基于机器学习的智慧政务系统

摘要：本文主要对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进行分析，通过分类、聚类、热度排序、实体识别、文本摘要、文本相关性分析等中文自然语言处理手段得到了用于群众留言分类、问题聚类、热度评价和答复意见评价等有益于智慧政务的模型。

针对问题一“群众留言分类”，我们首先对留言数据进行分析，选择用于分类的文本，经过数据预处理后，生成了计算机可直接处理的词向量。由于深度学习在文本分类领域应用广泛，综合考虑各种深度学习模型的速度和分类效果等因素后，分别利用基于深度学习的 FastText 模型和 TextCNN 模型设计了两套方案，预测留言内容对应的一级标签，然后通过调整相关参数，训练得到两者的最优模型。计算并比较两个模型的评价指标，最终选择 FastText 模型作为留言内容的一级标签分类模型。

针对问题二“热点问题挖掘”，我们根据留言主题文本，进行聚类，对聚类的结果进行完善后，进行实体识别和文本摘要，得到其反应的地点和人群以及一类留言问题的总体描述，再利用留言频率、点赞数、反对数等信息，计算出留言频率和用户活跃度指标，利用基于信息熵条件下的带有权值的 TOPSIS 方法综合评估留言类的热度，最后依据热度对所有留言类进行排序，取热度前五的留言类生成热点问题表以及热点问题明细表。

针对问题三“答复意见评价”，我们根据答复意见的特征与社情留言的特殊性，确定对答复意见的评价从相关性、完整性、及时性与可解释性四个方面来构建答复意见评价体系；并在评价体系中，使用中文分句手段、TextRank 算法与正则表达，分别从四个方面得到对应的特性指标；对四个特性指标进行加权与归一化处理，最终形成完整的答复意见评分。

关键词：中文自然语言处理 FastText TextCNN 文本分类 短文本聚类 热度评价模型 信息熵 TOPSIS 答复意见评价模型

Intelligent government system based on machine learning

Abstract: This article mainly analyzes the records of public questioning messages collected from public sources on the Internet, and the relevant departments' responses to some of the public messages, through classification, clustering, popularity ranking, entity recognition, text summary, text correlation analysis, etc. Chinese natural language processing methods have been used for smart government affairs models such as mass message classification, question clustering, popularity evaluation, and response opinion evaluation.

In response to the question 1, "classification of mass message", we first analyze the message data, select the text for classification, and after data preprocessing, generate a word vector that can be directly processed by the computer. Since deep learning is widely used in the field of text classification, after comprehensively considering the factors such as the speed and classification effects of various deep learning models, two sets of schemes were designed using the FastText model and TextCNN model based on deep learning to predict the level of the message content. Label, and then adjust the relevant parameters, training to get the best model of the two. Calculate and compare the evaluation indexes of the two models, and finally choose the FastText model as the first-level label classification model of the message content.

In response to the question 2, "hotspot mining", we clustered according to the subject text of the message, and after improving the results of the clustering, we performed entity recognition and text summarization to obtain the overall description of the location and the group of people who responded and a type of message problem. , Then use the message frequency, likes, oppositions and other information to calculate the message frequency and user activity index, use the TOPSIS method with weights based on the information entropy to comprehensively evaluate the popularity of the message category, and finally based on the heat All message categories are sorted, and the top five message categories are selected to generate a hotspot question list and a list of hotspot questions.

In response to question 3, "Review Opinion Evaluation", based on the characteristics of the response opinion and the particularity of the social comment, we determined that the evaluation of the response opinion should be based on four aspects: relevance, completeness, timeliness and interpretability. ; And in the evaluation system, using Chinese clause means, TextRank algorithm and regular

expression, get the corresponding characteristic indicators from four aspects; weighting and normalizing the four characteristic indicators, and finally forming a complete response opinion score .

Keywords： Chinese natural language processing FastText TextCNN text classification short text clustering thermal evaluation model information entropy TOPSIS regular expression response evaluation model

目录

一、挖掘目标

（1）通过深度学习的方法提取文本特征，自动学习特征和类别标签之间的关联，解决文本分类问题，最终高效准确地实现“智慧政务”中的群众留言分类功能。

（2）根据留言主题对留言进行归类，定义并生成这类留言的属性信息，再根据热度这一属性进行留言类的排序，准确挖掘群众留言中的热点问题。

（3）确定评价标准，从多个方面对答复意见形成特征指标，建立评价方案，根据多个特征指标得到答复意见评分，实现从多角度客观评价答复意见质量的功能。

二、群众留言分类的建模与解答

2.1 群众留言分类方案

留言分类方案的总体流程图如图 1：

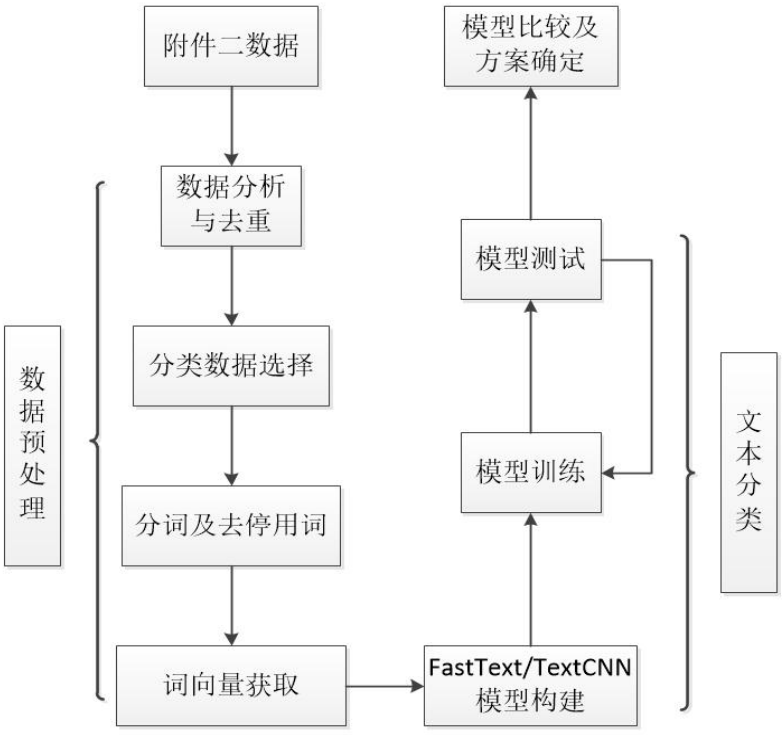

图 1 留言分类流程图

具体步骤如下：

①数据预处理。通过对数据的分析与处理，实现从中文文本到直接用于分类

算法的词向量的转化。

②文本分类。通过利用 FastText 模型和 TextCNN 模型实现两套设计方案，进行基于深度学习的群众留言分类模型的构建。分别通过优化参数，得到两套方案下的最优模型。

③通过比较两套方案的评价指标，确定最终用于群众留言的一级标签分类模型。

2.2 数据预处理

2.2.1 数据分析与清洗

1. 数据分析

附件 2 给出了留言分类的数据，考虑到类别分布的不均衡会对深度学习模型的效果造成较大影响，因此我们对数据进行了分析，留言内容总计 7 个类别，共 9210 条。具体分布如图 2：

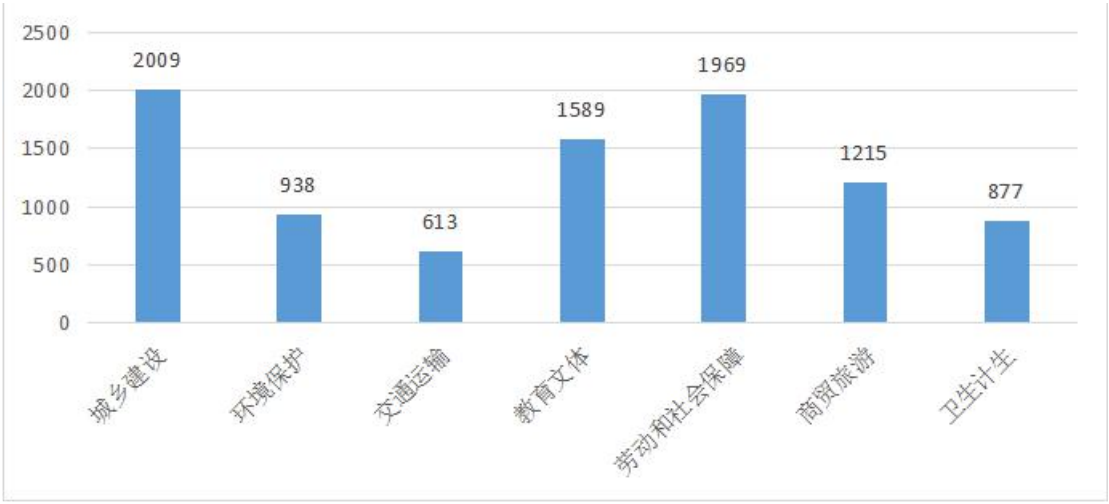

图 2 留言类别分布图

通过上图我们可以看到，类别分布没有太大差异，比较均衡，因此不需要对数据进行采样处理。

2. 数据清洗

观察到留言详情中有多余空字符，我们首先对数据进行了清洗，消除冗余字符及字符串的干扰，以避免对模型的性能造成影响。

此外，数据中还存在重复的数据，为防止训练集与测试集中混入相同数据，避免正确率虚高的情况，我们对数据进行去重处理。去重过程中，发现文件中数据还有异常重复，即同一内容被冠上不同标签和分类标签明显错误的情况，如图

3：

1	商贸旅游	目前B市发K市物流由三家物流单位控制分别是鸿泰，平安，鹏飞现已联合经营，价格是原来3_5倍的涨价。原发货到K市10
2	交通运输	目前B市发K市物流由三家物流单位控制分别是鸿泰，平安，鹏飞现已联合经营，价格是原来3_5倍的涨价。原发货到K市10
3	劳动和社会保障	尊敬的书记：你好!我在工作好几年，随着年龄的增长，渴望有自己住房，有自己的家，在亲戚朋友的东拼西借的情况下，
4	城乡建设	尊敬的书记：你好!我在工作好几年，随着年龄的增长，渴望有自己住房，有自己的家，在亲戚朋友的东拼西借的情况下，
5	劳动和社会保障	我本人的工作地在A市,国家规定的各项社保和公积金都是缴纳在A市，　缴纳公积金已经三年,今年想通过省直公积金中心在
6	城乡建设	我本人的工作地在A市,国家规定的各项社保和公积金都是缴纳在A市，　缴纳公积金已经三年,今年想通过省直公积金中心在
7	商贸旅游	M5市涟水名城55栋电梯经常出问题，多次发生突然下坠，颤抖的现象，一到下雨天，电梯井内就有漏水声，井底积水严重，
8	商贸旅游	M5市涟水名城55栋电梯经常出问题，多次发生突然下坠，颤抖的现象，一到下雨天，电梯井内就有漏水声，井底积水严重，

图 3 重复数据示例图

考虑到一条留言只能被分派至一个职能部门处理，对于内容重复，标签不同但没有明显错误的情况，如图 3 第 1、2 行数据，我们根据人工判断，取其中一个标签，对于标签明显错误的情况，如图 3 第 7、8 行数据，此内容理应属于城乡建设类，但标签却是商贸旅游，我们对其标签进行了修正。

最后，由于附件 2 是按标签顺序排列的，不利于训练集和测试集的划分。因此我们对数据的顺序进行了打乱。经过上述操作，得到 0.xlsx（详见 0.xlsx）。

2.2.2 分类数据选择

针对本题数据，考虑到数据中包含留言主题列，留言详情列，我们很容易想到用①留言主题、②留言详情、③主题+详情（后面称为重组详情）作为分类算法的对象。重组详情由拼接留言详情列与留言主题列内容而得到，为避免与原始数据混淆，重组详情与对应的一级标签单独放在 1.xlsx（详见 1.xlsx）中。

我们使用单层全连接神经网络（详见单全连接层.ipynb）简单测试了一下效果，数据的各项指标如下表 2 所示：

表 1 三种分类文本的指标详情

指标详情			
	①留言主题	②留言详情	③主题+详情
precision	0.85	0.86	0.9
recall	0.84	0.86	0.9
f1-score	0.84	0.86	0.9

由表 2 可以看出使用重组详情作为分类的文本对象，在同一网络结构中，在每一个评价指标下的表现均更好。因此我们后续的数据均采用重组详情，其中含有效数据 9020 条（详见 1.xlsx）。

2.2.3 分词与去停用词

深度学习模型以文本的向量作为输入，在我们在讨论模型之前，要先对文本进行处理。我们首先进行分词与去停用词。

目前主流的分词工具有 LTP-3.2.0、ICTCLAS、jieba 等等。由于 jieba 处理速度快，分词效果好，且易于使用，所以我们选择 python 的 jieba 库来分词。利用停用词表（详见附件 stopwords.txt）和 jieba 实现分词和去停用词。结果如图 4：

['慧开 故里 国家 aaaaa 级 旅游 景区 建设 缓慢 相比 本土 浔 龙河 A6 区 铜 官窑 A8 县 炭 河里 古城 差距 很大 提出 建议 县 全域 旅游 游客 接待 中心 集 停车 导览 纪念品 特产 一体 建议 选址 慧开 镇慧开 陵园 道路交通 横向 纵向 道路 成网 黄伸 杨慧开 纪念馆 华星 通用 机场 加快 建设 青山 铺 107 国道 至慧开 道路 建设 楚女 多情 剧院 打造 一场 楚女 多情 演出 爱情故事 主题 优惠 政策 引进 大型 文旅 项目 万达 西地省 广电 华侨城 恒大 国内外 知名 投资商 对接 建设 影视城 旅游 度

图 4 分词结果示例图

经过去停用词和分词操作后，重组详情的总体词数分布如图 5：

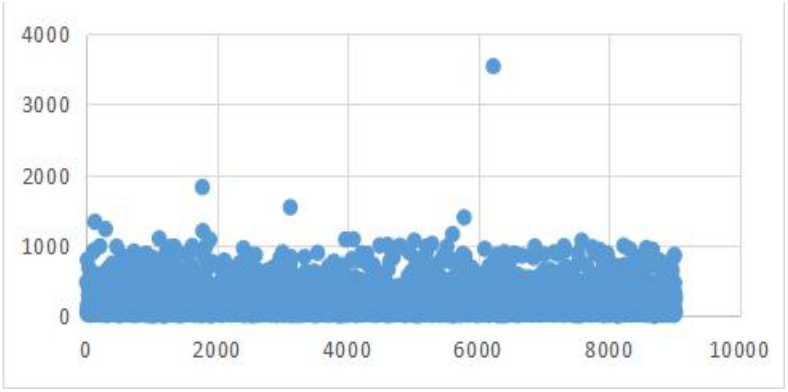

图 5 总体词数分布图

局部词数分布图如图 6：

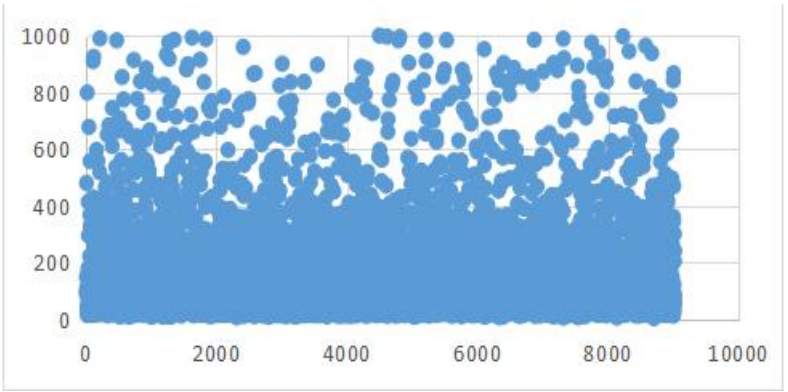

图 6 局部词数分布图

通过统计，500 词以内的留言条数占 96.2%。我们综合考虑，在后续的模型构建过程中，我们选定模型中的最大序列长度为 500。

2.2.4 词向量获取

深度学习模型的输入为向量，我们需要将分词的结果转化为向量，主要有三种方法，分别是独热编码，整数编码以及 Word Embedding（词嵌入）方法。从左至右如图 7：

图 7 三种编码示例图

独热编码是二进制的、向量稀疏的（主要由零组成）、高维度的（通常与词汇表中的单词数相当的维度），这给存储和运算带来了困难。而整数编码无法表达词语之间的关系，不便于模型解释。Word Embedding 既节省空间又能表达词语之间的语义关系。因此我们采用 Word Embedding 实现文本从到向量序列的转化。

通过 Word Embedding，我们得到了每条文本对应的词向量序列，也就得到了可以直接用于深度学习模型的输入数据。我们将得到的数据按照 9：1 的比例划分为训练集和测试集，分别用于模型的训练和测试。

2.3 文本分类

文本分类主要有基于传统机器学习和基于深度学习的两种方法。传统机器学习方法的特征提取工程任务繁重，且受主观因素影响大，不便于实现。而深度学习可以自动提取文本特征和类别标签之间的关联，还能从简单特征中提取复杂的特征，任务量小且准确率高。因此我们考虑采用基于深度学习的文本分类方法来解决群众留言分类。

基于深度学习的文本分类模型主要有 FastText，TextCNN，TextRNN 等模型。考虑到各种深度学习模型的运行速度和分类效果，我们设计了基于 FastText 模型和 TextCNN 模型两个方案来解决留言分类，测试两个模型的效果并进行比较，最终选定合适的群众留言分类的一级标签分类模型。

2.3.1 FastText 分类模型

FastText 模型的原理图如图 8：

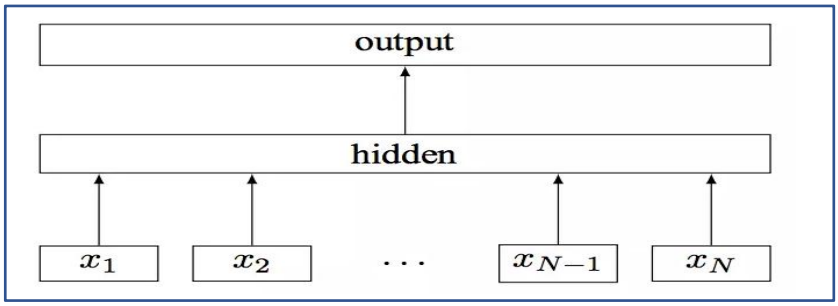

图 8 FastText 原理图

FastText 模型最大的特点就是简单高效，模型只有三层，以文本中的词向量做为输入，取平均得到隐藏层文本的向量表示，输出层为一个全连接层，并使用层次 softmax 计算文本属于每个类别的概率，从而得到文本的分类标签。FastText 模型能在保持准确率较高的情况下提高训练速度。

我们使用深度学习框架 Keras 搭建了 FastText 模型，将我们划分好的训练数据和测试数据输入模型，对模型进行训练和测试。结果如表 2 所示：

表 2 训练和测试结果

accuracy: 92.24%

	precision	recall	f1-score	support
0	0.88	0.86	0.87	95
1	0.96	0.92	0.94	192
2	0.87	0.90	0.88	186
3	0.94	0.97	0.96	167
4	0.95	0.96	0.96	102
5	0.92	0.89	0.91	66
6	0.94	0.93	0.93	94
accuracy			0.92	902
macro avg	0.92	0.92	0.92	902
weighted avg	0.92	0.92	0.92	902

最终 f1-score 为 0.92，准确率和召回率都比较高，模型分类效果良好。

附：FastText 实现代码详见 FastText.ipynb。

2.3.2 TextCNN分类模型

TextCNN 模型的原理图如图 9：

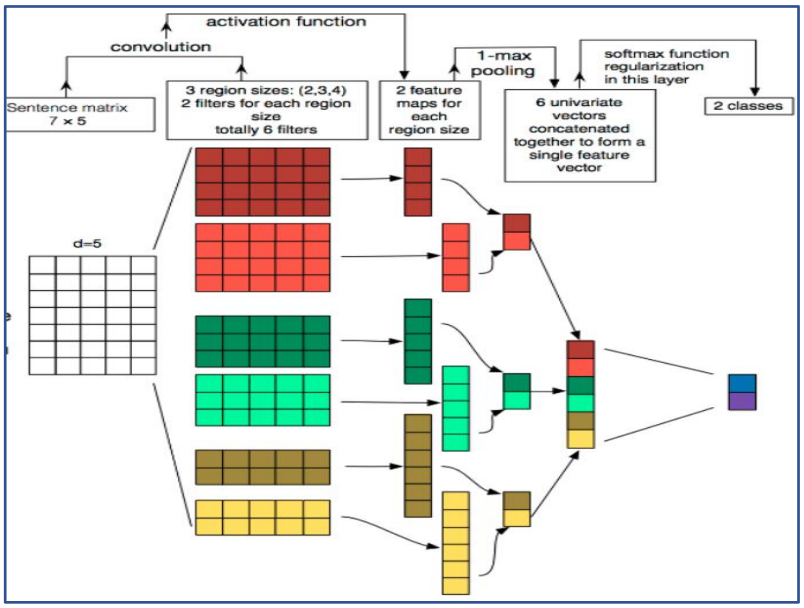

图 9 TextCNN 原理图

主要步骤为：（1）输入文本的词向量（2）卷积层使用多个卷积核以及多种卷积核窗口实现文本的局部特征提取。（3）通过池化层进一步提取文本特征（4）将提取到的特征输入全连接层，通过 softmax 函数激活得到类别标签。

我们同样使用深度学习框架 Keras 搭建了 TextCNN 模型，经过参数调试后及对应效果比较后，最终采用使用[2,3,4,5]四个卷积核窗口来构建我们的 TextCNN 模型。此外，我们还引入了 dropout 用来防止模型过拟合。我们同样对 TextCNN 模型进行训练和测试。结果如表 3 所示：

表 3 TextCNN 模型具体指标

accuracy: 91.35%

	precision	recall	f1-score	support
0	0.85	0.85	0.85	95
1	0.96	0.94	0.95	192
2	0.84	0.89	0.87	186
3	0.95	0.98	0.96	167
4	0.94	0.97	0.96	102
5	0.90	0.71	0.80	66
6	0.95	0.93	0.94	94
accuracy			0.91	902
macro avg	0.91	0.90	0.90	902
weighted avg	0.91	0.91	0.91	902

最终 f1-score 为 0.90，模型总体分类情况良好。

附：TextCNN 实现代码详见 TextCNN.ipynb。

2.4 群众留言分类总结

FastText 模型的正确率、精确率、召回率和 f1-score 均更高，分类效果更

好。

值得一提的是，FastText 模型单次迭代速度快，单次迭代时间基本稳定在 10 秒之内，但达到收敛的所需的迭代次数多。TextCNN 模型收敛较快，达到收敛的迭代次数在 10 次以内，但单次迭代时间基本为 1 分钟左右，花费的总时间稍长。

综上所述，我们分别设计了基于 FastText 模型和基于 TextCNN 模型的文本分类模型解决了群众留言分类问题。兼顾效率和准确率，我们最终选定 FastText 模型作为群众留言分类的一级标签分类模型。

三、热点问题挖掘的建模与解答

3.1 热点问题挖掘方案

解题流程图如图 10 所示：

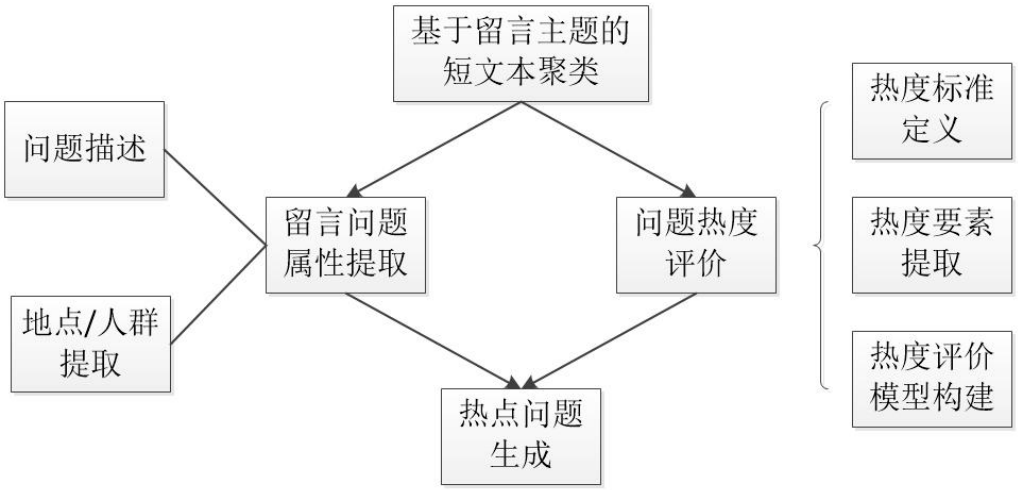

图 10 热点问题流程图

步骤①：基于留言主题的短文本聚类。对题目的文本进行预处理，建立停用词，根据关键词对留言进行聚类，然后对聚类的结果进行检验和矫正。

步骤②：留言问题属性提取。按照题目要求需要提取类的两个属性：问题描述、地点或人群。对于问题描述，可以从聚类结果中使用 TextRank 算法提取反映文本主题的句子作为问题的描述。对于地点和人群，可以进一步在问题描述的基础上用正则表达式提取。

步骤③：问题热度评价。在分析留言属性特征后，综合选取留言频率和用户

互动情况作为热度评价标准。并根据步骤①的结果，抽取出问题的热度要素，进行标准化处理，并运用信息熵进行赋权，最后得到归一化 TOPSIS 指标，对问题进行热度评价和排序。

步骤④：热点问题生成。依照指定的格式输出成 Excel 文件结果。

3.2 留言文本聚类

题目要求对反映特定问题的留言文本进行归类，本质上就是对留言根据反映问题聚类。对于文本聚类我们有以下思路：

文本聚类的流程图如图 11 所示：

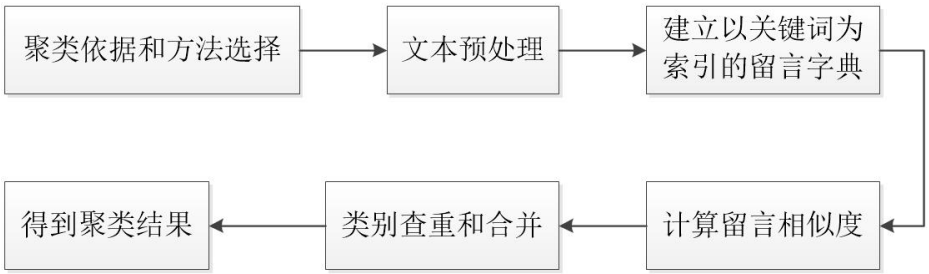

图 11 文本聚类流程图

3.2.1 聚类依据选定和方法选择

聚类依据：从附件 3 所给的信息可以看到，对于问题的归类主要依据留言的两个要素：留言主题和留言详情。考虑到留言主题已经可以很好的反应留言详情信息，所以我们主要根据留言主题对文本进行聚类，留言详情作为补充。当无法从留言主题中提取信息时，我们会转到留言详情去提取摘要作为主题。

聚类方法：因为不清楚留言一共反映了多少个问题，即类的个数不定，所以我们采用类数不定的聚类方法；根据聚类的对象都是短文本，且反映相同问题的留言中具有很大比例相同的关键词这个规律，我们决定以句子之间关键词的重复比例作为判定句子是否在同一类的标准。

3.2.2 建立专用停用词库

停用词作为对题目聚类意义不大的词语，考虑到题目作为政务系统的特殊性，其停用词库不能只用基础的常用停用词，比如“市”作为一个地点的标志，在一般的分类中是需要识别的，但在本题中，基本上问题都是关于 A 市的问题，所以

“市”应该被列为停用词。

大致观察给出的留言主题数据后，我们发现以下两类词适合作为停用词：

一类是出现频率低的成语或者感叹词等词，这类词只代表极少部分问题，且不是其主要特征。比如：

A2 区暮云街道西湖集资办下石岭塘组盗采砂矿屡禁不止。

这个留言主题的主要特征是采砂，屡禁不止是其特征并不是主要特征

另一类是出现频率较高的副词或者名词等，这类词能区分问题特征低。比如以下两个留言主题：

投诉 A2 区丽发新城附近建搅拌站噪音扰民。

A3 区中海国际社区空地夜间施工噪音太大了。

这两句都是噪音问题，但其发生的地点不同，不属于一类问题。

我们通过遍历文本进行 jieba 分词后，统计词频，从中简单的筛选词频靠前和靠后的词建立了包含 7418 个元素的停用词表。两类的示例类型如表 7、8：

表 4 低频停用词

分类	低频	频率
数字	1	1
语气词	千万别	1
成语	屡禁不止	1
专用指代	朱某	1
副词	大力	1
介词	由于	1

表 5 高频停用词

分类	高频频率
地点后缀	小区 549
问题区分度低的动词	建议 167
问题区分度低的名词	噪音 147
无意义形容词	无法 26
介词	关于 78

我们将挑选的 7418 个停用词表储存到 stop_words 文件中。

另外，由于 jieba 分词水平的限制，存在诸如“市丽发”这样的一半是没有意义的“市”，一半是有意义的地点名词“丽发”，这样的词我们需要在具体的程序中通过词语的匹配，去除没有意义的部分将词语转化成非停用词。

3.2.3 聚类过程与结果

1.建立以关键词为索引的留言文本字典

我们将留言主题使用 python 的 pandas 模块从附件 3 中提取出来后，设计了一个类，在类中调用现有的 jieba 分词包进行文本的预处理。逐条调用类，生成分词实例，对比停用词并进行删除，得到以关键词为索引的留言文本字典，字典的 key 为关键词，value 为包含该关键词的句子。

我们生成了包含 7032 个关键词的词典，从字典中可以看到其出现的句子：

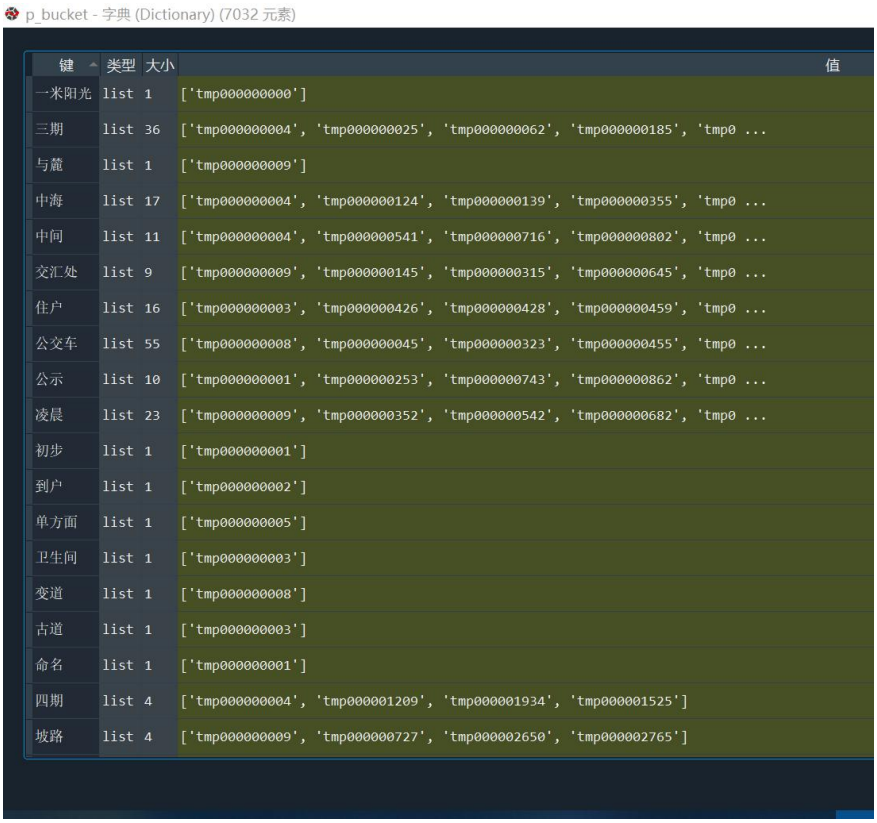

图 12 以关键词为索引的留言文本字典

2.计算留言相似度

通过字典，把这些关键词按照句子为单位储存为二维数组，通过一个简单的函数计算它们之间关键词的重复的部分和非重复部分的比值，高于某个值就被视为反映同一问题的留言，并写入这个类所在文件里。当所有的文件都匹配完毕，更改文件的名称为特定格式，我们就完成了句子的聚类。经过多次调试试验，我们发现将这个值定义在 0.4 时，结果中基本不会出现不同问题被聚类到同一类的情况，类的数目也还可以接受。

最后我们得到了 3689 个类，参考 output 文件夹：

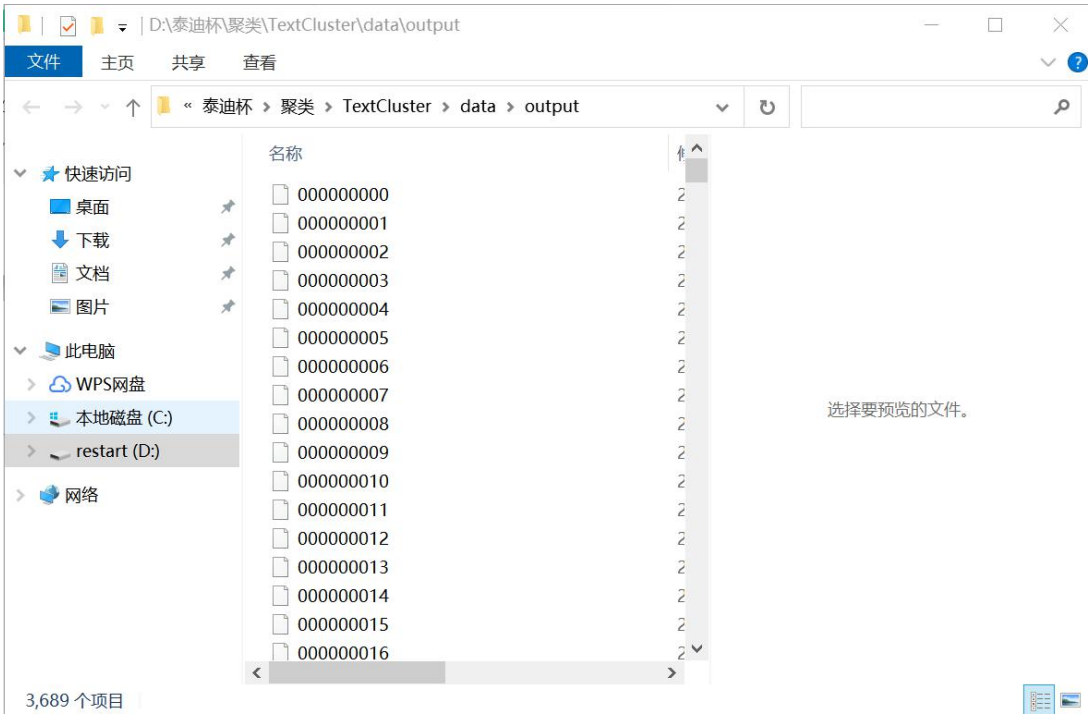

图 13 初步聚类文件

选取第一个类内容进行查看，可以看到初步效果基本符合题意：

表 6 初步聚类后的文件 000000000

留言序号	留言主题
89	关于伊景园滨河苑捆绑销售车位的维权投诉
319	投诉 A 市伊景园滨河苑捆绑车位销售
703	伊景园滨河苑捆绑车位销售合法吗？！
731	坚决反对伊景园滨河苑强制捆绑销售车位
787	伊景园滨河苑强行捆绑车位销售给业主
1296	A 市伊景园滨河苑捆绑销售车位
1445	A 市伊景园滨河苑定向限价商品房项目违规捆绑销售车位
1483	投诉 A 市伊景园滨河苑捆绑销售车位
1545	伊景园滨河苑车位捆绑销售！广铁集团做个人吧！
1778	投诉 A 市伊景园滨河苑捆绑车位销售
2014	和谐社会背景下的 A 市伊景园滨河苑车位捆绑销售

2346	关于伊景园滨河苑捆绑销售车位的投诉
2659	投诉伊景园滨河苑项目违法捆绑车位销售
2830	违反自由买卖的 A 市伊景园滨河苑车位捆绑销售行为
2937	投诉伊景园滨河苑捆绑销售车位问题
3051	投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位
3389	惊！！A 市伊景园滨河苑商品房竟然捆绑销售车位
3732	A 市伊景园滨河苑捆绑销售车位是否合理？
3840	投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位
4039	A 市伊景园滨河苑项目捆绑销售车位
4138	武广新城伊景园滨河苑违法捆绑销售车位,求解决
4273	反对滨河苑房子和车位捆绑销售
4298	投诉 A 市伊景园滨河院捆绑销售车位

3.类别查重和合并

我们在后续的查看中发现存在一些不同的类中留言反映的是一个问题的聚类瑕疵，主要原因有两个：一个是上文已经提到分词水平不足，另一个是建立的停用词库并不完全适应题目。其很难通过再次聚类的方法完善，要想进一步得到聚类的准确结果，需要对类进行查重和合并。

在聚类的过程中，我们发现基本上类里占比最高的关键词是其主要特征，如“搅拌站”，占比第二高的关键词则是问题地点，如“丽发”。所以可以依据类的这两类关键词进行查重和合并类，即遍历所有类，统计各类的关键词的占比，取占比最高的非停用词作为第一关键词，取占比第二高的非停用词作为第二关键词。如果类别的第一停用词和第二关键词均相同，那其反应的问题很可能是同一类，我们就把其合并，输出成一个新的文件，然后删除那些被合并的文件。最终我们得到一共 3617 个聚类结果，参考 finalout 文件夹：

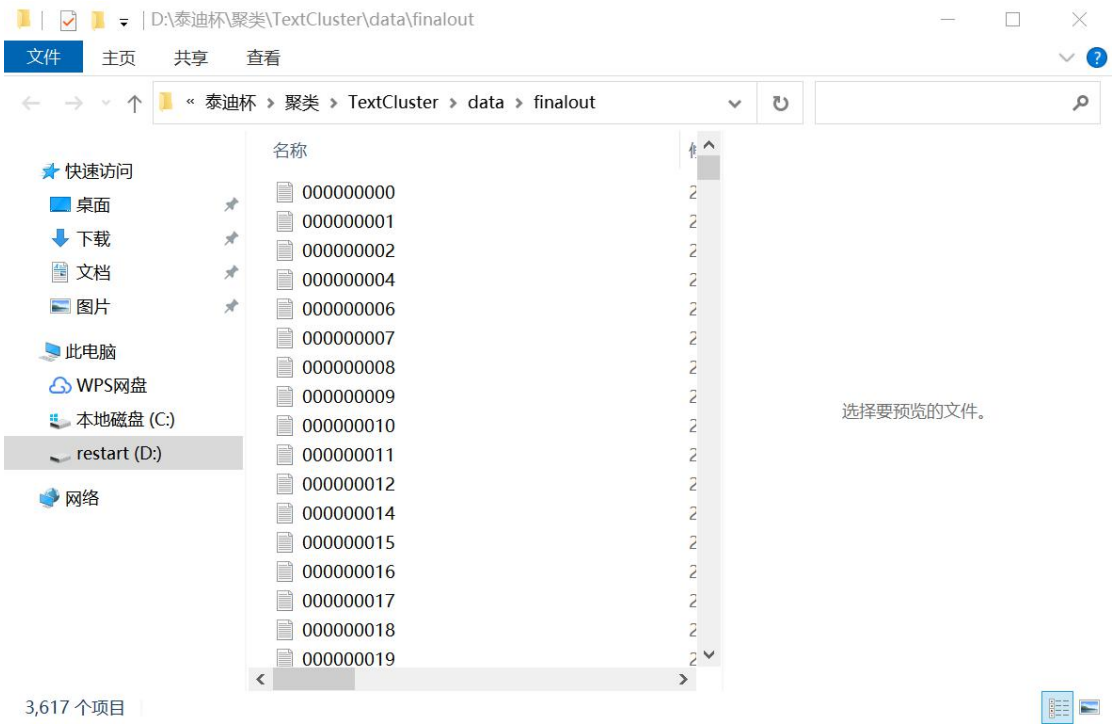

图 14 合并后的聚类文件

取之前的类查看，结果比较符合题意：

表 7 查重合并后的文件 000000000

留言序号	留言主题
89	关于伊景园滨河苑捆绑销售车位的维权投诉
319	投诉 A 市伊景园滨河苑捆绑车位销售
703	伊景园滨河苑捆绑车位销售合法吗？！
731	坚决反对伊景园滨河苑强制捆绑销售车位
787	伊景园滨河苑强行捆绑车位销售给业主
1296	A 市伊景园滨河苑捆绑销售车位
1445	A 市伊景园滨河苑定向限价商品房项目违规捆绑销售车位
1483	投诉 A 市伊景园滨河苑捆绑销售车位
1545	伊景园滨河苑车位捆绑销售！广铁集团做个人吧！
1778	投诉 A 市伊景园滨河苑捆绑车位销售
1952	无视消费者权益的 A 市伊景园滨河苑车位捆绑销售行为
2014	和谐社会背景下的 A 市伊景园滨河苑车位捆绑销售

2122	请维护铁路职工权益取消伊景园滨河苑捆绑销售车位的要求
2346	关于伊景园滨河苑捆绑销售车位的投诉
2659	投诉伊景园滨河苑项目违法捆绑车位销售
2937	投诉伊景园滨河苑捆绑销售车位问题
3051	投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位
3389	惊！！A 市伊景园滨河苑商品房竟然捆绑销售车位
3732	A 市伊景园滨河苑捆绑销售车位是否合理？
4039	A 市伊景园滨河苑项目捆绑销售车位
4138	武广新城伊景园滨河苑违法捆绑销售车位,求解决
4148	无视职工意愿、职工权益的 A 市伊景园滨河苑车位捆绑销售行为
4273	反对滨河苑房子和车位捆绑销售
2830	违反自由买卖的 A 市伊景园滨河苑车位捆绑销售行为
3840	投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位
4298	投诉 A 市伊景园滨河院捆绑销售车位
890	伊景园滨河苑项目绑定车位出售是否合法合规
2445	举报广铁集团在伊景园滨河苑项目非法绑定车位出售
2230	强行要求捆绑车位，请有关政府部门为民做主
3717	车位属于业主所有，不应该被捆绑销售！

通过表 9 和表 10 两次结果对比，可以看到聚类结果得到了完善。

3.3 留言问题属性提取

3.3.1 问题描述

我们已经获得了一个类的所有单条留言的留言主题，那么其描述问题，就变成了文本的摘要问题。这里我们选择了 TextRank[1]算法，其是利用重要性选取文章关键句的常用算法。我们先对文本进行分词，去停用词预处理后，得到句子和词汇的集合，并建立了一个字典，用字典记录一个文件中词的出现次数，并通过相邻关系构建出一个有向图。然后把集合看成一张有向图，每个单词作为图中

的一个节点，而能成为关键词的节点是那些度高的节点。我们使用词语同现的关系来控制词节点之间是否连接：把一个句子划分成多个窗口，如果在一个最大的窗口N个词语里两个相关的词语同现，那么这两个顶点是连接的。这样，我们就获得了一个有向无权图，对于这个无权图我们需要根据节点度数计算出和留言文本内容关联程度较高的词作为关键词，计算方法如下公式（2）：

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (1)$$

说明：$S(V_i)$是节点 i 的重要性；d 是阻尼系数，一般设置为 0.85；$\ln(V_i)$是存在指向网页 i 的链接的网页集合。$Out(V_i)$是网页 j 中的链接存在的链接指向的网页集合。初始时，可以设置每个节点的重要性为 1。

然后，我们对最终的到节点的重要性进行排序，取一定的比例得到了一类留言文本的关键词。利用这些关键词，我们就可以通过组合得到关键短语，甚至是主题语句。组合方法：若原文本中存在若干个关键词相邻的情况，那么这些关键词可以构成一个关键短语。

最后，当关键短语到达一定的数量，包含这个短语的句子就可以作为这类留言的问题描述。提取结果示例如下：

表 8 得到的问题描述示例

类序号	问题描述
1	投诉 A 市伊景园滨河苑捆绑车位销售
2	丽发新城小区附近的搅拌站噪音严重扰民
3	咨询 A3 区西湖街道茶场村五组的拆迁规划
4	咨询 A 市人才购房补贴通知问题
5	建议进一步优化 A 市经开区泉星公园项目规划
6	请加快 A 市国家中心城市建设刻不容缓
7	坚决反对在 A7 县诺亚山林小区门口设置医院
8	A7 县新国道 107 距我家仅 3 米，相关政府部门为何不同意拆迁
9	询问 A 市住房公积金贷款的相关问题
10	希望带着情怀提升 A 市规划建设水平带动经济发展
11	A 市 A3 区兰亭湾畔小区违法开餐厅

12	A 市江山帝景新房有严重安全隐患
13	对 A 市公交线路的建议
14	按照当前的高铁规划，A 市绿地海外滩小区会饱受噪音困扰
15	A2 区龙湾国际社区别墅区违建现象严重
16	A 市教师招聘考试不公平
17	A 市绿地城际空间站建筑质量堪忧
18	再次反映 A 市金茂府一房二卖
19	A 市 805 路公交车能不能改道走木莲中路
20	强烈建议将地铁 7 号线南延至 A 市生态动物园

3.3.2 地点，人群提取

题目表格中存在地点和人群这一项，所以对于问题中的地点和人群我们需要进行地点和人群的识别和提取。因为在之前的工作中我们已经获得的问题的描述中含有相应的地点和人群，所以我们直接从每个类生成的问题描述中进行地点和人群识别。

观察文本中地点和人群的关键词特征，我们发现问题地点是虚拟且开头含有字母和数字，即如果用机器学习的实体识别，在分词阶段，地点字段就可能被破坏，所以我们需要先分析地点和人群的格式特点。

对于地点，我们总结出了表 12 中所示的三种格式特点：

表 9 地点格式

格式	举例	识别结果
字母或数字+形容性名词+地点标志词	A7 县未来漫城物业不作为，还能评为五星级	A7 县未来漫城
字母或数字+地点标识词	咨询 A7 县道路规划的问题	A7 县道路
形容性名词+地点性标志词	丽发新城小区附近的搅拌站噪音严重扰民	丽发新城小区附近的搅拌站

对于人群，我们发现其绝大多数是通过地点+常见人群名称（如“职工”，“市长”，“考生”等）构成的，所以可以套用地点识别的格式。

在分析出实体识别的格式后，我们发现可以通过简单却高效的正则表达式匹配的方法从问题描述中提取出相应的地点或人群，识别结果如表 12 所示。

3.4 问题热度评价

构建流程图：

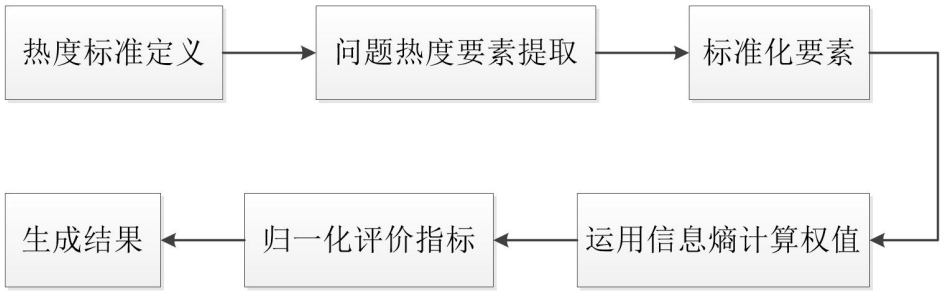

图 15 问题热度评价流程图

3.4.1 热度标准定义

在计算热度之前，我们需要对热度有一个明确的定义。根据题目的定义，从传播学角度，我们认为，一个事件的热度应该和用户的关注度和参与度相关，即在一定的时间内，反映的用户数越多，用户的表态越多，则问题的热度越高。考虑到题目没有提供时事背景以及群众在政府工作方面需求多样，所以我们忽略时事热点带来差异，假设每个类的问题初始热度都为 0。为了量化指标，通过观察留言文本，我们定义问题的热度与两个指标相关：留言频率，点赞数和反对数。

3.4.2 问题热度要素提取

我们先把点赞数和反对数分别以数组的方式从附件 3 中提取出来，并使用聚类结果中的序号从时间、点赞数和反对数数组中抽取对应的时间，点赞数和反对数，这样就获得了评价热度的基本要素。

1.计算留言频率

若类中存在两条或以上留言，则可能存在在同一个类中一部分留言是同一个用户发表的，其不能反映整体群众对问题的关注度，需要在频率计算时统计并去除，然后通过公式（3）得到其留言频率：

$$F_i = \frac{m_i - r_i}{t_{i,\max} - t_{i,\min}} \tag{2}$$

说明：其中F_i表示第 i 类的留言频率，m_i表示第 i 类的留言条数，r_i表示第 i 类内同一个用户的留言数减 1 的总和，$t_{i,\max}$和$t_{i,\min}$分别表示第 i 类中留言时间的最大和最小值。

若一个类中只有一条留言，则我们认为这类留言的该要素为所有非单条留言类最小值频率的一半，即频率最小的含有两条留言类的一半。这么做的原因是因为如果设定为 0，这样对于那些虽然只有单条留言的高点赞数或反对数的类不公平，而如果设定为平均值，则对于其他不是单条留言的类同样不公平。

2．提取点赞数和反对数

对于点赞数和反对数，我们认为两者都是用户对话题关注的体现，地位和作用均等，所以我们直接通过编号提取并相加。

通过以上这些处理，我们成功提取到了热度评价的要素。

3.4.3 热度评价模型的构建

优劣解距离法（TOPSIS 方法）是一种成熟的综合评价方法，其能充分利用原始数据的信息，结果能精确地反映各评价方案之间的差距，生成较为理想的排序结果。

标准 TOPSIS 方法的过程简述如下：先将原始数据矩阵统一指标类型（一般正向化处理） 得到正向化的矩阵，再对正向化的矩阵进行标准化处理以消除各指标量纲的影响，并找到有限方案中的最优方案和最劣方案，然后分别计算各评价对象与最优方案和最劣方案间的距离，获得各评价对象与最优方案的相对接近程度，以此作为评价优劣的依据。

本题中，为了让热度评价结果更准确，我们对 TOPSIS 方法进行了一些改进，根据题目特性，采用带有权值的 TOPSIS 方法。

1．转化指标类型

我们首先根据热度评价要素构建矩阵，式中的 m 为类数，n 为指标数：

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}$$

然后我们需要对矩阵进行正向化，在 TOPSIS 方法中，指标简单分为 4 类：极大型指标、极小型指标、中间型指标、范围型指标，其特征如表 13 所示：

表 10 指标类别及特征

指标名称	指标特点	例子
极大型（效益型）指标	越大（多）越好	企业利润
极小型（成本型）指标	越小（少）越好	坏品率
中间型指标	越接近某个值越好	水质量评估时的 PH 值
范围型指标	落在某个区间最好	体温

显然，在本题需要的三个热度评价要素中，留言频率、点赞数和反对数都是极大型指标。对于极大型指标我们采用如下公式（4）正向化数据：

$$y_{i,j} = \frac{x_{i,j} - \min(x_{i,j})}{\max(x_{i,j}) - \min(x_{i,j})} \tag{3}$$

说明：式中 $y_{i,j}$ 表示正向化后第 j 项指标下第 i 类的数据的值，$x_{i,j}$ 表示正向化前第 j 项指标下第 i 类的数据的值。

2. 矩阵标准化

在得到正向化矩阵后，我们还需要对矩阵进行标准化，目的是为了去除量纲对指标的影响，我们采用如下公式（4）进行矩阵标准化：

$$z_{i,j} = y_{i,j} / \sqrt{\sum_{i=1}^{n} y_{i,j}^2} \tag{4}$$

说明：式中 $z_{i,j}$ 表示标准化后第 j 项指标下第 i 类的数据的值，$y_{i,j}$ 表示标准化前第 j 项指标下第 i 类的数据的值。

3. 计算指标权值

在标准化后，接下来应该进行矩阵的归一化，而简单的无权归一化在解决实际问题中有失严谨，所以我们需要计算每个指标的权重。我们采用计算方法为基于信息熵[2]的文本权重计算。

信息熵是香农提出的概念，在后来的解释中被认为是信息的特征之一，其意义之一是反映信息价值的规律，即信息的数量越大，则其单位信息的价值越低，信息产生的概率越低，其价值越大。

在本题中，我们可以理解为，当一个指标，其指标内的差距越大，则这个指标反映的问题特征的价值越低，即权重越低，相反则越高，这样就有了一个可以计算的权重指标。

我们采用如下方法计算：

在一个指标内，计算其所占百分比：

$$p_{i,j} = \frac{z_{i,j}}{\sum_{i=1}^{m} z_{i,j}} \tag{5}$$

计算指标的熵值：

$$e_j = -k\sum_{i=1}^{m} p_{i,j} \ln p_{i,j}, k = \frac{1}{\ln m}\left(k > 0, 0 \leqslant e_j \leqslant 1\right) \tag{6}$$

对熵值进行归一化：

$$w_j = \frac{1-e_j}{\sum_{j=1}^{n}(1-e_j)} = \frac{1-e_j}{n-\sum_{j=1}^{n} e_j} \tag{7}$$

（$0 \leq w_j \leq 1$ 且 $\sum_{j=1}^{n} w_j = 1$）

说明：式中的下标 j 均表示第 j 个指标，i 均表示第 i 个类，n 表示指标数。

4. 归一化评价指标

进行归一化处理，我们需要求出每一项指标内的最值：

$$\begin{aligned} Z^+ &= \left(Z_1^+, Z_2^+, \cdots, Z_m^+\right) \\ &= (\max\{z_{11}, z_{21}, \cdots, z_{n1}\}, \max\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \max\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \end{aligned} \tag{8}$$

$$\begin{aligned} Z^- &= \left(Z_1^-, Z_2^-, \cdots, Z_m^-\right) \\ &= (\min\{z_{11}, z_{21}, \cdots, z_{n1}\}, \min\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \min\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \end{aligned} \tag{9}$$

然后计算其每一项赋权的欧拉距离：

$$D_i^+ = \sqrt{\sum_{j=1}^{m} w_j \left(Z_j^+ - z_{ij}\right)^2} \tag{10}$$

$$D_i = \sqrt{\sum_{j=1}^{m} w_j \left(Z_j^- - z_{jj}\right)^2} \tag{11}$$

最后求平均值得到最终的热度：

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{12}$$

说明：式中的下标 j 均表示第 j 个指标，i 均表示第 i 个类，m 表示指标数，n 表示类数。

3.4.4 热点排序算法

在得到地点或人群、热度等要素后，我们还根据聚类结果统计出题目要求的问题时间范围，并创建出带有题目要求属性的类后，依次生成类的实例，并在实例中按照热度属性使用 python 自带的 sorted 函数排序，得到了热点问题的排名。

3.5 热点问题生成

3.5.1 生成热点问题表

我们按照题目要求取前五名按指定的格式将结果输出成 excel 文件，结果如表 14 所示：

表 11 热点问题表

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	0.18908	2019-07-07 至 2019-09-01	A 市伊景园	投诉 A 市伊景园滨河苑捆绑车位销售
2	2	0.12009	2019-01-11 至 2019-03-01	A 市 58 车	西地省 A 市 58 车贷恶性退出，A4 区立案已近半年毫无进展
3	3	0.11171	2019-05-05 至 2019-09-19	A 市 A5 区汇金路五矿万境 K9 县	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
4	4	0.099	2019-04-11	A 市金毛湾	反映 A 市金毛湾配套入学的问题
5	5	0.04365	2019-09-05	A4 区绿地海外滩小区	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗

3.5.2 生成热点问题留言明细表

在实现热点问题排序后，我们只需要找的留言类里的留言编号就能从附件 3 中抽取出热点问题明细表，抽取的到的结果如表 15 所示：

表 12 热点问题留言明细表

问题ID	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
1	190337	A00090519	于伊景园滨河苑捆绑销售车位的维权扌	2019-08-23 12:22:00	E集团下发文件，强制要求职工再交	0	0
1	196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019-08-07 19:52:14	国家三令五申禁止捆绑车位销售，	0	0
1	205277	A909234	尹景园滨河苑捆绑车位销售合法吗？！	2019-08-14 09:28:31	盘时捆绑购买12万一个的车位，不买	0	1
1	205982	A909168	决反对伊景园滨河苑强制捆绑销售车	2019-08-03 10:03:10	这是明显的违法捆绑销售！通过购买	0	2
1	207243	A909175	尹景园滨河苑强行捆绑车位销售给业主	2019-08-23 12:16:03	购买车位，不买车位就取消购买资格	0	0
1	218709	A00010669	A市伊景园滨河苑捆绑销售车位	2019-08-01 22:42:21	签订正规购房合同，强制收取18万	0	1
1	222209	A00017171	滨河苑定向限价商品房项目违规捆绑	2019-08-28 10:06:03	取消购房资格相要挟，逼迫职工交钱	0	0
1	223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019-07-23 17:06:03	取消车位捆绑。2.希望这个房子对耶	0	0
1	224767	A909176	滨河苑车位捆绑销售！广铁集团做个	2019-07-30 14:20:08	我合同，说什么预购不用！后面就	0	0
1	230554	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019-08-19 10:22:44	是伤筋动骨的，购房者中有一部分	0	0
1	234633	A909194	者权益的A市伊景园滨河苑车位捆绑	2019-08-20 12:34:20	职工购买房子的同时一对一购买所	0	0
1	236301	A909197	会背景下的A市伊景园滨河苑车位捆	2019-08-30 16:32:12	，这个还是和谐社会么？政府的正	0	0
1	239032	A909169	职工权益取消伊景园滨河苑捆绑销售	2019-09-01 10:03:10	同，现在还要求一户一车位，捆绑	0	1
1	244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	2019-08-24 18:23:12	于业主共有，但属于城镇公共绿地和	0	0
1	251844	A909167	诉伊景园滨河苑项目违法捆绑车位销	2019-08-20 13:34:12	子，违法捆绑车位销售，难道就没地	0	1
1	258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019-08-23 11:46:03	盘还未建成,广铁集团却要求捆绑购	0	0
1	260254	A909173	伊景园滨河苑开发商违法捆绑销售无	2019-08-30 18:10:23	文整治房地产市场乱象，A市七部委	0	0
1	268299	A909193	A市伊景园滨河苑商品房竟然捆绑销售	2019-08-21 15:32:33	要求职工购买房子的同时一对一购买	0	0
1	276460	A909170	伊景园滨河苑捆绑销售车位是否合理	2019-08-24 17:23:11	没有车的并没有购买车位的需求。网	0	0
1	283879	A00044759	A市伊景园滨河苑项目捆绑销售车位	2019-07-18 20:27:40	多是新入职的员工，经济条件负担不	0	0
1	285897	A909191	城伊景园滨河苑违法捆绑销售车位,	2019-08-01 20:06:52	广铁集团就要求捆绑购买车位,还说	0	0
1	286304	A909196	、职工权益的A市伊景园滨河苑车位	2019-08-23 10:23:23	的车位需求么？有考虑过职工能否承	0	0

这样，我们就完成了第二问的所有任务。

四、答复意见评价方案的建模与解答

4.1 问题分析

问题三旨在对留言的答复意见进行综合全面的评价，并形成一套较为完整的评价方案。为完成这个目标，我们将从相关性、完整性、及时性与可解释性四个方面来对留言进行评价。

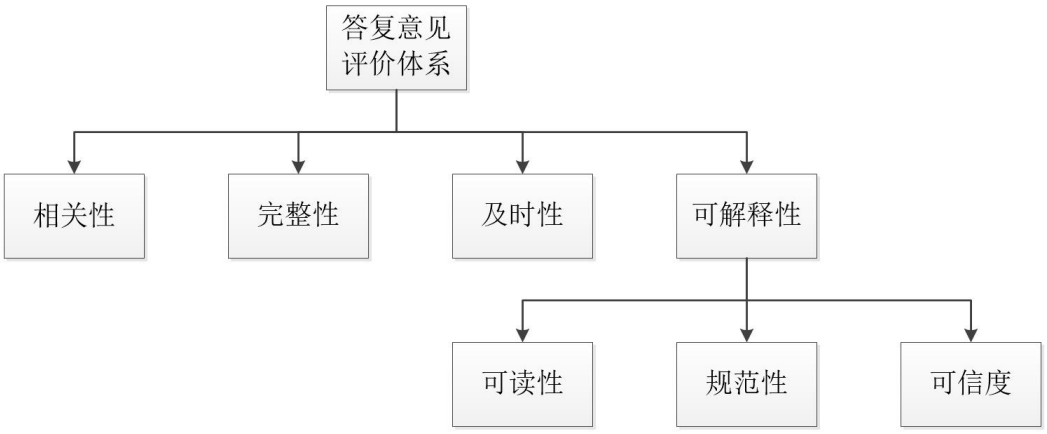

图 16 答复意见评价体系构架图

相关性是指答复意见与留言的关联程度，通过相关性评价判断答复意见与留言的主题是否相同，同时形成相关性指标，对答复意见的相关性做出评价。

完整性指标，用于评价答复意见的完备程度，具体是指答复意见对留言所提出的问题是否进行了全面的回答。

及时性，用于评价相关部门对留言的处理效率，主要考虑留言与回复之间的时间差。

可解释性是一个较为宽泛的概念，首先我们对可解释性这个概念进行细化和分解。

在本题当中，可解释性是指答复意见是否易于市民的阅读和理解，对政府而言该答复意见是否具有代表性，对市民而言该答复意见是否具有说服力和可信度。因此，我们将可解释性这个概念，细化分解为可读性，规范性和可信度三个方面，对答复意见的质量进行评价。

可读性用于评价答复意见是否易于市民更好地了解该答复的核心内容，可读性强的答复意见具有语句通顺，逻辑结构清晰，核心关键词突出的特点。规范性，指答复意见的格式和用词是否遵照一定的标准。可信度则是指答复意见是否具有真实性和说服力，是否能够代表相关部门对该留言的意见和看法。

从以上四个方面对答复意见进行评价，并形成指标。

4.2 相关性指标

4.2.1 总体思路

相关性即为答复意见与留言的关联程度，在此用两者间的文本相似度来表示。

单纯用余弦相似度计算的文本相似度，简单方便，但可靠性不高，与人工识别得到的相似度相差较大，不能满足此任务的需要。

在此对余弦向量算法进行优化，具体做法是：选用留言、留言主题与答复意见数据，在对数据进行预处理后，进行中文分词，对分词结果使用 dictionary 方法与 doc2bow 函数进行文本特征向量表示，筛选掉特征值较低的特征向量，最终使用基于余弦向量算法的 gensim 包分别计算出留言、留言主题两者与答复意见的文本相似度。

相似度计算流程如下：

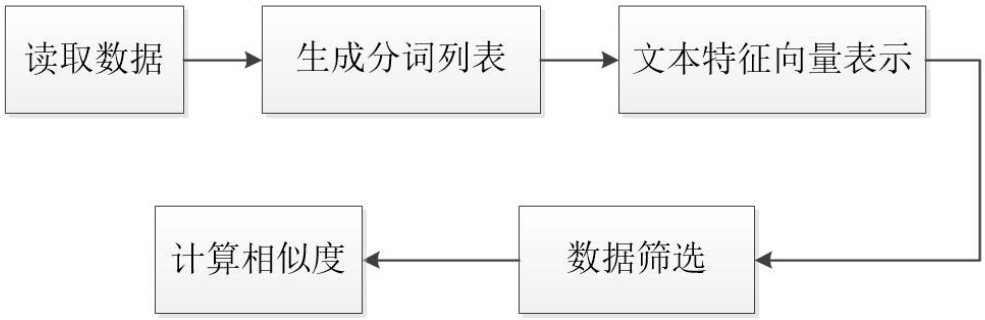

图 17 相似度计算流程图

4.2.2 具体步骤

1. 中文分词

目前，jieba 分词是使用最广泛的中文分词方式，具有简便易用、可自定义词表等优点。此处使用 jieba，对答复意见、留言、留言主题分别进行分词。

原文与分词后的结果示例如下：

2019年4月以来，位于A市A2区桂花坪街道的A2区公安分局宿舍区（景蓉华苑)出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买，水4.23一吨，电0.64一度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交20万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪?

['2019', '年', '4', '月', '以来', ',', ',', '位于', 'A', '市', 'A2', '区', '桂花', '坪', '街道', '的', 'A2', '区', '公安分局', '宿舍区', '（', '景蓉华苑', '）', '出现', '了', '一番', '乱象', ',', ',', '该', '小区', '的', '物业公司', '美顺', '物业', '扬言', '要', '退出', '小区', ',', ',', '因为', '小区', '水电', '改造', '造成', '物业公司', '的', '高昂', '水电费', '收取', '不了', '(', '原', '水电', '在', '小区', '买', ',', ',', '水', '4.23', '一吨', ',', ',', '电', '0.64', '一度', '）', '所以', '要', '通过', '征收', '小区', '停车费', '增加收入', ',', ',', '小区', '业委会', '不知', '处于', '何种', '理由', '对', '该', '物业公司', '一再', '挽留', ',', ',', '而', '对', '业主', '提出', '的', '新', '应聘', '的', '物业公司', '却', '以交', '20', '万', '保证金', ',', ',', '不能', '提高', '收费', '的', '苛刻', '条件', '拒之门外', ',', ',', '业委会', '在', '未', '召开', '全体', '业主大会', '的', '情况', '下', ',', ',', '制定', '了', '一', '高昂', '收费', '方案', '要', '各', '业主', '投票', ',', ',', '而', '投票', '不', '采用', '投票箱', '只', '制定', '表格', '要', '物业公司', '人员', '这一', '利害关系', '机构', '负责', '组织', ',', ',', '对', '投票', '业主', '隐私权', '没有', '任何', '保护', ',', ',', '还', '对', '投', '反对票', '的', '业主', '以', '领导', '做', '工作', '等', '方式', '要求', '改变', '为', '同意', '票', ',', ',', '这种', '投票', '何来', '公平', '公正', '公开', ',', ',', '面对', '公安干警', '采用', '这种', '方式', '投票', '合法性', '在', '哪', '？', '']

图 18 分词结果示意图

2. 文本特征向量表示

余弦相似度算法的作用对象是向量，因此需要把用于计算的文本表示为特征向量。

Gensim 库中的 doc2bow 函数可以无监督地实现将分词列表转换为稀疏向量，并存放在列表中，并为每个分词分配一个 ID。其中，向量中的每个元素都是一个二元组，分别对应该分词地编号与频次数。

使用 Gensim 库实现文本特征向量表示过程如下：

（1） 用 dictionary 方法对留言与留言主题进行处理，获取词袋，同时得到特征值。

（2） 基于以上词袋，使用 doc2bow 制作稀疏向量。

转换完成的稀疏向量示例图 19 所示：

[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)], [(0, 2), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 4), (25, 1), (26, 2), (27, 1), (28, 1), (29, 1), (30, 2), (31, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 1), (37, 1), (38, 1), (39, 1), (40, 1), (41, 1), (42, 1), (43, 1), (44, 1), (45, 1), (46, 1), (47, 1), (48, 1), (49, 1), (50, 1), (51, 2), (52, 2), (53, 1), (54, 1), (55, 1), (56, 1), (57, 1), (58, 1), (59, 1), (60, 1), (61, 1), (62, 1), (63, 3), (64, 1), (65, 1), (66, 1), (67, 1), (68, 4), (69, 6), (70, 1), (71, 1), (72, 1), (73, 1), (74, 1), (75, 1), (76, 1), (77, 1), (78, 1), (79, 5), (80, 1), (81, 1), (82, 1), (83, 1), (84, 1), (85, 1), (86, 2), (87, 1), (88, 1), (89, 1), (90, 2), (91, 1), (92, 1), (93, 1), (94, 1), (95, 1), (96, 1), (97, 1), (98, 1), (99, 2), (100, 1), (101, 1), (102, 1), (103, 5), (104, 1), (105, 1), (106, 8), (107, 1), (108, 1), (109, 1), (110, 1), (111, 2), (112, 1), (113, 1), (114, 1), (115, 4), (116, 1), (117, 2), (118, 1), (119, 1), (120, 1), (121, 2), (122, 1), (123, 1), (124, 1), (125, 2), (126, 1), (127, 1), (128, 1), (129, 2), (130, 1), (131, 1), (132, 15), (133, 1)]]

图 19 稀疏向量示例图

为使相似度结果更加准确可靠，使用 TextRank 算法，只筛选出特征值排序前 20 的分词进行相似度计算。

相似度计算结果如下：

图 20 相似度计算结果图

4.2.3 相关性指标建立

留言主题文本较短，但包含信息重要性高；留言文本长度较大，描述更加细致，特征更多，但与主题无关的信息也更多。因此取两者与答复间相似度的权值比为 0.5：0.5。

分别得到答复意见与留言主题和留言内容的相似度后，求得两个相似度数值间的平均数 M，M 即为相关性指标。

经过处理，得到相关性指标如图 21 所示。

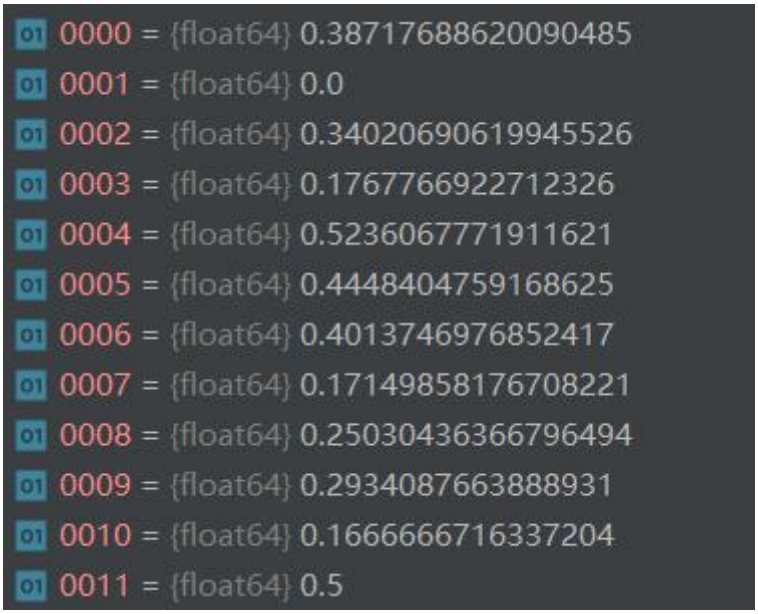

图 21 相关性指标示例图

4.3 完整性指标

4.3.1 总体思路

完整性是指答复意见的完备程度。答复意见越能全面回答留言所提出的问题，也就具备越高的完整性。为评价答复意见的完整性，主要判断留言中提出的多个主题与问题，答复中是否都进行了提及与回答。

在本任务中，首先对留言与答复意见数据分别进行预处理。使用 TextRank 算法，从留言中提取出三个核心句并记录关键值，这三个核心句可以代表本条留言的主要内容以及提出的多个问题。而后在经过分句的答复意见中依次进行模糊匹配，并记录下相似度。根据关键值和相似度，确定最终的完整性指标。

流程图如图 22 所示：

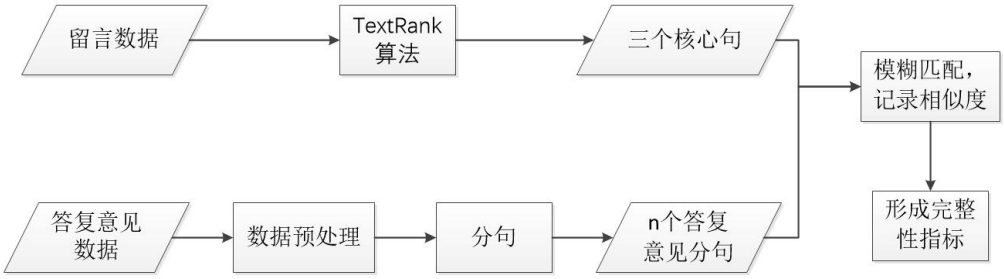

图 22 完整性指标形成流程图

4.3.2 对留言进行的处理

通过对文本的统计与分析，得到留言中所提出的问题大多为 1 个，最多为 3 个，因此我们设置提取关键句的数量为 3。

TextRank 算法简便易用，在 python 中，TextRank 算法可使用 textrank4zh 库进行方便的实现与使用。在此使用该算法找出具有代表性的 3 个关键句，并记录关键值。关键值代表某句话在文本中的重要程度。

图 23 是留言原文与经过筛选后的三条关键句的对比，可以发现经提取后 3 个关键句可以代表留言原文的主题和内容。

```
'\n\t\t\t\t\t\n\t\t\t\t\t早段时间看新闻说A市要开通50条社区巴士，我看到公众号的文章推送
，然后进行投票，我们怡海星城有过强烈建议设一条线路方便一下我们进城坐公共交通工具，
避免我们更多私家车队对城中心拥堵。但是没有听到动静。看了一下A市brt2014的规划，感觉
A市发展近几年发展太快了，城市扩张很快，公共交通地铁发展也迅速。我们近郊居民众多，所
以线路也很难完善。在brt公交系统，没有最终规划好的前提之下，提两点建议。是否可以brt
原来定在中信新城的终点站延伸到怡海新城或者途径在修的高云路通向动物园。万家丽路上面
的brt，原来定在广电中心设终点站，是否可以南延到五矿哈佛小镇，末站到动物园。\n\t\t\t
\t\t\n\t\t\t\t\t\n\t\t\t\t'

['早段时间看新闻说A市要开通50条社区巴士，我看到公众号的文章推送，然后进行投票，我们
怡海星城有过强烈建议设一条线路方便一下我们进城坐公共交通工具，避免我们更多私家车队
对城中心拥堵',
 '万家丽路上面的brt，原来定在广电中心设终点站，是否可以南延到五矿哈佛小镇，末站到动
物园',
 '是否可以brt原来定在中信新城的终点站延伸到怡海新城或者途径在修的高云路通向动物园']
```

图 23 关键句提取示例图

4.3.3对答复意见数据进行的处理

1. 数据预处理

答复意见中包含一些与答复主题不相关的句子，例如：“网友……，您好您的留言已收悉，现将有关情况回复如下……。”、“感谢您对……的理解与支持”。如果不对这些句子进行预处理，在模糊匹配阶段将对结果产生很大的影响。

下图分别展示了未经过预处理的模糊匹配结果，与经过预处理的模糊匹配结果。可以看出，经过预处理的答复意见，在模糊匹配阶段具有更高的准确性。

```
In[13]: process.extractOne(sen[66][0], data1[66][0])
Out[13]: ('感谢您对我们工作的支持、理解与监督！', 22)
In[14]: process.extractOne(sen[66][1], data1[66][0])
Out[14]: ('您的留言已收悉。', 13)
In[15]: process.extractOne(sen[66][2], data1[66][0])
Out[15]: ('您的留言已收悉。', 13)
```

图 24 未经预处理的模糊匹配结果图

```
In[12]: process.extractOne(sen[66][0], data1[66][0])
Out[12]: ('您的留言已现将有关情况回复如下：对于您的意见和建议，市城乡规划局将充分论证研究。', 12)
In[13]: process.extractOne(sen[66][1], data1[66][0])
Out[13]: ('您的留言已现将有关情况回复如下：对于您的意见和建议，市城乡规划局将充分论证研究。', 7)
In[14]: process.extractOne(sen[66][2], data1[66][0])
Out[14]: ('您的留言已现将有关情况回复如下：对于您的意见和建议，市城乡规划局将充分论证研究。', 5)
```

图 25 经过预处理的模糊匹配结果图

在此，对无关句子利用正则表达式筛选的方式，进行去除与替换。

预处理前的答复意见文本：

现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于2019年4月10日至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5月5日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019年5月9日

图 26 预处理前文本示例图

经过预处理的答复意见文本：

['，首先关于您在平台栏目给胡华衡书记留言，反映“A2区景蓉花苑物业管理有问题”的情况已现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5月5日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。']

图 27 预处理后文本示例图

2. 中文分句

对于简单的中文分句，只需要找到“？”、“！”、“。”等典型的断句符断开即可，但这类分句具有一定的局限性。现将分句模型进行调整与优化，确保分句的准确性。

经过分析与对比，分局任务分为以下三种情况。在优化模型中，对这三种情况分别进行处理。

情况一，单字符断句符：包括“？”，“！”，“。”，“””，“’”。

情况二，多字符断句符：包括英文省略号，中文省略号。

情况三，双引号：如果双引号前有单字符断句符，那么双引号才是句子的终点。

分句结果示例如下：

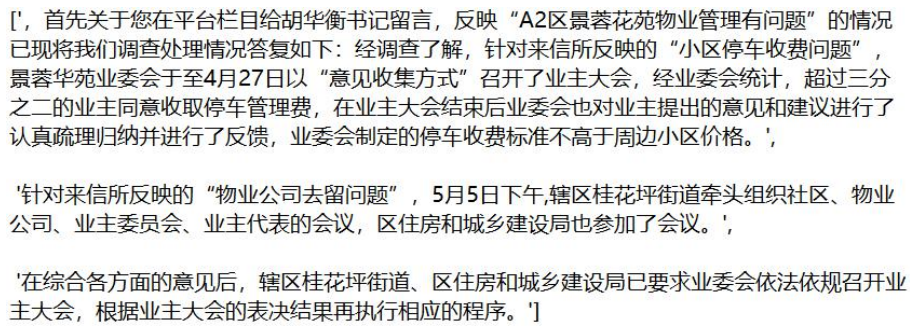

图 28 分句结果图

4.3.4 模糊匹配

上一阶段挑选出的三个核心句，能够充分代表留言的主题，包含了留言所提出的多个问题。经过分句的答复意见也更便于模糊匹配。

将第 i 条留言的三个核心句，依次与第 i 条答复意见的多个分句进行模糊匹配，针对每个核心句，筛选出相似度最高的一个答复意见分句，记录相似度。

在此使用基于编辑距离算法的 FuzzyWuzzy 工具包实现模糊匹配。编辑距离算法具有准确率高、扩展方便等优点，与其他相似度算法具有优越性。

4.3.5 完整性指标建立

第 i 条留言的三个核心句分别对应一个相似度最高的答复意见分句，并记录有百分数相似度S_1、S_2、S_3。将三条核心句对应的关键值进行归一化处理，用作三个相似度的权值。

计算S_1、S_2、S_3的加权和 S，S 即为该答复意见的完整性指标。

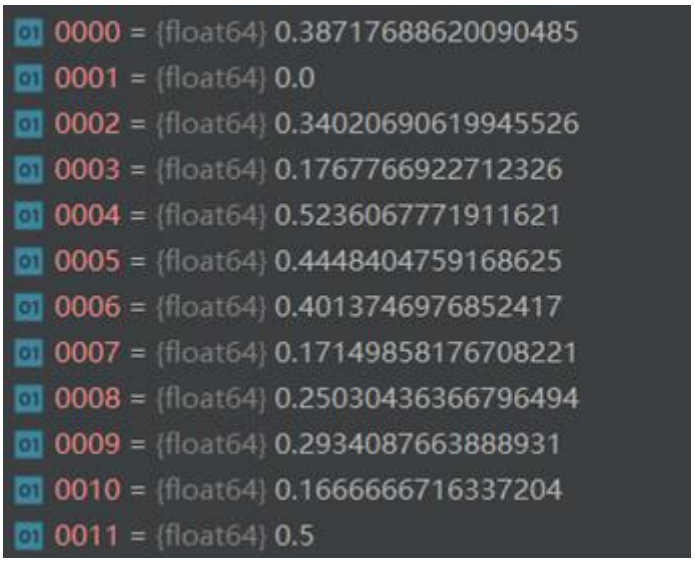

图 29 完整性指标计算结果图

4.4 及时性指标

及时性指标的形成，需要考虑留言时间与回复时间两者的时间差。

4.4.1 数据预处理

原始数据，提供了详细的留言时间与答复时间，但数据格式有较大不同。大多数数据为 str 格式，部分数据为 datetime 格式；多数时间详尽到秒，少数时间详尽到分钟或详尽到日。在此，我们将数据统一处理为 datetime 格式，方便下一步的处理与计算。

下图为经过预处理后的时间数据：

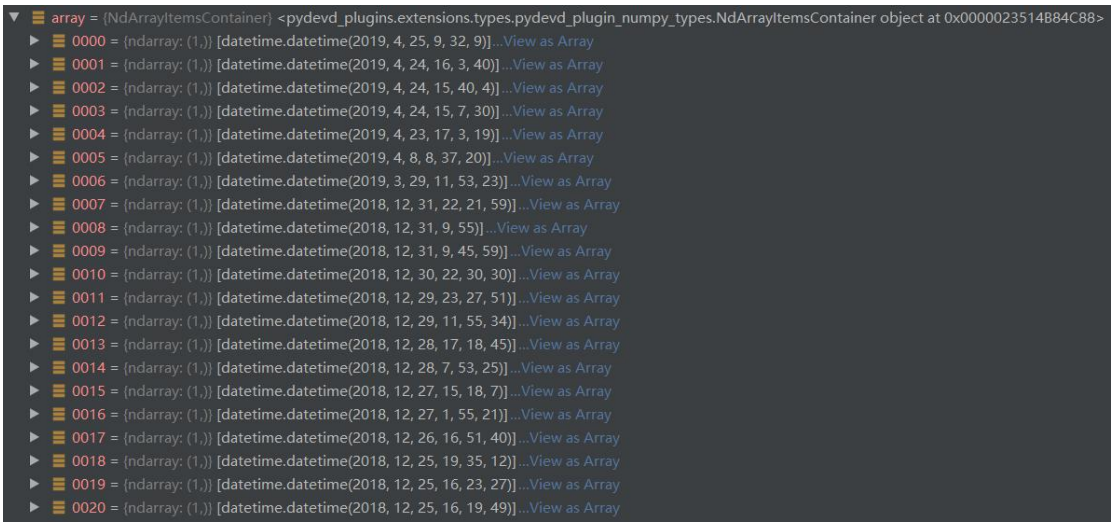

图 30 经过预处理的时间数据图

处理后的数据格式规范统一，可以方便地计算时间差。

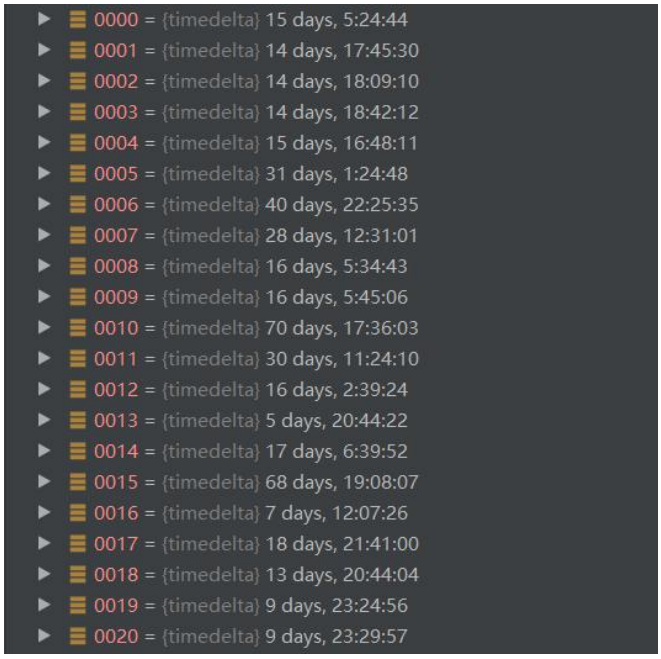

图 31 时间差计算结果图

4.4.2 及时性指标形成

根据答复时间差对答复意见的及时性进行评分，接下来确定评分标准。

留言所提出的问题具有时效性，答复时间过晚的留言不能为市民提供帮助；考虑到答复意见的形成需要调查情况或查阅文件，将评分标准定得太高也不切实际。因此要经过分析与对比，确定合理的指标评价方案。

首先对答复时间差进行分析。经过统计，总的答复意见数量为 2816 条，答复时间差大于一个月（即 30 天）的答复数量为 461 条。由于答复意见的时效性，

答复时间差过长，已经无法提供实质性的意见与帮助。因此，将答复时间差大于一个月的答复意见设定为最低评分。

为保证分级的科学性，应使每个等级的答复意见数量大致相等，因此确定分级细则如下：

表 13 及时性评分表

T/天	及时性评分
$T < 3$	1
$3 \leq T < 7$	0.8
$7 \leq T < 15$	0.6
$15 \leq T < 30$	0.4
$T \geq 30$	0.2

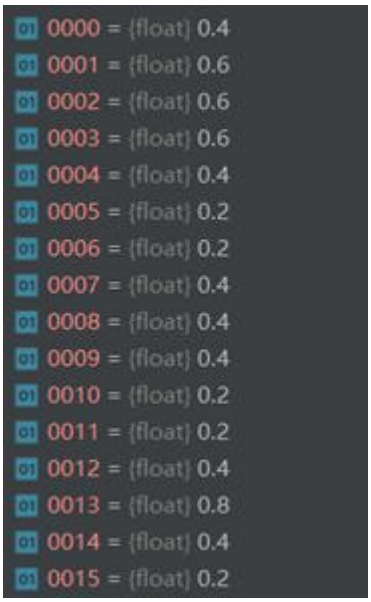

图 32 及时性指标计算结果图

4.5 可解释性指标

从可读性、规范性、可信度三方面综合考虑，首先形成三方面的指标，而后建立模型，确定每个指标的权重，形成可解释性指标。

4.5.1 可读性

为形成可读性指标，应考虑答复意见的行文逻辑，判断答复意见条理是否清晰。

在文本中使用编号与分点叙述，可使文本层次分明、结构性更强，对读者而言条理也更加清晰，可读性也更好。因此，我们通过判断答复意见中是否含有“1、

2、3”或“一、二、三”等特征标号，来确定答复意见的可读性。

正则表达式可以快速、方便地对文本进行匹配，考虑到任务并不复杂，在此使用正则表达式对答复意见进行匹配。

若存在特征标号，则该答复意见可读性好；若不存在，则可读性差。并按照此标准对可读性进行赋分。如下表所示：

表 14 可读性评分表

可读性	得分
可读性好	1
可读性差	0

可读性处理结果：

图 33 可读性指标计算结果图

4.5.2 规范性

经过分析，示例中答复一般分为三部分：

a) 问题重述（网友“UU0082211” 您好！您的留言已收悉。现回复如下：）

b) 解决方案

c) 致谢（感谢您对我们工作的支持、理解与监督！）。

我们将具有以上三个部分的答复意见称为一个完整的、具有规范性的标准答复。通过判断以上三个部分，在答复意见中是否都存在，来形成规范性指标。

具体细则如下：

表 15 规范性评分表

答复意见中含有标准答复三部分的个数	得分
3	3
2	2
1	1
0	0

规范性处理结果示例：

图 34 规范性指标计算结果图

4.5.3 可信度

一个可信的答复意见，需要经过核实调查，需要查阅相关规定。经过核实和调查的答复，意见中会含有“经……部门核实”，“经调查”等特征句子。查阅相关规定得到的答复意见中，含有“根据《……》规定”特征句子。

对以上两种句子使用正则表达式进行特征比对，同时经过核实调查并查阅相关规定的答复意见，可信度得分为“2”；经过核实调查或查阅相关规定的答复意见，可信度指标为“1”；未经过核实调查与查阅相关规定的答复意见，可信度指标为“0”。

可信度处理结果示意：

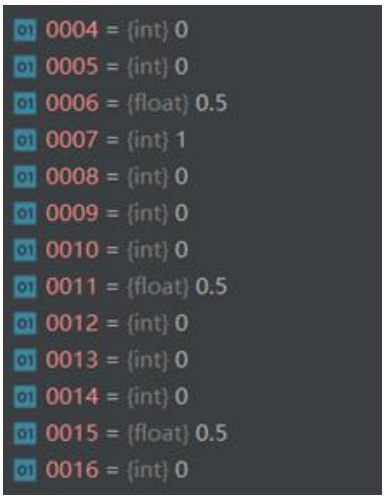

图 35 可信度指标计算结果图

4.5.4 可解释性指标建立

答复意见的最终目标是对市民提出的问题进行回复与解答。在以上三个特性中，可读性与规范性影响市民的阅读效率与阅读感受，并不直接影响答复内容的实际质量；而可信度反映了答复意见的真实性，起着关键的作用。

基于以上标准，我们确定可读性与规范性的权重大致相等，且均应小于可信度指标的权重。在这里，我们取三个指标的权重比为 0.25：0.25：0.5。

对这三个指标赋权值相加，并进行归一化处理，最终得到可解释性指标。

可解释性评分示例：

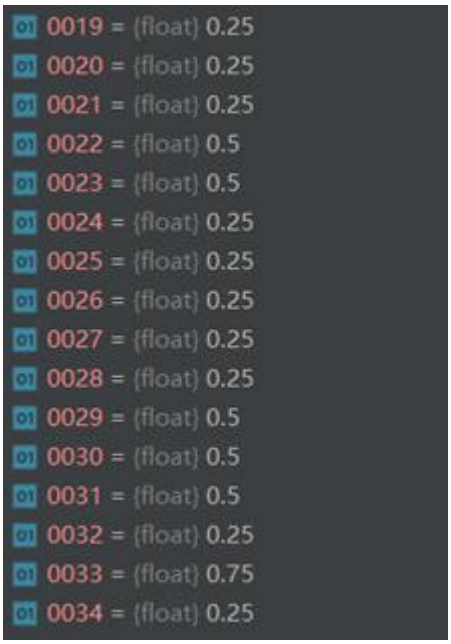

图 36 可解释性指标计算结果图

4.6 完整的评价方案

答复意见最重要的目的是回答市民的疑惑，解决市民的问题。答复是否反映了真实情况、是否具备解决问题的能力，应是评价答复质量的重中之重。

及时性反映了相关部门对市民留言的重视程度，以及查证问题的能力与效率。留言具有时效性，高质量的答复意见应具备良好的及时性。但真正能解决问题的答复意见，通常需要更多的时间去调查与证实。因此，及时性指标所占权重不应过高。

可解释性主要影响答复的逻辑性和阅读体验，对答复解决问题的能力影响较小，权重也不应过高。

相关性和完整性共同体现了答复意见对留言的回答情况，这两方面的指标权重占比应大一些。

根据以上讨论，相关性与完整性指标的权重大致相等，且高于另外两个特性，及时性次之，可解释性指标权重最小。

在此，取四者权重比如下：

表 16 答复质量权重表

指标名称	权重比
相关性	0.35
完整性	0.35
及时性	0.2
可解释性	0.1

将赋权后的四个指标赋权相加，即可得到总的答复意见质量评分，该评分可以较为全面地反映相关部门对某条留言的重视程度以及答复质量。

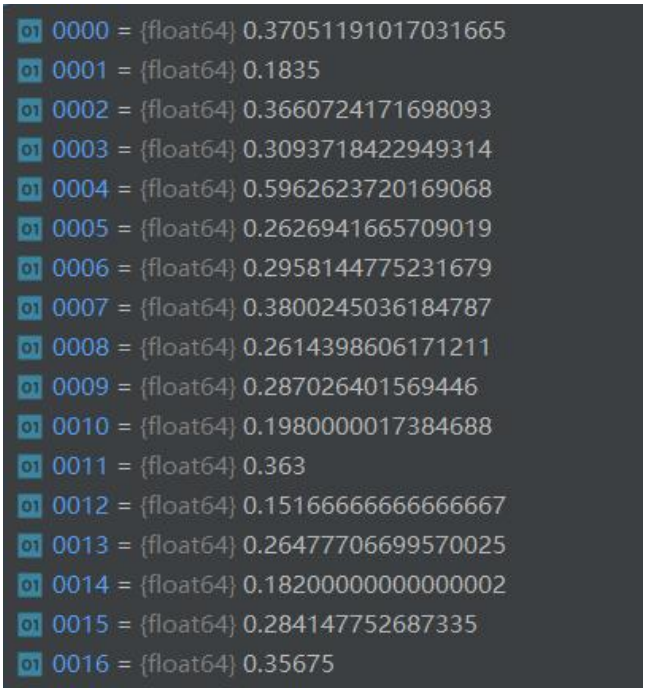

图 37 答复意见质量评分计算结果图

经过查找，附件数据中评分最低的答复意见分值仅为 0.04，详情如下：

表 17 答复详情表

留言主题	留言时间	答复意见	答复时间
关于请求开通暮云段 A1 区南路路灯及地下通道的请求	2014/4/21 1:40:59	“UU0081182”	2014/5/28 15:11:52

表 18 答复指标详情表

及时性指标	完整性指标	相关性指标	可解释性指标
0.2	0.0	0.0	0.0

可以得到，在该条答复意见回复时间差大于一个月，回复内容不完整，逻辑性差，与本评价方案得到的结果一致。

评分最高的答复意见分值为 0.685，详情如下，在此对答复意见详情内容：

表 19 答复详情表

留言主题	留言时间	答复意见	答复时间
咨询 D7 县残疾人扶助的有关政策	2018/12/9 13:04:14	UU0081788:您在……栏目中留言咨询……一事，D7 县残联十分重视，现就您反映的相关政策情况回复如下：1、……2、按省民政厅文件要求，……3、……以上回复，若有不明之处，欢迎到 D7 县残联具体咨询核实。D7 县残联联系电话 0734-0000-00000000 年 12 月 13 日	2018/12/13 16:53:20

表 20 答复指标详情表

及时性指标	完整性指标	相关性指标	可解释性指标
0.8	0.86	0.485	0.65

该条答复意见回复时间较快，回复内容完整，逻辑性强，与评价方案得到的结果一致。

五、参考文献

[1] 王宇．短文本聚类簇描述及标签生成方法，信息系统协会中国分会第六届学术年会，2015．

[2] Houtman E, Makos A, Meacock H L. The Intersection of Social Presence and Impression Management in Online Learning Environ-ments [J].E-Learning and Digital Media,2014,11(4):419-430.

[3] Yoon Kim. Convolutional Neural Networks for Sentence Classifification, 2014.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation

of Word Representations in Vector Space,2013.

[5] Tomas Mikolov. Bag of Tricks for Efficient Text Classification, 2017