技术交底书撰写

发明名称：	一种基于词嵌入空间的专业词库搜索方法		
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：

本发明专利可用于在特定专业领域下的内容分析和自然语言处理工作，通过构想等方式获取少量的专业词汇种子词（seed words），在词嵌入空间中搜索其语义近义词以获取更丰富的专业词库。一方面，专业词库的构建有助于理解领域中热点和重点的组成，另一方面，专业分词库能够反哺提升专业领域文本分词准确率，有助于对专业领域的文档数据进行深化分析。

术语解释：

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。

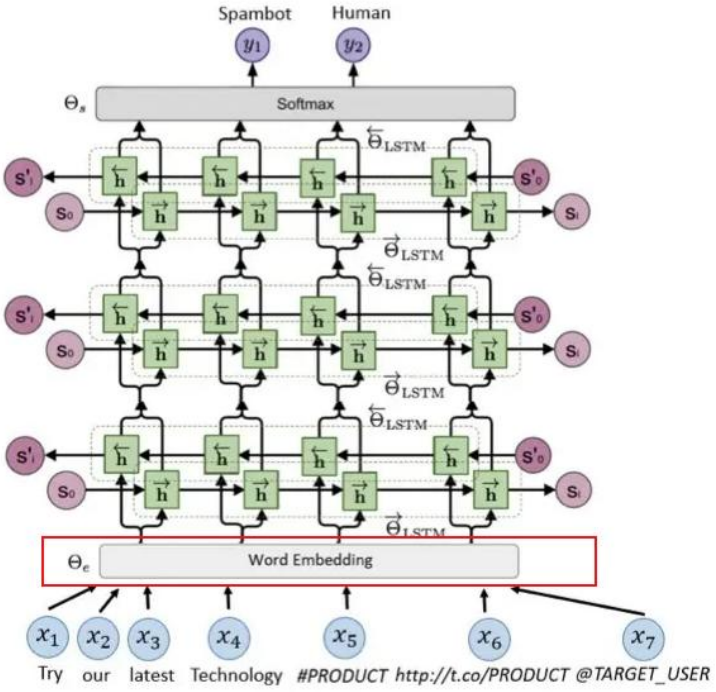
通俗的解释：词嵌入是根据大量文本根据其上下文信息训练出来的一种用向量表示词的方法。如简单的词嵌入假设是，根据人类自然语言中，相似词的意思上下文可能是一致的，如“渴”和“饿”：

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：



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）

目前，科学技术革命的快速发展主要是以计算的形式对数据进行处理。随着基础设施包括计算机网络、个人设备的迅速普及，大量信息在各种设备间传播，需要处理的信息量空前绝后。而如何从海量的信息中迅速找到我们关注的信息并加以分析，是现在研究人员面临的难题。

随着我国各专业的发展，专业性的文本数量的不断增加，采用纯人工方式对专业文本，特别是专业词汇，进行整理归纳和分析解读变得越来越困难。运用文本挖掘等计算机技术辅助开展专业领域的词汇发掘和内容分析具有重要意义。

词库是词汇的集合体，通常包括基本词库（通用词库）以及专业词库，如暴力词库、敏感词库、法政词库，危险品词库等，而本发明所构建的词库所针对的是某一领域的专业词库。其过程是，由人工构想少量的特定专业词库的种子词，通过在专业语料库上训练的词嵌入层中得到的分词嵌入坐标，搜索获取更深层次的专业词，形成专业词库。

目前最为接近的方法有：

巴桑卓玛,李苗苗,高定国.基于词嵌入的藏文情感词典的构建方法研究[J].电子技术与软件工程,2017(20):132-134.

该论文中，同样使用词嵌入作为词搜索空间，与本专利的区别如下：

1. 该文中使用的是 KNN 分类方法作为搜索的算法，仅仅只针对种子词寻找近邻，而未关注找到的近义词仍可进行递归查找下一代的专业词。

2. 该文中仅以特定的相似度阈值的大小作为判断是否收录进词库依据，忽略了对于空间中不同位置的分词，其相似度阈值应该加以变化和自适应的过程。

3. 该文中仅以特定的相似度阈值的大小作为判断是否收录进词库依据，忽略了词库搜索中对于分词搜索个数的参数，应该设置以单个分词进行近义词搜索的个数参数，并对该参数进行约束，以获取更多的专业词库。

4. 相比之下，本专利将搜索扩展到词库的更深层，并设置了阈值提高函数（用于每一代词库搜索时，提高收录阈值）、阈值适配函数（用于每一代词库搜索时，自适应调整阈值的变化）、约束函数（用于每一代词库搜索时，限制收录的词数）、约束适配函数（用于每一代词库搜索时，自适应要找多少个近义词）。

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）

现有研究构建的特定领域词库存在以下问题：耗费大量时间、依赖于现有词库、方法适应性较差、不能很好地移植到其他领域。 为解决上述不足，本文利用词嵌入计算构建适应于某一专业领域的词库， 本文构建专业词库的方法主要具有以下优势：

（1）耗时短，本发明通过收集少部分的种子词，并通过少量的编码标注工作，经过大语料库的词嵌入技术，构建丰富的同类型词库；

（2）精准率高，经由阈值提高函数、阈值适配函数和约束函数控制得到的词库相比于一般基于词嵌入的词库搜索办法更加精确，对于每一个分词都能找到较为合适的近义的专业分词。

（2）方法可以很好地应用到其他领域，现有的词库构建技术大部分集中于情感词库以及敏感词库等，未能迁移到其他领域，而本发明提出的基于词嵌入技术构建词库的方式能够很好的运用到其他专业领域；

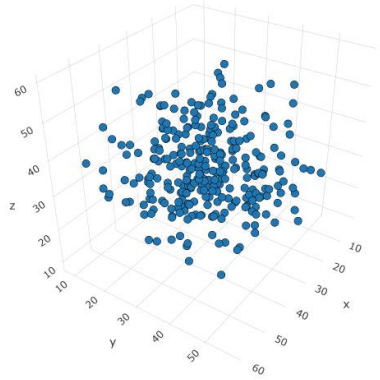
3、本发明技术方案的详细阐述，应该结合示意图进行说明（越详细越好，至少要提供 2 页；发明中每一功能的实现都要有相应的技术实现方案；所有英文缩写都应有中文注释；**所有附图都应该有详细的文字描述，以别人不看附图即可明白技术方案为准**；同时附图中的关键词或方框图中的注释都尽量用中文；方法专利都应该提供流程图，并提供相关的系统装置图；附图中各相关部件都要提供名称）。

有以下几个模块组成：（下文以暴力语言专业词库为例）

1.基于词嵌入的词库发现模块

- **第一步：获取所要构建专业词库的种子词**。如想要构建暴力专业词库，则通过观察、构想、浏览社交聊天文字、翻阅字典等方式寻找若干个暴力种子词，这一步骤中，寻找的种子词范围要尽可能地多，以暴力词的种子词为例：既选取和性歧视相关的暴力词，又要选取若干暴力的英文首字母缩写词，又选取和民族有关的若干暴力词。如想要构建情感词库：既要选择正向情感词，又要选取负向的情感词。这是为了在后续的词嵌入空间中尽可能地获取更广泛的词类。
- **第二步：选取或训练基于实际应用场景的词嵌入库**。这一步骤采用自己训练或采

用公开的、合适于具体应用的词嵌入库，如：如果是想要对社交语境下对暴力词进行分类或分级，可以选择一个有尽可能多的社交语言和暴力词语料参与的词嵌入库。词嵌入的构建方式有多种多样，既可以针对不同的语料库进行选择，也可以采用公开的中文词嵌入库，作为如图所示的**词嵌入搜索空间**，其中每一个节点代表一个分词。



- **第三步：定位种子词在词搜索空间中的位置和向量信息，删除不存在词嵌入空间中的分词。**有些构想、观察获取到的词可能受限于分词工具的限制，无法在词嵌入中找到对应的分词，无法作为种子在词搜索空间中开展搜索，因此删除这些人为构想的种子词。

例如：“混蛋”这一种子词在词搜索空间中存在，则保留。“八嘎”这一人为构想的种子词在词搜索空间中不存在，则删除。

- **第四步：词库的扩展。根据具体的种子词个数、词嵌入词量大小级，进行词库的搜索与扩充，最终形成专业词库。**计算与所选种子词余弦相似度，根据以下的几种方式综合进行词的扩展形成所需要的词库：

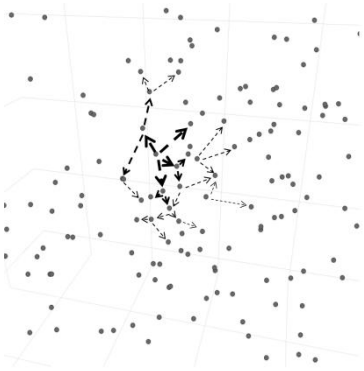
  方法一、二适用于种子词较少的情况下，期望获取尽可能多地近义词时：

  - 方法一：
    - 设一个定值 n 并给定一个迭代次数 i，
    - 找到当前种子词的 n 个近义词后，将这些词保存为“一代词”
    - 迭代这些“一代词”的 n 个近义词，保存为“二代词”，迭代查找这些二代词的 n 个近义词，保存为“三代词”，以此类推。当 n 为 3 时，该过程表示如下图所示。

- 当找到第 i 代词或专业词库足够充裕（词量到达某一定值）的时候停止

- 使用方法一在不排除重复词的情况下能找到$\sum_{k=0}^{i} a \times n^k$个词，a 是种子词的个数

- 方法二：
  - 设一个定值p作为余弦相似度的阈值，$p \in (0,1)$
  - 给定一个迭代次数 i，计算当前种子词余弦相似度>p 的那些近义词，将这些词保存为“一代词”
  - 迭代这些“一代词”余弦相似度>p 的近义词，保存为“二代词”，利用余弦相似度查找这些二代词的近义词，保存为“三代词”，以此类推
  - 当找到第 i 代词或专业词库足够充裕（词量到达某一定值）的时候停止，其扩展过程如图所示。



与方法一相比，这种方法找到的某一代的词数量是不确定的，更加考虑到了词和词之间的相似性。

以下的两种方法适用于种子词较充足的情况下，期望获取准确度更高或更精确的同类型词时：

- 方法三：

地址：广州市黄埔大道西 100 号富力盈泰广场 A 塔 910
联系人：谭秀芳　　电话：183 2008 3865　　Email：tanxiufang@jiaquanip.cn
网址：　www.jmpatent.com.cn

- 方法三可以理解为方法二的一个改进
- 给定一个值 p 作为余弦相似度的阈值，$p \in [0,1]$。
- 给定一个迭代次数 i

给定一个**阈值提高函数**f，该函数能逐代提高阈值（阈值提高函数的目的：查找词越到后面，相似度要求越高），这一函数的设计是不固定的，根据具体的案例来设置，本专利举出两个例子，因此这种函数的设计不是一个定式，不限于下述的两个函数：

- 例如用 logistics 函数表示的阈值提高函数，n 是第几代词，n 越大，p 值就会越大，用以提高

$$f_{increase}(p) = \frac{1}{1+e^{-n}},\ n \le i$$

- 另一种能够根据上一代的阈值 p 值提高的阈值提高函数f，展示如下：

$$f_{increase}(p) = p + \frac{1-p}{n},\ n \le i$$

阈值提高函数的效果展示为下面这个例子：

种子词	一代词个数	二代词个数	三代词个数	四代词个数	五代词个数	六代词个数
A	23	11	7	4	3	0
	一代阈值	二代阈值	三代阈值	四代阈值	五代阈值	六代阈值
	0.5	0.75	0.833333333	0.875	0.9	0.916666667

- 阈值提高函数的另一种变体是**阈值适配函数**$f_{adaptive}(p)$，可以根据一个词与周围词的相似度进行匹配。适用于下面所述的场景：

  其中,p'是某一代相似度最高的 k 个词与当前父亲词的平均相似度：

$$p' = \frac{\sum_1^k p}{k}$$

- 设计平均相似度的目的是：用于判断这个词与周围词的位置是否更加相似，如果某个词最近的一些近义词相似度本身就不是很高，阈值适配函数应该进行阈值降低以找到更多的近义词；相反，如果某个词最近的一些近义词相似度本身过高（有大量相同含义的重复的分词），阈值适配函数应该进行阈值的**提高**以找到更精确的近义词。
- 因此，这一函数的设计是不固定的，根据具体的案例来设置，本专利给出了一个适配函数的例子，但不应限于下述的函数：

  假设当前代词最终设计的相似度为 p，则需要先计算 p 的增量Δp，用以表示阈值 p 在当前代应该产生变化趋势，然后作用在当前代的阈值 p 上。

$$\Delta p = \frac{(p' - p)}{p}$$

$$f_{adaptive}(p) = \Delta p + p = \frac{p' - p + p^2}{p}$$

种子词	一代词个数	二代词个数	三代词个数	四代词个数	五代词个数	六代词个数
A	7	15	8	3	8	7
每一代最近 10 个词的平均相似度	0.85	0.65	0.75	0.93	0.78	0.81
初始阈值	一代阈值	二代阈值	三代阈值	四代阈值	五代阈值	六代阈值
0.75	0.88	0.61918239	0.830457149	0.950322275	0.77109647	0.821548692

- 计算当前种子词余弦相似度>p 的近义词，将这些词保存为“一代词”，计算该代的余弦相似度阈值 p，带入阈值提高函数或阈值适配函数f，f函数的返回值与词的代数增长可以在部分阶段呈现相关的关系，用以控制专业词搜索的比例和数量。
- 逐步减小并精确在词嵌入空间中的搜索数量。迭代这些“一代词”中大于新的阈值 p 的近义词，保存为“二代词”，循环迭代，以此类推。
- 当找到第 i 代词或专业词库足够充裕（词量到达某一定值）的时候停止，或词语都因为设定的阈值太高而无法再找到新的词时停止迭代。

- 方法四：
  - 方法四在方法一基础上进行修正和约束
  - 设一个值c_0作为初始近义词搜索个数，给定一个约束函数$f_{restriction}$，计算离当前种子词最近的c_0个近义词，将这些词保存为“一代词”，输入函数$f_{restriction}$计算该代的近义词个数c_1，该 f 能逐步减小和约束c的大小，以逐步减小并精确在词嵌入空间中的搜索数量。本专利展示了构想的一个约束函数如下，但不限于该函数：其中，Q 是自定义的约束系数，控制了将原有的搜索个数减小的程度，$0<Q<c_{m-1}$

$$f_{restriction}(c_m) = c_{m-1} - Q，\ m \geq 1$$

即，每一代词都比上一代词少找C个，或另一种约束函数的方式：

$$f_{restriction}(c_m) = \frac{c_{m-1}}{n}$$

下表展示了上述约束函数的效果：

种子　一代词要找　二代词要找的　三代词要找的　四代词要找的　五代词要找的

词	的个数	个数	个数	个数	个数
A	50	25	8.333333333	2.083333333	0

另外，如方法三的阈值适配函数类似地，约束函数也可以构建其**约束适配函数**，即考虑到当代词最高相似度关系以适配最优的c大小。本专利展示了一个适配函数如下：

p_i'是某一代相似度最高的 k 个词与当前词的平均相似度，表示如下：

$$p' = \frac{\sum_1^k p}{k}$$

先使用方法一，设置一个定值 n 找寻种子词的 i 代词中，计算这些代前 k 个近义词的平均相似度得到 p'，用于度量近 i 代的阈值变化情况。

$$Z = \frac{\sum_1^i p'}{i}$$

Z 是设置的相似度观察阈值，用以判断当前代的最高阈值或平均阈值到达何值时，需要多收录一些词。

g_i是约束系数，用以控制当前代要多找几个词还是少找几个词，定义如下：

$g_m = |p_m' - p_{m-1}'| * r + g_{m-1}$,r 是控制当前相似度要增长的权重，默认为 10

最后将该约束系数作用在当前词所要找的 c 个词上

$$f_{adaptive}(c_m) = \begin{cases} c_m - g_m, p' \geq z \\ c_m + g_m, p' < z \end{cases}$$

- 迭代搜索这些“m 代词”的c_m个近义词，保存为“m+1 代词”，
- 当找到第 i 代词或词库足够充裕（词量到达某一定值）的时候，或者直到所有词语都因为设定的阈值太高而无法再找到新的词时停止迭代。

以数据为例：

	C0	C1	C2	C3	C4	C5	C6	C7
专业词 1	5	6.18	4.18	9.294201672	7.294201672	9.508661796	7.508661796	5.508661796
判断 Z 和 p’的关系		p’<z	p’>z	p’<z	p’>z	p’<z	p’>z	p’>z
约束系数 g		1.18	2.93	5.114201672	3.30194499	2.214460124	0.962104888	3.989101542
Z	0.59	0.63	0.76	0.69	0.50	0.98	0.60	0.54
p’	0.68	0.73	0.95	0.53	0.81	0.67	0.64	0.98

另外，亦可以使用上述方法进行组合，结合相似度阈值和近义词个数两个指标进行精确地词库构建，最后，将重复收录的词删去。

至此，获取了一个给定特定种子词的专业词库，以暴力词库为例。以上的工作通过收集少部分的种子词而经由大语料库训练的词嵌入技术，构建了丰富的同类型专业词库。