摘 要

21世纪信息是行业竞争的核心，数据是行业发展的重要基础。当前，数据分析和数据挖掘成为当下商业活动的重要组成。大数据技术、云计算及物联网等信息通信技术的出现，难以满足海量数据的价值挖掘和内在分析。面对复杂、海量、低密度及快速生成的数据，必须深度分析才能够获取所需信息，以构架数据分析平台及系统快速完成数据

文学家主观虚构的文学世界,来源于现实,反映了现实世界,因此每个时期的作品或多或少都有着那个时期特有的特点。在文学作品中,人物的优缺点会被放大,这是作者为了将对现实世界的想法赋予到作品中的人物身上，通过对文学作品进行分析不仅能够了解文学家主观思想还能了解当时的时代特色。

本文对路遥、陈忠实以及贾平凹等作家的作品进行研究分析，具体来说，本文的主要研究内容包括：

1. **分词：**利用开源框架Jieba提供的接口对任意作品进行分词。
2. **分词和标注：**利用Jieba相关的接口对分词后的词语进行词性标注。
3. **构建用户字典，进行精确分词，并与缺少字典的分词结果进行对比：**利用斯坦福大学提供的NLP框架，对任意文学作品构建字典，并依据构建后的字典进行分词，最后与缺少用户字典的额分词结果进行对比。研究结果表明存在用户字典的分词结果根据准确。
4. **利用已有的工具对分词结果进行量化：**利用Quita，Mattr，Mawatatarad等工具对分词后的结果进行图形化展示。

**关键字：Python、文学作品、自然语言处理、分词、词性标注、字典**

Abstract

In the 21st century, information is the core of industry competition and data is the important foundation of industry development. Currently, data analysis and data mining have become an important part of current business activities. The emergence of big data technology, cloud computing, Internet of things and other information and communication technologies cannot satisfy the value mining and internal analysis of massive data. In the face of complex, massive, low-density and rapidly generated data, in-depth analysis is necessary to obtain the required information, so as to construct a data analysis platform and system to quickly complete the data

The literary world created by the writer is derived from the real world and reflects the real world.Therefore, the works of each period have more or less unique characteristics of that period.

In literary works, the advantages and disadvantages of the characters will be magnified. This is toendow the author with the idea of the real world to the characters in the works. Through the analysis of the literary works, the author can not only understand the subjective thoughts of the writers, but also understand the characteristics of The Times at that time.

This paper studies and analyzes the works of lu yao, Chen zhongshi and jia pingwa.

Specifically, the main research contents of this paper include:

1. segmentation: word segmentation of any work using the interface provided by the open source

framework Jieba.

1. participle and annotation: the word after the participle was annotated using the interface

associated with Jieba.

1. construct user dictionaries, conduct precise word segmentation, and compare with word

segmentation results in the absence of dictionaries: use the NLP framework provided by Stanford university to construct dictionaries for any literary works, and conduct word segmentation according to the constructed dictionaries, and finally compare with word segmentation results in the absence of user dictionaries. The results show that the segmentation results of user dictionaries are accurate.

1. quantization of word segmentation results with existing tools: graphical display of word

segmentation results with Quita, Mattr, Mawatatarad and other tools.

1. 绪论

语言是最重要的交流工具，在人类的历史上扮演者重要的角色。自从语言发明以来，语言始终是信息的载体，绝大多数文学作品都是以语言或者语言文字的形式流传下来，因此，语言或者以语言为载体的作品蕴含着丰富的信息。

用自然语言[1]与计算机进行通信是自计算机发明以来，人类与计算机交互的重要途径。但是与计算机交互的语言具有难以理解且不易学习的特点。用自然语言与计算机交互是一直以来科学家努力的目标。如果人类可以用自然语言与计算机交互不仅可以为人类节省大量的学习各种高级语言（Java、Python、Shell等），还可以让人类本省对自己的语言有更深刻的认识。

另一方面，计算机与人类通过自然语言的方式进行通信，意味着计算机需要理解自然语言背后的蕴含的意义，同时将处理后的任务以自然语言的形式表达出来。上述两个方面涉及到自然语言理解与自然语言处理。但是自然语言在人类数千年的演化过程中，往往相同的语言在不同的情景下具有不同的含义，因此给计算机理解自然语言带来了严峻的挑战。由于计算机难以理解自然语言的真正意图，导致自然语言生成更加困难。

不论计算机理解人类的自然语言或者计算机生成含义恰当的自然语言，都是十分困难的任务。基于当前学术领域与工业界的情况，恰当的、精准的自然语言处理系统，仍处于初级阶段。在一些特定的领域中，自然语言处理系统的表现尚可：多种语言的翻译系统、检索系统、实施语音翻译、信息检索系统等。自然语言处理的最大阻碍在于自然语言的随意性、多意性。

与基于英文的自然语言处理过程中，基于中文的额自然语言处理更加具有挑战性，也一直以来是自然语言出来领域的研究热点，具体来说：（1）中文中蕴含的信息量大。汉子作为一种历史最悠久的语言，在历史的演化过程中，其意义与含义不断得到扩展。以文言文为例，往往简单几个汉子，遍拥有大量的信息。而现在的白话文，语言灵动，相同的意义可以由各种方式表达出来。另一方面，基于我国研学家的研究，汉字的容量极限是12366个汉字。如此多的汉子经过各种组合其含义具有爆炸性的增大。（2）计算机难以识别。自计算机发明以来，基于键盘的出入系统，没有考虑到中文的输入，因此这种机制为计算机接受中文带来了挑战。另一方面，中文作为一种“方块”语言，很多字体具有相似性，却含义复杂。更进一步，基于语音的输入，要求中文的读音必须准确，但是中文的汉子具有多音的特点，即一个汉子具有多个读音，另一方面，多个汉字具有相同的读音。这就要求计算机必须基于上下文，在理解句子的含义的基础上进行推理演化才能输入正确地句子。上述特点增加了汉字输入计算机的难度。（3）具有多意性。中华民族上下五千年的历史演变过程中，为汉语附上了丰富多彩的含义，相同的语句在不听的情境下表示的意义是不同的，这为计算机理解中文内在的含义增加了难度。

近年来，随着自然语言处理技术与计算机硬件与软件的不断发展更新，人工智能领域得到了长足的进步[2]。在羽然语言处理方面，由于英文与计算机的语言更加相近，因而，基于英文的自然语言处理发展要快于基于中文的自然语言处理。中文的形式与含义的差异性太大、多异性更强，导致了基于中文的额自然语言处理一直以来是一个亟待解决的重要问题，然而随着各种高效自然语言处理方法的提出，为基于中文的自然余力插上了腾飞的翅膀。

* 1. 研究背景与意义

自改革开放以来，我国经济迅猛发展，企业已经成为推动我国经济发展的重要力量。然而随着时代的进步，自然语言处理得到了研究领域的广泛重视。随着一系列技术和方法的提出，自然语言处理的精确度得到了人们的认可，并在工业界得到了认可。

目前我们正在进入一个智能化时代（互联网+及工业4.0时代），大数据、云计算和人工智能已经成为这个时代进步的三驾马车(（图1-1所示）)，它们分别为智能化时代提供数据、算力和算法层面的支持，从而成为各行各业技术革新和社会发展的重要引擎。

在智能化时代的背景下，云计算为文学作品相关的数据提供了大量的存储空间，使得海量、多维度的文学数据存储成为可能。研究者可以从大量存储的数据中识别、提取感兴趣的信息，为文学作品的分析、情感分析等领域带来巨大的益处，具体来说：（1）目前我们已经有非常好的语音识别系统了，现在基本上达到了人类的水平，在理想环境里可以达到95%以上的正确率。同样我们也有比较正确的机器翻译系统，正确率换算过来也可以有70%到80%，虽然离人的水平还有一定的差距，但是已经是可用的状态。（2）自然语言处理在教育领域中的应用，帮助老师修改作业、批阅试卷等。

图 1-1 拉动智能时代的三驾马车



Figure 1-1 Pulling the Troika of the Intelligent Age

基于上述分析，自然语言处理是智能化时代的主要特色之一，但是达到理想的结果不是一件容易的事情，主要存在两个方面的问题：一方面，迄今为止的语法都限于分析一个孤立的句子，上下文关系和谈话环境对本句的约束和影响还缺乏系统的研究，因此分析歧义、词语省略、代词所指、同一句话在不同场合或由不同的人说出来所具有的不同含义等问题，尚无明确规律可循，需要加强语用学的研究才能逐步解决。另一方面，人理解一个句子不是单凭语法，还运用了大量的有关知识，包括生活知识和专门知识，这些知识无法全部贮存在计算机里。因此一个书面理解系统只能建立在有限的词汇、句型和特定的主题范围内；计算机的贮存量和运转速度大大提高之后，才有可能适当扩大范围。

* 1. 研究内容与成果

21世纪信息是行业竞争的核心，数据是行业发展的重要基础。当前，数据分析和数据挖掘成为当下商业活动的重要组成。大数据技术、云计算及物联网等信息通信技术的出现，难以满足海量数据的价值挖掘和内在分析。面对复杂、海量、低密度及快速生成的数据，必须深度分析才能够获取所需信息，以构架数据分析平台及系统快速完成数据

文学家主观虚构的文学世界,来源于现实,反映了现实世界,因此每个时期的作品或多或少都有着那个时期特有的特点。在文学作品中,人物的优缺点会被放大,这是作者为了将对现实世界的想法赋予到作品中的人物身上，通过对文学作品进行分析不仅能够了解文学家主观思想还能了解当时的时代特色。

本文对路遥、陈忠实以及贾平凹等作家的作品进行研究分析，具体来说，本文的主要研究内容包括：

1. **分词：**利用开源框架Jieba提供的接口对任意作品进行分词。
2. **分词和标注：**利用Jieba相关的接口对分词后的词语进行词性标注。
3. **构建用户字典，进行精确分词，并与缺少字典的分词结果进行对比：**利用斯坦福大学提供的NLP框架，对任意文学作品构建字典，并依据构建后的字典进行分词，最后与缺少用户字典的额分词结果进行对比。研究结果表明存在用户字典的分词结果根据准确。
4. **利用已有的工具对分词结果进行量化：**利用Quita，Mattr，Mawatatarad等工具对分词后的结果进行图形化展示。
   1. 论文组织结构

本文的组织结构如下：

1. 绪论，主要包括课题的研究背景与意义、研究内容与成果以及论文的组织结构。
2. 背景介绍，主要包括神经网络和词袋模型的基础概念与理论，以及目标识别、卷积神经网络的国内外研究现状。
3. 深入讨论本文提出的基于文本挖掘的神经网络企业信用评级方法的框架。
4. 在开源的数据集上验证所提框架的有效性与准确性，主要包括研究问题阐述、实验对象说明、实验流程。
5. 通过图、表的方式回答所提的问题，并深入分析实验结果。
6. 研究工作的总结以及未来的展望。
7. 背景介绍

本章介绍自然语言处理的发展、Jieba框架、以及斯坦福大学的开源NLP框架。

* 1. 自然语言处理的国内外研究现状

20世纪中期，自然语言开始进入历史的潮流。虽然发展时间不长，但是人类

在自然语言处理方面取得的成绩有目共睹，特别是近20年来，随着人工智能算法、算力的不断提升，自然语言处理也得到了长足的进步。自然语言处理是一个交叉学科，往往涉及到数学、统计学、计算机等领域。当今社会处于信息的时代，各种数据中蕴含着大量的信息，云计算为这些数据提供了大量的存贮空间，并且自然语言处理算法的不断改进为自然语言处理的发展提供了便利。目前，自然语言处理已经成为人们在日常生活中获取信息的重要手段。接下来，在自然语言处理方面的主要技术进行简要分析。

Luhn等人首次将统计学知识运用到自然语言处理中，并通过计算词语频率的方式来推测关键字[3]。Hulth等人将词语信息放入语法中进行分析来提取关键字[4]。Turney等人将机器学习中的算法引入到自然语言处理中，并利用当时先进的C4.5分类算法提取文本中的关键字[5]。2004 年，Mihalcea 和 Tarau 利用特征词在共现窗格中出现的关系构造词与词的图模型，并结合 Google 公词提出的 PageRank 算法，提出 TextRank 关键词提取算法[6]。

2012 年，Google 率先提出知识图谱 (Knowledge Graph)的概念，将知识图谱应用到搜索引擎中。运用自然语言处理提取实体和实体间的关系，并以此分别作为图谱的节点和边，构建知识图谱。知识图谱的关键技术主要有知识抽取、知识表示、知识融合以及知识推理技术等。文献 [7] 运用 KNN 算法与条件随机场模型，对 Twitter 文本数据中实体进行抽取。陈立玮等人 [8] 提出了一种协同训练方法，通过向传统模型中引入 N-Gram 特征进行协同训练，改善了弱监督关系抽取模型的效果。Li等人根据给定的少量种子链接，利用概念标

注方法，发现新的链接，最终实现了知识的扩充 [9]。

* 1. Jieba

Jieba[10-14]是一个广泛使用的开源自然语言处理框架，支持三种分词模式：

（1）精确模式，试图将句子最精确地切开，适合文本分析；（2）全模式，把句子中所有的可以成词的词语都扫描出来,速度非常快，但是不能解决歧义；（3）搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。Jieba分词包在这里运用到了数据结构里的trie(前缀树或字典树)对词语进行高效的分类如图2.1，便于查找。然而trie的从上至下搜索、每一次只判定一个字母、如果某个特定的节点(node)的下一个节点(child node)不再符合搜索要求，那么搜索就会停止，这样效率就会大大的提高。此外，trie也结合了计算机领域另一个知识，名为有向无环图（DAG），trie与有向无环图的结合很高效的解决了第一段所提到的双重理解词语组合的问题。

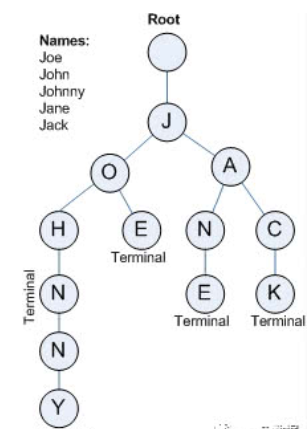


图2.1 Trie树

Jieba利用TF-IDF(Term Frequency-Inverse Document Frequency, 词频-逆文件频率)算法进行关键词的提取，该算法是一种统计方法，用以评估一个词语对于一个文件集或一个语料库中的一份文件的重要程度，其原理可概括为：一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章。

* 1. 斯坦福NLP

Stanford CoreNLP[15-18]是用处理自然语言的工具集合。它可以给出词语的基本形式：词性（它们是公司名、人名等，规范化得日期，时间，和数字），根据短语和语法依赖来标记句子的结构，发现实体之间的关系、情感以及人们所说的话等。

该框架具有：（1）它是一个众多语法分析工具集成的工具包；（2）对任意的文本来说它具有快和鲁棒性强的特点，并且还广泛的用于生产中；（3）它还具有时用性，及时性；（4）支持数量众多的（主要）自然语言；（5）支持编程语言的接口丰富；（6）能够作为简单的web服务运行。

Stanford CoreNLP的目的是为了让一系列语言分析工具应用到一段文本上变得更容易。用两行代码就可以让一段原生的文本在一个工具管道（一系列处理文本的操作）中流过。CoreNLP 具有很强的灵活性和扩展性。通过一个选项的设置你就可以选择那些语言处理的工具可以应用到这段文本。Stanford CoreNLP包含许多的斯坦福的nlp工具，包括：词性标注器、命名实体的识别器、解析器（句子与语法结构）、指代消解器（就是在篇章中确定代词指向哪个名词短语的问题）、情感分析器、引导模式学习器、开放信息提取器。而且，一个解释管道还可以包含其他自定义或者第三方的解释器。本文利用该工具为每一个文学作品创建对应的字典。

3 基于Python的文学作品量化分析

3.1 动机

语言是最重要的交流工具，在人类的历史上扮演者重要的角色。自从语言发明以来，语言始终是信息的载体，绝大多数文学作品都是以语言或者语言文字的形式流传下来，因此，语言或者以语言为载体的作品蕴含着丰富的信息。

用自然语言[1]与计算机进行通信是自计算机发明以来，人类与计算机交互的重要途径。但是与计算机交互的语言具有难以理解且不易学习的特点。用自然语言与计算机交互是一直以来科学家努力的目标。如果人类可以用自然语言与计算机交互不仅可以为人类节省大量的学习各种高级语言（Java、Python、Shell等），还可以让人类本省对自己的语言有更深刻的认识。

另一方面，计算机与人类通过自然语言的方式进行通信，意味着计算机需要理解自然语言背后的蕴含的意义，同时将处理后的任务以自然语言的形式表达出来。上述两个方面涉及到自然语言理解与自然语言处理。但是自然语言在人类数千年的演化过程中，往往相同的语言在不同的情景下具有不同的含义，因此给计算机理解自然语言带来了严峻的挑战。由于计算机难以理解自然语言的真正意图，导致自然语言生成更加困难。

不论计算机理解人类的自然语言或者计算机生成含义恰当的自然语言，都是十分困难的任务。基于当前学术领域与工业界的情况，恰当的、精准的自然语言处理系统，仍处于初级阶段。在一些特定的领域中，自然语言处理系统的表现尚可：多种语言的翻译系统、检索系统、实施语音翻译、信息检索系统等。自然语言处理的最大阻碍在于自然语言的随意性、多意性。

与基于英文的自然语言处理过程中，基于中文的额自然语言处理更加具有挑战性，也一直以来是自然语言出来领域的研究热点，具体来说：（1）中文中蕴含的信息量大。汉子作为一种历史最悠久的语言，在历史的演化过程中，其意义与含义不断得到扩展。以文言文为例，往往简单几个汉子，遍拥有大量的信息。而现在的白话文，语言灵动，相同的意义可以由各种方式表达出来。另一方面，基于我国研学家的研究，汉字的容量极限是12366个汉字。如此多的汉子经过各种组合其含义具有爆炸性的增大。（2）计算机难以识别。自计算机发明以来，基于键盘的出入系统，没有考虑到中文的输入，因此这种机制为计算机接受中文带来了挑战。另一方面，中文作为一种“方块”语言，很多字体具有相似性，却含义复杂。更进一步，基于语音的输入，要求中文的读音必须准确，但是中文的汉子具有多音的特点，即一个汉子具有多个读音，另一方面，多个汉字具有相同的读音。这就要求计算机必须基于上下文，在理解句子的含义的基础上进行推理演化才能输入正确地句子。上述特点增加了汉字输入计算机的难度。（3）具有多意性。中华民族上下五千年的历史演变过程中，为汉语附上了丰富多彩的含义，相同的语句在不听的情境下表示的意义是不同的，这为计算机理解中文内在的含义增加了难度。

近年来，随着自然语言处理技术与计算机硬件与软件的不断发展更新，人工智能领域得到了长足的进步。在羽然语言处理方面，由于英文与计算机的语言更加相近，因而，基于英文的自然语言处理发展要快于基于中文的自然语言处理。中文的形式与含义的差异性太大、多异性更强，导致了基于中文的额自然语言处理一直以来是一个亟待解决的重要问题，然而随着各种高效自然语言处理方法的提出，为基于中文的自然余力插上了腾飞的翅膀。

本文的利用python对文学作品进行分析，主要步骤介绍如下。

3.2 实验对象

本文利用路遥、陈忠实、贾平凹作品进行分析，首先从网站下载作者的txt格式的作品集，并利用UtraEdit软件对下载的txt文件进行整理，便于下一步的分析。

3.3 实验环境

本文使用 python 语言编写测试脚本， 运行在 Windos10 64 位操作系统上。该系统有 4 个 CPU 和 16GB 内存。

3.4 数据整理

中文自然语言处理的第一步是获取语料，即语言材料。语料是语言学研究的内容。语料是构成语料库的基本单元。所以，人们简单地用文本作为替代，并把文本中的上下文关系作为现实世界中语言的上下文关系的替代品。我们把一个文本集合称为语料库（Corpus），当有几个这样的文本集合的时候，我们称之为语料库集合(Corpora)。然后对语料进行预处理，在一个完整的中文自然语言处理工程应用中，语料预处理大概会占到整个50%-70%的工作量，所以开发人员大部分时间就在进行语料预处理。下面通过数据洗清、分词、词性标注、去停用词四个大的方面来完成语料的预处理工作。

清理数据应该是数据科学或者自然语言处理工作流程的第一步。如果数据没有清理干净，将很难在探索中的看到实际重要的部分。在数据科学和自然语言处理的环境中，数据清理意味着过滤和修改数据，使数据更容易探索，理解和建模。过滤掉不想要或不需要的部分，这样就不需要查看或处理它们。通过修改不需要的格式，不仅可以正确使用它们，而且可以提高算法的准确度。

本文的处理对象是txt格式的文学作品，为了使得后期处理更加方便，我们去掉作品名、空白行、作者信息、章节信息等不需要的信息。为了高效地完成这一工作，本文利用UtraEdit工具，该工具的界面如图3.1。

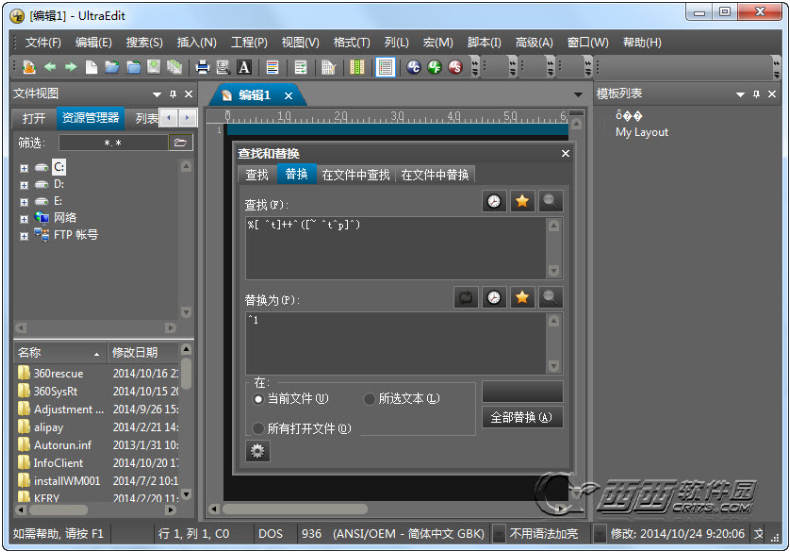


图3.1 UtraEdit工具界面

UltraEdit 是一套功能强大的文本编辑器，可以编辑文本、十六进制、ASCII 码，完全可以取代记事本（如果电脑配置足够强大），内建英文单字检查、C++ 及 VB 指令突显，可同时编辑多个文件，而且即使开启很大的文件速度也不会慢。

正则表达式，又称规则表达式，是计算机科学的一个概念。正则表达式是一种字符串的匹配模式，描述的是某一类字符串的共同特征，通常被用来检索、替换那些符合某个模式(规则)的文本。所谓模式，就是模板样式或模具样式。正如符合某种样式的模板或模具，可以用来生产符合这种样式的同一类产品一样；反过来，也可以用某种样式的模板或模具，来检验或框定哪些产品才是符合这种样式的同一类产品。正则表达式正是类似于这样的模板或模具，用来检验或框定哪些字符串是符合正则表达式所描述的字符串共同特征的同一类字符串；而这个检验或框定的过程，就称之为匹配。我们平时所使用的自然语言中，可以用“漂亮”、“坚固”、“挺拔”等高度抽象性词语来描述事物的共同特征一样，一个正则表达式正是某一类字符串的高度抽象，用来描述这类字符串的共同特征。也就是说，一个正则表达式代表了某类字符串的一个集合，而正则表达式相当于对该字符串集合的特征性质描述。正则表达式还可看作是对字符串操作的一种逻辑公式，其构造方法和创建数学表达式的方法差不多，也就是用普通字符(如字母a到z、数字0到9等)和事先定义好的一些特定字符(专业术语称之为元字符)，以及这些字符的组合，组成一个特定的规则字符串。而所谓特定的规则，即是正则；因此特定的规则字符串，即是正则表达式。

因此，正则表达式是一种特殊的字符串(即正则表达式字符串，往往直接简称为正则表达式或正则式)，用来描述、匹配、过滤符合某些特征的其它字符串(即输入字符串、源字符串、被测试的字符串、被匹配的字符串，往往直接简称为字符串)。一般而言，典型的简单搜索和替换操作，可通过直接提供与预期的搜索结果相匹配的字面文本来实现。虽然这种方法对于文本执行简单的、静态的搜索和替换任务可能已经足够了，然而却缺乏足够的灵活性和动态性。若通过使用正则表达式，则可以：

1. **查找文本。**查找符合某一正则表达式的文本，尤其是查找符合某一正则表达式的非固定文本，比如查找符合某一种模式(甚至长度不定)的文本。
2. **提取文本。**可以查找字符串内符合某个文本模式的文本(子字符串)，然后将其提取出来以备他用。
3. **验证文本。**是指检査文本能否完全由正则表达式匹配，主要用来测试和保证数据文本的合法性。
4. **替换文本。**可以使用正则表达式所表示的文本模式来识别、匹配文档中符合该文本模式的所有文本(即符合该文本模式的文本的集合)，而不只是识别、匹配某个特定的、确切的文本(比如0XXX-XXXXXXXX就是电话号码模式，而0755-88888888就是某个特定的、确切的电话号码），然后可以完全删除匹配该文本模式的所有文本(相当于用空字符串替换)或者用其他文本逐一进行替换。
5. **切分文本。**切分也是正则表达式的常见操作之一，切分操作一般以正则表达式匹配的文本作为间隔，将字符串切分成多个片段。

本文利用UtraEdit工具首先对txt文件进行一下操作：

1. **删除行首空行：**利用正则表达式“%[ ^t]+”来搜索文中所有的行首空行，然后全部替换成空字符串。
2. **删除行尾空格：**利用正则表达式“[ ^t]+$”来搜索文中所有的行尾空行，然后全部替换成空字符串。
3. **删除章节名称：**利用正则表达式“第\*节”来搜索文本中所有的章节信息，然后全部替换成空字符串。
4. **删除空行：**利用正则表达式“%[ ^t]++^p”来搜索文本中所有的章节信息，然后全部替换成空字符串。

整理后的txt文件不仅保留了所有原著中的信息，而且删除了多余的空格、

空行便于下一步的进一步分析。以路遥的《匆匆过客》为例，图3.2展示了例前的本文状态。图3.3展示了处理后的文本状态。

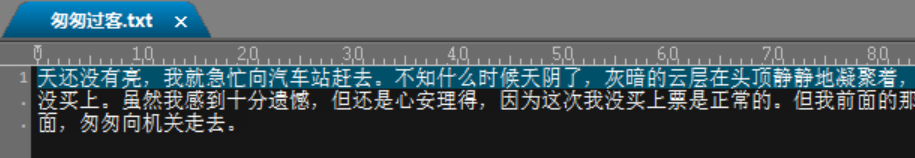


图3.3 《匆匆过客》文本处理后

从图3.2和3.3可以明显地看出，处理后的文本删除了空格、空行、行首空格等无用信息，完整地保留了原著中的核心信息，便于下一步进行文本处理和分析。



图3.2 《匆匆过客》文本处理前

3.5 分词和标注

在利用UtraEdit工具对作家的著作进行初步整理后，得到相对规范的本文，然后对规范的文本进行分词和词性标注。

**3.5.1** **Jieba框架对文本进行分词的原理**

jieba自带了一个叫做dict.txt的词典,里面有2万多条词,包含了词条出现的次数(这个次数是于作者自己基于人民日报语料等资源训练得出来的)和词性。该框架具有如下特征。

1. 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)，其主要过程分为两步：（1）根据dict.txt生成trie树，字典在生成trie树的同时, 把每个词的出现次数转换为频率（jieba自带一个dict.txt的词典, 里面有2万多条词, 包含了词条出现的次数和词性(作者基于人民日报语料等资源训练得出来)。trie树结构的词图扫描, 说的就是把这2万多条词语, 放到一个trie树中, trie树是有名的前缀树, 也就是说一个词语的前面几个字一样, 就表示他们具有相同的前缀, 就可以使用trie树来存储, 具有查找速度快的优势）；（2）对待分词句子, 根据dict.txt生成的trie树, 生成DAG, 通俗的讲, 就是将句子根据给定的词典进行查词典操作, 生成所有可能的句子切分。jieba在DAG中记录的是句子中某个词的开始位置, 从0到n-1(n为句子的长度), 每个开始位置作为字典的键, value是个list, 其中保存了可能的词语的结束位置(通过查字典得到词, 开始位置+词语的长度得到结束位置）。
2. 动态规划查找最大概率路径, 找出基于词频的最大切分组合。首先查找待分词句子中已经切分好的词语（全模式下的分词list）,得出查找该词语出现的频率(次数/总数),如果没有该词(基于词典一般都是有的),就把词典中出现频率最小的那个词语的频率作为该词的频率。然后根据动态规划查找最大概率路径的方法,对句子从右往左反向计算最大概率(这里反向是因为汉语句子的重心经常落在后面（右边）, 因为通常情况下形容词太多, 后面的才是主干。因此, 从右往左计算, 正确率要高于从左往右计算, 这里类似于逆向最大匹配), P(NodeN)=1.0, P(NodeN-1)=P(NodeN)\*Max(P(倒数第一个词)依次类推,最后得到最大概率路径,得到最大概率的切分组合。
3. 采用了基于汉字成词能力的 HMM 模型对未登录的词进行分词。利用HMM模型将中文词汇按照BEMS四个状态来标记, B是开始begin位置, E是end结束位置, M是middle中间位置, S是singgle单独成词的位置。jieba采用(B,E,M,S)这四种状态来标记中文词语, 比如北京可以标注为BE,即北/B京/E, 表示北是开始位置, 京是结束位置, 中华民族可以标注为BMME, 就是开始, 中间, 中间, 结束。

**3.5.2 Jieba框架对文本进行分词的主要代码**

本文的所有均利用Python语言实现，通过调用Jieba分词中的相关接口（图3.4），实现在指定包名下的所有作品进行分词，并将分词后的结果记录在特定的包中。

def cut\_words(dir):

"""

分词

:param dir: data包下的作者名字

:return: NULL

"""

data\_path = os.path.abspath(os.path.join(os.path.join(os.getcwd(),".."),"data"))

#根据不同的dir名字对文件进行分词

parent\_path = os.path.join(data\_path,dir)

write\_parent\_path = os.path.join(parent\_path, "participle")

if os.path.exists(write\_parent\_path):

pass

else:

os.mkdir(write\_parent\_path)

#遍历指定目录下的所有文件

for file\_name in os.listdir(parent\_path):

file\_path = os.path.join(parent\_path,file\_name)

if os.path.isdir(file\_path):

continue

output = ''

with open(file\_path, 'r') as f:

content = f.read()

word\_list = jieba.cut(content, cut\_all=False)

output = "/ ".join(word\_list)

f.close()

#将分词后的内容写入到相同文件名的participle目录中

write\_file\_path = os.path.join(write\_parent\_path, file\_name)

with open(write\_file\_path, 'w') as f:

f.write(output)

f.close()

图3.4 自动分词的主要逻辑

**3.5.3** **Jieba框架对文本进行标注的原理**

jieba分词的词性标注过程非常类似于jieba分词的分词流程，同时进行分词和词性标注。在词性标注的时候，首先基于正则表达式（汉字）进行判断，如果是汉字，则会基于前缀词典构建有向无环图，然后基于有向图计算最大概率路径，同时在前缀词典中查找所分出的词的词性，如果没有找到，则将其词性标注为“x”（非语素字 非语素字只是一个符号，字母x通常用于代表未知数、符号）；如果HMM标志位置位，并且该词为未登录词，则通过隐马尔科夫模型对其进行词性标注；如果是其它，则根据正则表达式判断其类型，分别赋予“x”，“m”（数词 取英语numeral的第3个字母，n，u已有他用），“eng”（英文）。

**3.5.2 Jieba框架对文本进行标注的主要代码**

本文的所有均利用Python语言实现，通过调用Jieba词性标注中的相关接口（图3.5），实现在指定包名下的所有作品进行词性标注，并将分词后的结果记录在特定的包中

def cut\_word\_and\_label(dir):

"""

分词和标注

:param dir: data包下的作者名字

:return: NULL

"""

data\_path = os.path.abspath(os.path.join(os.path.join(os.getcwd(), ".."), "data"))

# 根据不同的dir名字对文件进行分词

parent\_path = os.path.join(data\_path, dir)

write\_parent\_path = os.path.join(parent\_path, "participleAndLabel")

if os.path.exists(write\_parent\_path):

pass

else:

os.mkdir(write\_parent\_path)

# 遍历指定目录下的所有文件

for file\_name in os.listdir(parent\_path):

file\_path = os.path.join(parent\_path, file\_name)

if os.path.isdir(file\_path):

continue

output = ''

with open(file\_path, 'r') as f:

content = f.read()

word\_list = pseg.cut(content)

for w in word\_list:

output += w.word + "/" + w.flag + " "

# 将分词后的内容写入到相同文件名的participleAndLabel目录中

write\_file\_path = os.path.join(write\_parent\_path, file\_name)

with open(write\_file\_path, 'w') as f:

f.write(output)

f.close()

图3.5 词性标注的主要逻辑

3.6 创建自定义字典

3.6.1 斯坦福大学开源CoreNLP介绍

几乎之前所有的神经机器翻译（NMT）使用的词汇都受限，随后可能用一个方法来修补未知的单词。该框架展示了一个全新的能实现开放词汇神经机器翻译（open vocabulary NMT）的词-字符解决方法。建立了一个混合的系统，能够实现大部分的词级（word level）翻译，并可查阅罕见词的字母组成。字符级的循环神经网络能计算源词的表征，并能在需要时恢复未知的目标词。这种混合的方法还有一个双重优点是，与基于字符的网络相比，它更快且更容易训练；同时，它不像基于词的模型那样会产生未知的词。

斯坦福大学的CoreNLP是用处理自然语言的工具集合。它可以给出词语的基本形式：词性（它们是公司名、人名等，规范化得日期，时间，和数字），根据短语和语法依赖来标记句子的结构，发现实体之间的关系、情感以及人们所说的话等。该框架具有以下特点：（1）它是一个众多语法分析工具集成的工具包；（2）对任意的文本来说它具有快和鲁棒性强的特点，并且还广泛的用于生产中；（3）它还具有时用性，及时性；（4）支持数量众多的（主要）自然语言；（5）支持编程语言的接口丰富；（6）能够作为简单的web服务运行。

CoreNLP的目的是为了让一系列语言分析工具应用到一段文本上变得更容易。用两行代码就可以让一段原生的文本在一个工具管道（一系列处理文本的操作）中流过。CoreNLP 具有很强的灵活性和扩展性。通过一个选项的设置你就可以选择那些语言处理的工具可以应用到这段文本。CoreNLP包含许多的斯坦福的nlp工具，包括：词性标注器、命名实体的识别器、解析器（句子与语法结构）、指代消解器（就是在篇章中确定代词指向哪个名词短语的问题）、情感分析器、 引导模式学习器、开放信息提取器。而且，一个解释管道还可以包含其他自定义或者第三方的解释器。

3.6.2 **创建用户自定义字典的主要代码**

通过调用NLP创建自定义字典中的相关接口（图3.6），实现在指定包名下的所有作品进行词性标注，并将分词后的结果记录在特定的包中。

def create\_dicts(dir):

nlp = StanfordCoreNLP(r'G:\stanford\_nlp', lang='zh')

data\_path = os.path.abspath(os.path.join(os.path.join(os.getcwd(), ".."), "data"))

# 根据不同的dir名字对文件进行分词

parent\_path = os.path.join(data\_path, dir)

write\_parent\_path = os.path.join(parent\_path, "dicts")

if os.path.exists(write\_parent\_path):

pass

else:

os.mkdir(write\_parent\_path)

# 遍历指定目录下的所有文件

for file\_name in os.listdir(parent\_path):

file\_path = os.path.join(parent\_path, file\_name)

if os.path.isdir(file\_path):

continue

output = ''

with open(file\_path, 'r') as f:

content = f.read()

list = nlp.ner(content)

for word, flag in list:

output += word + '\n'

output = output.encode('utf-8')

write\_file\_path = os.path.join(write\_parent\_path, file\_name)

with open(write\_file\_path, 'wb') as f:

f.write(output)

f.close()

图3.6 自动创建自定义字典的主要逻辑

3.7 依据自定义字典进行分词

3.7.1 基于字典的Jieba分词

开发者可以指定自己自定义的词典，以便包含jieba词库里没有的词。虽然jieba有新词识别能力，但是自行添加新词可以保证更高的正确率。具体的用法是调用相关的接口jieba.load\_userdict(file\_name)（file\_name为自定义词典的路径）。词典格式和dict.txt一样，一个词占一行；每一行分三部分，一部分为词语，另一部分为词频，最后为词性（可省略），用空格隔开。

3.7.2 **基于用户自定义字典的分词代码**

通过调用Jieba根据自定义字典中的相关接口（图3.6），实现在指定包名下的所有作品进行分词，并将分词后的结果记录在特定的包中。

def cut\_word\_with\_dict(dir):

data\_path = os.path.abspath(os.path.join(os.path.join(os.getcwd(), ".."), "data"))

# 根据不同的dir名字对文件进行分词

parent\_path = os.path.join(data\_path, dir)

write\_parent\_path = os.path.join(parent\_path, "participleWithDict")

# get dit

dict\_path = os.path.join(parent\_path, "dicts")

if os.path.exists(write\_parent\_path):

pass

else:

os.mkdir(write\_parent\_path)

# 遍历指定目录下的所有文件

for file\_name in os.listdir(parent\_path):

file\_path = os.path.join(parent\_path, file\_name)

if os.path.isdir(file\_path):

continue

dict\_file\_path = os.path.join(dict\_path, file\_name)

jieba.load\_userdict(dict\_file\_path)

output = ''

with open(file\_path, 'r') as f:

content = f.read()

word\_list = jieba.cut(content, cut\_all=False)

output = "/ ".join(word\_list)

f.close()

write\_file\_path = os.path.join(write\_parent\_path, file\_name)

with open(write\_file\_path, 'w') as f:

f.write(output)

f.close()

图3.7 自动创建自定义字典的主要逻辑

* 1. **数据的你和以及可视化**
     1. **Quita**
     2. **MATTR**
     3. **mawatatarad2**

1. 工作总结与展望

21世纪信息是行业竞争的核心，数据是行业发展的重要基础。当前，数据分析和数据挖掘成为当下商业活动的重要组成。大数据技术、云计算及物联网等信息通信技术的出现，难以满足海量数据的价值挖掘和内在分析。面对复杂、海量、低密度及快速生成的数据，必须深度分析才能够获取所需信息，以构架数据分析平台及系统快速完成数据

文学家主观虚构的文学世界,来源于现实,反映了现实世界,因此每个时期的作品或多或少都有着那个时期特有的特点。在文学作品中,人物的优缺点会被放大,这是作者为了将对现实世界的想法赋予到作品中的人物身上，通过对文学作品进行分析不仅能够了解文学家主观思想还能了解当时的时代特色。

本文对路遥、陈忠实以及贾平凹等作家的作品进行研究分析，具体来说，本文的主要研究内容包括：

分词：利用开源框架Jieba提供的接口对任意作品进行分词。

分词和标注：利用Jieba相关的接口对分词后的词语进行词性标注。

构建用户字典，进行精确分词，并与缺少字典的分词结果进行对比：利用斯坦福大学提供的NLP框架，对任意文学作品构建字典，并依据构建后的字典进行分词，最后与缺少用户字典的额分词结果进行对比。研究结果表明存在用户字典的分词结果根据准确。

利用已有的工具对分词结果进行量化：利用Quita，Mattr，Mawatatarad等工具对分词后的结果进行图形化展示。

参考文献

[1] 妮鲁帕尔·艾山江. 自然语言处理技术综述[J]. 商情, 2013(39):326-326.

[2] 张妮, 徐文尚, 王文文. 人工智能技术发展及应用研究综述[J]. 煤矿机械, 2009, 30(2):4-7.

[3] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development.1958,2(2): 159-165

[4] Hulth A.Improved Automatic Key word Ectraction Given More Linguistic Knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing[C]. Sapporo,Japen,2003:216-223.

[5] T u r n e y P D . L e a r n i n g a l g o r i t h m s f o r k e y p h r a s e e x t r a c t i o n [ J ] . I n f o r m a t i o n Retrieval,2000,2(4):303-336.

[6] Mihalcea R,Tarau P.TextRank:Bringing Order into Texts.In:Proceedings of Empirical Methods

in Natural Language Processing,Barcelona,Spain[C].2004:404-411.

[7] LIU Xiao-hua，ZHANG Shao-dian，WEI Fu-ru，et al.Recognizing named entities in

tweets［C］//Proc of the 49th Annual Meeting of the Association for

Computational Linguistics：Human Language Technologies. Stroudsburg，PA：ACL，2011：

359-367.

[8] ChenLiwei,Feng Yansong,Zhao Dongyan.Extracting relations from the Web via

weakly super vised learning[J]. Journal of Computer Research and Development,2013,50(9):1825-1835（in Chinese）.

[9] Wang Z, Li J, Tang J. Boosting cross-lingual knowledge linking via concept annotation[C] ∥ Proc of the 23rd Inter-national Joint Conference on Artificial Intelligence,2013: 2733-2739.

[10] Tian J G H Z. Impaction of Climate Change and Human Activities to Land Desertification in Jieba Mountain Over Last 50 Years[J]. Bulletin of Soil & Water Conservation, 2007.

[11] Cheng F, Jing L I, Chang D D, et al. Characterizing Spatial Distribution Patterns of Direct Solar Radiation in Jieba Mountain Using GIS[J]. Bulletin of Soil & Water Conservation, 2011.

[12] 邢彪, 根绒切机多吉. 基于jieba分词搜索与SSM框架的电子商城购物系统[J]. 信息与电脑(理论版), 2018(7):104-105.

[13] 佚名. 基于深度学习的中文分词模型应用研究[D]. 重庆大学, 2017.

[14] 基于信息可视化与数据挖掘的高校图书馆推荐系统的设计与实现[D]. 北京交通大学, 2014.

[15] Aprosio A P, Moretti G. Italy goes to Stanford: a collection of CoreNLP modules for Italian[J]. 2016.

[16] Min S, Chambers T. Text Mining with the Stanford CoreNLP[M]// Measuring Scholarly Impact. 2014.

[17] Solaimani M, Gopalan R, Khan L, et al. Spark-Based Political Event Coding[C]// IEEE Second International Conference on Big Data Computing Service & Applications. 2016.

[18] Steedman M. Information-Structural Semantics for English Intonation[M]// Topic and Focus. 2008.

致谢

在xx大学度过了四年的紧张学习时光，系统地学习了工商管理的各方面知识，深深的佩服各位专业老师的学识，从中我不仅学习到管理知识，而且学到很多做人、做事、做学问的道理，在此表示真挚的谢意。在论文即将完成之际，我要感谢我的导师xxx教授。在论文撰写的整个过程中，从论文选题、到撰写开题报告、最后到正文撰写，x老师都提出了很多宝贵意见。x老师指出的每一个问题，指导的每一个思路，都使我有醍醐灌顶之感。给我感受最深的是x老师严谨治学的态度，无论从格式规范、论文要点、还是文章结构，于老师都不厌其烦，给予我及时的帮助，使我能够最后顺利完成论文写作工作。

　　在此我要感谢xx大学xx系的所有老师，你们无私的奉献精神和爱岗敬业的治学态度，不仅使我对管理理论有了更进一步的理解，将理论和自己的工作互相印证，受益匪浅。而且使我能够将所学理论应用于对现实问题的分析和解决，继而提高自己的管理水平。

　　感谢我的各位同学，是你们的无私帮助让我感受到校园的温暖，在我的论文写作过程当中，多位同学为我提供了信息支持，在此一并表示感谢。最后再次感谢xx大学为我提供了宝贵的学习机会，使我能够走上一个新的平台，幵始一段新的人生!

父书空满筐，母线萦我襦。父亲母亲是我们最温暖的港湾，他们的爱是最无私的，不断地鼓励我继续前进。 真心希望他们身体健康、笑口常开。