**技术交底书撰写**

|  |  |  |  |
| --- | --- | --- | --- |
| 发明名称： | 一种基于云模型的专业词库分类和语料分析方法 | | |
| 本专利发明人： | 丁亦刚 | | |
| 专利申请人： | 杭州中策职业学校钱塘学校 | | |
| 交底书撰写人： | 丁亦刚 | 电话：19854819020 | 传真： |
| E-MAIL：909975853@qq.com | |
| 技术联系人： | 丁亦刚 | 电话：19854819020 | 传真： |
| E-MAIL：909975853@qq.com | |

**该技术应用产品：**

对于专业领域文本通过专业词库分类，从而对专业文本进行量化分析的方法。

本发明专利可用于专业词库中，对分词进行自动分类和分级的标注工作，首先通过少量的人为标注实现大语料库的分词分类，并将这种分级和分类过程转变为句子粒度、文章粒度和对象级粒度的分数，实现对文本内容的量化评价。

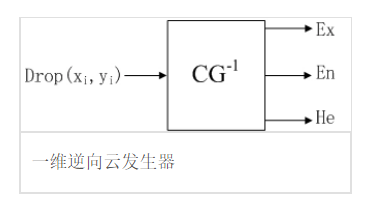
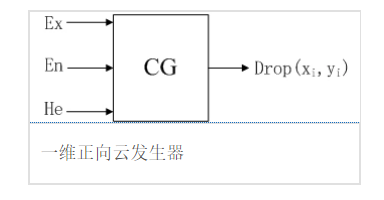
以下举例2个案例：

以敏感词库为例，设置一级敏感词，二级敏感词，三级敏感词后，可经由少量的标注扩展词库标注，而后使用该词库能对文本内容进行量化的分析，分析文本的敏感程度（如某一句话的一级敏感词指数较多，则其敏感程度更高）。

以计算机专业词库为例，标注划分并细致分类（软件工程类词库、网络词库、人工智能词库）后，可经由少量的标注扩展词库标注，而后使用该词库能对文本内容进行量化的分析，分析在不同粒度下的文本或对象偏向于哪一类型的分类倾向（如某一句话更接近软件工程类的句子，某篇文章是更接近人工智能的文章，某个人发送的东西更倾向于软件工程，说明他更关注这一方面的信息）。

**术语解释：**

**云模型**



云模型是云的具体实现方法，也是基于云的运算、推理和控制等的基础。它可以表示由定性概念到定量表示的过程（正向云发生器），也可表示由定量表示到定性概念的过程（逆向云发生器）。

该模型是由中国工程院院士李德毅于1995年提出云模型的概念，是处理定性概念与定量描述的不确定转换模型。自提出，已成功地应用到自然语言处理、数据挖掘、决策分析、智能控制、图像处理等领域 。

[具体公式可见](https://baike.baidu.com/item/%E4%BA%91%E6%A8%A1%E5%9E%8B/1198972?fr=aladdin)。

解释：可以将云模型理解为一个分类器，通过输入不同类别的特征数据，云模型会将每一个分类的分布以Ex、En和He（期望、熵、超熵）三个参数进行表示，反过来，对于每一个分类，都能由Ex、En和He来进行数据的模拟生成。

**词嵌入：**

词嵌入（Word embedding），又叫词嵌入技术。是自然语言处理（NLP）中的一组语言建模和特征学习技术的统称，其中来自词汇表的单词或短语被映射为一个实数的向量。 从概念上讲，它涉及从每个单词一维的空间到具有更低维度的连续向量空间的数学嵌入。生成这种映射的方法包括神经网络，单词共生矩阵的降维，概率模型，可解释的知识库方法。词嵌入技术可以用作一个神经网络训练器的底层输入表示，即先将文本转化为向量空间中的坐标，再输入到分类模型中，以实现对文档的决策和分析。

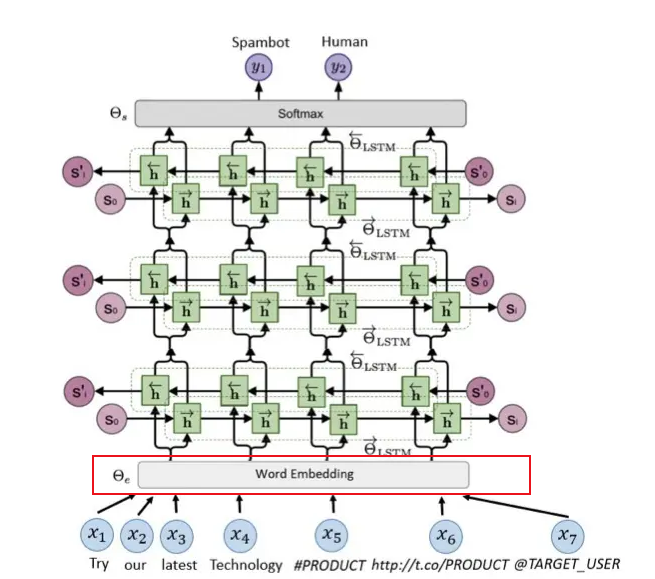
通俗的解释：词嵌入是根据大量文本根据其上下文信息训练出来的一种用向量表示词的方法。如简单的词嵌入假设是，根据人类自然语言中，相似词的意思上下文可能是一致的，如“渴”和“饿”：

*我现在很渴。*

*我现在很饿。*

可以基于不同的算法表示上述“渴”和“饿”的向量，他们在嵌入空间中应该是相似的，也就是近义词。但是需要极大量样本的训练才能准确地表示上述两个词之间的近义关系，且不同的词嵌入的特点不同。如词嵌入可能具有不同的维度数，所包含的分词大小不同，所富含的上下文信息量不同等，由于神经网络的可解释性原因，词嵌入可以包含上下文信息，语义信息，情感信息等。

**词嵌入层的过程如图中红色框所示：**

IMG_256IMG_256

**词库分类/分级：**

包括但不限于以下对应用场景：

词库分类可以理解为对分词所属的种类，如情感词库可以分类为积极的还是消极的，也可以为分类为喜怒哀乐苦悲等多种类别，或将已分类的情感词库指标再做细分，这种水平的分类意味着每一个分词所代表的分数和权重都相等。

需要说明的是，词库分级也可以理解为一种词库分类：如：敏感词库的敏感程度，暴力词的暴力程度等，这种垂直方向的分级意味着每一个分词具有特定的分数和权重因素，如敏感程度越高的分词，其权重越大。

因此，词库分类/分级即将一个词库按照一定的逻辑依据，采用定性的方法进行分类，以预备为一个多级别或多类别的词库，应用于情感识别、自然语言处理或数据分析中。

**应用场景：**

1. 如：LIWC和文心（textmind）工具，构建了一个中文词库分类成了（脏话词、第一人称词、宗教词、积极情感词、消极情感词），然后使用这种词库计算词频占比，以达到内容分析的效果。
2. 如snowNLP工具，将中文的积极情感词分为喜、乐、惊、恶等模块，然后对文本对词进行匹配以分析情感倾向。
3. 例如：构建敏感词的分等级词库，以求对不同等级的词做不同的决策。屏蔽

目前上述的这些词都需要大量的人工编码，即，先获取一个大规模的相关的词库，然后通过人为编码的方式，对每一个词加以甄别。

本方法仅需要不同类别少量的词的数据，将定量的分析转化为定性的内容，发现每个词在词嵌入空间中所处的相对位置，以获取大量的同类别或同级别的词。

1. **详细介绍技术背景，并描述已有的与本发明最相近似的实现方案**（包括两部分：背景技术及现有技术方案［大的技术背景和小的技术背景］，应详细介绍，以不需再去看文献即可领会该技术内容为准，如果现有技术出自专利、期刊、书籍，则提供出处）

本发明是一种：对于专业领域文本通过专业词库分类，从而对专业文本进行量化分析的方法。

技术背景介绍：

1.其中使用的技术背景是基于一个专业词库，对其进行初步的人工标注，形成分类或分级。

2.将分词对应的词嵌入坐标进行读取，以不同类别的词及其嵌入坐标导入到云模型中，云模型会得到每一个类别的Ex、En和He（期望、熵、超熵），可以理解为每一类词在词向量空间中的分布，将定性的分类转换为定量的特征。

3.通过正向云模型，随机生成各词类型的n个词坐标，以这些词作为各个词类的词坐标，利用生成的虚拟词汇作为未标注新词的坐标指引点，进行新词的分类和分级依据。

4.将分好类别和级别的词转化为特定的权重分数，在句子级、文章级和对象级上展示出来。

**2、现有技术的缺点是什么？针对这些缺点，说明本发明的目的。**（客观评价，现有技术的缺点是针对于本发明的优点来说的，本发明不能解决的缺点不必写；基于本发明能解决的问题写出发明的目的）

相似专利：

[1]崔路男,尹存祥,吴伟佳,韦庭,潘旭,雍倩,李云聪. 句子的情感极性分析方法、装置及存储介质[P]. 北京市：CN110738220B,2022-09-30.

目前类似的发明专利主要以分析句子情感极性为主，即，其本质是将句子的情感作分类，而后通过训练获取模型参数，输出句子的情感分数。

本专利相比之下具有以下的特点：

将句子的量化分析从情感词扩展到专业词库上，通过对专业词库进行分类或分级，可以获取诸如句子的暴力等级、敏感等级等一系列的句子量化方法。

云模型是可解释的，由于云模型的目的是实现定性评价和定量评价的转换，因此，云模型得到的每一个类别的词/级别的词的Ex、En和He三个指标，可以用作分析该类型的特点而加以分析和解释。

相比之下，极性分析属于对分词进行分类，而非分级的方法。本发明既可以作为极性分析的一种方法，也可以拓展为对不同分词设置权重指标，用以对句子级、文章级、对象级的文本内容开展特定内容的量化分析。

**3、本发明技术方案的详细阐述，应该结合示意图进行说明**（越详细越好，至少要提供2页；发明中每一功能的实现都要有相应的技术实现方案；所有英文缩写都应有中文注释；**所有附图都应该有详细的文字描述，以别人不看附图即可明白技术方案为准；**同时附图中的关键词或方框图中的注释都尽量用中文；方法专利都应该提供流程图，并提供相关的系统装置图；附图中各相关部件都要提供名称）。

本发明专利有以下几个模块组成：

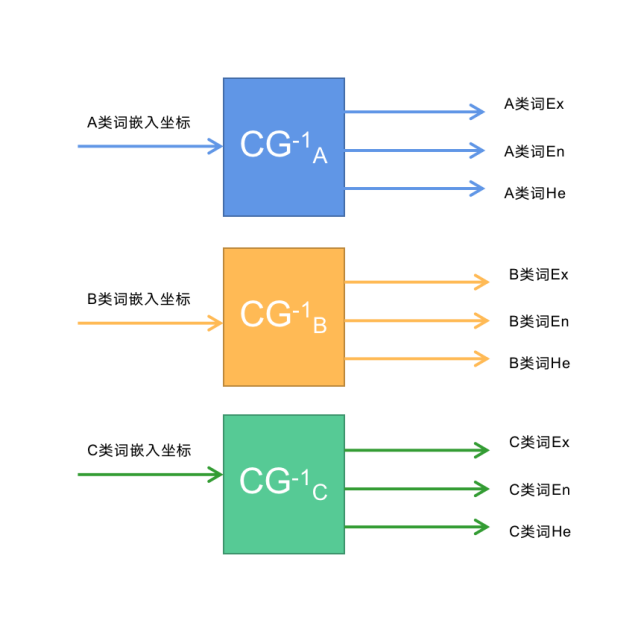
1. 基于云模型的词库分类模块。

读取准备词库和他们的词嵌入坐标。

人为设定一个词库分类或分级标准，如将敏感词分为：政治敏感词、历史敏感词、色情敏感词等类别，或将敏感词分为一级敏感、二级敏感、三级敏感词的级别，将暴力词分为隐形暴力词、显性暴力词，或一级暴力词、二级暴力词、三级暴力词的级别。

根据词库，人工编码若干各级别/类别的词语。以对暴力词的等级为例：对每一级别的暴力词汇标注和编码其所属的级别，这里要设置一个最小标注量。根据词库的大小不同，设置的最小标注量也不同，该标注量受到两个因素的影响，第一是准备词库的大小，第二是该类别可能的词量。如果准备词库越大，那么这个标注量也应该扩大，以获取该类别最精准的向量特征。

假定标注了3个级别的词。则原来的准备词库可以看成分割成了3个词库，分别为A、B、C级暴力词。将各个级别的暴力词输入逆向云模型中，如图所示。



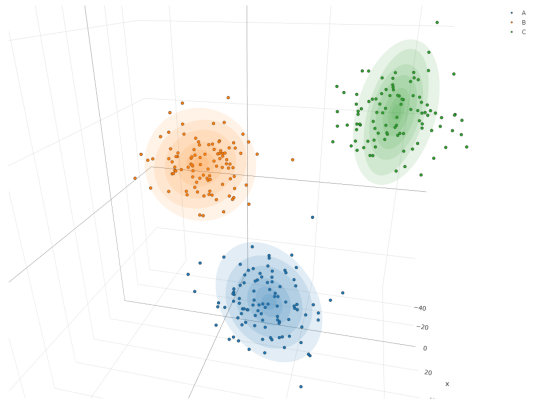
将每个类别中的词当作云滴，获取类别云在词嵌入空间的定性关系，即获得每个类别词的期望Ex、熵En和超熵He。在空间中描述三个类别暴力词的关系。

如：

A级暴力词

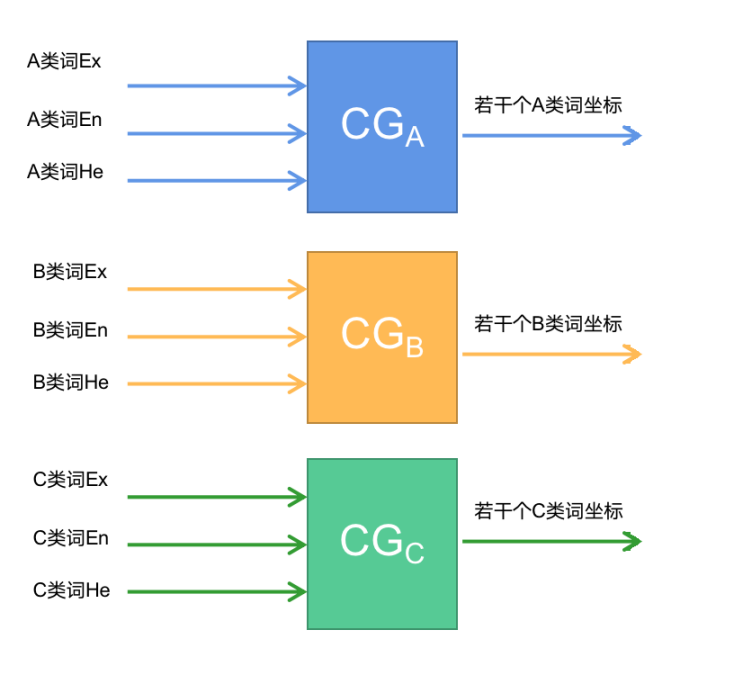
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 词1 | 0.1515213 | 0.81231 |  | -0.2139 |
| 词2 | -0.21559 | 0.854943 |  | -0.8547 |
| 词3 | 1.45412 | 0.543123 |  | 0.91326 |
| 词4 | -1.92351 | -0.23516 |  | 1.53341 |
| 词n |  |  |  |  |
|  | 0.68519 | -0.150234 |  | 0.855345 |
|  | -0.13514 | 0.53141 |  | 0.251234 |
|  | 0.512353214 | 0.52134231 |  | 0.651234123 |

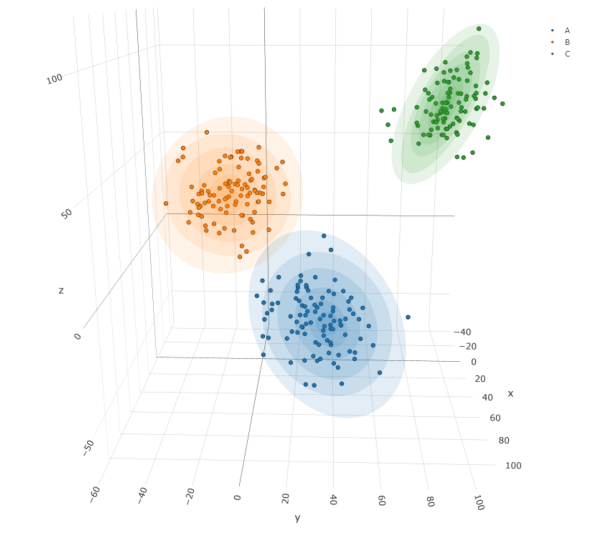
保存A词类别的期望Ex、熵En和超熵He，作为三个词类的定性特征。对于其他的词类也以此类推。生成各自的期望Ex、熵En和超熵He用以标度该类词在空间中的分布，下图展示了人工编码的部分词计算后在空间中的分布。



1. 新词标注模块。

* 对于原有准备词库中未标注词类型的新词K。
* 首先通过正向云模型，生成各词类型的n个词，以这些词作为各个词类的词坐标。下图展示了对A、B、C三类通过正向云模型**生成**的一些词。





* 用如下的方式，计算其与三个类别词之间的距离如图所示。C是正向云生成的词簇，K’是C中的词，|K-K’|是新词K与K’的距离。m是C词的均值点，n是C中词的个数。距离可以采用欧式距离或余弦距离的方式，需要根据语料库和词嵌入的训练过程做出选择。

最小距离：

最大距离：

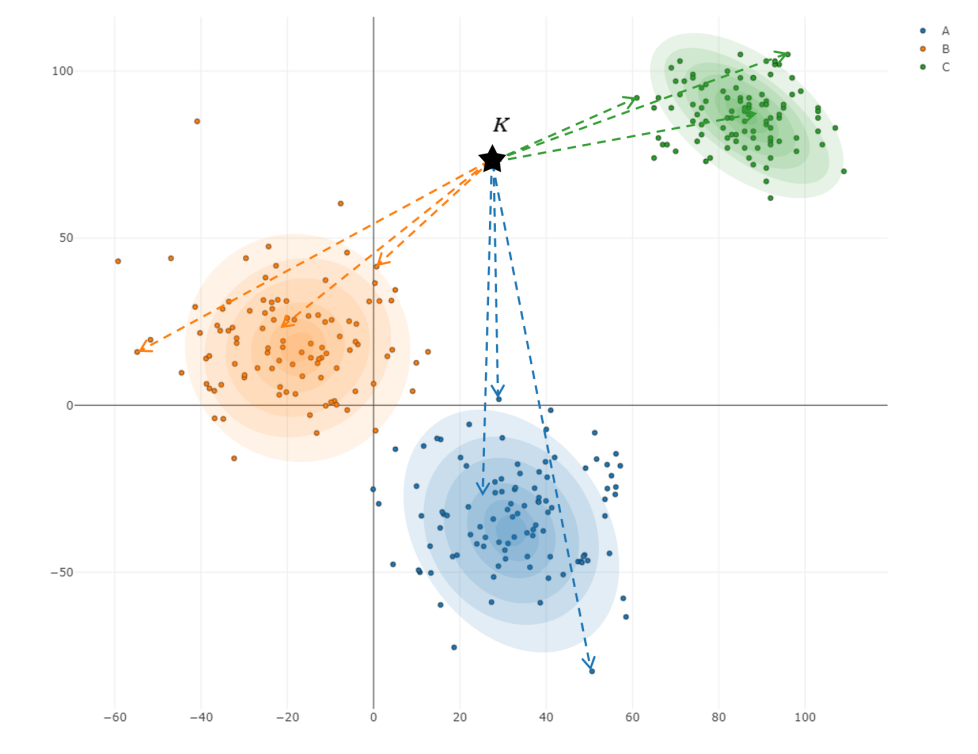
均值距离：

平均距离：

将上述四个距离组合成距离存储矩阵D：

可以使用上述单独某一个距离指标进行新词的位置判断，也可以将上述四个距离以线性方式组合成综合距离( Distance)，其中X是上述每个距离的权重形成的矩阵，根据具体应用场景完成定义,表示如下：

如图所示计算新词K与各个词类中的综合距离，将离K词最近的词类作为其词类自动标注，以判断新词属于哪一种级别的词。



基于此方法，可以实现对所有未做标注的词进行自动标注。

1. 文本内容评价（这一模块针对不同的词类或不同等级的词计算其分数，以暴力词库为例，可以分为不同等级的暴力词，这些词具有等级和程度上的差异，可以计算词汇级、句子级、文章级、对象级的暴力程度分数；也可以分为显性暴力词和隐形暴力词，这两类词没有等级或程度上的差异，可以计算词汇级、句子级、文章级、对象级的显性暴力分数和隐性暴力分数）

下面以暴力词库的分级为例子：

* 1. 词汇级的暴力分数计算

**n是词库的等级数，可以表示为下面的内容：**

* + 1. 固定值法：根据定性的编码，对每次级别的词各设定一个分数，：（如果n是分类的类别的画，则每一个级别固定同一个分数。）
    2. 随机值法（为了更准确地描述不同级之间的严重程度关系）

使用正态分布随机初始化n个的初始值，用以表示该词的分数：

其权重可以记为下面的矩阵：

保证矩阵中上三角部分的元素值大于等于1，自然地有。调整这些位置的元素值>1，即可定义不同级别词严重程度的关系，

* 1. 句子级的暴力分数计算

句子级的暴力分数计算可以表示为：

设有n个级类，是该句子中A级词出现的次数，是词汇级A类的暴力分数，是句子长度，是A词的词库个数，是词库的总个数。

以A类词的严重程度系数为例，可以表示为：

其中，

* 代表了这句话中A类暴力词词在某句子中出现的比例，如果该值越大，说明这句话的A类越严重
* 代表了这句话的A类词占到所有A类词库的比例，如果该值越大，说明这句话用了更多的A级暴力词词汇，暴力越严重
* 代表了A类词本身占所有暴力此库的比例，如果这个词越小，说明A类暴力词在所有词中比例很小，但还仍然被使用到，表示暴力越严重

对于其他的某一类词在句子中的严重程度系数，可以表示成下面的一个集合：

对于每一类词的严重程度系数，都有：

将某一类词在句子中出现的严重程度系数，与第I级词本身的指标输入Softmax函数，得到一个权重系数，表示如下：

是该级别词汇级的暴力分数，见**（1）词汇级的暴力分数计算，将该类词的**严重程度系数与词本身的暴力分数相乘，再除以完整句子的长度即可得到句子级的暴力分数。

以下表为例：

* 1. 文章级的暴力分数计算

将文章整体使用句子级的暴力分数计算，或采用以下的方式：

先计算文章的句子数为，使用句子级的暴力分数计算出计算单个句子的暴力分数。

* 1. 对象级的暴力分数计算

将文章级的暴力分数计算扩展到对象级，如：个人、群体或地区的暴力分数计算。先按照文章级的暴力分数进行计算，然后除以文本或文章个数，以社交平台某个用户的暴力分数计算如下

是用户发布的文章或评论数，计算每一个评论或文章的暴力分数，再求平均，以此类推，更获得更上层对象级的暴力分数，如用户集群、地区等。

非暴力（和平）分数：用以标识某一个句子的非暴力分数，以句子级的为例：

* 1. 比较分数

可以通过比较不同的，得到哪句话/哪篇文章/哪个用户更加暴力，另外，如果此方法针对的专业词库不是等级而是类别，就可以通过比较类别下不同粒度的S，来确定该句子/文章/对象更倾向于什么。

举例：以计算机专业词库为例，标注划分并细致分类（软件工程类词库、网络词库、人工智能词库），可以基于本发明，利用软件工程类的词来计算某句话与软件工程主题的倾向，利用网络词库来计算某句话与计算机网络主题的倾向，利用人工智能词库来计算某句话与人工智能主题的倾向，最后比较三个S值的大小，确定该句子最终更倾向与哪个主题。（由于这时候是一个对词库进行的分类操作，每个分词的权重分数值设置为相同的值带入词汇级的暴力分数计算的固定值法中。）