

# 网络书籍抓取器使用说明

本程序是用Python编写，封装了多个包所以文件尺寸较大！

界面

程序只有一个窗体，界面如下：



操作基本上按 ①~⑤顺序来就可以完成 (红框为关键步骤), 操作有误时会有提示。

【主要功能】可以提取指定小说目录页的章节信息并调整, 再根据章节顺序抓取小说内容, 最后再进行合并。抓取过程可随时中断, 程序关闭后也能继续上次任务。

## 特色功能

1、**章节调整**: 提取目录后, 可以进行移动、删除、倒序等调整操作, 调整将直接影最终书籍也将按调整后的章节顺序进行输出。

2、**自动重试**：在抓取时因网络因素可能会出现抓取失败情况，本程序可能自动重试直致成功，也可以暂时中断抓取（中断后关闭程序不影响进度），待网络良好后再重试。

3、**停止和恢复**：抓取过程中可随时停止，退出程序后仍能保证进度不受影响（章节信息将保存在记录中，可在下次运行程序后恢复抓取。**注意：需要先用停止按键中断再退出程序，若直接退出将无法恢复**）。

4、**一键抓取**：也称为“傻瓜模式”，基本可实现全自动抓取及合并功能，直接输出最终的文本文件。前面可能需要输入最基本的网址、保存位等信息（会有明显的操作提示），一键抓取也可以调整完章节后使用，将自动完成抓取及合并操作。

5、**适用网站**：已例入10个适用的网站（选择后能快速打开该网站查找所需书籍），也能自动套用合适的编码，也可对其他小说网站进行测试，如合用可手动添加到设置文件中备用。

6、**方便制作电子书**：可以在设置文件中加入每个章节名的前缀、后缀，给后期制作电子书的目录编排带来及大的便利。

---

## 操作步骤

先指定小说的目录页，如下图



## 我是至尊

作者：风凌天下

动

作：加入书架 直达底部

最后更新：2020-02-10 18:23:19

药不成丹只是毒；人不成神终成空！..... 天道有缺，人间不平；红尘世外，魑魅横行；哀尔不幸，恨而不争；冷眼红尘，无憾今生！..... 时值乱世，群雄并起；烽烟处处，山河破碎。九尊智囊云扬大难不死，潜心复仇；惊天智谋，踏破国仇家恨；铁骨柔肠，演绎爱恨情仇；绝世神功，屠尽人间不平；丹心碧血，谱写兄弟千秋！.....

推荐阅读：圣墟 龙王传说 三寸人间 天下第九 飞剑问道 我是至尊 凡人修仙之仙界篇 元尊 大道朝天 大龟甲师

《我是至尊》最新章节（提示：已启用缓存技术，最新章节可能会延时显示，登录书架即可实时查看。）

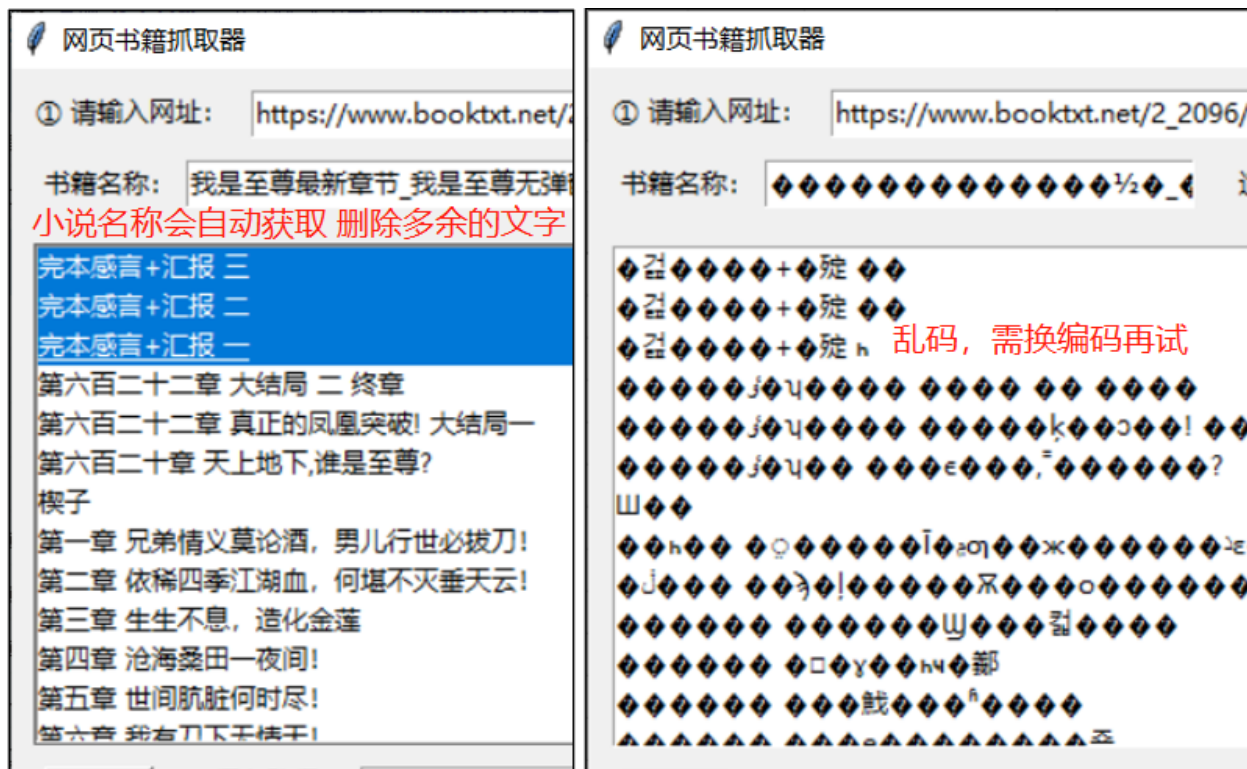
完本感言+汇报 三	完本感言+汇报 二	完本感言+汇报 一
第六百二十二章 大结局 二 终章	第六百二十二章 真正的凤凰突破！大结局一	第六百二十章 天上地下，谁是至尊？
《我是至尊》正文		
楔子	第一章 兄弟情义莫论酒，男儿行世必拔刀！	第二章 依稀四季江湖血，何堪不灭垂天云！
第三章 生生不息，造化金莲	第四章 沧海桑田一夜间！	第五章 世间肮脏何时尽！
第六章 我有刀下无情天！	第七章 王法不杀我来杀！	第八章 莲开一叶，千幻灵猴！
第九章 天意如刀，天意之刀！	第十章 好姓、好名字！	第十一章 你是个大麻烦！
第十二章 条件、成交！	第十三章 夜变、警告、纨绔	第十四章 这不是闪电猫！
第十五章 可敢与我一赌？	第十六章 西门万代、收网！	第十七章 你的陷阱，我的设计。
第十八章 我给你选更好的！	第十九章 家贼、绿绿！	第二十章 有钱大家赚，布局
第二十一章 做人不能天真！	第二十二章 丹心玉剑，震住！反震！	第二十三章 四大公子
第二十四章 我也没见过！	第二十五章 可以开始了。	第二十六章 伤心、离开、巡视
第二十七章 守护、绿衣、纨绔、路遇	第二十八章 谁跟你讲理？	第二十九章 你是我的偶像！
第三十章 惊了、狗不错、不明白	第三十一章 森罗庭	第三十二章 留下的方，头痛的云
第三十三章 一卷定生死！	第三十四章 惊吓、报恩、刺杀！	第三十五章 刺客、下令、赴宴！
第三十六章 这是摸了个什么东西！	第三十七章 失望、核查、出动！	第三十八章 门庭若市！
第三十九章 我真他么贱啊.....	第四十章 消息、驯狮、去战！	第四十章 消息、驯狮、去战吧
第四十一章 输赢、找人、蹊跷、踪迹	第四十二章 心中天火早燎原！	第四十三章 痴心，下毒，动手！
第四十四章 轻描淡写搞强敌！	第四十五章 谈谈心，做知己。	第四十六章 绝世忽悠

## 1、指定网址

先用浏览器在适用的小说网上搜索想下载的书籍，将该书籍目录页的网址（如：[https://www.booktxt.net/2\\_2096](https://www.booktxt.net/2_2096)）复制到最上面的文本框内，选择编码（该网站的编码为 gb18030，如不清楚的可先不管，在目录提取后发现是乱码后再换，反正编码也不多 ^\_^）。

## 2、提取章节目录

按【目录提取】将章节的信息提到到列表框中，如下图：



修整书籍名称，将多余的文本删除，当然不改也行，最后导出的书籍会以些作为文件名。

### 3、查看内容

可查看所选的章节内容，用于检测编码是否正确，也用来检测是否能正确抓取章节内容，如多个章节内容为空说明本程序不适用于该网站。

未选择章节时将查看第一个章节的内容，多选时则查看第一个选择的章节。

### 4、章节调整

章节列表右侧的一排按钮为调整章节用，能进行移动、删除、倒序等操作，可进行多选，个按钮作用为：

↑↑：置顶，将选择的章节全部移到最前面，可按住【Ctrl】键来跳选、多选；

↑：上移，将选择的章节全部上移一格，可多选但必须是连续的；

↓：下移，将选择的章节全部下移一格，可多选但必须是连续的；

↓↓：置底，将选择的章节全部移到最后面，可按住【Ctrl】键来跳选、多选；

Del：删除，将选择的章节删除，可按住【Ctrl】键来跳选、多选；

Clr：清空列表中的所有章节；

**倒序：**将列表中的章节进行倒序，即反向排序。

5、指定存放路径、抓取方式

选指定抓取章节的存放目录，用【浏览】选择目录，建议手动创建一新的空目录，抓取时所有章节都将存放在该目录中，如有其他文件容易造成麻烦。

抓取记录和出错记录也将存放于此目录中，要继续前面中断的抓时，需要再次指定存放目录（重启程序后不会保存指存放路径，需重新指定）。

程序可能会因网络问题出现抓取失败的情况（抓取时状态栏会有提示），针对抓取失败程序有3种方式供选择：

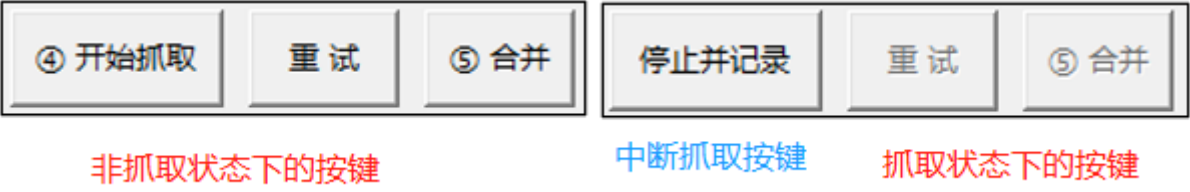
- （1）自动重试直至成功：抓取完后自动跳回到失败的章节重试，若还存在失败自动重试，直至全部抓取成功，也称为挂机模式 ^\_^；
- （2）间隔 n 秒后重试：以上模式类似，但每次重试会间隔 n 秒后再试，防止个别网站有时间访问限制，秒数可在设置文件中手动修改，默认为 3 秒。
- （3）手动重试：每一轮抓取完后会询问是否重试，需要人工干预，但有利于分析抓取情况再做决定，如抓取失败次数较多的建议先停止，过段时间再重试。

注：抓取过程中可以随时改变重试模式

6、开始抓取、重试

按【开始抓取】键开始抓取，抓取后该键会变成【停止抓取】按键，可中断抓取（不影响抓取进度），要恢复时注意要用【重试】按键进行，若按【开始抓取】将又全部重新开始抓取。

停止抓取后，可以退出本程序，在下次运行时，重新指定存放目录后，可接着继续抓取（注意还是要用【重试】按键进行恢复抓取）。未抓取章节的信息会存在原指定目录的“@ error.log”文件内，不同书籍目录的未抓取信息将不会相互影响，因此可以成片成片地中断抓取 ^\_^。





## 7、合并

全部章节抓取成功后，可以用【合并】按键将所有章节合并到单一文件中。其实本功能为比较简单的文本文件合并功用，即把指定目录中的所有文本文件 (\*.txt) 文件按字符顺序进行合并，因此合并功能可以抓取完章节后的任意时间进行，也可用来临时合并其他文本文件用。

合并完后将询问是否删除章节文件，**建议删除**（留着也没啥用，删除完了也容易找到正主 ^\_^）

合并完后可以对其进行二次加工，如去空行、删除或增加段前空格、过滤广告文字等等，这里建议大家用 EmEditor 工具进行，制作电子书建议大家用“calibre”或“Sigil”。

## 8、一键抓取

也称为“**傻瓜模式**”，基本可实现全自动抓取及合并功能，直接输出最终的文本文件。一键抓取前复制小说目录页的网址，再指定保存的位置后即可（也可以直接进行一键抓取，会有明确的操作提示）。如章节顺序较乱的小说，也可以先【目录提取】进行调整后再进行一键抓取，此时本程序将自动抓取所有章节后再进行合并（中途仍可以随时停止，但停止后就只能用【重试】接着下载抓取后再【合并】了）。

---

## 设置

要工具的设置全在 setup.ini 文件中（同本程序目录），第一次运行时工具会自动生成该文件，可用记事本打开进行修改，里面有较详细的说明。下面对一些细节进行说明：

**DIGITAL\_WIDTH=章节序号位数（整数）**

即保存章节时文件前缀数字的位数，如“0001 第1章 XXXX.txt”，前面的0001 就表现为 4 位数，要根据小说的章节总数来定，如不超过1000章节的可定为3，如超过10000章节的小说（怎么可能）就需要设成 5 了。**目的是为了保证章节能按正确的顺序来合并。**

**CONNECT\_TIMEOUT=抓取网页超时时间（可以为小数，单位：秒）**

为了防止抓网时卡死，可以设置连接读取的超时，此处为设定超时的秒数，可在1~120之间设定（超出将自动取极值），可以为小数，默认值为15。

CHAPTER\_PREFIX=章节名前缀

CHAPTER\_SUFFIX=章节名后缀

这两项设置主要是为了方便后期制作电子书用的，如不制作则要无视。如习惯用 calibre 软件的，则前缀可设为：CHAPTER\_PREFIX=# 或者

CHAPTER\_PREFIX=##

该软件可根据 #、## 字符来生成一、二级目录，后缀为空即可。

习惯用 Sigil 软件的，前缀可设为

CHAPTER\_PREFIX=<hr class="sigil\_split\_marker"/><h1> （一级目录）

或

CHAPTER\_PREFIX=<hr class="sigil\_split\_marker"/><h2> （二级目录）

相应的后缀为

CHAPTER\_SUFFIX=</h1> （一级目录）

或

CHAPTER\_SUFFIX=</h2> （二级目录）

合并后方便在 Sigil 软件中进行章节切割和生成目录。

name=网站名称（编码）

site=网址

手动加入适用的网站，添加时需要成对出现，建议在文件最后面插入，注意编码要填写正确，到时方便自动套用。

要查看新网站是否能适用本工具，方法也很简单，符合两条件即可：1、粘贴入目录页的网址，能用【提取目录】获取章节信息；2、能用【查看】抓到章节内容。只要符合这两条件的就可以将网站添加到设置文件中。

注：每次修改保存后再运行程序方能生效，数值型的只能填写数字，否则可能出错

---

## 常见问题

在获取章节信息或抓取章节内容时，可能会出现卡顿情况，若出现这种情况说明网络连接不稳定造成，由于网络问题较为复杂，出现此情况也不为鲜。建议过段时间来尝试抓取（抓取过程中出现卡顿可按【停止并记录】，过段时间后再用【重试】恢复继续抓取）。

抓取过程中程序窗口操作延时，这种情况属正常现象，本程序用 Python 编写，运行效率上会慢很多，抓取中存在大量的字符运算过程，也将导致延时现象，占 CPU 率较高。

---

版本 1.40    2020年5月9日