**西安电子科技大学**

**安全前沿讨论班（I） 课程实验报告**

**实验名称 简单文本处理**

学 院 网络与信息安全学院 班 级 1618019

姓 名 曹寅峰 学 号 16020610025

实验日期 年 月 日

|  |
| --- |
| 指导教师评语：  指导教师：  年 月 日 |
| **实验报告内容基本要求及参考格式**   1. 实验目的   应用已经掌握的Python程序设计语言的相关知识，对文本进行处理。学习怎样统计文本中的单词频率，怎样通过文件路径获取文件内容，如何利用pandas库建立数据表格，以及如何利用matplotlib库画图。   1. 实验环境   Python 3.5 jupyter   1. 实验基本原理及步骤（或方案设计及理论计算）   将2个压缩包，Books\_EngFr.zip和Books\_GerPort.zip解压，然后将所有的文件放在同一个名为Books的文件夹中。Books文件夹要和所运行的.ipynb文件在同一个目录。接下来的实验步骤、要求已经在.ipynb文件中进行了说明，请按照要求完成。  首先可以用传统方法统计频率，创建一个结果字典，如果一个键不存在则创建它；存在的话则计数加一  也有比较方便的方法，利用collections里面的counter类，可以自动根据列表创建计数字典    通过open可以打开文件获取文件描述符，同时去掉里面的换行回车符，最后关闭文件  利用切片可以快速截取想要的部分字符串  统计字典里的唯一单词和单词总数，可以分别对字典求长度和利用.values()方法获取数量列表。  获取路径注意自己的实际路径就可以了。  这里注意么一层循环的意思，分别获取语言，作者，标题，然后返回    对其中的某些元素处理，利用strip去掉制定字符串，利用capitalize()转换为首字母大写格式   1. 实验结果分析及回答问题（或测试环境及测试结果）     对比发现两次统计结果一致  得出了正确的字母数量与位置    唯一的单词与总数也是正确的  路径获取正确    利用pd库进行多维统计  首字母大写，去掉txt  利用mplot库画图，制定x坐标为文本长度，y坐标为唯一单词数量，bo代表蓝色实心点    同理，可以同时画出多个情况同时存在的对比图，从而得出统计规律 |