**成都理工大学本科毕业论文**

**《平凡的世界》小说文本分析**

**Text Analysis of Novel《The Mundane World》**

**作 者： 刘建军**

**学 院： 管理科学学院**

**专 业： 信息管理与信息系统**

**年 级： 2018级**

**学 号： 201807030126**

**指导教师： 彭立**

**论文成绩：**

**完成日期：**

成都理工大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经标注引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

论文作者签名：

日期： 年 月 日

学位论文使用授权说明

本人完全了解成都理工大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

论文作者签名：

指导教师签名：

日期： 年 月 日

摘要

这里是中文摘要。在对论文进行总结的基础上，用简单、明确、易懂、精辟的语言对全文内容加以概括，提取论文的主要信息。

进入21世纪以来，全球进入数字化信息时代，随之而来的产物是庞大、海量的数据文本信息。随着计算机硬件的不断发展，自然语言处理能够有效且准确的挖掘蕴涵在文本中的信息，帮助人们在信息关联、文学研究、支持决策、人机交互等领域更加有效地进行信息的沟通交流。文本分析与可视化通过关键的文字和简单的图形界面帮助人们快速抓住文本中的重要信息和主题思想，从而达到掌握文本的内容和特征。

本文通过使用自然语言处理中较为流行技术，对每个章节进行文本摘要，以简短的语句概括每个章节的内容，建立故事情节发展的脉络，分析小说人物的情感走向，使用聚类方法对章节进行可视化分析，生成相应的LDA主题模型。同时，提取小说中人物命名实体，同时计算不同人物之间的相关系数，得到人物之间关系的强弱知识，以可视化的方式呈现关系图谱等

为了对本文分析的小说原著有全面且清晰的掌握，本文通过词频统计，词云生成，文本摘要，文本聚类，降维可视化，文本聚类，神经网络等技术来对小说原著进行人物命名实体提取，相关性计算及可视化，构建角色词频图，词云图，关系发展图，感情变化图等图形，帮助读者掌握小说的全文内容及其表达的感情，更好地传达小说作者所表达的思想，增强用户的理解和从中收获的启发，提出小说阅读的新的创新性阅读方式

关键词：自然语言处理，小说文本分析，文本可视化

1. 绪论

1.1研究背景及意义

在数字化、网络化、全球化的科技迅猛发展的时代下，互联网上人与人之间的沟通交流方式越来越来丰富多彩，产生的数据总量及其复杂度都呈现出“爆炸式增长”的模式。面对海量、结构复杂的文本数据，通过以往的人工浏览的方式从大量的文本中获取有效的信息变得十分困难。近些年来，自然语言处理(Natural Language Processing)逐渐兴起，融合了计算机科学，统计学，代数学，语言学等多门学科，其目的是为了让计算机学会认知、理解、运用自然语言，从而帮助人们快速、准确的获取蕴含在文本中的信息。这里的“自然语言”是指自然世界中存在的各种人类语言，如中文、英文、德语、日语等自然界中人们沟通交流所使用的语言，并非指c、c++、java、goland等计算编程语言。由于其高效，快速，准确的特性，自然语言处理广泛应用于信息检索(Information Search)、数据挖掘(Data Mining)、机器阅读理解(Machine Reading Comprehension)、机器翻译(Machine Translation)、情感分析(Sentiment Analysis)、文本分类(Text Classification)、语音识别(Speech Recognition)等领域。随着计算机的算力不断提升以及存贮容量的增加，自然语言处理取得的效果和准确率也在不断的提升，能够成为科研人员或者课堂老师分析文本数据的有效利器。

由于长篇小说《平凡的世界》的篇幅过长，其中的人物关系复杂，故事情节跌宕起伏，对其感兴趣的读者在阅读的过程中没有足够的时间时，难以将整本书读完。即便读过的读者也需要往复阅读这本小说，才能理清其中的故事脉络和人物关系网络。而《平凡的世界》这本小说其中所蕴含的人情事故的处事原则和人生哲理也将会使读者受益。为帮助《平凡的世界》的读者能够掌握其中的知识，且鉴于自然语言处理技术的强大，本文利用自然语言处理中较为流行的技术帮助对《平凡的世界》长篇小说感兴趣的读者进行文本分析，帮助其掌握小说的主题思想，故事情节发展线以及小说人物之间的关系网络，从而有效的利用现代计算机强大的运算能力，为人类的科学生产和智能生活带来便捷和效率，这便是文本分析及其可视化的意义所在。

1.2研究现状

文本分析属于自然语言处理下的一个重要领域，而自然语言处理属于深度学习和机器学习的范畴。本小节主要将自然言语的发展历程划分为5个阶段进行介绍。

1.2.1 萌芽期（20世纪40年代-20世纪50年代末）

在第二次世界大战后，图灵(Turing)于1936年首次提出“图灵机”得概念。图灵模型的诞生带动了自动化的萌芽，这被认为是当今计算机科学发展的根基。图灵在自动化研究上的工作引领了同时期的科学家们在该领域的不断思考和研究。其中数学家沃尔特皮茨(Pitts)与医学博士麦卡洛克(McCulloch)的共同工作使得神经网络得最初模型问世；以及美国数学家、逻辑学家Kleene在有限自动机和正则表达式上的研究为后来的机器学习奠定了基础。1956年，香农(Shannon)将离散马尔可夫过程的概率模型应用在描述语言的自动机上。1956年，语言学家乔姆斯基（Chomsky）博士在香农的工作上，提出使用有限状态机来识别句法以及定义有限状态机过程产生的语言为有限状态语言，包括上下文无关语法。这些工作形成了前期机器语言处理的理论基础。

该时期，基于概率算法的语音和语言处理也得到了发展，香农提出了在信道中传输的信息的编解码方式以及噪声的存在对信道的干扰。香农将热力学中的“熵”的概念应用到统计信道中信息容量的大小，开创了将概率方法应用到英文信息测度领域的先河。

1952年，贝尔实验室创建了第一个基于统计的语音辨识系统，该系统通过说话人的语音来辨识输出0到9之间的数字。在特定的条件下，该系统能够达到97%至99%的正确率（Davis et al., 1952）。

1.2.2 快速发展期（1957年-1970年）

20世纪50年代末至60年代初，自然语言处理形成了两种研究范式，符号派研究学者认为应当将基于规则的符号表用于NLP中的语言处理，而随机派学者则主张使用概率方法进行研究。以乔姆斯基为代表的学派在形式语言理论和生成句法上进行不断研究，形成了一套解析算法理论，由最初的由上至下以及自下而上到动态编程技术。1958年至1959年，哈里斯（Harris，1962）在宾夕法尼亚大学的实验室创造了最早的完整解析系统Transformations and Discourse Analysis Project（TDAP）；以香农为代表的学术流派提出了人工智能的概念，他们注重研究推理和逻辑问题的研究。在这一时期，科研人员将模式匹配（Patten Matching）与关键词查找等启发式的研究方法应用在问题回答（Question Answeing）领域上，建立了自然语言理解系统的最初模型。

20世纪50年代末，贝叶斯方法理论开始被应用到光学字符识别问题上。布莱索（Bledsoe）和布朗宁（Browning）建造了一个基于贝叶斯理论的文本识别系统，该系统在一个项目庞大的词典中查找给定的词概率，并将它们乘起来计算特定字符序列出现的概率。同时期形成一些可以用于科学研究的语料库。

1.2.3 低速发展期（1971年-1993年）

由于语料库的缺乏，以及计算机的算力不高，当时自然语言处理的应用不能取得较好的结果。

但相关的研究仍然在继续着，进行随机范式研究的学者提出了隐式马尔科夫模型（Hidden Markov Model，HMM）；基于逻辑推理的研究学者提出了变形语法(Metamorphosis grammar )、固定子句语法（Definite clause grammar）、功能语法（functional grammar）和词汇功能语法（lexical functional grammar）等语法规则；自然语言理解中诞生了SHRDLU系统，它能够接受自然语言的文本命令并作出相应的行动。基于逻辑推理的范式与自然语言理解范式相结合，推动了语义表示的发展，譬如LUNAR问答系统的诞生。

1.2.4 复苏融合期（1994年至今）

这一时期的科学家们的研究方式逐渐从以经验为导向向数据驱动学习过渡，着得以与计算机的计算能力的大幅度提高。各种研究方式相互借鉴，汲取他人所长，概率方法论与数据建模模型的结合使得信息提取有了较好的效果。

这一时期，NLP研究人员可以在互联网上获取种类丰富，数据量庞大的预料库来训练自己的模型。统计机器学习社区的活跃度保持较高的水平：支持向量机的提出（Boser et al.，1992；Vapnik，1995）、最大熵技术以及多项式回归（Berger et al.，1996）、贝叶斯图论模型（Pearl，1998）、非监督统计方法受到广泛的关注。

1.3论文的工作内容和贡献

本文旨在使用自然语言处理和文本分析的一些有效技术和方法，来对小说原著进行全面而准确的文本分析，帮助该小说的读者了解全书的原貌，包括角色、情节、情感、故事发展脉络等等，增强小说读者的直观感受和理性认识，完善读者的阅读体验。

本文的工作内容和贡献如下：

（1）使用LDA主题模型对小说各个章节进行主题模型分析。

（2）使用神经网络技术对小说人物角色之间的关系进行建模和可视化分析，形成小说人物关系知识图谱。

（3）使用交互的方式，让用户选择感兴趣的的内容进行分析。

本文使用自然语言处理和文本分析中的一些流行技术包括不限于词频、词云、文本聚类、LDA主题模型、文本摘要、神经网络等技术手段来对小说原著进行主题模型分析和人物关系网络分析。

1.4论文的组织结构

本文对小说原著的文本分析从以下5个章节进行开展：

第一章，绪论。主要介绍自然语言处理和文本分析的研究背景和意义，国内外的研究现状，以及本论文的工作内容和重点。

第二章，相关技术研究现状。本章介绍了自然语言处理和文本分析中的一些流行技术，包括文本的预处理、命名实体的提取，常见的文本分析模型LDA和词嵌入向量，以及其他流行的神经网络技术。

第三章，小说文本的主题提取。本章使用隐式狄利克雷模型来对各个章节的主题进行建模和分析，同时使用文本摘要技术对各个章节的内容进行简要的概述生成，且对各个章节进行文本聚类。

第四章，小说人物关系图谱。本章使用分词技术对小说原著中的重要人物实体进行抽取，并根据词语的共现性来计算人物之间的相关系数，并进行可视化的分析和呈现。

第五章，通过网络查询资料或咨询相关专家的意见来验证以上技术对小说分析的准确性和合理性。评价本文使用的技术的优缺点，并阐述了未来的改进的研究方向，讨论其他研究技术的可能性。

1. 相关技术研究现状

2.1 自然语言处理

自然语言处理是指利用计算机等拥有强大计算能力的机器设备来对人类中通常沟通或书面表达的语言形成的文本文字进行建模和不断的收敛训练，最终使得结果稳定，并在新文本上进行计算以及相关的分析和操作，从而帮助一方或者双方更好更有效的沟通和表达彼此所传达的信息，常用于机器翻译，文本挖掘，阅读理解等领域。

2.2 命名实体识别

在一篇小说中，其必然有人名或地名或机构名或专有名词的存在。命名实体识别就是要找出这些词语的边界，确定实体的类型，从非结构化的文本中抽取出上述实体。这是进行文本分析的前预处理的一项基本性的关键任务，是关系抽取、事件抽取、知识图谱、信息提取、问答系统、句法分析、机器翻译等其他NLP任务的前提。

对于命名实体识别这项工作，目前有以下两种经典的方式

A 基于规则的方法。该方法使用基于文本存在的潜在性规律特征，其中的命名实体满足一定的范式规则，因此可以通过正则表达式的方法进行实体的抽取，但是面对实体命名形式复杂，变化不定的情况，单一的正则表达式不能够很好的从文本中抽取出命名实体，会带来命名实体抽取错误或者准确率低的情况。一般适用于半结构化或者命名实体简单的文本语料。

B 基于模型的方法。该方法一般为使用神经网络模型在特定类型的文本上进行训练，直到结果能够大致收敛，稳定在一定的阈值上下，就可以用于实际的命名实体抽取和词性标注。该种方法一般对于具有不同特征的语料要求不同的网络模型，并且训练的周期一般较长

2.3 word2vec模型

在计算机的世界中，只存在二进制，即0和1。为了让计算机能够读取文本，输入到计算机的内存中，需要将一系列的文字转化为机器能够识别的数字。早期的办法是使用独热向量来表示，即每个字都用一维来表示，这虽然将文本转化为了数字，但是其存在的问题也是显然的，中文的常见字有上万个，如果每个词都使用一个万维向量来表示，那么对于计算机的内存来说将是一个严重的灾难。其存在的问题的本质是忽略的字与字之间是存在相互联系的，例如称呼一个母亲，可以称做“娘亲”和“母亲”，它们所表达的含义是非常相似的，因此它们在转化为向量之后应当也保持较近的距离。而word2vec正是将不同的词语投影到隐式空间，使得原本意思相近的单词在特征空间中根据某种距离度量公式的计算后也具有相近距离，意思不同的词语在投影后有相对较远的距离，而不是像独热向量一样每个词语都是正交的。同时word2vec能够指定根据不同的场景选择特征空间的维数，更好的拟合模型所需要的特征空间容量，能够定量的计算两个词语之间的相似度。

常见的word2vec的统计模型有CBOW和skip-gram模型。前者通过给出词语的上下文来预测当前位置最大可能性的词语，而后者则是通过给出当前词语，来预测可能性最大的上下文。

2.4 LDA主题模型

LDA主题模型即隐含狄利克雷分布模型。LDA模型通过词袋模型统计文本语料中的不同词汇的出现频率，同时根据最终需要分类的主题个数，以不同的概率模型将不同的文本归结于不同的主题，学习到一定准确率的先验知识后在对实际中需要分类的文本进行主题分类，常见的先验知识有二项分布、Gamma函数、Beta分布、多项分布、Dirichlet分布、马尔科夫链等。LDA一般选择其中的一个或几个先验计算方式来在特定的文本上进行训练，通过反馈机制来告诉模型应该如何正确分类，最终达到较高的分类效果。

LDA的训练步骤一般如下：

1 对预料库中的每篇文档中的每个词汇w，随机的赋予一个topic编号z；

2 重新烧苗语料库，对每个词w，使用Gibbs Sampling公式对其进行采样，求出它的topic，在语料库中更新；

3 重复步骤2，直到Gibbs Sampling收敛；

4 统计语料库的topic-word共现频率矩阵，该矩阵就是LDA模型；

通常在LDA训练过程中，是使用Gibbs Sampling收敛后的n个迭代的结果平均来进行参数估计，这样会使得模型的质量更高。

2.5 其他网络模型

出了上述介绍的技术外， 下面是21世纪以来应用在NLP领域具有里程碑的重点技术

A、神经语言模型（Neural language models）。该模型旨在根据给定的文本中的词语或前n个词语预测下个词语的可能情况，将概率最大的词语作为下一个词语输出（Bengio et al.，2001）

B、多任务学习（Multi-task learning）。该模型是为了解决单一预料库不足导致训练样本不充分的问题，它将多个任务一起训练，共享任务模型之间的参数，能够较好使每个任务模型得到较好的训练（Collobert，2008）

C、用于自然语言处理的神经网络（Neural networks for NLP）。NLP中流行的神经网络包括卷积神经网络（Convolutional neural networks）其通过输入-编码-激活-卷积-池化-归一化等操作输出期望值、循环神经网络（Recurrent neural networks）该网络能够记住一定长度的句子信息，避免局部局限的问题。

D、序列到序列模型（Sequence-to-sequence models）。该模型用于机器翻译，对话代理等领域，取得较好的实际效果（Sutskever，2014）

E、注意力机制（Attention）。该机制能够学习较长范围的语句信息，得到比较综合全面的文本信息，适合于阅读理解等方向（Bahdanau，2015）

F、预训练语言模型（Pretrain language models）。由于现在的网络模型的庞大与学习的复杂性，且认为一般问题遵循基本的表示规律，科研人员提出了预训练方法，用于学习输入的特征表示，将其作为下游任务的输入，并且根据特殊问题进行微调（fine-tuning）能够缩短训练时间和降低训练复杂度，且取得较好的实验结果（Conneau et al.，2017；McCann et al.，2017；Subramanian et al.，2018）

理论模型的蓬勃发展，促进了相；关软件的诞生，帮助人们将注意力放在行业务领域问题的分析上，减轻了人们的学习和应用负担。

1. 小说文本的主题提取

3.1 主题模型介绍

LDA（Latent Dirichlet Allocation）是一种文档生成模型。它认为一篇文章是有多个主题的，而每个主题又对应着不同的词。一篇文章的构造过程，首先是以一定的概率选择某个主题，然后在这个主题下以一定的概率并选出某一个词，这样就生成了这篇文章的第一个词。不断重复这个过程，就生成了整篇文章，其中假定词与词之间是没有顺序的。

LDA的使用是上述文档生成的逆过程，它将根据一篇得到的文章，取寻找出这篇文章的主题，以及这些主题对应的词。一般来说，每篇文章都会有一个中心思想，那么一些特定的词语会很频繁的出现。这样我们只需要统计每个词出现的频数，计算出每个词对应主题的频数，就能得到该偏文章的主题分布和比率。

LDA主题模型可以通过以下公式进行描述

其中w表示生成的词语，topic为文档可能的主题，d表示文档，t为某个可能的主题。

3.2 文本预处理

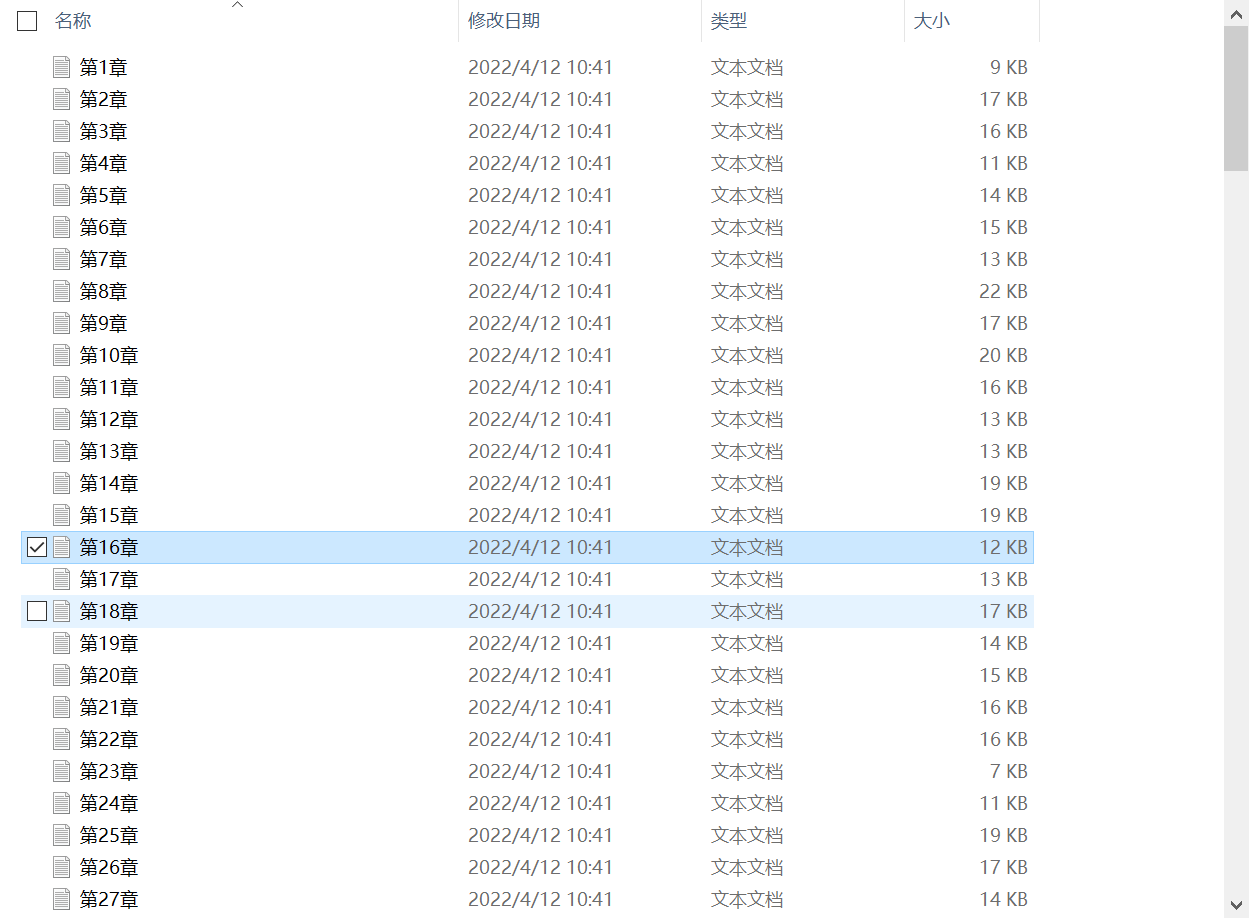
在正式进行主题模型抽取之前，由于小说文本的非结构特点，我们需要对小说原著进行预处理，使预料规范化。例如对文本进行分词，停用词过滤和词性标注等。

3.3.1 数据获取

从网络或者图书馆数据库中上获取《平凡得世界》小说得全文文本，以.txt格式保存到本地磁盘上。

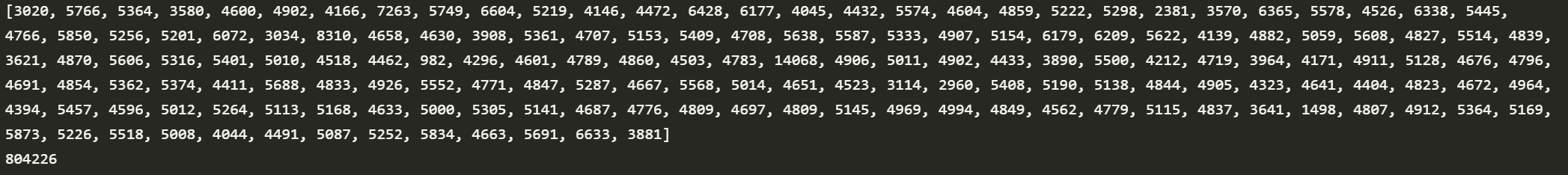
3.3.2 小说预处理

第一步，通过使用正则表达式提取出形如“第×章”的段落，并以此为依据将整篇小说划分为一个个的章节，生成后的文件以“第×章.txt”的格式保存在同一个目录下。如下图所示：

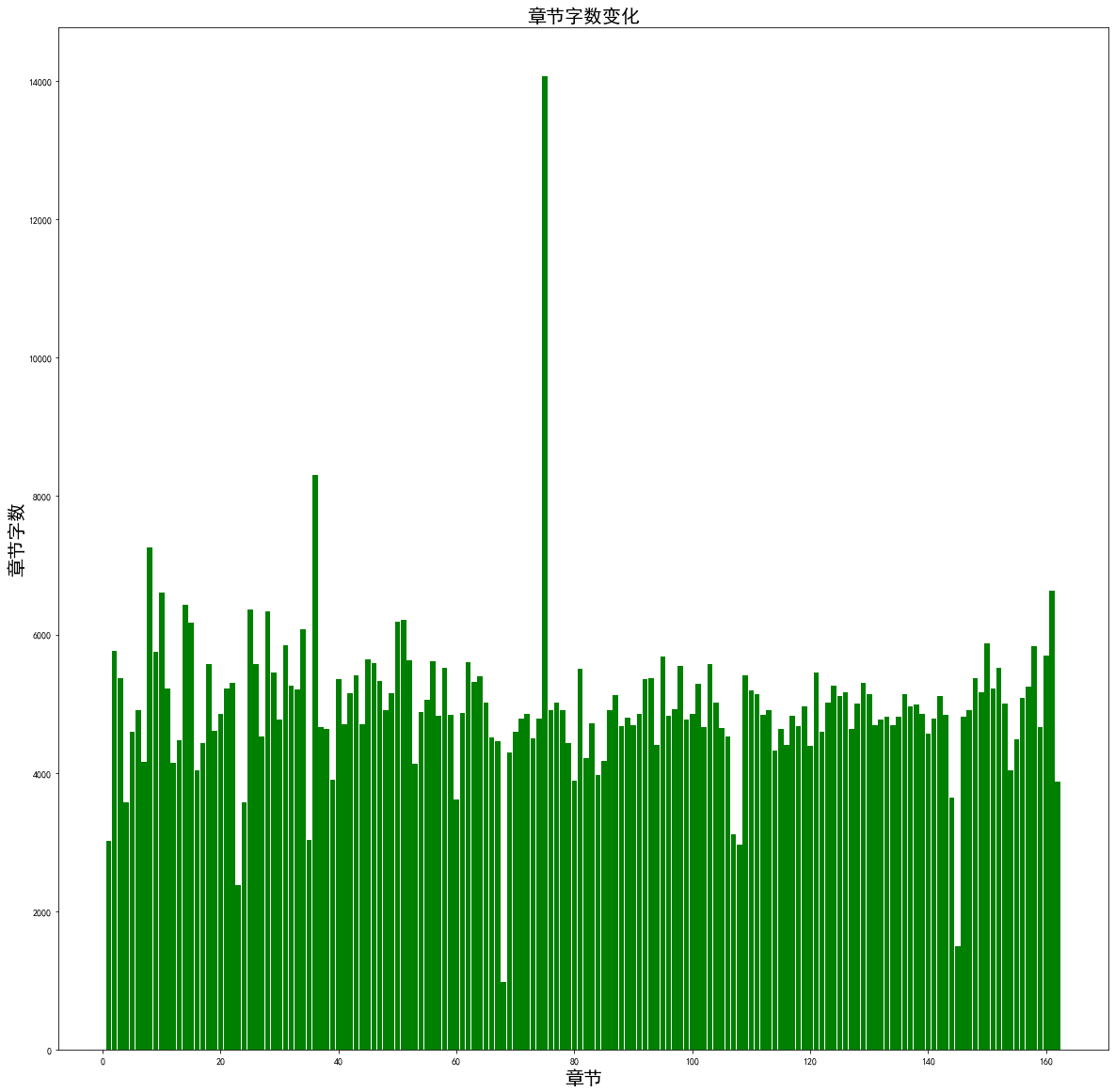


得到《平凡的世界》全文共162章，包括上中下三部。

第二步，运用python脚本的便捷和易用性，统计《平凡的世界》每个章节的字数，从第1章到第162章字数依次如下图示：



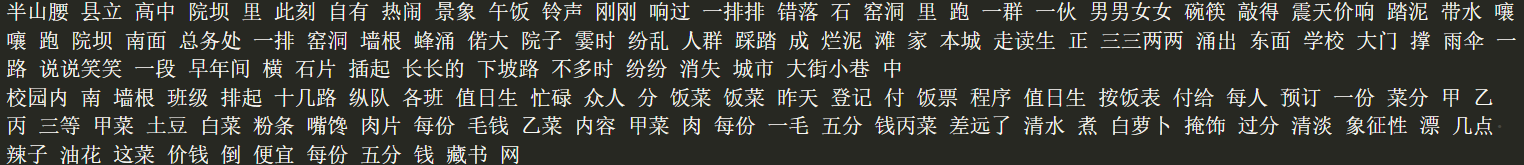
绘制成如下图



可以看到第74章的篇幅描述较多，达到了14068个字，其余大部分章节的字数均匀分布在3500到7000字的范围。

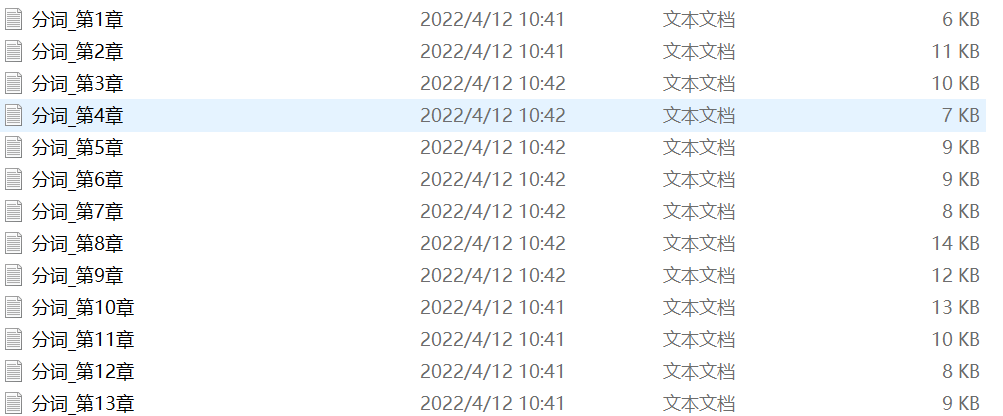
第三步：对小说进行编码转化，分词和取停用词，和词性标注。将每个章节的文本的编码格式转化为UTF-8编码格式并保存。本文使用的分词工具式jieba，由于英文的词和中文的词存在一定程度的差异，且为了保证一些专有名词能够顺利的被保留下来，我们需要自定义小说中的专有词的字典，如人名字典，地名字典等，当工具在分词时遇到了这些字典中定义的词时就不会将该名词分开，而是作为一个整体进行保留，同时过滤掉一些无用的无实际意义的词，如“的”，“吧”，“呀”，“哦”，“哇”等语气助词和停用词以及标点符号，因为这些词会降低后续文本分析的效率，去除这些词后，可以净化后续文本分析的结果，大大提高文本分析 的准确性和效率，最后进行词性的标注。

分词后的部分结果如下：

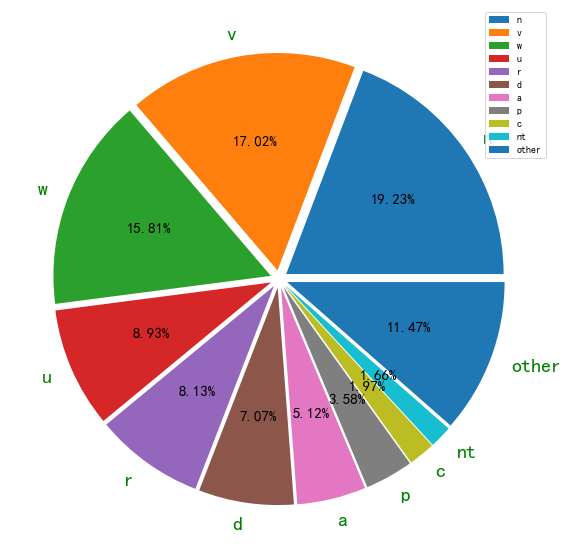


详细的章节分词结果见[kkl-oop/thesis: works on thesis (github.com)](https://github.com/kkl-oop/thesis)

各章节分词后也同样放在同一个目录下



对各个章节进行分词之后，对每个分词结果使用jiagu库进行词性，统计每个词性的出现次数，能够得到如下图：



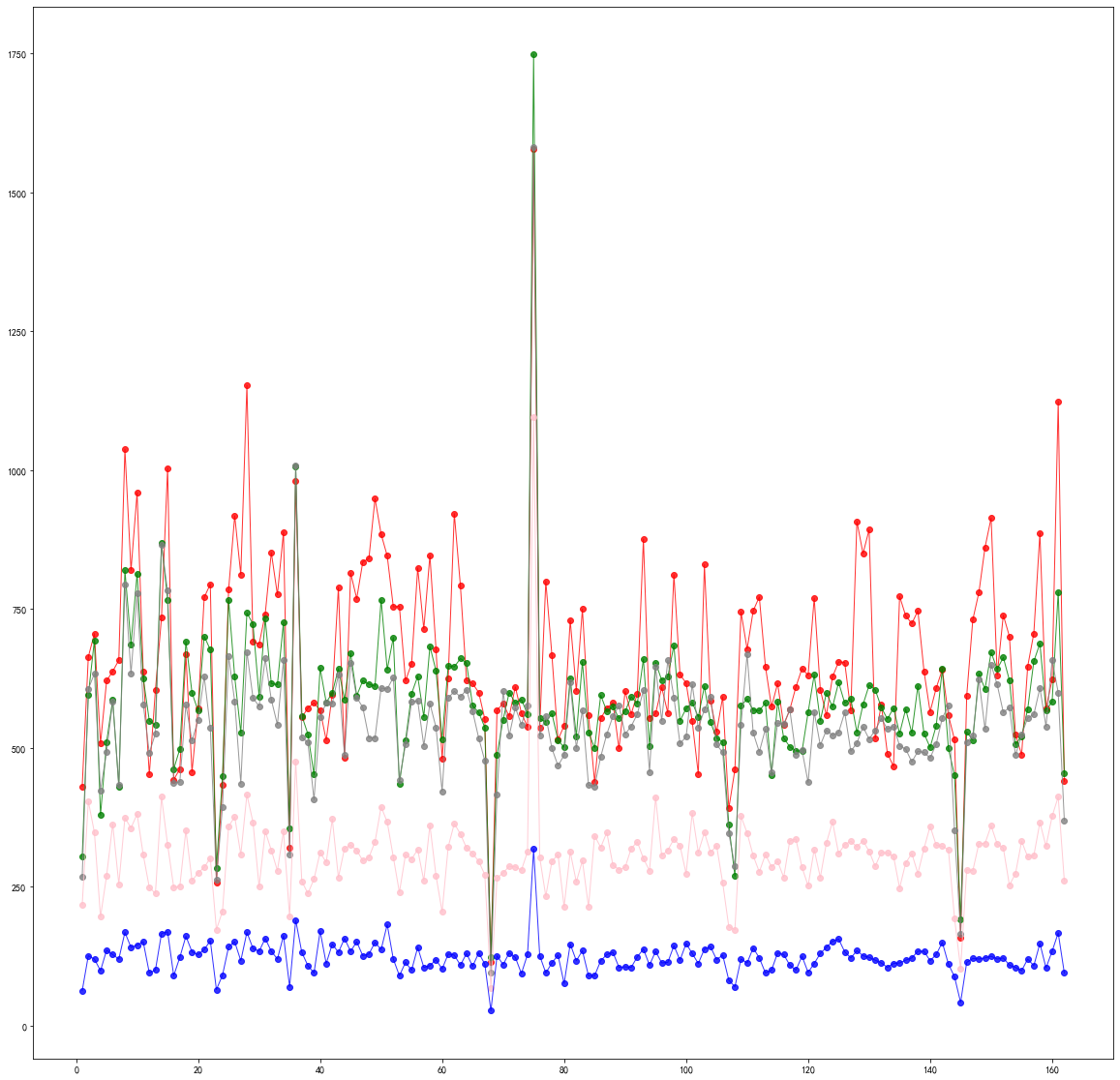
改图为小说全文的词性占比图，其中n，v、w、u、r、d、a、p、c、nt、other依次表示为普通名词、动词、标点符号、助词、代词、副词、形容词、介词、连词、时间名词和其他词性的词，前10种词性为全文种出现最多的词性。

接着通过使用jiagu库的词性标注功能获取了小说种出现频率最高的前20个词汇，如下图：



可以看到词频排名前20的词汇，名词类型的词语占了18个，说明本小说种涉及的专有名词相对其他类型词语的出现频次较多，其中“少平“词语的出现频率最高。

除此之外，本论文的工作还统计了小说的几种代表性的词语的词频随章节变化的折线图，如下图所示：

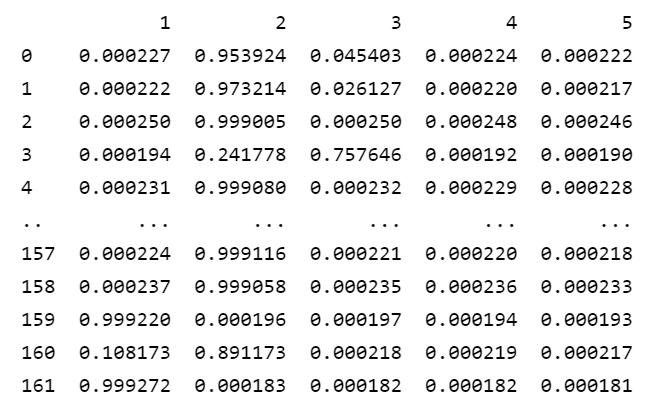


横坐标表示章节数，纵坐标表示某种类型的词语的词频数量，图中红色、绿色、灰色、粉红色、蓝色的折线依次表示名词、动词、标点符号、助词、介词。从图中可以看出，这几种词性的词频都分布在一定的数量区间，名词、动词的重叠度较高。这几种词性的词频随章节波动的趋势大致相同。

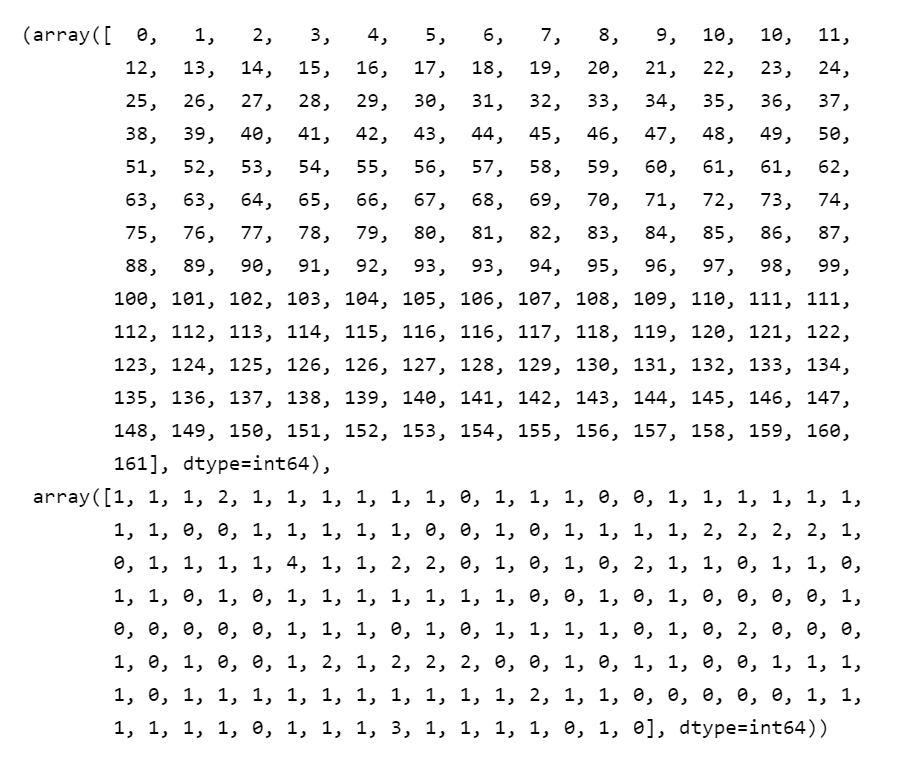
3.3 主题建模

本论文使用LDA主题模型来对小说进行主题分析，一般该主题模型为三层贝叶斯模型，通过计算主题的概率分布来给出每篇文本的主题，是一种非监督的机器学习技术。LDA模型使用了词袋模型，将每篇文章看成是一个词频向量，每篇文章都会存在一个概率分布。LDA具体的具体操作是，给定一个主题，反复的按照在该主题概率下进行词语的不断采样，直到最终的文本文章生成。

在对每章进行分词后，本文使用gensim库来全文构建词频矩阵，然后LDA模型对文本进行训练，本文对小说定义了5种主题，依次为主题一、主题二，主题三、主题四、主题五，生成的结果图下图：

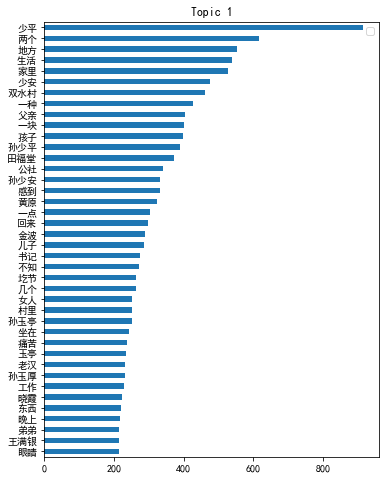


其中左侧表示每个章节的序号，顶部每列表示5种主题，矩阵种的每个数字表示该章节属于该主题的几率，除了个别章节属于其他主题的权重较大外，可以看到大部分章节为主题二的概率比较大，说明几乎所有章节的主题都是相同的。

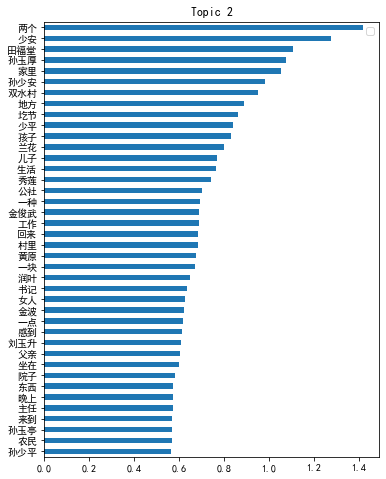


上图是使用LDA主题模型对各个章节进行归类得到的结果图，从上图的输出结果来看，大部分章节可以归类为主题二，少量章节归结为主题4、主题0和主题三，据此判断，全文小说的故事内容是比较紧凑的，主题思想集中，故事情节初步判断是一致的，没有出现大幅度离散的情况。这与上一张图得出的结论是一致的。

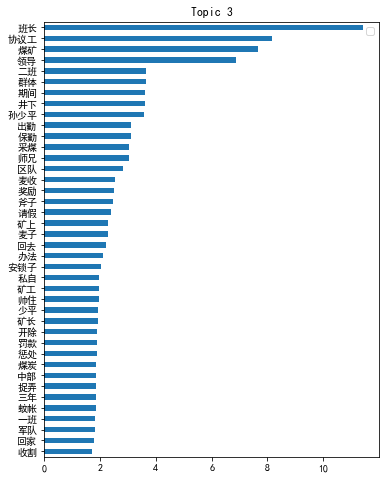
通过LDA模型，笔者还分析出了每个主题中关键词语的组成成分，如下图：



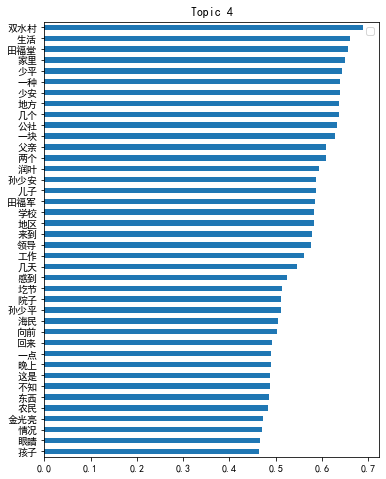
由上图可见，主题一中，能够代表主题一的前5个关键词语依次为：“少平”、“两个”、“地方”、“生活”，“家里”。其中“少平”为全文小说的主人公之一。”



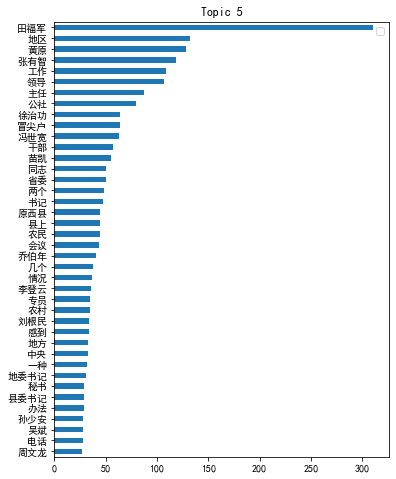
由上图可见，主题二中，能够代表主题二的前5个关键词语依次为：“两个”、“少安”、“田福堂”、“孙玉厚”、“家里”。其中“少安”、“田福堂”、“孙玉厚”为小说的人物角色。”



由上图可见，主题三中，能够代表主题三的前5个关键词语依次为：“班长”、“协议工”、“煤矿”、“领导”、“二班”。



由上图可见，主题三中，能够代表主题三的前5个关键词语依次为：“双水村”、“生活”、“田福堂”、“家里”、“少平”。



由上图可见，主题三中，能够代表主题三的前5个关键词语依次为：“田福军”、“”、“煤矿”、“领导”、“二班”。

通过以上的LDA主题模型分析，我们可以获取小说中的重点主题，以及该主题下所代表的关键性词语，帮助读者掌握小说中提及的重点词汇，增强读者的阅读体验，给小说读者后续精读一个深刻鲜明的印象。

1. 小说人物关系图谱

4.1 知识图谱介绍

4.2 文本预处理

4.3 建立人物网络关系

1. 实验结果分析

5.1 实验结果准确性分析

5.2 有缺点分析

5.3 改进的地方和未来的工作

1. 附录 参考文献
2. 致谢