# 文本处理

1. 文本格式转换

通过使用VSCode分别以GB2312打开文本,并以UTF-8重新保存

1. 段落合并
2. 首先，对于日期较少的作品，手动去除标题、作者、日期以及后记部分，对于章节号和空换行，则通过VSCode文本替换功能配合正则表达式，完成批量替换。
3. 对于日期较多的作品集，日期多使用写/草(于)$日期$（于$地点$）的方式进行记录，因此使用正则表达式[写|草]?.?\w{4}年\w{1,2}月\w{1,2}日.{0,5}进行替换。
4. 对于章节号替换为空字符串。具体如下:

* 对于<天狗>中，章节格式为第一章 井(1)，正则表达式为第\W章\s+\W+\(\w+\)。
* 对于<朋友(散文集)>，章节名称不做替换。
* 对于除<天狗><朋友(散文集)>外的其他作品，对于数字章节中，数字部分只有一位或两位，则对应正则表达式为第\w{1,2}章，与此类似，相应节为第\w{1,2}节，合并后的正则表达式为第\w{1,2}[章|节]{1}；对于中文章节,则数字分为类似 一，十一，二十一 三种，相对应的章节数正则表达式为第[\W]?[十]?[\W]?[章|节]{1}，最终总体合并后的正则表达式为 第[\W]?[十]?[\W]?[章|节]{1}|第\w{1,2}[章|节]{1}，以此替换文本中的章节文本.

1. 将空行和多余空格替换为空字符串。具体如下:

* 因为不需要保留任何空行，因此使用正则表达式\n+进行替换；多余空格则是指至少两个空格的情况，因此使用\s\s\*确保不会使文本中类似于“1 2 3 4”这种正常空格也进行替换。合并后的正则表达式为: \n+|\s\s\*。