一种基于机器学习的智慧政务信息处理方案

摘要：

在如今的信息时代，网上问政已经成为了群众参与政治生活的重要方式，但随着信息的爆炸性增长，传统的政务系统渐渐无法满足现有需求。近年来，随着自然语言处理技术研究的不断深入，许多“人工智能+政务”的新模式也随之出现，人工处理难以企及的效率和精度如今成为了可能。本文从群众留言分类、热点问题挖掘、答复信息评价三类问题着手，对政务系统中涉及的自然语言处理技术展开研究，为智慧政务系统的构造提供了一些解决方案。

在数据预处理阶段，先对数据集结构展开分析，然后基于数据在分析中表现出的数据特点，进行针对性的数据清洗、记录去重、汉语分词和停用词去除。在文本特征抽取的相关工作中，分别使用 TF-IDF 模型和 Word2Vec 模型对词与文本进行了向量化。

在群众留言分类问题中，选用 kNN 分类模型，在 8977 条带标注语料的基础上建立了分类器。经过一系列优化后，分类器的平均F_1值达到 0.8835，与其他同类的解决方案相比，达到了更好的效果。

在热点问题挖掘中，选用 Mean-Shift 聚类模型对 4222 条有效留言进行聚类，在筛去了簇中留言较少的离群点后，得到了 63 个热点问题簇。针对热点的热度评估问题，建立合理的热度评价体系，并通过层次分析法与人工评估混合的方式优化了指标权重。为了提高热点问题的可解释性，使用 LDA 主题模型对热点问题簇进行了主题抽取，作为热点问题的摘要，方便针对热点问题的理解与决策。

在答复信息评价问题中，引入了人工标注与半监督学习结合的模式，对那些计算机不能直接判断的指标，先使用人工评价并标注部分数据，再交由半监督学习算法 Label Propagation 从已标注数据和未标注数据中学习评价的潜在模式。评价体系的计分采用了常模参照计分法，具有更优良的健壮性和可解释性。

最后的实验部分呈现了方案构造中的技术细节，评估了上述几种解决方案的效能，并与其它同类解决方案作出对比。

关键词：k 最近邻，均值漂移，半监督学习，自然语言处理

目录

一、引言 3
二、技术路线 3
三、智慧政务信息处理方案 5
3.1 数据分析与预处理 5
3.1.1 数据特征分析 5
3.1.2 数据预处理 7
3.2 文本特征抽取 8
3.2.1 词的向量化表示与词频-逆向文档频率模型 8
3.2.2 词嵌入与 Word2Vec 模型 11
3.3 留言分类器的建立 12
3.3.1 短文本的向量化表示 13
3.3.2 K 最近邻分类模型 13
3.3.3 分类器评价指标 13
3.3.4 超参数优化 14
3.4 热点问题聚类 15
3.4.1 Mean-Shift 聚类模型 15
3.4.2 聚类效果评价与优化 16
3.4.3 热度评价指标 16
3.4.4 热点问题摘要 17
3.5 答复意见评价体系 18
3.5.1 评价指标的设置 18
3.5.2 标记传播模型 18
3.5.3 常模参照计分法 19
四、实验结果分析 19
4.1 预处理与特征抽取实验 19
4.2 分类器的建立与评价 20
4.3 聚类与热度排序实验 21
4.4 答复信息评价 23
4.4.1 时效性指标与相关性指标 23

4.4.2 完整性指标与可解释性指标 23
五、总结与展望 25
参考文献 26
附录 26

一、引言

网络的极速发展，为公民提供了更多的政治参与的机会。一项调查[1]显示，87.9%的网民非常关注网络监督；当遇到社会不良现象时，99.3%的网民会选择网络曝光；2010 年，江西省省长和其在任的省级领导通过人民网、中国江西网等与全球网民在线交流，点击量突破了 1000 万人次。由此可看出，微信、微博、市长信箱、阳光热线等网络问政平台已经逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，但与此同时，各类社情民情相关的文本数据量不断攀升。百湖民生问政平台自开通一年半后收到的问政留言共 11110 条，有效留言 9300 条，15 年 2 月份，问政平台共收到 613 条留言，与去年同期相比增长一倍。海量的信息也给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大的挑战。

同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。而信息检索为上述问题提供了解决方案，利用划分重要词汇来快速筛选出民生热点，不仅减轻人工留言划分的工作量，还能快速提高答政率。

基于自然语言处理技术，本文针对不同任务与不同的数据特征建立合适的模型。在完成指定任务之后，还将所用的方案与其他同类方案作出对比，在操作的便利性、结果的准确性上进行分析和讨论，进一步验证了方案的有效性和可行性。

二、技术路线

为了提高工作的效率，我们依照软件工程的基本思想将处理系统分解成了数个相对独立的处理模块，模块内部独立完成一部分工作，模块间使用公开接口通信，如下图所示。

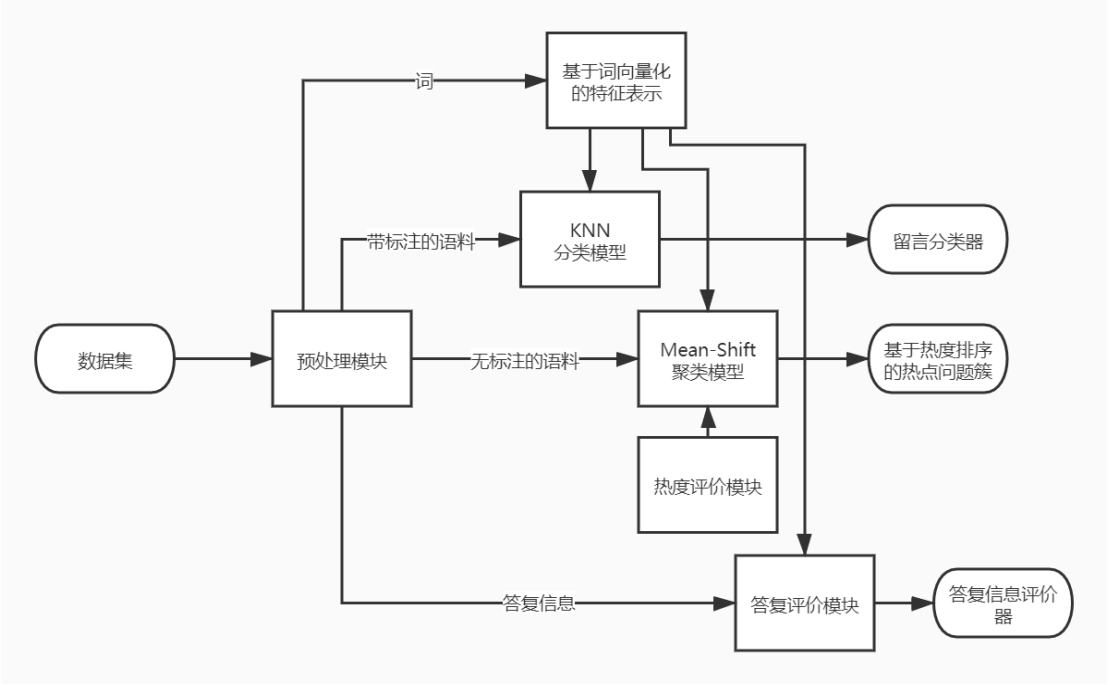


图 1 技术路线图

与数据集直接相连的是预处理模块，该模块具有各类数据接口，内部进行数据清洗等预处理工作，并建立 TF-IDF 和 Word2Vec 两种词向量化模型，以供后续模块使用。

留言分类模块完成留言的分类问题，它从预处理模块中获取带有标注的语料，基于 kNN 分类模型与词的 Word2Vec 向量表示，最终得到一个面向留言的多分类器。

聚类模块完成热点问题的挖掘问题，它从预处理模块获取无标注的语料，基于 Mean-Shift 聚类模型与词的 Word2Vec 向量表示。聚类模型会给出若干问题簇，而热点问题簇的排序功能由热度评价模块完成。该模块还具有子模块主题抽取模块，基于 LDA 主题模型与词的 TF-IDF 表示，从热点问题簇中抽取主题作为问题摘要。

最后，我们通过分析答复信息与其他数据的关系，建立了答复信息评价体系，封装在答复评价模块中，最终形成一个答复信息评价器。

三、智慧政务信息处理方案

3.1 数据分析与预处理

3.1.1 数据特征分析

为了对数据集有更加全面的了解、方便后续的数据处理，需要对数据集进行数据分析。我们使用 Python 工具读取了附件二、三、四中的记录，针对表属性中的数值分布、附件与附件间的结构差异等做出分析。通过分析，发现三个附件的前五个属性（编号、用户、主题、时间、详情）是共有的，而“一级标签”属性为附件二独有，“点赞数”与“反对数”属性为附件三独有，“答复意见”与“答复时间”为附件四独有。

在共有的属性中，“留言主题”和“留言详情”是研究的重点，对这两个属性的具体分析如下表所示。

表 1 数据特征信息表

	原始记录条数	“留言主题”长度/字			“留言详情”长度/字		
		最大	最小	平均	最大	最小	平均
附件二	9210	48	2	19.55	12420	12	401.74
附件三	4326	37	1	20.19	6999	9	351.624
附件四	2816	36	8	18.57	3191	12	318.6

观察分析结果可得，三个附件的“留言主题”属性长度差异较小，而在“留言详情”属性的长度上存在明显差异，其中附件二中的“留言详情”字数最多，附件三次之，附件四最少。在附件三和附件四中，“留言时间”的分布（如下图所示，以附件三为例）无明显差异，留言发布最集中的时间段均为 2018 年 4 月至 2020 年 1 月。

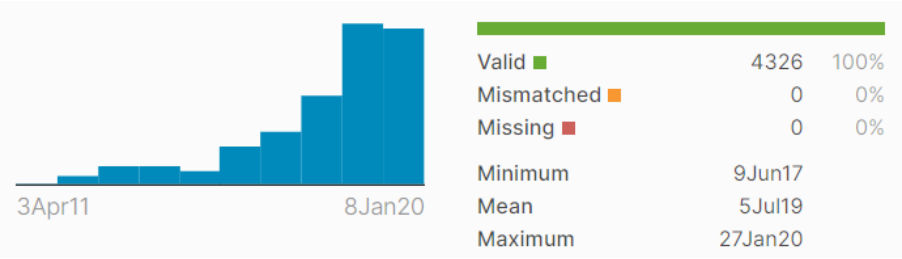
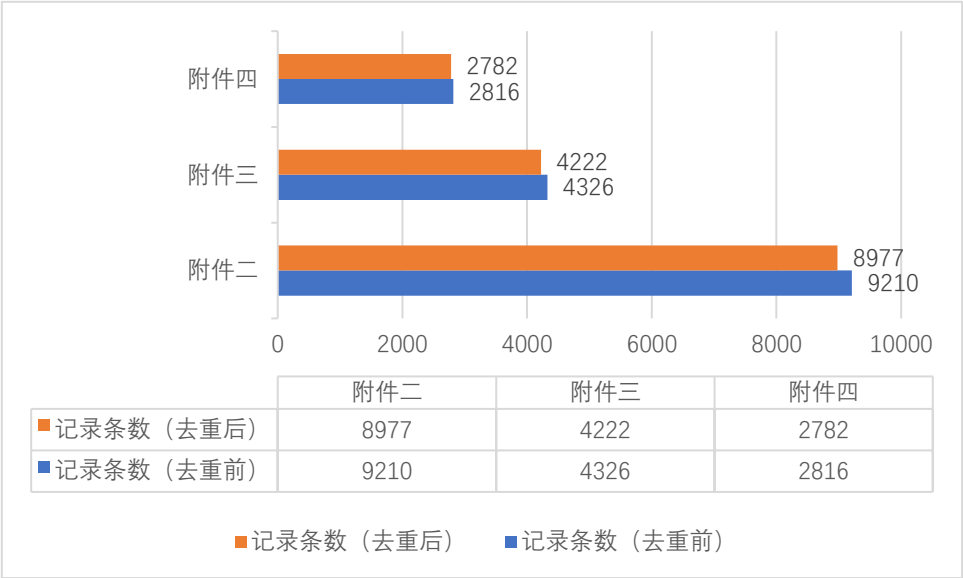


图 2 附件三中“留言时间”属性的时间分布

经过分析，我们还发现在附件二、三、四的共有属性中，“留言用户”、“留言主题”、“留言时间”、“留言详情”四个属性在记录间存在重复数据。考虑到同一个用户可以多次提交同一类型的问题的情况和留言的发布时间具有一定的密集性的特点，前三个属性出现重复是合理的；但如果“留言详情”的内容完全一致，则说明出现了重复提交或者记录错误，这类重复出现的记录将对我们后续的数据处理产生影响，因此需要针对“留言详情”属性对记录去重。附件二、三、四中“留言详情”重复情况见下图。



	附件二	附件三	附件四
记录条数（去重后）	8977	4222	2782
记录条数（去重前）	9210	4326	2816

图 3 附件二、三、四中记录的重复情况

另外，在三个附件中都没有出现数据缺失的状况，因此可以省去数据填充的工作。我们还对“点赞数”、“答复意见”等某个附件特有的属性进行了数据分析，在本文实验部分相应的章节还会提到，在此不再赘述。

3.1.2 数据预处理

预处理的第一步是数据清洗，针对数据特征分析中得到的结论，在使用Python 工具读取留言详情时,抛弃文本前后用于格式输出的空白字符（空格、制表符、换行符等），并针对留言详情进行了去重，对于留言详情完全一致的记录仅保留第一条。

第二步是进行分词。分词是几乎所有自然语言处理任务必不可少的一步，即将连续的句子分割成词语的组合,同时保持原有的语义不丢失。只有经过了分词，才能进行语义分析等后续工作。由于自然语言多样性和语义不确定性的特点，分词在自然语言处理中一直是一个经典的议题。而中文分词与英文分词的工作重心又有所不同。在英语写作习惯中，句子中的每个单词之间由空格分开，所以分词工作相对简单，工作重心一般在词型的转换上。而在汉语语言中每个词语之间没有固定的分隔，同时又受到汉字语义多样性的影响，会出现同一句话在不同的语境下有不同的语义的情况，可能产生歧义。例如语句“希望西门的小吃店能有多少就有多少”可能表达对小吃店的便利的赞赏，也可能表达的是对小吃店扰民的困扰。另外，不同的划分方式也可能产生截然不同的语义。因此中文的分词工作变得相对复杂。为此，我们选用了 Python 平台的中文分词工具 jieba[2]，该工具的“全模式”分词能将语句中词汇可能的拆分组合全部保留,并存储到词链表中。下表为使用 jieba 工具的“全模式”分词举例。

表 2 “全模式”分词示例

处理前	处理后
‘我来到北京清华大学’	’我’，’来到’，’北京’，’清华’，’清华大学’，’华大’，’大学’

第三步是去除停用词。在分词工作的过程中，常常会切分出一些功能词，这些功能词十分常见，但大多没有实际含义，如“啊”“呀”等语气词，“一个”“一只”等数量词和“不但”“不管”一类的连词等。这些词称作停用词，停用词不仅会占用存储空间，降低搜索效率，而且对文本处理没有实际意义，甚至可能作为噪声产生干扰，因此需要对停用词进行检索和去除。针对汉语停用词

的研究在国内已经有了较丰富的公开成果，我们采用哈工大停用词表，百度停用词表和四川大学机器智能实验室停用词库，将其进行整合，生成模型中所运用的停用词库，共有 2312 个词。下表为随机抽取停用词库中的 5 个停用词。

表 3 停用词举例

’与’	’亦’	’。’	’③’	’$’

我们使用 Python 工具在文本分词得到词链表中检索停用词库中的停用词，将它们从词链表中剔除。

3.2 文本特征抽取

3.2.1 词的向量化表示与词频-逆向文档频率模型

要在计算机系统中处理汉字，就必须对其特征进行编码，而在自然语言处理中，大部分的汉字特征可以被舍弃以提高处理效率，例如汉字字形特征就对大多数自然语言处理系统没有意义。目前主流的方法是用词向量来表示词的特征。

独热模型（one-hot）是最为经典的词向量化模型，使用 one-hot 方法进行词向量化的具体演示如下。

假设语料库中有两句话：

1. 小明喜欢阅读
2. 小李喜欢跑步

首先，对每句话进行分词，并编号：

表 4 独热模型示例 1

词	编号
小明	1
小李	2
喜欢	3
阅读	4
跑步	5

然后，为每个词分配一个 one-hot 向量，使得该向量仅在一个分量上为 1，其他位置上为 0，由此该向量可以唯一标识这个词，即完成了词的向量化表示，举例如下：

表 5 独热模型示例 2

词	编号	One-hot 向量
小明	1	（1，0，0，0，0）
小李	2	（0，1，0，0，0）
喜欢	3	（0，0，1，0，0）
阅读	4	（0，0，0，1，0）
跑步	5	（0，0，0，0，1）

若要进行语句的向量化，可以考察语句中出现的词语，若某个词语出现，则其编号对应的向量分量为 1，否则为 0，这样得到的句向量也能够唯一标识一个句子，即完成了语句的向量化：

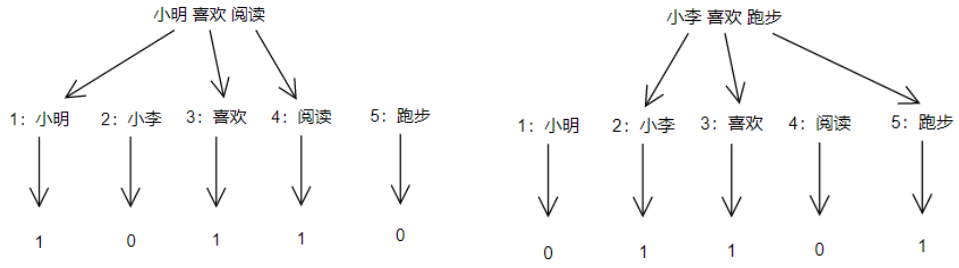


图 4 独热模型示例 3

即：

1.“小明喜欢阅读”表示为（1，0，1，1，0）

2.“小李喜欢跑步”表示为（0，1，1，0，1）

虽然 one-hot 编码方法处理简单，但这种处理得到的特征向量是稀疏的。由于词向量的维数与词典的大小相等，当词的数量增多时，词向量的维数也会增加，计算量会呈指数性增长，最终导致“维数灾难”。稀疏的向量不仅浪费了存储空间，模型的运算速度还会因此受到限制。此外，one-hot 编码方法中，每个词向量都是正交于其他向量的，并未考虑词之间的语义关联。因此基于该方法的模型得到的结果准确度较低，已经较少使用。

词频-逆向文档频率模型（TF-IDF）[3]改进自 one-hot 模型，该模型认为某个字词的重要性与它在一篇文章中出现的次数成正相关，同时与它在语料库中出现的频率成负相关；即如果某个词在当前文本中频繁出现，而在语料库的其他文本中很少出现，则说明该词很能代表当前文本。其中词频（Term Frequency，TF）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数(Term Count)的归一化，以防止它偏向较长的文本。对于在某一特定文本里的词来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

式子中的分子是指某一选定词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和。

逆向文件频率（Inverse Document Frequency，IDF）是一个衡量词是否具有普遍重要性的重要指标。某一特定词语的逆向文件频率，可以由所有文件的数目除以包含该词语的文件的数目，再将得到的值取以 10 为底的对数得到：

$$idf_i = lg\frac{|D|}{|\{j: t_i \in d_j\}|}$$

其中|D|指的是语料库中的文件总数。

TF-IDF 值最终由以下方法计算得到：

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

3.2.2 词嵌入与 Word2Vec 模型

在向量化表示过程中，自然语言中的两个词语可能在拼写上毫无关联，但词义却有联系。在上述的 TF-IDF 模型中，重要的词语能够得到高权重，一定程度上表现了词与词之间、词与文本之间的联系。但此模型产生的向量也是稀疏的，维数亦会随语料库的增长而增大，并没有解决 one-hot 模型中，“维数爆炸”的问题。于是我们引入降维的概念，即把特征向量的维度降低，并尽可能保留原有的语义信息。

词嵌入（Word Embedding）是一类将词的语义映射到低维向量空间中的降维技术[4]，它能把表示在高维空间中的词嵌入到一个维数低得多的连续向量空间中，每个词语将被映射为实数域上的稠密向量，向量之间的欧式距离一定程度上表征了词之间的语义关系，即两个词语义越相近，在向量空间中的位置也越相近（如下图所示）。这种方法可以很好地解决稀疏向量和“维数爆炸”的问题。

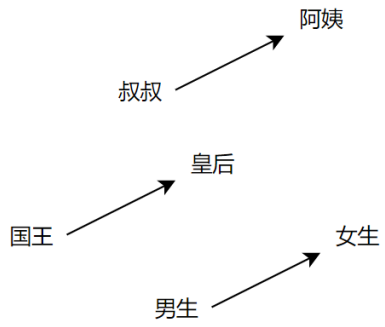


图 5 向量空间中临近的词

我们选用 Python 平台中 gensim 工具[5]封装的 Word2Vec 模型进行词嵌入训

练，该模型先将语料转化为 one-hot 编码，再输入一个单层神经网络模型，得到指定维数的低维词向量，具体实验流程见 4.1 小节。

下表为使用附件二、三、四中的语料训练 Word2Vec 模型处理之后，检索与“火车站”词义最为相近的前 10 个词（降序排列）的结果。可以发现，该 10 个词中大部分为“火车站”这个词的近义词或关联词，且极具政务类文本的风格，说明该模型的运用较为成功。

表 6 在向量空间中与“火车站”最相近的 10 个词

词	相似度
换乘	0.9867
K1	0.9864
地级	0.9829
城区	0.9810
火车站	0.9803
楚江	0.9796
地级市	0.9791
火车	0.9775
桥	0.9766
A3	0.9757

TF-IDF 模型与 Word2Vec 模型都是较为实用的词向量化模型，本文大部分词向量化方案选用 Word2Vec 模型，在 LDA 主题建模中使用 TF-IDF 模型。

3.3 留言分类器的建立

随着群众留言信息的爆炸性增长，传统的人工分类留言的模式不再适用，建立快速、精确的留言分类器成为了迫在眉睫的需求。

3.3.1 短文本的向量化表示

在前面处理中已经得到了词的向量表示，但当研究对象从词变换到由若干词组成的文本上时，同样需要一个将文本向量化表示的方法，该方法得出的文本向量需要能够集中体现文本中的词的特征。在主流的短文本向量化方法中，我们选择了基于词嵌入的 Embedding Average 方法，即得到词嵌入向量后，用短文本中包含的所有词的词向量的平均值作为文本向量；该方法实现简单，在较短文本中性能良好，符合处理任务中留言文本的特征，且继承了词向量中稠密向量、语义相关性等优点。

3.3.2 K 最近邻分类模型

经过数据集结构的分析，我们发现语料库（附件二）中每条记录都带有明确的标注，因此群众留言分类问题是有监督学习中经典的分类问题。有监督学习的研究开始较早，如今已有非常丰富的成果，常用的模型有 k 最近邻、支持向量机、朴素贝叶斯等。

这里我们使用 k 最近邻模型（k-Nearest Neighbor，kNN）[6]构造留言分类器。kNN 模型认为，若当前样本周围的 k 个已标注样本中属于类别 A 的样本最多，则可以认为当前样本属于类别 A。kNN 模型在类域重叠较多的样本空间中表现突出，很适合本任务中彼此重叠的词向量空间。

3.3.3 分类器评价指标

对于分类器的评价，我们使用的是$\mathrm{F}_1 - Score$方法。该评价指标演化自两个子指标：精确率与召回率。精确率（Precision）指的是某类别正确分类样本数与总样本数之比，计算公式为：

$$\mathrm{P} = \frac{TP}{TP + FP}$$

召回率（Recall）指的是正确判断某类的样本个数与属于该类的实际样本总数之比，计算公式为：

$$\mathrm{R} = \frac{TP}{TP + FN}$$

其中 TP（真阳性）指的是正样本中被正确预测为正样本的个数，FP（假阳性）为负样本中被错误预测为正样本的个数，TN（真阴性）指的是负样本中被正确预测为负样本的个数，FN（假阴性）为正样本中被错误预测为负样本的个数。

单独考虑这两个子指标都不能得到很好的分类效果。如果单独提高召回率，即希望更多的相关文档被检索到，就要放宽检索的条件，最终导致一些不相关的样本的出现，导致准确率下降；而若只提高准确率，会使检索策略严格，导致一些相关文档无法检索到，导致召回率下降。因此，很多情况下我们需要综合权衡这两个指标，于是诞生了综合指标$\mathrm{F}_1 - \mathrm{score}$，即综合考虑精确率与召回率的调和值：

$$\mathrm{F}_1 = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 \times (Precision + Recall)}$$

在分类任务中，精确率和召回率都很重要，权重相同，因此我们取$\beta = 1$。同时，由于本次任务是多分类问题，对分类器的评价应当综合考察其在各个类别的表现，这可以通过评价的平均值来实现，由此我们得到最终$\mathrm{F}_1 - \mathrm{score}$的计算公式：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

其中P_i为第i类的精确率，R_i为第i类的召回率。

3.3.4 超参数优化

在机器学习中，开始学习之前设置好、而不是在训练过程中得到的参数称为超参数，超参数定义了关于模型的更高层次的概念，例如模型复杂性或学习能力。为了提高学习的性能和效果，需要对超参数进行优化，给学习机选择一组最优超参数。常见的超参数优化算法有网格搜索、贝叶斯优化、随机搜索、梯度下降算法等。我们先选用网格搜索，即对指定的参数空间子集进行离散扫描，我们以分

类器的F_1值作为优化的评价函数，在 Word2Vec 和 kNN 组合模型超参数的参数空间中搜索了一万余种参数组合，最终得出了一组较优的超参数。

然而由于网格搜索笨重、效率低下的缺点，在参数增多时搜索耗时会指数性上升，不适合进一步优化。于是我们进一步使用了贝叶斯优化算法[7]对分类器进行优化。贝叶斯优化算法采用高斯过程，在选取参数时会考虑之前的参数信息，不断更新先验，从而大大减少参数迭代的次数。另外，贝叶斯优化器还能避免参数组合落入局部最优。

具体实验过程见本文 4.2 小节。

3.4 热点问题聚类

在群众留言中，常常会出现多条留言信息指向同一个问题的情况，像这样在一段时间内被集中反映的问题可以称为热点问题。热点问题的挖掘可以帮助决策者快速了解民意，一定程度上提高了问题处理的效率。

3.4.1 Mean-Shift 聚类模型

考察任务需求与数据分析的结果，我们发现需要对语料库中的无标注数据进行不确定类别数的划分，这是无监督学习中经典的聚类问题。聚类过程与分类相似，都是将同类的样本划分到一起，但前者不同之处在于其没有预设的类别和带标注的数据，需要算法从数据中学习。常用的聚类算法有基于划分的 K-Means 算法[8]、基于密度的 Mean-Shift 算法[9]等。

基于划分的 K-Means 模型在开始学习前需要指定待划分的类别数，但在本任务中，留言信息众多且可能存在大量离群点，类别数即问题簇的个数难以估计，因此我们选用了 Mean-Shift 模型。

Mean-Shift 聚类模型也称为均值漂移模型，该模型的运行过程即是均值点不断迭代的过程，初始均值点被散布到样本空间中的随机位置，每次迭代时先算出当前点周围一定半径内样本点的均值，称为偏移均值，然后移动当前点到其偏移均值，然后以此为新的起始点，继续移动。若干次迭代后，均值点会趋于稳定，当均值偏移小于某个临界值时迭代结束。因此我们只关心簇半径参数 bandwidth，而无需关心具体的类别数。

3.4.2 聚类效果评价与优化

聚类过程属于无监督学习，由于其数据无标注的特点，不像有监督学习算法那样具有预设置的学习目标和精确的测试集，也就缺乏统一的衡量标准。在本次聚类任务中，我们使用抽样调查与人工评估结合的方法对聚类效果进行评估。在单次聚类的结果中，随机抽取 10 个问题簇，阅读其中的留言文本，考察其内部相关性与问题簇的大小。同一个问题簇中的留言相关度越高，则说明聚类的效果越好，反之说明聚类的效果越差。

通过上述的评价方法，我们逐步调整了 Mean-Shift 模型中的簇半径参数 bandwidth，直至得到令人满意的聚类效果。

3.4.3 热度评价指标

对留言进行聚类处理后，得到了一些含若干留言的热点问题。但是不同问题簇对决策者的价值是不同的，若能按照问题热度将问题排序，就能让价值高的问题优先被处理，大大提高决策的效率。而若要对热点问题的价值进行排序，就需要制定热度评价指标。

通过分析热度影响因素，将热度评价问题分解为 4 个衡量评价指标：问题内的留言条数、留言发布时间密集程度、热点问题内总点赞数和总反对数。前两个指标分别通过热点问题内的留言计数、热点问题内的留言的发布时间方差的倒数进行计算，后两个指标已在附件三中给出，只需在热点问题内部求和即可得到。将问题簇内留言条数少于 3 条的问题簇视为聚类中的离群点，在热度评价前剔除，避免其作为数据噪声产生干扰。

为了确定上述四个指标的对热度影响的权重，我们引入了层次分析法[10]，将四个指标的重要性进行分级与赋值，制作问卷并收集问卷填写者所打出的分值，对分值进行加权平均计算出各个指标的权重值，初步确定四个衡量指标之间的权重关系。四个评价指标的加权求和即为最终该热点问题的热度得分，也就是我们进行热点问题排序的依据。

初步确定指标权重后，将其输入热点评价模块，对热度排序结果进行人工评估，经过若干次微调与重新评估后，逐步得到最终的指标权重。

3.4.4 热点问题摘要

至此已经完成了对同类留言的归类和热度评估，但每个热点问题中的留言可能多达几十条甚至上百条，逐条阅读留言来归纳问题的性质十分耗费时间，不符合解决问题的初衷。如果程序能够自动生成摘要来归纳留言集中反映的问题，无疑大大提高了人工复核及有关部门工作人员的处理效率。

针对这个问题，我们引入了隐迪利克雷分布主题模型（Latent Dirichlet Allocation，LDA）[11]。LDA 模型是一种基于假设分布（隐迪利克雷分布）的文档主题生成模型，核心是由词、主题和文档三层结构组成的三层贝叶斯概率模型。该模型认为，每个文档由若干主题按照一定的概率分布组成，每个主题又由若干词按照一定概率分布组成，可以认为组成一个文档的主题中概率最大者最能体现该文档的核心信息。

我们对每个热点问题（内含若干条留言）进行了 LDA 主题建模，使用模型抽取出的主题中的概率最大者作为该热点问题的问题摘要，该主题具体表现为若干关键词的序列。

表 7 使用 LDA 模型抽取摘要示例

簇号	留言主题	摘要
1	’居住在地铁 3 号线 A7 县松雅西地省站西北方向 10 万民众的心声’	″地铁″″号″″线″″东四″″四路″″小区″
	’反映 A 市地铁 3 号线松雅西地省站地下通道建设问题’	
	’反映 A 市地铁 3 号线松雅西地省站西北方向 10 万民众安全问题’	
	’反映 A 市地铁 3 号线松雅湖站点附近地下通道问题’	

实验表明，这些关键词能够较好地反映热点问题中的留言信息。

3.5 答复意见评价体系

答复意见评价体系的建立有利于监督体系的完善，提高市民在决策中的参与感和满意度。但若要对答复意见进行逐条人工评价，不仅耗时耗力，还会大大增加监督体系的运营成本，因此制定合理的评价指标、进行答复信息自动化评价具有现实意义。

3.5.1 评价指标的设置

针对答复信息的评价，我们将其分解为以下四个子评价指标：答复时效性、答复相关性、答复完整性与可解释性。其中时效性指标具体通过计算留言发布至收到答复所经过的时间得到，相关性指标通过计算留言文本与答复文本之间的语义相关度，即计算两个文本向量在向量空间中的距离得到。

相比前两个指标，答复完整性与可解释性的概念较为模糊，难以得到定量评价，于是将这两个指标进一步分解。我们将答复完整性指标又分解为以下三个子指标：答复中是否有对问题定性、答复中是否有解释问题发生的原因、答复中是否有说明将实施的措施；可解释性指标分解为以下若干子指标：答复中是否引用了正式文件作为证明、答复中是否引用了具体数据作为证明以及其他的合理证据。

针对以上指标及细化指标的定义，我们在数据集中选取了一部分答复文本，对完整性和可解释性这两个指标进行手工标注；若答复文本满足完整性子指标三项中的两项或以上，则该指标记 1 分，否则记 0 分；若满足可解释性子指标中的任一项，则该指标记 1 分，否则记 0 分。

3.5.2 标记传播模型

评价指标建立以后，完整性和可解释性这两个指标的评价被转化成分类问题。要训练针对这两个指标的分类器，就要有相关的数据集。但由于答复文本的数量较大，要对其完整性和可解释性全部进行手工标注是不现实的，因此引入了半监督学习模型：标记传播模型。

标记传播算法（Label Propagation Algorithm，LPA）[12]是一种基于图的半监督学习模型，它基于少量已标注的数据和大量无标注的数据进行学习，因此只需

我们手工标注答复信息集中的一小部分，就足以学习机完成学习。另一方面，我们仅对这两个较模糊的指标作定性评价，实际上是对每个指标构造一个二分类器，系统的复杂度得到简化。

3.5.3 常模参照计分法

至此我们已经得到了答复信息每一项指标的得分，根据预设的权重将它们加权求和即可得到总分。但答复评价得分的可解释性上仍存问题：分数分布存在可能的不均匀情况，数值相近的分数能否说明答复的质量也相近？前两个指标（时效性和相关性）的取值范围难以评估，导致最终得分的上限和标准参考线难以划分，这一定程度地影响了分数的可解释性。

借鉴标准化考试的计分方式，我们引入了常模参照计分法。该方法基于正态分布理论，以个体在团体中的相对位置作为计分的标准，适用于个体差异较大的团体。在本次任务中，采用附件四语料中全体答复信息的得分作为样本总体，每个个体的标准得分通过基于样本总体的常模参照计分得出。

四、实验结果分析

4.1 预处理与特征抽取实验

使用 Python 平台的开源 Excel 读写工具 openpyxl 读取附件二、三、四，使用 Python 内置方法完成初步的字符串清洗，并使用开源中文分词工具 jieba 完成汉语分词工作。三个附件中“留言详情”属性分词后所含的词数与词长情况如下表所示。

表 8 “留言详情”属性分词后的数据特征

词数/词	最大	最小	平均
附件二	3766	3	128.8
附件三	2013	3	110.84
附件四	1097	3	101.1

表 9 分词后词的数据特征

词长/字	最大	最小	平均
附件二	51	1	2.04
附件三	22	1	2.08
附件四	60	1	2.05

下一步是停用词的去除，我们将停用词表存储在文本文件中，使用 Python 内置方法将其读取到列表。经过分词的“留言主题”和“留言详情”属性被存储在另一个列表中，我们使用停用词表对其进行词匹配，如检索到停用词则从词列表中剔除。

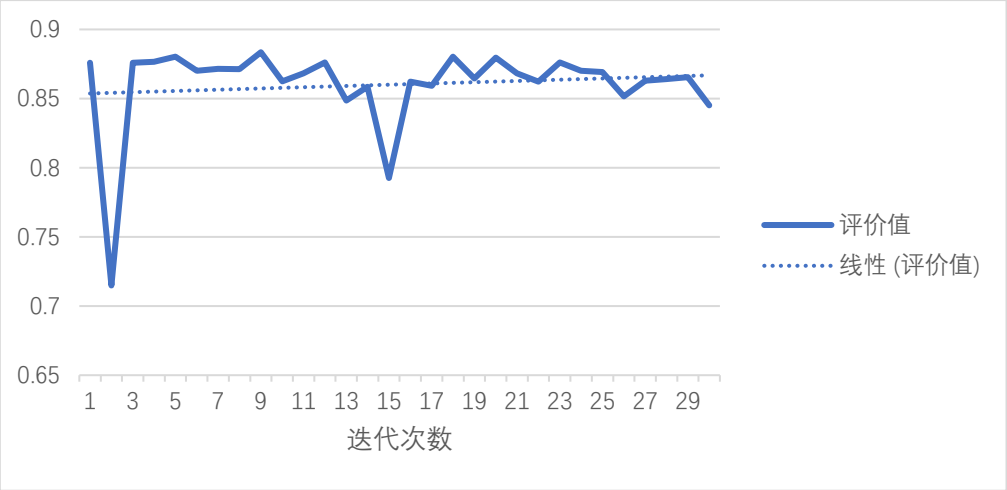
在特征抽取中，使用 Python 工具 gensim 的 Word2Vec 模块中封装的 Word2Vec 模型提取词向量，设置超参数 sg=1，以使用更适合当前语料库的 skip-gram 变体模型。使用来自附件二、三、四的全部 16,352 条记录的“留言主题”和“留言详情”文本作为语料，进行 Word2Vec 模型的训练，并将得到的模型存储到磁盘，以使后续处理流程在使用词向量时无需再次训练，节省大量时间。

4.2 分类器的建立与评价

使用 Python 工具 sklearn 中 KNeighborsClassifier 模块封装的 kNN 分类模型构造分类器，使用附件二中的 9210 条带标注的留言记录作为语料库。词向量化模型使用上述的建立在全部 16,352 条记录上的 Word2Vec 模型，文本向量使用 Embedding Average 方案。单次训练时从语料库中随机抽取 70%的留言记录作为训练集，剩余的 30%记录作为测试集，以分类器的 5 次交叉验证F_1均值作为评价函数。

使用 Python 工具 bayes_opt 中 BayesianOptimization 模块封装的贝叶斯优化器对分类器的超参数进行优化，将分类器评价函数与参数空间嵌入到贝叶斯优化器中，优化过程评价值的迭代如下图所示。

图 6 贝叶斯优化器的迭代过程



经过贝叶斯优化器的 30 次迭代，最终得到的分类器的F_1值为 0.8835。

4.3 聚类与热度排序实验

针对附件三中的 4222 条有效记录，使用 Python 工具 sklearn 中 MeanShift 模块封装的均值漂移模型进行聚类。先将每条记录的“留言主题”和“留言详情”属性进行拼接，使用前面得到的 Word2Vec 模型完成词向量化，通过 Embedding Average 方法计算出文本向量，作为每条留言记录的特征向量。

多次调整簇半径参数 bandwidth 后，发现其取值为 4 时能够得到令人满意的聚类效果。此时部分问题簇的内容如下表所示。

表 10 聚类结果举例

问题簇号	留言主题
1	’A3 区西湖街道茶场村五组什么时候能启动征地拆迁’
	’A3 区西湖街道茶场村五组何时启动拆迁？’
	’A3 区西湖街道茶场村五组什么时候能拆迁’
	’咨询 A3 区西湖街道茶场村五组的拆迁规划’
2	’A 市限卖房产政策一刀切’
	’A 市限卖房产是否侵犯了普通公民的正当权益？’

经过上述过程，得到了 3923 个问题簇，去除那些留言条数少于 3 条的簇（视为离群点）后，剩余有效问题簇 63 个。使用热度评价模块对它们进行评分，并按照热度降序排序，得到的前三个问题簇如下图所示（每个簇仅展示前三个留言的留言主题）。

表 11 热度前三的问题簇中的部分留言

簇号	留言主题
1	’请问 A4 区公安派出所对 58 车货一案办案的进度如何了’
	’严惩 A 市 58 车贷特大集资诈骗案保护伞’
	’再次请求过问 A 市 58 车贷案件进展情况’
2	’居住在地铁 3 号线 A7 县松雅西地省站西北方向 10 万民众的心声’
	’反映 A 市地铁 3 号线松雅西地省站地下通道建设问题’
	’反映 A 市地铁 3 号线松雅西地省站西北方向 10 万民众安全问题’
3	’问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划’
	’问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划’
	’问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划’

使用 Python 工具 gensim 的 LdaModel 模块中封装的 LDA 主题模型对每个有效的热点问题簇进行主题建模，其中词的向量表示使用 TF-IDF 模型，设置超参数 num_topics 为 1，以便 LDA 模型抽取出一个对簇内留言普适的主题，得到热度排名前三的热点问题簇的主题如下表所示。

表 12 热度排名前三的问题摘要

簇号	主题（关键词摘要）
1	58，受害，毛，车，贷，说
2	地铁，号，线，东四，小区，四路
3	商业，厂房，昌，中心，闲置，塘

最后使用 Python 工具将热点问题的相关信息输出到 Excel 表格中，根据问题摘要手工填写问题涉及的“地点/人群”，并优化“问题描述”一项的表述，其它表项由程序自动生成。

4.4 答复信息评价

4.4.1 时效性指标与相关性指标

在答复信息评价问题中，为了避免丢失答复文本，没有根据“留言详情”去重，而是保留了附件四中的全部 2816 条原始记录。

使用 Python 内置工具将“留言时间”“答复时间”属性中的字符串转化为 datetime 内置对象，两者的时间差值即为时效性指标的评价值。导入在特征抽取一步生成的 Word2Vec 模型，将每条记录的“留言详情”与“答复意见”词列表向量化，文本向量化使用 Embedding Average 方案，使用科学计算工具 numpy 计算“留言详情”与“答复意见”文本向量在向量空间中的欧氏距离，将其作为相关性指标的评价值。

4.4.2 完整性指标与可解释性指标

这两个指标由于主观性较强，要求计算机直接评价比较困难，于是我们根据指标标准人工标注了 100 条留言文本，标注数据的得分分布如下图所示。

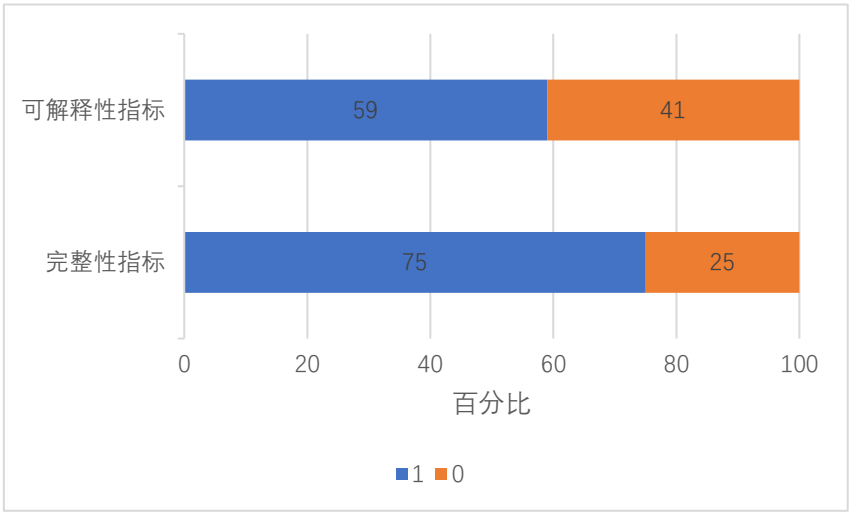


图 7 人工标注数据的得分分布

我们从标注的 100 条留言文本中随机抽取了 70 条文本，与 2716 条未标注文本组成训练集，剩下的 30 条标注文本作为测试集，词向量化方案使用 Word2Vec，文本向量化方案使用 Embedding Average。通过 Python 工具 sklearn 中 LabelPropagation 模块封装的标记传播模型进行训练，多次随机测试得到“完整性”指标预测的平均准确率平均为 0.82，而“可解释性”指标预测的平均准确率在 0.69 左右波动。针对后者准确率较低的情况，我们更换了对噪声更具健壮性的变体模型 LabelSpreading 训练“可解释性”指标分类器，多次随机测试得到“可解释性”指标的预测平均准确率提升到了 0.74。

为了验证标记传播模型确实利用无标注数据提升了自身的预测准确率，我们选用了有监督学习中的 kNN 分类模型进行对比。我们随机从 100 条已标注数据中抽取了 70 条作为 kNN 模型的训练集，剩下作为测试集，无标注数据对 kNN 模型没有意义，因此不做考虑，其他环境变量控制不变。多次随机测试得到 kNN 模型对“完整性指标”的平均预测准确率在 0.71 左右，低于标签传播模型的 0.82，因此可以认为半监督学习的方案具有更良好的性能。

计算附件四中所有答复信息四个指标加权评价值的均值与方差，并以此作为常模，设置满分为 5 分，使用参照常模计分法生成每个答复样本的标准分。

五、总结与展望

为了满足人民群众日益增长的网上问政的需求，文本基于自然语言处理技术进行了论述与实践，设计了一个面向未来的政务信息处理解决方案，让留言信息直达决策者，也让决策者更直观高效地看到热点问题。在群众留言分类问题中，本文设计的 kNN 分类器方案取得了较好的精确性和稳定性；在热点问题挖掘中，本文构建的聚类与热度评价体系在准确高效的基础上，还具有优良的可解释性；在答复意见评价问题中，本文给出了一种低成本的、可操作性强的解决方案。总体上看，本文专注于传统机器学习方法，在前人研究的成果上，寻找适用于新时代政务信息处理的解决方案。

自然语言处理领域的研究成果浩如烟海，一定还有更多优秀的方案我们无法涉及，希望我们的工作能够成为自然语言处理这顶桂冠上的微小点缀。

参考文献

[1]. 翟慧慧.网络走进百姓政治生活 87.9%的网民关注网络监督[J/OL].http://media.people.com.cn/GB/8766731.html,2009-02-09.

[2]. fxsjy. "Jieba" Chinese text segmentation: built to be the best Python Chinese word segmentation module.[EB/OL].https://github.com/fxsjy/jieba,2020-02-15.

[3]. E, Campus, Mislove, Alan, Marcon, Massimiliano. Measurement and Analysis of Online Social Networks[J]. 2007.

[4]. D Zhang, H Xu, Z Su, Y Xu. Chinese comments sentiment classification based on word2vec and SVMperf[J].Expert Systems with Applications, 2015.

[5]. RaRe-Technologies.gensim – Topic Modelling in Python [EB/OL].https://github.com/RaRe-Technologies/gensim,2020-05-07.

[6]. Dudani S A . The Distance-Weighted k-Nearest-Neighbor Rule[J]. Systems, Man and Cybernetics, IEEE Transactions on, 1976, SMC-6(4).

[7]. Jasper Snoek, Hugo Larochelle, Ryan P. Adams.Practical Bayesian Optimization of Machine Learning Algorithms[J].Advances in Neural Information Processing Systems 25 (NIPS 2012),2012.

[8]. Wong J A H A . Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1):100-108.

[9]. Comaniciu D , Member, IEEE, et al. Mean Shift: A Robust Approach Toward Feature Space Analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(5):603-619.

[10]. 郭金玉, 张忠彬, 孙庆云. 层次分析法的研究与应用[J]. 中国安全科学学报, 2008, 018(005):148-153.

[11]. Blei D M , Ng A Y , Jordan M I , et al. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[12]. Cordasco G , Gargano L . Label propagation algorithm: a semi-synchronous approach[J]. International Journal of Social Network Mining, 2012.

附录

附件一 项目源代码（包含模型文件等过程数据）

附件二 项目代码说明文档（代码实现流程及输入输出说明）

附件三 论文正文（Word 版）

附件四 热点问题表

附件五 热点问题留言明细表