文本复制检测报告单(全文标明引文)

№:ADBD2018R_201805301531272018053015484744017425326 9 检测时间:2018-05-30 15:48:47

检测文献: 53141332_黄紫宁_计算机科学与技术（网络与信息安全）_经侦案件人工智能定性系统设计与实现(1)

作者: 黄紫宁

检测范围:
- 中国学术期刊网络出版总库
- 中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
- 中国重要会议论文全文数据库
- 中国重要报纸全文数据库
- 中国专利全文数据库
- 图书资源
- 优先出版文献库
- 大学生论文联合比对库
- 互联网资源(包含贴吧等论坛资源)
- 英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
- 港澳台学术文献库
- 互联网文档资源
- CNKI大成编客-原创作品库
- 个人比对库

时间范围: 1900-01-01至2018-05-30

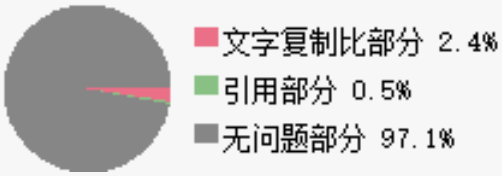
检测结果

总文字复制比：2.9% 跨语言检测结果：0%

去除引用文献复制比：2.4% 去除本人已发表文献复制比：2.9%

单篇最大文字复制比：0.7%（本科毕业论文致谢模板-百度文库）

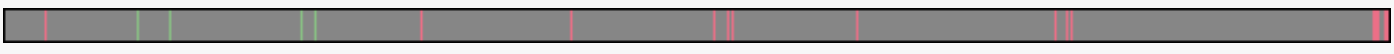
重复字数：	[895]	总段落数：	[7]
总字数：	[30992]	疑似段落数：	[6]
单篇最大重复字数：	[207]	前部重合字数：	[153]
疑似段落最大重合字数：	[315]	后部重合字数：	[742]
疑似段落最小重合字数：	[38]		



指 标：☐疑似剽窃观点 ☑疑似剽窃文字表述 ☐疑似自我剽窃 ☐疑似整体剽窃 ☐过度引用

表格：0 公式：0 疑似文字的图片：0 脚注与尾注：0

- 2.3%（66） 中英文摘要等（总2819字）
- 2.7%（155） 第1章绪论（总5838字）
- 2.2%（199） 第2章中文文本分类的关键技术_第1部分（总8957字）
- 1.4%（38） 第2章中文文本分类的关键技术_第2部分（总2682字）
- 2%（122） 第3章基于CNN的经侦案件文本分类模型（总6021字）
- 0%（0） 第4章使用基于CNN的文本分类模型对案件文本分类（总3414字）
- 25%（315） 第5章总结与展望（总1261字）



（注释：无问题部分 文字复制比部分 引用部分）

1. 中英文摘要等 总字数：2819

相似文献列表 文字复制比：2.3%(66) 疑似剽窃观点：(0)

1	论我国居住权制度的构建 解斐斐(导师：徐涤宇) - 《湖南大学硕士论文》- 2009-04-20	2.3%（66） 是否引证：否

原文内容 红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

吉林大学学士学位论文（设计）承诺书

本人郑重承诺：所呈交的学士学位毕业论文（设计），是本人在指导教师的指导下，独立进行实验、设计、调研等工作基础上取得的成果。除文中已经注明引用的内容外，本论文（设计）不包含任何其他个人或集体已经发表或撰写的作品成果。对本人实验或设计中做出重要贡献的个人或集体，均已在文中以明确的方式注明。本人完全意识到本承诺书的法律结果由本人承担。

学士学位论文（设计）作者签名：

2018年5月20日

摘要

经侦案件人工智能定性系统设计与实现

经济犯罪侦查是一项致力于预防以及打击经济犯罪的专门调查活动，对国家经济安全、保障社会稳定有重要作用。但伴随社会经济发展，当前的经侦活动也有诸多掣肘之处。人工智能可以通过算法实现程序的自主学习并进行决策，尤其在解决知识密集型问题上有很好的效果。近几年神经网络技术在图像和文本的处理上都取得了较好的成绩，而在自然语言处理（NLP）、文本分类中，卷积神经网络（Comvolutional Neural Network,CNN）有较为突出的成果。另外绝大多数的法学活动都是可以找到相应可计量的标准规律的，定量法学研究也是当前社会科学发展得意一个必然趋势。所以将人工智能技术用于法学方面的案件侦查定性的研究，是确实可行，顺应时代潮流，且又具有实用价值的。本研究涵盖的主要的工作内容如下：

收集用于模型训练的训练和测试用的数据，选取合同诈骗罪和信用卡犯罪两类经济犯罪类型为例，以记录合同诈骗与信用卡诈骗犯罪事实的叙述性中文文本为语料。

研究了传统文本分类方法。将对原始犯罪案件描述文本进行分词，并标注词性，分词后对数据进行清洗，降低训练维度。

对清洗后的数据进行结构化处理，建立字典，将中文文字映射成数字。

搭建CNN神经网络模型，设置并调整超参数，用带标签的训练和验证数据集训练该神经网络模型，直到模型可以输出较高的准确率为止。

使用未参与训练的测试数据集测试上一步获得的神经网络模型。分析研究测试结果。

实验结果表明，CNN神经网络模型对犯罪描述文本有较好的分类效果，可以用于经济侦查案件的定性。

关键词：经侦定性,卷积神经网络,文本分类,自然语言处理

Abstract

Economic crime investigation is a special investigation activity dedicated to preventing and cracking down on economic crimes. It plays an important role in national economic security and social stability. However, with the socio-economic development, the current economic investigation activities also have many limitations. Artificial intelligence can realize self-learning and decision-making through algorithms, especially in solving knowledge-intensive problems. In recent years, neural network technology has achieved good results in the processing of images and texts. In NLP and text classification, the Convolutional Neural Network (CNN) has made outstanding achievements. . In addition, the vast majority of legal activities can find corresponding quantifiable standard laws, and quantitative law research is an inevitable trend in the development of social science. The main tasks covered in this study are as follows:

(1) Collect training and test data for model training, and select two kinds of economic crimes such as contract fraud crimes and credit card crimes as examples to record narrative Chinese texts of contract fraud and credit card fraud crimes as corpus.

(2) Research on traditional text classification methods. Segmentation of the original complete corpus and word tagging. Then clean the data sample after word segmentation to reduce the training dimension.

(3) Structured processing of cleaned data, creating a dictionary to map Chinese text into numbers.

(4) Build a CNN neural network model, set and adjust hyperparameters, train the neural network model with labeled training and validation data sets until the model can output a high accuracy.

(5) Test the neural network model obtained in the previous step using test data sets that are not involved in training. Analyze the study test results.

The experimental results show that the CNN neural network model has a good classification effect on crime description texts and can be used to determine the nature of economic investigation cases.

Keywords: detective , CNN , text classification , NLP

目录

2. 第1章绪论

总字数：5838

相似文献列表　文字复制比：2.7%(155)　疑似剽窃观点：(0)

1	量化法学及人工智能在民商法学中的应用 张妮;杨亘; -《民商法争鸣》- 2015-06-15	2.7%（155） 是否引证：是

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第1章绪论

1.1 课题背景及意义

1.1.1. 论文选题背景

“经侦”全名“经济犯罪侦查”，是一项致力于预防以及打击经济犯罪的专门调查活动，是维护国家经济安全、保障社会稳定的重要举措。随着社会经济的发展，目前我国已经进入了一个斗争复杂的新时期。一方面是现在已经进入了法律信息大爆炸的时代，旧有的基于关键字的法律检索早已不能满足当前时期现实对法律知识管理的要求，法学***发展迫切需要新技术加入进来，需要新技术来提高法律使用的效率***。另一方面，是由于经济犯罪与人民日常经济生活息息相关，有强关联性，并且在现实中涉及的经济领域广、经济环节多，经济犯罪常常假扮作合法的经济活动出现，同时具有复杂性和隐蔽性。尤其是现当下，伴随新型科学技术及新兴产业的发展，我国也出现了许多全新经济犯罪类型，违法犯罪者的犯罪手法变得越来越多样化。

以至于，当前时期的经侦工作虽然已经有了显著发展，但是在复杂的经济犯罪新形势面前人们仍然面临着许多难点，比如经济犯罪案件的侦查定性通常关系多部法律，涉及包含经济、法律、侦查等多方面，关联的知识领域宽广，侦查的案件复杂，且犯罪手段更新快，这样的现实发展状况对经侦从业人员无论在知识储备总量还是知识更新频率两方面都有较高要求，并且经济犯罪案件的多发使得针对经济犯罪案件的相关处理工作量大，现有经济犯罪侦查员队伍人员素质和人员数量往往还是难以满足现实斗争需要，有较高的案件处理成本。

绝大多数的法学活动都是可以找到相应可计量的标准规律的，研究定量法学研究是社会科学发展的一个必然趋势，案件定性本质上也是根据相关标准进行的标准比对与分类活动，和人工智能技术相性良好。近几年神经网络技术在图像和文本的处理上都取得了优秀的成绩。在自然语言处理（NLP）、文本分类中，卷积神经网络（Comvolutional Neural Network,CNN）有较为突出的成果。人工智能可以通过算法实现程序的自主分析决策，并能同时随着发展不断更新学习，可以高效低成本的紧跟现实发展不断进行更新。尤其在解决知识密集型问题的方面，人工智能往往能取得很好的效果，例如经侦定性系统这样***专门供政府部门使用为执行某项任务产生的法律本体系统,功能比较简单,涉及的领域比较集中,在实践中可以取得较好的实施效果***，可以极大的节约成本。

所以将人工智能技术用于法学方面的案件侦查定性的研究，是确实可行，顺应时代潮流，且又具有实用价值的。

1.1.2. 目的和意义

现在对经济犯罪案件的定性普遍都是由人来人工进行的，需要消耗一定的人力成本和经费成本，且案件处理效率不高。使用人工智能技术搭建经济犯罪案件侦查系统有利于增强调查者的能力，可以有效缩短经济犯罪侦查程序的简约进路，可以极大的提高案件处理效率。而且神经网络技术赋予了计算机系统优秀的更新学习能力，可以便利的随着现实发展而及时快速的进行更新。最后犯罪调查的自动化和支持还可以有效削减人力成本和经费成本，使潜在的节约成为可能。甚至还可以为经验不足的调查人员提供培训以及规范调查程序，从而，并且使各种调查人员都可以处理同一起案件。

将人工智能运用于犯罪案件定性是对法律理论、文本、案例增加趋势的顺应，也是对国家大力发展政府机构信息化号召的积极响应。

所以，基于以上原因，本论文拟研究一种针对犯罪案件描述文本的文本分类系统，利用计算机结合人工智能技术对经济犯罪案件的描述文本进行犯罪类型定性，使计算机能自主识别出犯罪事实是何种犯罪类型，对经济犯罪案件进行定性。

1.2 研究现状

1.2.1. 文本数据分析发展现状

人脑可以快速、自然高效的理解语言、文字里蕴含的意义，但是理解语言文字的意义对计算机来说是困难的，本研究也是建立在计算机对中文语句“理解”的基础上的。自然语言处理的研究已经有50多年的历史，并随着计算机的兴起而成长为语言学领域。文本数据分析是自然语言识别中一个非常重要的环节。

传统的文本是识别（数据分析）是有人工来完成的，需要人工手动分类文档或手工制作自动分类规则。但在计算机互联网普及，人们步入大数据时代之后，以计算机网络、大型数据库的建立以及强大计算功能为背景, 文本数据具有数据量大的特点，文本数据信息量大的特点使得传统的分本分类方法处处掣肘。直到二十世纪末，基于统计的机器学习算法和全新统计方法出现，并且最终把基于规则的语言方法取而代之。统计文本分类使用机器学习方法来学习基于人类标记的培训文档的自动分类规则，而不是传统的手动分类文档或手工制作自动分类规则。主要是因为它们具有更好的结果，具有更好的速度和稳健性，研究自然语言的统计方法在这个领域占主导地位。

近十年以来，伴随着人工智能技术的发展，文本数据分析技术更是有了突飞猛进的发展，包括分本表示算法方面和机器学习算法方面两个方面。

在文本表示算法方面，理论技术已经从最初的辞典表示，已经经历了到基于上下文的表示，到词向量的表示的一系列发展。而中文环境吓得文本数据由于汉语和英语不同的独特语言特性，也发展处了针对中文文本数据语言处理的一些方法。

在机器的自然语言学习方面，则可以分为监督和无监督学习，特别是最近研究活动的一个领域是使用自动学习技术对文本文档进行分类。基本的NLP任务包括标记和解析，词形/词干，词性标注，语言检测和语义关系识别。NLP任务将语言分解成更短的元素，尝试理解片断之间的关系，并探究片断如何协同工作以创造意义。LP有助于解决语言中的歧义问题，并为许多下游应用程序（如语音识别或文本分析）的数据添加有用的数字结构。这些方法对于解决包括但不限于关键词标记，词义消歧，信息过滤和路由，句子分析，在解决相关文档聚类和将文档分类成预定义主题的问题是有用的，而本研究也正是要用到将文档分类成预定义主题的方法，而深度学习的概念提出后，深度学习现在也被广泛用于建模人类语言，2017年里Fcebook使用PyTorch为机器学习做出了巨大贡献，PyTorch框架多被用于NLP处理，十分受人工智能方面研究者追捧，同年Tensorflow也发出了能够稳定向后兼容的API，新API发布同时还更新有多个伴随库，伴随库的更新使Tensoflow框架的动态性也随之增强，变得更加利于自然语言处理。除谷歌和脸书两家公司外，许多其他公司也投身机器学习热潮中，包括苹果公司于同年发布CoreML移动机器学习库；Uber的团队发布了一种深度概率编程语言Pyro；优步公司公开了米开朗基罗机器学习基础设施平台的详情。另外Google公司还宣布将在北京开设一个新实验室，因为中国在机器学习领域备受世界关注。由于资金充沛，人才丰富，并且大量潜在易于提供的政府数据，中国在人工智能发展、应用还有推广方面都有巨大潜力，另外中国政府本身也致力于推动大数据计划的发展，提出了AI计划。

目前国内外较为流行的较为成熟的文本数据分析方法有很多，比如有：SVM方法，K-近邻方法（K-NN）、Rocchio算法、朴素贝叶斯分类方法、决策树、遗传方法、神经网络。

而在本分数据分析的应用方面。每天从医疗记录到社交媒体产生的非结构化数据数量惊人，从医学到教育到法律。随着非结构化信息的数量继续呈指数级增长，催生出自然语言识别和文本数据分析的许多新应用方向，其中包括：识别电子邮件或书面报告中的模式和线索，以帮助发现并解决犯罪；将内容分类为有意义的主题，以便用户可以发现趋势；还有将文本数据分析技术用于社交媒体分析等等。而在基础NLP的基础上更进一步的NLP的一个称为自然语言理解（NLU）的子领域由于其在认知和人工智能应用领域的潜力而开始受到欢迎。现在一个热门关注点便是使用深度学习神经网络来执行文本数据分析的特定推

理任务以及开发强大的端到端系统。

NLP向NLU的演变对企业和消费者都产生了很多重要的影响。算法的强大功能可以在许多情况下理解人类语言的含义和细微差别，

大多数分析师似乎都认为，IT领域的下一个重大事件将涉及语义搜索。这将是一件大事，因为它将允许非主题专家使用自然语言来提出问题的答案。这种魔力将包含在搜索中的分析中，从而产生既相关又富有洞察力的答案。

1.2.2. 经济犯罪侦查发展现状

1.国外研究状况

国外早在00年以前就一些研究者曾讨论过法律专家系统（Gabrovsky，1988; Garner 1987; Kowalski和Sergot，1989; Susskind，1987; Rissland，1988; Rissland，1989; Wagner，1992），他们的工作讨论了如何进行法律推理并陈述了案例。并已经有过许多尝试尝试包括尝试建立专家系统来模拟选定的法律依据（Ciampi，1982），模拟税法（McCarthy，1977）和模拟案例准备（Popp & Schlink，1975）[1]。

之后亦有一些研究者讨论过经济犯罪警察调查专家系统（James E. Bowen，1994）[1]，还有利用AI技术进行欺诈检测的前瞻性评估和案例研究。

2.国内研究状况

国内的法律方面学者和从事刑侦行业的专业人员就法律知识运用和经济犯罪侦查两个方面都有过很多相关努力。随着现在进入法律信息爆炸的时代，研究定量法学研究已经成为了社会科学发展的一个必然趋势。旧有的***基于关键字的法律检索早已不能满足当前时期经济犯罪侦查活动对法律知识运用的要求***。目前我国经济犯罪案件已经普遍变得越发复杂，常常同时具有复杂性和隐蔽性双重特性，尤其是伴随新型科学技术及新兴产业的发展，我国也出现了许多全新经济犯罪类型，违法犯罪者的犯罪手法多样化，经济犯罪侦查活动迫切需要新技术来提高法律使用的效率。我国在当前时期的经侦工作虽然已经有了显著发展，但是在复杂的经济犯罪新形势面前人们仍然面临着许多丞待解决的困难，有极大的革新改良空间。

现如今国内人工智能在案件侦查和法律方面的研究讨论更多的是在刑事案件的侦查方面，应用也主要是应用于利用数据挖掘辅助案件侦破，如基于数据挖掘的刑事犯罪侦查***系统，还有一些基于评价规则和案件推理、针对司法裁定任务的系统***——如基于机器学习的计算机辅助量刑。

而在民商法和经济侦查方面，目前人工智能的运用尚还较少，虽然也有研究者讨论过量化法学及人工智能在民商法学中的应用，但暂时还没有具体针对经济犯罪案件侦查定性的系统实现。

1.3 本文研究内容

1.3.1. 研究方法

本文以合同诈骗罪和信用卡诈骗罪两类犯罪案件为样本，拟设计出能够对案情描述文本自主学习并定性分类的系统。将尝试使用卷积神经网络（全名Convolutional Natural Network, 后文将全部简写为CNN）模型为基础完成对两种经济犯罪案件描述分本进行分类的任务，通过对描述犯罪案件案情的中文文本分类实现经济犯罪侦查定性。本文中，课题将集中在以下的两个重点方面：

（1）获得将用于训练、验证和测试的文本数据，然后进行文本数据预处理。“文本”即文字、语言，是语言书面化的表示形式，本课题使用的合同诈骗罪和信用卡诈骗罪的案件描述文本是不含结构化字段的纯文本。为使计算机可以识别，要对该非结构性、形态自由的自然语言进行自然语言处理，使原本计算机无法理解的中文文本处理成计算机可以理解并可以进行计算的数据。先人工对训练和测试用的两个数据集合进行人工的类型表述，使单纯的非结构化文本转化为半结构化文本。其次将半结构化文本内部，非结构化的文字部分，进行分词分词操作，然后标注上词性。中间还同步进行数据清洗，去除文本中的歧意词，停用词以及语义助词也都要去除，还有标点符号。然后将所有词都统一到同一个词向量空间（包括训练、测试、验证三个数据集合，日后所有将输入该系统的文本中的词都将会被添加入该词向量空间中）中。通过建立词向量空间，将空间中的词隐射成数字，将半结构化的数据转化为结构化的数据，以期获得能输入进下游CNN模型进行训练、测试的结构化数据。

另外因为案件描述文本的一些具体特性，和中文文本的独特特性，将会有一些额外的优化需要，比如犯罪案件复杂性带来的长句多，句段长，描述文本长的特点，将会带来词向量空间过大，高纬度且稀疏的问题，高纬度稀疏性使得机器学习计算量大，学习计算时间长，计算机负担重，需要在文本预处理的过程中注意降维处理，并注意词向量空间搭建过程中的映射方法选择。

（2）以CNN模型及相关组成部件为组成成分，搭建神经网络模型，设计并且实现分类器，将上文中经过了预处理已经转化为结构化数据的数据集中带有人工标记的的训练集和验证集输入搭建的模型，进行反复训练和验证，并在训练过程中相应的调整神经网络模型结构和各超参数，直到获得一个较高的准确率为止，结束训练，获得可以用于分本犯罪案件案情描述文本分类的分类器模型，并最终用未参与训练过程的测试集对该系统进行测试，测试本系统的准确率。

本课题使用了相对简单的二分类分类器进行分类，通过对通过对描述犯罪案件案情的中文文本分类并且由于中文语言背景下的犯罪案件描述文本文长，数据维度高的特点而特别选用深度学习方法，使用卷积神经网络模型进行文本的分类，实际上达成经济犯罪侦查定性的目的。

1.3.2. 研究内容

本文的组织结构如下：

第一大章将主要概括本论文写作背景以及选择该课题的原因，将会从我国当今的经济犯罪侦查的发展形势、法学发展形

势、国内外人工智能技术和文本数据分析技术的发展形势、目的与技术结合的应用发展情况等多个方面，介绍了本课题的背景、目的和意义，介绍了本课题研究涵盖的主要内容，并列举文章章节安排。

第二大章则会详细解说本课题将涉及到的关键技术。主要是文本分类方面的专业技术，主要包含：汉语言环境洗的分词、文本数据结构化处理、文本分类方法几个方面。

第三章将简单介绍本课题使用的语料库的形成。然后着重针对CNN模型的设计和搭建进行介绍，并将对卷积神经网络的原理进行周详描述，并分析模型训练和验证的结果。

第四章使用训练所得的基于CNN的经侦案件文本分类模型对为参与模型训练的测试集进行案件文本分类测试，分类完成后还将分析测试结果以及对本系统的性能做出总结评价。

第五大章为终章，将在第五章对论文全文进行归纳总结，并总结本课题的项目成果，并对本课题未来工作方向进行展望。

3. 第2章中文文本分类的关键技术_第1部分

总字数：8957

相似文献列表　文字复制比：2.2%(199)　疑似剽窃观点：(0)

1	基于内容过滤的局域网防泄密系统的研究与实现 龙浩(导师：朱培栋) -《国防科学技术大学硕士论文》- 2009-03-01	0.5%（48） 是否引证：否
2	基于函数依赖改进隐含朴素贝叶斯的性能和鲁棒性 覃事东(导师：王利民) -《吉林大学博士论文》- 2014-05-01	0.5%（45） 是否引证：否
3	苏里格气田致密砂岩储层流动单元研究 李丁(导师：潘保芝) -《吉林大学博士论文》- 2014-05-01	0.4%（40） 是否引证：否
4	数据挖掘在城镇基本医疗保险中的应用分析与设计 张莎莎(导师：陈建中) -《贵州财经大学博士论文》- 2013-04-01	0.4%（34） 是否引证：否
5	基于Web的新词语发现研究 盛启东(导师：徐超;谭守标) -《安徽大学硕士论文》- 2010-05-01	0.4%（33） 是否引证：否
6	基于词典的文本情感计算系统的设计与实现 李其达(导师：陈少华) -《华中科技大学博士论文》- 2014-01-01	0.4%（33） 是否引证：否
7	垂直搜索引擎关键技术的研究与实现 贾岩峰(导师：赵志滨) -《东北大学博士论文》- 2013-06-01	0.4%（32） 是否引证：否
8	基于PCA-NBC算法的股票分类研究 王志(导师：焦桂梅) -《兰州大学博士论文》- 2014-04-01	0.3%（31） 是否引证：否
9	基于CUDA和深度置信网络的手写字符处理应用 陆军建(导师：林家骏) -《华东理工大学博士论文》- 2014-12-12	0.3%（29） 是否引证：否
10	基于Teradata数据仓库的银行系统的研究与服务提升 张 -《大学生论文联合比对库》- 2014-04-14	0.3%（29） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第2章中文文本分类的关键技术

2.1 引言

本章节将主要对本课题涉及的重要相关技术进行一个系统介绍，包括：中文分词方法；对文本数据进行清洗的过程即在预处理过程中降低数据维度的方法；文本数据的结构化处理，生成词向量空间；和具体的文本分类方法。

2.2 中文分词

要先使机器能理解人类的语言文本，才能使用机器对语言分本进行分析并对其进行分类。而人类语言是非结构化的文本，把非结构化文本转化为结构化文本要经历一个从对语言文本预处理到具体文本数据结构化的过程。语言文本预处理，即自然语言处理，在步骤上一般包含划分词、对词性进行标注、对命名实体进行识别、对句子进行句法分析这几个步骤。而其中分词是其他一切步骤的基础，合理的分词结果十分重要，分词的效果可能影响到后续文本数据分析的效果。

英文分词因为英文文本独特的语言特性，每个单词为一词汇，词汇与词汇之间由空格作为自然分节符分隔语句，句子与句子间再由标点分割组成大篇幅的文本。中文分词则不然，形式上每一段中文文本数据都可以看作是有多个汉字与标点符号组成的字符串。中文语句有字与字组成词，然后词与词可以组成词组，也可以词与词或词与字组成句子，更进一步句子与句子组成段、节、章、篇。

人可以很容易根据场景或者语境来判断一个句子的组成成分，容易找出句子中有实际意义且符合语境需要的词语，能够理解不同文本的语义，但是这对于计算机而言把句子分成词是一项复杂的工作。

对汉语进行分词根据划分词汇的原理不同可以大致分为三种方法，分别是基于字符串匹配的词汇划分方法，基于统计的词汇划分方法和基于理解的词汇划分方法，其中基于字符串匹配的词汇划分方法又常被称为机械分词。

基于字符串匹配的分词方法的原理是依靠一定的匹配规则，再配合一个庞大的词汇表，使用待分词语料依据相关匹配规

则与词汇表内词汇匹配，若匹配成功则切割出匹配成功的这个词汇。依据匹配规则不同，基于字符串匹配的分词方法又可以划分成四类：正向最大匹配方法（指通过对待分词的文字进行从左向右或朝前向后的扫描进行正则匹配），逆向最大匹配方法（指对待分词的文字进行从右往左或从后向前扫描，与正向最大匹配法扫描方向相反，但同样进行正则匹配），双最大匹配法（同时使用正向最大匹配方法和逆向最大匹配方法两种匹配方法），最后还有最小切分法。

基于统计的分词方法是利用中文中组成词汇的两个字因为词汇的存在而常常一同出现的现象。该分词方法讲统计统计上下文中字符与字符相邻出现的概率，两字符相邻出现的概率越高则它们联合起来恰好组成一个词语的概率也越大。该方法不适用字典，而主要使用概率统计方法，而依据概率统计方法的不同，基于统计的分词方法又可以划分为互信息统计法、N-Gram方法、基于组合度的方法等多种算法。

基于理解的分词方法是拟将分词任务与对句子进行语义分析的任务结合，同时分词，同时句法分析。该方法将构建一个由分词子系统、句法句意系统、还有总控部分，一种三块组成的系统。这种分词方法对解决歧义问题可以做出较好判断，歧义问题向来是困扰分词的一个难处。但是因为该类分词方法普遍需要极大数量的数据基地，且增加了把非结构化的文本数据转化为结构化数据的难度，所以该方法还处于探索实验阶段。

迄今为止已经有许多研究人员为中文分词问题做出了努力，现在已有一些可以使用的可用于中文分词的成熟工具包，比如jieba、Yaha、finalseg、pynlpir、scseg、Genius等等。

本课题拟使用字符集表示构建词汇表，使用训练集数据作为素材构建词汇表，将添加一个 <PAD> 将所有文本pad统一为同一长度，获得的字典将保存在文件'case.train.txt'中。设置了字典大小上限为5000，若训练规模更大使用更丰富的数据进行训练则还可扩大字典上限。

2.4 文本数据结构化处理

在文本数据分析中，文本数据通常表现为自由的、非结构的形态，或是一串语言文字，或是由许多符合特定计算机语言语法规则的文字字符串。如果直接以非结构化的语言文本数据为素材，是无法用现有数据挖掘方法进行分析的。文本数据挖掘需要建立在把文本转化为结构化数据的基础处伤得，文本数据分析必须在结构化框架下进行，所首先要在进行数据挖掘前对待分析数据进行数据结构化处理。

常规的数据结构化处理过程为：分词、进行结构化处理，后续由于部分文本的不同特点可能还会涉及到比如高维矩阵处理、降低维度等操作。

而其中文本数据结构化处理长常用的方法有：文档-词项矩阵（DTM）、词频-逆向文件频率（TF-IDF）、词向量。常用的库有sklearn库、gensim库、和word2vec工具包等。

2.3.1. 文档-词项矩阵

“文档-词项矩阵” 通常被简称DTM，转置后可记为TDM，它的原名是“Document-Term Matrix”，直观上，矩阵以行代表文档，列代表文档中含有的词汇，矩阵中的元素为一片文档中某一词汇出现的次数。一个文档-词项矩阵的列向量应包含矩阵中所含的全部文档内含有的词汇。

举例，以下有两个分此后的文本，文本一[经济，犯罪，信用卡，诈骗，调查]，文本二[经济，犯罪，合同诈骗，研究]，则基于这两个文本构建出的辞典包含7个不同的特征词汇，{1：经济，2：犯罪，3：信用卡，4：诈骗，5：调查，6：合同诈骗。7：研究}，而假设以one-hot形式表示以上两个文档，可以表示为：（1,1,1,1,1,0,0,）和(1,1,0,0,0,1,1),其中向量元素表示对应词汇在该句子中出现的次数，两向量合并则获得文档-词项矩阵。

构建文档-词项矩阵的基本思想是源于“词袋模型”。“词袋模型”的英文本名叫做“Bag of words”，词袋模型通常会被简写为BOW。这个模型的思想是把文本集合想象成一个口袋而口袋内装满了词汇，一份文字数据相当于一个装满词汇的口袋，亦即一份文字数据相当于一个由彼此相互独立、又没有谁许的的词汇共通组合在一起形成的的集合。其中文本可以是一个短句、一条长句，一条长段落又或者是一份完整的文档。词袋模型假设了词汇的独立性，简化了文本数据结构化过程中的计算，这个优点使文档-词项矩阵模型被广泛使用，但同时该模型亦有其缺陷，即该模型忽略了词汇间的顺序和依赖关系，可能降低获得的结构化数据对素材文本的代表性。但文档-词项矩阵仍然常被应用于很多数据分析过程，比如计算文档间的相关性、用于文档分类、用于文本聚类分析等等。

常用的可以构建文档-词频矩阵的库有scikit-learn库和gensim库。

Scikit-learn库将其中数据结构化处理的工具统一称为“特征抽取（Feature Extraction）”，包含从语言数据文本中进行特征提取和从图像数据中中进行特征提取。Sklearn.feature_extraction数据包中使用当中的text模块进行文本数据结构化处理，当中的CountVectorizer类可以同时实现分词处理和词频统计两项功能，将调用两次文档集合，一次调用文档集合创建词典，一次调用文档集合创建对应于每个文档的词频向量，从而最终得到文档-词频矩阵。

Gensim库则在corpora包的dictionary模块中提供用于文本结构化处理的工具。Dictionary模块下定义了类Dictionary，使用Dictionary类实现词汇和词汇id之间的映射，建立词典，另外Dictionary类下的doc2bow方法是词典方法，能够将文本词汇集合转化为词袋模型，并将获得的词袋模型以列表形式返回，以形式为(词汇id,词频)的元组为列表元素。

2.3.2. 词频-逆向文档矩阵

词频-逆向文档矩阵是在文档-词项矩阵基础上更进一步的发展。词频-逆向文档矩频率是一种根据词汇出现频率而对文档-词项矩阵进行调整的一种方法。

“词频-逆向文档频率”的英文原名全称叫做“Term Frequency-Inverse Document Frequency”，它通常会被简写，本文中后

文都将用简写进行称呼，简写称呼为TF-IDF。TF-IDF算法有一个设想的前提，该假设是：如果对一个文档最有代表性或者对区别一个文档和其他文档最有意义的词汇的存在，这个词汇应该在它对应的文字文本中所有词汇里拥有最高出现概率的，而同时这个词汇在整个文档集合内的所有词汇中，应该是拥有最低出现概率的——这意味着该词汇只在对应文档中大量出现。

在实验文本数据量很大的情况下，常常会出现一些出现频率高但却缺乏实际意义的词汇包括语气助词、表达停顿的词、另外还有标点还有其他不能称之为文字的符号，这些出现概率高的无意义符号可能会对文本数据分析的效果产生一定影响，比如使出现平率低于它的真正有标志性的重要词汇被忽略。

在分词过程中通过建立停用词表清洗掉停用词是一种对停用词的处理方法，而另一种处理方法就是利用词频-逆向文档频率对词频矩阵进行核心信息提取，使词汇能突出所代表文档的特点。

具体的操作是对矩阵内各个词汇的频率（tf）做进一步调整，通过降低停用词在DTM矩阵中的系数，从而降低这些词汇对文本数据分析的影响。如果单词项在它所代表的语言文本中出现概率高而在其他语言文本数据中只要很低的出场概率，那么将给这个词汇赋予较高重量（idf），通过按照逆文本频率（idf）对词频加权，将词频的绝对大小转化为相对大小，最终的特征空间将由tf*idf的值——即加权后的词频构成。

TF-IDF算法中的权重矩阵是二维的，其中元素[x,y]表示第y个词在第x个类别中的IF-IDF值。需要特别注意的是，在中文文本的数据分类任务中，往往有训练集、验证集和测试集三个集合。训练集和验证集用于模型训练，而使用模型对集合内数据进行分析的测试集，测试集的数据往往和训练集和验证集中包含的词向量有出入，假如测试集内包含新词，即使通过了数据清理环节新词也仍然保留着，这个新词仍然会被抛弃，因为新词只要不被训练集生成的TF-IDF词向量空间包含就会被抛弃，不这样做将无法保证所有输入的数据集合使用同样的词向量空间，保证了各集合公用相同的词向量空间各集合才能共用同一个运算模型使用相同参数。

TF-IDF方法可以一定程度上解决常用词和停用词被划分为关键特征的问题，使词项对文档的重要程度得到更好区分。

TF-IDF作为重要文本特征表示方法在文本分类聚类、对语义进行识别的操作、对信息检索进行优化和搭建推荐系统的任务中都有许多运用。

2.3.3. 词向量

所谓词向量，即把自然语言中的词汇转化成数值向量进行表示的方法。它经历了从one-hot representation到distributed representation的发展过程。

1. One-hot representation 方法

One-hot representation 方法指的是使用一个长度很长的长向量映射单个词汇。向量长度或者说是向量维度将等同于辞典尺寸，长向量内部的分量元素全部由数字零和数字一构成，其中数字一是唯一的，数字零却可以有无数个——零的个数将具体取决于向量长。其中并且由值等于一的分量在向量中所处的位置，表示的是这个词汇在辞典中的所处的位置。One-hot representation表示方法非常直观可见容易理解，这个表示方法运用十分广泛。但是对于数据量大词典尺寸很大的情况，One-hot表示方法往往适应不良，比如在对文章任务时，就会出现因为文章样本涉及词汇多而产生的维度灾难，高维度伴随着巨大的向量个数和向量长度，为存贮大量的超长向量将耗费许多空间资源，而对这些超长的向量结构进行计算也导致了繁重的运算压力，尤其是将One-hot representation 方法应用于深度学习的一些算法中时。另外一个方面是，One-hot representation方法无法刻画词与词之间的相似性，在One-hot representation表示方法下表示的词汇彼此之间没有联系，这种现象也被称作“词汇沟壑”。

2. Distributed representation方法

由上文可知One-hot representation方法有许多不尽人意的地方，为了弥补它的缺陷一位名叫Hinton的研究者研究出了一种新方法。这种新方法名为Distributed representation，中文可翻译为分布表示方法，该方法通过训练神经网络，利用训练后的神经网络把语言文本中的每一个词转换成一个比较短的向量，这些短向量统一有相同的长度，虽然向量的实际长度可能其实并不短，但是相对于同样数据用One-hot representation方法表示时的长度则短得多，这些向量的维度大小通常在50-100之间，训练方法常常是使用某种神经网络方法。这种方法把词汇表示从高度空间映射到低维空间，降低了实数向量维度。

通常称通过分布表示法获得的长度固定又长度相对较短的向量命名为“词向量”，同时把包含所有这些向量的集合叫做“词向量空间”，在词向量空间中，每一个词汇代表空间中的一个点。将字典和词汇转化成了空间和空间中的点的相对形式，空间中点与点的距离即可以用来表示字典中词汇与词汇间的关联程度。可以根据词汇间的距离判断词汇与词汇在语法、语义上的相似性。

词汇在空间上距离的衡量方法现在已有多种理论方法，比如使用欧氏距离（一种最传统的方法）来衡量；使用cos夹角衡量（两个词的词向量夹角余弦值越大，则意味着夹角越小，即等同于两个词汇在语义上越接近）等等。

另一方面，Distributed representation方法获得的词向量具有“可加性”，经过训练的词向量可以得到效果例如：“国王-男=女王-女”、“伦敦-英国=巴黎-伦敦”

总而言之，建立词向量空间可以通过建立词向量空间提取出词与词之间的深层语义关系。分布表示方法在语义和维度两个层面都填补了One-hot 的缺陷。

3. word2vec工具包

word2vec工具包是2013年Google公司发布的一个用于获取词向量的工具包，该工具包简单高效，可以极大提升获得使用Distributed representation方法获得词向量的训练速度。

Word2vec包包含两种可以用于获取词向量的模型，在算法实现上两种模型相似。两个模型都通过构建人工神经网络达成分类的目的。通过给每个单词都初始分配一个内部数值随机的多维向量，然后在训练中使用模型学习获得每个词汇的最佳向量。两种模型：其中一种是连续词袋模型（英文原名全名为Continunous Bag of Words，通常简写作CBOW），这个模型会根据上下文情况来预测当前词汇的概率。

第二种是Skip-Gram模型，Skip-Gram模型和前者正好背道而驰，Skip-Gram模型依靠当前词汇来推理前后文的可能情况。Skip-Gram模型在处理大规模数据集的情形下通常能获得更准确的结果。

这两个模型可以用不同的方法训练分别是等级sofmax方法和负面抽样方法。

sensim库的models包提供word2vec模块下定义了word2Vec类，该类最初训练算法移植自Google的word2vec C语言包，并拓展了功能。

1.5 文本分类方法

文本分类是数据分析领域内的一大重点任务，本课题也将使用语言文本分类方法，通过对对犯罪案件案情进行描述的中文文本进行分类从而实现经侦定性的目的。

对文本分类和对文本进行聚类不同，它们最大的区别在于是否标注训练文本。用于训练分类模型的训练样本需要带有类别标注。用于数据分类的机器学习属于有监督的学习，每个作为训练样本的数据对象都应带有其对应的数据标识。分类模型通常可以通过分类学习算法学习训练成一个针对特定任务的可用于分类的模型或函数（所得的这个分类用函数通常也被成为“分类器”）。分类器可以把数据项（也是文本数据中提取出来的属性）隐射到预先设定的可选类别中的某个类中，从而达成对测试数据进行分类的目的。

本质上，文本分类就是根据提取出的文本特征或文本属性，还有人工打上的文本类别标签，通过设计和训练构造分类器，并最终将训练所得的分类器去将不含类别标签的文本划分到已有的可选择的类别中。

经典的文本分类过程可以划分为以下几个步骤：对文本进行类型标注、对文本数据进行结构化处理、构造分类器和对分类效果进行分析评价。对文本进行类型标注通常是利用人工对一批文档进行分类，给文档标上类别标签；对文本数据进行结构化处理是为了是文本数据可以被计算机理解识别，并且可以提高训练效果，主要方法即上述提到的：文档-词项矩阵、词频逆向文档矩阵、构建词向量空间。构造分类器通常需要先对分类特征进行提取筛选，然后根据任务需求和可能的目标运行环境选择合适的构建方法和算法，分类器将会构建在训练集上。为了对分类效果进行分析评价，通常需要设置专门的验证集合，将会用从训练集上学习训练获得的分类器对验证集数据进行分类，通常验证集数据也是需要人工标记类别标签的，将可以利用分类器输出的预测的类别结果和真实的人工标记上的类别标签进行比对，进而达到对分类器分类效果进行评价的目的。常使用的评价标准数据项包括准确率、查全率和F1值等。

通常的分类方法都可以用于文本分类，当下常用的分类算法包括：朴素贝叶斯方法、线性判别分类器（LDA）、最邻近分类方法、支持向量机方法（Suppor Vector Machine，SVM）、神经网络、决策树方法（Decision Tree）等等。

Scikit-learn库能提供丰富的分类算法，包括以上提到的各分类器算法。

1. 朴素贝叶斯方法

贝叶斯公式最早由18世纪一位英格兰长老会牧师提出，贝叶斯公式的提出极大地推进了概率论和数理统计的发展，贝叶斯分类器是一种机器学习分类方法，该方法建立在贝叶斯公式的基础上，是有监督的学习算法。

贝叶斯公式：设c1,c2,c3,…,cn是样本空间中的划分，其中以P(ci)(其中i=1,2,3,…,n)表示事件发生的概率，且有P(ci)>0。则对于任意的时间x,P(x)>0,有：

贝叶斯分类中也包含多种分类，而朴素贝叶斯分类是贝叶斯分类中最简单的，虽然最简单但贝叶斯分类也有十分广泛的运用，常见的被用于一些针对业界基础文本的文本分类场景中。朴素贝叶斯分类方法的思想是：对于一个尚未分类的文本，我们可以依次求出文本再其对应特征条件下每个类别出现的概率，哪个类别出现概率最高，则认为该样本属于哪个类别。在文本数据分析中我们通常当做样本的类别特征可以和文本数据结构化处理后的文本特征项一一对应。

朴素贝叶斯分类方法的运用过程通常可以分为几步：

确定特征项的划分：使用之前预处理完成的训练集数据，确定类别划分c={y1,y2,y3,…,yn}和文本特征项x = {a1,a2,a3,…,an}；

训练分类器：计算训练集中所有特征项对应的先验概率P(a1), P(a2), P(a3),…, P(an)，并统计出在每个类别出现的条件下的各数据特征项目的条件概率P(a1|y1), P(a2|y2), P(a3|y3),…, P(an|yn)；

验证分类器：取用一个未参与分类器训练的新文本样本数据，并根据贝叶斯公式计算出该样本其文本特征项出现条件下每个类别的出现概率（这个概率也称为后验概率）,然后通过比较每个类别对应后验概率大小，得到新文本样本对应的类别y’。

但另外需要注意的是 ：朴素贝叶斯分类以一个重要假设为基础，即假设样本文本的特征彼此相互独立。但显而易见显示中文本内字词前后具有关联，而非完全相互独立，因而在应用朴素贝叶斯分类器进行分类时，应注意在第一阶段特别做好文本特征的提取工作，若能提取出相互独立的文本特征项，将能显著提高贝叶斯分类器的准确率。

2. 最邻近分类方法

最邻近分类方法又常被称为K最邻近分类（K-Nearest Neighbor Classification，简写为K-NN）。

最近邻分类算法也是一种较为简单的分类算法。核心原理，是通过对样本相邻样本属性的观察来，来判断最初样本所属的类别。最近邻分类算法通常按照以下步骤进行：

测量样本点间的距离：其中“距离”表示的是样本点之间相似的程度，即词向量空间中“距离”的含义，常用距离测量法有包括：欧氏距离法、马氏距离法、另外还有cos余弦距离法等等。当中对于文本分类任务，夹角余弦距离法较为合适。

确定邻近点：计算出新样本点与原有所有的样本点间的相似度后，将从所有原始样本点中找出一个或多个与新样本点最相似的样本点，这样找到的样本点可称为“邻近点”。邻接点个数通常用K来表示，通常K是一个可以人为设置的超参数，但要注意的是：小K可能使分类结果受噪点影响；大K又可能使领近的代表性被冲淡。通常K的值会低于训练样本数目的平方根。

获得类别：设置一定的规则，根据规则从邻居点的类别中获得样本类别。常用方法有通过加权或等权进行选择等等。

K邻近分类器的优点是原理简单易于实现，对参数的设置需求少，即对设计人员经验的需求更少，可以说它是“无需训练”的。但是它的缺点也是显著的，比如不适合分类类别中包含一些稀有类别的场景，也不适合运用在样本规模过大的样本数据集合上，计算量大、内存占用率高、效率低下。

3. 支持向量机方法

支持向量机方法简称SVM算法，是当今数学方法和最优化技术在机器学习和文本分类中的经典应用。SVM算法于1995年优俄罗斯统计学家Vldimir Vapnik领导的AT&T Ball研究实验小组首先提出，在90年代末，它在生物信息学、手写识别和文本识别等方面都有很多成功应用。

支持向量机方法认为属于不同类别的样本点间可能存在类别边界。支持向量机方法通过构建一个用于分割样本类别空间的超平面，使用规划思维对样本类别进行判断。

指　标
疑似剽窃文字表述

1. 贝叶斯分类中也包含多种分类，而朴素贝叶斯分类是贝叶斯分类中最简单的，虽然最简单但贝叶斯分类
2. 对应特征条件下每个类别出现的概率，哪个类别出现概率最高，则认为该样本属于哪个类别。

4. 第2章中文文本分类的关键技术_第2部分

总字数：2682

相似文献列表　文字复制比：1.4%(38)　疑似剽窃观点：(0)

1	基于证据理论的知识发现分类算法 李芳,韩元杰 -《桂林电子工业学院学报》- 2004-06-25	1.4%（38） 是否引证：否
2	基于多层B/S结构的全国计算机等级考试网上报名系统设计 张琳(导师：姜昱明;李波) -《西安电子科技大学博士论文》- 2009-04-10	1.1%（30） 是否引证：否
3	基于多层结构的物业服务系统的设计与实现 李玮瑶(导师：姜建国;吕海莲) -《西安电子科技大学博士论文》- 2010-12-09	1.1%（30） 是否引证：否
4	基于数据挖掘的网络学习评价 李盛瑜(导师：廖晓峰) -《重庆大学硕士论文》- 2008-04-01	1.1%（30） 是否引证：否
5	基于数据挖掘技术的智能信息处理 李敬有(导师：张昕) -《哈尔滨工程大学硕士论文》- 2007-05-01	1.1%（30） 是否引证：否
6	数字校园学生综合信息管理系统的设计与开发 杨臣(导师：杨宗凯;刘三（女牙）) -《华中师范大学硕士论文》- 2008-05-01	1.1%（30） 是否引证：否
7	决策树分类器在分析基因微阵列数据中的应用 项婧;任劼; -《计算机工程与设计》- 2006-08-16	1.1%（29） 是否引证：否
8	决策树算法在高校成绩分析中的应用 杨莅沅; -《潍坊学院学报》- 2008-07-15	1.1%（29） 是否引证：否
9	数据挖掘的决策树技术在高校毕业生管理中的应用 瞿花斌(导师：李学庆) -《山东大学博士论文》- 2014-06-30	1.1%（29） 是否引证：否
10	单元机组性能在线监测系统开发及电站运行数据的知识发现研究 李琳(导师：王培红) -《东南大学硕士论文》- 2005-03-01	1.1%（29） 是否引证：否
11	数据挖掘在初中学生管理中的应用 李国徽(导师：王莉) -《辽宁科技大学硕士论文》- 2007-04-26	1.1%（29） 是否引证：否
12	数据挖掘在考试系统中的应用 王永生(导师：张书杰;宋群) -《北京工业大学硕士论文》- 2004-04-29	1.1%（29） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

支持向量机由线性分类器、核函数和软间隔三个组成成分构成。

线性分类器以函数间隔>1作为条件，通过把样本点最大化到分类平面的几何间隔来实现最优化。

使用适当的核可以得到应用于维度数据空间的分类器。因为把低维度数据空间的向量集映射到高维度数据空间，可以使得更好划分难以划分的低维度向量集合，而核函数可以解决空间维数提高后计算复杂度增加的负面效应。

软间隔可以解决规则化和不可分情况，因为设置软间隔将允许一些词向量点在划分范围外，并且允许一些点在模型中又

同时不满足限制条件。

支持向量机算法往往将构建超平面，可以很好地解决低维空间中难以划分类别的情况，可以比其他分类算法获得更稳定的性能和效率。但是文本分类问题中，文本数据的特征向量大多维度高、较为稀疏的特点，且文本数据的特征向量特征间关联性较强，这使得文本分类任务中文本数据内噪音可能较大影响训练速度和分类准确度。

4. 决策树方法

决策树方法产生于上世纪70年代，在洛思坤（J.Ross Quinlan）提出ID3算法后决策树算法的正式成形。决策树算法的核心思想模型是是建立一套类似于树杈形状的逻辑策略，是一个非参数、同时又有监督的学习过程。

决策树算法以特定的树状数据结构构建分类器模型，其中树状数据结构可以是二叉树，也可以也是多分叉的树状结构。同时，决策树算法将对每个树杈上的判断条件都进行具体量化。一颗树形数据结构包括结点，分支和叶子。处于树状结构最上端的点被称作“根”，每个分叉点出衍生出的分支都可以衔接下一个包含分叉的结点或者叶子点。本算法中将用每一个数据结构结点代表一次决策判断，通常一个判断点将对应一个待分类对象属性，二分叉可代表这项属性在这个输入对象上是存在或不存在，而每个叶子都意味着一种可能的类型结果。

决策树将进行从根开始开始遍历，一般便利方向从上往下，遍历过程中每经过一个点就进行一次判断，通过在每个结点设置不同判断标准，不同判断结果导致不同走向，最后到达的叶子就等同于分析得出的输入对象的对应类别标签。

决策树主要在逼近离散函数值的方法上构建，常用算法又ID3、C4.5和CART等等。

用决策树算法构建分类模型，可以首先选出一个单独的店作为根结点，从根出发：根据节点上设置的判断标准对样本做出分析，如果对样本中所有元素的判断结果都是一样的，那么当前的这个点就是叶子，这个样本中所有元素可以算作同类型，它们共有的类型可以作为当前叶子的标记；如果输入样本中元素就此获得了不同的判断结果，则样本集合中的元素就被分为不同类型，当前节点就此衍生出不同分支。

结分类过程中，不同算法可能选择不同划分准则，但无论选择哪种算法，理论上算法会选择最易于分类且划分类型最具有代表性的属性项，这些属性将会用于指定当前点上的划分判断标准，目标是划分次数尽可能少，每次划分都能越快的是划分的集合类别纯度越高。

如果数据属性中有离散类型的属性决策判断标准将容易得多，决策树会用这些离散值和类别一一对应，有多少个离散值就对应多少个子节点，有多少取值就将生成多少分支；如果数据属性是连续类型的属性，算法往往会根据任务目的和划分准则在值域中选择分裂点，通过分裂点划分出子集。

决策树划分将重复取结点和划分这两个步骤多次。如果出现了以下情况：节点内所有样本都属于同一类;没有剩余属性可以进一步划分样本，通常没有剩余属性可以进一步划分样本时会通过一定方法把这些无法再划分的集合都归为一类，并且用样本中占主体的的类别作为类别标记。

是用决策树构建分类器时，通常还会根据实际应用情况对“树枝”进行修剪，剪枝可以处理数据中有噪点或者存在游离于主体外的点导致的过拟合问题。剪枝可以分成先后两种：先剪枝是指在构造树的过程中若有结点满足修剪条件，就直接恰掉这个满足条件点的后续划分活动；后剪枝是指在形成了完整的树形决策模型后，再设置一些条件，然后重新遍历一遍这个树形决策模型然后根据条件去除不要的分支走向。

决策树可以构建出有较高分类精度的分类器，并且分类模式容易理解，另外决策树对于有噪声的环境也表现出较好的稳健性。在文本分类任务中使用决策树可以减少训练过程中出现过拟合现象的概率。

5. 神经网络

神经网络算法也称为人工神经网络，全名Artifical Neural Network，通常缩写为ANN。如今在人工智能领域广受追捧的深度学习算法就是以神经网络为基础的。神经网络算法也是文本数据分析中常用的分类算法。神经网络思想起源于生物的神经网络运作原理。

图2-5 人工神经网络中的一个神经元

在人工神经网络中会由人工设置结点，并把最简单的人工结点称为神经元(neurons)，人工设置的神经元也像生物神经网络中的神经元一样成网形结构。其中f()是神经元的传递函数，t是神经元的输出。

算法中神经元活动的数学表示通常是计算输入向量和权重向量的内积，然后经历一个非线性传递函数，并最后获得一个标量结果。这种从输入向量到输出标量的映射通常被称为激励函数。

输入层是即将输入神经元的大量特征输入，输入特征通常被称作“输入向量”，在文本数据分析任务中，输入向量可以等同于文本数据样本的特征项。

隐藏层通常由多个相互连接的神经元组成，一个神经网络模型内部可以有多层隐藏层，且各隐藏层间的神经元个数不必须相同。对于隐藏层数和隐藏层内神经元个数这两超参数值的设置通常依赖于经验。一般来说，神经元个数越多的神经网络的非线性越显著，越增强神经网络的稳定性。但通常会将单个隐藏层内神经元个数这职位输入节点的1.2~1.5倍。

输出层将输出输出向量，输出向量的每一维度将对应输入数据对应各分类类别的可能性。

神经网络模型通常需要经过训练集训练，训练过程则实际是一个反复对网络中每一层和所有神经元的权重即偏置参数调整更新的过程。常用神经网络训练方法有梯度下降法等等。通常的训练方法需要设计人员人工置顶权重参数和偏置值更新的学习规则（Learn Rule）。在学习规则的要求下，训练函数将计算损失值，通常将更新各参数值使损失值尽可能的小，并最终实现模型拟合。而权重向量和偏置值的初始值通常可以初始化为0或随机数。

神经网络技术广泛运用于人工感知领域，而其在文本数据分析方面的应用在学术界也较受重视。独特的神经元权重结构将在对文本数据进行分析的同时体现对文本数据的提取过程。并且该技术通常会配合分类器评价技术一起使用。

本课题将使用的就是神经网络方法构建分类器，将构建并训练一个基于卷积神经网络的分类模型。

5. 第3章基于CNN的经侦案件文本分类模型

总字数：6021

相似文献列表　文字复制比：2%(122)　疑似剽窃观点：(0)

1	21105058429680114_王点_基于深度增强学习的五子棋人工智能实现 王点 -《大学生论文联合比对库》- 2017-05-20	1.2%（71） 是否引证：否
2	20130042512 -《大学生论文联合比对库》- 2017-04-25	0.8%（49） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第3章基于CNN的经侦案件文本分类模型

1.1 引言

本课题旨在设计并实现一个可以对经济犯罪案件自动进行定性的系统。当中能对犯罪案情进行描述的文本进行分类的分类器将是本课题的核心，将通过该分类器实现案情描述和经济犯罪案件犯罪性质的映射。拟使信用诈骗案件和合同诈骗罪两类经济犯罪案件的案件描述文本作为样本数据，进行训练，训练出一个可以对犯罪案情描述文本进行识别并按犯罪类型分类的分类器。将使用神经网络中的卷积神经网络模型进行分类器的搭建。

卷积神经网络是深度学习中重要的一部分，虽然卷积神经网络最早是针对解决图像识别任务而被提出的，但随着研究人员们的不懈努力，卷积神经网络模型在文本数据分析方面的运用也日益成熟。图像识别任务和文本数据分析任务最主要的不同在于输入的数据格式不同，图像数据可表示为二维数组且是结构化的数据，而文本是一维数据且非结构。但是对文本进行结构化处理后，文本数据也可以转化为数字组成的结构化数据，可以作为卷积神经网络的输入。

1.2 建立语料库

本课题将通过分类器实现案情描述和经济犯罪案件犯罪性质的映射，旨在设计并实现一个可以对经济犯罪案件自动进行定性的系统。拟使用信用诈骗案件和合同诈骗罪两类经济犯罪案件的案件描述文本作为样本数据。各法院的官方网站上往往有案件判决信息公示，包含认为定性的文本标签和犯罪案件案情描述文本，将使用爬虫工具从官网上爬取数据，获得数据后将经过一系列人工筛选和整理，使用信用卡诈骗案件的文本作为正样本，而合同诈骗罪为负样本。以一件案件的描述文本作为一段，并将所有文本数据划分为三个集合，三个集合分别是训练集、验证集和测试集。

当中训练集和验证集中的数据是犯罪案件类别标签加上案情描述文本一起为一段，当中标签和文本本体以tab作为分隔。训练集是将用于训练模型的数据集合，验证集则是用于在训练过程测试训练过程中准确率、验证训练情况的数据集。测试集不参与训练，测试集以单纯的犯罪案件描述文本为一段，当中不含类别标签，测试集将用于对最终获得模型的分类准确率进行测试，测试结果将用于系统性能的分析评价。

经过整理其中训练集训练集含数据1600条，当中正负样本各800条；验证集一共200条，当中正负样本各100条；测试集一共200条，当中也正负样本各100条。

语料库建成后将通过程序对输入语料做预处理，使训练用的语料结构化。

为达成数据数据结构化的目的，设置了专门的文件case_loader.py。该用于对数据进行结构化等处理，当中包含多个功能函数：

使用read_file()函数从“case/case.train.txt” “case/case.val.txt” “case/case.test.txt”中获得要处理的数据，读取数据的时候将对数据进行一些与处理，包括获取数据类别标签、去除多余的标点符号等等。

使用buld_vocab()函数构建词汇表，本课题使用的是字符集表示，故不进行专门分词，直接将文本数据以中文文字为单位分为一个一个的独立字符存入字符库。并且该函数内将通过添加pad使所有文本pad保持相同长度。由于本课题的使用的样本数据数据量有限，故设置词典的最大长度为5000。

每次训练或进行测试时都使用同一张词汇表，该词汇表将在训练模型时就建立，且建立的词汇表将会保存在目录“case/case.vocabe.txt”中，减少了重复处理，但是下一次进行训练或每次重新进行模型训练的时候，该词汇表都会重新建立，旧词汇表将会被覆盖。另外需要注意的是：测试集的数据往往和训练集和验证集中包含的词向量有出入，假如测试集中出现了非停用词的新词汇，只要它不是训练集生成的词典中的词，需会都将其抛弃，保证所有集合使用的都是同一个词向量空间。

read_vocab()函数：将从buld_vocab()输出的“case/case.vocabe.txt”中读取词汇表，并且将词汇表转换成{词:id}的表示形式，所转化的id为每个中文文字符号对应的unicode码。

read_category()函数：从训练集和验证集中读取预先人工标记好的犯罪类型标记，固定分类目录，并且把类别标签也转化成一一对应的id表示。

to_words()函数：用于进行id和文字间的转换，输入id可以返回id对应的文字，主要用于在输出显示分配的类别标签。

process_file()函数：给数据集合中的每段数据，即每一起案件标记序列id,每个id和案件一一对应。

batch_iter()函数：用于为神经网络按设定的批次大小，按批次提供数据。

经过case_loader.py文件处理非结构化的中文文本数据将被转化为结构化的数据，模型函数将通过调用batch_iter()函数获得处理好的数据，投入训练或者测试活动。

经过数据预处理，数据的格式如下：

表3-2 预处理后的数据结构

Data Shape Data Shape

X_tarin [1600,600] Y_train [1600,2]

X_val [200,600] Y_val [200,2]

X_test [200,600] Y_test [200,2]

1.3 搭建基于CNN的神经网络模型

本课题将通过搭建卷积神经网络模型构造分类器。神经网络的特性是程序可以通过“自学”来接近实验目的。卷积神经网络是在神经网络基础上的进一步发展，增加了卷积和池化的概念，在文本分类任务中卷积神经网络的优点是拥有较快的运算速度，且可以通过GPU进行计算。

结构上卷积神经网络和基础神经网络不同的地方在于，卷积神经网络中多添加了卷积层和池化层。

一个卷积神经网络可能由许多层卷积层和池化层组成，数据通过输入层被读入神经网络中后，将经过几次阶段性特征提取，通常每个阶段都会包含有至少一层卷积层和至少一层池化层组合，这两个层可以交错轮流出现，它们将共同对数据特征进行反复的精炼，每个阶段的池化层对数据进行池化操作也能降低数据维度。

本课题中搭建的卷积神经网络结构如图：

图3-3 卷积神经网络结构

包括Embedding层一层，CNN层和Polling层各一层，以及两层全连接层。其中Pooling层将选用Max Pooling方法。

输入输入层的数据是经过结构化处理后的文本数据，所有词有相同的词向量长度，设定词向量的长为64，经过结构化处理后的文本数据是形如：[所含案件样本条数，序列长度]的二维矩阵。

因为考虑到不同的案件描述文本结构化处理后所携带的特征个数（分词后的的词数）不同，所以通常会把矩阵的宽度统一为最长的文本的特征向量个数。本卷积神经网络模型中，序列长度人为设定为600，即设定每个案件由600个特征向量描述，不足时都用0补齐，保证所有的样本项目序列长度都相同。

1.3.1 卷积层

卷积是一种可以简写为：s(t)=(x*w)(t) 的函数操作，其中x是输入的矩阵形式的结构化数据，w是神经网络中常用的权重向量，在卷积神经网络中也可以称作“卷积核”，输出s是特征向量，通常在变量都是二维矩阵的情况话，K中应给越接近当前节点的点（词汇）分配越大的权值，且因为K的矩阵大小通常小于输入的矩阵I的大小，所以为了运算效率和内存占用率的考虑，往往先遍历K中元素再寻找对应的I[i-m,j-m]，卷积函数通常也写成如下形式：

其中I是输入矩阵，K是权重矩阵。

使用卷积是为了提取输入的矩阵形数据的特点。卷积可以通过从输入矩阵内的部分数据内学到输入样本的特点，同时，卷积可以在提取特征的同时可以保留元素间的空间关系。

权重矩阵K作为特征检测工具对每个输入的矩阵进行处理操作。

卷积神经网络中使用卷积替代一般的矩阵乘法。卷积层中，卷积核携带一组固定权重，形如一个滑动窗口，一轮窗口滑动将获得多个卷积层输出，卷积神经网络将一层卷积结果作为一个输出层。

不同卷积核在同一个输入矩阵上进行卷积将获得不同的特征图。卷积活动中中“深度”是指卷积核的个数，等同于输出的特征图个数也等同于厚度，每个卷积核将有一个输出。“步长”是在输入矩阵上滑动卷积核一次滑过的元素数，当步长为 n时，每次就移动卷积核滑过n个元素的位置，设置越大步长将获得越小的特征图。

在对模型进行训练前需要人为指定核的个数、核的大小、网络架构等，CNN 会在训练过程中通过学习调整权重矩阵中的值。使用的卷积核个数越多，提取到的图像特征就越多，网络所能在未数据上识别的效果将越好。本模型中设置卷积核个数为128个，卷积核大小为5。

卷积神经网络具有稀疏关联的特性，它使得输出可以一次受更少的输入点影响，但随着网络层数不断增加，一个输出点将受更多输入结点影响。相比之下矩阵乘法将使得每一个输出都会同时受所有输入点的影响，哪怕只有一层网络，这会带来更大的计算量。

同时卷积还有参数共享、权数共享的特点，可以缩短运算时间提高运算性能，并且这样可以使得需要学习的参数个数更少。另外，卷积层还有等变性质，每当输入发生改变时，输出也会以相同方式个改变。

1.3.2 池化层

卷积神经网络中池化层通常和卷积层一起使用，当中池化函数被用来修正输出。通常两者一起完成特征提取任务。

空间池化可以保留大多数重要特点的同时降低特征图维数，池化函数通常会用一些运算方法计算各个区域区域内位置接近的输出，然后用所得计算值代表这一整块区域。常有池化函数有两种，包括最大池和平均池，本课题使用的是最大池。最大池方法返回各个区域内部最大的那个输出值代表这块区域的输出。使用最大池，可以从上一层卷积后获得的特征图中提取出对文本语义贡献最大的部分。这样操作可以降低我们特征图的维度。其他的计算方法还有L2正则方法，计算基于距离矩形中心

距离的加权均值方法等等。

池化层除了降低纬度外还可以增强模型鲁棒性，缓解过拟合。

1.3.3 全连接层

全连接层即经典神经网络中的多层感知器。

“全连接”意味着该层使得全面层的全部神经元都将和下一层的所有神经元相连。前文所提到的设置在全连接层之前的卷积层和池化层以及提取了输入矩阵型数据的高级特征，这些高级特征将作为全连接层的输入，在全连接层阶段将利用这些特征完成对输入数据的分类。

本模型中在池化层后使用了两层全连接层。设置每层全连接层都含有128个全连接层神经元，且学习率设置为1e-3。

设置每批训练大小为64，全连接层将分批对输入的特征数据进行遍历，每批次计算都将经过dropout方法处理，并且用relu激活函数激活。函数将通过矩阵乘法生成预测值，预测值最终将通过softmax激活函数被转化成标准数值。

其中dropout是指通过随机禁用神经网络中的部分神经元，这种禁用是暂时的，目的是为了使每个mini-batch都训练不同网络，这样做可以预防神经元出现共通适应的情况，保证神经元学习的特征都是单独有用的。dropout方法可以提升训练速度并缓解过拟合现象。其中dropout保留比率设置为0.5。

训练中把对所有批次遍历一遍叫做一轮，设置最高将循环10轮，其中每循环10批次将输出一次结果，并保存一次卷积神经网络模型，所得卷积神经网络模型将保存在目录“/tensorboard/textcnn”中。如果多批次训练准确率都没有变化将在不足10轮时就提前结束遍历。

1.4 搭建基于CNN的经侦案件文本分类系统

本课题中使用cnn_model.py定义卷积神经网络模型，使用case_loader.py对输入的犯罪案件描述中文文本进行清洗和结构化处理。并通过带参数的运行run_cnn.py文件来对卷积神经网络模型进行训练或使用训练后的模型对测试集进行分类。

read_vocab()函数：将从buld_vocab()输出的“case/case.vocabe.txt”中读取词汇表，并且将词汇表转换成{词:id}的表示形式，所转化的id为每个中文文字符号对应的unicode码。

read_category()函数：从训练集和验证集中读取预先人工标记好的犯罪类型标记，固定分类目录，并且把类别标签也转化成一一对应的id表示。

to_words()函数：用于进行id和文字间的转换，输入id可以返回id对应的文字，主要用于在输出显示分配的类别标签。

process_file()函数：给数据集合中的每段数据，即每一起案件标记序列id,每个id和案件一一对应。

batch_iter()函数：用于为神经网络按设定的批次大小，按批次提供数据。

经过case_loader.py文件处理非结构化的中文文本数据将被转化为结构化的数据，模型函数将通过调用batch_iter()函数获得处理好的数据，投入训练或者测试活动。

当中在命令行输入指令“python run_cnn.py train ”将对卷积神经网络模型进行训练。

输入指令“python run_cnn.py test”将利用之前训练的模型对case.test.txt文件内的犯罪案情描述文本进行分类。

训练所得的模型将保存在目录“/tensorboard/textcnn”下，对文件分类的分类结果将通过命令行输出，且会保存在文件目录data/case下，名为“输出时间+ case_result.txt”通过时间来标记分类结果对应的原文件。

输入通过将先经过人工整理，然后命名为“case.test.txt”存放在目录“/data/case”下。

1.4.1 训练模型

通过命令行输入：“python run_cnn.py train ”，训练卷积神经网络模型，训练流程如图所示。

图3-4 卷积神经网络模型训练过程(上)

图3-4 卷积神经网络模型训练过程(下)

1.4.2 实验结果与分析

将使用正样本800条，负样本800条，总计1600条训练样本投入模型进行训练，并使用正负各100条样本一起组成的验证集用于在训练过程中检测训练模型的准确率。

在对其训练成果进行评价时选用了计算准确率并求准确率平均值的方法。

其中准确率随着训练轮数而变化的变化曲线如图。

图3-4 卷积神经网络模型训练过程中的准确率和损失值

根据变化曲线可得：本模型的超参数设置使训练获得了较好效果，损失值在前400个批次的运算中有显著的降低，同时在前200个批次运算中准确率有显著提升，两百个批次运算后准确率接近百分之九十，之后准确率的提升逐渐趋于平缓，直到运行结束为止，准确率基本在百分之九十五和百分之百之前浮动。训练过程中有过拟合现象的出现，但是过拟合的程度比较低，这个现象对训练的影响较弱，没有干扰模型的分类效果。

指　标
疑似剽窃文字表述
1.　步长”是在输入矩阵上滑动卷积核一次滑过的元素数，当步长为 n时，每次就移动卷积核滑过n个元素的位置，

第4章使用基于CNN的文本分类模型对案件文本分类

1.1 使用基于CNN的文本分类模型对案件文本分类

本系统使用卷积神经网络模型搭建分类器。将通过分类器实现案情描述和经济犯罪案件犯罪性质的映射，旨在设计并实现一个可以对经济犯罪案件自动进行定性的系统。拟使用信用诈骗案件和合同诈骗罪两类经济犯罪案件的案件描述文本作为样本数据。各法院的官方网站上往往有案件判决信息公示，包含认为定性的文本标签和犯罪案件案情描述文本，将使用爬虫工具从官网上爬取数据，获得数据后将经过一系列人工筛选和整理，使用信用卡诈骗案件的文本作为正样本，而合同诈骗罪为负样本。以一件案件的描述文本作为一段，并将所有文本数据划分为三个集合，三个集合分别是训练集、验证集和测试集。

当中训练集和验证集中的数据是犯罪案件类别标签加上案情描述文本一起为一段，当中标签和文本本体以tab作为分隔。训练集是将用于训练模型的数据集合，验证集则是用于在训练过程测试训练过程中准确率、验证训练情况的数据集。测试集不参与训练，测试集以单纯的犯罪案件描述文本为一段，当中不含类别标签，测试集将用于对最终获得模型的分类准确率进行测试，测试结果将用于系统性能的分析评价。

数据预处理过程中分词阶段使用字符集的分词粒度，按单个中文字符对汉语言文本数据进行分词构建尺寸为5000字大小的辞典，然后将汉字转化为unicode编码形式完成对非结构化数据的结构化处理。

分类器部分本课题中使用卷积神经网络模型搭建，其中卷积神经网络的结构包括包括Embedding层一层，CNN层和Polling层各一层，以及两层全连接层。其中Pooling层将选用Max Pooling方法。分类模型定义实现在文件cnn_model.py中，使用case_loader.py对输入的犯罪案件描述中文文本进行清洗和结构化处理。输入输入层的数据是经过结构化处理后的文本数据，所有词有相同的词向量长度，设定词向量的长为64，经过结构化处理后的文本数据是形如：[所含案件样本条数，序列长度]的二维矩阵。本卷积神经网络模型中，序列长度人为设定为600，即设定每个案件由600个特征向量描述，不足时都用0补齐，保证所有的样本项目序列长度都相同。本模型中设置卷积核个数为128个，卷积核大小为5。

输入通过将先经过人工整理，然后命名为“case.test.txt”存放在目录“/data/case”下。准备好将用于分类的数据后将将通过带参数的运行run_cnn.py文件来对卷积神经网络模型进行训练或使用训练后的模型对测试集进行分类。

在对测试集进行测试之前，要首先对模型进行训练。当中在命令行输入指令“python run_cnn.py train ”将对卷积神经网络模型进行训练。

本模型中在池化层后使用了两层全连接层。设置每层全连接层都含有128个全连接层神经元，且学习率设置为1e-3。

设置每批训练大小为64，全连接层将分批对输入的特征数据进行遍历，每批次计算都将经过dropout方法处理，并且用relu激活函数激活。函数将通过矩阵乘法生成预测值，预测值最终将通过softmax激活函数被转化成标准数值。

其中dropout是指通过随机禁用神经网络中的部分神经元，这种禁用是暂时的，目的是为了使每个mini-batch都训练不同网络，这样做可以预防神经元出现共通适应的情况，保证神经元学习的特征都是单独有用的。dropout方法可以提升训练速度并缓解过拟合现象。其中dropout保留比率设置为0.5。

训练中把对所有批次遍历一遍叫做一轮，设置最高将循环10轮，其中每循环10批次将输出一次结果，并保存一次卷积神经网络模型，所得卷积神经网络模型将保存在目录“/tensorboard/textcnn”中。如果多批次训练准确率都没有变化将在不足10轮时就提前结束遍历。

模型训练完成后将可以对测试集进行分类测试。

输入指令“python run_cnn.py test”将利用之前训练的模型对case.test.txt文件内的犯罪案情描述文本进行分类。

使用模型进行分类数，数据同样要依次进过分词、构建字典、数值转化等结构化操作，结构化数据输入分类器后同样依次通过CNN层和Polling层各一层、全连接层两层并通过激活函数，不过在测试分类的过程中数值变化的传播是单向的，获得分类结果后将直接输出，不再会计算损失值和参数变化率，也不再会改变模型参数。

训练所得的模型将保存在目录“/tensorboard/textcnn”下，对文件分类的分类结果将通过命令行输出，且会保存在文件目录data/case下，名为“输出时间+ case_result.txt”通过时间来标记分类结果对应的原文件。

程序运行过程如图。

图4-1 分类模型对案件分类(上)

图4-1 分类模型对案件分类(上)

测试完成后将输出检测结果，并且会将测试结果以txt文档形式保存在/data/case/目录下，文档名称形如“2018.05.20_20.53.55case.result.txt”其中日期和时间是获得当次检测结果的时间。

图4-1 分类检测结果的输出

图4-1 分类检测结果的输出

1.2 实验结果和分析

用于测试的测试样本为：正样本100对，负样本100对，一共200对训练样本。为便于分类结果的测评和分析，在系统搭建的同时就包含了对测试集也进行准确率分析的步骤。在对其测试成果进行评价时选用了计算准确率并求准确率平均值的方法

，在分类测试结果输出时同时也对测试分类的准确率进行统计比对，对测试结果准确率的分析统计如图。

图4-2 分类模型对案件分类准确率及结果分析

在多次的训练并测试尝试中，准确率大多在百分之九十五到百分之一百之间浮动。

根据对神经网络学习训练过程的可视化过程可知，损失值在前400个批次的运算中有显著的降低，同时在前200个批次运算中准确率有显著提升，两百个批次运算后准确率接近百分之九十，之后准确率的提升逐渐趋于平缓，直到运行结束为止，准确率基本在百分之九十五和百分之百之前浮动。训练过程中有过拟合现象的出现，但是过拟合的程度比较低，这个现象对训练的影响较弱，没有干扰模型的分类效果。dropout方法可以提升训练速度并缓解过拟合现象。本课题中dropout保留比率设置为0.5，若要再更进一步减轻过过拟合现象还可对去他的dropout保留比率进行选择尝试。

在多次对数据进行训练测试，并用额外的测试集进行测试的过程中，模型也曾获得不同的结果，分类结果如图。

图4-2 分类模型对案件分类准确率及结果分析

获得百分之百正确分类结果是偶然事件，普遍只能获得较高准确率的分类结果。获得本次运行结果测试获得百分之百的准确率的原因，概因为本次实验用于测试的测试集数据集合较小，两类测试数据一共只有200条，虽然相比于用于训练和验证的的共计1800条数据，测试集占大约九分之一，但从数据数量总量上看还是比较小，所以导致了出现百分之百这样的低概率事件。

若增大测试集合大小，系统测试分类结果会更接近真实数值，但真实数值也在一个较好分类效果的区间内。由CNN模型的训练过程曲线可知本系统的文本分类准确率主要在百分之九十五以上。本系统可以达到对经济犯罪案件描述文本进行定性分类的目的。

如果要在当前基础上进一步提升准确率，在数据预处理方面：可以增加对中文语言文本数据的提前分词操作，去除停用词等意义不大的词汇之后再使用字符集的分词搭建字典；其次可以换用其他其他构建词向量空间的方法，比如利用词频-逆向文档矩阵实现字典中字符和数字的映射，又或是换用word2vec工具包工具包，通过调用连续词袋模型来构建词向量空间。在分类器的构建方面，可以增加卷积神经网络模型当中CNN层和Polling层的层数，使用两组甚至三组的CNN层和Polling层；或可以设置多种核的大小，比如每层CNN层中设置三到四个大小分别依次为2345的卷积核，获得多张特征图，然后结合多层卷基层搭建神经网络分类器。

有更大量的可用于训练的数据的话神经网络分类器也可以获得更好的分类效果。本系统在实际运用中的话也可以不断更新训练集合，从而通过累计获得更多大的训练集，提升该神经网络分类器的准确度。

13	博山云行山道教与地方社会	9.6%（121）
	宋建明(导师：赵卫东) -《山东师范大学博士论文》- 2016	是否引证：否
14	我国最早的阶级社会	9.1%（115）
	何艺培(导师：杜斗成;郭永利) -《兰州大学博士论文》- 2016	是否引证：否
15	1101170201-何兆鑫-刺参养殖池塘生态环境典型重金属含量的分析	9.1%（115）
	何兆鑫 -《大学生论文联合比对库》- 2015	是否引证：否
16	电信_11424057_师清_数字语音识别系统的实现	8.6%（109）
	电信 -《大学生论文联合比对库》- 2015	是否引证：否
17	初中生羞怯、压力知觉与心理健康之间的关系	8.2%（103）
	于华颖(导师：陈英敏) -《山东师范大学博士论文》- 2016	是否引证：否
18	抽象雕塑中的哲学	8.2%（103）
	王彦文(导师：刘大顺) -《沈阳大学博士论文》- 2017	是否引证：否
19	240897613_丘莉琴_浅析日本公益广告创意及启示	7.8%（98）
	丘莉琴 -《大学生论文联合比对库》- 2012	是否引证：否
20	邓雅萍-10153129-经济学	7.8%（98）
	邓雅萍 -《大学生论文联合比对库》- 2014	是否引证：否
21	营销082班抽检毕业论文 冯佳翔	7.7%（97）
	-《大学生论文联合比对库》- 2012	是否引证：否
22	应用文理学院_杨梦迪_试论新版电视剧《红楼梦》的成与败	7.7%（97）
	-《大学生论文联合比对库》- 2013	是否引证：否
23	防酒驾控制系统设计	6.7%（84）
	郭力 -《大学生论文联合比对库》- 2013	是否引证：否
24	2008110823-刘献美-数据挖掘技术在超市销售中的应用研究	6.3%（79）
	刘献美 -《大学生论文联合比对库》- 2012	是否引证：否
25	研究生网络自主学习行为现状调查及对策研究	5.7%（72）
	马晶晶(导师：张萍) -《陕西师范大学硕士论文》- 2009	是否引证：否
26	罗婷婷	4.0%（50）
	罗婷婷 -《大学生论文联合比对库》- 2013	是否引证：否
27	PVC聚合间歇过程控制	2.5%（32）
	荆胜南(导师：杨小健) -《南京工业大学硕士论文》- 2004	是否引证：否
28	10kV开关站设备与运行	2.3%（29）
	蒋丽娟(导师：周浩;余虹云) -《浙江大学博士论文》- 2011	是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第5章总结与展望

5.1 总结

但人工智能可以通过算法实现程序的自主学习并进行决策，尤其在解决知识密集型问题上有很好的效果，虽然在复杂的经济犯罪新形势面前人们仍然面临着许多难点，但是当前时期的经侦工作虽然已经有了显著发展，定量法学研究也是当前社会科学发展得意一个必然趋势。所以将人工智能技术用于法学方面的案件侦查定性的研究确实可行，并且正顺应了时代发展的潮流。虽然本实验因可用样本数据种类和数目都有限，只能实现对两类犯罪案件进行分类的任务，但是如果要对更丰富的犯罪类型进行定性，只需要对各超参数进行适当的修改调整，将使用的模型仍然是相通的。所以实验结果表明，CNN神经网络模型对犯罪描述文本有较好的分类效果，可以用于经济侦查案件的定性。

5.2 展望

现在对经济犯罪案件的定性普遍都是由人来人工进行的，需要消耗一定的人力成本和经费成本，且案件处理效率不高。使用人工智能技术搭建经济犯罪案件侦查系统有利于增强调查者的能力，可以有效缩短经济犯罪侦查程序的简约进路，可以极大的提高案件处理效率。而且神经网络技术赋予了计算机系统优秀的更新学习能力，可以便利的随着现实发展而及时快速的进行更新。最后犯罪调查的自动化和支持还可以有效削减人力成本和经费成本，使潜在的节约成为可能。甚至还可以为经验不足的调查人员提供培训以及规范调查程序，从而，并且使各种调查人员都可以处理同一起案件。

将人工智能运用于犯罪案件定性是对法律理论、文本、案例增加趋势的顺应，也是对国家大力发展政府机构信息化号召的积极响应，将人工智能技术用于法学方面的案件侦查定性的研究和应用是时代发展下的必然产物，且未来必将有更好的实现和更普遍的应用，发展潜力无限。

参考文献

[1] Petter Gottschalk,Geoff Dean. Stages of knowledge management systems in policing financial crime[J]. International Journal of Law, Crime and Justice,2010,38(3).

[2] James E. Bowen.An expert system for police investigators of economic crimes,Expert Systems with

Application.Volume 7, Issue 2,1994,Pages 235-248,ISSN 0957-4174,https://doi.org/10.1016/0957-4174(94)90040-X.(http://www.sciencedirect.com/science/article/pii/095741749490040X)
[3] 张妮,杨亘.量化法学及人工智能在民商法学中的应用[J].民商法争鸣,2015(00):13-20.
[4] 唐稷尧.中国当前经济犯罪的界定和分析[J].四川师范大学学报(社会科学版),2000(01):16-22.
[5] 杨维林. 经济犯罪的法律规制[D].吉林大学,2012.
[6] 房军,张宝瑞,王全.谈经济犯罪案件的分析方法[J].辽宁警专学报,2005(06):22-25.
[7] 程小白.论经济犯罪案件的侦查学分类与构成[J].江西公安专科学校学报,2009(01):10-14.
[8] 张栩华.浅谈经侦工作中的定性难点及对策[J].法制博览,2016(05):128-129.
[9] 王俊家.经济犯罪侦查中的难点问题研究[J].中国人民公安大学学报(社会科学版),2008(04):115-118.
[10]刘红岩,陈剑,陈国青.数据挖掘中的数据分类算法综述[J].清华大学学报(自然科学版),2002(06):727-730.
[11]刘莹,王宁,李保华,罗强.模糊语法方法在犯罪文本分类中的应用[J].计算机工程与设计,2017,38(07):1965-1971.
[12]苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报,2006(09):1848-1859.
[13]陶林润德.机器学习方法在文本分类中的应用[J].中国战略新兴产业,2017(40):134-135.
[14]高菲. 基于机器学习的计算机辅助量刑初探[D].华东政法学院,2005.
[15]李旭东. 公安部门警情研判系统设计与实现[D].辽宁科技大学,2015.
[16]崔剑非. 面向刑侦应用的数据挖掘问题研究[D].国防科学技术大学,2008.
[17]陈巍.基于数据挖掘的刑事犯罪侦查系统研究[J].山西警官高等专科学校学报,2010,18(04):71-75.

致谢

四年的大学生涯即将结束，而于我的人生却只是一个逗号，我将面对又一次征程的开始。四年的求学生涯在师长、亲友的大力支持下，有遗憾也同时收获满囊，这些经历必将成为我人生中弥足珍贵的记忆。在论文即将付梓之际，要特别感谢求学过程给予我无限支持和帮助的老师、朋友和亲人们。

尤其是要感谢我的指导老师，邹淑雪老师。在毕业设计这个大学最后的学习阶段里，从最初的定题到资料收集，从课题项目的研究到搭建，再到最后论文写作、修改到论文定稿，这段历时半年的时光里，邹老师全程给予了我耐心的指导和无私的帮助，给我鼓励和动力，既为我打开了人工智能这个更进一步的专业方向的大门，又为我在课题研究与文献写作方面提出了许多富有建设性的有益意见。也正是在她的指导和督促下论文才得以如期完成，无论是日常的学习，还是课题研究，甚至未来发展规划，邹老师都给予我悉心的关怀与细心的指导，在此谨向邹老师致以诚挚的谢意和崇高的敬意。

同时还要感谢所有任课老师这四年来给自己的指导和帮助，你们无私的奉献精神和爱岗敬业的治学态度，使我受益匪浅。感谢我们一起在学校努力的同学，我们彼此关心、互相支持和帮助，留下了许多难忘的回忆。

最后再一次感谢所有在毕业设计中曾经帮助过我的良师益友和同学，以及在设计中被我引用或参考的论著的作者。