



bc 的时候可以只改一边的文档（即：以左侧/上方的文档为基准，如果左侧/上方是对的，则不进行任何修改，最终关闭的时候会提示问你要不要存到右侧/下方，点“是”即可，或者直接另存为一份文件——**注意存的是左侧/上方的文件**——所以如果你改的是右侧/下方的文件那就……），这样可以少改很多处，但如果不熟练的话有可能会遗漏，所以……自己看着办吧。（懒人大法 2。）

上面这些都看不懂也没关系，直接看视频吧（群文件——BC 操作演示.MP4）。

记得关闭或者离开电脑之前 Ctrl+S。不然……

第六步 合并段落 (TextForever)

bc 完的文本是按照 PDF 断行的（如前面 bc 界面内所示），所以需要合并回去，此处采用@捕鱼 的办法，用 TextForever 合并段落（当然也可以用别的）：



此处是根据前面 ab 识别后段落间留空行的设置来合并的，然鹅机器毕竟是机器，如果遇到 PDF 里面一排字最后一个符号是句号感叹号省略号问号之类的很可能被 ab 认为该处换段而在后面加了空行，此时合并之后也是错误分段，所以，需要依照 PDF 再检查一遍段落是否分对（如果两三句一段那种书可能工作量就很大了）。之后即可进行排版（前面说了这个步骤省略，想学排版我就帮不上忙了，大神也没法几句话讲清楚，建议自己上网搜教程学习）并读校——也就是二次校对——以读为主，发现错误标注一下，有空到电脑上对照 PDF 修改即可（bc 工作做好的话其实读校错误已经非常少了，所以并不麻烦。但机器不是万能的，要保证文本质量一定要进行二次校对。）

结束。祝愉快。（弄完一本书会很有成就感的，所以，加油你是最胖的！）