自然语言处理实验报告

实验一

学 院： 信息工程学院

指导教师： 孙 媛

班 级： 19计算机2班

学生姓名：王 君 、高 雯 慧、路 琪

学 号：19040119、19040126、19040127

日期：2020 年9月30日

目 录

一、实验内容

二、具体步骤

三、主要程序模块和框架

四、实验结果

五、总结

六、程序亮点

摘 要

历史的长河带走了时间的碎片，留下的却是永恒的记忆。如今网络小说的发展前景良好，从1991年4月，北美留学生创办第一份中文电子周刊《华夏文摘》，同月，少君的《奋斗与平等》在华夏文摘发表，成为最早在网络上发表的小说，

到如今的起点中文网、晋江文学网、黑岩网等等很多知名的网络小说的网站，不得不说网络小说的发展十分迅速。而此次实验，我们小组通过Python爬虫爬取晋江文学网，来获取网站上小说的一系列的信息，可帮助读者更多的了解某本小说，以免“踩雷”，也可以让读者认识更多的小说作者；同时，小说作者也可以通过爬取的小说简介和内容等信息来确定当下读者最喜欢的小说题材，具有特殊的意义。

**实验一 编写网络爬虫软件爬取文本语料**

**一、实验内容**

学会使用python（或其他编程语言）编写网络爬虫软件，能对网站格式进行解析，爬取相关内容并过滤，建立语料库。

要求：能够自动爬取多个网页，过滤掉图片、视频信息，存取文中中的title，context，URL等关键文本字段，存取到不同的txt文本和文件夹中。

**二、具体步骤**

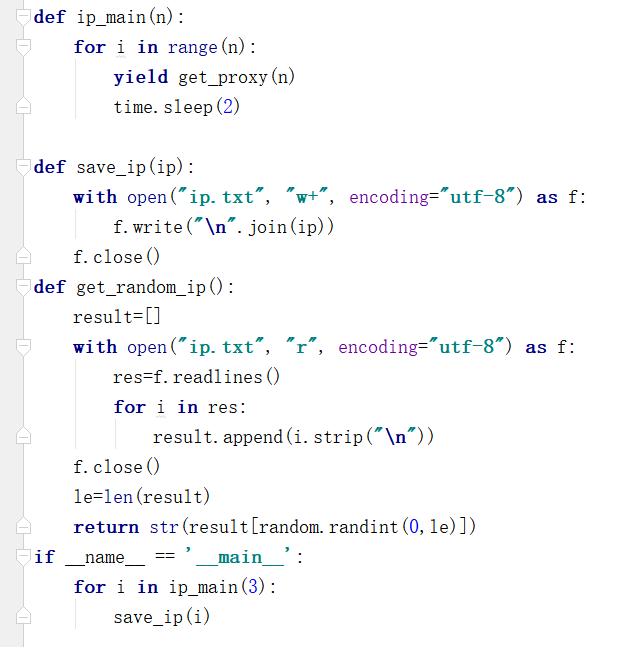
（1）观察网页：

打开晋江文学网页面**http://www.jjwxc.net/**和**http://www.jjwxc.net/bookbase\_slave.php?page=**

在页面空白处点击右键，点选最下方的“查看页面源代码”选项，获取网页的源代码，并采用xpath进行解析。

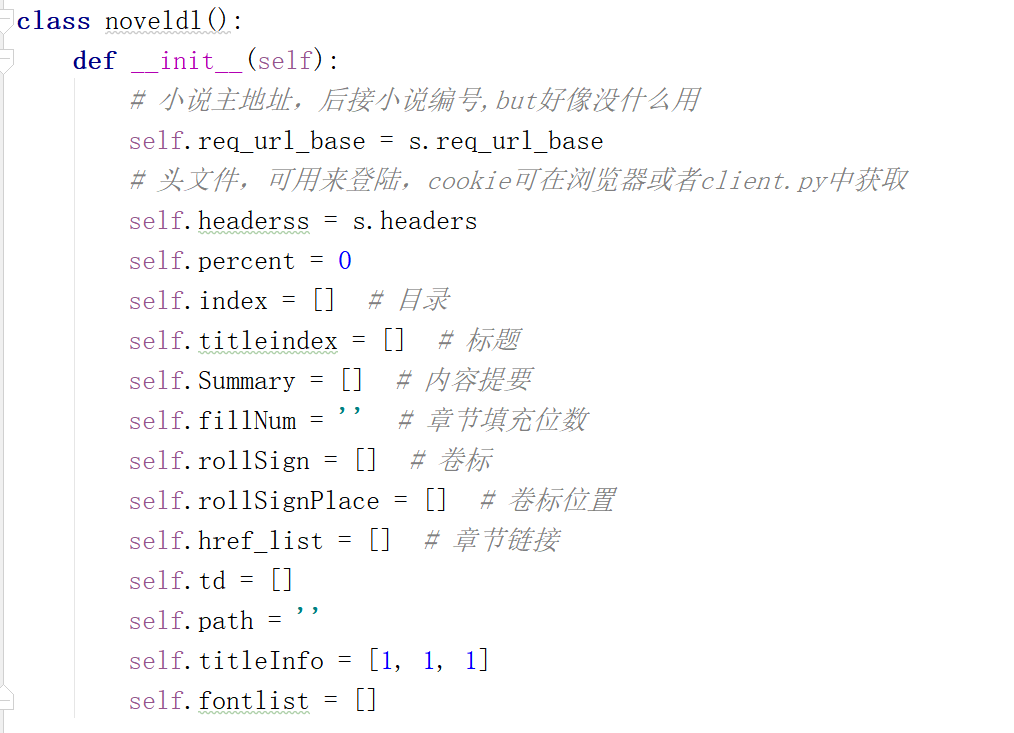
（2）编写获取代理ip的代码：

获取免费代理ip，并将其存储，然后再随机抽取一代理ip，最后爬虫网页的工作是使用其代理ip完成的。

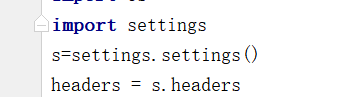
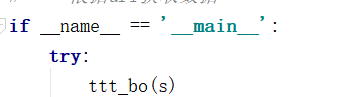


（3）编写爬取小说信息代码：

首先创建类setting()，其中包括晋江文学网页面的链接以及小说各种信息变量的定义。其次在代码main\_txt中使用模块settings来调用函数settings()，并创建了另一个类noveldl()，其中定义了小说具体章节各种变量的定义，如小说目录、标题、内容提要等变量。而且定义了一系列函数来完成小说标题、内容、URL的爬取以及存储。最后在代码free\_books中完成调用，实现最终的内容爬取与存储。如图：





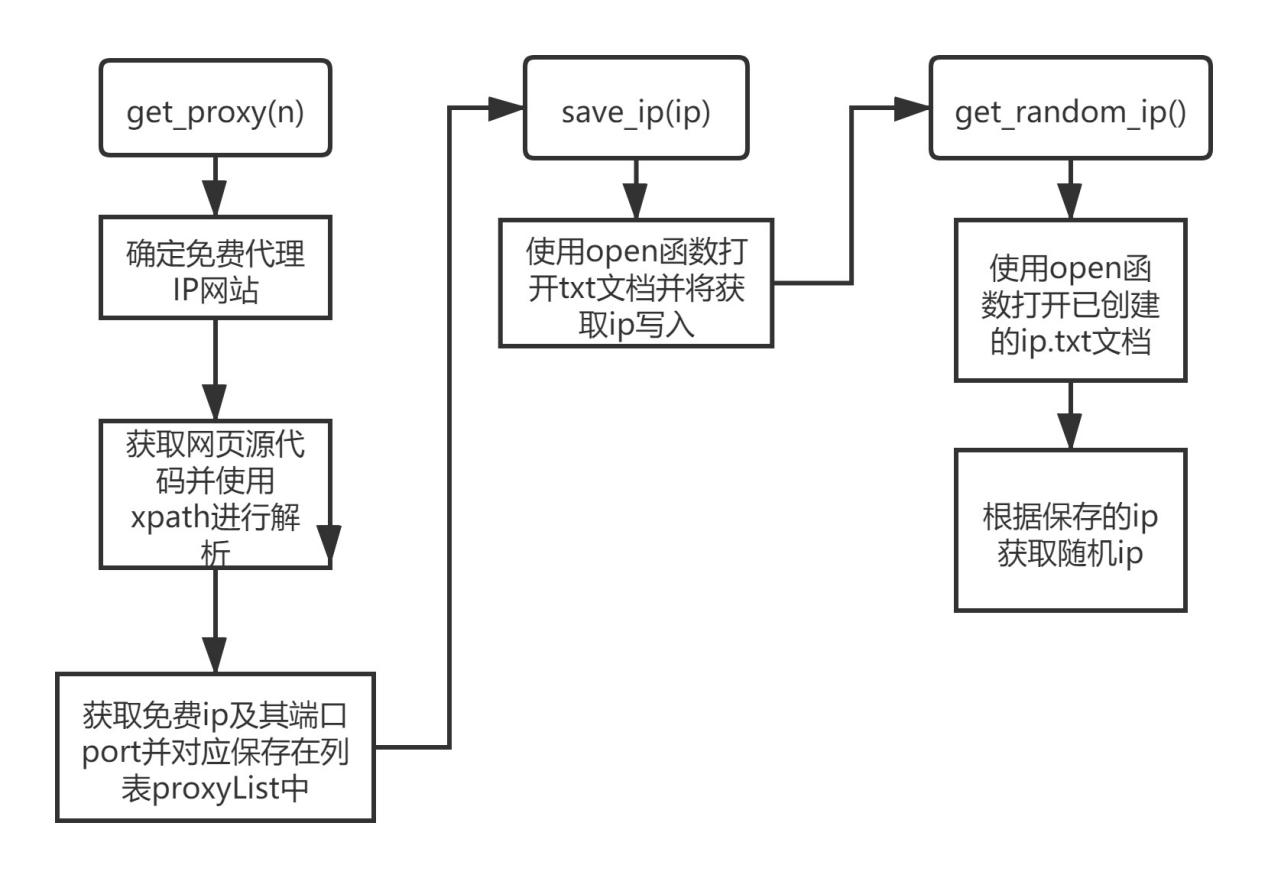


（4）整合代码，爬取网页信息:

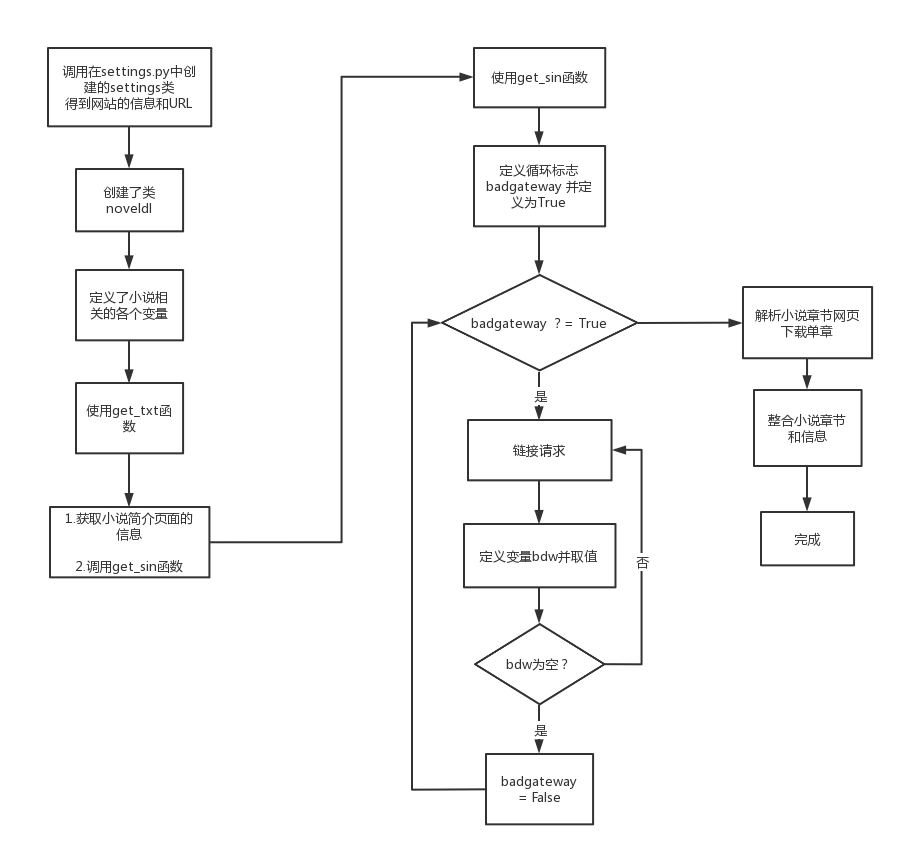
首先进入晋江文学网页面，看到该页的小说标题，代码可以依次爬取各个小说的信息包括URL（保存在文件夹txt中），并可以爬取各小说的免费章节，最后将其整合到文件夹result\_txt中。

**三、主要程序模块和框架**

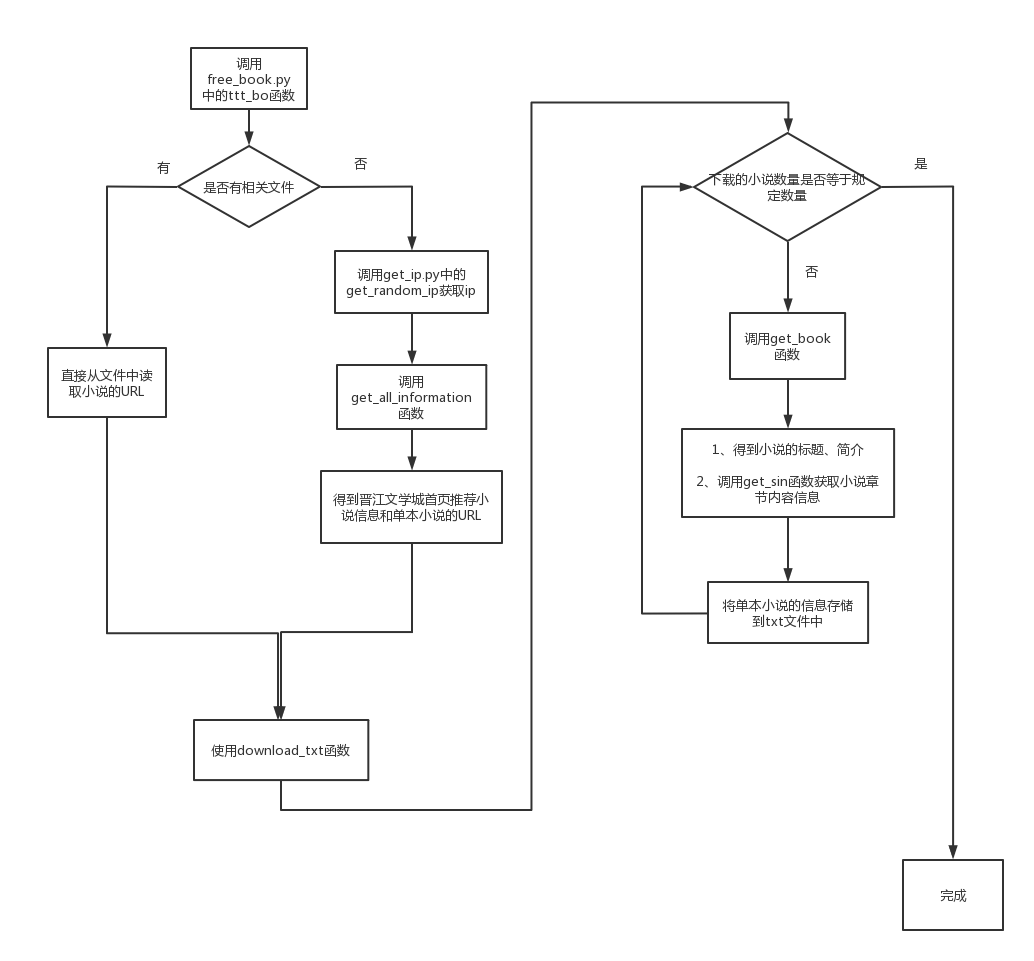
该程序主要包括三个大的模块，一个用来爬取晋江文学城首页推荐小说的各种信息，包括小说名、作者名、类型、风格、进度、字数等等；一个用来爬取单本小说的标题、简介和小说章节内容等；另一个用于ip代理。

①对于get\_ip.py程序，其主要作用是获取免费代理ip，并将其存储，然后再随机抽取一代理ip，最后爬虫网页的工作是使用其代理ip完成的。

②对于main\_txt.py程序，首先调用在程序settings.py中创建的settings类，得到网站的信息和URL。然后创建了类noveldl，在类中又创建了函数，其中在\_\_init\_\_函数中定义了小说相关的各个变量，在get\_sin函数中首先定义循环标志badgateway = True，在while循环中定义一变量bdw，链接请求。如果bdw为空，则标志badgateway = False，退出循环，然后下载小说单章内容。get\_txt函数第一步是获取小说简介页面的信息，其次调用了get\_sin函数来完成小说章节的下载，最后将小说的内容整合保存在了txt文档中。

****

③首先调用free\_book.py中的ttt\_bo函数，然后判断是否有相关文件，如果有,说明已经调用过get\_all\_information函数，直接从文件中读取小说的URL；如果没有相关文件，调用get\_all\_information函数，在get\_all\_information函数中调用get\_ip.py中的get\_random\_ip获取ip，得到晋江文学城首页推荐小说信息和单本小说的URL。最后使用download\_txt函数，循环调用get\_book函数，在get\_book函数中使用main\_txt.py中的get\_txt函数得到小说的标题、简介并且调用get\_sin函数获取小说章节内容信息，最后将小说相关信息存储到txt文件中。

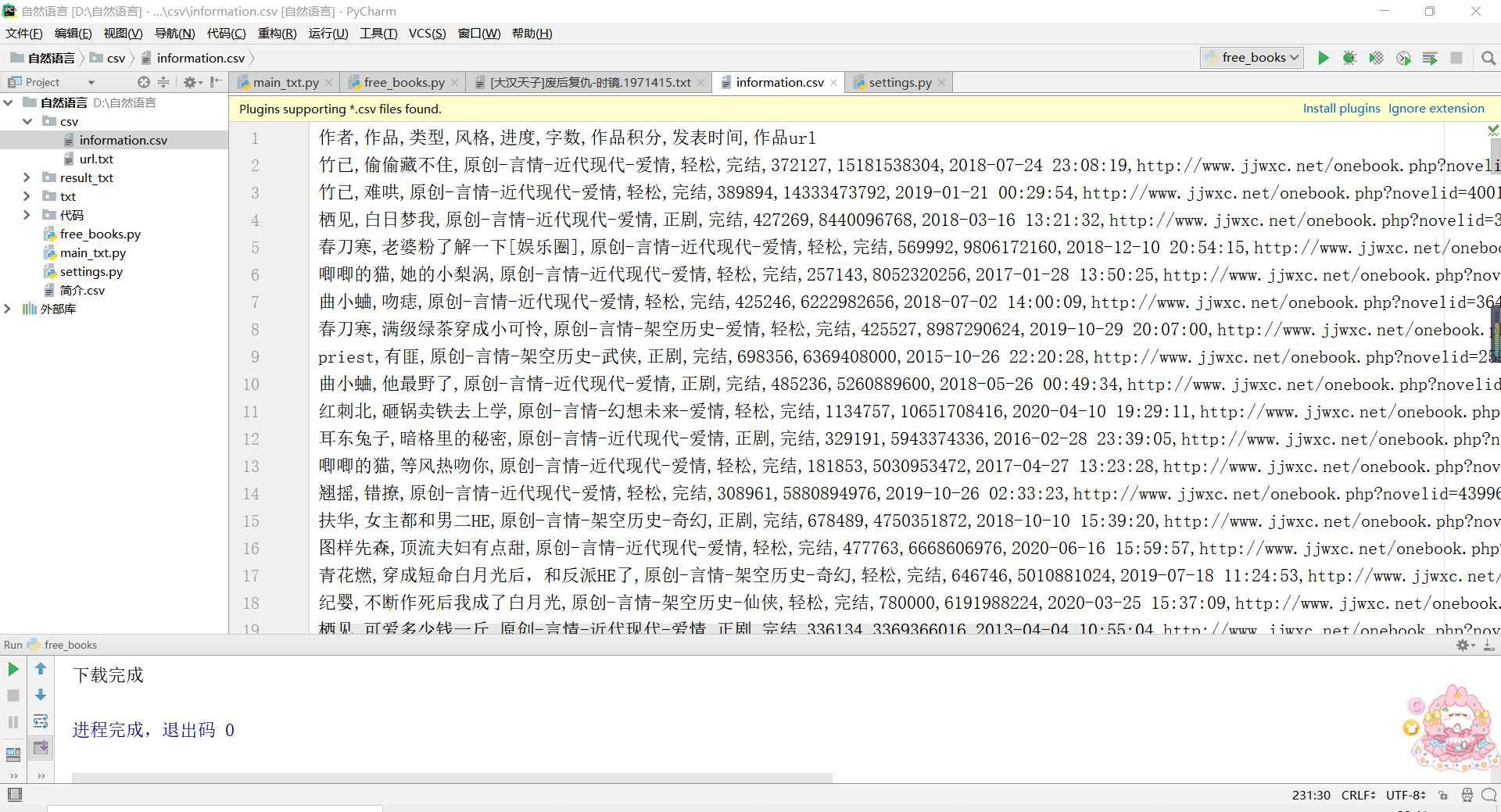


**四、实验结果**





所有的小说信息（作者、作品、类型、风格、进度、字数、作品积分、发表时间和作品url）存储在简介.csv文件中。



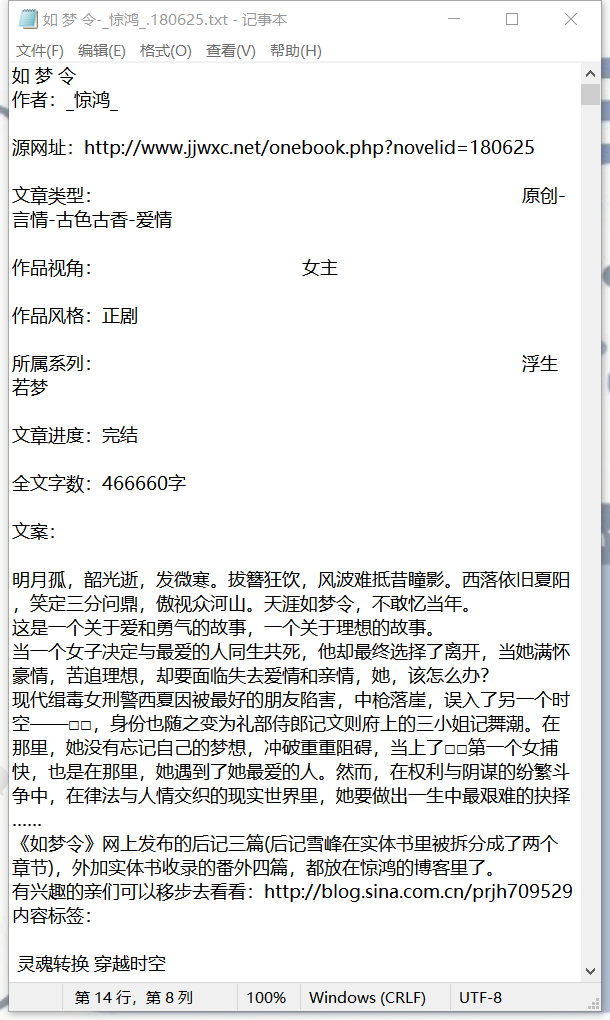
**爬取内容存txt文档中(在result\_txt文件夹中)**

（由于小说过多，爬取小说章节的时候只爬取了一部分。）



**文档中的内容如下（以《如梦令》为例）：**

（由于版权原因无法爬取到付费章节）



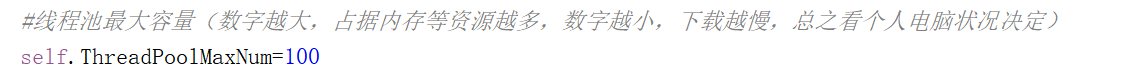
**五、总结**

通过此次试验，学会使用python编写网络爬虫软件，能对晋江文学网格式进行解析，爬取相关内容并过滤，建立语料库。同时该爬虫软件可以使用代理ip自动爬取多个网页，过滤掉小说封面图片，存取文中的小说标题、简介、章节内容和URL等关键文本字段，存取到不同的txt文本和文件夹中。

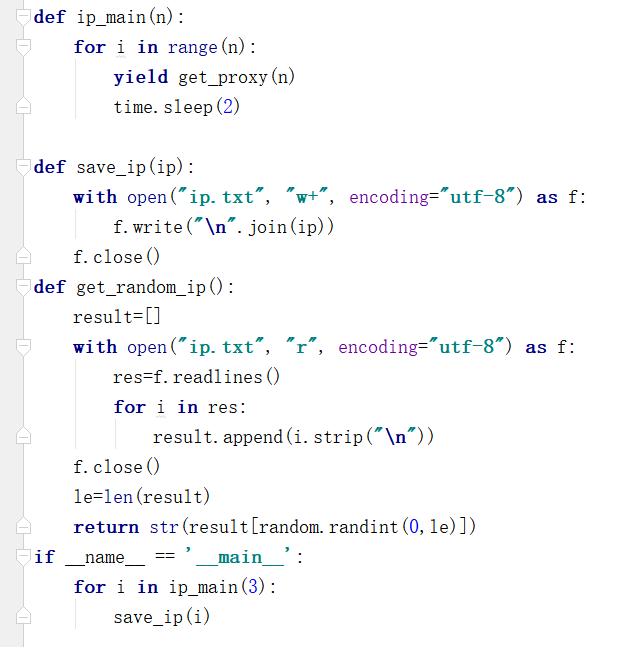
**六、程序亮点**

（1）在爬取章节内容的时候使用了线程池，让不同请求的等待同时进行，可以大大提高爬虫运行效率。

（2）可以根据个人电脑状况调节ThreadPoolMaxNum的大小，来控制下载速度。



（3）使用了ip代理，使用新的IP地址来代替自身IP地址进行爬虫操作，保护隐私，可以防止被黑客攻击，起到保护个人信息的目的。



**附网页图：**

晋江文学城首页：



以《偷偷藏不住》为例，小说页面：



免费IP代理页面：

