“智慧政务”民意留言综合评价方案

摘 要

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，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。

基于组委会给出的收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，利用自然语言处理和文本挖掘的方法，我们对“群众留言分类”、“热点问题挖掘”以及“答复意见评价”等三个问题分别进行了分析和解决。

针对问题一，我们提取了附件 2 中的“留言主题”，“一级标签”，“留言详情”，提取了排名数量最多的七个一级标签并可视化降序排序，随后进行数据分层，然后对留言主题和留言详情进行分词和去停用词处理，文本向量化，最后选取了支持向量机和随机森林分类模型，并通过 k 折交叉验证与网格搜索技术找出最优分类模型及最优参数，藉此得出了关于留言内容的一级标签分类模型。

针对问题二，在对附件 3 数据进行去重、异常值处理和分词、去停用词等预处理操作后，利用 LDA 主题模型，挖掘出群众留言的 200 个话题，挑选包含群众留言数量最多的 10 个话题，在话题内进行 K-Means 聚类，同时，考虑群众留言的点赞数和反对数，将每个话题下所包含的群众留言数量、点赞数和反对数进行加权汇总，得出群众留言所关心问题的热度评价指数，并筛选出排名前五的热门话题，从而实现热点问题挖掘。

针对问题三，提出了现在政府治理存下的问题和回复意见质量的不足，针对这些问题，首先提出评价指标的原则，分别是相关性原则，完整性原则和可解释性原则，并针对这些原则作出详细的说明和解释。最后提出对留言答复意见质量评价指标方案，并构造信息制造系统模型来衡量答复意见质量的高低，从而帮助政府等管理职能机构提高回复的质量，更好的解决民众所反映的问题。

关键词：智慧政务；word2vec；LDA 主题模型；支持向量机；随机森林

Abstract

With the development of the times, the Internet and mobile Internet technology are more and more developed, and people's lives have changed dramatically. Among them, the former government's understanding of public opinion has changed from email, petition and other ways to wechat, mailbox, telephone and other forms of reflection through the network. At the same time, the number of various public opinion texts and information is growing, which is a huge reading volume and workload compared with the previous manual reading and reply one by one. Therefore, with the development of artificial intelligence, big data, deep learning and other technologies In the past, the manual processing method will be divided and processed by computer, which greatly improves the work efficiency, and at the same time, it is more targeted for dealing with all kinds of public opinion problems. It is of great help for the government to go deep into the grass-roots level through the Internet, truly understand the social situation and public opinion, and formulate corresponding measures for the benefit of the people. In this regard, we first solve the first and second questions in the "smart government" solution, and then through the message details in Annex 2, Annex 3 and Annex 4, and then combined with the reply opinions in Annex 4, we give three different evaluation schemes.

To solve the problem 1, we extract the "message subject", "first level tag", "message details" in Annex 2, extract the seven first level tags with the largest number of rankings and visually rank them in descending order, then carry out data stratification, then carry out word segmentation and de stop word processing for the message subject and message details, quantify the text, and finally use the model support vector machine, random Sen Lin, grid search, compare different models to improve accuracy and optimize parameters, get the first level classification label model about message content.

To solve the second problem, the text is transformed into a word bag model, then the TF-IDF model is used to quantify the text, then K-means clustering is used to measure the similarity relationship between data, then the top five hot topics are selected according to the similarity index, and finally the heat evaluation index is defined according to the likes and anti logarithm of the message, so as to get the evaluation results.

In view of the third problem, this paper puts forward the existing problems in the current government governance and the shortcomings of the quality of the response opinions. In view of these problems, first of all, it puts forward the principles of the evaluation indicators, which are the principle of relevance, the principle of integrity and the principle of interpretability, and makes a detailed explanation and explanation for these principles. In the end, the paper puts forward the evaluation index scheme of the quality of reply comments, and constructs the information manufacturing system model to measure the quality of reply comments, so as to help the political and other management functions improve the quality of reply and better solve the problems

reflected by the public.

Keywords: Wisdom government affairs; word2vec; Latent Dirichlet Allocation ； support vector machine; random forest

目录

1.绪论

1.1 背景、目的及意义

在当今互联网时代，数据库、大数据等技术已经广泛应用在政府部门的日常管理系统中。政务系统是政府部门的重要工作平台，一方面向公众开放并展示公共政务信息与资讯，另一方面进行内部人员和机构流程管理，协调政府各机构系统信息共享，提升政府系统的政务效率。智慧政务系统进一步融入了云计算、人工智能等先进技术，使得政府部门对信息的采集更加广泛，处理能力更加强大，因此政府资源整合效得到了极大提升，凸显了政务的智慧决策价值。同时各类社情民意相关的文本数据量不断攀升，依靠人工进行留言的分类和处理面临难题。同时智慧政务系统采集到的民生数据和政府内部数据杂乱无章，从战略规划上来看，智慧政务建设在统一的战略规划上，主要体现在顶层设计上，不但没有 实现全面统一的规划，也没有建立专门的智慧政务管理中心，缺乏前瞻性的发展战略规划，从而造成数据标准不统一和共享性差等问题。若不及时调整会导致各部门之间政务协同困难，无法完成预期的标准。基础建设是政府统一规划制定的核心支撑，因此，需要提高智慧政务处理数据和采集数据的水平。

智慧政务对数据的标准如下；第一，易获取，用户可以无障碍的从政府部门获取数据，第二，组织效率较高，指政务部门为实现智慧政务所采集的数据符合低成本的原则。如果政府在智慧政务方面能够与公众紧密结合，就可以在移动终端上与市民联系， 发布注意事项，提醒市民的出行安全，将损失降到最低。如果因为人工的失误造成一些来自人民的好的建议和留言被忽略，那可能会直接影响政府的决策，不能真正的进入到群众，不能了解群众所反映的内容，严重将不可避免地对国计民生造成严重的危害。综上，基于自然语言处理技术的智慧政务非常重要但也面临技术的难题和挑战，基于自然语言处理技术的智慧政务具有重要性与必要性。

智慧政务除了将政府内部的数据共享给民众，还要将信息提供给政府，为了保证智慧政务顺利实施，提出相应的对策，第一，法律保障，为保障智慧政务的顺利实行，建立法律体系，保护信息安全。第二，绩效考评体系，智慧政务的建立不是一蹴而就的，所以在实践的过程中要不断的调整，以此促进智慧政务达到最理想的状态。

在智慧政务建设的过程中，要尊重公众 的需求，政府要将公众的需求放在第一位，从整体出发，在实践的过程中，仅依靠政府的力量是不够的，政府应该鼓励多方面投资，培养更多的人才，挖掘公众的需求，有效提升智慧政务的服务水平。

智慧政务的建设是一种新型的治理理念，需要民众参与政府的治理，利用开放性和多样化的建设特点推动政务建设发展。研究结果表明，要想 促进智慧政务建设的发展，就需要掌握社会的信息资源，挖掘公众的需求，从根源上落实智慧政务的发展策略，最重要的是要利用好文本挖掘技术，对民众的留言进行充分的处理，将民众的留言信息划分给不同的职能部门，针对性处理帮助政府有效决策，服务人民。

1.2 解决问题分析

问题 1：群众留言问题分类。建立一个文本分类模型，类别变量是一级分类标签，流程是把留言进行分词、去停用词之后，将文本转成向量，然后拆分训练集和测试集，选择不同的分类模型训练即可。常用的文本分类大体上为基于传统机器学习的文本分类模型，基于深度学习的文本分类模型。同时问题一可能存在文本语义带来的词语交叉，多分类问题带来的难度，数据不平衡的影响，长文本无意义表达太多。

问题 2：热点问题挖掘。分为三步：问题识别，如何从众多留言识别出相似的留言；问题归类：把特定地点或人群的数据归并，即把相似的留言归为同一问题；热度评价，制定一个关于答复意见的评价指标，并作出热度评价的定义和计算方法，对指标排名之后得出对应表。可能存在的难点，地点和人群难以识别，相似的计算复杂（特征多，两两之间计算相似计算量大）

问题 3：答复意见的评价。针对答复意见作出评价，如积极的评价，消极的评价等，并制定出衡量该评价的方案，评价的标准等，分析时考虑答复意见的内容是否与问题相关，是否满足某种规范，答复意见中内容的相关解释。

2.数据预处理

数据预处理是指在主要的处理以前对数据进行的一些处理。对于题目中给定的数据集，包括留言主题、留言详情、留言时间和答复意见等。对于这些原始数据，我们要做一些处理，将原始数据转换为预测模型易于处理的数据类型。如 DataFrame 类型。

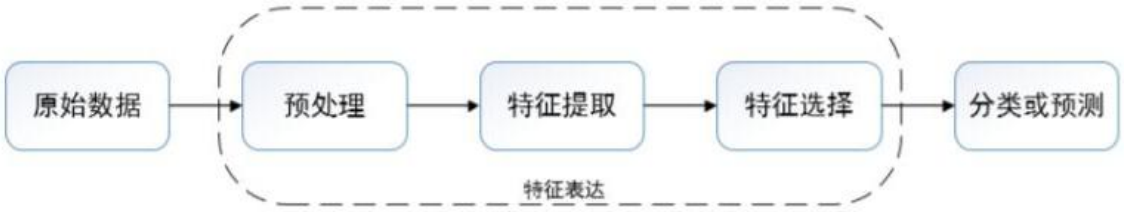

图 2-1 分类预测流程图

2.1 数据转换

我们先提取并查看了'留言主题','一级标签','留言详情'前 5 行的数据，并将它们转换成数据框形式，数据处理的目的是提高数据质量方便后续的数据分析。

2.2 数据清洗

数据清洗是指发现并纠正数据文件中可识别的错误，包括检查数据一致性，处理无效值和缺失值等。我们对所有附件中的文本数据进行统计分析，初步了解数据的特点，其中附件 2 一共有 9210 条数据，附件 3 有 4326 条，附件 4 有 2816 条。首先进行了异常值检测，对留言时间等字段中出现的异常数据进行检测，并确定异常处理的方法；对于存在缺失值的留言用户和留言详情字段进一步诊断缺

失值产生的原因，从而确定缺失值的处理方法。

2.3 文本分词

文本分词是预处理过程中必不可少的一个操作，因为后续的分类操作需要使用文本中的单词来表征文本。文本分词包括两个主要步骤，第一个是词典的构造，第二个是分词算法的操作。本次是使用 python 中的 jieba 库，导入特定的中文词库对留言详情进行分词处理。

2.4 去停用词

去停用词也是预处理过程中不可缺少的一部分，因为并不是文本中每一个单词或字符都能够表征这个文本，比如说“这个”、“的”、“一二三四”、“我 你 他”、“0 1 2 ……9”等等，那么这些词就应当从文本中清除掉。本次问题导入常规停用词库，将附件中出现的词库中的停用词去掉。

2.5 词频统计

我们对附件 2 留言详情中的内容进行词频的统计，并通过数据可视化的方法以词云的方式展示出来。

图 2-2 留言详情词云图

如图 2-2 所示，在对留言详情进行词频统计之后，除去一些人称代词和动词和一些无意义名词意外，人们关注更多的是“学校”，“工作”，“学生”，“政府”一系列的问题，这些也许是政府应该重点关注的目标，也是应该在相应处理部门投入更多的人力物力来解决相应的问题。

2.6 文本向量化

2.6.1 词袋模型

最早的一种比较直观的词向量生成方式称为 One-hot Representation，这种映射方式是通过先将语料库中的所有词汇汇总得到 N 个词汇，并将语料库中的每个文档个体生成一个 N 维的向量，在每个维度就体现了该文档中存在多少个特定词汇。这种方式是一种较为简单的映射方式，其产生的向量表示体现了词频的信息。

2.6.2 TF-IDF 模型

上述方式的模型仅考虑了词频，并且会造成长句子和短句子的向量长度不一致的情况，因此又有一种考 虑了文档词汇中的逆文档频率的映射方式：TF-IDF（term frequency–inverse document frequency）模型，在这 种方式中，首先对词频进行了归一化，即使用词出现的频率而非次数代表词频，表示为公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

另一处改进为统计了每个词的逆文档频率指标，并使用该指标作为词罕见程度的度量值，以更好地 刻画文档的生成向量。逆文档频率的模型如下：

$$\mathrm{idf}_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

这两种模型的共同的缺点在于其二者的向量长度都非常大，对于一个有着 30W 词汇量的语料，每个文档 的映射向量长度将都是 30W，这意味着产出的矩阵非常稀疏，并且在计算时也会非常复杂。

2.6.3 潜语义分析模型

因此，我们引入降维的方式来对高维度的文档向量进行处理，其主要的模型为潜语义分析模型(Latent Semantic Analysis)，这种模型通过数学方法，将文档之间的关系、词之间的关系和文档与词之间的关系都 纳入考虑中(Deerwester,S.,Dumais,S.T.,Furnas,G.W.,Landauer,T.K.,&Harshman,R.(1990).Indexing By Latent Semantic Analysis.Journal of the American Society For Information Science,41,391-407.10)。具体来讲，潜语义分析模型使用了主成分分析的方式来进行降维，即通过抽取向量空间内分布方差最大的若干个正交方向来作为最后的表示方向，并对其余的方向的内容进行丢弃即得到了每个样本的低维表示，该表示是有损的，即丢失了在丢失维度上的分布细节。

2.6.4 Word2Vec 模型

Word2vec 模型是在 Log-Bilinea 及 NNLM 两个模型的基础上由 Tomas 等人开发的工具 Word2vec 可以将词从高维空间分布式映射到低维空间且保留了词向量之间的位置关系，从而解决了 向量稀疏和语义联系两个问题。Word2vec 分为 CBOW(continuous bag-of-words)和 Skip-gram 两种方式。本文主要是基于 Skip-gram 方法进行词向量处理，因为它可以在大型数据集上产生更准确的结果。词向量表示是指在对词语按照其表达的意思 进行分类时，对词语进行向量化处理。在现实中，词语所表达的意思通常是向多个方向发散的，因此需要将词语映射为到多维向量。这样做一方面解决了语意的多方向发散问题，另一方面多维向量能够 使用较小的数字来表征词语。Word2vec 类似于自动 编码器，它在向量中编码每个单词，但是 Word2vec 不会像受限玻尔兹曼机那样通过重构来针对输入的单词进行训练，而是针对与输入语料库中与它们相邻的其他单词进行训练。

3. 数据分析与模型构建

3.1 数据可视化

首先对一级标签数量排名前七的详细留言可视化展示。

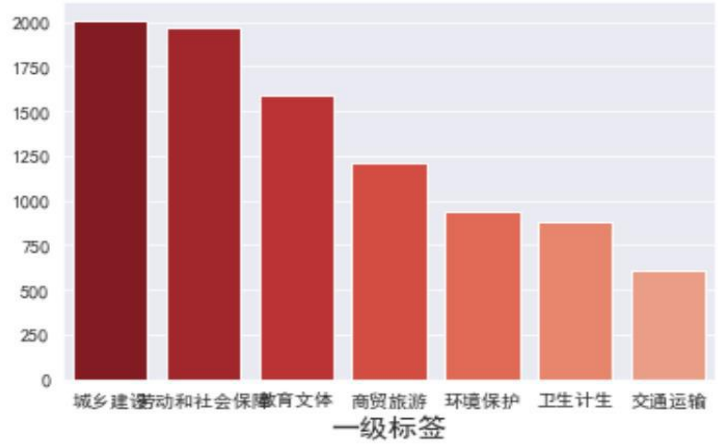

图 3-1 一级标签分布图

如图 3-1 所示，排名前三的一级标签的分类分别是“城乡建设”，“劳动和社会保障”，“教育文体”，其中“城乡建设”和“劳动和社会保障”的数量接近 2000 条，说明人们的留言和所反映的问题大多集中于“城乡建设”和“劳动和社会保障”，关于“城乡建设”，留言内容大致为施工安全隐患，物业问题等，关于“劳动和社会保障”则是劳动保险，工作保障，退休和补助涉及一定金额的问题，政府应该重点关注这些方面的问题，优化和完善智慧政务系统，对这些问题给民众满意答复的同时还要做到工作的公开透明，制定相应的政策来解决该类问题。

将“留言详情”划分为训练集和测试集进行模型训练，然后对利用模型对数据进行训练，同时优化模型，得出结果。

起初我们利用随机森林对数据进行搭建和训练，得出准确率只有 50%，因此模型并不理想，于是采用支持向量机模型，准确率达到 85%，后来继续优化，在支

持向量机的基础上用网格搜索，准确率，召回率，精确率都有所提高。（由于随机森林并不理想，以下便不介绍该模型）

3.2 支持向量机（SVM）

SVM 是基于统计学的一种监督式学习方法，普遍应用于数据分类和回归分析。SVM 用于回归分析的基本思路为：对于 n 个输入变量和 m 组数据的训练样本，即 T={(x11, y1),…(xij, yj),…(xmm, ym)}。设支持向量回归的超平面的拟合函数为 y=b+WTX，W 为权重系数向量，b 为偏置量。利用拟合函数所得的预测值和实际值之间有一定的差值，若差值大于ε，则对损失函数有贡献，若小于ε，则无贡献。损失函数为：

$$L=\sum_{j=1}^{m}\left[\max\left(0,\left|y_j-\bar{y}_j\right|\right)-\varepsilon\right]^2 \tag{3}$$

求解回归方程的参数，即求解损失函数最小值时的参数，此处引入拉格朗日函数，求得支持向量机回归函数为：

$$f(x)=\sum_{j=1}^{m}(a_j-a_j^*)K(x_j,x)+b \tag{4}$$

3.3 支持向量机（SVM）参数寻优

在使用 SVM 算法建模时，存在惩罚参数 C、核函数参数 g 等可调参数会对建模结果产生较大影响。其中惩罚参数 C 影响模型的拟合程度，而核函数参数 g 影响支持向量的个数。确定最佳的 C、g 参数在 SVM 算法建模中显得尤为重要，本文通过 交叉验证法与网格搜索法对 C、g 参数进行寻优，以实现对 SVM 的优化。N 折交叉验证法的基本原理：轮流 N 次将数据 集划分为大小一致的 N 部分，用其中的 N-1 部分作为训练集，剩余的 1 部分作为验证集，N 次验证结果的精度的平均值作为对建模精度的估计值。 网格搜索法优化 SVM 模型参数的基本思路：(1) 利用网格搜索法找出用于建模的所有可调参数并 进行参数组合；(2)依次对所有参数组合进行支持向量机建模；(3)以 N 折交叉验证法下的建模精度为判断依据得出最佳模型和可调参数。

3.4 网格搜索优化 PCA-SVM 算法流程

整个流程分为 4 个部分(1)数据预处理：对原始数据集进行缺失值及异常值处理，分词和去停用词处理；(2)文本向量化：利用 TF-IDF 模型对文档处理；(3)构建模型：将得到新变量的数据按照 4:1 的比例构建训练数据和测试数据，利用训练数据训练出支持向量机模型；(4)优化模型：利用 10 折交叉验证和网格搜索法得到最优的惩罚参数 C、核函数参数 g 的取值，同时得到最优的支持向量机模型。最后计算出准确率为：88.28%，加权分值为：88.13%。

```
1 print_evaluation_scores(y_test,prediction_3)
```

```
accuracy:  0.8827616152844117
f1_score_macro:  0.8708807963015641
f1_score_micro:  0.8827616152844117
f1_score_weighted:  0.8812541692670898
```

图 3-2 F1-score 评价结果图

4. 热点问题分析

在对附件 3 数据进行去重、异常值处理和分词、去停用词等预处理操作后，首先利用 LDA 主题模型，挖掘出群众留言中蕴涵的 200 个 LDA 话题，挑选包含群众留言数量最多的 10 个话题，在话题内进行 K-Means 聚类，同时，考虑群众留言的点赞数和反对数，将每个聚类类别下所包含的群众留言数量、点赞数和反对数进行加权汇总，得出群众留言所关心问题的热度评价指数，并筛选出排名前五的热门话题，从而实现热点问题挖掘。

4.1 LDA 主题分析模型

LDA 由 Blei D M、Ng A Y 和 Jordan 于 2003 年 提出。它是一种文档主题生成模型，本质上是一个三层贝叶斯概率模型。LDA 模型认为一篇文档包含若干主题，而每个主题又有若干词语体现，因此 生成一篇文档的过程也就包含两个阶段：由文档到 主题，再由主题到词语。生成一篇文档之前，应该先确定这篇文档包含 哪些主题，这些主题的分布是什么，先假设在文档中包含 K 个主题，这些主题在文档中的分布为 Om。主题分布符合公式(5)所示的 Dirichlet 分布。

$$f(x_1,x_2,\cdots,x_k;x\propto_1\propto_2\cdots\propto_n)=\frac{1}{B(\propto)}\prod_{i=1}^{k}x_i^{a^i-1} \tag{5}$$

其中参数 B(∝)如公式(6)所示。

$$B(\propto)=\frac{\prod_{i=1}^{k}\Gamma(a^i)}{\Gamma(\sum_{i=1}^{k}a^i)}\sum x_i=1 \tag{6}$$

其中的Γ是 Gamma 函数，满足Γ(x)=(x-1)!，在 Dirichlet 分布中，向量∝是一个 K 维向量，与文档中包含的主题个数 K 相同。 在主题的多项式分布θ_m。中取样生成文档 m 第 n 个词语的主题$Z_{m,n}$。扣多项式分布是 Dirichlet 的共轭分布。重复上述的步骤，从 Dirichlet 分布β中取样生成主题$Z_{m,n}$对应的词语分布$\varphi_{Zm,n}$，然后在根据$\varphi_{Zm,n}$。采样得到最终的词语$W_{m,n}$。

在文档主题模型生成过程中，两个 Dirichlet 分 布分别有一个超参数α和

β，参数α控制了文档中包含主题的多少，α越高，文档包含的主题更多。β表示一个主题中词语的多少，β越大表示主题中包含的词语越多。 LDA 模型生成文档的过程可以用图 4-4 表示。

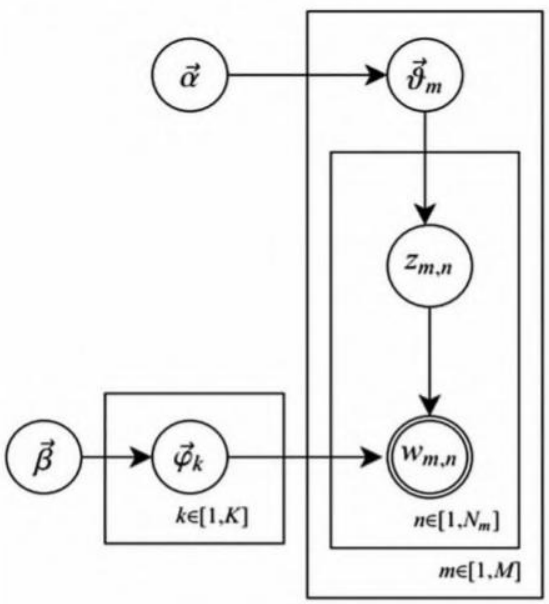

图 4-4 文档生成过程图

基于 LDA 模型求解每一篇文档的主题分布和 每一个主题中词的分布可以使用两种方法，第一种是基于 Gibbs 采样算法求解，第二种是基于变分推断 EM 算法求解。

利用 LDA 主题分析模型，我们首先挖掘得到附件 3 的群众留言中所蕴涵的 200 个话题，分析代码及主要结论如图 4-5 和图 4-6 所示：

```
n_topics = 200
lda_model = LatentDirichletAllocation(n_components=n_topics, learning_method='online',
                                      random_state=0, verbose=0)
```

```
lda_model.fit(small_document_term_matrix)
```

图 4-5 LDA 主题分析模型代码图

```
n_top_words = 10
tf_feature_names = small_count_vectorizer.get_feature_names()
print_top_words(lda_model,tf_feature_names,n_top_words)
```

```
Topic1: 定损 接待室 来访 嘴路 计算机技术 技术资格 接待日 面对面 面见 停错
Topic2: 过道 一十八日 下雪天 耐久 出血热 今夏 窗等 站稳 过路人 为费志刚
Topic3: 购车者 任顺祥 亲取 诈骗犯 状态栏 甄别 延缴 代表性 m7市 流水帐
Topic4: 便利店 拉帘拉 暗点 光想 拉上 减暗 路万坤图 无照经营 瞎眼 透光
Topic5: 打零工 东来 杨锋玮 二合一 发旺 杨桦 处分权 及下 陈愈及 杨锋
Topic6: 泉星 游戏机 物流园 篮球 平面图 楼前 亮之星 美食街 羽毛球 米业
Topic7: a3区 拆迁 西湖 大学 财富 科技 金色 视频 广电 红线
Topic8: 铂金 拜谢 洗砂 填充 车友们 冒领 b9zm 矿井 冒名 检点
Topic9: 折磨 谷林语 问话 矿泉水 催办 南院 宝怡 周建 加家 证是
Topic10: 轻奢 山才苑 东苑 机都 第六十条 少管闲事 宋姓 救援车辆 狐假虎威 抬着
Topic11: 公司 停水 自来水 解决 供水 用水 村民 水费 水表 水管
Topic12: 冲站 建树 未清 两枚 对栋间 结决 图定 第二份 四米 涂家
Topic13: 校名 水域 及澡 自来 利为民 姜坚 同名 特殊教育 请托 停约
Topic14: 托词 iptv 失信于人 吃干饭 一封 星沙住 机顶盒 电视机 数字电视 传送
Topic15: 省辉 远鑫 玲珑 期为 尚细时 h52 骗标 二起 过审 花果
Topic16: a市 建设 城市 发展 中心 加快 市政府 国家 市委 希望
Topic17: 体检中心 行成 圈里 警备区 给华畋 不累 精选 军队 附图片 退转
Topic18: 路霸 试试看 分行 点半多 强拉强 毕业设计 因受 修车 价格比 首善之区
Topic19: 工龄 石碑 年休假 试点 入编 对省 冷敷 皮肤科 十一五 民诉法
Topic20: 退铺 有钱 格子 贸城 星沙中 铺子 中贸城 退换 一笔 穿城
```

图 4-6　LDA 主题模型分析结果图（前 20 个主题）

```
Topic181: 伪证 多设 君尚嘉筑 事故责任 齐头并进 淌水 使多人 林羽果 雨果 河床
Topic182: c2区 江千 游行 居住权 谭虎德 提车 决战 王昕 法律义务 贯彻落实
Topic183: 自付 拯救 太苦 癌症病人 地狱 钱来 开药 病友 或血滤 偷空
Topic184: 刘楠 雨果 预缴 废料 批后 花木 串供 农作物 江家组 过够
Topic185: 护窗 三考 木板 上月底 不晴 平装 出会 防盗窗 推拉 监督机构
Topic186: 姜坚群 特校 井喷 开口子 要待 严重错误 寥寥无几 户有 零时 酸甜苦辣
Topic187: 管辖权 向左转 恶果 电信局 错乱 因一 书证 控辩 景雅苑 青线
Topic188: 湖路 数年 格林 破旧 七八 隔墙 五六 恒丰 乘以 凯通
Topic189: 临建 首层 恒鑫澜 整条 群情激愤 车价 解析 令人心寒 宁花星 礼堂
Topic190: 改造 提质 水电 升级 煤气 香江 棚户 购水 68 停用
Topic191: 车辆 a市 公交车 路口 出行 建议 马路 车道 停车 交通
Topic192: 洞天福地 耿波彦 行舟 之战 李斐 风云际会 周来 渡江 欧阳询 江而治
Topic193: 法院 执行局 三星公司 牛角 找过 被执行人 小微 打过 12345 砍头
Topic194: a1区 大厦 公开 信息 罚款 成本 贵局 依法 申请 自然资源
Topic195: 王震 伴江 吴划 遥想 风云际会 率军 无名之辈 关云长 余秋雨 无法估量
Topic196: 深夜 开设 ktv 扰民 噪音 隔音 场所 外来 每天晚上 丝毫
Topic197: 苏晖 审判员 利好 挑动 杨三人 分案 直销 中众 发改 142
Topic198: 驾驶员 出租车 营运 强硬 敢怒不敢言 自封 新车 出租汽车 难进 悄无声息
Topic199: 筹委会 比选 东部 谷山站 灵堂 阴性 监狱 四块 招商网 铁道
Topic200: 各镇 盗用 欣创 电业局 沸沸扬扬 傲慢 轰响 街道 冰柜 天顶
```

图 4-7 LDA 主题模型分析结果图（后 20 个主题）

经过汇总，整理出 10 个包含群众留言数量最多的话题，结果如表 4-1 所示：

表 4-1 包含留言数最多的前 10 个 LDA 主题表

序号	主题编号	主题词	包含留言数量
1	Topic107	公园 周边 医疗 搭乱建 一圈 周边环境 三道 配套 破损 体育馆	1740
2	Topic62	业主 开发商 户型 小区 房产证 办理 房屋 出售 baidu https	1263
3	Topic33	奥林 招标人 向冬云 投标法 截止期 时禁 错时 分禁 汉联润屋	320
4	Topic162	广花 姜楠 荒谬 僵尸 姜平建 137 特教 洪灾 一百元 车停	298
5	Topic190	改造 提质 水电 升级 煤气 香江 棚户 购水 68 停用	246
6	Topic44	岳鞍阁 退伍军人 牧工商 罗明 李玉 从宁华路 运渣车 玫瑰中遭 享誉	97
7	Topic15	省辉 远鑫 玲珑 期为 尚细时 h52 骗标 二起 过审 花果	78
8	Topic53	四方 a4 区 景城苑 回执 窗帘 七队 国有资产 国资 有色 甩锅	33
9	Topic70	省望 王洋 佳县 继任 沈以 刘龙 iv 睁眼瞎 三宝 还灵	25
10	Topic10	轻奢 山才苑 东苑 机都 第六十条 少管闲事 宋姓 救援车辆 狐假虎威 抬着	23

可以发现，前 5 个话题包含的群众留言数量较多，其反应的问题有一定的共性，但又有细微的差别，前 5 个话题的群众留言情况如图 4-8 至图 4-9 所示：

留言编	留言用	留言主题	留言时间	留言详情	反对	点赞	LDA主题类
188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	是表面进行清扫。没有解决根本问题	0	1	107
188059	A00028571	区中海国际社区三期与四期中间空地夜间施工噪	2019/11/22 16:54:42	打电话给投诉业主，态度强硬恶劣充	0	0	107
188249	A00084085	麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点	2019/9/17 4:25:00	不行，周边邻居也是苦不堪言，请相	0	0	107
188260	A00053484	青青家园小区乐果果零食炒货公共通道摆放空调	2019/5/31 17:06:13	乐果果零食炒货公共通道摆放空调/冰	0	0	107
188399	A00097934	A市利保查号公馆项目夜间噪声扰民	2019/7/3 6:23:25	30日凌晨2点还在施工中，且噪声极大	0	0	107
188414	A00096844	A4区北辰小区非法住改商问题何时能解决？	2019/8/1 7:20:31	多数温和处理、不做太多干预；近两年	0	0	107
188475	A00055810	A6区乾源国际广场停车场违章乱建现象严重	2019/12/3 15:04:58	商铺买了车位然后把车位砌围墙再把里	0	0	107
188535	A00061775	A7县时代星城4幢有非法经营的家庭旅馆	2019/6/13 15:28:44	暗藏着20几间旅馆房间，进出人员混杂	0	0	107
188546	A0006817	A2区佳兆业水新都小区垃圾无人处理	2019/1/23 13:09:19	存在一年有余，周边小区业主反映多次	0	0	107
188774	A00048792	政府东门至万芙路段经常有改装车飙车，真的很	2019/6/18 23:03:31	多次拨打A2区交警大队的电话（0000-	0	0	107
188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019/11/19 18:07:54	小区旁50米处建搅拌站，运渣车吵得人	0	1	107
188876	A00013435	咨询A7县榔梨龙华安置区外围马路修复问题	2019/2/26 11:29:49	一到下雨天裤子鞋子上都是泥巴 出太	0	0	107
188930	A00035285	A市C5市中路798号恩皇建材店私自占用消防通道	2019/4/9 19:14:41	建材店私自占用消防通道，影响居民安	0	0	107
188941	A00018142	中建嘉和城存在严重设计问题	2019/1/7 10:16:45	如此设计，如何保障老人和孩子出行	0	1	107
189029	A00080952	A4区楚江北路能走大货车吗	2019/4/3 12:23:04	的居民生活，特别是夜间货车更加密集	0	0	107
189176	A00093880	区威嘉湖西路车辆违停及占道经营乱象为何久治无	2019/7/24 11:04:12	肆意任性的穿梭于马路上。附近商贩、	0	0	107
189245	A00042948	市北辰三角洲奥城E4区1栋4003房被改建成群租房	2019/3/19 11:22:30	全隐患等，且分别出租给不同人员，还	0	0	107
189345	A00077163	东风街道蚌塘社区自来水管网改造设计存在严重	2019/7/31 6:52:29	次找自来水公司，才得以修复。这件事	0	0	107
189381	A000109815	万科魅力之城商铺无排烟管道，小区内到处油烟	2019/12/4 16:25:06	物业置若罔闻，几年了都没见处理.	0	0	107
189635	A00029819	A1区桐阴里小区一直夜间施工	2019/7/15 21:27:02	期以来一直夜间施工，对周边的住户	0	0	107
189663	A00083693	三期复式楼34层高楼用1.2的栏杆作为生命的屏障	2019/5/11 19:03:58	但是开发商的某些工作人员回答，如	0	0	107
189856	A00073717	）	2019/7/3 11:53:35	”这个太阳一等就不知道是多久，最	0	1	107
189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	百米的地方建搅拌站。可想而知，一个	0	0	107
190033	A000106961	塘泉里社区力都大厦旁新规划的菜市场什么时候	2019/11/3 20:19:18	一直没有听见启用，泉里社区楼盘多	0	5	107
190077	A000112913	A3区欣胜园小区急需水改电改和车位改造	2019/3/14 15:27:28	车需求。而小区一些树下未硬化的地面	0	0	107
190090	A000108593	举报A1区华海3c朝阳菜市场附近夜宵摊门店口摆满	2019/3/26 11:00:05	来碗碰碗的刺耳声，人群的吵闹声。	0	0	107
190108	A909240	丽发新城小区旁边建搅拌站	2019-12-21 15:11:29	尘严重影响几千名学生的健康，很多	0	1	107
190170	A00089880	A9市金刚镇挖埋污水管道，导致我家住房成危房	2019/12/7 16:42:05	鉴定机构鉴定为危房。检测结果分析：	0	0	107

图 4-8 话题 107 所包含的群众留言详情

留言编	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	LDA主题类别
188007	A00074795	询A6区道路命名规划初步成果公示和城乡门牌问	2019/2/14 20:00:00	的门牌号10年都未曾更换过，什么时候	0	1	62
188073	A909164	麓泉社区单方面改变麓谷明珠小区6栋架空层使用	2019/3/11 11:40:42	更没有任何政府调规、改建的流程文件	0	0	62
188074	A909092	A2区富绿新村房产的性质是什么？	2019/1/31 20:17:32	定合同转让给业主了，然而因为湖开的	0	0	62
188416	A00029753	请给K3县乡村医生发卫生室执业许可证	2019/06/20 20:38:47	个新村的是证件下来啊。有些老村医反	0	0	62
188455	A00035902	咨询异地办理出国签证的问题	2019/5/16 15:20:43	需要回到原来的户籍所在地办理呢？还	0	0	62
188467	A00050188	投诉A市温斯顿英语培训学校拖延退费！	2019/3/28 19:57:19	次去学校都是推辞的态度！去了3次每	0	1	62
188592	A00039456	A市长房云时代小区三期后面要建垃圾站？	2019/6/18 10:38:44	宅都只有10多米，垃圾站朝向小区，请	0	0	62
188691	A00036601	举报A市仕弘教育等培训机构涉嫌欺诈	2019/7/11 15:21:58	面试培训费近7000元，并信誓旦旦能确	0	1	62
188780	A00094754	请依法解决A7县黄花镇梁坪村黄泥岭山地建房问题	2019/10/20 14:59:13	林权办的工作人员拒不接收材料，无	0	0	62
188829	A00026141	A2区先锋派出所办个签证，拒收现金	2019/2/15 13:17:06	由于没带银行卡，我说用人民币支付，	0	2	62
189066	A00073590	A3区积溪路颐美幼儿园门口路段收费停车合法吗	2019/10/22 8:06:58	要收费。根据“法无授权不可为”原则	0	0	62
189113	A000112050	融圣国际长城物业动用水枪、灭火器在小区内斗	2019/7/1 17:29:02	小区曝光。近2年内，业委会多次找街	0	1	62
189180	A000106515	A市人才购房补贴申请是否与单位注册地有关？	2019/6/18 9:51:36	现在长工作，单位出具的在职证明也写	0	3	62
189247	A000107866	建议在“我的A市”app中尽快接入法律服务的意见	2019/12/21 9:45:46	证有三个律师解答（全国其它平台一个	0	0	62
189278	A00048293	咨询A市对残疾人有什么扶持政策？	2019/7/29 9:24:16	么扶持证策，我现在房子都是住的父母	0	0	62
189587	A00018292	区复兴-艾家冲Ⅰ、Ⅱ线500kV线路杆迁工程临近居	2019/1/13 18:44:38	用基本农田，质疑其未通过相关用地	0	0	62
189733	A0009754	A市限卖房产政策一刀切	2019/10/10 17:09:47	套现，导致我无法启动我的创业计划。	1	0	62
189866	A000109850	A市美联嘉园物业私下变卖产权不明的地下车库	2019/10/22 14:40:59	1日晚有不明人士统一着装阻挡小区2个	0	0	62
189992	A00086181	A市南郡明珠业主8年了都没有房产证	2019/7/26 11:16:26	就打发了，我们老百姓辛辛苦苦攒的钱	0	0	62
190019	A000104285	A7县星沙恒基凯旋门房屋漏水问题一直没解决	2019/6/4 18:08:56	现漏水现象，墙皮发霉生虫，联系物	0	0	62
190021	A00098570	A市民年龄与身份证不符，可以办理退休吗？	2019/4/16 7:53:13	有69年的也有72年的，如果按69年的，	0	0	62
190087	A00052076	A市新奥燃气服务态度差	2019/7/22 0:01:47	要挟，蛮不讲理地必须先缴费，再能买	0	0	62
190156	A00040576	A7县星沙中贸城欺诈业主、拖欠业主资金不退还	2019/3/25 11:53:00	019年3月30之前退换全部款项、到现在	0	0	62
190171	A00021818	A8县科目三补考费要600一次，合理吗？	2019/11/20 19:58:15	一个人是少，A8县每天几百的补考群众	0	2	62
190261	A0007341	A市楚江技工工贸学校乱收补考费	2019/10/14 16:38:38	私下决定的，在期中考试对不合格学生	0	0	62
190266	A00094438	A市高铁南站出租车用假钱	2019/2/26 17:04:35	的问题。同住一个酒店的旅客也都反应	0	2	62
190301	A00031131	A市蓝的出租车营运到期，迟迟不发新车	2019/6/12 15:50:29	，毫无音讯。其他公司车辆交车，基本	1	1	62
190355	A00034724	区白箬铺镇光明村田坪组顺祥养老院涉嫌非法集	2019/8/15 18:32:14	下来回奔波讨要养老费用，恐发生人	0	0	62

图 4-9 话题 62 所包含的群众留言详情

留言编	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	LDA主题类别
188119	A00035029	对A市地铁违规用工问题的质疑	2019/5/27 16:04:44	加班，不加班还扣钱，扣身份证，一天	0	0	33
188665	A000106234	A市松雅湖东方航标2栋2楼有传销窝点	2019/3/26 0:18:26	寒嘘问暖一下后　就开始讲起来他们的	0	1	33
188887	A00085665	于A7县恒基凯旋门万婴格林幼儿园办普惠园的咨	2019/7/17 10:19:08	文件要求，我县全面开展城镇小区配套	1	4	33
190192	A00094992	试管做出基因缺陷女婴，不幸家庭雪上加霜	2019/9/1 22:30:53	于这种特殊家族背景情况，在划生二胎	0	0	33
190522	A00056153	A5区洞井镇目前仅有一所民办的同升湖高中学校	2019/10/9 16:27:50	一所搞质量的全日制公办高级中学，缓	0	6	33
191394	A000106333	A市中南林业科技大学违反研究生培养计划	2019/4/25 11:38:43	奔波，又要兼顾学业。学校为了自己方	0	0	33
191572	A000106425	科技园园区振兴工业实体经济扶植奖励未按标准	2019/5/22 9:18:07	奖励并已发放，在申请扶植奖励的时	0	0	33
191792	A00096212	A市A5区稻田中学午餐伙食太差劲	2019/12/3 19:09:59	做好点，做得色香味俱全点，做卫生	0	0	33
192029	A00024840	县星沙水痘疫苗及腮腺炎疫苗是入学必须补种的	2019/8/18 18:24:25	家孩子补种后，医院开具的证明上，	0	0	33
192500	A00058135	A市盈泉国际美容养生会所欺骗消费者	2019/5/13 16:10:06	而且收据还写着，特价不予退款）。但	0	0	33
192774	A00039164	A3区第十幼儿园数百名学龄前儿童即将无学可上	2019/6/13 12:20:43	招收大班和中班生的古塘幼儿园、以	0	2	33
193385	A909144	请A市坚决取缔校园贷！	2019/10/30 14:07:24	。要坚决取缔。信任卡发放也乱七八	0	6	33
194573	A00076207	区教师招聘报名只给报考考生一次修改资料的机	2019/6/18 12:16:01	有一次，A3区这样做，完全没有具备一	0	0	33
195005	A00041623	A6区二中强制学生暑假补课且收费	2019/7/12 17:02:03	假，西地省A市A6区，A6区二中学校违	0	0	33
195013	A00078096	寻找A7县退伍人员的下落	2019/3/4 20:42:07	在54师医院工作。现在战友们都在找	0	0	33
195124	A00091862	A8县喻家坳中心幼儿园突然砍掉学生上课时间	2019/9/18 10:17:32	候再来这么一出，不解决甚至推脱，这	0	0	33
195453	A000109415	A2区天悦幼儿园的社团收费太贵了	2019/9/13 9:41:52	用也应该看情况适当的收取我们家长还	0	1	33
195845	A00095816	A2区天悦幼儿园的老师为什么全部换了	2019/9/8 21:27:54	让原来的老师全部离开了天悦幼儿园？	0	0	33
195917	A909119	A市涉外经济学院组织学生外出打工合理吗？	2019/11/05 10:31:38	作十个小时以上，（晚班时间是20：3	0	1	33
197099	A00031618	请A市提高民办幼儿园的师资待遇水平	2019/2/22 16:02:54	在超负荷工作的前提下，工资待遇始	0	1	33
197185	A00041597	西地省邮电职业学院校园网掉线率高、网速慢	2019/4/2 20:36:21	了，像在寝室连着WIFI，出门打个水就	0	0	33
197463	A00059726	A3区山景区的一些路标标错了吧	2019/12/26 16:39:45	亭，看见路标往台阶上爬。这种害人	0	0	33
197504	A00078335	A市晚报普通老百姓想看，却订不起	2019/2/13 13:39:16	什么涨价了这么多，　普通老百姓想看	0	11	33
198132	A00063808	区中海国际社区幼儿园无法满足实龄幼儿就读需	2019/8/12 15:32:27	我家二胎宝宝目前3岁半，正是就读年	0	0	33
199581	A00031618	长郡月亮岛二小招生公告要求新生考试分班，合理	2019/8/13 16:02:21	新生考试分班，严重违背国家要求的	0	0	33
200858	A00062596	请问强制封闭煤球厂是国家政策吗	2019/12/6 0:16:05	合理吗？冬天来了，老百姓靠煤球取	0	0	33
201059	A00060044	廉洁过教师节，亟待教育主管部门的重视	2019/9/10 9:59:42	工礼物、或者多少金额以下的，如果超	0	0	33
201727	A00026322	咨询A4区万国三期幼儿园的学位问题	2019/5/22 15:37:25	远大于现有学位数量。为什么幼儿园不	0	0	33

图 4-10 话题 33 所包含的群众留言详情

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	LDA主题类别
188409	A0003274	市地铁3号线星沙大道站地铁出入口设置极不合理	2019/6/19 10:14:39	区、星沙三区、星沙四区、开源鑫阁、	0	4	162
190978	A00025654	A7县东六路附近的空地做什么用途？	2019/4/8 9:24:39	以西,凉塘路以南，北斗路以南中间这	0	1	162
191993	A00083527	希望A7县城区增加智能红绿灯和过街天桥	2019/5/7 9:57:30	少车等人的时间，增加路口车流通过率	0	1	162
192976	A00049731	星沙公交X2路车换成中型车体客车是不合理的	2019/9/6 10:20:31	**********路本就是客流大	0	0	162
193056	A00032672	对A7县“雅韵星沙周末有戏”演出活动的补充建议	2019/12/9 21:33:26	电子化，既可以节省纸张， 也可以让市	0	3	162
193286	A000103197	主地铁3号线A7县松雅西地省站西北方向10万民众自	2019/4/17 11:13:12	号：开元路北侧穿越东四路和东四路西	0	32	162
193337	A00080343	请问A7县泉星公园何时开工，工期多长？	2019/10/16 13:03:13	了5年之久， 人民群众非常期待，请问	0	12	162
193678	A000104070	反映A市城市交通存在的诸多问题	2019/3/24 20:40:57	也不积极作为。为何轨道基团不能和公交	0	8	162
193893	A000103884	我觉得A市地铁四号线六沟垅站出口设计不合理	2019/1/27 13:45:39	垅站最合理，但是六沟垅站设计出口全	0	0	162
194022	A00042107	坚决反对在A7县诸亚山林小区门口设置医院	2019/7/8 10:39:54	份，危害我们的安宁居住环境和可能出	0	1	162
194260	A00077323	A市地铁五一广场站一二号线换乘的建议	2019/1/18 10:25:27	下到地下三层来乘坐二号线。原本地下	0	2	162
194554	A00075486	7县三一大道下匝道左拐进锦绣路绿灯通行时间太	2019/11/4 21:07:28	下时间。还有经过改造后左拐车辆人为	0	0	162
195198	A00026503	对远期A市地铁的建议	2019/3/22 18:03:51	和欣盛路交汇处）-曾家冲（麓延路和	2	0	162
195915	A00018309	建议A市规划局东延地铁7号至泉塘片区，便民出行	2019/2/26 8:22:42	建有众多楼盘，企业，十几万人口，出	1	21	162
196092	A00049639	县黄兴镇主干道黄江大道路边围档严重影响居民出	2019/9/27 8:25:47	余，路边都挡了围档，严重影响居民出	0	0	162
196144	A0005821	A市北二环与A4区大道交界附近有三处超高减速带	2019/7/9 10:29:17	安全驾驶没问题，但问题在于，（该减	0	0	162
196282	A00039390	反映A市地铁3号线松雅西地省站地下通道建设问题	2019/9/6 13:48:12	侧穿越东四路和东四路西侧穿越开元路	0	16	162
196878	A00090150	A7县插盘问题何时解决？	2019/8/29 9:33:10	星园，楚天世纪城西苑，等楼盘何时	0	2	162
197479	A00063376	县高铁站和火车站的道路金水西路西延什么时候	2019/3/6 17:05:22	建路金水西路西延什么时候动工？常益	0	0	162
197662	A00085667	询A7县东四路远大路到人民路南延拆迁进度等问	2019/1/22 21:08:46	市地铁站也正在施工,后续是否能争取增	0	18	162
197665	A00072517	县星沙开元路与东四路西南角闲置地块有什么规划	2019/5/23 17:23:27	东小区处是一个大水坑，靠家和院附近	0	1	162
198012	A0009769	A市地铁4号线月亮岛地铁口出行交通不便	2019/7/9 10:26:38	来还有共享单车，现在共享单车都没有	0	0	162
198247	A00060228	A市公交公司的清扫令人寒心	2019/3/26 10:47:51	雨天都如此。有两个问题想问问：第	0	0	162
198331	A000112983	A7县星沙东六路下穿长浏高速工程进展如何？	2019/5/7 20:40:15	路的修建，已启动东向西分流，但受国	0	4	162
198370	A00074236	A市田函剧院前地下工程是什么？	2019/1/12 18:14:50	工程进度相当慢，而且是在地铁1号线	0	0	162
198975	A00051906	咨询A7县道路规划的问题	2019/8/1 17:13:39	是怎么规划的，会从东七路一直往东修	9	4	162
199202	A00077844	希望A7县x117公交末班车时间能延长	2019/8/22 18:33:29	南站的**********公交车确7点	0	1	162
199345	A00075471	希望L827路停一下A7县机场大道人民东路站牌	2019/11/26 14:44:18	场，距离不远，却没有直到的公交，不	0	1	162

图 4-11 话题 162 所包含的群众留言详情

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	LDA主题类别
188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	通行时，该司机并未按地面车辆指示导	0	0	190
188251	A00013092	特立路与东四路口晚高峰太堵，建议调整信号灯	2019/10/19 11:02:40	一般情况下至少两到三个信号灯才能通	0	0	190
188922	A00039738	A市A1区纬二路上有两种处罚标准？	2019/6/20 13:13:17	王堆路为界线，路东边马路两边停满了	0	0	190
189055	A00081390	A2区猴子石大桥下公交车经常乱停乱放	2019/7/31 14:49:58	员正常安全通行。本来道路只有2个车	0	1	190
189294	A00083527	A市南站里面的12306客服虚设	2020/1/1 18:00:16	字来的醒目。2、西出入口客服中心规	0	0	190
189456	A00028347	A3区云栖路云栖谷小区辅道好多乱停车的	2019/11/1 17:45:07	常大，违停路段导致严重堵车和安全隐	0	0	190
189864	A000100372	县路口镇黄兴大道与路青线交叉路口的红绿灯错	2019/8/1 12:45:07	行同时亮，存在交通隐患，且该路口	0	0	190
190538	A00051614	建议在A3区旺龙路增加人行斑马线	2019/1/30 11:30:34	原本就在对面，却因为增加护栏而绕很	0	0	190
190754	A00089983	A7县万家丽北路楚龙街道长期拥堵问题求解决	2019/3/13 18:03:53	熙实验中学，人流量非常大。特别早晚	0	0	190
191951	A00041448	A4区绿地海外滩小区距渝长厦高铁太近了	2019/8/23 14:21:38	是轻轨也不是火车，这是高铁，每天	0	1	190
192440	A00063205	A市地铁7号线何时能开工建设？	2019/1/26 22:36:53	府的规划，我们了解到7号线正好经过	0	3	190
192521	A00083527	希望A市城区增加智能红绿灯和过街天桥	2019/5/7 9:52:44	时长，让每个方向都能够一路绿灯开下	0	0	190
192633	A00018151	市高新区谷苑路的渣土车为何如此疯狂飞驰？？？	2019/4/28 15:48:35	无视行人，无视小车，无视红灯。是	0	0	190
193723	A00080323	7县809路方略潇邦站无站牌，公交站牌更新不及时	2019/6/20 23:44:24	这边的公交站亭却没有704路的站牌信	0	1	190
193788	A00053352	县开元东路城东与睿城小区之间能否设置人行天桥	2019/9/9 11:16:56	区之间没有设置人行天桥，而过往车	0	1	190
193835	A00031421	期盼A市星沙15路车在上班高峰期能准时发车	2019/5/14 11:08:53	短短几分钟，但是影响很多人上班上	0	1	190
194025	A000102813	捞刀河镇星沙联络线罗涵庄段辅道口什么时候修	2019/3/31 17:41:19	接下桥就逆行掉头，交通事故频繁发生	0	8	190
195609	A00048849	王家湾西二环辅路与含浦路交叉口红绿灯设置不	2019/4/17 8:13:42	拥堵。之前左转直行一起放行，时间大	0	0	190
196434	A00025475	山南路上左家陇西公交站旁的红绿灯损坏，盼及时	2019/6/15 10:10:15	灯损坏，一侧的通过按钮损坏，无法	0	0	190
196503	A0008801	A7县毛塘铺到楚龙的公交车何时可以开通啊？	2019/2/19 21:56:57	大道---万家丽北路到楚龙区域的，目前	0	4	190
197195	A00035697	希望A市社区巴士5号线能增加车次	2019/12/19 10:27:12	写明的是高峰期10分钟一趟，低峰时段	0	0	190
197437	A00013625	县万家丽路星沙机电学院段有没有规划过街天桥	2019/4/16 10:12:51	交通安全隐患。修地道是不可能的了，	0	0	190
197964	A00052075	请解决A7县海伦春天、当代广场居民出行难的问题	2019/2/20 20:54:05	星大道隔档，设人行道，增加红绿灯	0	3	190
198225	A00056682	请增加A市276路运行线路或增加发车频率	2019/3/29 10:39:26	门上车，挤得车门也关不上，上车后	0	5	190
198393	A00046173	A市X115等待时间太久了	2019/2/7 11:00:42	人是赶不上的，去高铁站的线路肯定	0	0	190
198680	A00050238	A6区汽车站公交站站牌需要合理整治规划	2019/7/4 14:25:59	又等，一样的，没办法只能叫个摩的	0	0	190
199271	A00062677	A市中海国际居民出行不方便	2019/10/3 22:33:35	附二，附三的公交车，希望有关部门能	0	0	190
199330	A00092277	A7县海伦春天、当代广场居民出行困难	2019/2/15 11:08:34	站点路段，打开黄星大道隔档，设人行	0	0	190

图 4-12 话题 190 所包含的群众留言详情

由于在每一个 LDA 话题中又包含多个具体的问题，因此，接下来我们在前 5 个话题内部对群众留言进行 K-Means 聚类，以更准确地挖掘出群众关心的热点问题。

4.2 K-Means 聚类

4.2.1 算法原理

K-Means 算法是一种无监督学习，同时也是基于划分的聚类算法，一般用欧式距离作为衡量数据对象间相似度的指标，相似度与数据对象间的距离成反比，相似度越大，距离越小。算法需要预先指定初始聚类数以及 k 个初始聚类中心，根据数据对象与聚类中心之间的相似度，不断更新聚类中心的位置，不断降低类簇 的误差平方和（Sum of Squared Error，SSE）当 SSE 不再变化或目标函数收

敛时，聚类结束，得到最终结果。

K-Means 算法的核心思想是：首先从数据集中随机选取 k 个初始聚类中心 Ci($1 \le i \le k$)，计算其余数据对象 与聚类中心 Ci 的欧氏距离，找出离目标数据对象最近的聚类中心 Ci，并将数据对象分配到聚类中心 Ci 所对应的簇中。然后计算每个簇中数据对象的平均值作为 新的聚类中心，进行下一次迭代，直到聚类中心不再变化或达到最大的迭代次数停止。 空间中数据对象与聚类中心间的欧式距离计算公式为：

$$d(x,C_i)=\sqrt{\sum_{j=1}^{m}(xj-C_{ij})^2} \tag{7}$$

其中，x 为数据对象，Ci 为第 i 个聚类中心，m 为数据对象的维度，xj ，Cij 为 x 和 Ci 的第 j 个属性值。整个数据集的误差平方和 SSE 计算公式为：

$$SSE=\sum_{i=1}^{k}\sum_{x\in C_i}\left|d(x,C_i)\right|^2 \tag{8}$$

其中，SSE 的大小表示聚类结果的好坏，k 为簇的个数。

4.2.2 K-Means 算法流程

K-Mean 聚类算法是一个不断迭代的过程，如图 4-13 所示，原始数据集有 4 个簇，图中 x 和 y 分别代表数据点的横纵坐标值，使用 K-Means 算法对数据集进行聚类，在对数据集经过两次迭代后得到最终的聚类结果。

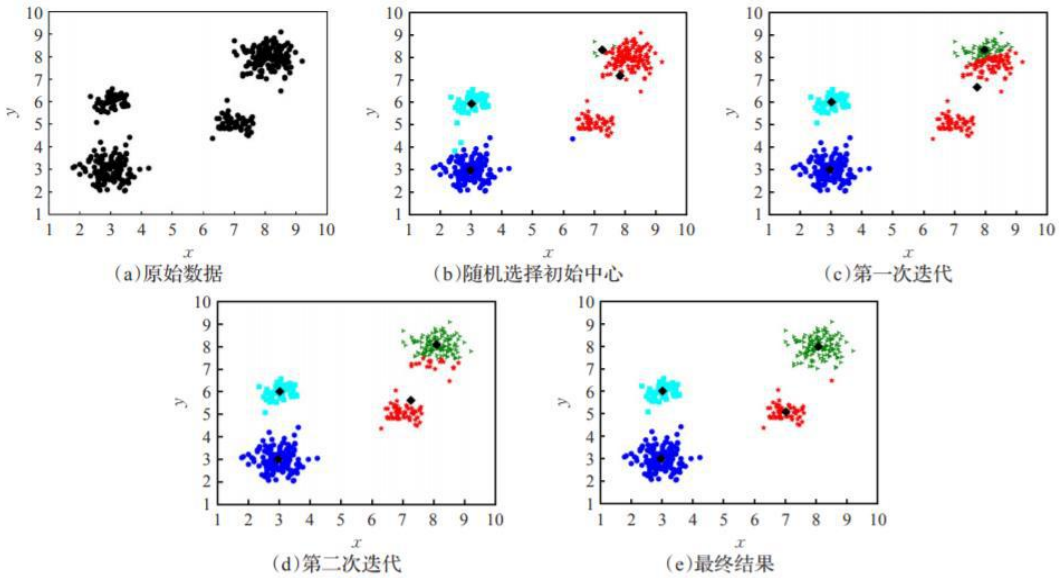

图 4-13 K-Means 算法迭代过程

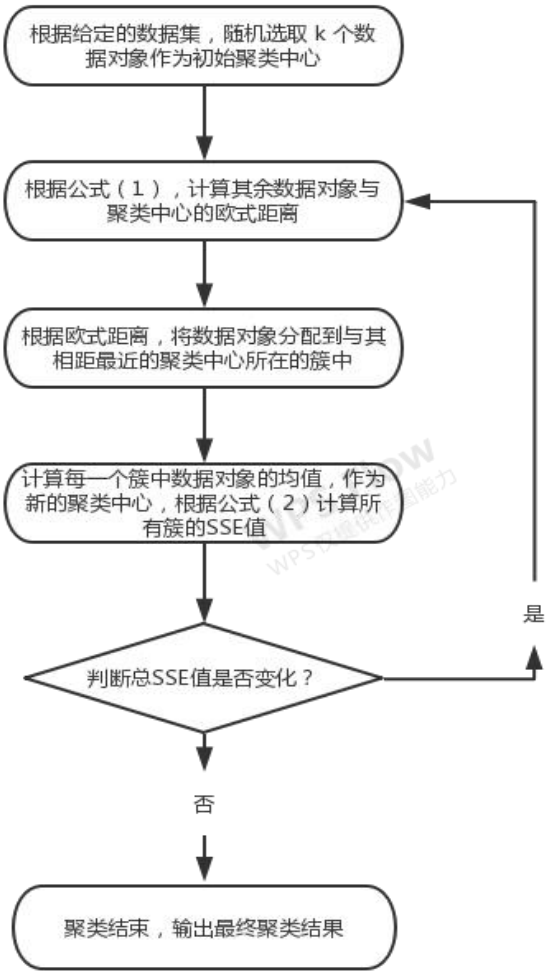

图 4-14 K-Means 算法流程图

K-Means 聚类算法对于大数据集有高效的聚类效果，其算法复杂度为 O(nmkT)。其中，n 为数据集大小，m 为数据对象特征维数，k 为指定的簇的数目，T 为总的迭代次数。

4.2.3 Top5 话题聚类

接下来我们对 LDA 主题模型中挖掘出的包含留言数量最多的 5 个话题进行 K-Means 聚类，其中聚类的数量通过计算相应类别数的轮廓系数得到。留言文本的向量化采用 word2vec 模型。

Topic107 的留言文本转向量化后的结果如图 4-15 所示。

```
1  word2vec_topic107
```

```
array([[ 0.62382908,  0.04324519,  0.37953743, ...,  0.13201197,
         0.00664324,  0.02717257],
       [ 0.61023589,  0.02499241,  0.36512788, ...,  0.12611472,
        -0.00777654, -0.00793427],
       [ 0.70867805,  0.06111381,  0.39197442, ...,  0.09468329,
        -0.00941421,  0.08909263],
       ...,
       [ 0.54360093,  0.03248362,  0.36548937, ...,  0.17098385,
         0.0216388 , -0.01253172],
       [ 0.61868521,  0.0439753 ,  0.38208625, ...,  0.11600109,
         0.01389987,  0.04287588],
       [ 0.53902518,  0.02703422,  0.35242887, ...,  0.14885499,
         0.03256417,  0.01402078]])
```

图 4-15　Topic107 的留言文本 Word2Vec 向量化结果

对 Topic107 进行 K-Means 聚类，聚类主要过程及轮廓系数图如图 4-16 和 4-17 所示。

```
1  from sklearn.cluster import KMeans
2  from sklearn import metrics
3  silhouette = []
4  for k in range(3,30):
5      kmeans = KMeans(n_clusters=k)
6      kmeans.fit(word2vec_topic107)
7      silhouette.append(metrics.silhouette_score(word2vec_topic107,kmeans.labels_))
8  print("轮廓系数最大的聚类数为：",silhouette.index(max(silhouette))+3)
9  print(silhouette)
10 import matplotlib.pyplot as plt
11 %matplotlib inline
12 plt.plot(silhouette,'r*-')
```

```
轮廓系数最大的聚类数为： 3
[0.40034563068861895, 0.3369528451454747, 0.27998696512631194, 0.25737542864684254, 0.2612000213495766, 0.22117297504125089, 0.2135271
642579227, 0.19659665769801848, 0.1820753267191418, 0.1776598320937542 2, 0.18264182384522135, 0.17698322829129562, 0.1729345806143568
8, 0.1750303241638214, 0.17814957598259423, 0.16328911623848383, 0.16656862319737248, 0.16709228793220646, 0.16962963152209934, 0.1618
1179409275298, 0.15731301139269674, 0.15106407232602903, 0.1577343866658308, 0.1457257190474701, 0.14660044400049294, 0.14868362304693
53, 0.15751132767426515]
```

图 4-16　Topic107 的 K-Means 聚类过程及结果

由图 4-16 可以的结果得出轮廓系数最大的聚类为 3，且聚类后结果如图 4-16 效果也比较明显，由此我们可以进行下一步分析。

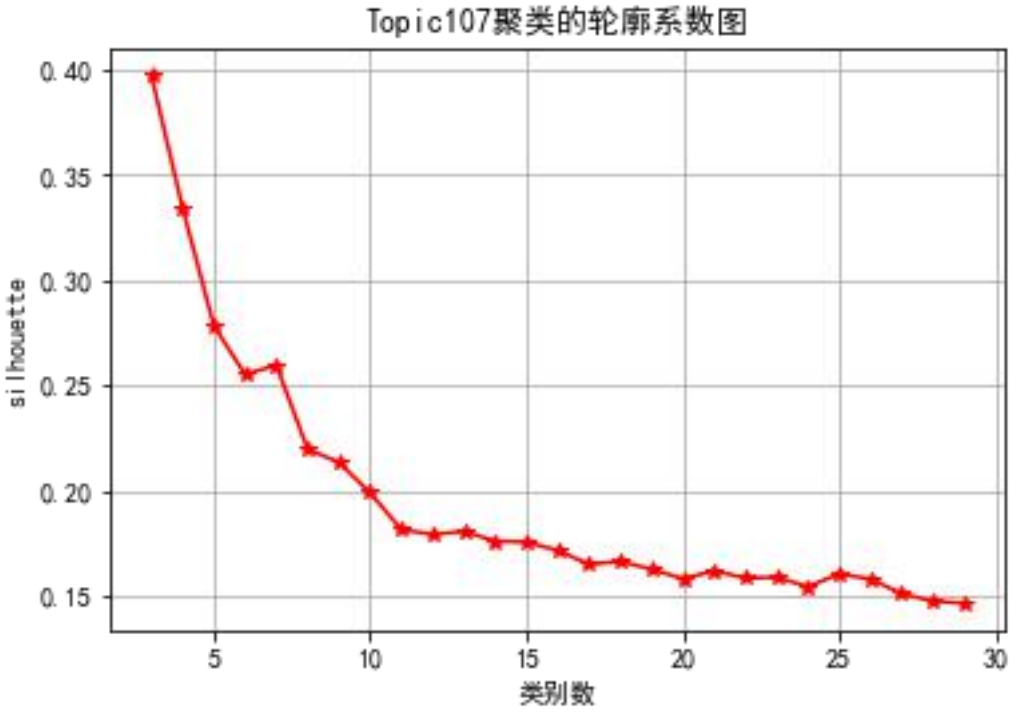

图 4-17　Topic107 聚类的轮廓系数图

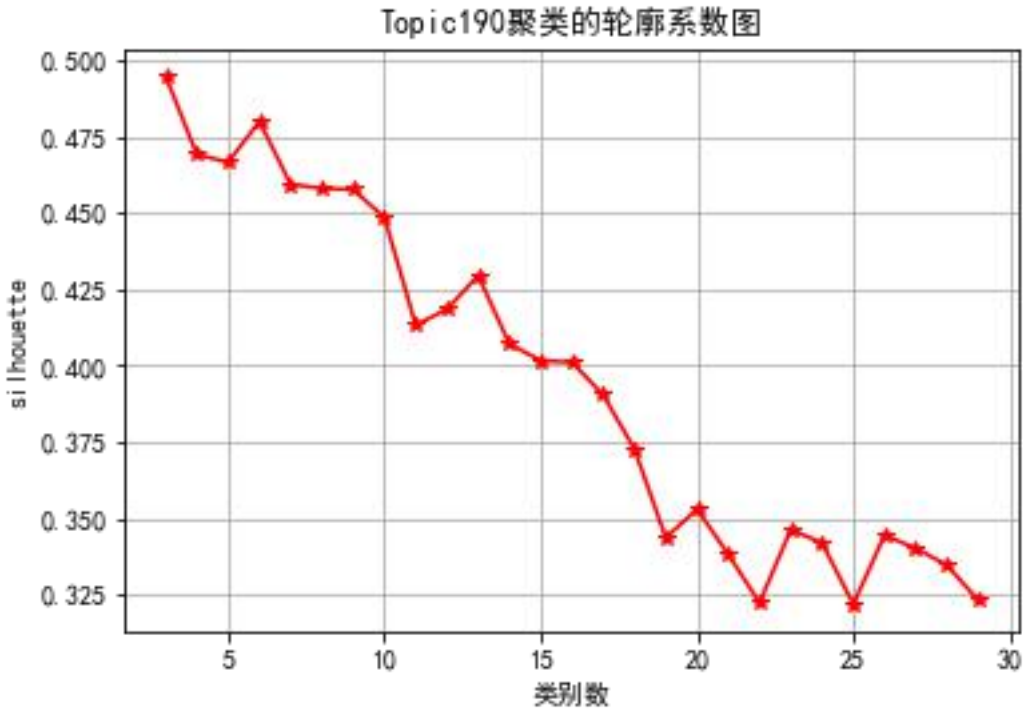

图 4-18　Topic190 聚类的轮廓系数图

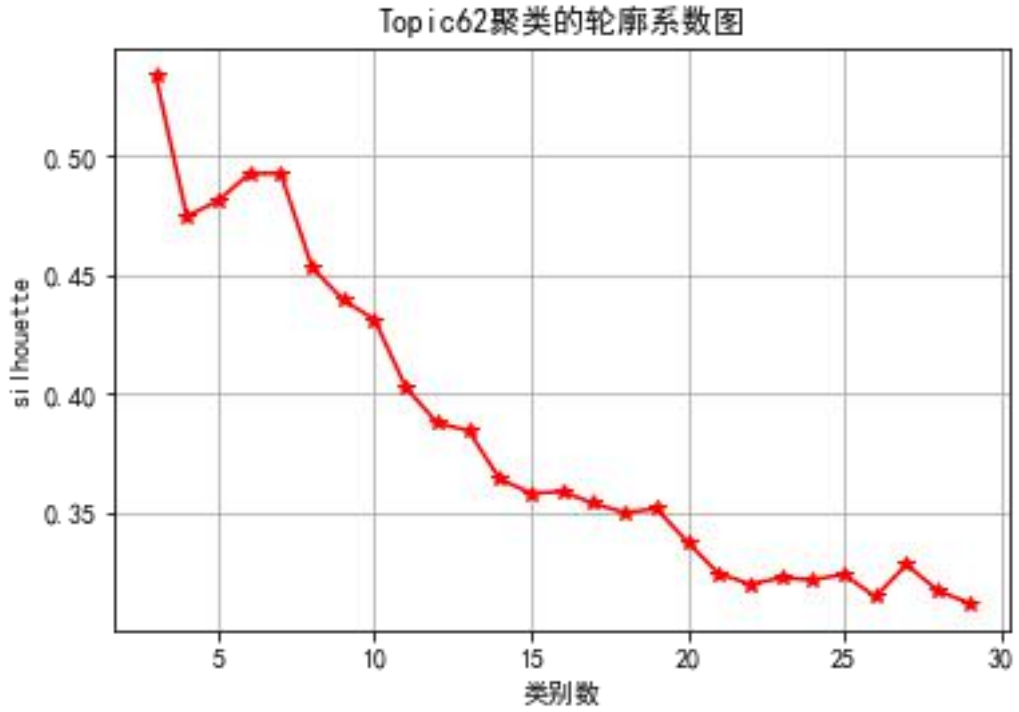

图 4-19　Topic62 聚类的轮廓系数图

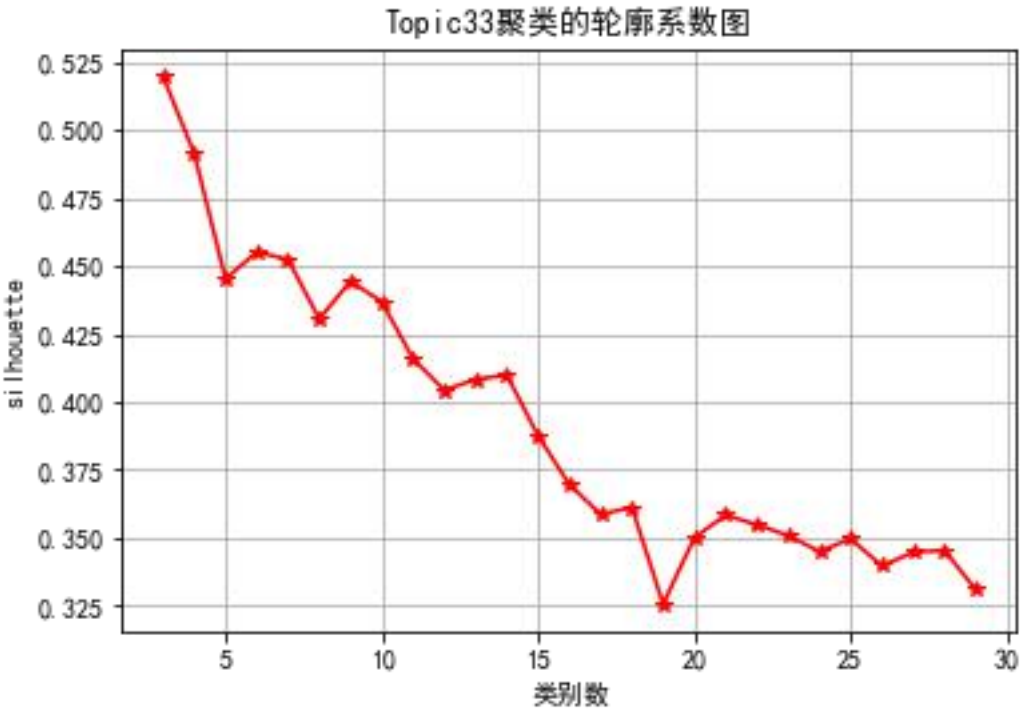

图 4-20　Topic33 聚类的轮廓系数图

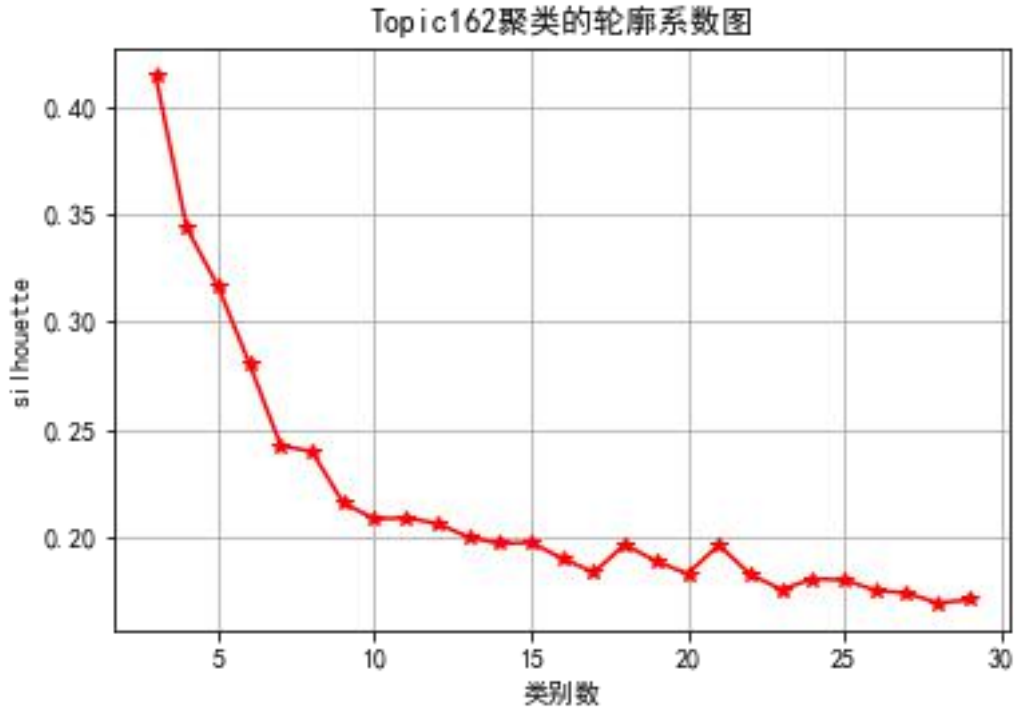

图 4-21　Topic162 聚类的轮廓系数图

4.3 热度评价指标

根据已有留言内容，结合留言主题，留言时间，留言详情，点赞数和反对数特点，设计了热点问题热度评价指标体系。

为便于描述各评价因素的权重计算过程,在表 4-2 中,我们将针对留言详情热度评价指标体系中的评价结果、评价指标和度量方案分别 用变量 R、Cj(1≤j≤7)和 Mk(1≤k≤7) 表示。 最终评价结果 R 的大小由评价指标 C1~C7 的度量值加权计算得到。在本文中,权重包括:组合权重(combination weight,记 cw),即 C1~C7 相对于 R 的权重;单一权重(signle weight,记为 sw),即 C1~C7 相对于 R、M1～M7 相对于 C1~C7 的权重。

表 4-2 针对留言详情热度评价体系及度量方案

热点问题热度评价指标	评价指标	度量方案
针对留言热度评价结果 R	点赞数量 C1	点赞数量是否大于 200 M1
	反对数量 C2	反对数量是否大于 9 M2

	留言时间范围 C3	该月留言数占该月天数的比值　M3
	受众地域分布 C4	留言所覆盖地理区域数量 M4
	留言正负面 C5	留言是否属于负面反映 M5
	同话题留言数 C6	同一反映的留言数量 M6
	留言用户反映数量 C7	某一用户发留言的数量 M7

4.4 热点问题排名

问题ID	热度指数	时间范围	地点/人群	问题描述
1	155	2019/1/7 14:27:13-2020/1/2 10:41:53	A7县海伦春天、当代广场居民	不便，车辆调头的问题，建议A7县交警严管海伦春天与当代广场黄星大道沿线
2	109	2019/1/6 14:29:42-2020/1/7 21:31:53	A市南站乘客	指引前往服务中心；西出入口客服中心规模很大，没人在值班；东出入口客服
3	109	2019/1/4 22:58:13-2019/12/17 17:24:29	中铁十五局一公司A9市蒙华铁路民工	中铁十五局一公司A9市蒙华铁路工程拖欠工程款
4	95	2019/1/26 22:36:53-2019/12/23 16:56:09	A7县万家丽北路楚龙街道	A7县万家丽北路楚龙街道长期拥堵问题
5	56	2019/1/16 11:19:35-2019/12/23 8:50:24	A市6路公交车乘客	A市6路公交车随意变道通行

图 4-22 热点问题排名 TOP5

如图 4-22 所示，经过统计分析，得出了热点排名前 5 的结果，排名第一的是位于 A7 县海伦春天、当代广场居民，问题描述是黄星大道出行交通不便，车辆调头的问题，建议 A7 县交警严管海伦春天与当代广场黄星大道沿线辅路的违章停车；排名第二的是 A 市南站乘客，问题描述是候车室没有明显的地标或墙标或者大型指引牌指引前往服务中心;西出入口客服中心规模很大,没人在值班;东出入口客服有两个人在，紧急的车票改签服务无法处理；第三是中铁十五局一公司 A9 市蒙华铁路民工反映中铁十五局一公司 A9 市蒙华铁路工程拖欠工程款；第四是 A7 县万家丽北路楚龙街道居民，反映 A7 县万家丽北路楚龙街道长期拥堵问题；第五是 A 市 6 路公交车乘客，问题描述是 A 市 6 路公交车随意变道通行。政府可以根据热点问题制定相应的政策，投入相对更多的人力和物力来解决这些问题。

5.答复质量评价方案

大数据与人工智能技术的快速发展为政府治理效能的提升提供了新的可能。目前,政府部门在运用大数据提升治理能力方面已经取得了一定进展。但是。 随着越来越多的政府开始运用大数据提升治理效能,一些问题也显现出来。 由于缺乏清晰的评价标准,在政府治理过程中存在着:(1)大数据运用目标不够清晰;(2)策略选择不够得当;(3)对效果的判断缺乏参考依据等问题。为了更好地引导政府运用大数据提升治理效能,本研究拟运用价值焦点思考法(VFT)构建大数据提升政

府答复意见的评价指标体系,为后续实证评价奠定基础。

5.1 评价指标体系构建的原则

5.11 评价指标体系构建的原则

运用自然语言处理技术优化政府答复意见是一个正在深化的过程,构建评价指标体系需要遵循相关性原则、完整性原则、可解释性原则。

（1）相关性原则

这里是指信息集合中信息元素之间的相关性。当信息系统将一个信息集合提供给用户时"其中的信息元素之间应该具有较强的相关性。毫无关联的信息元素所组成的信息集合将使得用户无法使用,还有的信息相关性是指信息与用户的使用目的相关，在我们的指标体系中将其归于信息的有用性。

政府答复意见涉及多个方面,一些方面可以通过自然语言处理的运用加以提升,另一些方面则需要依靠制度的改进与优化。处理后可以保证答复意见的内容是否与留言有关，是否对该留言作出合理的解释，以及能否解决和解决情况给出答复意见。在设计评价指标时,需要聚焦在留言详情热点内容可以提升的政府治理效能的方面,同时要保证答复意见相关性的质量有针对性地设计和甄选评价指标。

（2）完整性原则

作为一个信息集合必然具有一定的结构。这里的完整性就是指信息集合结构的完整。一个具有完整结构的信息集合能够完整地表述一个思想和事实，描述一个事物。比如要提供一份职工简历"那么"其中姓名，性别，出生，民族，现职别，主要工作经历及年份等信息元素必须具备，他们组成了一个完整的信息结构。如果缺少了其中的某一项或几项，则其信息结构不完整，不能够完整地表达一个思想描述一个事物。

评价指标体系的完整性取决于留言的可读性和完整性。 一些指标在理论上有很好的表现力,但是针对一些计算机无法处理的数据的存在较大难度。 在选择具体指标时,需要将留言的完整性性作为一个取舍依据。若留言的内容本身不够完整，计算机无法识别，人工也无法处理，答复意见就应依照完整性原则指出该条留言的不足和漏洞。

（3）可解释性原则

信息是通过信息符号来表达的，信息用户通过信息符号来理解和使用信息，因而，信息符号必须能够理解且易于理解，如果用户看不懂信息符号，那么信息的用途就会丧失。可解释性，首先要看信息本身的表达方法：其次，对于不同的用户群，其解释力和知识前提不同。同样的信息他们的理解程度也是不同的。因而，在对信息进行表达时，对用户群进行分析是非常必要的，在考虑到信息用户的理解力因素之后,可解释性要求表达信息符号的编码格式简写形式等要有明确的解释，以便于用户理解。在必要的时候，还要提供一定的元信息以增强信息符号的可解释性。

该原则意指政府对于留言的回复应当清晰明了，便于大众理解。答复评价的目的是促进管理目标的实现。在事物发展的不同阶段,对于一些民众难以理解的

专业术语和名词,在答复时要尽可能详细解释，做到通俗易懂，让广大老百姓真正了解政府回复的内容和意义，有利于引导政府治理向着预期目标发展。 因此,在指标设置时,需要考虑在当前发展阶段自然语言处理技术提升政府治理效能的侧重点和主要任务。

5.12 评价指标选择的要求

第一，指标需要便于理解、有着较高的明确性，不会让用户产生误解；第二，指标要具有可测性特点，能够被测量，同时测量难度小；第三，指标尽量避免出现包含或者相互冲突现象；第四，指标需要划分重点，明确指标的主体；第五，指标需要具有层次特点；第六，指标的选择需要适应信息共享发展阶段当前的质量水平；第七，指标的选择要能将政府答复意见的质量特征全面的具体的分析出来。

5.2 答复质量评价

在对附件 4 数据进行类似于前述几个问题的预处理基础上，我们首先对附件 4 中的群众留言文本和政府答复文本进行了 word2vec 向量转化，并通过余弦相似度计算出每条群众留言与其回复文本的相似度指数，其中 word2vec 向量取 400 维。具体计算过程及结果如图 5-1 至 5-2 所示。

```
# 训练word2vec模型
from gensim.models.word2vec import Word2Vec
data['cutt'] = data['cut_problem'].apply(jieba.lcut)
x = data["cutt"]
#print(x)
model = Word2Vec(x,size=400, min_count=5)
import numpy as np
word2vec = np.zeros((len(data),400))

#文本转word2vec向量
for i in range(len(data)):
    for j in range(len(data.cutt[i])):
        if data.cutt[i][j] in model.wv.vocab.keys():
            word2vec[i] += model.wv[data.cutt[i][j]]
for i in range(len(data)):
    word2vec[i] /= len(data.cutt[i])
```

图 5-1 word2vec 向量转化图

```
from sklearn.metrics.pairwise import cosine_similarity
cosine = cosine_similarity(word2vec,word2vec_answer)
cosine_sim = pd.DataFrame(cosine)
cosine_sim.to_csv("cosine_sim.csv")
```

图 5-2 余弦相似度指数图

```
print(data['simi'])
```

```
0       0.063778
1       0.065426
2      -0.071477
3       0.161823
4       0.183601
          ...
2811   -0.018770
2812    0.047037
2813    0.112878
2814    0.113224
2815    0.113432
Name: simi, Length: 2816, dtype: float64
```

图 5-3 余弦相似度结果图

留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	余弦相似度	回复字数	相似度指数	及时度	答复质量指数
UU0082390	咨询J9县的油茶种植政策	2019/1/5 15:52:14	机农产品品牌建设的决定和农	为目标，以品牌建设为动力，以“原生态、	2019/1/7 17:00:37	0.069221755	7881	69	2.047488	80.48
UU0082390	咨询J9县的油茶种植政策	2019/1/5 15:56:23	机农产品品牌建设的决定和农	以新型组织为主体，以产业发展为目标，以	2019/1/16 8:38:52	0.056041529	7878	56	10.69617	80.23
UU0082390	农产品品牌建设扶持政策的	2018/12/15 23:28:40	生态有机农产品品牌建设的决	组织为主体，以产业发展为目标，以品牌建设	2019/1/3 10:53:46	-0.018517373	7883	-19	18.47576	79.46
UU0082352	07年无处理结果，盼望领导	2018/5/25 16:56:36	奔跑，想尽快的得到解决，	2007年10月19日，K8县三中建房协调小组	2018/6/4 21:13:23	-0.023136033	5117	-23	10.17832	51.84
UU008720	万明村征收问题回复的质疑	2015/2/7 12:59:42	2 土地问题 给我们	持，我们各相关单位经认真梳理，查阅原始	2015/8/24 15:19:53	0.006188591	3885	6	198.0973	37.93
UU008758	平交易分局局长邓斌文面	2016/9/29 18:39:41	商局的领导是否能够反省自己	言我一语一直哭诉到中午12点半，廖组长看	2017/1/9 9:25:47	0.01500669	3721	15	101.6153	37.34
UU0081257	县龙门镇居民请求医疗费减	2019/1/15 17:10:59	花去医疗费用12万余元，现	救助”的情况说明。 吴根福是一位木工，	2019/2/22 16:26:56	0.010238385	3561	10	37.96941	36.33
UU0081871	镇建安亭村多年来横行鱼	2018/8/8 12:56:50	……因为通过其家族三代用	生汉，又名徐廷汉）于1986年因犯盗窃罪被	2018/10/10 20:16:14	0.01244901	3529	12	63.30514	35.78
UU0081149	对K6县小升初政策的质疑	2018/8/13 11:30:13	对于决策方案有重大分歧，	别组建专职队伍服务方便学生入学相关手续	2018/9/30 15:11:41	0.020315746	3472	20	48.1538	35.44
UU0081457	关闭F5县东山镇关山村采	2018/6/4 8:25:49	头的车每天络绎不绝，道路被	场检查，该公司无工业废水产生，已采取雨	2018/7/2 11:14:54	0.05693594	3094	57	28.11742	32.23
UU0081286	A7县春华镇下一步的发展	2016/3/10 22:14:21	往那个方面发展，发展的力	品通过网络和物流进入市场，激发农村市场	2016/3/23 16:19:02	0.043214882	2735	43	12.75325	28.65
UU0081304	于“三类人员”考试一事回	2018/6/7 10:14:07	加分；（三）由于省市计生	站长毛群刚提拔为冷水井乡副乡长），由县	2018/6/26 12:19:03	0.093782112	2681	94	19.08676	28.56
UU0081304	于“三类人员”考试一事回	2018/6/7 10:27:47	加分；（三）由于省市计生	各乡镇，未竞争上岗的经管员（约30人）进	2018/6/26 11:09:22	0.043309655	2649	43	19.02888	27.73
UU008985	十里洞村民希望喝上健康	2016/11/7 14:11:23	段时间天天放水，搞的我们	值班，污水处理人员是雷喜外。上午10时	2017/1/6 15:56:08	0.022087449	2636	22	60.07274	26.98
UU0081049	6县尿毒症患者大病报销的	2019/5/10 9:50:38	一些老人，不懂互联网，只能	0%的比例报销，起付线为8000元（建档立卡	2019/5/15 15:15:10	-0.009476271	2571	-9	5.22537	26.56
UU0081647	溪市镇移民房屋修建一事回	2017/12/19 12:03:03	交付使用，和坑蒙移民的合法	4月8日，2013年2月22日，K市建设工程质量	2017/12/21 16:52:49	0.049455487	2456	49	2.201227	26.03
UU0081096	投诉K市潇楚源学校乱收费	2019/10/7 17:19:55	量差，问题目不教学生这样的	管理中存在的不足进行了改进，加快了管理	2019/11/18 15:05:27	0.029290867	2315	29	41.90662	24.02
UU008830	处理不当，导致患者错过	2019/7/27 22:40:09	在等待检查结果同时，我去	节脱位。 4、右足第3跖骨中段骨折。入院初	2019/10/21 17:21:57	0.005970503	2315	6	85.77903	23.35
UU0081179	0县金盆圩骆铭孙村强制办	2016/10/13 15:34:44	1800多亩林地办养猪场。已经	项目要求、环境保护作了详细介绍。与会人	2016/11/7 16:26:21	0.04026458	2212	40	25.03584	23.27
UU0081514	银行应取消短信收费，多次	2018/12/22 11:07:16	卡，免收短信费，人社部门	发放全省PSAM卡，实行全省统一初始化；指	2019/1/10 11:05:39	0.059097651	2066	59	18.99888	22.06
UU0081514	市人社部门加强对社保卡的	2018/12/12 16:51:22	收费；因为垄断，就不会让一	发放全省PSAM卡，实行全省统一初始化；指	2018/12/25 9:07:12	0.02340106	2077	25	12.67766	21.90
UU008593	融圣国际小区的几点问题	2017/5/11 21:47:17	注意根本看不见），西大门	定给予优惠。国家鼓励、支持企业事业组织	2017/6/12 16:23:36	0.053751181	1972	54	31.77522	20.94
UU00815	部门整治井湾子山水庭院	2018/7/13 10:58:24	进行改革，长期停水；小区	层和19-21栋负一层，均为人防工程，出售	2018/8/6 14:12:17	-0.043761177	1984	-44	24.13464	20.16
UU0082338	部的装修价格以及[政府发	2018/10/24 10:55:47	160号文件的计算表在直接费	。你反映的该单位工程费用计算表列有1.4	2018/11/6 10:21:48	0.034978709	1877	35	12.9764	19.99
UU008344	询L3县的异地搬迁有关政	2018/11/10 18:38:00	地搬迁政策，仅剩我一户没有	栋**房，镇人民政府按工作流程向县联席办	2018/12/7 11:27:08	0.016794963	1885	17	26.70079	19.75
UU008437	咨询A市户口迁移问题	2017/10/31 10:28:50	方式又有多种不同说法，不知	及其户口迁入的书面意见；6.无处挂靠户口	2017/11/13 14:56:20	0.061061617	1767	61	13.18576	19.15
UU008777	打造文明畅通新A市的十条	2018/6/21 15:23:49	路更窄的香港，A市要拥堵得	工作重心，坚持交通违法整治的强日常、	2018/7/6 14:57:36	0.027733288	1785	28	14.98179	18.98

图 5-4 群众留言与其回复文本的相似度图

到此，根据余弦相似度计算出每条群众留言与其回复文本的相似度指数，根据留言时间和答复时间计算出及时率，再计算出答复意见的字数，加权最后得到答复质量指数，依据答复质量指数的大小来决定答复质量的高低，答复质量指数越大，答复质量越好，反之则越差。从而实现关于答复意见的自动评价方案。

6. 总结与展望

本文主要使用大数据相关的自然语言处理技术，利用 python 工具来解决问题和分析数据，对留言的内容进行处理，转化为计算机能够识别并且分析，接着建立关于留言内容的一级标签分类模型，我们搭建了三种不同的模型才使参数最优，使模型更加理想；随后我们将留言详情向量化，并根据向量化后的结果聚类得出群众集中反映和关注的热点问题。群众主要集中在 A 市，反映最多的问题围绕“住房保障与房地产”，“城市建设和市政管理”等相关内容，我们以点赞数和反对数作为主要热点问题的评价指标，能够充分反映人们所关注的问题，对此类问题作为管理者应当引起重视，为群众作出一个满意的答复的同时也应尽快落实

实处，解决群众面临的问题。最后针对回复意见我们从三个原则进行论述和解释，并阐述了关于留言答复质量评价指标的详细方案。

在接下来的工作中，文本挖掘和自然语言处理技术将帮助社会各类组织机构处理繁杂的文本问题，减轻人工整理负担的同时提高效率和准确率，这将会是未来发展的一个趋势。

致谢：非常感谢泰迪杯官方能够设计文本挖掘相关的赛题，真正解决民生问题，也帮助政府机构更加深入了解到群众，我们都知道，在当今网络飞速发展今天，我国的网民规模超 9 亿，网络已经离不开人们的生活，如何有效处理文本信息是至关重要的课题，希望我们的初步尝试能够让民众所表达的心声得到真正的回复。

7. 参考文献

[1] 王晓东.朝阳市智慧政务现状与对策[J].中共朝阳市委党校，2019, (17):199-200
[2] 吴宏胜.基于可信计算和 UEBA 的智慧政务系统[J].信息网络安全,2020, (1):89-93
[3] 冀宇轩.文本向量化表示方法的总结与分析[J].探索与观察,2018, (22):10-12
[4] 周志华.机器学习[M].北京: 清华大学出版社,2016: 229-229
[5] 王智锐,唐汝宁.基于支持向量机算法的空调负荷预测及实验研究[J].制冷技术,2013,33(4):28-31
[6] XIAO F,FAN C.Data mining in building automation system for improving building operational performance [J].Energy and Buildings,2014, 75(11): 109-118
[7] 刘峥,黄真银,徐成良,陈焕新,李昱瑾.基于网格搜索优化的主成分分析-支持向量机算法的冷水机组能耗预测[J].制冷技术 ,2019, 39(6):15-20
[8] 曹瑞昌,吴建明.信息质量及其评价指标体系[J].情报探索.2019, 39(6):15-20
[9] 王芳,张百慧,杨灵芝,李晓阳,刘汪洋,张建光,赵洪.基于大数据应用的政府治理效能评价指标体系构建研究[J].信息资源管理学报.2020, 10(2):17-28
[10] 赵彦华,王桂珠,杜海燕.农业科技信息共享中信息质量评价指标体系研究[J].农业网络信息.2018, (5):22-24
[11] 苏强,梁冰.信息质量及其评价指标.计算机系统应用.2000, (7):63-65
[12] 杨俊，闯赵超.K-Means 聚类算法研究综述.计算机工程与应用.2019, 55(23):7-14,63
[13] 孙飞显 程世辉 靳晓婷 倪天林.政府负面网络舆情热度定量评价方法--以新浪微博为例[J].舆情研究.2015-09-14:137-141