

****

信 息 学 院

软件工程系

《开源软件》

实验六 实验报告

**实验题目 如何利用Python实现电子书自动注解数据爬取**

**组 员 学号19720172203875 姓名 宋澳龙**

# 一、实验内容

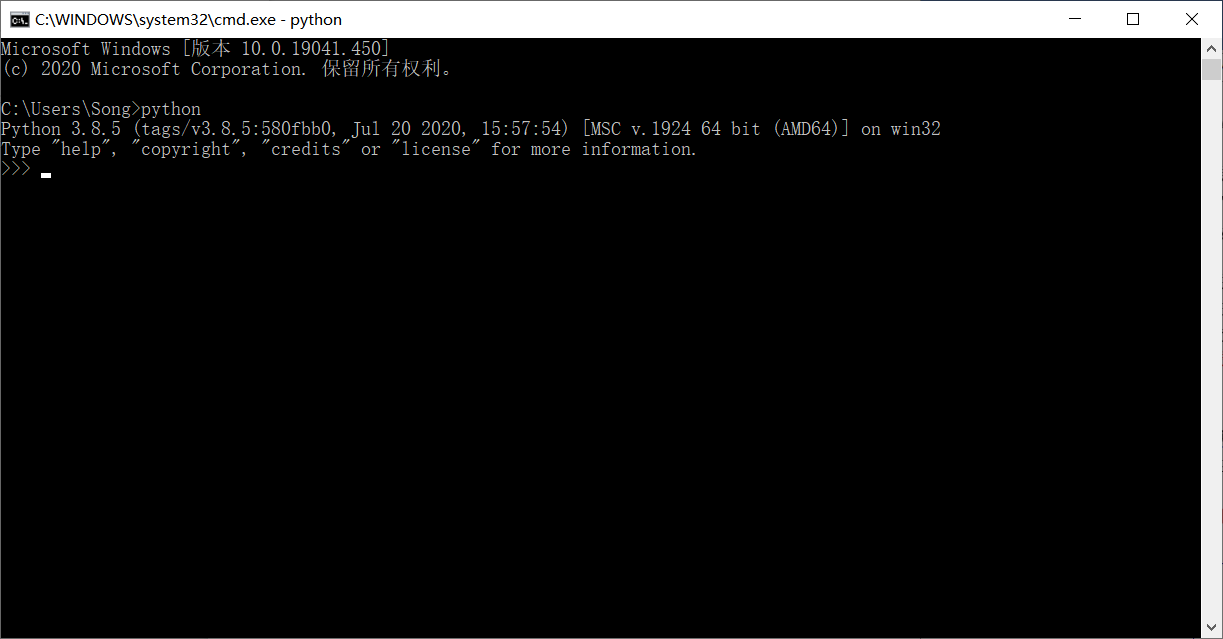
1、下载实验测试模块；

从大数据平台的开源软件课程页面下进入“实验”-》“提交”-》“实验附件”，附件当中有一个“lab6.zip”的压缩包，该压缩包里面包含了所有实验内容相关的测试代码：“jiebaTest.py”，“data.txt”和“weiboSpiderDayKWS.py”。

2、安装Python并测试相关代码，具体步骤如下

1）安装Python3最新版本并测试是否安装成功，具体可参考“L12-社交阅读系统电子书自动注解的实现之数据爬取”课件解决此问题。

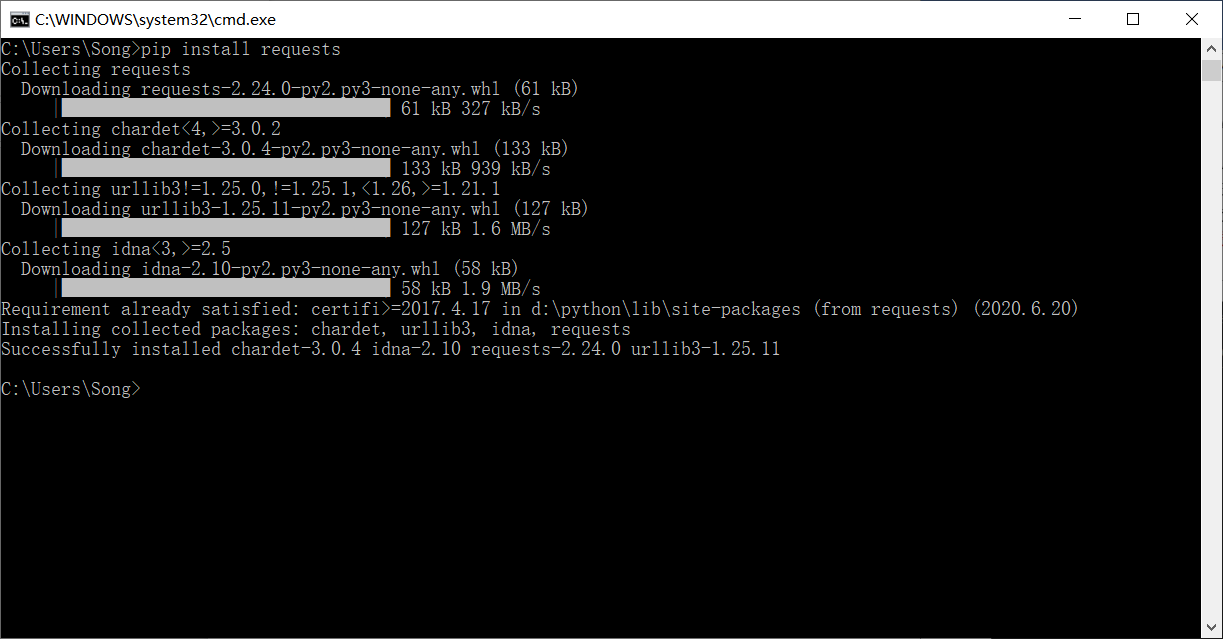
（请具体描述安装和测试步骤，并添加显示结果截图）



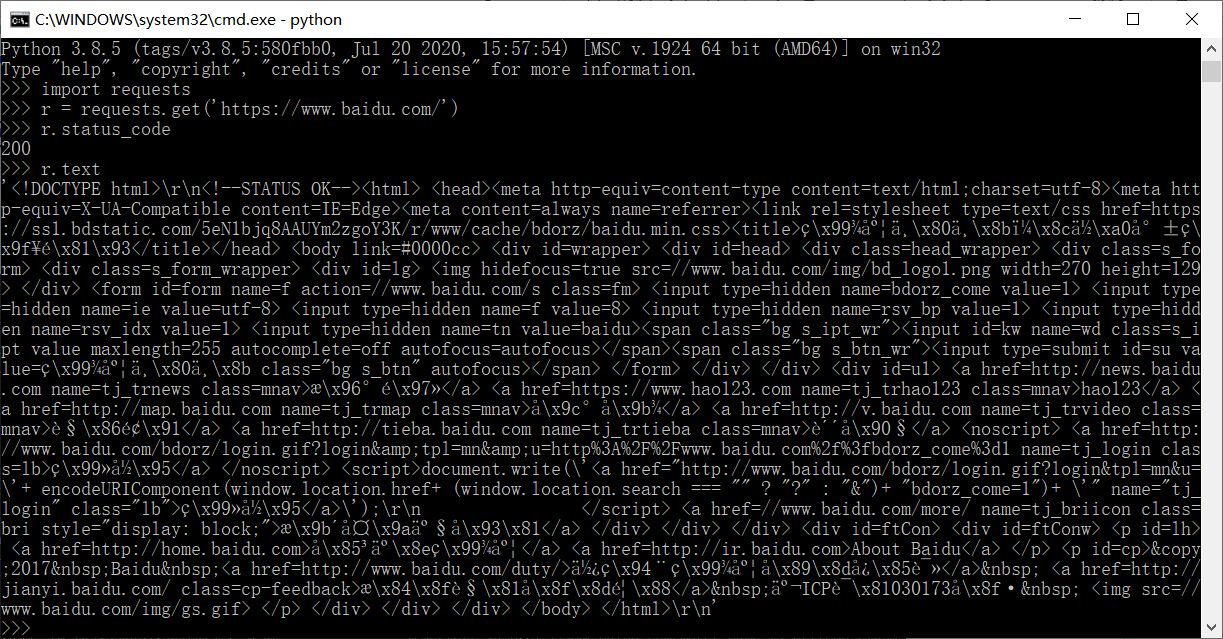
2）安装requests第三方库并测试是否安装成功，测试如何爬取一个给定URL地址的pdf文件并存储成本地pdf文件。具体可参考“L12-社交阅读系统电子书自动注解的实现之数据爬取”课件解决此问题

（请具体描述安装、实现和测试步骤，并添加显示结果截图）

安装requests库



测试requests库是否能够正确运行



3）实现大作业选定的主题的相关电子书数据爬取（数据量不小于1000），通过修改提供的样本代码爬取你们课程大作业所需的电子书数据。具体可参考“L12-社交阅读系统电子书自动注解的实现之数据爬取”课件解决此问题。

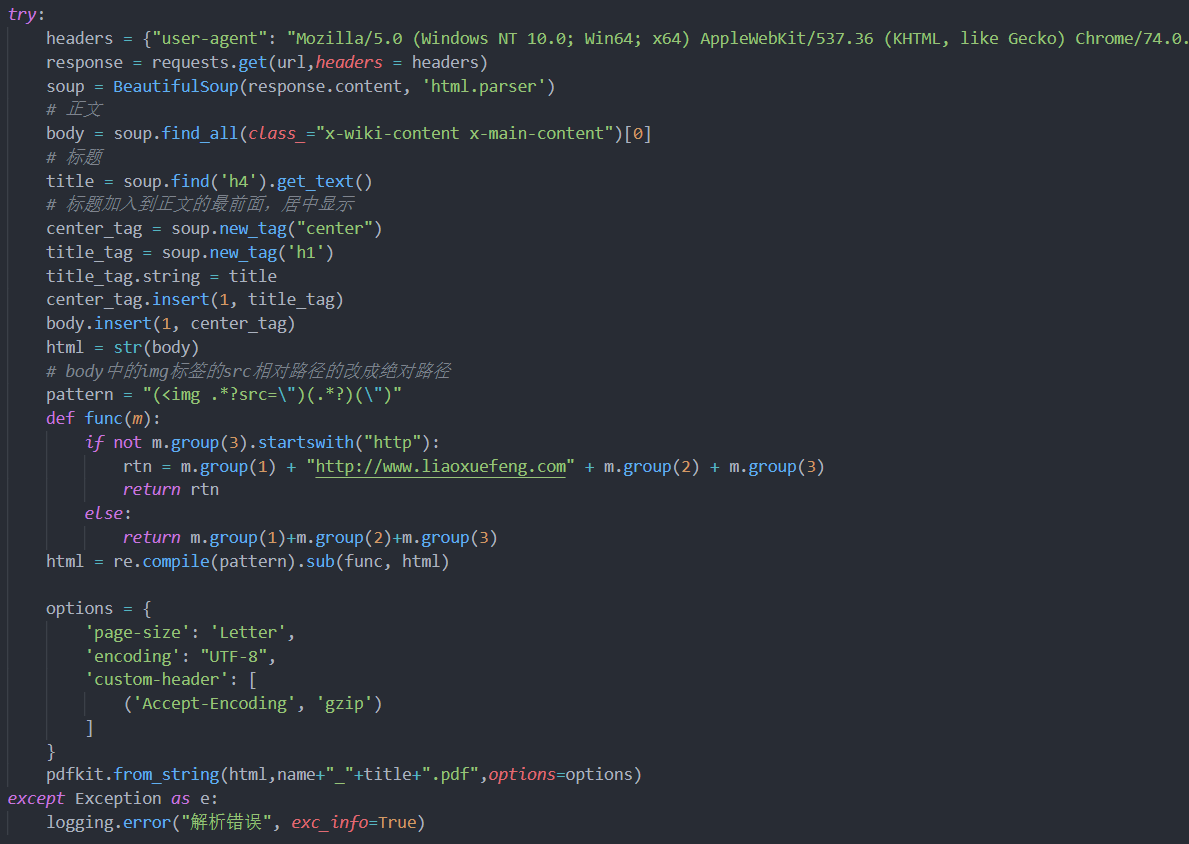
（请具体描述实现和测试步骤，并添加显示结果截图）

选择使用廖雪峰的官方网站中的教程内容作为爬取内容

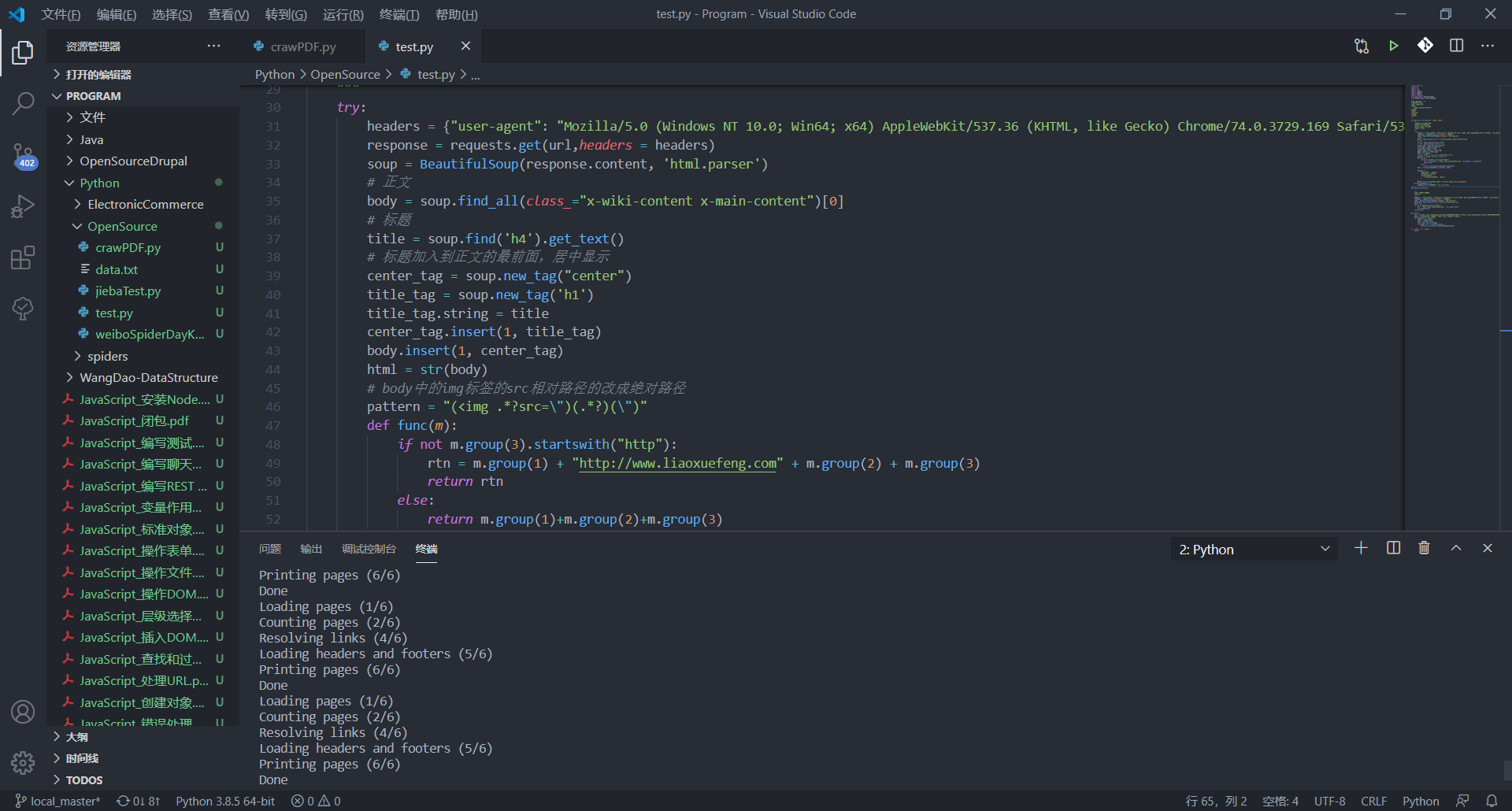


选择Java、JavaScript、区块链、python、SQL和Git的教程

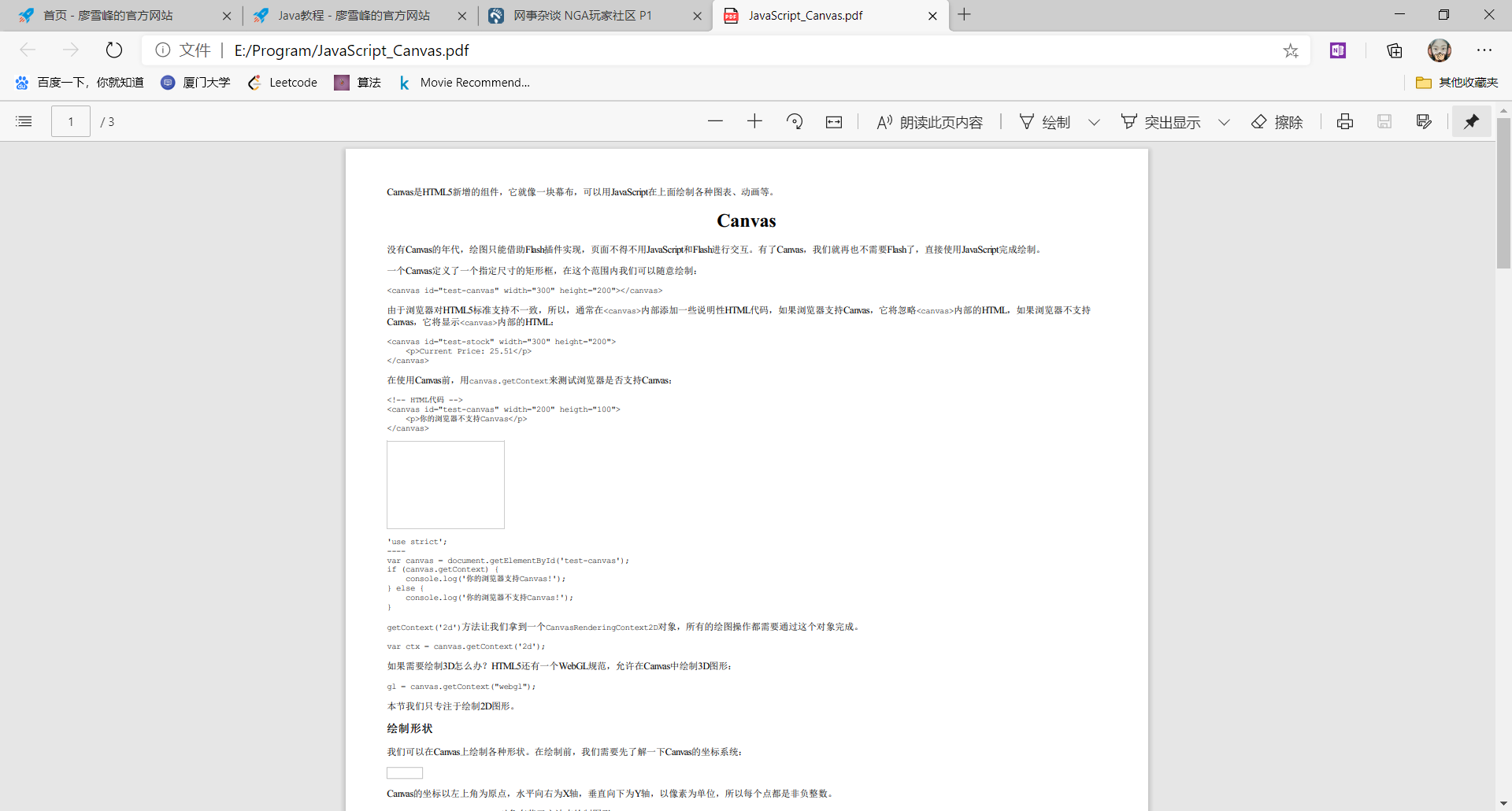
打开每个网页内容之后选择教程内容进行保存



文件执行过程：



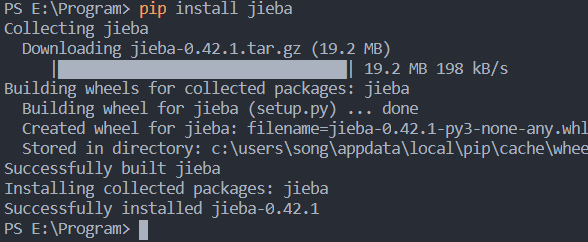
爬取PDF文件查看



4）安装中文分词所需的第三方库并测试是否安装成功，具体可参考“L12-社交阅读系统电子书自动注解的实现之数据爬取”课件解决此问题。从下载的lab6中查看“jiebaTest.py”和“data.txt”文件，测试如何基于一个已下载的电子书pdf文件进行中文分词，并将分词结果存储到本地文件中。

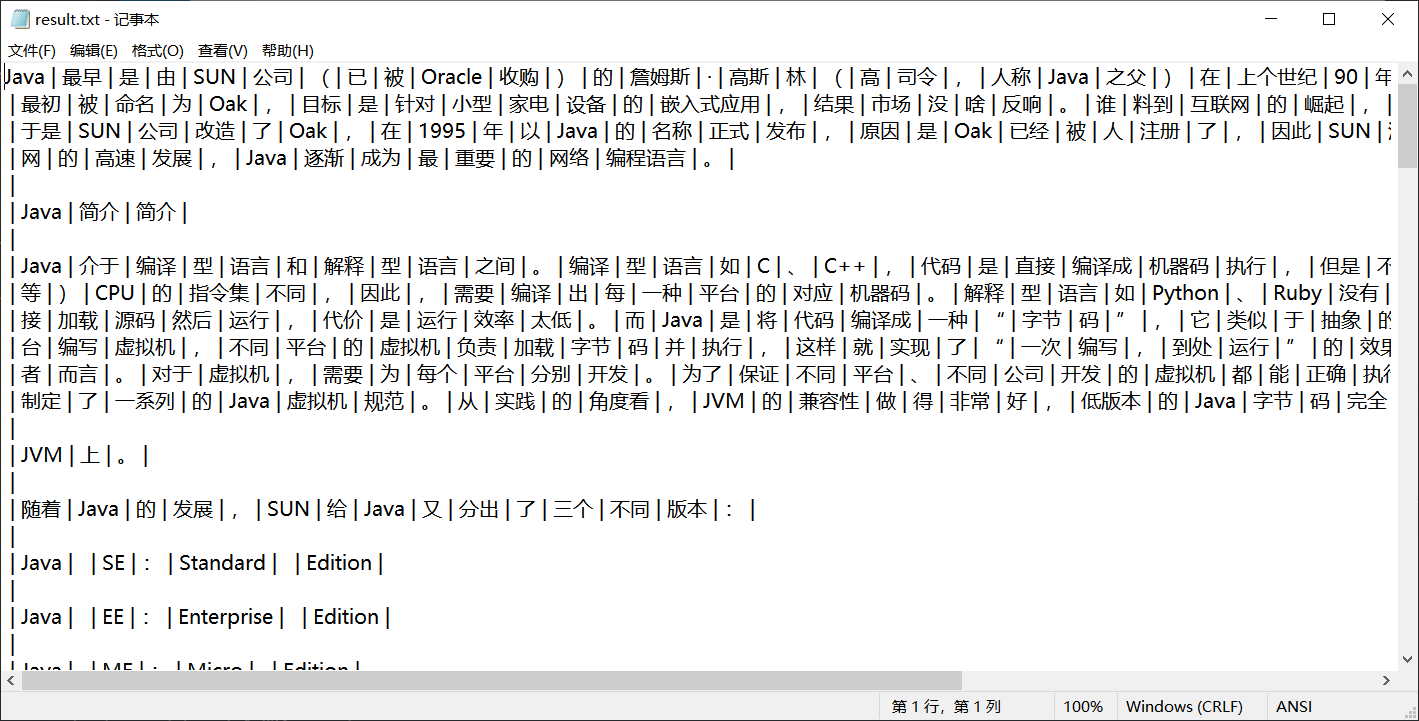
（请具体描述安装、实现和测试步骤，并添加显示结果截图）

安装jieba分词库



测试jieba分词效果

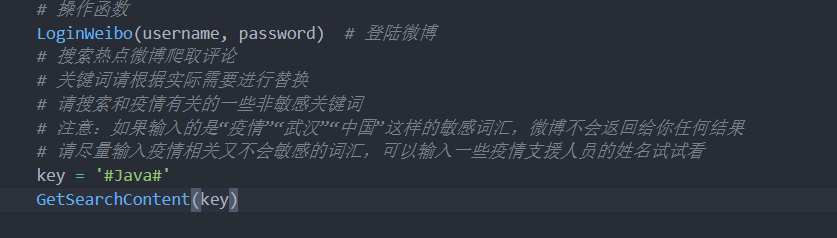


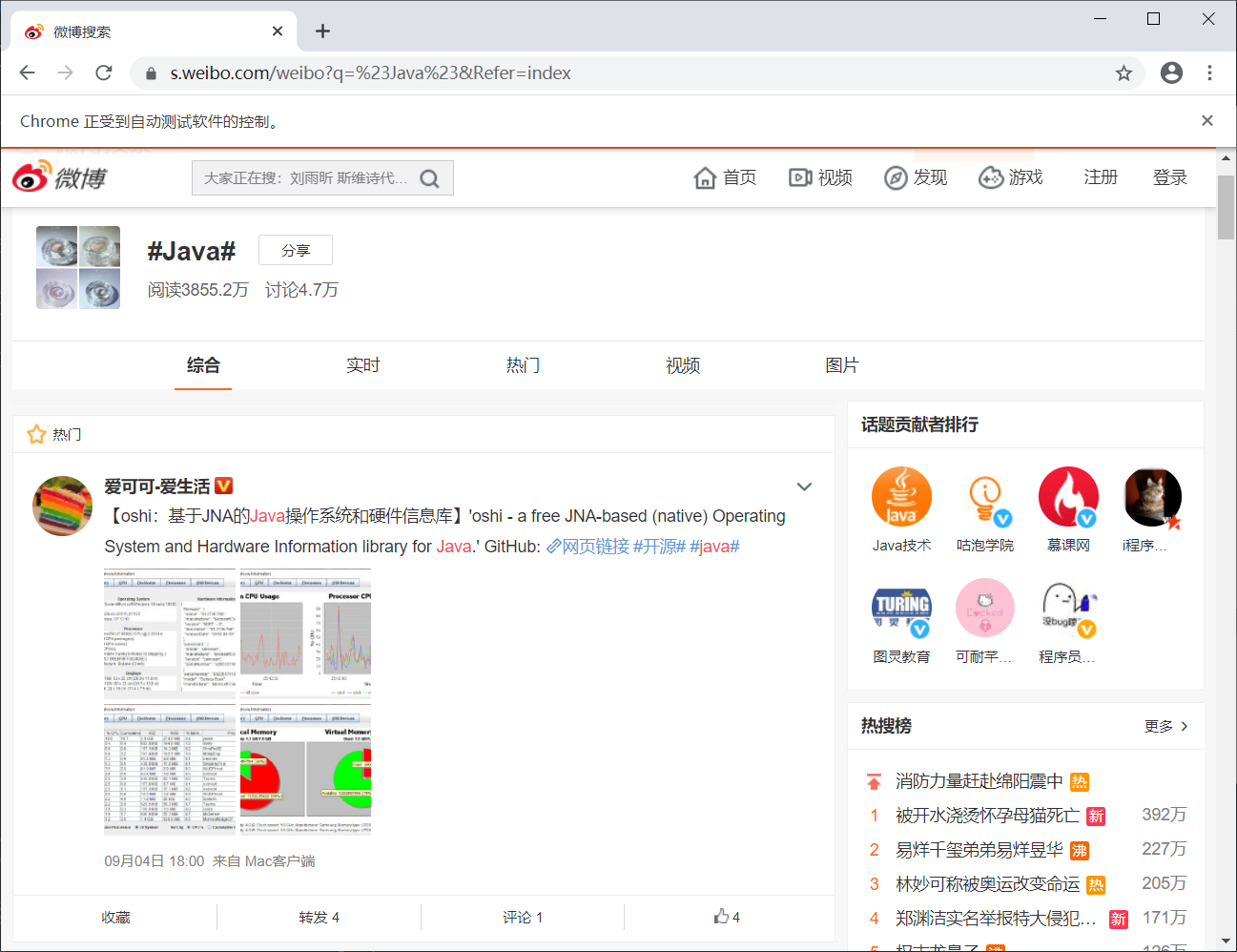


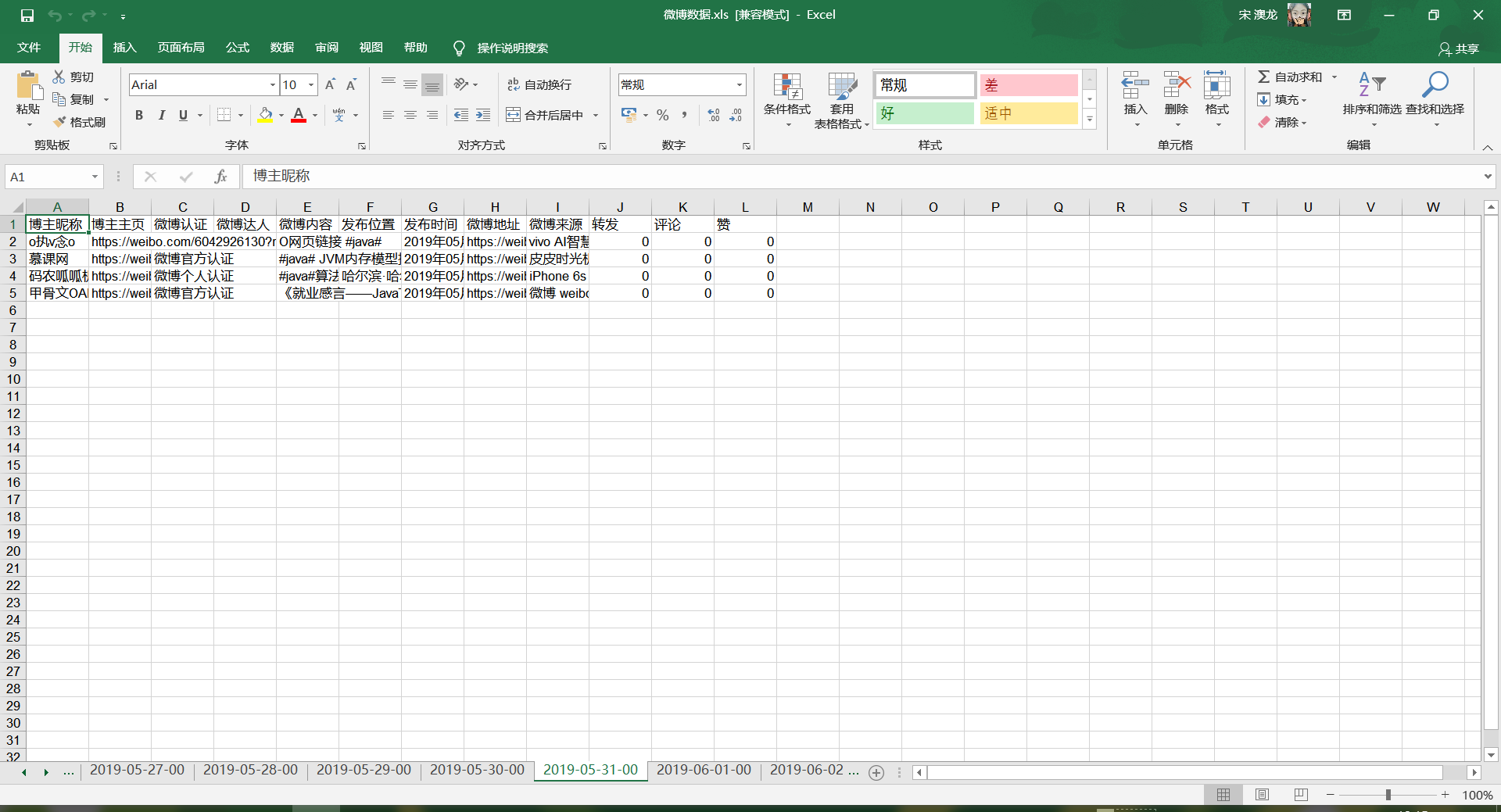
5）安装微博数据爬取所需的第三方库并测试是否安装成功，具体可参考“L12-社交阅读系统电子书自动注解的实现之数据爬取”课件解决此问题。查看“WeiboSpiderDayKWS.py”文件，测试如何基于关键词爬取微博数据并存储到本地文件中。

（请具体描述安装、实现和测试步骤，并添加显示结果截图）

修改已有代码，实现对指定关键词的爬取



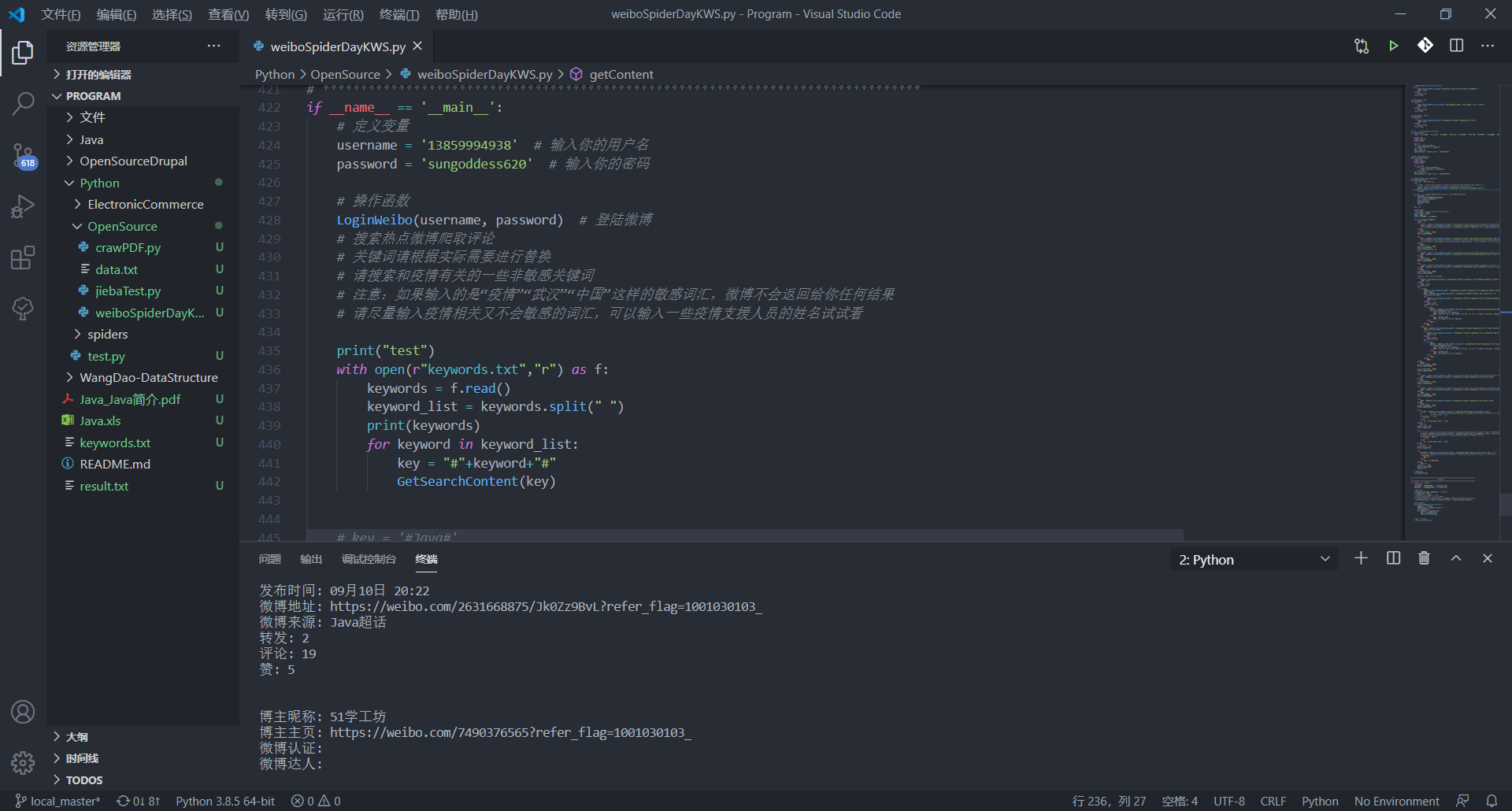




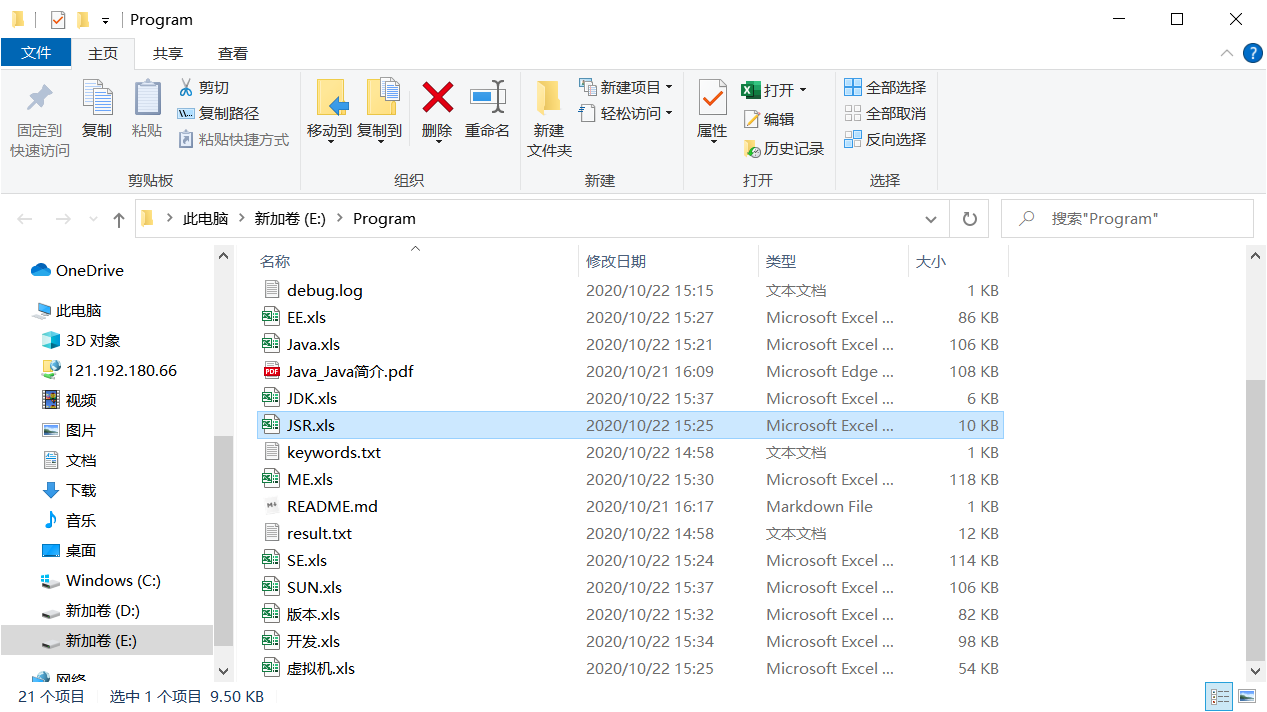
6）修改“WeiboSpiderDayKWS.py”文件中的源代码，选定一本已下载的电子书，实现基于电子书的分词结果提取重要关键词并实现基于关键词的相关注解数据爬取。

（请具体描述实现和测试步骤，并添加显示结果截图）

修改代码整合实现根据电子书完成对关键词的相关注解爬取



循环查找分词后的关键词为关键词查找微博相关注解并分别保存



3、备份Python源代码和数据文件

把相关Python文件和数据文件上传到Github项目（注意是个人项目）；

Github项目创建指南：https://jingyan.baidu.com/article/8cdccae9269b1f315413cde2.html

**注意：请按照要求完成步骤描述、截图和报告，10月28日23:59pm之前每个人把报告上传到大数据平台“开源软件实验报告提交”课程中的“实验六”(请将“报告+代码+数据文件”放入同一个文件夹打包成“姓名+学号+实验六.zip”)。**