实 验 报 告

**实验课程名称　　 Python程序设计**

**专业班级　 数据科学与大数据技术2班**

**学　　号　　　 22022402430**

**学生姓名　　　 　　樊宗豪**

**指导教师　　　　　 张辉辉**

**2023至 2024学年第 二 学期**

**潍坊学院计算机工程学院**

实 验 报 告

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **实验项目**  **名　　称** | **中文分词** | | **实验**  **类型** | **演示□　验证□**  **综合□　设计√** | |
| **实验室名称** | | **7325** | **实验日期** | | **2024.5.7** |

**一、实验目的**

1．掌握jieba分词模块；

2. 知道如何对文本进行分词并提取词语；

3. 学会使用分词模块进行程序设计。

**二、实验仪器设备**

一台配置好Python环境的PC机

PyCharm

**三、实验内容（步骤）**

[实验题目]

《西游记》中主要有四个角色：唐僧、孙悟空、猪八戒和沙僧，这些角色中哪个才是男主角呢？本实验案例需统计角色的出场次数，再按出场次数对角色排序，查看哪个角色排在首位。

[代码实现]

import jieba

# 打开并读取“西游记.txt”

txt = open(r"西游记.txt", "rb").read()

# 构建排除词库

excludes = {"一个", "那里", "怎么", "我们", "不知", "两个", "甚么",

           "只见", "不是", "原来", "不敢", "闻言", "如何", "什么"}

# 使用jieba分词

words = jieba.lcut(txt)

# 对划分的单词计数

counts = {}

for word in words:

   if len(word) == 1:

       continue

   elif word == "行者" or word == "大圣" or word == "老孙":

       rword = "悟空"

   elif word == "师父" or word == "三藏" or word == "长老":

       rword = "唐僧"

   elif word == "悟净" or word == "沙和尚":

       rword = "沙僧"

   else:

       rword = word

   counts[rword] = counts.get(rword, 0) + 1

# 删除无意义的词语

for word in excludes: del counts[word]

# 按词语出现的次数排序

items = list(counts.items())

items.sort(key=lambda x: x[1], reverse=True)

# 采用固定的格式进行输出

for i in range(9):

   print("{0:<5}{1:>5}次".format(\*items[i]))

**四、实验数据记录**



**五、实验体会、收获及及建议**

实验体会：在这次设计性实验中，我有机会学习和使用jieba分词模块，这是一个非常强大的工具，可以帮助我们理解和处理中文文本。通过对《西游记》中的文本进行分词和词语提取，我不仅掌握了如何使用jieba模块，还学会了如何对文本数据进行分析和处理。这个实验让我对自然语言处理有了更加深刻的理解。

收获：我的主要收获是对中文分词技术的实际应用。在统计《西游记》中角色出场次数的过程中，我学会了如何有效地使用分词技术来提取关键信息，并对数据进行排序和分析。这个过程不仅提升了我的编程技能，也锻炼了我的数据分析能力。此外，我也意识到了在文本处理中算法选择的重要性，以及它对结果准确性的影响。

建议：对于这类设计性实验，我建议同学们在编程时多考虑代码的可读性和可维护性。例如，我们可以尝试编写清晰的注释和文档，这样不仅有助于他人理解我们的代码，也方便我们自己日后的修改和优化。此外，我认为我们应该多尝试使用不同的自然语言处理库，这样可以拓宽我们的技术视野。最后，我觉得我们可以定期进行小型项目，比如分析不同文本的内容，这将有助于我们更好地理解和应用所学的知识。

**六、指导教师评分**

**成绩： 　　　　签名（电子）：　　 日期：**