“智慧政务”中的文本挖掘应用

摘　要：当下，随着网络问政平台的兴起，各类类社情民意相关的文本数据量急剧增加，数据的增多使得相关部门可以从中获取更多的信息，但是如果继续使用以往人工挖掘数据的方法，势必会造成“数据灾难”。但随着当下大数据、NLP(Natural Language Processing)技术的发展使得在各类留言中智能分析处理的“智慧政务”成为了可能。

赛题中的留言为非结构性的自然语言，我们采用了以深度学习为基础的 word2vec 模型。我们首先通过使用 jieba 分词对于文本进行分词处理，使用 word2vec 进行词向量模型的构建并通过词向量实现对于留言文本的向量化处理，然后我们使用 SVM 对于所有的留言进行分类实现对于一级标签的生成，其判断的准确率使用 F-Score 进行评价，可以达到 97.6%以上，具有较高的准确率。而对于热点问题的分析，我们基于上述模型中的句向量，实现对于留言的种类预测，尔后对于属于同一一级标签下的留言进行聚类分析，该过程使用 Meanshift 算法实现无监督下的聚类，并通过爬取全国所有住宅小区的名称数据，对于留言中指出问题的发生地进行识别，通过选取地点重合且分类与子类均相同的留言生成了热点问题的挖掘。最终我们通过对于留言与答复的模式识别与匹配，实现对于答复的评价系统。

我们的模型在构造的线下测试集上得到了有利的验证，反映出该系统在实际中能够实现预期的效果。也说明了模型的鲁棒性，实用性。

关键词：自然语言处理，词向量，SVM，Meanshift

Application of text mining in “intelligent government affairs”

Abstract: At present, with the rise of seeking public opinion on units, various types of public opinion related text data volume increase sharply, the increasing of the data that relevant departments can obtain more information, but if you continue to use artificial mining data, is bound to cause “data disaster”, but as the mega data NLP (Natural Language Processing) the development of technology makes the wisdom of the intelligence analysis Processing affairs can be used in all kinds of messages .

The problem of the message is unstructured natural language, we adopted on word2vec model on the basis of the deep learning. First, we use the “jieba participle word segmentation” for text preprocessing, use word2vec build word vector model and use the word vector to realize the vectorization of text messages, and then we use SVM to classify all messages and implement level 1 label generation, use the F - Score evaluation to judge its accuracy, its F number can reach more than 97.6%, it’s a very high level. For the analysis of the hot issues, we based on the sentence vector in the model, implement the message types of prediction, and then clustering analysis the different message under the same level label. This process uses Meanshift algorithm under the unsupervised clustering, and we crawl the data of the name of all residential area across the country. Identify the field of the message, by choosing location overlap and classification and subclasses which are of the same message to generate the hot issues of digging. In the end, we recognized and matched the pattern of the message and the reply of your message, realize the evaluation system for a reply.

Our model has been verified in the constructed offline test set, which shows that the system can achieve the expected effect in practice and also shows the robustness and practicability of the mode.

Key words: natural language processing, word embedding, SVM, Meanshift

目　　录

1. 挖掘目标

1.1 挖掘意义

近年来，随着各类网络问政平台的兴起，各类社情民意相关的文本数据量不断攀升。对于社情民意的分析，以及基于分析结果妥当回复、解决问题，很大程度上决定了政府的公信力与效率。[1]目前相关部门依靠人工来进行留言划分和热点整理，庞大的数据量以及对于各类问题的归类、精确分析，使得通过人工分析数据、挖掘数据难度极高且公务人员的压力极大。因此从使用计算机对于文本进行挖掘，从中给出分类、热点信息，有着重要的社会意义。

1.2 挖掘目标

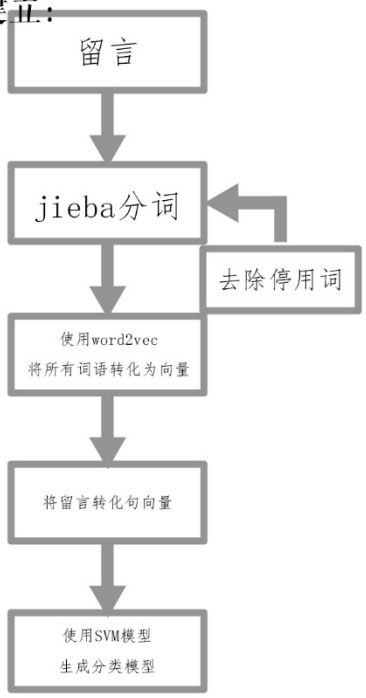
本次建模针对网络问政平台中的大量自然语言写成的文本信息进行分类以及热点问题的提出，主要分为建模，识别和分析三个过程，首先通过监督下学习算法建立词向量模型，然后根据这一模型对于留言进行从文本到数字的转换。在对于转换后留言信息，通过 SVM 算法进行监督下学习，从而实现对于输入留言的归类，基于留言信息获得热点问题。同时针对留言的答复信息，建立起一套对于答复信息的相关性、完整性、可解释性的评价标准。

2. 分析方法与过程

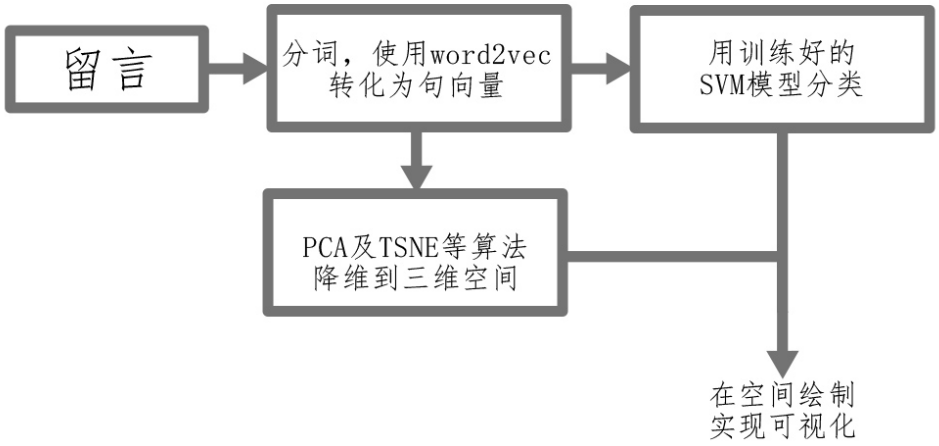
2.1 问题 1 的分析方法与过程

2.1.1 问题 1 基本流程图

问题1模型建立：



问题1标签预测及可视化：



2.1.2 留言的预处理

由于涉及到对于自然语言的分析处理，我们就要先对于留言进行预处理，在题目给出的数据中每一条留言，基本上都会出现无用的停词以及无关的数字。同时，留言中涉及到大量的地点信息，以及针对某一具体地点的描述。而对于地点的分析，我们就需要识别出这些词语是表示某一具体地点。为此，我们爬取了全国主要城市的小区、社区的名称，这些数据

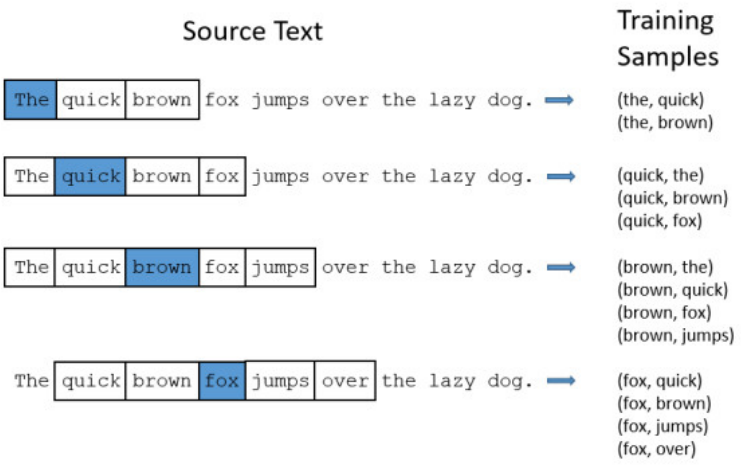
分别保存在。

同时，由于留言为自然语言属于非结构化文本信息，因此在对于留言信息的挖掘前，必须要将这些文本信息转化为计算机能够识别的结构化信息，这里为便于转换为就需要要对于这些信息进行中文分词。这里采用了 jieba 中文分词包进行分词。

2.1.3 通过词向量生成句向量

使用词向量模型的基本思路就是通过训练，将每个词都映射到一个较短的词向量上来。所有的这些词向量就构成了向量空间，进而可以用普通的统计学的方法来研究词与词之间的关系，此处在多次试验后词向量维度我们指定为 1000。[2]

生成词向量我们采用的是 gensim 库中的 word2vec 模型实现。传统的神经网络词向量语言模型，里面一般有三层输入层（词向量），隐藏层和输出层（softmax 层）。而 word2vec 是基于这一思路的改进：首先，对于从输入层到隐藏层的映射，采用简单的对所有输入词向量求和并取平均的方法。第二个改进就是从隐藏层到输出的 softmax 层这里的计算量改进，避免要计算所有词的 softmax 概率。基于上述特点 word2vec 在对于该问题的建模时实现速度快且准确率高。对于向量的生成，选择的是对于低频词敏感 Skip-Gram 模型。



而本题中要求对于句子进行分类，就需要对于句子的特征进行识别，

根据这一思路，可以通过对于词向量的叠加来构建一个句子的向量表示，而相比较更为复杂的 dec2vec 算法，该算法运行效率高在比较后准确率也比 doc2vec 高。

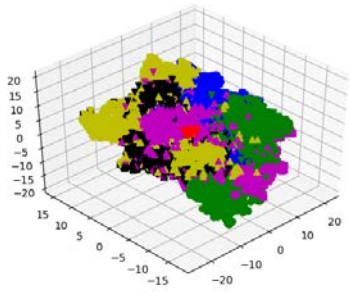
基于上述分析与实验结果，我们通过对于文本进行分词，利用 gensim 库实现词向量的训练生成，并利用该模型进行句向量的构建。

由于生成的向量存在方向性且句子的长度不一致，因此在生成最终的句向量时，将所有的句向量进行了归一化处理。

2.1.4 建立监督学习聚类模型

基于题目数据，实现对于留言的种类进行分析，就需要通过对于分类器训练来得到合适的数据结构，而经过对比前期使用朴素贝叶斯模型（Naive Bayesian Model，NBM）[3]进行聚类与分类时，正确率在 60%左右，分类能力不强，因此改进模型使用支持支持向量机（support vector machines, SVM）[4]分类器来实现分类。

由于数据已经存在了一级标签这个字段，因此为保证模型训练时准确且不过敏感，就需要对于对数据进行切分，分为训练集和测试集，并将标签信息加入训练集，生成训练后的 SVM 模型，此时的分类正确率可达 90%以上。



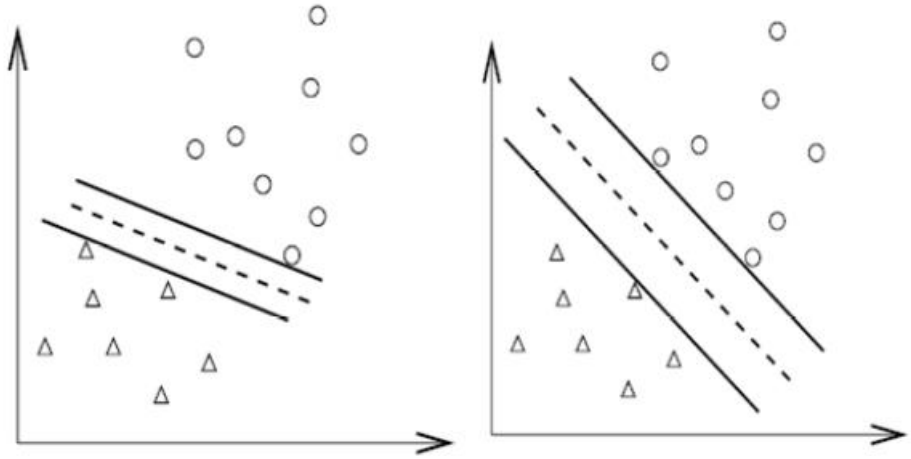
SVM 算法的具体原理如下：

在上述步骤中，通过使用 word2vec 实现将非结构化的文本信息转换为了一个向量，而这一向量具有一个标记（一级标签），表示为：

$$D_i = (X_i\ ,\ t_i)$$

其中，i 表示为第 i 个句子向量，X_i表示该句的句向量，t_i为该句的一级标签。

在超平面上可以想象上述向量存在多个分类面实现分类，而 SVM 寻找这些最优面的算法。



首先一个样本点到某个超平面的间隔表示为：

$$\delta_i = Y_i\ (wX_i + b)$$

归一化后间隔表示为：

$$\delta_i = \frac{1}{||w||}|g(X_i)|$$

其中||w||表示做向量 w 的范数。

当时用线性 SVM 时，约束条件是样本点到决策边界的距离大于等于 1，这使得问题变为：

$$\max_{w,b}\ \frac{2}{\|\boldsymbol{w}\|} \quad \text{s.t.}\quad y_i\left(\boldsymbol{w}^\top \boldsymbol{X}_i + b\right) \geq 1 \quad \Longleftrightarrow \quad \min_{w,b}\ \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.}\quad y_i\left(\boldsymbol{w}^\top \boldsymbol{X}_i + b\right) \geq 1$$

按照这一思路将数据分为训练集和测试集，将给出的数据随机打乱后建立训练与测试集，其大致比例为 3:1。通过训练获得模型，进行测试验证模型的准确率，并根据预测结果，使用 F-Score 进行评价，再进行优化直到其 F-Score 最大。

为了使得上述结果直观且便于改进参数，因此使用 PCA(Principal

Component Analysis)[5-7]以及 TSNE（Stochastic neighbour Embedding）方法将句向量降至三维进行分析，便于对于该过程的合理性分析以及直观可视化。

2.1.5 通过训练模型进行预测

留言通过 word2vec 后生成的向量在 SVM 模型中，将其分到最接近的类簇并贴上标签，这样便得到了使用模型分类处理后的标签。

2.1.6 使用 F-score 方法对模型评分

使用 F-Score 对分类方法进行评价

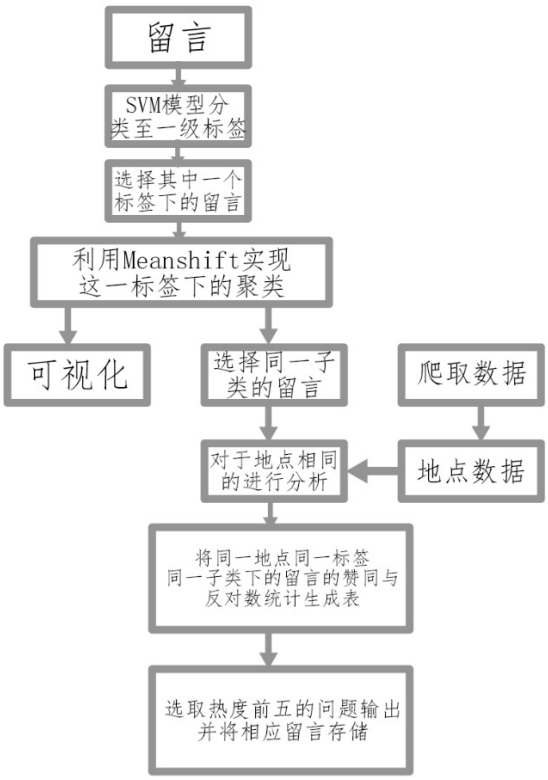
$$\boldsymbol{F}=\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}\frac{\mathbf{2}\boldsymbol{P_i}\boldsymbol{R_i}}{\boldsymbol{P_i}+\boldsymbol{R_i}}$$

其中，$\boldsymbol{P_i}$为第 i 类的查准率，$\boldsymbol{R_i}$为第 i 类的查全率，查准率为 SVC 分类与原始附件二分类中的对比。

2.2 问题 2 的分析方法与过程

2.2.1 问题 2 流程图

问题2：



2.2.2 数据预处理

由于数据中包含大量的地理位置信息，因此需要对于地名与街道名称进行爬取，爬取结果见附件“道路街区名称.txt”。

使用附件中的地理位置信息，实现对于留言的地理位置进行标记，此后依据这一信息来实现对于热点问题的分析。

2.2.3 Meanshift 算法

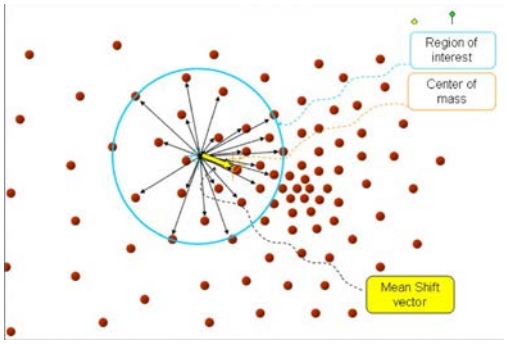
meanshift 算法的核心即均值（mean）与偏移（shift），简而言之，就是随机的选取点，计算需要的偏移量之和并求平均，就得到平均偏移量，（该偏移量的方向是周围点分布密集的方向）该偏移量是包含大小和方向的。然后点就往平均偏移量方向移动，再以此为新的起点不断迭代直到满足一定条件结束。其过程如下：

1、在句向量中随机选择一个点作为中心 center；

2、找出以 center 为核心时，搜索半径距离在 bandwidth 之内的所有点，记做集合 M，认为这些点属于同一簇。同时，把这些求内点属于这个类的概率加 1，这个参数将用于最后步骤的分类；

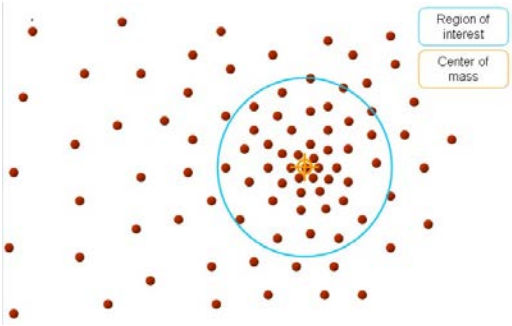
3、以 center 为中心点，计算从 center 开始到集合 M 中每个元素的向量，将这些向量相加，得到向量 shift。

4、center = center+shift。即 center 沿着 shift 的方向移动，移动距离按照均值偏移函数进行||shift||。



5、迭代上述过程直到收敛，将此时的 center 作为新的中心点。同时，这个迭代过程中遇到的点都应该归类到簇 c。

6、如果收敛时当前簇 c 的 center 与其它已经存在的簇 c2 中心的距离小于阈值，那么把 c2 和 c 合并。否则，把 c 作为新的聚类，增加 1 类。



6、重复 1、2、3、4、5 直到所有的点都被标记访问。

7、分类：根据每个类，对每个点的访问频率，取访问频率最大的那个类，作为当前点集的所属类。

简单的说，mean shift 就是沿着密度上升的方向寻找同属一个簇的数据点。

数学推导：

给定 d 维空间 Rd 的 n 个样本点 ，i=1,…,n,在空间中任选一点 x，那么 mean shift 向量的基本定义如下：

$$M_{\hbar} = \frac{1}{K}\sum_{x_i S_k}(x_i - x)$$

其中 S_k是一个半径为 $\hbar$ 的高维度区域。定义如下：

$$S_h(x) = \{y: (y - x_x i)_T (y - x_x i) < h^2\}$$

k 表示在这 n 个样本点x_i中,有k个点落入 S_k 区域中. 然后，我们对 mean shift 向量进行升级，加入核函数（比如高斯核），则 mean shift 算法变为:

$$f_{h,K(x)} = \frac{C_{k,d}}{nh^d}\sum_{i=1}^{n} k(||(x - x_i)/h||^2)$$

其中 K(x)为核函数，h 为半径，$\frac{C_{k,d}}{nh^d}$为单位密度，要使得上式 f 得到最大，就要对上式进行求导

$$\hat{\nabla} f_{h,K(x)} \equiv \nabla \hat{f}_{h,K(x)} = \frac{2C_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(x - x_i)k'(||(x - x_i)/h||^2)$$

令 $g(x) = -k'(x)$, 则我们可以得到：

$$\hat{\nabla} f_{h,K(x)} = \frac{2C_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(x - x_i)g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)$$

$$= \frac{2C_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)\right]\left[\frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x\right]$$

由于我们使用的是高斯核，所以第一项等于$f_{h,K(x)}$

$$\hat{f}_{h,G}(x) = \frac{C_{g,d}}{nh^d}\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)$$

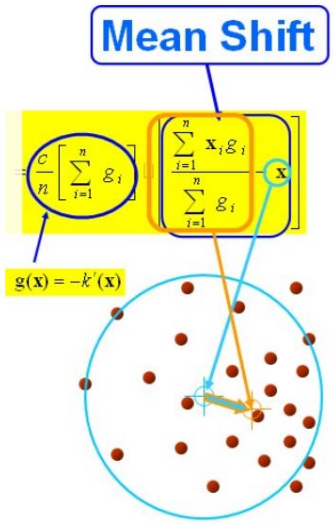
第二项就相当于一个 mean shift 向量的式子:

$$m_{h,G}(x)=\frac{\sum_{i=1}^{n}x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n}g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}-x$$

则上述公式可以表示为：

$$\hat{\nabla} f_{h,K(x)}=\hat{f}_{h,G}(x)\frac{2C_{k,d}}{h^2C_{g,d}}m_{h,G}(x)$$

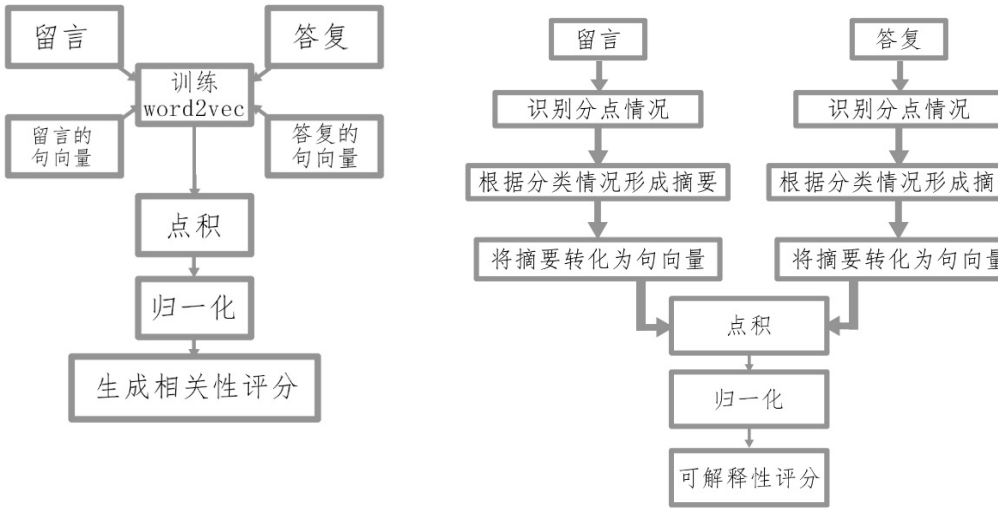
图解公式的结构如下：

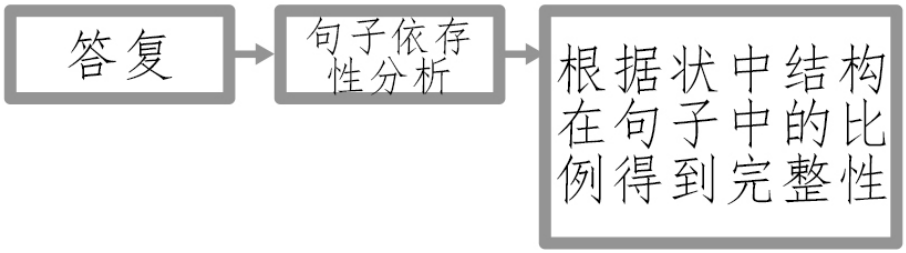


当然，我们求得 mean shift 向量的时候，其密度最大的地方也就是极值点的地方，此时梯度为 0，也就是 $\nabla f_{h,K(x)}=0$,当且仅当 $m_{h,G}(x)=0$ 的时候成立，此时我们就可以得到新的原点坐标：

$$x=\frac{\sum_{i=1}^{n}x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n}g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}$$

2.3 问题 3 的分析方法与过程





2.3.1 有效性、完整性与可解释性

答复的有效性，即为留言与答复二者之间的相关性，对于相关性的计算则仍然可以通过词向量的模型来进行衡量。通过将留言与答复转换为句向量，再将转换得到的向量进行转置，点乘，即可获得二者的相似度。

其中两个部分的句向量分别为$X_i = (a_1, a_2, a_3, \ldots, a_{1000})$与$Y_i = (b_1, b_2, b_3, \ldots, b_{1000})$，而二者的相似度计算即为：

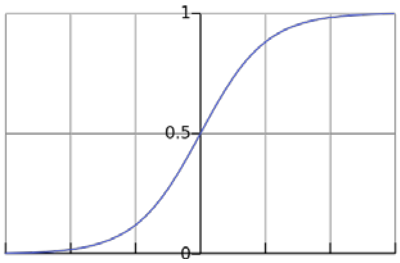
$$relevant_i = X_i \cdot Y_i = a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_{1000}b_{1000}$$

答复的完整性，就是问题与回答的对应关系，可以对于留言与答复中的分点的数量进行统计，同时，对于统计得到的数字，通过 pyhanlp 库进

行对于摘要的提取，此后，将摘要转为句向量的列表，将留言与答复的列表进行点乘，点乘的原理同有效性计算。通过计算后将数值返回，即为答复的完整性指标$complete$。

对于可解释性进行分析，我们从中文的语法上进行分析，通过对于大量样本与理论指导下，得出结论是一个句子中状中结构占比越高，则答复的意见更加主观与不确定。直观的例子就是，当回复中出现“可能”、“原则上”、“应该”或“或许”这类词语时，其不确定性更加大，直接导致文本的可解释性下降。基于这一思路，我们通过对于文本进行句子成分分析，分析其中状中结构与句子中的主要结构关系的比例来获得$explain_i$指标。

综合上述指标，可以对于答复进行综合评价。首先将三个指标的集合进行归一化处理，其中完整性与可解释性进行简单的线性归一化处理，而对于可解释性归一化处理，采用的是 sigmoid 归一化方法，即引入：



使用阈值函数$\frac{1}{1+e^{-x}}$进行归一化处理。

对于归一化的结果，按照加权计算的方法即可以获得最终对于留言答复的评价指标。其生成结果见附件“附件 4_result(1).csv”。

2.3.2 pyhanlp

在对句子成分进行分析时对于 pyhanlp 库进行了调用，而该库对于依存句法的分析使用的是最大生成树模型(maximum spanning trees, MST)以及最大熵模型作为衡量一条依存关系存在的概率的工具[8-9]。

3. 结果分析

3.1 问题 1 的结果分析

3.1.1 一级标签识别结果

经过调整参数，以 F-Score 作为评价标准不断优化模型，此时得到的一级标签识别准确率约为 95%，但是此时的效果已经比较理想，基本可以达到将让其作为模型的要求。

对于数据完全的判别结果见附件“附件 2 分标签结果.xlsx”，其计算得到的 F-Score 为：

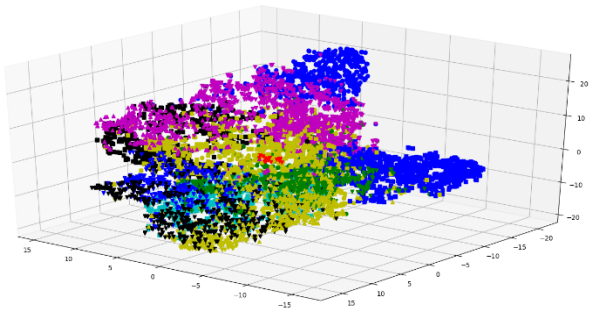
$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} = 0.976781$$

而算法的识别准确性还可以通过加大训练数据进行优化，当数据量越大时，通过该模型获得的一级标签准确率越高。

同时结合测试结果，使用 word2vec 的模型效果优于直接使用 doc2vec 模型的结果。其主要原因是留言的文本长度较短、训练集数量较少。

3.1.2 可视化聚类结果

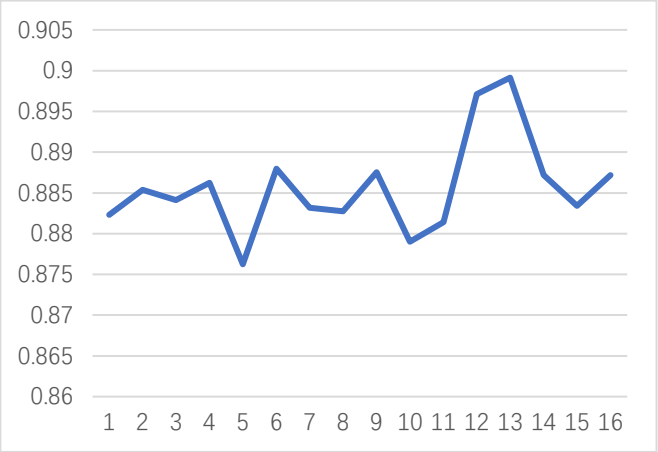
为将上述过程更为直观的体现出来，通过降维的方式将句向量在三维空间中展示：



其中不同标识、不同颜色的点标识某一类数据，可以看出其与中心的距离不同且分布于一块区域中，表明在空间中同类句子的句向量实现了较好的聚类，表明该识别模型对于一级标签的识别准确性较高。

3.1.3 模型的稳定性分析

由于在训练前对于训练集进行了随机的打乱，同时由于各类超参数的选取可能不是最优，因此对于模型的超参数在合理区间进行小幅度的修改，得到的其模型的局部最优化结果，其中为保证运行效率，采用 AUC 值进行衡量。



在选取图中的局部最优点后，AUC[11-12]在 0.89 附近波动，证明模型预测效果很好。

3.2 问题 2 的结果分析

3.2.1 对热点问题的分析

根据题目要求，生成了相应的热点问题表以及热点问题留言明细表，这两张表可以有效的帮助相关部门进行针对性处理，提升服务效率。

其中热点问题表如下表所示：

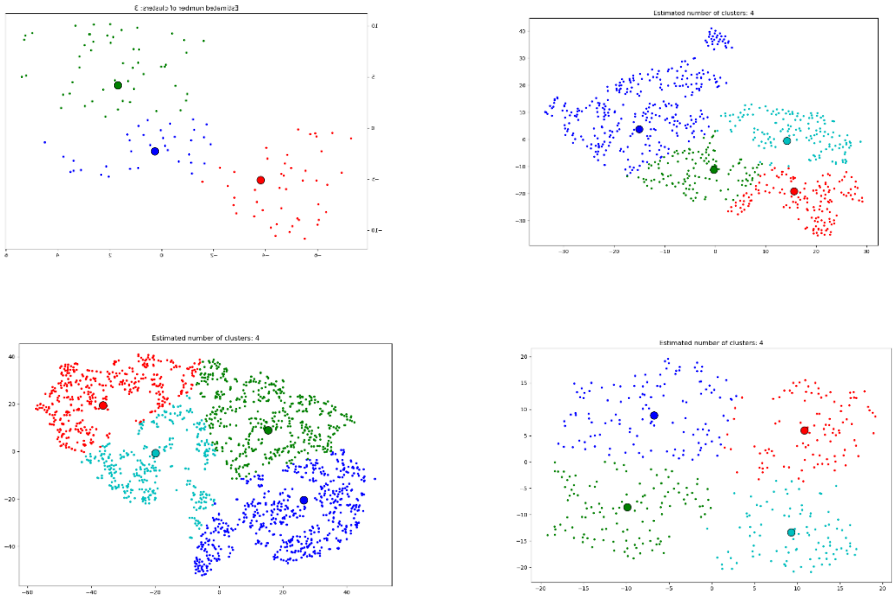
热度排名	问题 ID	热度指数	时间范围	地点\人群	问题描述
1	1	100	2019/1/8-2019/7/8	A 市 58 车贷	西地省 A 市 58 车贷恶性退出，A4 区立案已近半年毫无进展
2	2	72.2	2019/5/5-2019/9/12	A 市 A5 区汇金路五矿万境 K9 县业主	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
3	3	47.9	2019/4/11	A 市附近居民，入学问题	反映 A 市金毛湾配套入学的问题

4	4	35.8	2019/8/23 -2019/9/6	A4 区绿地海外滩	高贴线路过近，噪声影响
5	5	26.1	2019/7/21 - 2019/12/4	A 市万科魅力之城居民	魅力之城小区临街门面油烟与噪声扰民

这其中由于竞赛所给出的数据中将真实的地名进行了修改，导致爬取数据时缺乏方向性，全国的道路数据过于庞大，会导致识别过程极长，而在现实情况中，各地区是确定的，数据量可以有效降低会使得对于地区热点问题的分析更为精确。对于地点的识别中，采用的字典巨大，因此对于地名的识别过于精细导致识别出来的地点存在分散的问题，基于此我们对于识别出来的地点重新排序、去重，最后选取地点中的前三条作为识别的具体地点。

3.2.2 热点问题可视化结果

由于采取的是无监督 Meanshift 进行聚类分析，此后再进行的地点匹配，同样采取降维的方式实现可视化。结果如下：



上述聚类结果是对于同一一级标签前提下进行的聚类分析，而可视化

结果表明，聚类后同一标签下存在 3、4 种问题，而对于该类进行语义的分析与总结，就可以得到这一热点问题的具体内容。由于该系统是在同一一级标签前提下进行的聚类分析，因此样本数量较为缺乏，当样本数据更多时，代码的运行结果会更好，这也表现出该模型在处理更为庞大的数据时会具有更高的适应性。

3.3 问题 3 的结果分析

答复的相关性、完整性和可解释性仍与句向量有着较大的关系，因此采用训练后的 word2vec 模型转化句向量并分析句向量，可得到答复的答复的相关性、完整性和可解释性评分。基本上可以得出对答复的比较客观的评价，且评分较为合理。

通过对生成文件中的数据进行分析，表明该方法能够完成对于答复的综合性评价，以得分前五的答复与得分后五的答复进行比较,可以发现综合性评价较低的数值与综合性评价较高的数值差别很大，因此可以把综合性评价作为指标来综合判断答复的相关性、完整性和可解释性。

综合评分前五	综合评分后五
0.78	0.0287
0.79	0.0356
0.79	0.0368
0.8	0.0708
0.81	0.0735

4. 总结与展望

综上所述，我们提出了“智慧政务”的文本挖掘模型，这套模型可以实现对于留言的分类标识，并在生成模型的基础上利用网络爬取的地点信息，实现对于某一地区的热点问题的提取并生成热点表。我们的模型在线下测试集上取得了良好的结果，达到了题目的要求，总体而言，我们的工作在于实现词向量到句向量的转换，并实现利用 SVM 以及 meanshift 算法实现聚类。在接下来的工作中，我们会就我们的模型进行优化，解决目前运行时间长的问题，实现更加高效的识别。

参考文献

[1] 大数据技术在智慧城市管理中的应用探究[J]. 任志孟. 信息系统工程. 2019(11)

[2]Rumelhart D E, Hinton G E, Williams R J. [J].Learning representations by back-propagating errors. nature, 1986, 323(6088): 533~534

[3] V. MURALIDHARAN, V. SUGUMARAN. A comparative study of Naive Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis[J]. 2012,12(8).

[4] T. joachins[J]. Making large - scale svm learning practical. 1999.

[5] JOLLIFFE I T.Principal component analysis [M].2nd Edition.New York：Springer，2002.

[6] LI W，SHI T，LIAO G，et al.Feature extraction and classification of gear faults using principal component analysis [J].Journal of Quality in Maintenance Engineering，2003，9(2)：132- 143.

[7] WANG P，LONG Z，DANG N.Multi-model switching based fault detection for the suspension system of maglev train [J].IEEE Access，2019，7：6831- 6841.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv.org, vol. cs.CL. 10-Oct-2018.bert

[9]辛霄,范士喜,王轩, 等.基于最大熵的依存句法分析[J].中文信息学报,2009,23(2):18-22. DOI:10.3969/j.issn.1003-0077.2009.02.002.

[10] 周德懋, 李舟军. 高性能网络爬虫:研究综述[J]. 计算机科学, 2009(08):26-29

[11]HSIEH F, TURNBULL B W. Nonparametric and semiparametric estimation of the receiver operating characteristic curve[J]. The annals of statistics, 1996, 24(1): 25–40.

[12]ELKAN C. The foundations of cost-sensitive learning[C]//Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, WA, 2001: 973–978.