**基于python的《天龙八部》文本可视化分析**

**摘要：**为了让读者能够更直观、简洁、高效的了解到小说中的基本信息，本文以《天龙八部》为例，通过python中的jieba库、pandas库等实现文本分词、去停用词及高频词统计，通过WordCloud、matplotlib 库绘制词云图、柱状图、饼图等实现可视化图表展示，能够帮助读者快速了解小说基本信息。

**关键词：**数据分析; 文本可视化; python; jieba库;

**1研究背景**

大多数读者在阅读小说前都会通过文章简介、前言等了解小说的主要角色、故事发生地点、故事背景等基本信息，进而选择自己感兴趣的小说。为了让小说的基本信息更利于查看，使用python对小说进行文本数据分析可让读者更直观、简洁、高效的了解小说的基本信息，进而对读者提供阅读建议。本文以小说《天龙八部》为例，通过具体的工具与算法，实现文本的数据可视化设计。

文本可视化技术综合了文本分析、数据挖掘、数据可视化、计算机图形学的理论和方法，为人们理解复杂的文本内容、结构和内在的规律等信息的有效手段。将文本信息以图形化的方式呈现出来，以提供更加直观的视觉效果和交互性。它可以帮助用户更快速地理解和分析大量的文本数据，并发现其中的规律和关系，使文本信息更加易于理解，可以将抽象的文本信息转化为可视化的图表或图形，使信息更加清晰易懂，更容易被人们接受和理解。提高效率，可以帮助人们更快速地分析和处理大量的文本信息，从而提高工作效率。文本可视化一般包括数据收集、数据预处理、文本分词与统计、可视化处理。

**2 数据和方法**

**2.1数据来源**

起点中文网是国内领先的原创文学网站，隶属于引领行业的正版数字阅读平台和文学IP培育平台——阅文集团旗下。本文所选用的《天龙八部》数据文本可在在起点中文网中下载并保存为txt文件。中文停用词表可在giuhub上下载。

**2.2分析方法**

jieba库是优秀的中文分词第三方库，中文文本可通过分词获得单个的词语。jieba库的分词原理是：利用一个中文词库，确定汉字之间的关联概率，汉字间概率大的组成词组，形成分词结果。jieba库分词有3种模式：精确模式、全模式、搜索引擎模式。本实验采用jieba库精确模式，精确模式试图将语句最精确的切分，不存在冗余数据，适合做文本分析。

词云（Word Cloud)又称文字云，是python的一个三方库，根据文本中的词频，对内容进行可视化的汇总是文本数据的视觉表示，由词汇组成类似云的彩色图形，用于展示大量文本数据。每个词的重要性以字体大小或颜色显示。主要用来做文本内容关键词出现的频率分析，适合文本内容挖掘的可视化。词云中出现频率较高的词会以较大的形式呈现出来，出现频率较低的词会以较小的形式呈现，词云的本质是点图，是在相应坐标点绘制具有特定样式的文字的结果。

Matplotlib是一个Python中广泛使用的绘图库，可以用于创建各种类型的静态、动态或交互式图表和可视化。它提供了一系列函数和工具，使得用户可以很方便地对数据进行处理和分析，并将其以图表形化呈现出来。Matplotlib库常常与Numpy库进行联合使用，这种组合广泛用于替代MatLab，是一个强大的科学计算环境，有助于我们通过python实现数据挖掘与数据可视化分析。

**3可视化设计与实现**

**3.1 初步分词**

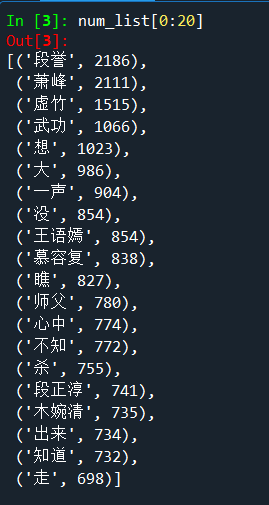
****

图1 全文词频统计

通过jieba分词和函数FreqDist（）可对小说全文进行初步分词与词频统计如图1。通过matplotlib库可将数据列表转化为词云，所谓词云就是对文本中出现频率比较高的“关键词”予以视觉上的突出，形成关键词渲染，从而滤掉大量文本信息，使得浏览的人可以快速、简洁的领略文本的主旨，如图2 所示。



图2 文本词云效果

由图2词云图可以清晰看出为萧峰、段誉和虚竹是最为突出的，说明萧峰、段誉和虚竹是《天龙八部》中的主要角色。紧随其后的王语嫣、慕容复、段正淳、游坦之、木婉清和鸠摩智即为《天龙八部》主要出场人物。

**3.2 主要角色分析**

通过遍历分词数据和FreqDist()函数可将出现频率较高的人物及出现次数存放于列表中，如图3。将列表数据转化为柱状图可更为清晰简洁的看出小说主要出场人物，如图4，段誉、萧峰出现次数超过2000次，其后为虚竹出现1515次，其余人物出场次数均不超过1000次。

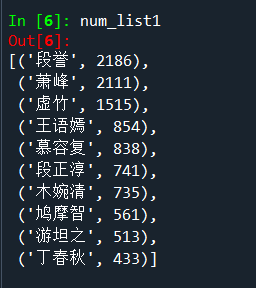
****

图3 出场人物词频统计

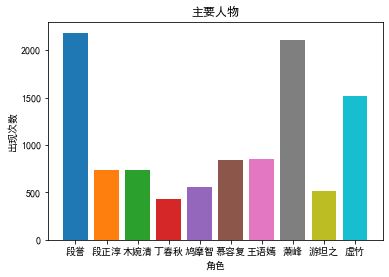


图4 主要出场人物

**3.3 主要势力分析**

通过遍历分词数据和FreqDist()函数可将出现频率较高的门派势力及出现次数存放于列表中，将列表数据转化为柱状图，如图5。

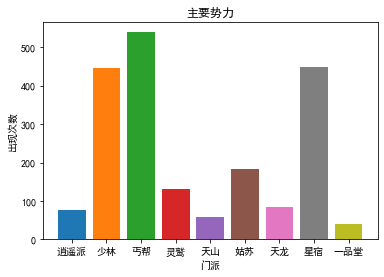
****

图5 主要出场势力

由图5柱状图可清晰看出”丐帮”出场次数最高，超过500次。其次为“少林”和“星宿”均超过400次。“姑苏”和“灵鹫”均超过100次。其余势力出现次数均少于100次。表明丐帮、少林、星宿为主要出场势力，与主要出场人物关系密切。

通过遍历分词数据和FreqDist()函数可将出现频率较高的国家及出现次数存放于列表中，将列表数据转化为饼图，如图6。

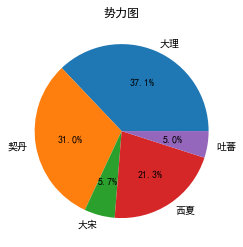


图6 出场国家

由饼图可清晰看出大理和契丹出场次数最多，均超过30%。其次为西夏，超过20%。最后为大宋与吐蕃，均不超过10%。由次可看出小说故事发生地主要为契丹、大理、西夏。

**4结论与展望**

本篇论文围绕小说文本分词与可视化设计进行研究，能为解决中文文本分析与可视化设计问题提供参考与借鉴。本篇论文仍有不足之处，如小说中的出场人物具有各色各样的称号与头衔，本篇论文并未做完全统计，且jieba库分词可能会影响部分角色人物分析，因此本篇论文中小说出场人物次数统计略有差错。本篇论文缺少对具体角色的情感分析与各个角色之间的关系分析，可能会影响读者对小说全文主旨的理解与领略。此问题解决方法可使用nltk.classify模板中的NaiveBayesClassifier类实现朴素贝叶斯分类算法，使用train()方法训练模型，将待训练文本经过预处理后设定情感分值，即将积极情感的分值设为1，将消极情感设为-1，构建一个训练模型，根据训练的模型就能够辨识出部分带有情感色彩的固定单词。

通过本篇论文研究可以看出小说《天龙八部》中主要出场人物为萧峰、段誉、虚竹、王语嫣、慕容复、段正淳、游坦之和木婉清。主要势力为丐帮、少林、星宿。故事发生地主要位于契丹、大理、西夏。《天龙八部》主要围绕大理国段誉、丐帮萧峰、少林寺虚竹等角色人物来展开，通过[宋](https://baike.baidu.com/item/%E5%AE%8B/4762756?fromModule=lemma_inlink)、[辽](https://baike.baidu.com/item/%E8%BE%BD/9050148?fromModule=lemma_inlink)、大理、[西夏](https://baike.baidu.com/item/%E8%A5%BF%E5%A4%8F?fromModule=lemma_inlink)、[吐蕃](https://baike.baidu.com/item/%E5%90%90%E8%95%83?fromModule=lemma_inlink)等王国之间的武林恩怨和民族矛盾，展示了一幅波澜壮阔的江湖生活画卷。

本篇文章基于Python并结合jieba分词，对《天龙八部》小说文本实现了柱状图、饼图、词云图等多种可视化形式。对小说文本进行可视化分析，可以帮助读者通过图表、词云图等更直观、全面、简洁地了解小说主要角色，主要故事发生地，进而了解故事情节梗概。对读者能够挑选到自己感兴趣的小说提供了极大的帮助

**参考文献**

[1]穆翠霞.基于Python和Echarts的商品评价文本的可视化设计[J].电脑知识与技术,2020,16(35):11-14.

[2]唐婷.基于Python的词云生成技术分析[J].科学技术创新,2021(23):77-78.

[3]徐博龙应用Jieba和Wordcloud库的词云设计与优化[J].福建电脑,2019,35(06):25-28.DOI:10.16707/j.cnki.fjpc.2019.06.006.