附件1

Python编程 实验指导书

学院名称： 计算机学院

教师姓名： 杨大利，郝保水

适用专业： 计类

编写日期：

一、课程性质：专业选修课

实验学时：8学时

实验类型：课程实验

实验批次、每组人数：1批，每组1人

二、实验目的及要求

本课程是一门实践性强的课程，上机实验是学习和掌握本课程的重要环节。要学好本课程，应在掌握必要的python语言程序设计基础知识基础上，通过上机实验，将课堂所学理论知识与实际应用结合起来，熟练掌握调试程序的方法和编写简单程序的初步能力。

通过本实践教学环节的锻炼，学生应该能够：

1、熟练运用python语言的基本语法；

2、熟练运用典型的python语言程序运行环境，编写、调试和运行；

3、基本掌握结构化程序设计的思想；

4、能够运用python语言解决简单问题。

三、实验主要仪器、设备

实验环境：普通PC机，Windows操作系统，python3编程环境

四、实验内容

《红楼梦》前八十回与后四十回的语言特点分析：掌握需求分析方法，学习第三方库jieba的安装和使用。用字符串、列表、字典操作完成任务。

五、实验步骤

实验的基本思路和步骤

1. 预处理，把回目编号删除，然后根据标点符号等，把每一句话标记出来；
2. 利用开源分词组件“结巴分词”(<https://github.com/fxsjy/jieba>)对红楼梦全书进行分词，统计词频，把出现100次以上的高频词语标记出来；
3. 统计每一章高频词语出现的频次；
4. 前80回、后40回各选15回分别作为训练数据，通过机器学习得出用词特征；
5. 根据机器学习训练出的用词特征，计算每章是否和所属前80回或后40回相同，如果相同，则说明前80回和后40回很可能是两个人写的。换句话说，如果给出前80章的某一章，机器能算出它属于前80回而不是后40回；给出后40回的某一章，机器算出它属于后40回而不是前80回，则显然前80回和后0回遣词造句方面不同，从而得出不是同一个作者。
6. 利用相同算法分析《三国演义》、《水浒传》、《西游记》会得出什么结论呢？

参考文献

[1]黎晨(知乎).用机器学习判定红楼梦后40回是否曹雪芹所写[EB/OL].(2016-6-27)[2018-3-28].https://zhuanlan.zhihu.com/p/21421723

[2]楼宇(知乎). 用 Python 分析《红楼梦》[EB/OL].(2017-11-23)[2018-3-28] .https://zhuanlan.zhihu.com/p/29209681

[3]. 小智(知乎). 通过数据挖掘能分析《红楼梦》各回的真伪吗[EB/OL].(2016-6-24)[2018-3-28] . <https://www.zhihu.com/question/19768898/answer/107358050>

[4]. 永琳(知乎).有哪些现代分析方法可以用于解决“红楼梦续写争议”？有哪些例子？[EB/OL].(2016-6-5)[2018-3-28] . <https://www.zhihu.com/question/47059344/answer/104326214>

六、实验报告要求

实验报告：

在实验报告中应包括以下内容：实验目的、实验内容、实验原理、实验步骤、程序结果截图、源代码等。

七、实验成绩评定办法

实验考核：

根据运行结果、答辩情况、实验报告以及平时成绩综合评定。