Computer Science and Application 计算机科学与应用, 2019, 9(12), 2191-2207
Published Online December 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.912245

Hans汉斯

Research and Implementation of Character Analysis Algorithm Based on Text Information

Yi Kong

Beijing University of Posts and Telecommunications, Beijing
Email: 18500096915@163.com

Received: Nov. 11th, 2019; accepted: Nov. 22nd, 2019; published: Nov. 29th, 2019

Abstract

Based on the text information, this research focuses on the static text to obtain the characters' characters in novels, scripts and other literary works. Using the novel "the ordinary world" as the sample analysis training model, combined with the big five personality algorithm in psychology, it mainly adopts the combination of neural network and traditional machine learning. By comparing the model effects of the two models doc2Vec and word2Vec + CNN, it is found that the former has a better performance when predicting the character of an unknown person. Therefore, the model makes the idea of intelligent text analysis possible, and the prediction score can be mapped out the characters' character vocabulary through the big five personality scale, so that the future machine can "read" the semantics, and the user's portrait, intelligent machine and psychological development is of great significance.

Keywords

Big Five Personality, Natural Language Processing, Machine Learning, Character Prediction

基于文本信息的人物性格分析算法的研究与实现

孔　仪

北京邮电大学，北京
Email: 18500096915@163.com

收稿日期：2019年11月11日；录用日期：2019年11月22日；发布日期：2019年11月29日

文章引用: 孔仪. 基于文本信息的人物性格分析算法的研究与实现[J]. 计算机科学与应用, 2019, 9(12): 2191-2207.
DOI: 10.12677/csa.2019.912245

摘　要

本课题研究基于文本信息分析人物性格，聚焦于静态文本来获取小说、剧本等文学作品中的人物性格，使用《平凡的世界》这一小说著作作为样本分析训练模型，结合心理学中的大五人格算法，主要采用神经网络与传统机器学习相结合的方式，通过对比doc2Vec和word2Vec + CNN两个模型的模型效果发现在预测未知人物性格时前者有着更好的表现，因此，该模型将智能分析文本这一想法变为可能，并且可通过大五人格量表将预测得分映射出人物的性格词汇，使得未来机器能够“读懂”语义，对用户画像、智能机器以及心理学的发展具有重要意义。

关键词

大五人格，自然语言处理，机器学习，性格预测

1. 背景

1.1. 自然语言处理

作为近年来最火热的一个研究方向，自然语言处理在自然科学乃至社会科学等多个领域都发挥着其极其重要的作用。自然语言处理作为计算机科学与人工智能领域相结合的产物，其主要研究人类语言与计算机之间的通信理论和通信方法。由于其涉及到统计学、概率论等数学领域，同时为了研究人类语言中的语言特性，所以自然语言处理也包含了语言学、社会学等社会学科，此外，为了将自然语言处理算法得以实现，需要应用相关编程语言和程序设计，总之，自然语言处理是一门包含数学科学、社会语言学和计算机等多门学科于一体的综合性学科交叉产物，这也正解释了为什么自然语言处理能在当今人工智能时代大放异彩。

经过近几年的发展，自然语言处理技术发展出很多技术路线，见图 1，比如分词、文档分析、机器翻译、词性标注、命名实体识别、文本语义相似度分析、句法成分分析、词向量语义相似度分析等。同时，依据这些技术，自然语言处理有着多种多样的研究方向，比如信息检索、文档分类、自动文摘、语音识别、舆情分析等。这些技术都慢慢地在我们所处的社会生活中得以应用。比如，在现在的智能手机上大多会有各大厂牌自主研发的语音助手，通过语音识别、语义分析以及其他自然语言处理技术，可以让手机快速分析用户所说的话中的语法结构，从而获取关键信息，使得系统得以理解用户所发出的命令从而进行响应。目前市面上流行的天猫精灵等人工智能音箱均是以自然语言处理为技术基础所研发的，从长远角度来看，自然语言处理是人工智能领域的重要基石，是人与机器“对话”或者信息交互的重要工具，通俗点说，要想让机器听懂“人话”说“人话”，自然语言处理技术必不可少。

同时，自然语言处理也存在着许多瓶颈，首先，目前自然语言处理大多分析的是单纯的某一段文本或者某一个句子，这样很容易造成信息缺失。就从简单的机器翻译来说，一句话的语义并不只是由文本所决定的，通常也会由很多其他信息所决定。比如把一篇文章里面的某一句话截取出来做翻译，该句话里的人称代词、物主代词在没有上下文辅助的情况下是无法被机器所识别的，这样就会造成语义模糊或

者歧义的问题，因此，自然语言处理目前仍然受到输入数据的信息维度限制。其次，由于很多国家的语言博大精深，比如我们汉语，一词多义、一义多词的情况十分常见，简单的一句“你就等着吧”，就可能有两种截然不同的意思，同时语法结构也多种多样，比如“中国队大胜美国队”和“中国队大败美国队”表示的都是“中国队赢了”的含义，所以语言学的复杂性决定了自然语言处理有相当多的难点问题需要解决，汉语如此，其他国家的语言也面临着同样的问题。因为相比于自然语言处理技术，语言的诞生有几万年的历史，如果需要把人类语言完全应用于计算机科学，让机器像人类大脑一样思考问题，那么就需要把语言本身的起源研究透彻，究其原理，然而这是目前语言学家所面临的最大难题之一。

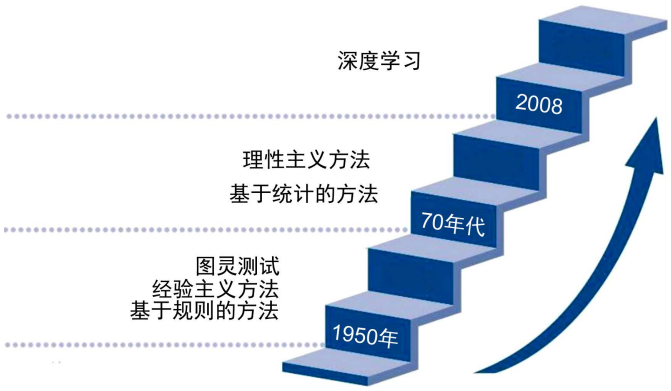

Figure 1. Development of natural language processing technology
图 1. 自然语言处理技术发展现状

在自然语言处理大力发展的情况下，词向量[1] [2]以及文档向量[3]的表示有助于文档向量化表示，同时在以往的研究中，基于词向量[4]或文档向量[5]的用户画像或用户标签分类系统。

对于自然语言处理的前景发展，目前大多应用在原理研究与应用研究方向上。由于自然语言处理是让机器去处理人类语言，那么一切基于语言、文字的学科都可以与自然语言处理相结合，针对本篇论文的研究方向，是基于文本信息去分析人物性格，而对于某一个人的性格分析往往是涉及到心理学的研究范畴，所以本论文是将自然语言处理技术与社会心理学做一个学科交叉，为的是拓展自然语言处理技术的无限可能。

1.2. 用户画像

用户画像[6]也被叫做用户角色，它是一种描述用户、了解用户的工具。通过运用多种数据分析方式以及产品设计思路去全方位地了解一个企业所拥有的用户群体，同时用户画像也分为群体画像和个人画像，群体画像旨在为用户群体分类，掌握用户画像的群体属性，个人画像旨在为每个用户个人生成其自身特有的“个性化”标签。

用户画像作为近年的新兴领域显示出了蓬勃的生机，在各大领域都有着广泛的应用。用户画像的八要素主要有基本性、同理性、真实性、独特性、目标性、数量性、应用性和长久性。从这八要素出发，我们才能建立一个完好、实用的用户画像模型，才能将一个实际的用户抽象成一个虚拟的、具有代表性的用户画像。

在以往的研究中，主要的工作在于针对移动互联网[7]、手机设备[8]、社交媒体[9]、用户个人信息[10]的用户画像的分析，也有通过分析手机 app [11]、用户长期的搜索记录[12]来个性化推荐[13]或者识别用户兴趣[14]，也有一部分研究基于已有的用户兴趣来达到兴趣化地搜索推荐的效果[15]或者采用一种隐式的[16]潜在的[17]思路去构建用户模型，或者根据移动互联网的广告效果挖掘目标用户[18]。

随着互联网大数据时代的不断发展，企业拥有越来越多的关于用户的数据，用户的行为在企业面前是完全“可视化”的。因此，越来越多的企业聚焦于自己所拥有的海量用户数据，并想以此来改变或重塑用户的行为，以此来做到精准营销等策略，全面提升客户的核心影响力，因此，众多互联网企业一方面希望能在用户使用产品的过程中，利用多维度多渠道的用户行为数据采集方式记录用户尽可能多的数据；另一方面，针对如此规模的用户大数据，企业也很难对用户的数据进行准确的分析，从而导致无法有效的将最优的服务投放到最合适的用户人群手中。近年来兴起的用户画像分析技术正是为了解决这一难题而成为当前用户行为分析的热点技术。

目前在国内的各个领域，用户画像分析都有一定的应用，包括用于精准营销，通过分析潜在用户，针对特定用户群体进行广告投放，减少不必要的广告费用；或者通过分析用户数据的关联性，构建面向用户的个性化推荐系统，对服务或产品做到千人千面的定制化部署；或者进行企业经营效果评估，完善产品运营，提升服务质量，其实这也就相当于市场调研、用户调研，迅速定位服务群体，提供高水平的服务。

对于本课题来说，性格分析是用户画像分析的一个重要方面，相比于用户的其他的静态属性，比如人口属性、信用属性等，性格属性往往难以直接获取得到，所以要应用到自然语言处理技术来获取用户的性格属性。同时为什么选择针对文本信息，第一是考虑到目前的互联网时代，文本信息作为互联网信息的基础，大多数互联网用户目前还是很大程度上依赖于文本信息的输入和输出，以新浪微博为例，大多数用户在其网页端或客户端发微博，并且用户可以在微博下评论，这样就存在大量的文本数据可供分析，所以基于文本信息可以获取到更多的数据源。第二，以文本信息为研究方向所产生的研究成果可以被更广泛地拓展使用，目前自然语言处理技术发展得很成熟，之后基于文本的分析技术可以拓展到基于语音的用户画像分析，只需要把由语音转换成文字，再进行用户画像分析即可，同样地也可以改进聊天机器人以及客服人员的回复等，因为只要实行了用户画像等其他爱好的分析技术后，就可以做到对用户的个性化，国外有一名科学家通过使用其过世好友的聊天记录个性化训练了一个机器人来缅怀其好友，并达到了不错的效果。

由此可见，本课题所研究的方向在现在这个大数据时代有着商业前景和研究价值。

1.3. 人物性格

性格是一个心理学名词，在心理学的定义中，性格是指一个人对现实的稳定的态度，以及与这种态度相应地，习惯化了的行为方式中表现出来的人格特征。也就是说，针对某一个人或者某一个人物来说，性格可以从他的习惯性的行为方式中体现出来。但同时，性格虽然具有稳定性，但在一定程度上也会发生变化。人与人之间的性格往往各不相同。

对于一个人来说，性格因素相当重要。有一个好的性格才能让人喜欢，比如友善积极性格的人往往让人愿意与其相处，而孤立自傲的性格往往会被疏远。在当今社会中，一个人的性格会成为这个人的软实力之一，也会成为其能不能在一个集体中获得认可的关键因素。从职场学的角度来说，性格特征决定的行为方式往往能影响一个人的职业发展方向以及职场关系，因此一个人的性格是极其重要的。

同时，性格是极其复杂的。心理学将性格分为静态结构和动态结构，静态结构中包含了态度特征、意志特征、情绪特征和理智特征；动态结构指的是性格静态结构中的各个特征的相互作用性。人的性格分析是心理学的一个重要研究方向，在本论文中，不去过多地去探究心理学针对性格的分析原理，只选择其中较为广泛应用的大五人格理论，该理论在之后的算法设计环节会详细介绍。

正如上文所提到的，人物的性格是复杂难以分析的，所以针对本课题而言，有以往的论文是针对小说文本基于统计的方法进行人物性格分析[19]，单纯以人物的文本信息来分析其复杂的性格会有很多的难点和瓶颈。首先，文本信息的质量决定了一个人的性格分析的准确率，如果关于这个人物的文本信息都

是十分客观、没有任何主观情感词，那么该人物的性格就无法从文本信息中得以体现。其次，人物的文本信息往往不足以完全表现人物的性格，人物的性格是多维度多层次多方面的，仅仅通过文本去分析一个人物复杂的性格，是很容易不准确的。通过与心理医生沟通调查发现，一个人的神态、动作等肢体语言是他们判断一个来访者心理状态以及人格特征的重要依据，因此当我们只有文本信息时，我们往往会忽略一个人的其他重要信息。最后，为了保证分析的准确性，文本信息的数据量必须够多，作为算法的输入必须要有一个合适的文本数量，才能保证算法对未知数据的有效预测，提高泛化能力。

通过以上对心理学的研究分析，我们了解到人物的各种行为都是人物性格特征的表象体现，所以对于其文本信息，也是与其性格特征相关联的。所以只要我们获取足够多、质量足够高的文本信息数据，加以针对人物性格分析所设计的自然语言处理算法，就可以达到良好的预测效果。

2. 算法分析

针对本课题——基于文本信息的人物性格分析算法，本章节主要谈论如何选择适用于本课题的算法，并且对于算法的选择上，分析各个算法的优势和劣势，并且阐述了为何该算法更适用于本课题的研究。本章节所列出的算法都将成为下一章节算法框架的重要组成元素，本章节的阐述主要是为基于文本信息的人物性格分析算法打下理论基础。

2.1. 自然语言处理算法

1) 文档分类算法

文档分类是自然语言处理技术中常见的一项应用。它旨在解决在给定一个文档或者一段文本的情况下，如何准确地将其分类的问题。确切来说，文档分类称不上是一个具体的“算法”，更像是一个“算法思路”。因为它常常需要将传统的文本处理技术和其他传统的机器学习或深度学习算法相结合，以此来完成它对文档“分类”的任务目标。

具体来说，文档分类算法的主要思路是文档标注、文档格式转换、生成词典、计算词语权重、主题建模、训练分类模型、预测等步骤。在训练分类模型的过程中，可以应用机器学习模型，比如支持向量机、逻辑斯蒂回归模型等，或者应用深度学习构建神经网络进行分类器的训练。

举例来说，近年比较热门的情感分类问题，通过对一段已标注正负向情感倾向的文本进行模型训练和预测，从而达到自动分析一段文字的情感极性的目的。这也是一种典型的文档分类问题。

针对本课题来说，可以应用文档分类算法的算法思路，因为从本质上来说，本课题所研究的主要问题可以类比成“文档分类”。在输入层面上，文档分类算法与本课题所研究的都是文本。在输出层面上，文档分类算法主要注重的是多分类问题，也就是在有限的类数上通过算法预测出该文档的确切类别，而本课题旨在分析人物性格，在之后的算法框架描述上会具体说明预测人物性格并不是简单的分类问题。

除了算法思路之外，从文档分类算法的实现步骤上来具体谈谈一些其他可被本课题借鉴的思路。首先是文档标注和文档格式转换，文档标注主要涉及到文本的标签问题，这会在之后的数据收集方面具体谈到，而在本课题中不会存在文档格式转换这个问题，因为我们分析的就是纯文本，输入的就是文字，不会输入其他pdf、excel 等格式的文件，同时不会有其他的图片等非文字格式的元素。其次是生成词典部分，这会成为本课题的重要步骤之一。余下的其他步骤都是针对“分类”这一目标服务的，对本课题的借鉴意义不大。

2) 分词算法

对于现在大多数自然语言处理算法模型来说，分词算法是一个基础性工作。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。中文与英文的区别就是英文的单词之间是有空格的，而中文只在语句之间有标点符号，字词之间是没办法划分的。以此来看，中文分词的难度很高。

中文分词算法大概分为三大类，第一类是基于字符串匹配，即扫描字符串，如果发现字符串的子串和词典中的词相同，就算匹配。第二类是基于统计以及机器学习的分词方法，它们基于人工标注的词性和统计特征，对中文进行建模，即根据观测到的数据(标注好的语料)对模型参数进行训练，在分词阶段再通过模型计算各种分词出现的概率，将概率最大的分词结果作为最终结果。常见的序列标注模型有 HMM 和 CRF。第三类是通过让计算机模拟人对句子的理解，达到识别词的效果，由于汉语语义的复杂性，难以将各种语言信息组织成机器能够识别的形式，目前这种分词系统还处于试验阶段。

对于本课题来说，采用基于统计以及机器学习的分词方法 n-gram 模型，模型基于这样一种假设，第 n 个词的出现只与前面 N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。并且将 n 设为 3，采用 Trigram 模型，因为高阶 n-gram 虽然对更多的上下文敏感，但是数据拥有更多的稀疏性，而低阶 n-gram 考虑非常有限的上下文信息，但是具有更强的鲁棒性，综合考量选择使用 Trigram 模型。

同样地，作为对比，也使用中文分词工具 jieba。Jieba 作为当下最常用的中文分词工具自带了一个叫做 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。把这 2 万多条词语，放到一个 trie 树中进行词图扫描，而 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。

2.2. 传统机器学习算法

机器学习是多种理论学科交叉生成的产物，其中包括统计学、概率论、高等数学多门学科。机器学习的主要目的是为了分析一些已知的数据，针对不同的数据类型、数据量的多少以及不同的研究目的，来选择并使用相应有效的算法来对未知的数据进行预测或决定，或者探究已知数据自身所存在的一些潜在的、未被人们发现的规律等。通过这样的方式，计算机就可以像人类一样在“经验”数据中获取解决某些任务的方法，从而“学习”到相应的算法，这也正是机器学习的含义所在。

机器学习目前主要分为三种类型：监督学习、无监督学习和强化学习。监督学习与无监督学习的主要区别在于是否已知标签，对于一个明确的任务目标来说，如果你所拥有的数据已经含有任务目标数据，那么这种机器学习任务被称为监督学习。同样地，如果你所拥有的数据不含有你的任务目标，或者你并不需要分析某个任务目标而只是探索已拥有的数据之间的联系，或者你只需要对数据进行一些变换等预处理操作，那么这样的机器学习任务被称为无监督学习。而强化学习与标签关系不大，它的运作机制与监督学习和无监督学习有本质上的差别。强化学习主要是使用机器自身的个人历史经验来决定它的算法走向的。

对于监督学习来说主要分为两大类，分类和回归。分类的一个经典例子是垃圾邮件的分类，对于我们已有的数据来说，是否为垃圾邮件就成了分类算法的任务目标，这就要求在我们已有的邮件信息里就已经包含了其是否为垃圾邮件的标签。那么邮件的其他信息，包括发件人地址、邮件信息内容等都将成为为了预测一封邮件是否为垃圾邮件的重要输入，选择合适的监督学习算法，从而在输入的邮件特征中学习如何区分是否为垃圾邮件。回归问题与分类问题的主要区别在于分类问题的最终目的是要把一条数据归为某一类当中，而回归问题主要是为了预测一条数据的任务目标的值。比较经典的例子是天气的预测，通过选择相应地回归算法来对天气的历史数据进行回归学习，从而拟合出回归曲线，进而达到对未来的未知天气的预测，比如预测温度、湿度和降水量等，从这个例子中可以看出，回归任务与分类任务最大的不同在于预测结果的“类型”不同。

由于在实际的生产生活中，常常没有办法获取到我们想要的数据标签，所以无监督学习也获得了广

泛的应用。对于无监督学习来说，常分为两大方法，聚类和降维。聚类顾名思义主要是把相似数据归为一类，从而达到把数据分为一簇一簇的效果，这与分类有着本质差别，因为分类任务的目标类别是指定好的已知的，而聚类完全是根据所选择的不同算法来生成的，也就是聚类的结果，数据所生成的类别具体是什么类型是人为不可控的，只能根据数据各个属性的相似度来衡量一些数据是否为同一类别，常见的应用比如一些营销手段，通过使用聚类算法来把相似顾客归为同一类，从而使用相同的营销手段，从而达到减少成本、提高效率的目的。无监督学习的另一个应用降维的主要目的是减少数据的维度，因为对于大多数机器学习来说，数据常常有着大量的属性标签，在进行特征工程处理之后一条数据的维度将成指数爆炸级别的增长，这对于计算机来说处理起来就会有很大的困难，而降维的主要目的是对输入标签太多的数据进行预处理，从而寻求其高维数据结构的低维向量表示，将其作为高维数据的特征表达向量，这样就使得高维向量问题转化为低维特征向量问题，大大降低了计算的复杂程度，减少了额外维度的附加信息所造成的误差。

对于本课题而言，输入数据是文本信息，输出信息是人物性格。为了使得传统机器学习算法能够处理本课题的输入输出数据，我们首先要对输入输出数据进行量化处理。首先是文本输入信息，如果我们将文本拆分成字，通过自然语言处理算法直接对字进行处理，那么面临的问题有：一、对于一段文本来说，如果将文本切割成字符形式或者单词形式，那么输入向量的维度就是整段文本的字个数或者词个数，这就意味着输入向量将面临指数级别的维度增长问题，作为训练数据会使得模型的处理能力变弱，这样对于输入向量的表示是不推荐的。二、从自然语言处理的角度分析，如果将文本切割成字，那么就会忽略掉上下文的信息，因为一段文本之所以能表达含义，是通过文本的上下文语法结构来实现的，切割成字必然会打乱文本原本想要表达的信息。总之，通过以上两点原因的分析，输入向量不能使用简单地将文本切割成字的方式去实现，具体的实现方法在上一节中有提到自然语言处理中的分词算法，而将分好的词变成向量，需要用到后面章节中的深度学习算法，深度学习算法是如何应用的在之后的章节会提及，所以在此先假设我们已经获取了关于一段文本的特征向量表示，这就是我们的输入信息。

对于输出信息而言，本课题研究的是人物性格。为了量化人物性格，本课题采用大五人格理论作为人物性格的量化标准，大五人格理论在本课题中如何应用的具体说明会在之后章节中阐述，假设我们在此已经获得了人物性格的量化表示，同样的，也是一个向量。

那么在输入和输出都是一个多维向量的情况下，传统机器学习算法在本课题中是如何得以应用的。通过分析输入输出向量的条件发现，本质上这是一个输出向量为连续值的预测问题，也就是回归问题，那么可选择的传统机器学习算法有感知机算法，同样的，如果将性格的向量输出看作是一个有“阈值”的分类问题，那么可应用的机器学习算法有逻辑斯蒂回归算法、决策树算法等。此外，为了降低输入向量的维度，减小模型复杂度，非监督学习的降维和聚类算法也会在本课题中有所应用。关于这些机器学习算法如何在本课题中应用，将在后续算法框架设计中具体阐明。

2.3. 人物性格算法

如背景中所提及的，人物性格可从人物的语言文本中分析得到，但是如何量化复杂的人物性格，从而使算法能够加以预测，这是我们需要考察研究的。

人物性格的量化方法经历了一百多年的历史，首先心理学家、语言学家便提出词汇学假设，人格词汇学假说指的是人类特性的个体差异可用一套适用于世界上多种语言甚至所有语言的简单术语表达出来。该假设使得人格特质从可词汇使用中推断出来这一想法得以发展，从奥尔波特到卡特尔，先后做了很多基于该假设的工作，主要是在语料词典中挑选出所有能够描述人格差异的词汇，经过一层一层筛选，并对比了自我评定、同伴评定和心理咨询师的评定，最终发现所有与人格相关的词汇大多会出现在五个因

素中，在之后的心理学家的研究中也发现这五个因素的普遍性，因此大五人格理论被提出，且成为心理学权威的人格模型。

大五人格理论指的是人的人格描述主要由五种特质涵盖——开放性(Openness)、责任心(Conscientiousness)、外倾性(Extroversion)、宜人性(Agreeableness)和神经质性(Neuroticism)。因此大五人格理论也常被称为 OCEAN 人格的海洋。开放性主要反映人物个体对于新鲜事物的接受程度，开放性得分高的人往往用于探索未知领域，对从未接触过的知识或生活状态保持着开放的态度，因此可能具有兴趣广泛的特征，而得分低的个体一般情况下是墨守成规的、性格保守且比较传统的。责任心主要反映了个体的自控力，责任心得分比较高的人通常给人的感觉是靠谱的、值得信赖的，而得分比较低的个体通常是比较冲动的、自控能力比较差的，而冲动的个体有时会被当作有趣的朋友，虽然这有时会带来各种麻烦，尽管责任心高的人会在工作中可能会取得领导更高的信任，但极端情况下会出现事事完美主义这一情况，会让人觉得单调乏味。外倾性主要反映的是与人交往从而获得快乐的能力，在外倾性下得分高的个体通常会有较强的交际能力，喜欢与人接触并很热情，而得分低的个体通常比较谨慎内向，喜欢独处但并不一定抑郁。宜人性反映的是人物个体对其他个体所持的态度，与外倾性是不同的，宜人性高的人通常是有同情心的、宽容的、容易信任他人的等等，反之宜人性得分低的个体通常是有心机的、愤世嫉俗的、无情的等等，可以看出宜人性高的人通常是无私的、具有奉献精神的，而宜人性低的人是总会把自己的利益高于别人。神经质主要反映的是人物个体的情感调节情况，反映出情绪的不稳定程度。比如神经质性得分高的人经常会有更多的心理压力，对外界的刺激反应会比较强烈，而得分低的人一般会比较少出现情绪化的状况。

因此，在本课题中，我们采用大五人格理论作为人物性格的考量标准。将文本中的人物的人格量化成开放性、宜人性、外倾性、责任心和神经质性五个维度上的得分，以此来使得本课题所研究的输出向量为一个五维向量，同时该五维向量的各个维度是连续的，那么本课题的研究目标也就转化成如何训练一个算法模型来拟合一个五维的输出连续向量。

2.4. 深度学习、神经网络算法

正如背景所说，语言具有动态性、歧义性和不规范性，我们通常需要一定的推理能力和足够的知识储备才能够推断出一句话中所包含的含义内容。随着互联网时代的快速发展，网络语料库成爆炸式的增长，这为自然语言处理技术的发展提供了丰厚的土壤，同时也让人们能够更深刻地理解语言学的机理机制。

由于语言的复杂性，使得传统机器学习在直接应对自然语言处理这一复杂问题时略显“吃力”。统计机器学习在自然语言处理上的应用被称为统计自然语言处理。统计自然语言处理的主要设计思路是由训练数据和统计模型两部分构成的，然而统计自然语言处理在这两部分上都存在严重的问题。首先是获取训练数据环节，对于文本语料库而言，如果想要对文本进行处理，就要获取标注数据，无论是输入向量还是输出向量，都需要大量的人工工作进行繁琐的标注工作，这样使得获取大量的高质量的训练数据需要更多的财力物力成本。然后是统计模型也就是算法设计环节，由于传统机器学习中并不存在与语言学相对应的机器学习算法，也就是说在机器学习算法中，数据并不具有语言学特征，这使得在输入进算法模型中之前就要进行语言学特征的人工设定，从而保证在算法训练环节考虑语言学因素，从而保证整个算法模型的对文本输入的泛化能力。这样的前期人工特征设计以及预处理更加加大了模型训练的时间以及人力成本，因此，统计自然语言处理存在诸多短板，不适合广泛应用。

近年来，深度学习的快速发展为统计自然语言处理所存在的问题提供了一个解决方案。因为深度学习算法一般建立在复杂神经网络结构上，这样的神经网络结构可以对多种数据进行抽象表示从而进行学

习，神经网络算法已经在图像等多种领域大放异彩，同样地，深度学习算法也可以在自然语言处理任务中得以应用。

针对本课题而言，深度学习算法的主要应用点在于处理人物的输入文本，将输入文本抽象处理成文本向量，同时要保证获取到的文本向量能够尽量表示文本的语法结构以及语义信息，这样才能够尽量保证信息的不丢失，增强模型的预测能力。

具体而言，主要应用的神经网络模型的深度学习或神经网络算法有 Word2Vec、Doc2Vec、LSTM 算法，以及为了处理文档向量而采用的 CNN 算法。

1) Word2Vec 算法

为了解决如何计算一段文本序列在某种语言下出现的概率这一问题，研究者们提出了 Ngram 模型，即根据要预测的词的前 n-1 个词来获取该词的条件概率，常见的如 bigram 模型即 N = 2 和 trigram 模型即 N = 3。

然而由于 Ngram 模型的参数会随 N 而产生爆炸式的增长，所以 N 的取值一般不超过 3。此外，Ngram 模型只考虑了每个预测词的上文，而忽略了词与词之间的内在联系以及相似性，比如 women 和 man 在语句的语法结构和语义表示上有极大的可替换性，所以对于两个语句 The women is sleeping 和 The man is sleeping 而言，如果能计算出前者的概率，那么即可通过 women 和 man 的相似性来推断出后者的概率，而这是 Ngram 模型无法做到的。究其根源而言，主要是因为 Ngram 模型将词当作成一个个孤立的原子单元去处理的，对于每个词而言其都对应一个 one-hot 向量，比如 The women is sleeping 对应一个大小为 4 的词典，而 women 对应的向量为[0, 1, 0, 0]，也就是说，在 Ngram 模型中，每个词对应的向量长度为词库的大小，如果语料库有成千上万的词，那么这样就面临着词向量维度爆炸的问题。

为了解决 Ngram 的问题，2003 年，Bengio 等人提出了 Neural Network Language Model，并首次提出了 word embedding 的概念，具体原理不加赘述，其解决了统计语言模型中条件概率的计算和向量空间模型中的词向量的表达着两大问题。

然而该模型存在难以处理变长序列和训练速度太慢两大问题，因此 Mikolov 提出了改进方案，也就是 Word2Vec，简单阐述一下 Word2Vec 的原理，其主要分为两种模型——CBOW 和 Skip-gram，前者是根据中心词的上下文来预测中心词，后者相反，利用中心词来预测上下文。同时，Word2Vec 还有两种优化算法——Hierarchical Softmax 和 Negative Sampling，前者是将复杂的归一化概率分解为一系列条件概率的乘积，并利用哈弗曼树来提升训练效率，后者是为了解决那些无法归一化的概率模型的参数预估问题而改造模型的似然函数。

针对本课题而言，Word2Vec 的主要作用在于将输入文本向量化，具体来说是将输入文本的“词”向量化，得到可表征语义的稠密词向量，从而为该算法的后续步骤做好基础性工作。

2) Doc2Vec 算法

在通过 Word2Vec 算法获取到文本的各个词向量之后，我们面临着一个问题，就是如何将词向量与输入文本的文档向量相关联，也就是说如何用获取到的词向量来生成文档向量，并且能保持 Word2Vec 词向量的语义丰富性和高质量。

以往的主要思路有 bag of words、LDA、平均词向量、tfidf 加权词向量等。首先 bag of words 不适用，因为其单纯根据词频来建立文档向量忽略了单词的顺序和语义。然后 LDA 模型主要应用在于计算文本的主题分布，在原理上与 Doc2Vec 有所不同。其次平均词向量和 tfidf 加权词向量都是对文档中所有的词向量进行处理，前者直接求平均值来代表文档向量，后者通过计算 tfidf 的值来让句子中更重要的词权重更大，然而两者都存在一个缺陷就是并没有考虑单词的顺序。

为了解决以上问题，Doc2Vec 算法在 2014 年被提出，是一种非监督学习算法，可以获得句子、段落、

文档的向量表示，是 Word2Vec 的拓展。简单阐述一下原理，Doc2Vec 即是在输入层输入词向量时，把文档向量添加成一个额外的输入，与其他词向量一起映射到投影层进行级联或求均值，这被称作为 Distributed Memory Model of Paragraph Vectors (PV-DM)，与 Word2Vec 类似，Doc2Vec 也有另一种训练方式——PV-DBOW，即忽略输入的上下文，让模型预测段落中随机的一个单词，而模型的输入就是段落向量，对于大多数任务而言，PV-DM 表现更好。

针对本课题而言，可用 Doc2Vec 模型直接获取文档主题向量表示，也就是针对每个人物的每段文本都可以利用 Doc2Vec 算法获取到该人物的文本特征向量集作为输入向量集来进行训练。

3) CNN 算法

CNN 即卷积神经网络是近几年发展起来得到广泛应用的神经网络模型，20 世纪 60 年代，Hubel 和 Wiesel 在研究猫脑皮层中用于局部敏感和方向选择的神经元时发现其独特的网络结构可以降低反馈神经网络的复杂性，卷积神经网络(Convolutional Neural Networks——简称 CNN)因此被提出。就目前而言，卷积神经网络主要应用于图像处理问题，包括边界检测等应用。

CNN 的结构构成主要有 3 个层构成：Convolutional layer (卷积层——CONV)、Pooling layer (池化层——POOL)、Fully Connected layer (全连接层——FC)。卷积层主要由滤波器和激活函数构成，池化层主要去设定模型的池化方法是 Max pooling 还是 Average pooling，最终全连接层即神经网络中最普通的一排神经元。

CNN 中有一些基本概念需要阐明。首先是 padding 填白，因为 CNN 的卷积操作会使输入矩阵经过每次卷积之后都会缩小，这样就会存在矩阵消失的问题，而且输入向量边缘的值会因此而无法进行多次运算，从而使得信息缺失，所以可以采用 padding 填白操作来补 0。其次是 stride 步长，这控制了滤波器 kernel 的卷积步长，从而影响了每次卷积操作之后的输出矩阵大小。然后是 pooling 池化，池化的主要作用是为了提取一定区域的主要特征，从而减少参数数量，防止过拟合，方式主要有 Max Pooling 和 Average pooling，一般使用前者。最后是对多通道图片的卷积，这一概念一般使用在彩色图像 RGB 的处理上，与本课题相关性不大，不加赘述。

针对本课题而言，CNN 主要为了处理文档矩阵或词向量所拼接成的文档矩阵，目的是降低模型复杂度，提取文档中的关键特征，提高训练效率和模型的泛化能力，并且通过 CNN 之后，能获得更低维度、更高凝练度的文档向量，也为了之后的全连接层或机器学习层做准备。

3. 实验论证

3.1. 数据准备

针对本课题而言，针对人物的分析主要分析“静态”人物，输入文本主要选取“静态”文本，所谓“静态”指的是人物与文本是不随时间的变化而改变的。因此，小说、剧本等文学创作作品都属于本课题的研究范畴，因为在此类文学作品中，人物在一个有限的时间线内活动，人物性格的变化是固定的，但是一般情况下根据作者的文笔，所描绘的主要人物性格一般是鲜明的，这为本课题所研究的基于文本分析人物性格这一目标提供了优质且不冗杂的文本数据集，有利于将研究成果推广至现实世界中的人物“动态”文本，比如微博博文、搜索记录等。

在输入文本的选取上，我们选择《平凡的世界》作为主要分析文本，《平凡的世界》是中国作家路遥创作的一部百万字的小说，这部长篇小说全景式地表现了中国当代城乡社会生活，以中国七十年代中期到八十年代中期作为时间背景，以孙少安和孙少平两兄弟作为人物核心，刻画了复杂的矛盾和人物心理的纠葛，也反映出了当时社会环境下各个阶层人民的形象，该小说获得中国第三届茅盾文学奖。

可以看出，《平凡的世界》作为文本分析具有一定的权威性。首先，这部文学作品语料丰富，充分

地刻画了其所描绘的人物形象，每个人物都有充足的语言对话可供分析，这样就保证了模型的输入文本的文本质量。其次，这部文学作品在许多论文文献中都作为训练样本集出现，这也充分说明了其在自然语言处理相关问题上具有一定的实践意义和参考意义，同时，也可以让本课题所研究的任务目标有一定的参照目标，从而有所对比。

据了解，在《平凡的世界》中，作品刻画的主要人物有孙少平、孙少安、田润叶、田晓霞和贺秀莲。孙少平作为男主角，虽然家庭环境极其贫困，但是他有不为命运低头的志气，即便物质上有所匮乏，但其求真务实的态度让他的骨子里有着一股傲气，具有忍辱负重、踏实肯干的特征，在爱情的选择上也可以看出他的责任和担当。孙少安作为孙少平的哥哥，因为家庭上的种种原因，性格上就更加内敛，不能像孙少平那样自由追求自己的理想。田润叶在作品中虽然是一个普通农民的女儿，但是却因为接收了开放程度更高的大城市教育而敢于追求自己想要的生活，在与孙少安的感情线中，田润叶表现出勇敢大胆的特质，不介意孙少安的家庭环境等其他因素，始终认为两人要在一起那么感情才是第一位的。田晓霞在作品中与其他人不同的是，她是在城市中长大，并没有经历农村的贫苦生活，但是她却不娇弱，反而有着朴实、善良、单纯、勇敢等特性，同时城市的教育让她接触了更高的文化，让她有着更高的理想抱负，她天真、热情并且坚强。最后是贺秀莲，她性格泼辣、踏实肯干，嫁给了孙少安，成为孙少安家中的好帮手，在孙少安家中为了家庭勤勤恳恳，把家中打理得井井有条，让孙少安可以全身心地在外忙事业而无后顾之忧。

首先使用 jieba 的 posseg 工具来统计获取文本中所有词性为 nr 即人名的出现次数，并将“孙少平”、“少平”、“孙少安”、“少安”、“田润叶”、“润叶”、“田晓霞”、“晓霞”加入自定义词库，通过 jieba 分词工具来加载自定义特征词库来提升分词效果，降序排列可得到下图结果，见图 2。

Figure 2. The number of appearances of characters in the ordinary world
图 2. 《平凡的世界》人物出场次数

通过以上对五个人物的简单介绍以及词频统计分析，我们可以看出这些主要人物出现次数较高，并都具有鲜明的人物特征，甚至其中对人物的特征描述可直接映射到大五人格量表中的词汇，因此在此我们选取孙少平、孙少安、田润叶和田晓霞作为人物分析样本进行人物性格分析。

3.2. 文本预处理

在文本预处理环节，首先需要将孙少平、孙少安、田润叶和田晓霞的人物描述抽取出来，人物描述包括其肖像描写、语言描写、动作描写和心理描写，因此在本课题中，首先需要把要研究的主要人物的相关描写语句抽取出来，通过将每行中出现该人名的文本读取出来，我们获取到了四位主要人物的分析样本，如下图所示，见图 3。

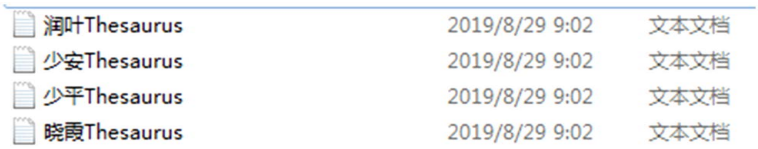

Figure 3. Sample analysis of characters in the ordinary world
图 3. 《平凡的世界》人物分析样本

每行语句中都或多或少有着对人物的神情、形态、动作、语言的描述，如下图所示，见图 4。

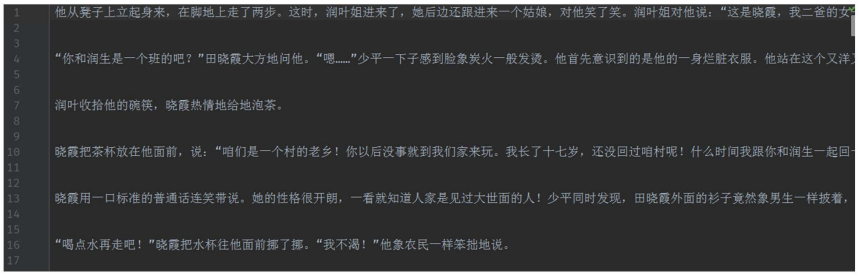

Figure 4. Excerpt from the description of characters in the ordinary world
图 4. 《平凡的世界》人物描述摘选

在输入模型之前，先去除空行，以免输入空向量影响模型效果。结果如下图所示，见图 5。

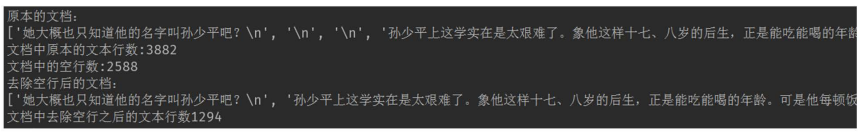

Figure 5. Analysis input text
图 5. 分析输入文本

3.3. 生成文档向量

在本环节中，我们主要采用两个思路来生成文档向量，直接使用 doc2Vec 或使用 word2Vec 生成词向量再拼接成文档向量。

1) doc2Vec 生成

doc2Vec 的参数设置为：min_count = 1，min_count 可以对字典做截断。词频少于 min_count 次数的单词会被丢弃掉，此处设置为 1，说明原则上在训练文档向量时不放弃任何一个出现的词；window = 3，window 指的是窗口大小，表示当前词与预测词在一个句子中的最大距离是多少；vector_size = size，是指特征向量的维度，此处设置值为 200，虽然大的 size 需要更多的训练数据，但是效果会更好，一般情况下推荐值为几十到几；sample = 1e−3，指的是高频词汇的随机降采样的配置阈值，默认为 1e−3，官网给的解释 1e−5 效果比较好。设置为 0 时是词最少的时候，即不进行降采样，结果词少，当设置 1e−5，相应的词展现更丰富；negative = 5，如果设置该值，则会采用 doc2Vec 中的 negative sampling 方法，此参数用于设置多少个 noise words (一般是 5~20)；workers = 4，用于控制训练的并行数。

通过 doc2Vec 生成四个主要人物的向量模型以及向量表示如下图所示，见图 6 和图 7。

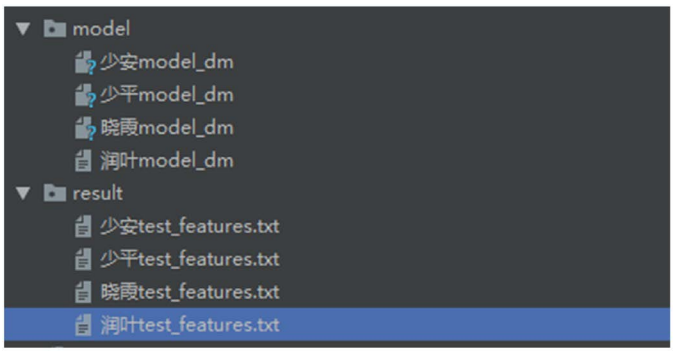

Figure 6. Doc2vec character vector model
图 6. Doc2Vec 人物向量模型

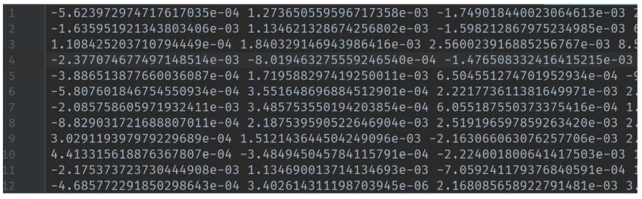

Figure 7. Doc2vec character vector representation
图 7. Doc2Vec 人物向量表示

2) Word2Vec 生成

word2Vec 的参数设置为 min_count = 3、size = 50、window = 5、workers = 4，由于 word2Vec 中的这些参数与 doc2Vec 类似，此处不加赘述。

在使用 word2Vec 获取到词向量之后，针对该人物的每一条描述将所有的词向量拼接成文档矩阵，使用高度为 2，3 和 4 的 300 个卷积核进行卷积处理，再将得到的结果最大池化，可得 300 维的特征文档向量，如下图所示，见图 8。

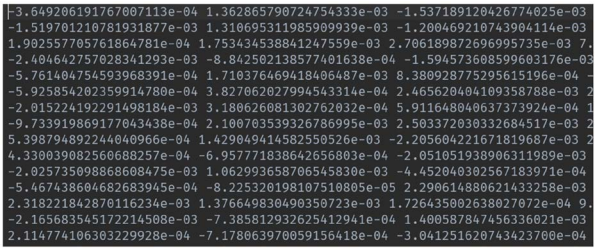

Figure 8. Word2vec character feature document vector
图 8. Word2Vec 人物特征文档向量

3.4. 模型训练

通过对文学作品的分析，我们知道，孙少安慷慨待人、热情洋溢，这样的品性使得孙少安具有极高的外倾性(Extroversion)，同时田晓霞相比于孙少安而言，不像孙少安外倾性那么高，因此我们将孙少安

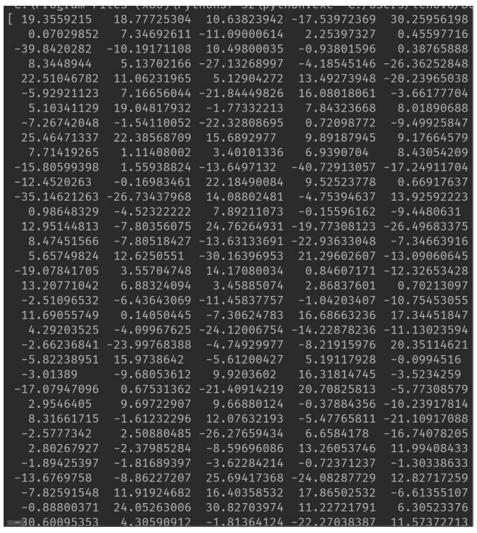

Figure 9. Vectors from doc2vec training
图 9. Doc2Vec 训练所得向量

和田晓霞的大五人格中的外向性作为输出样本进行分析，将孙少安的输入文档向量集所对应的得分控制在 93~95 分之间，将田晓霞的输入文档向量集所对应的得分控制在 83~85 分之间，因此，整体训练集规模为孙少安的语句文档向量总数，即 1130 加 421 共 1551 条数据。

采用 sklearn.linear_model 中的 LinearRegression 来进行线性拟合，训练文档向量的模型参数。

Doc2Vec 模型所训练出的向量如下图所示，见图 9。

均方根误差为 0.6916464343383234。

Word2Vec 结合 CNN 所训练出来的向量如下图所示，见图 10。

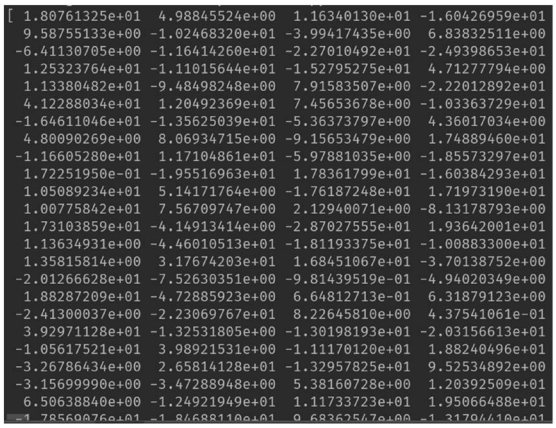

Figure 10. Word2vec training vectors
图 10. Word2Vec 训练所得向量

均方根误差为 0.6483908566794765。

虽然两个向量在数值表现上有一定的差异，但从均方根误差而言，两者相差不多。

3.5. 模型预测

通过使用两个模型所训练出的针对外向性这一维度的预测向量来预测同样具有外倾性特征田润叶，根据对作品的分析，可以看出在外倾性上田润叶与孙少安具有相似的得分程度，因此我们将田润叶的外倾性得分设为 95 来比较两个预测向量的效果，并验证该算法的实用性。

Doc2Vec 向量输出结果如下，见图 11。

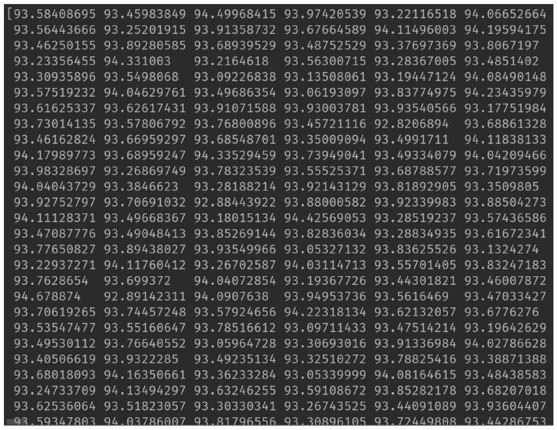

Figure 11. Doc2vec vector output results
图 11. Doc2Vec 向量输出结果

均方根误差为 1.4312380302855083。

Word2Vec 结合 CNN 输出的结果向量如图，见图 12。

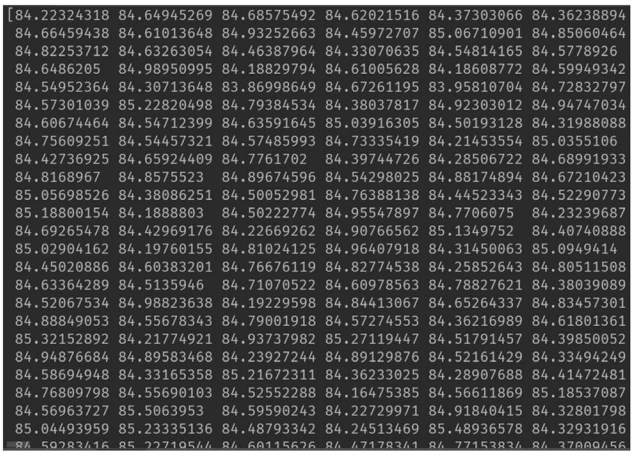

Figure 12. Word2vec vector output results
图 12. Word2Vec 向量输出结果

均方根误差为 10.657833760771707。

3.6. 实验结果分析

通过以上实验，我们可以看出在大五人格外倾性这一维度上，doc2Vec 和 word2Vec + CNN 都可以训练出各自的特征向量，并且都能够保证其实验误差较小，保证了对训练数据的拟合能力，确保较小的偏差。

同时，在针对训练集所划分出的验证集的表现上，我们可以看出两者的均方根误差都比较小，也就是说在针对本训练集上的未知数据，两者模型都存在较好的泛化能力，能够较准确地预测出人物在外倾性上的大五人格得分。

但是在对未知数据的预测上，两个模型的表现却有一定的差异性，当我们选择使用该模型去预测一个新的人物——田润叶的外倾性得分时，doc2Vec 的预测效果要优于 word2Vec + CNN 的预测效果，doc2Vec 的所预测的得分为 93.70167173504245，而 word2Vec + CNN 的得分为 84.72613665200262，这对于田润叶这一人物来说，其在作品中的表现具有极强的外倾性，显然相比于与田晓霞而言，在外倾性上的得分应高于她，因此 doc2Vec 的模型表现要优于 word2Vec + CNN。

从算法结构上分析，我们可以推测，doc2Vec 模型直接使用文本作为输入所得到的文档向量更具有普适性，任何词语不会被赋予更高的关注度，相比于 word2Vec + CNN 所训练出来的模型在本训练集上表现更优而泛化能力不强，主要原因在于 word2Vec 获取到词向量后拼接成的文档向量在经过 CNN 处理之后更容易突出某些关键词的作用，因为 CNN 有着突出局部特征的效果，因此在分析未知人物性格时，该模型容易受之前的人物训练集所影响，难以达到客观评分的目的。

4. 总结

针对本课题而言，基于文本信息分析人物性格，通过对比 doc2Vec 和 word2Vec + CNN 发现，前者在预测未知人物性格时有着更好的表现，该课题使用《平凡的世界》作为分析样本，采用神经网络与传统机器学习相结合的方式来训练模型，以此我们可以看出，该模型可以用来智能化地分析静态文本来获取小说、剧本等文学作品中的人物性格，通过大五人格量表，我们可以通过预测得分来映射出人物的性格词汇，从而使得智能阅读理解成为可能。

同时，本课题也存在一些不足，在文档向量与人物性格得分训练的环节，考虑到任务目标，本课题只使用了线性回归，同时可以考虑使用深度神经网络等模型进行参考。此外，在设置人物性格得分时，是靠人为去分析文本设定的，显得不够智能。

本课题在未来可拓展到动态文本的分析当中，分析微博博文、百度搜索词条等来获取实际用户的性格评分，来进行用户画像和推荐，或者应用于心理学分析访谈对话的内容来获取受访者的心理状态来辅助心理咨询师进行心理评判等，都具有重要意义。

参考文献

[1] Mikolov, T., Sutskever, I., Chen, K., *et al*. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Neural Information Processing Systems*, 3111-3119.

[2] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. Arxiv: Computation and Language.

[3] Le, Q.V. and Mikolov, T. (2014) Distributed Representations of Sentences and Documents. *International Conference on Machine Learning*, 1188-1196.

[4] Hu, J., Jin, F., Zhang, G., Wang, J. and Yang, Y. (2017) A User Profile Modeling Method Based on Word2Vec. 2017 *IEEE International Conference on Software Quality*, *Reliability and Security Companion*, Prague, Czech Republic, 25-29 July 2017, 410-414. https://doi.org/10.1109/QRS-C.2017.74

[5] 李恒超, 林鸿飞, 杨亮, 等. 一种用于构建用户画像的二级融合算法框架[J]. 计算机科学, 2018, 45(1): 157-161.

[6] 郭炜. 架构师特刊: 用户画像实践[Z]. 2017.

[7] Yin, Z.Y., Jiang, Y. and He, J. (2017) Analysis of Mobile Internet Multi-Context User Preference. *Advances in Engineering Research* (*AER*), **130**, 1175-1180. https://doi.org/10.2991/fmsmt-17.2017.232

[8] Tang, T.T., Yin, J.Y. and Zou, Y. (2017) A Method for Telecom User Portrait Modeling. *Advances in Engineering Research* (*AER*), **130**, 1181-1187. https://doi.org/10.2991/fmsmt-17.2017.231

[9] Gu, H., Wang, J., Wang, Z., Zhuang, B. and Su, F. (2018) Modeling of User Portrait Through Social Media. 2018 *IEEE International Conference on Multimedia and Expo*, San Diego, CA, 23-27 July 2018, 1-6. https://doi.org/10.1109/ICME.2018.8486595

[10] Ouaftouh, S., Zellou, A. and Idri, A. (2015) User Profile Model: A User Dimension Based Classification. 2015 10*th International Conference on Intelligent Systems*: *Theories and Applications*, Rabat, Morocco, 20-21 October 2015, 1-5. https://doi.org/10.1109/SITA.2015.7358378

[11] Ma, Q., Muthukrishnan, S. and Simpson, W. (2016) App2Vec: Vector Modeling of Mobile Apps and Applications. 2016 *IEEE*/*ACM International Conference on Advances in Social Networks Analysis and Mining*, San Francisco, CA, 18-21 August 2016, 599-606. https://doi.org/10.1109/ASONAM.2016.7752297

[12] Tan, B., Shen, X. and Zhai, C. (2006) Mining Long-Term Search History to Improve Search Accuracy. *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, 20-23 August 2006, 718-723. https://doi.org/10.1145/1150402.1150493

[13] Teevan, J., Dumais, S.T. and Horvitz, E. (2005) Personalizing Search via Automated Analysis of Interests and Activities. *Proceedings of the* 28*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, **51**, 449-456. https://doi.org/10.1145/1076034.1076111

[14] Qiu, F. and Cho, J. (2006) Automatic Identification of User Interest for Personalized Search. *Proceedings of the* 15*th International Conference on World Wide Web*, Edinburgh, Scotland, 23-26 May 2006, 727-736. https://doi.org/10.1145/1135777.1135883

[15] Ma, Z., Pant, G. and Sheng, O.R. (2007) Interest-Based Personalized Search. *ACM Transactions on Information Systems*, **25**, Article No. 5. https://doi.org/10.1145/1198296.1198301

[16] Shen, X., Tan, B. and Zhai, C. (2005) Implicit User Modeling for Personalized Search. *Proceedings of the* 14*th ACM International Conference on Information and Knowledge Management*, Bremen, 31 October-5 November 2005, 824-831. https://doi.org/10.1145/1099554.1099747

[17] Teevan, J., Dumais, S.T. and Horvitz, E. (2010) Potential for Personalization. *ACM Transactions on Computer-Human Interaction*, **17**, Article No. 4. https://doi.org/10.1145/1721831.1721835

[18] Zhang, T., Cheng, X., Yuan, M., *et al*. (2016) Mining Target Users for Mobile Advertising Based on Telecom Big Data.

2016 16*th International Symposium on Communications and Information Technologies*, Qingdao, 26-28 September 2016, 296-301. https://doi.org/10.1109/ISCIT.2016.7751639

[19] 吴育锋, 吴胜涛, 朱廷劭, 等. 小说人物性格的文学智能分析: 以“平凡的世界”为例[J]. 中文信息学报, 2018, 32(7): 128-136.