Computer Science and Application 计算机科学与应用, 2020, 10(6), 1194-1208
Published Online June 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.106124

Hans汉斯

Document Embedding Dimension Homeomorphism

—Government Work Report Analysis

Hualun Xie[1], Xun Liang[1], Mengdie Li[2]

[1]School of Information, Renmin University of China, Beijing
[2]Government Service Center of Kaifu District, Changsha Hunan
Email: xhlgogo@foxmail.com

Received: May 27th, 2020; accepted: Jun. 5th, 2020; published: Jun. 12th, 2020

Abstract

The intelligent analysis of big data in government work reports can quickly and fully grasp the correlation of various internal factors and support decision maker to complete reasonable judgment. Taking the 18-year government work report as research target, which consists of 31 provinces and municipalities at the county level and above after 2000, this paper first proposes the optimal embedding dimension analysis framework based on document embedding homeomorphic space. And then on this basis, the paper studies the separability and similarity of provincial documents of the best document embedded dimension homeomorphism model of government work report. Finally, the experimental results and analysis of the document clustering separability and similarity difference in the government work report are given. The optimal document embedding vector obtained by the model can effectively divide the provincial subspace of the government work report, and the differences in the similarity of the provincial time series of the local government work report highlight their gaps in politics, economy, education, culture and so on. The experiment also found that the Euclidean distance using the regularized government document vector is equivalent to the traditional cosine distance. The document embedding analysis framework not only has certain reference significance and application value for the construction of smart government affairs in China, but also can support deep information mining, report reinterpretation and intelligent decision making in multi-category documents such as public company announcements.

Keywords

Government Work Report, Document Embedding, Homeomorphism, Dimension

文档嵌入维度同胚

——政府工作报告实例分析

谢华伦[1]，梁　循[1]，李梦蝶[2]

文章引用: 谢华伦, 梁循, 李梦蝶. 文档嵌入维度同胚[J]. 计算机科学与应用, 2020, 10(6): 1194-1208.
DOI: 10.12677/csa.2020.106124

[1]中国人民大学信息学院，北京
[2]湖南省长沙市开福区政务服务中心，湖南 长沙
Email: xhlgogo@foxmail.com

收稿日期：2020年5月27日；录用日期：2020年6月5日；发布日期：2020年6月12日

摘 要

对政府工作报告大数据的智能分析，可以快速且充分地掌握其内在各因素的关联，支持决策者完成合理的决断。本文以2000年后共18年的全国31个省和直辖市县区级及以上的政府工作报告为分析对象，首先提出了基于文档嵌入同胚空间的最佳嵌入维度分析框架，在此基础上对政府工作报告的最佳文档嵌入维度同胚模型进行了省域文档可分性和省域文档相似性研究，最后给出了政府工作报告的文档聚类可分性和相似性差异的实验结果和分析。该模型得到的最佳文档嵌入向量能够有效地对政府工作报告的文档省域子空间进行划分，各地方政府工作报告的文档省域时间序列相似性差异凸显了它们在政治、经济、教育、文化等多方面的差距，实验同时发现在求解相似文档集上使用正则化后的政府文档向量的欧式距离能够等效于传统的余弦距离。本文提出的文档嵌入同胚分析框架不仅对我国智慧政务的建设具有一定的参考意义和应用价值，同时可以在上市公司公告等文档多分类任务中对深度信息挖掘、报告再解读和智能决策提供支持。

关键词

政府工作报告，文档嵌入，同胚，维度

1. 引言

中央集权制为我国当前的政治现状，这一现状决定了我国大陆地区 31 个省、直辖市、自治区的最高一级行政单位在各自管辖的行政区划内具有高度的政策权威，其辖区内的州、市、盟、县、旗、区在政府工作上与省部级的政策指令高度看齐，这为地方政府工作报告按省域一级文档可分奠定了现实基础，计算机算法的进步为分析我国县(区)级及以上的政府工作报告文档大数据奠定了技术基础。

对政府工作报告的已有研究中虽部分使用了自然语言处理及文本分析方法，但在其做后续分析时仍存在数据来源较小且仅限于中央政府工作报告、研究不深入、方法较单一以及内容分析上的单方面性和主观性等一些不足之处。结果反映的往往是一组政府报告在全部表达内容下，或全部政府工作报告在某些方面的相关性[1]。然而在政府公文中，很少有像政府工作报告这样一篇文档表达多重内容的综合性论述报告，使用基于人工智能的大数据智能分析来研究它们在时间尺度上的传承性、空间维度上的相似性和内在各因素的关联性，对于全面解读各地方政府的工作报告具有重要意义。

在拓扑学中，同胚(homeomorphism)是两个拓扑空间之间的双连续函数，拓扑空间范畴中的同构[2]。也就是说，它们是保持给定空间的所有拓扑性质的映射。如果两个空间之间存在同胚，这两个空间就称为同胚的，从拓扑学的观点来看，两个空间是相同的。本文中的同胚，是指同一文本使用相同的训练模型或经过相同的处理过程，得到的具有不同训练参数但具有相同特征数的文本的计算机表示形式。

本文首先获取除港澳台以外、2000 年以后的 31 个省、州、盟、市、旗、区和县级政府工作报告，在 Gensim 平台上使用 Doc2Vec 方法进行文档向量的分布记忆模型(distributed memory model of paragraph vectors, PV-DM)训练报告文档大数据，通过无监督的方式得到文档的深度学习(Deep Learning, DL)向量空间表示，进而使不同维度的政府工作报告具有相同的文档嵌入向量空间同胚表示，找到最优的政府工作报告文档嵌入维度，然后使用基于分类和回归树(classification and regression tree, CART)的 XGBoost 算法进行有监督的省域政府工作报告向量空间的划分，接着对最佳嵌入维度空间的政府工作报告进行相似性分析，最后通过流形学习(manifold learning, ML)中的 T 分布随机近邻嵌入(T-distributed stochastic neighbor embedding, T-SNE)非线性降维算法，将高维的政府文档向量降到二维，得到政府工作报告的文档可视化结果。

针对政府工作报告，本文首先通过文档嵌入维度同胚分析模型，结合无监督的深度学习将其向量化，然后使用有监督的机器学习对同一地方政府工作报告的时间序列传承性进行分析，接着得到不同地方政府工作报告之间的同地相似性、同市相似性和异省相似性，进而去分析产生该相似性差异的可能原因，如地方政府政治地位、政府官员的来源、实体经济产业联系、高等教育资源分配差异、语言文化用词风格差异等。这种先使用无监督的深度学习再使用有监督的机器学习的方式，既能够有效规避不同嵌入维度的政府文档嵌入矩阵不同、文档向量优劣性不好比较的问题，得到的政府文档最佳嵌入维度，又可以有效防止文档向量化过程中因嵌入维度过小而引起的信息丢失和因嵌入维度过大包含大类噪声。基于人工智能模型，本文对政府工作报告文本大数据进行了分析，其内在各因素的关联分析结果具有较好的经济和政策借鉴意义。同时该文档嵌入维度同胚分析框架可以应用在其他政府公文、行业分析报告、上市公司公告等文档多分类问题中，对文档的深度信息挖掘、分类再解读和智能决策提供基于人工作智能的自然语言处理支持。

2. 政府文档嵌入维度分析

自然语言处理中，需要将词和句子表示成计算机能够处理的形式。传统的 one-hot 独热表示仅仅将词符号化，不包含任何语义信息。Harris 在 1954 年提出的“分布假说”为这一设想提供了理论基础：上下文相似的词，其语义也相似。Firth 在 1957 年对分布式假说进行了进一步阐述和明确：词的语义由其上下文决定[3]。分布式表示的较大优点在于它具有非常强大的表征能力，比如 n 维向量每维 k 个值，可以表征 kn 个概念。

Bengio 提出一个自然语言处理的大规模深度学习模型，这个模型能够通过学习单词的分布化表示来获得上下文信息[4]。基于神经网络的分布式表示一般称为词向量、词嵌入(word embedding)或分布式表示(distributed representation)，核心依然是上下文的表示以及上下文与目标词之间的关系的建模(见图 1)。相比于词的独热表示，一是将词向量的每一个元素由整型改为浮点型，变为整个实数范围的表示，二是将原来稀疏的巨大维度压缩嵌入到一个更小维度的空间。

词嵌入是深度学习模型应用到自然语言处理中不可或缺的一环：神经网络的输入是连续空间中的向量，而自然语言运用的则是离散的字符，词嵌入将自然语言转换成空间中的向量，是连接两者不可或缺的桥梁。维度(dimensionality)是词嵌入模型里最重要的一个超参数。斯坦福大学的研究者 Yin 和 Shen 提出了一个理解词嵌入维度的理论框架[5]。

Milokov 等人提出了 Word2Vec 词嵌入算法，Google 将其于 2013 年开源，采用了 CBOW (continuous bag-of-words model)和 skip-gram (continuous skip-gram model)两种模型与负采样和层次损失函数两种方法[6]。随后的 2014 年，Le 和 Mikolov 提出了文本向量的深度学习算法 Doc2Vec 文档嵌入(document embedding)，借鉴 Word2Vec 思想的两种结构：PV-DM 和 PV-DBOW，分别对应 Word2Vec 中的 CBOW 和 skip-gram [7]。Doc2Vec 在神经网络的输入层增加了一个唯一的文档(段落)ID，在计算句子向量的同时算法也计算出词向量(见图 2)。在图 2 中，政府文档用矩阵 D 中一行唯一表示，词用 one-hot 词汇表 W 中一行唯一表示。

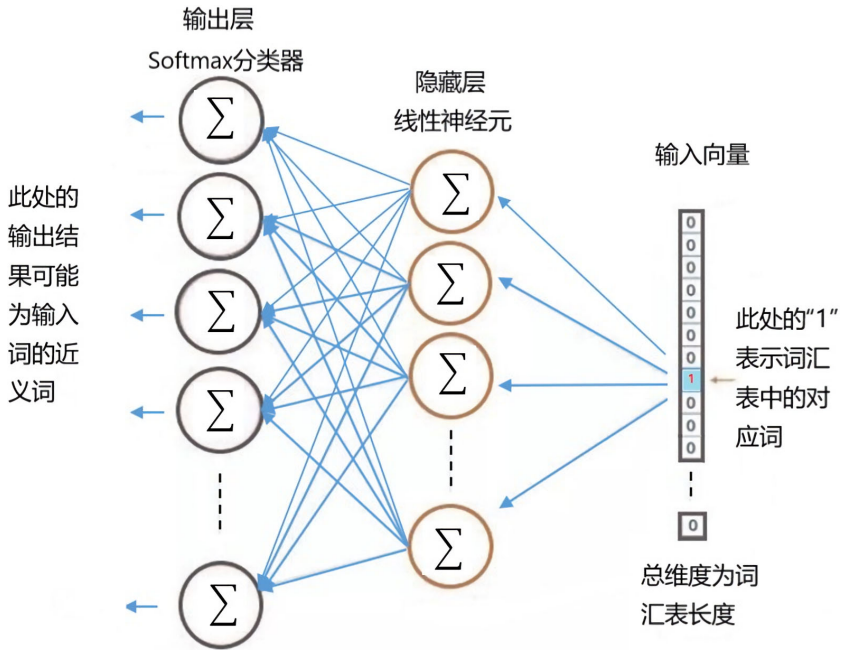

Figure 1. Word vectors are hidden layer parameters when training neural networks
图 1. 词向量是训练神经网络时的隐藏层参数

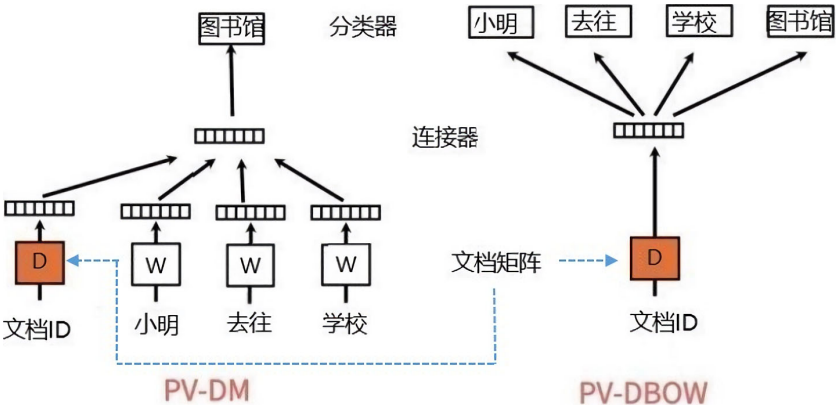

Figure 2. Two structures of Doc2Vec
图 2. Doc2Vec 两种结构

2.1. 文档嵌入的几何性质与几何不变性

几何模型中的文档向量的几何性质与文档的表达内容存在着对应关系[8]：

- 相似性。文档的相似性可以通过向量夹角的余弦值来表示。对于两个文档向量 V_i 和 V_j，它们夹角的余弦值 $\cos\left(V_i, V_j\right)$ 的范围在−1 到 1 之间。这个值越大，代表第 i 个和第 j 个文档之间的相似度越高；
- 类比性。类比性考察的是两篇文档之间作者是否相同。文档类比这一语义性质对应的几何性质为两对文档向量的差的平行度；
- 聚类性。相似文档向量之间的距离总是较小，因而这些文档会倾向于聚成一类。

文档嵌入向量的夹角关系、平行关系和距离关系具有酉不变性，文档嵌入向量空间的旋转、等量伸缩、镜像等整体变换并不影响这种几何关系。酉变换并不改变向量之间的相对位置。评价文档的相似、作者等语义学任务时，由于文档嵌入向量空间的几何性质具有酉不变性，因此将文档向量矩阵全部单位

化后向量夹角关系不变。

2.2. 政府文档嵌入的损失函数

定义 1 (文档嵌入同胚 DVM)：为使不同嵌入维度的相同文档集具有相同的同胚表示，给定文档嵌入向量集 $E \in R^{n\times d}$，定义其文档嵌入同胚为：

$$\mathrm{DVM}(E) = EE^{\mathrm{T}}$$

其中 n 为文档数，d 为嵌入维度，E^{T} 为 E 的转置矩阵。文档嵌入同胚 EE^{T} 的第 i 行是一个 n 维向量 $\left((v_i,v_1),(v_i,v_2),\cdots,(v_i,v_n)\right)$，该向量可以看成是 v_i 的相对坐标，其参照物是 v_1 至 v_n。

定义 2 (文档嵌入损失 DVML)：给定相同文档集、不同维度的文档嵌入 E_1 和 E_2，定义其文档嵌入损失为：

$$\mathrm{DVML}(E_1,E_2) = \left\|\mathrm{DVM}(E_1) - \mathrm{DVM}(E_2)\right\| = \left\|E_1E_1^{\mathrm{T}} - E_2E_2^{\mathrm{T}}\right\| = \sqrt{\sum_{i,j}\left(\left(v_i^{(1)},v_j^{(1)}\right) - \left(v_i^{(2)},v_j^{(2)}\right)\right)^2}$$

文档嵌入损失通过测量 E_1 和 E_2 中同一文档向量的相对坐标漂移，从而消除对于特定文档嵌入坐标系的依赖。注意到文档嵌入损失 DVML 的酉不变性，如 U 为一个 n 阶酉矩阵，I_n 为 n 阶单位阵，满足 $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_n$，$E_2$ 是 E_1 的一个酉变换，满足 $E_2 = E_1U$，则 $\left\|E_1E_1^{\mathrm{T}} - E_2E_2^{\mathrm{T}}\right\| = 0$。

3. 政府文档嵌入流形学习分析

在现实中高纬度的文档向量数据点可能并不分布在欧式空间，使得非线性高纬度文档嵌入数据难以在传统欧式空间中度量，因此引入新的数据分布流形空间用于文档嵌入数据的空间分布。流形(manifold)的局部具有欧几里得性质，是欧式空间中几何概念的推广，文档嵌入的向量空间通常具有在高纬度空间中实数“充盈”，传统的欧几里得度量方式不再适用于文档嵌入的向量空间，但完全摒弃欧几里得的数学度量方式、创造性地开创一种新的空间度量方式并未在数学上取得突破性进展，因此，在对高纬度、实数域的文档嵌入向量空间进行分析时，引入局部具有欧几里得性质的流形的数学概念，变得实际且可行。

3.1. 高维空间政府文档嵌入

高维度的政府工作报告文档嵌入向量空间中，任意两个文档向量之间的距离差异较小，欧式距离失效，基于欧式距离、建立最近邻算法图来近似流形的方法在高维空间失效[9]。由于数据内部特征的限制，一些高维中的数据会产生维度上的冗余，实际上只需要比较低的维度就能唯一地表示[10]。

政府工作报告文档嵌入本质上是文本在高纬度空间的实数域上的向量表示，其作为典型的几何模型，其具有流形学习所需要的数据特征，因此，将政府工作报告文档嵌入向量之间的夹角余弦值定义为它们在高维实数空间中的距离，即可对政府工作报告文档嵌入进行文档相似性评价，为政府工作报告省域文档相似性奠定了数学基础，并且这种文档向量流形空间所具有的局部欧几里得性质也为使用 T-SNE 进行流形学习降维从而实现可视化提供了依据。

3.2. 政府文档向量距离

定义 3 (文档余弦距离 DCOS)：为使在高维空间中表征不同文档嵌入向量之间的距离，定义其文档余弦距离为：

$$\mathrm{DCOS}(\boldsymbol{A},\boldsymbol{B}) = \frac{\sum_{i=1}^{n}\left(A_iB_i\right)}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}$$

其中， $\boldsymbol{A},\boldsymbol{B}$ 为不同文档嵌入向量。

定义 4 (文档欧式距离 DEUC)：为使在高维空间中表征不同文档嵌入向量之间的距离，定义其文档欧式距离为：

$$\mathrm{DEUC}(\boldsymbol{A},\boldsymbol{B})=\sqrt{\sum_{k=1}^{n}\left(x_k^*-y_k^*\right)^2},\quad x_k^*=\frac{x_k}{\sqrt{x_1^2+x_2^2+\cdots+x_n^2}}$$

其中， $\boldsymbol{A}(x_1,x_2,\cdots,x_n),\boldsymbol{B}(y_1,y_2,\cdots,y_n)$ 为不同文档嵌入向量， y_k^* 与 x_k^* 为 $\boldsymbol{A},\boldsymbol{B}$ 向量正则化后的第 k 个坐标。注意到：

$$\begin{aligned}\mathrm{DEUC}(\boldsymbol{A},\boldsymbol{B})&=\sqrt{\sum_{k=1}^{n}\left(x_k^*-y_k^*\right)^2}\\&=\sqrt{\sum_{k=1}^{n}\left(x_k^*\right)^2+\sum_{k=1}^{n}\left(y_k^*\right)^2-2\left(\sum_{k=1}^{n}x_k^*y_k^*\right)}\\&=\sqrt{2-2\mathrm{DCOS}(\boldsymbol{A},\boldsymbol{B})}\end{aligned}$$

任意两个 n 维文档嵌入向量 $\boldsymbol{A},\boldsymbol{B}$，其正则化后的标量之间的欧式距离区间为[0,2]，点坐标落到单位 $n-1$ 维球面 $S^{n-1}=\{x\in R^n:\|x\|=1\}$（$n$ 维空间内的 $n-1$ 维流形）上，其单位化后的向量之间的余弦距离区间为[−1,1]，位于 n 维曲面包含的空间里，欧式距离与余弦距离在半个 $n-1$ 维球面上 S^{n-1} 上具有一一映射关系且线性单调。n 维文档嵌入向量的欧式距离与余弦距离相同之处在于，都可以表征文档之间的相似度，而它们的不同之处在于，欧式距离越小则文档越相似，而余弦距离越小则文档越不相似。

4. 政府工作报告文档嵌入维度

从政府文档向量模型训练后得到的政府文档嵌入同胚，求解政府文档嵌入损失，对政府工作报告作出最佳嵌入维度的选择。

4.1. 政府文档嵌入滑动窗口大小的选择

滑动窗口是指在训练政府文档嵌入向量时，每次训练输入到神经网络的词的数量。通用文档嵌入参数同胚模型，是一个用于求解文档嵌入参数的计算机算法，它在事先计算好的文档嵌入同胚集合(也即多个文档嵌入同胚矩阵)上遍历每个同胚向量，调用文档嵌入损失数学公式，计算一个文档嵌入同胚和其他文档嵌入同胚之间的文档嵌入损失后累加，得到该文档嵌入同胚的文档嵌入损失并添加到字典，最后字典里最小的值所对应的键即为最优的文档嵌入同胚，该文档嵌入同胚所使用的参数即为最优参数。

模型 1(通用文档嵌入参数同胚模型)	**模型 2(政府文档嵌入维度同胚模型)**

输入：文档嵌入同胚集合{DVM}
输出：文档嵌入参数
result= {} //储存计算结果的字典
foreach $E_i\in\{\mathrm{DVM}\}$: //一层循环
temp=[] //单个 DVML 临时序列
foreach $E_j\in\{\mathrm{DVM}\}$://二层循环
add $\mathrm{DVML}(E_i,E_j)$ to temp
//添加 $\mathrm{DVML}(E_i,E_j)$ 计算结果到临时序列
add (E_i:sum(temp)) to window
//添加 E_i 为键、temp 序列和为值的到结果字典
parameter = min(result).key
//排序，得到最优文档嵌入参数
	输入：政府文档集合{Documents}，预估区间 Interval
	输出：政府文档嵌入维度
	{Mini-Doc}⊆{Documents}//抽取政府文档集的子集
	foreachvec-size in Interval//单层循环，
	E = Doc2vec({Mini-Doc},vec-size)
	//调用 Doc2vec 文档向量化算法训练不同维度政府文档向量
	$\mathrm{DVM}(E)=EE^{\mathrm{T}}$ //计算政府文档嵌入同胚
	add DVM to {DVM}
	//添加到政府文档同胚集合
	bestvec=function({DVM})
	//调用通用文档模型 1，计算预估区间内最佳文档嵌入维度

固定相同政府文档嵌入维度为 800，滑动窗口大小取[5,30]，得到政府文档嵌入同胚序列，在该序列上运用模型 1，得到政府工作报告文档嵌入的最优滑动窗口为 17。

4.2. 政府文档嵌入维度大小的选择

政府文档嵌入维度的选择是指 Doc2Vec 训练政府文档向量时神经网络的维数，即得到的政府文档向量的维度数，采用先粗后细的方式。政府文档嵌入维度同胚模型，是一个用于求解政府工作报告最优文档嵌入维度的计算机算法，它直接输入经过预处理(中文分词、去标点、去停用词、去数字、去人名地名等)的政府工作报告文本，在预先设定的维度区间内，调用 Doc2Vec 文档向量化算法和通用文档嵌入参数同胚模型，得到最优文档嵌入同胚，其对应的文档嵌入维度即为该预设区间内的政府工作报告最优文档嵌入维度。

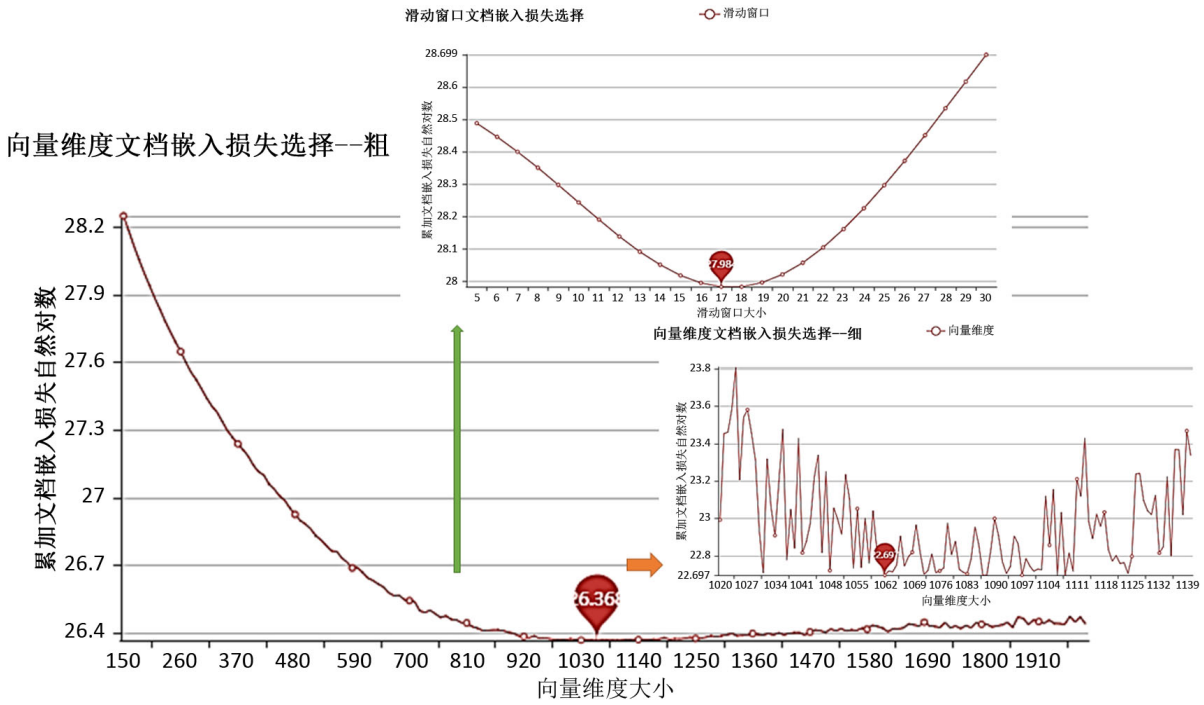

Figure 3. Government work report embedding dimension and sliding window selection
图 3. 作报告嵌入维度及滑动窗口选择

先训练出滑动窗口为 10、间隔为 10 的维度区间[100,2000](以[100,2000,10]表示，下同)的政府文档嵌入同胚集合$\{DVM\}_1$，在$\{DVM\}_1$上使用模型 1，求得最优政府文档嵌入维度区间为[1020,1140]，然后在该区间上间隔为 1 训练出政府文档嵌入集合$\{DVM\}_2$，在$\{DVM\}_2$上使用模型 1，最终选择最优政府文档嵌入维度为 1062，即使用 1062 维神经网络训练出的政府文档嵌入损失最小(见图 3)。

结论 1：全国 31 个省、直辖市、自治区的各级地方政府工作报告最佳文档嵌入维度在 1062 附近，最佳滑动窗口大小约为 17。

4.3. 文档嵌入维度大小设置及其影响

Yin 和 Shen 在其论文中提出了一个理解词嵌入维度的理论框架[5]。他们定义了 PIP 损失函数来测量词嵌入之间的差异，通过将词嵌入转化为带噪音的矩阵分解问题，开发了一套基于微扰理论(matrix perturbation theory)的分析技术，发现 PIP 损失函数有偏差-方差分解(bias-variance decomposition)进而直接证明了词嵌入最佳维度的存在，通过寻找偏差和方差的平衡点可以直接求解理论上最优的维度。

由于文本嵌入和词嵌入虽然在神经网络架构上存在相似性，但由于文本嵌入在神经网络的输入层上多了一个文档 ID，使其对于不同的文本集有不同的最佳嵌入维度，文本集的平均长度、长度的方差都是影响最佳嵌入维度的重要因素。对于文档嵌入维度大小选择的普遍做法是维度越大越好，这种做法通常会带来以下问题：

- 深度神经网络对于文本分词序列的过学习，使得文档向量具有较大方差夹杂大量杂音；
- 训练的时间成本和设备成本与文档的规模、维度的大小成正比，越大的文档集和越大的维度设置会带来越长的训练时间和越昂贵的硬件要求。

实际情况中，不同类型的文本集的最佳文档嵌入维度大小和滑动窗口大小是不同的，通常表现为：

- 文本越长，最佳嵌入维度越大；
- 文本的单句平均长度越长，滑动窗口越大；
- 滑动窗口通常取文档句子长度的平均词数。

针对上述情况，在实际的使用中，对于政府文本嵌入维度大小的设置可以采用预估-抽样实验-选择的三个步骤决定：

1) 根据待训练文档集的平均长度预估其最佳嵌入维度大小的区间；

2) 在预估区间上根据实际情况设置区间间隔，使用小批量文档集，快速训练，在得到的文档向量上使用通用模型 1 得到最优维度大小；

3) 绘制文档嵌入损失-嵌入维度图，若所选最优维度处于凹曲线内，决定使用更小的区间间隔再度优化或退出选择；若所选最优维度处于预估区间两端之一或无凹曲线，则应向该方向调整预估区间。

上述政府文档嵌入维度大小的选择步骤，关键在于使用小的文档集小步试错、快速迭代，从而找到整个待训练政府文档集的最佳训练维度。

文挡嵌入维度是深度学习运用在自然处理中，使用神经网络训练得到文档的向量表示首要考虑的重要参数，过小的嵌入维度会造成政府文档隐含信息的丢失，使得政府文档向量表示具有较大的偏差，而过大的嵌入维度会造成深度学习的过拟合夹杂大类噪音，政府文档向量表示具有较大的方差[11]。无论是过小还是过大的政府文档嵌入维度都不能很好的完成文档的矢量化，进而在下一步的文本分类、聚类和内容分析上表现不佳。

5. 政府工作报告省域可分性

5.1. 省域文档大数据可分性分析

政府工作报告的省域可分性是指，以报告所属省市为标签，以全国所有的报告文档为分析对象，通过自然语言处理、机器学习和深度学习的方式，使用部分已知标签的文本训练出分类模型，能够对未知标签的文本进行分类并达到一定的分类准确指标。本文通过将每个报告文本embedding为一个数学向量，这些向量在空间中的各个维度即为报告文本通过深度神经网络学习到的文本特征，具有相同维度特征的文本向量即可使用传统机器学习如分类回归树的方式，将带省市标签的报告文本向量组成的矩阵划分为训练、测试集，论证政府工作报告的省域可分性。

对政府工作报告文档嵌入向量进行分类的机器学习算法为 XGBoost。底层算法分类回归树(classification and regression tree，CART)使用决策二叉树，树节点的划分指标是平方损失函数，通过不断地基于特征值将树的叶结点进行分裂，最终得到一个可以正确分类的二叉树预测模型，如结点是基于第 j 个特征值分裂，数据点在该特征值上小于等于分界点 s 划为左子树，大于为右子树[12]，即：

$$R_1(j,s)=\left\{x \mid x^{(j)} \le s\right\},\quad R_2(j,s)=\left\{x \mid x^{(j)} > s\right\}$$

XGBoost 将许多 CART 集成联合得到总输出，这种许多使用同一模型、效果不佳的弱分类器，联合在一起使得它们组成一个分类效果超过它们中任一一个弱分类器的强分类器的算法思想为 Boosting [13]。XGBoost 支持在特征粒度上并行计算，这使得其对具有高纬度特征的文档向量非常友好，并且其支持列抽样。XGBoost 在机器学习领域的天池、Kaggle 和 DataCastle 的比赛中屡获大奖，模型的准确性得到实际检验。不同维度的文档嵌入同胚具有相同数目的相对坐标，对政府工作报告文档大数据进行标签文档训练、预测时，可利用 XGBoost 算法准确、快速、多线程并行计算的优点，研究不同省份的报告文档是否具有明显的可区分性。

5.2. 省域可分性实验结果与因素分析

以 31 个省和直辖市为标签，使用最佳嵌入维度的政府工作报告文档向量模型，3/4 训练、1/4 预测，评估各省市政府工作报告是否具有明显可分性，统计各省的预测指标见表 1。不同的省份具有不同的经济、文化、产业、自然环境、民族、宗教、饮食和方言等特色[14]，因此不同的地方政府在其政府工作报告中，具体到自身的特色会有相应的区分于其它报告的报告内容，例如西藏自治区的“藏药保护和开发”就不大可能出现在其他省市的政府工作报告中。

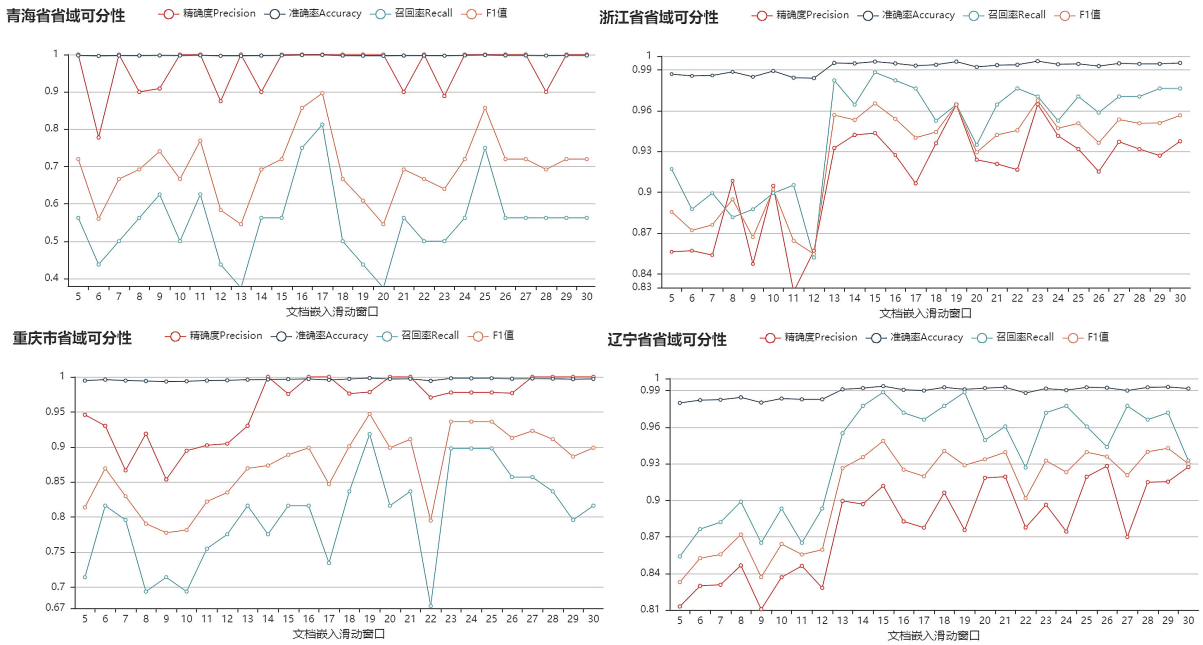

Figure 4. Provincial separability under different sliding window of government work report
图 4. 政府工作报告不同滑动窗口下的省域可分性

从表中可以看出，不同地区政府工作报告的省域区分度并不相同。这不仅与各省特色是否鲜明，还与各省省会首位度、省内经济发展阶段分布情况有关。在固定滑动窗口为 10、嵌入维度区间为[100,2000,10]，对各省市的省域区分度分类指标进行统计；在固定嵌入维度为 800、滑动窗口区间为[5,30,1]，对各省市的省域区分度分类指标进行统计。图 4 为选取的青海、浙江、重庆、辽宁四个省和直辖市(以下简称省市)相同嵌入维度不同滑动窗口的省域可分性示例，图 5 为相同滑动窗口不同嵌入维度的省域可分性。

通过深度学习的方式训练政府文档嵌入向量，得到政府文档嵌入同胚后通过文档嵌入损失，全国所有省市政府工作报告在维度为 1062 和滑动窗口为 17 上训练得到其最佳嵌入维度的文档嵌入向量，可将各省市的政府工作报告进行有效区分。文档嵌入流形空间的维数并不是指通常意义上的平面直角坐标系

的维数，它是拓扑意义下的和嵌入流形微分同胚的欧式空间维数，即流形的维数。本文实验得出的政府工作报告最佳嵌入维度的结论，指非线性流形的维数。

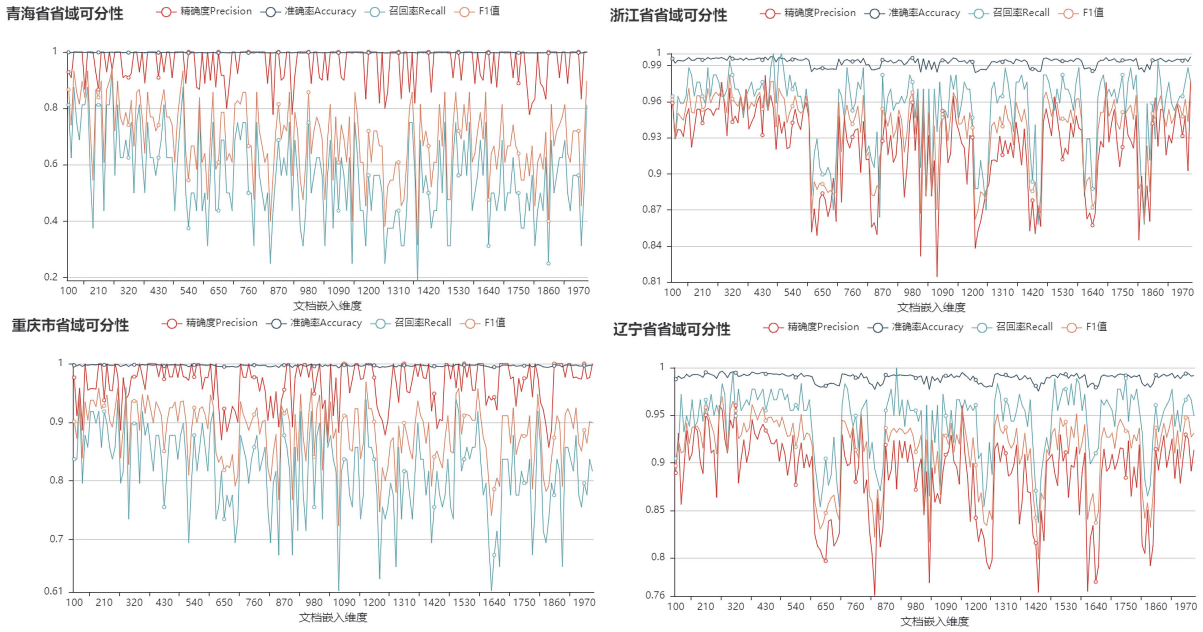

Figure 5. Provincial severability in different embedding dimensions of government work report
图 5. 政府工作报告不同嵌入维度下的省域可分性

Table 1. Provincial severability under the optimal dimension of government work report
表 1. 政府工作报告最优维度下的省域可分性

	F1 值	准确率 Accuracy	召回率 Recall	精确度 Precision		F1 值	准确率 Accuracy	召回率 Recall	精确度 Precision
中央	1.000	1.000	1.000	1.000	江西	0.974	0.997	0.980	0.967
上海	0.990	1.000	0.980	1.000	河北	0.939	0.995	0.950	0.927
云南	0.991	1.000	0.982	1.000	河南	0.881	0.997	0.804	0.974
内蒙	0.985	0.998	1.000	0.970	浙江	0.980	0.998	0.983	0.977
北京	0.939	0.997	0.939	0.939	海南	0.985	1.000	0.971	1.000
吉林	0.968	0.998	0.955	0.981	湖北	0.996	1.000	1.000	0.992
四川	0.982	0.998	0.985	0.978	湖南	0.857	0.998	0.750	1.000
天津	0.929	0.998	0.886	0.975	甘肃	0.969	0.998	0.951	0.987
宁夏	0.958	0.999	0.958	0.958	福建	0.989	0.999	1.000	0.979
安徽	0.978	0.998	0.969	0.987	西藏	1.000	1.000	1.000	1.000
山东	0.967	0.996	0.978	0.957	贵州	1.000	1.000	1.000	1.000
山西	0.980	0.999	0.974	0.987	辽宁	0.954	0.995	0.978	0.931
广东	0.981	0.997	0.981	0.981	重庆	1.000	1.000	1.000	1.000
广西	1.000	1.000	1.000	1.000	陕西	0.982	0.999	0.964	1.000
新疆	0.963	0.999	0.929	1.000	青海	1.000	1.000	1.000	1.000
江苏	0.966	0.994	0.977	0.955	黑龙江	0.971	0.997	0.975	0.968

从图 4、图 5 中可以看出，各省的省域区分度分类指标取得最优时的文档嵌入维度和滑动窗口大小并不相同，主要原因有：

- 各省省情不一，例如西藏、宁夏等作为少数民族自治区，政府工作侧重民族区域自治，报告内容自然与其他省市不同；
- 各省语言风格不同，为使政府工作报告接地气、入民心，各地官话、晋、湘、赣、吴、闽、粤、客语八大方言[15]本地特色的语料不同程度地进入到当地的政府工作报告中。

政府工作报告因其主要陈述政府一年来的工作政绩、来年的重点工作内容，这些结构性因素使得它们的行文分段往往在框架上具有相似性。从报告的结构和内容来看，虽然不同省市、不同行政级别的报告内容和结构相似，但因其填充的具体报告内容是结合当地一年来的工作回顾、当年工作任务和政府自身建设，如同在同一词牌名下的宋词，虽其韵律字数高度一致，但用词写意是不同的，因此在使用文档向量的方式表示各个政府工作报告后，报告的用词、造句、表文、达意这些细节就会映射到向量各维度的数值差异，因此使用深度神经网络解剖后的政府工作报告，可以通过传统机器学习分类回归树的方式将各省报告有效区分。

报告按照省(直辖市)进行省域可分实验而不是按照地级或市级实验，是由我国的行政划分决定的。当前除中央政府政策、指令和指示能覆盖全国以外，我国各地方政府的最高行政机关就属省部级的省(直辖市)一级了，因此在同一省(直辖市)下，各地的政府工作报告都会贯彻执行这一省部级的行政指令，因此全国所有省市的政府工作报告会以省域为分界线聚集成团。

结论 2：全国 31 个省、直辖市、自治区的各级地方政府工作报告可以通过深度学习结合机器学习的方式有效地区分开。

6. 政府工作报告省域相似性

6.1. 省域文档大数据相似性分析

定义 5 (同地相似率 SPSR)：为比较某一地方政府工作报告与它不同时间的报告文档的相关度，定义其同地相似率为：

$$\mathrm{SPSR}\left(D,\boldsymbol{W}_d\right)=\frac{\mathrm{SUM}\left(\left\{D=\boldsymbol{W}_{di}\right\}\right)}{\mathrm{SUM}\left\{\boldsymbol{W}_d\right\}}$$

其中，D 为目标文档，$\{\boldsymbol{W}_d\}$为 D 的最相似文档集合，这里取 10 个最相似文档(下同)，运算符=在这里表示 $\boldsymbol{W}_{di}$为 D 的同一地方政府、不同年份的文档，运算符 SUM 表示求集合元素个数。

定义 6 (同市相似率 SCSR)：为比较某一地方政府工作报告与它所处同一市级报告文档的相关度，定义其同市相似率为：

$$\mathrm{SCSR}\left(D,\boldsymbol{W}_d\right)=\frac{\mathrm{SUM}\left(\left\{D\cong\boldsymbol{W}_{di}\right\}\right)}{\mathrm{SUM}\left\{\boldsymbol{W}_d\right\}}$$

运算符≅在这里表示 $\boldsymbol{W}_{di}$为 D 的同一市级下的报告文档。

定义 7 (异省相似率 DPSR)：为比较某一地方政府工作报告与它不同省份报告文档的相关度，定义其异省相似率为：

$$\mathrm{DPSR}\left(D,\boldsymbol{W}_d\right)=\frac{\mathrm{SUM}\left(\left\{D\neq\boldsymbol{W}_{di}\right\}\right)}{\mathrm{SUM}\left\{\boldsymbol{W}_d\right\}}$$

运算符≅在这里表示 $\boldsymbol{W}_{di}$为 D 的不同省份下的报告文档。

对文档嵌入向量集 $\boldsymbol{E}^{n\times d}$ 和相对应的文档名列表，使用模型 3 求每篇文档的前 k 个最相似文档集。最相似政府文档发现模型是一个用于求解政府工作报告最相似文档集合的计算机算法，它输入经过向量化后的政府工作报告文档向量，在向量距离模块可以使用文档余弦距离或文档欧式距离公式。

模型 3 (最相似政府文档集发现模型)

输入：政府文档嵌入向量集合 $\boldsymbol{E}^{n\times d}$
输出：政府文档的最相似文档字典

1、求政府文档距离矩阵 $\boldsymbol{B}^{n\times n}$，使用文档余弦距离 DCOS 或文档欧式距离 DEUC；
2、求政府文档距离矩阵 $\boldsymbol{B}^{n\times n}$，使用文档余弦距离 DCOS 或文档欧式距离 DEUC；
3、对 $\boldsymbol{B}^{n\times n}$ 排序，使用 DEUC 时求最小的 k 个距离所在的文档名列表索引 $\boldsymbol{S}^{n\times k}$ 矩阵，使用 DCOS 时求最大的 k 个距离得到 $\boldsymbol{S}^{n\times k}$；
4、使用索引矩阵 $\boldsymbol{S}^{n\times k}$ 和文档名列表，求得所有政府文档的最相似文档集合。

6.2. 省域相似性实验结果与因素分析

对最优参数的政府工作报告文档嵌入 E，使用文档相似度 DCOS 获取所有文档的 10 个最相似文档列表，然后按照省市计算该省市下所有文档的 SPSR、SCSR、DPSR，取它们的平均值作为该省市的相应评估值，并统计该省市下 SPSR 最高的地方政府名(见表 2)。

Table 2. Provincial document similarity in the optimal dimension of government work report
表 2. 政府工作报告最优维度下的省域文档相似性

省/直辖市	同地相似率	同市相似率	异省相似率	最高同地相似率报告	省/直辖市	同地相似率	同市相似率	异省相似率	最高同地相似率报告
北京	0.804	0.862	0.278	北京市延庆区	江西	0.367	0.409	0.567	江西省
天津	0.689	0.781	0.340	天津市红桥区	山东	0.461	0.489	0.518	山东省青岛市崂山区
上海	0.679	0.915	0.281	上海市	河南	0.148	0.158	0.807	河南省
河北	0.277	0.329	0.668	河北省	湖北	0.450	0.498	0.465	湖北省
山西	0.367	0.420	0.534	山西省	湖南	0.180	0.195	0.789	湖南省
辽宁	0.414	0.439	0.553	辽宁省沈阳市沈河区	广东	0.544	0.664	0.431	广东省
吉林	0.449	0.485	0.525	吉林省通化市	海南	0.499	0.575	0.322	海南省
重庆	0.486	0.602	0.495	重庆市	四川	0.433	0.493	0.455	四川省
黑龙江	0.396	0.500	0.414	黑龙江省黑河市	贵州	0.379	0.475	0.511	贵州省遵义市桐梓县
江苏	0.527	0.588	0.259	江苏省南京市建邺区	云南	0.590	0.634	0.397	云南省
浙江	0.524	0.601	0.419	浙江省温州市洞头区	陕西	0.488	0.551	0.427	陕西省铜川市
安徽	0.378	0.405	0.556	安徽省阜阳市	甘肃	0.495	0.588	0.414	甘肃省甘南藏族自治州
福建	0.506	0.597	0.426	福建省福州市	青海	0.562	0.609	0.416	青海省
西藏	0.718	0.718	0.489	西藏自治区	内蒙古	0.316	0.446	0.513	内蒙古自治区
宁夏	0.431	0.529	0.485	宁夏回族自治区	广西	0.379	0.470	0.521	广西壮族自治区柳州市
新疆	0.581	0.619	0.311	新疆维吾尔自治区博尔塔拉蒙古自治州博乐市					

在求取所有政府文档的前 10 个最相似文档集时，分别使用文档余弦距离 DCOS 和文档欧式距离

DEUC，对所有的政府工作报告文档，两种方式得到的政府文档相似集一致。

T-SNE 是一种可用于高维数据降维的流形学习方法[16]。文档嵌入向量降维引起的信息损失会随着维数的增加而不能弥补维数优化带来的增益，因此，这里选取 200 维、训练迭代次数 40 的文档嵌入向量使用 T-SNE 降维后得到可视化结果(见图 6)。

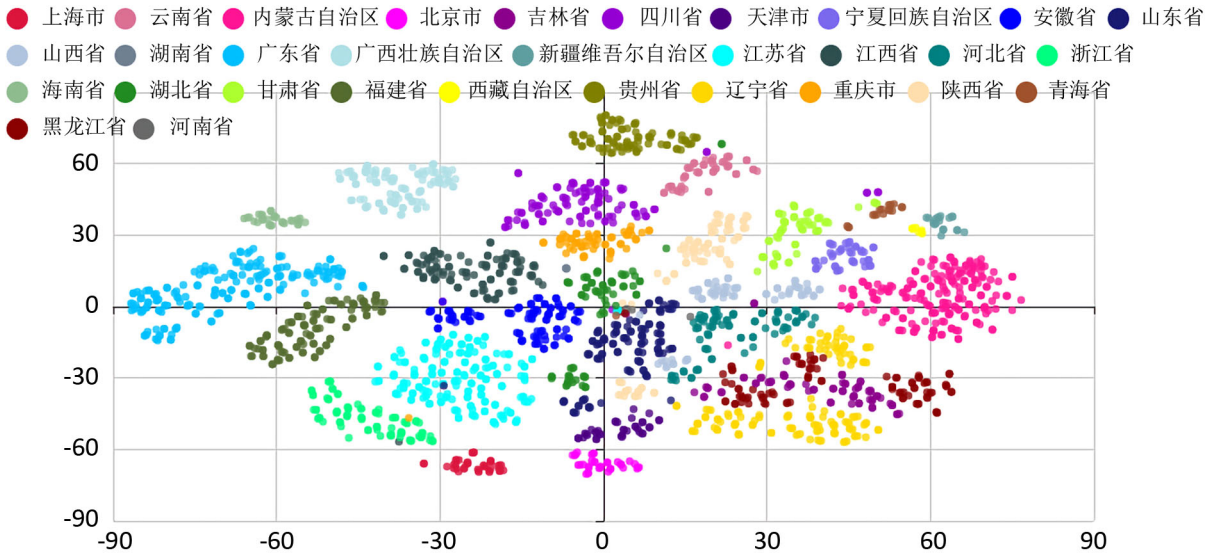

Figure 6. Two-dimensional visualization of embedding vectors of government work report after dimensionality reduction by T-SNE

图 6. T-SNE 降维后的政府工作报告文档嵌入向量二维可视化

政府工作报告文档嵌入向量的二维可视化，从侧面印证了省域相似性比较低的数据点，在高维空间中相对于本省市的数据点离异省市数据点文档距离更近，这是因为现实中各地的内在因素关联不同：

1) 经济活动联系并不是以省市行政区域为分割的，例如长三角经济带，江苏苏州、江苏昆山、浙江嘉兴相对本省省会来说，与上海市经济联系更加紧密，产业配套更加衔接；

2) 社会发展阶段并不是同省市就处于同一水平的，例如江苏无锡与湖南长沙经济总量相差不大，但长沙作为省会城市与同省非省会城市岳阳、衡阳等就不在一个发展阶段上。

政府工作报告在内容和结构上，不同省市的同地相似率、同市相似率和异省相似率差异较大，具体到地方政府的报告上，大致有以下几种情况：

1) 以北京市为代表的同地相似率和同市相似率显著高于异省相似率；

2) 以河南省为代表的同地相似率和同市相似率显著低于异省相似率；

3) 以湖北省为代表的同地相似率、同市相似率和异省相似率差别不大。

出现以上三种情况的主要原因，可以从地方政府的政治地位、政府官员的来源、经济教育文化差异等各方面因素做出如下分析：

1) 各地方政府官员的调动频繁程度不一，往往地方主要领导的升迁会相应的伴随着施政方向的变化，政府工作内容和政府工作报告执笔人的变化也相应的反应在各地方政府的同地相似率中；

2) 北京、上海等同地相似率和同市相似率显著高于异省相似率，由于经济地位和政治地位的塔尖位置，其政府的工作重心、行政风格、经济政治体量和卓越的公务员群体都充分渗透到其政府工作报告中显著区别于其他省市，行政风格的创新性和试点试新的领先性体现在政府工作报告的报告内容中，使得它们的政府工作报告表现为同地相似率和同市相似率很高，异省相似率低；

3) 河南和湖南等同地相似率和同市相似率显著低于异省相似率，与其产业链完整度、经济发展水平、地理位置、人口流动、教育资源及民间文化相关，湘楚文化重文教、向往外出闯荡，同时湖南紧邻广东、湖北、重庆，与广州、深圳、重庆、武汉的经济、人口往来密切，河南省内高等教育资源匮乏导致其地方政府的各级政府主要领导在外省市接受高等教育的比例大，同时河南身为人口大省又地处中原，省会城市对省内经济、文化、人口、产业的吸引力不是特别明显；

4) 湖北、四川、陕西同地相似率、同市相似率和异省相似率差别不大，由于武汉、成都、西安分别是中部、西南部、西北部地区的中心城市，其周边省市人口流入带来的文化多样性和省会城市教育、医疗、政治资源的集中性等因素都使得其同地、同时、异省相似率均比较平衡。

政府工作报告的同地、同市和异省相似率，在反映当地政府工作特色的同时，一定程度上折射了地方政府官员的文化、教育和籍贯带来的差异，这种差异体现在行文造句、语言风格上面，各省市的最高同地相似率基本为省部一级行政单位，也反映了省部级官员地域多样、执政为民、政策连续的工作特点。

结论 3：同一地方政府的历年工作报告具有时间传承性，表现为其最相似的文档为其往年的政府工作报告。

结论 4：高维文档嵌入求相似度时，使用向量余弦距离(余弦夹角)和使用向量正则化后的欧式距离(直线距离)几乎等效。

7. 结语

本文通过一种基于文档嵌入维度同胚空间的最佳嵌入维度人工智能分析模型，对同一政府工作报告在不同维数的深度学习向量表示转化为相同维数下的同胚向量表示，通过比较嵌入同胚损失得到政府工作报告的最佳维度，在最佳维度空间中进行基于机器学习的政府工作报告省域文档可分性和省域文档相似性的管理学分析。

本文通过省域可分性实验表明了各地方政府的政府工作和施政纲领一定程度上在省(直辖市)一级上呈聚类关系，通过省域相似性挖掘出了不同地方政府的政治地位、政府官员的构成、实体经济产业联系、高等教育资源分配差异、语言文化用词风格差异等内在多因素带来的政府工作报告同地、同市和异省相似率的差异。本文的实验方法对文本所属领域、语言类型、文本长度、文本多分类和文本时间序列特征均具备较好的鲁棒性，因此对非中文文本和其他领域的文本最佳嵌入维度分析与分类，均有一定借鉴意义。

本文的政府工作报告省域可分性和省域相似性的大数据分析具有一定的经济政策借鉴意义，对我国智慧电子政务建设具有一定的参考意义和应用价值，同时本文使用的文档嵌入维度同胚分析框架可以支持一些实际的管理学文本研究，如政府公文、行业分析报告、上市公司公告等。如对上市公司公告文本，将其在最佳文档嵌入维度上使用深度神经网络向量化后，使用机器学习的方式可以分析不同行业的公司产业关联性和同行业不同公司之间的竞争优势，也可将其应用于国情分析咨文，通过对不同国家、不同行业的竞争优势进行分析，得到报告的深层次经济、政治、文化等多方面再解读，从而为跨国公司提供国际化战略参考。

致　谢

本研究工作得到中国人民大学高性能计算校级公共平台支持。

基金项目

国家社会科学基金项目(18ZDA309)、国家自然科学基金项目(71531012)。

参考文献

[1] 魏伟, 郭崇慧, 陈静锋. 国务院政府工作报告(1954-2017)文本挖掘及社会变迁研究[J]. 情报学报, 2018, 37(4): 406-421.

[2] 何家莉. 覆盖近似空间的连续与同胚映射[J]. 纯粹数学与应用数学, 2019, 35(2): 229-234.

[3] Harris, Z. (1954) Distributional Structure. *Word*, **10**, 146-162. https://doi.org/10.1080/00437956.1954.11659520

[4] Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003) Neural Probabilistic Language Model. *Journal of Machine Learning Research*, **3**, 1137-1155.

[5] Yin, Z. and Shen, Y.Y. (2018) On the Dimensionality of Word Embedding. Neural Information Processing Sytems. ComputerScence, Montreal. http://arxiv.org/abs/1812.04224v1

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. http://arxiv.org/abs/1310.4546v1

[7] Le, Q. and Mikolov, T. (2014) Distributed Representations of Sentences and Documents. http://arxiv.org/abs/1405.4053

[8] 李欣, 李旸, 王素格. 面向情感聚类的文本相似度计算方法研究[J]. 中文信息学报, 2018, 32(5): 97-104.

[9] 朱小飞, 郭嘉丰, 程学旗, 杜攀. 基于流形排序的查询推荐方法[J]. 中文信息学报, 2011, 25(2): 38-43.

[10] 罗四维, 赵连伟. 基于谱图理论的流形学习算法[J]. 计算机研究与发展, 2006, 43(7): 1173-1179.

[11] Yin. (2018) Understand Functionality and Dimensionality of Vector Embeddings: The Distributional Hypothesis, the Pairwise Innerproduct Loss and Its Bias-Variance Trade-Off. https://arxiv.org/abs/1803.00502

[12] 张蕾, 崔勇, 刘静, 江勇, 吴建平. 机器学习在网络空间安全研究中的应用[J]. 计算机学报, 2018, 41(9): 1943-1975.

[13] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[14] 徐春华, 刘力. 省域市场潜力、产业结构升级与城乡收入差距——基于空间关联与空间异质性的视角[J]. 农业技术经济, 2015(5): 34-46.

[15] 李荣. 官话方言的分区[J]. 方言, 1985(1): 2-5.

[16] Maaten, L. and Hinton, G. (2008) Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605.