# 基于文本挖掘的智慧政务系统

## 摘 要

近年来，随着电子技术、信息技术和网络技术的发展，给政府部门了解民意、汇聚民智、凝聚民气为核心目标的智慧政务带来了发展的核心技术和推动力量。积极实施建立智慧政务系统在为公众提供信息化高效服务、提高政府工作效率、精准制定政府政策、透明化政务信息等方面具有重要的意义。本文主要内容如下：

针对问题 1，首先读取文件内容，取出“留言详情”列的数据进行数据预处理：依次进行文本去重、文本去噪操作。采用 textrank 算法对文本数据提取关键句。利用 Python 中的 jieba 库对数据清洗后的数据进行分词处理,导入停用词表去除停用词。之后采用 gensim 库下的 word2vec 模块训练词向量，将其作为 LSTM 模型的词嵌入层。最后，搭建 LSTM 模型对文本进行分类，采用 F1 分数对模型精度进行评估。

针对问题 2，首先对数据进行基于层次的文本聚类，对聚类好的分组提取关键词形成话题。然后建立评价话题热度的指标。我们基于话题持续时间，话题点赞和反对总数，话题总文本条数三个维度进行话题热度的评价。然后采用熵权法划分指标权重，并计算出热度，最后对话题的热度排序，得出排名前五的话题，生成“热点问题表”及“热点问题留言明细表”。

针对问题 3，我们建立了一套政务评价系统，对政府答复意见的及时性、相关性、规范性、可解释性四个方面进行评估，并按照不同权重赋予分数，最后综合各项指标对政府的整体政务水平进行评估。

**关键词:LSTM 神经网络,Word2vec 词向量,BRICH 层次聚类,熵权法,TextRank 算法**

# Intelligent government affairs system based on text mining

## Abstract

In recent years, with the development of electronic technology, information technology and network technology, it has brought the core technology and driving force for the development of intelligent government affairs with the goal of understanding public opinion, gathering people's wisdom and prople's spirit.

For question one, first read the contents of the file, take out the data in the column of "message details" for data preprocessing: carry out text de-duplication and text de-noising successively. Apply the TextRank algorithm to extract key sentences from text data. The jieba library in Python is used to segment the data after data cleaning and import the stop word list to remove stop words. After that, word2vec module in genism library is used to train the word vector, which is used as the word embedding layer of LSTM model. Finally, the LSTM model was built to classify the text, and F1 score was used to evaluate the accuracy of the model

For question two, firstly, hierarchical text clustering was conducted on the data, and keywords were extracted from the clusters to form topics. Then build an indicator to evaluate the topic's popularity. We evaluated the topic popularity based on three dimensions: the duration of the topic, the total number of thumb up and the total number of text bars. Next, the entropy weight method is used to divide the index weight, adnd the heat is calculated. Finally, the heat of the topic is sorted to get the top five topics, and the "hot issues table" and "hot issues message list" are generated.

For question three, we have established a set of government affairs evaluation system, which evalues the timelines, relevance, standardization and interpretability of the government's replies, and evaluates the overall level of government affairs according to different weights.

**Key words: LSTM neural network, Word2vec word vector, BRICH hierarchical clustering, entropy weight method, TextRank algorithm**

# 目 录

# 图表目录

# 1 挖掘目标

## 1.1 挖掘背景

近些年来，随着网络技术的高速发展和数据库系统的广泛应用，催生了人工智能、移动互联网、大数据等领域的高速发展。智慧政务是政府使用互联网技术，利用大数据手段推进智慧化城市建设的示范性举措。如今，在“互联网+政务服务”的不断推进下，全国各级政府加快推进政府服务全面转型。未来，应用“互联网+政务服务”模式将成为一大趋势。

其中，微信、微博和各种网站是人民大众每天都离不开的工具。各级政府利用互联网中微信、微博、市长信箱、阳光热线等建立起了了解民意、汇聚民智、凝聚民气的政务平台，越来越多网民通过发表帖子，微博评论，微信公众号发表自己的意见和投诉身边的问题。如果能及时处理热点问题对于树立政府形象、舆情分析、把握舆论走向等具有重要意义。然而，传统依靠人工进行处理的方式在面对爆炸式的信息时显得十分困难，因此运用数据挖掘手段处理这些平台的信息也因此成为了政府工作中的重要内容。基于文本挖掘的智慧政务系统能够很好地协调完成政府各项工作，大大提高施政效率和管理水平。

此题需要对群众在政府网络问政平台的留言建立分类模型并挖掘出群众集中反映的热点问题，最后针对相关部门对群众留言的答复意见建立一套评价方案并尝试实现。

随着“互联网+政务系统”逐渐普及，建立留言分类模型可以解决以往依靠人为经验分类的大量人力问题，从而把有限的人力资源集中到对群众留言尤其是其中反映的热点问题的答复上。对答复意见进行评价也可以更好地评判政府工作的效率和群众满意度。所以，建立留言分类，热点问题提取和答复评价为一体的智慧政务系统有着极为重要的意义。

## 1.2 挖掘目标

根据附件 2 里的数据提取，进行数据处理后，对所有留言建立关于留言内容的一级标签分类模型。然后通过网络爬虫获取更多数据对模型进行数据增强。最后使用 F-Score 对分类方法进行评价。

根据附件 3 的数据，将某一时段内反映特定地点或特定人群问题的留言进行

归类，定义合理的热度评价指标，并给出评价结果。建立模型对留言反映的问题进行热度排序，找出排名前 5 的热点问题。

根据附件 4 的数据，从答复意见的相关性、完整性、可解释性等角度对答复意见的质量建立起一个评价模型。

# 2 符号说明

## 2.1 问题一符号说明

| 符号 | 意义 |
|---|---|
| $Q$ | 查询语句 |
| $q_i$ | 关键字 |
| $D$ | 被检索的文档 |
| $S_i \quad S_j$ | 文本中任意两个句子 |
| $w_k$ | 当前句子中的词 |
| $w_t$ | 语料词典中的一个词 |
| $F1-Score$ | 对 LSTM 模型评估分数 |
| $precision$ | 评估的精确率 |
| $\mathrm{Re}call$ | 评估的召回率 |
| $Loss$ | 评估的丢失率 |

表格 1 问题一符号说明

## 2.2 问题二符号说明

| 符号 | 意义 |
|---|---|
| $t$ | 词条 |
| $d$ | 文档 |
| $df_t$ | 包含词条$t$的文档数量 |
| $x$ | 点赞数与反对数之差 |
| $t_s$ | 发布时间到目前的时间间隔 |
| $z$ | $x$和 1 的绝对值的最大值 |
| $y$ | 当$x$>0 取 1；$x$<0 取-1；$x$=0 取 0 |
| $x_{ij}$ | 矩阵第 i 行第 j 列的元素 |
| $P_{ij}$ | 第 j 项指标下第 i 个记录所占比重 |
| $e_j$ | 第 j 项熵值 |
| $g_j$ | 第 j 项差异系数 |
| $w_j$ | 第 j 项指标的权重 |

表格 2 问题二符号说明

# 3 问题一分析

## 3.1 问题一流程图

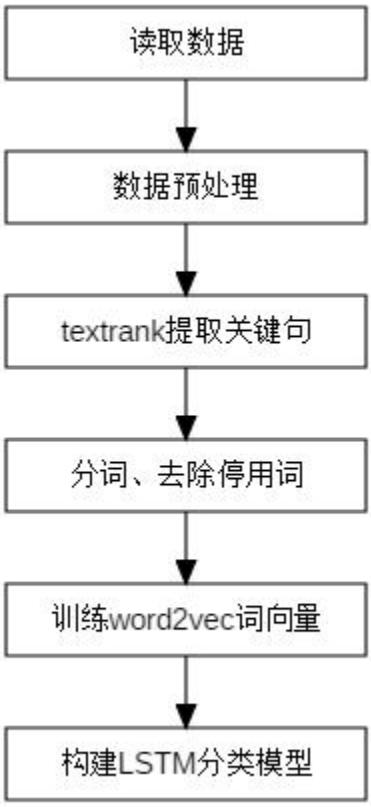


图 1 问题一流程图

## 3.2 数据读取

读取附件 2 的所给数据，使用“留言详情”和“一级标签”的数据进行分类。

## 3.3 数据预处理

### 3.3.1 文本去重

在题目所给附件 2 的数据中，出现了重复的留言内容。考虑到这些重复的留言内容只需回复其中之一即可，影响了工作量和工作效率，增加计算机工作量，因此用 Python 中 drop_duplicates 函数对数据进行去重处理，只保留其中一条数据。

### 3.3.2 文本去噪

由于留言内容中除中文内容外还有很多标点符号等非汉字内容，这些内容对于内容分类工作没有帮助，所以用正则表达式将换行符和空格替换成空字符串，即去除了文本中的非中文字符。

## 3.4 利用 TextRank 算法提取关键句

### 3.4.1 TextRank 算法介绍

在进行了以上操作后，为了减少后续工作的工作量，可以把现有文本数据中的关键句提取出来，以供后续分词处理。这里我们采用 TextRank 算法将文本中关键句提取出来。而 TextRank 算法则利用了 BM25 算法可以将一个句子视作查询语句，相邻的句子视作待查询的文档，就能得到它们之间的相似度的特点。其中，BM25 算法是 TF-IDF 的一种改进变种。TF-IDF 衡量的是单个词语在文档中的重要程度，而如何衡量多个词语和文档的关系[1]，正是 BM25 所解决的问题。BM25 的度量如下：

$$BM25(D,Q)=\sum_{i=1}^{n}IDF(q_i)\bullet\frac{TF(q_i,D)\bullet(k_1+1)}{TF(q_i,D)+k_1\bullet(1-b+b\bullet\frac{|D|}{avgDL})} \tag{3-1}$$

其中形式化的定义$Q$为查询语句，由关键字$q_1$到$q_n$组成，$D$为一个被检索的文档。

### 3.4.2 TextRank 算法实现

TextRank 算法的实现这里使用 pyhanlp 库，pyhanlp 是 Python 包装了 HanLP 的 java 接口。关于 TextRank 算法的具体原理如下：

第一步，将文本中每个句子分别看作一个节点。

第二步，如果有两个句子有相似性，则认为这两个句子对应的节点之间存在一条无向有权边，衡量句子相似性的公式如下：

$$Similarity(S_i,S_j)=\frac{|w_k|\{w_k\in S_i\cap w_k\in S_j\}}{\log(|S_i|)+\log(|S_j|)} \tag{3-2}$$

式中，$S_i$、$S_j$：两个句子，$w_k$：句子中的词。分子部分的意思是同时出现在两个句子中的同一个词的数量，分母是对句子中词的个数求对数后求和，这样设计可以遏制较长的句子在相似度计算上的优势。

第三步，根据以上相似度计算公式循环计算任意两个节点之间的相似度，设置阈值去掉两个节点之间相似度较低的边连接，构建出节点连接图。

第四步，迭代计算每个节点的 TextRank 值，排序后选出 TextRank 值最高的几个节点对应的句子作为关键句。

## 3.5 利用 jieba 进行中文分词

在对留言信息进行挖掘分析前，需要把已经提取的关键句这种非结构化的文本信息转换为计算机能够识别的结构化信息。在所给的数据中，都是以中文文本的方式给出数据，为了便于后续的转换，因此在此步对这些关键句进行中文分词。这里我们采用 Python 的中文分词包 jieba 进行分词。jieba 支持三种分词模式，采用了基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）。同时它采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，并且使用了 Viterbi 算法，使得其能充分实现良好的分词效果。部分分词后的结果如下图所示：

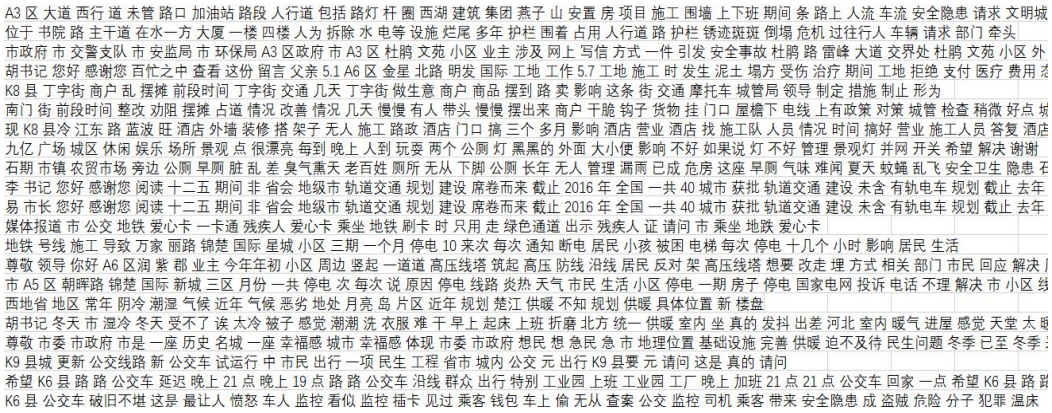


图 2 部分分词结果示意图

| | 留言详情 | 一级标签 |
|---|---|---|
| 0 | [整改, 文明, 路段, 安全隐患] | 城乡建设 |
| 1 | [护栏, 锈迹斑斑, 位于, 书院, 路, 主干道, 在水一方, 大厦, 一楼, 四楼, 人... | 城乡建设 |
| 2 | [程明, 物业, 小区, 停车, 收费, 标准, 上调, 小区, 物业, 市程明, 物业管理... | 城乡建设 |
| 3 | [健康, 生活, 环境, 请, 环保部门, 检测, 水是, 日常生活, 必不可少, 用品] | 城乡建设 |
| 4 | [2015, 年, 购买, 盛世, 耀凯, 小区, 17, 栋, 楼, 想, 问一问, 政府... | 城乡建设 |
| 5 | [不知, 规划, 供暖, 具体位置, 新, 楼盘] | 城乡建设 |
| 6 | [停水, 影响, 小区, 居民, 日常生活, 相关, 单位, 承诺, 给与, 解决, 恳请,... | 城乡建设 |
| 7 | [配合, 代收, 城市, 垃圾处理, 费, 小区, 对此, 拒交, 城市, 垃圾处理, 费,... | 城乡建设 |
| 8 | [小区, 绿化带, 破烂不堪, 小区, 绿化带, 垃圾场, 地方, 停车] | 城乡建设 |
| 9 | [滨江, 社区, 居委会, 收缴, 泰华, 一村, 小区, 第二届, 业主, 委员, 资料,... | 城乡建设 |

图 3 分词后结果整理

## 3.6 使用 word2vec 训练词向量

在对文本数据进行了预处理后，得到他们的语料库后，将它们的 one-hot 向量作为 word2vec 的输入，通过 word2vec 来训练词向量。word2vec 是 Geogle 在 2013 年推出的一款用于训练词向量的工具，word2vec 提出了一种用用分布式向量对文本进行表示的方法。与传统文本向量空间模型相比，使用 word2vec 模型来表示文本，既能解决传统向量空间模型的高维稀疏特征问题，还能引入传统模型不具有的语义特征，有助于短文本分类。word2vec 目前有两种训练模型：CBOW 模型

和 skip-gram 模型。CBOW 模型根据中心词周围的词来预测中心词，skip-gram 模型则根据中心词来预测周围的词。CBOW 算法对于很多分布式信息进行了平滑处理（例如将一整段上下文信息视为一个单一观察量）。很多情况下，对于小型的数据集，这一处理是有帮助的。[2]相比之下，skip-gram 模型将每个“上下文-目标词汇”的组合视为一个新观察量，因为这种做法语义准确率高,在大型数据集中会更为有效，故此题采用 skip-gram 模型。skip-gram 模型的数学表达式如下：

$$P(W_{t-k},W_{t-k+1},\cdots,W_{t+k-1},W_{t+k}|W_t) \tag{3-3}$$

其中 $W_t$ 为语料词典中的一个词，skip-gram 即通过词汇 $W_t$ 去预测相邻窗口 k 内词汇的概率。skip-gram 模型示意图如下：

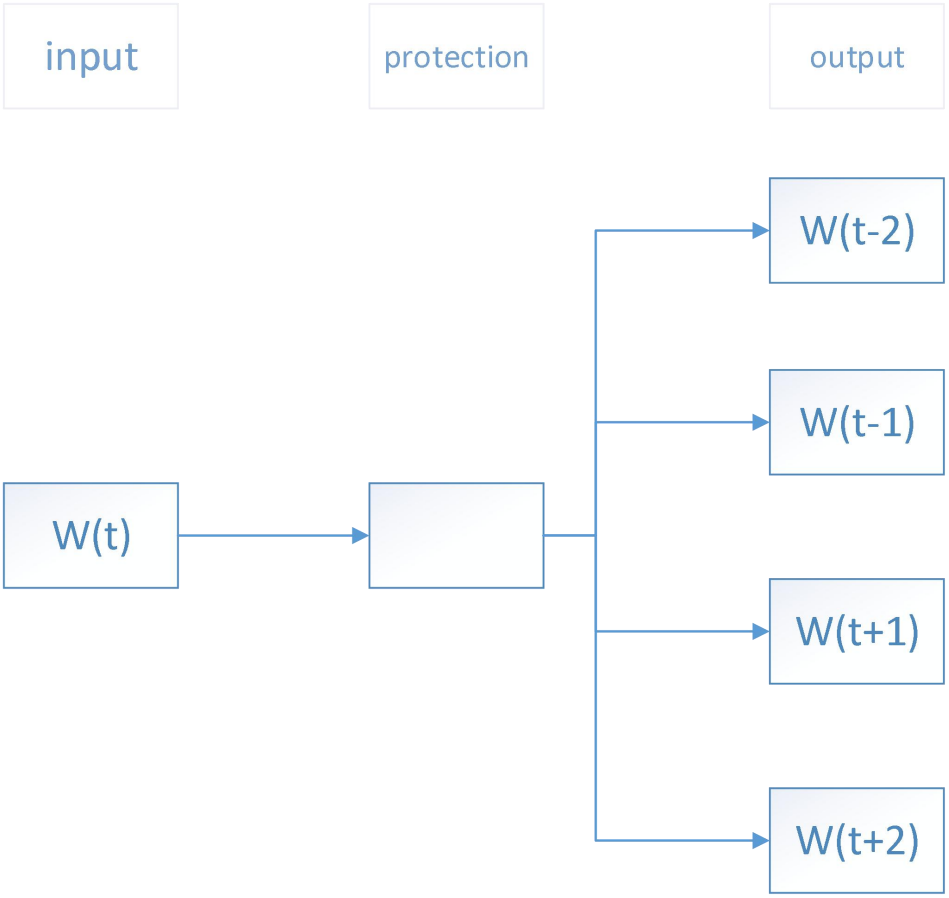


图 4skip-gram 模型

利用 Word2Vec 模型进行特征提取等最终将不同性质的词语归为一类并用高维的实型向量表示，构建的文本向量作为 LSTM 的初始输入。训练词向量后的部分数据如下图所示：

```
24834 200
市 0.0011541994 -0.50866365 -0.26532653 -0.25757673 0.17745684 -0.16984075 -0.20633756 -0.15888956 0.61311567 -0.30635977 0.20307061 0.23127571
-0.39754876 -0.30665025 0.062132787 -0.042427175 0.5193976 0.04040693 0.0108731445 -0.4582008 -0.06986257 -0.20314966 0.31461307 0.14810985
-0.3184278 0.047898203 -0.59545696 0.008922206 0.13651563 -0.15589295 0.26463622 0.25995153 -0.03754552 0.2383991 -0.24137367 0.047397885
0.081047334 0.04333339 -0.065006904 0.0065795267 -0.39554033 -0.063826956 0.5536419 -0.22746511 -0.2129853 0.1464262 -0.48670262 -0.115735
-0.363573 0.31545433 -0.08220515 -0.009169614 -0.1345496 -0.22969545 -0.30200884 0.017502856 0.19949456 0.0864428 -0.13004662 -0.071205474
-0.032084174 -0.26895618 -0.35362616 -0.1133779 -0.52752376 -0.1962252 -0.2737664 0.08431288 -0.3110843 -0.35768318 0.08987246 0.22265457
-0.04292654 0.2952187 0.23977187 0.2725057 0.25800166 0.26949248 -0.1487385 -0.017031047 -0.106325515 0.068614766 0.21526814 0.21726789 0.41021457
0.19122775 -0.05750315 0.16152233 -0.3223813 0.28259638 -0.33062595 -0.21271077 0.22001149 -0.5526863 0.058931954 0.17504622 -0.24571279 0.05204788
0.33514893 0.0071580913 -0.22795106 -0.20955332 0.051871818 0.23880215 0.18453549 -0.1309859 0.17567903 -0.41758567 0.38226855 0.30139965
0.11659394 0.18937705 -0.20513672 -0.0015067204 -0.08785473 -0.5844676 0.02378169 -0.19766283 0.36377802 0.07804472 0.28529468 0.065338545
0.59441054 0.30399185 0.055048097 -0.09249787 0.042541735 0.14266695 0.07510554 0.3752263 -0.18030412 0.3371574 -0.18126641 -0.18702042 0.037841998
-0.08469228 -0.21463375 -0.070177354 0.18180963 -0.4365021 -0.16585901 0.021690296 -0.084550545 0.20585787 0.19722109 -0.23248583 0.23550707
-0.33855167 0.021727653 0.37650397 -0.004195833 -0.21467614 0.11679957 0.43772873 -0.06344589 0.23602934 -0.026962114 0.24988186 0.26216304
-0.14946845 -0.07419161 -0.0109963445 0.12849176 -0.21779981 0.030820183 0.22328939 -0.11340799 0.02812851 0.27715957 0.35259917 0.19217709
0.20895788 -0.47781816 -0.057086386 -0.0052266596 -0.22578488 -0.21461853 0.017470153 0.01629018 -0.083549194 -0.12147934 0.11898941 0.17881198
0.19307889 -0.43627304 -0.018041192 0.07659592 0.3133433 -0.16956806 -0.2830825 0.23208547 -0.10295632 -0.34242517 0.012809032 0.05708022
-0.17654106 0.1594821 -0.26500842 -0.12573257 -0.1320404
县 0.116441995 -0.42771325 -0.1485908 -0.35913232 0.079642005 -0.24836843 -0.06608887 -0.27563918 0.5349626 0.054074343 0.22922441 0.28541243
0.13225628 -0.16984545 0.3799469 0.27410683 0.4915931 -0.00828232 0.026309285 0.07611979 0.43218353 -0.17690513 0.039082654 0.13325837 -0.47681195
-0.1331957 -0.42771184 0.08020639 0.18895786 -0.09807185 0.29221877 -0.06468131 0.10822363 0.20400938 -0.01999881 -0.05419126 0.0989445 0.067404225
-0.13008764 -0.14394067 -0.043946285 -0.16393064 0.19942322 -0.24597292 -0.04837835 0.2646962 -0.18314078 -0.31143755 -0.16110533 0.36052042
0.017482525 -0.10371117 0.047845375 0.086950004 -0.22582403 0.110175245 0.30273128 0.29993924 -0.245186 0.06307859 0.08840933 -0.050518468
-0.37634313 -0.14286956 -0.4075256 -0.08468685 -0.08031114 0.2748501 -0.34980568 -0.19317918 0.25355923 0.12149552 -0.085145056 0.1336375
0.35937333 0.46719912 0.39204034 0.23779032 0.12783776 0.056640204 -0.1281562 0.22968043 0.23405637 -0.026364204 0.06597225 0.21260339 0.19108258
0.47945663 -0.27927864 -0.040114228 -0.25995135 0.10289565 0.1545016 -0.34368628 0.20394535 0.05523166 0.09458025 -0.09251359 0.35906184
-0.005493915 -0.06069829 -0.05745646 0.012198964 0.2164975 -0.039948534 -0.3090834 0.25065175 -0.12754299 -0.036459047 0.08322037 0.18918285
0.23860848 0.04695689 0.2625955 0.11240388 -0.43227944 0.021735318 -0.064030446 0.16218473 -0.095860325 -0.028473234 0.053050116 0.32162815
0.034963768 0.16392191 0.09725316 -0.051549986 0.09048322 0.15049203 0.27741125 -0.115899116 0.039027378 -0.19902721 0.096514426 -0.003912111
-0.115338065 -0.23320258 0.14123869 0.23147145 -0.1870253 -0.22501339 -0.02881617 0.059097808 0.06482844 0.10344486 0.16724339 -0.030563986
-0.2981212 0.2540617 0.28060314 -0.033025097 -0.2546813 -0.025097357 0.081922226 -0.29844862 -0.25867558 -0.09632231 0.27226976 0.43700072
0.014450556 0.12853241 0.38458613 0.1898532 0.12123692 0.006832846 0.32476464 -0.12748885 -0.031012516 0.43947348 0.19327651 0.096440054 0.44856465
-0.0891027 0.058507215 0.16628562 -0.16409154 -0.060030956 0.042949047 0.052148074 0.1563007 0.20452417 0.20372427 0.4792601 -0.07937238
-0.15008931 -0.09829009 0.012925974 0.18205255 -0.08313745 -0.15640867 -0.06925336 -0.4336662 0.1113943 0.08296852 -0.16151743 -0.06962269
-0.1641773 0.13238269 0.037999023 0.2378688
```

图 5 部分词向量

# 3.7 使用 LSTM 模型对文本数据进行分类

## 3.7.1 LSTM 模型介绍

利用基于 word2vec 方法得到的词向量训练集来训练 LSTM 神经网络。LSTM 神经网络中文名是长短期记忆神经网络（long-short term memory），是一种特殊的 RNN 模型，是为了解决 RNN 模型梯度弥散的问题而提出的。

LSTM 神经网络的结构如下：

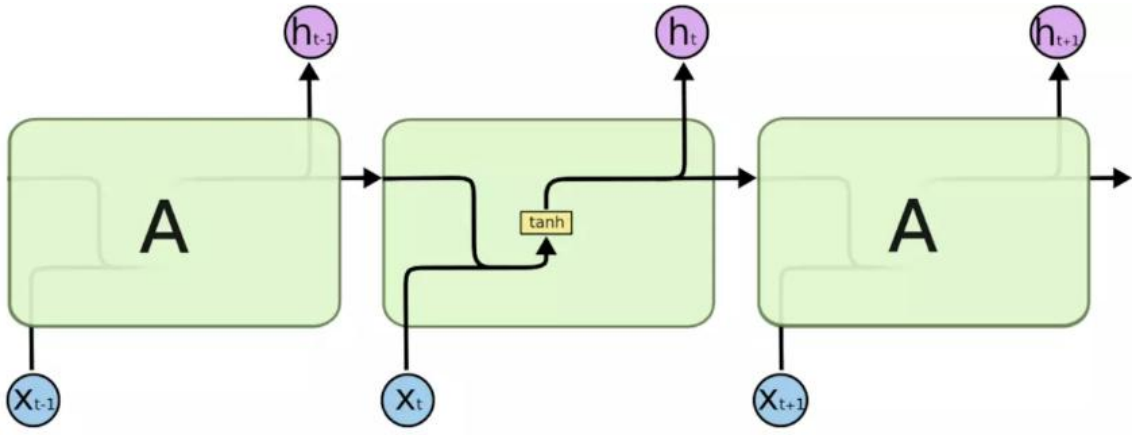


图 6LSTM 神经网络

如图所示，LSTM 神经网络与 RNN 相比，LSTM 的巧妙之处在于通过增加输入门限，遗忘门限和输出门限，使得自循环的权重是变化的，这样一来在模型参数固定的情况下，不同时刻的积分尺度可以动态改变，从而避免了梯度消失或者梯度膨胀的问题。[3]

### 3.7.2 分类数据初始化

第一步，我们统计分类标签的数量及其包含的留言分布情况。如下图所示：

| | 一级标签 | count |
|---|---|---|
| 0 | 城乡建设 | 2009 |
| 1 | 劳动和社会保障 | 1969 |
| 2 | 教育文体 | 1589 |
| 3 | 商贸旅游 | 1215 |
| 4 | 环境保护 | 938 |
| 5 | 卫生计生 | 877 |
| 6 | 交通运输 | 613 |

图 7 类目数量表

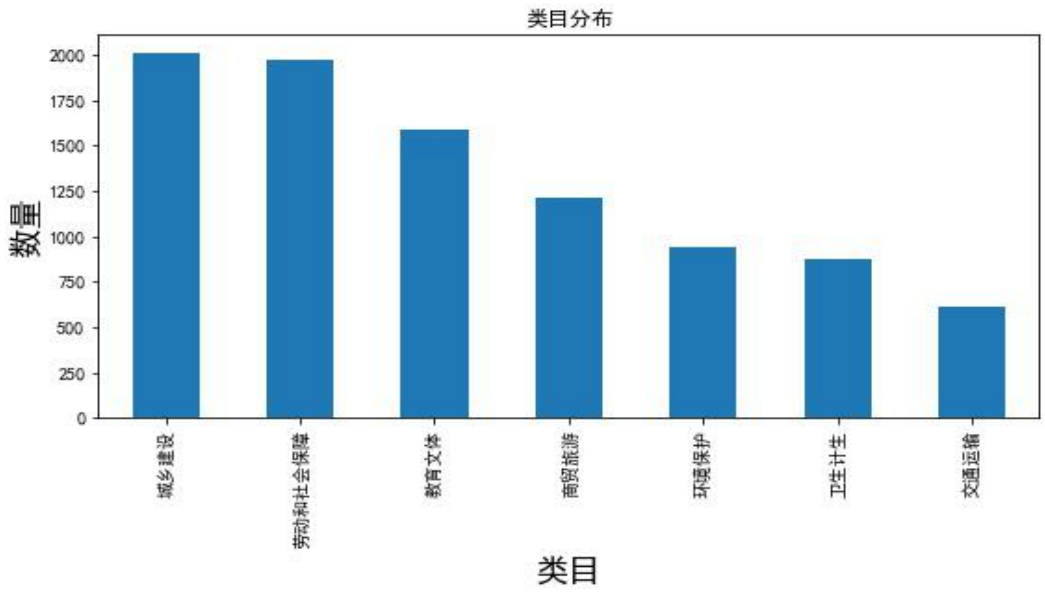


图 8 类目分类图

共有七类标签，但标签类目分布略有不均衡，采用　进行数据增强

第二步，将七类标签转化为 0-6 七个数字表示的形式，并进一步转化为独热编码。如图所示：

| | 留言详情 | 一级标签 | classify1_id |
|---|---|---|---|
| 0 | 护栏 锈迹斑斑 位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水 电等 设... | 城乡建设 | 0 |
| 1 | 程明 物业 小区 停车 收费 标准 上调 小区 物业 市程明 物业管理 有限公司 类 业主 ... | 城乡建设 | 0 |
| 2 | 健康 生活 环境 请 环保部门 检测 水是 日常生活 必不可少 用品 | 城乡建设 | 0 |
| 3 | 2015 年 购买 盛世 耀凯 小区 17 栋 楼 想 问一问 政府 职能部门 业主 投诉 ... | 城乡建设 | 0 |
| 4 | 不知 规划 供暖 具体位置 新 楼盘 | 城乡建设 | 0 |
| 5 | 停水 影响 小区 居民 日常生活 相关 单位 承诺 给与 解决 恳请 书记 出面 解决 民生 | 城乡建设 | 0 |
| 6 | 配合 代收 城市 垃圾处理 费 小区 对此 拒交 城市 垃圾处理 费 居民 业主 两个 确实... | 城乡建设 | 0 |
| 7 | 小区 绿化带 破烂不堪 小区 绿化带 垃圾场 地方 停车 | 城乡建设 | 0 |
| 8 | 滨江 社区 居委会 收缴 泰华 一村 小区 第二届 业主 委员 资料 公章 B 市 B4 区... | 城乡建设 | 0 |
| 9 | 市 自来水 公司 停留 梅 溪湖 新建 小区 生活 用水 保障 自来水 工作人员 楼 水 加... | 城乡建设 | 0 |

图 9 标签转化为数字形式

第三步，将分词好的文本进行向量化处理，将每一条文本转换成一个整数序列，即为每一条词语设置一个对应的编号。部分文本向量结果如下：

{'市': 1, '县': 2, '年': 3, '月': 4, '学校': 5, '领导': 6, '公司': 7, '区': 8, '西地省': 9, '小区': 10, '工作': 11, '政府': 12, '学生': 13, '部门': 14, '希望': 15, '人员': 16, '请': 17, '居民': 18, '相关': 19, '医院': 20, '业主': 21, '政策': 2
育': 105, '发展': 106, '发放': 107, '缴纳': 108, '物业': 109, '工程': 110, '中学': 111, '待遇': 112, '交': 113, '服务': 114, 'l': 115, '住房': 116, '地方': 117, '证明': 118, '回复': 119, '社区': 120, '新': 121, '违规': 122, '城市
查': 198, '排放': 199, '2015': 200, '11': 201, '厂': 202, '两个': 203, '信息': 204, '报名': 205, '搞': 206, '车': 207, '司机': 208, '咨询': 209, '景区': 210, '医疗': 211, '就业': 212, 'j': 213, '贷款': 214, 'a7': 215, '恳请': 216
90, '劳动者': 291, '报告': 292, '道路': 293, '西地': 294, '行政': 295, '发': 296, '失业': 297, '工龄': 298, '统一': 299, '增加': 300, '米': 301, '公开': 302, '改制': 303, '全国': 304, '同意': 305, '市政府': 306, '经营': 307, 'a
'k3': 384, '拒绝': 385, '条例': 386, '拖欠': 387, '家里': 388, '住院': 389, '历史': 390, '计划生育': 391, '整改': 392, '提出': 393, '取消': 394, '病人': 395, '成绩': 396, '承诺': 397, '建': 398, 'a5': 399, '正式': 400, '维护': 4
75, '万': 476, '劳动局': 477, '改革': 478, '领取': 479, '计算': 480, '社会保险': 481, '接受': 482, '区域': 483, '主任': 484, '廉租房': 485, 'g5': 486, '读书': 487, '就读': 488, '处': 489, '差': 490, '代表': 491, '此事': 492, '
567, '迟迟': 568, '80': 569, '16': 570, 'e8': 571, '上课': 572, '身份': 573, '学籍': 574, '出生': 575, '用人单位': 576, '意见': 577, '小时': 578, '连续': 579, '装修': 580, '城': 581, '工厂': 582, '申报': 583, '承担': 584, '客运
': 661, '款': 662, '办公室': 663, '补缴': 664, '人才': 665, '受理': 666, '60': 667, '生': 668, '关闭': 669, '许可证': 670, '儿童': 671, '破坏': 672, '同学': 673, '煤矿': 674, '假': 675, '丶': 676, '砖厂': 677, '否': 678, '计生办
, 'm6': 755, '一线': 756, '31': 757, '资格证': 758, '养老': 759, '仲裁': 760, '工伤保险': 761, '准生证': 762, '诊所': 763, '制定': 764, '办事处': 765, '无视': 766, '房产': 767, '保证': 768, '旁边': 769, '24': 770, '超市': 771
': 846, '早上': 847, '运行': 848, '办事': 849, '维权': 850, '大部分': 851, '交了': 852, '混凝土': 853, '位置': 854, '唯一': 855, '每人': 856, '老公': 857, 'l3': 858, '换': 859, 'e12': 860, '检验': 861, '早日': 862, '物流': 863,
937, '废气': 938, '打工': 939, '300': 940, '中标': 941, '县县': 942, '在职': 943, 'l12': 944, '非': 945, 'm7': 946, '停车': 947, '水电': 948, '2008': 949, '坐': 950, '对面': 951, '治理': 952, '建成': 953, '山': 954, '好多': 95
27, '单独': 1028, '名义': 1029, '加快': 1030, '公里': 1031, '场所': 1032, '委托': 1033, '桥': 1034, '营运': 1035, 'https': 1036, '多个': 1037, '区政府': 1038, '送': 1039, '上午': 1040, '开放': 1041, '一种': 1042, 'd9': 10
112, '做出': 1113, '河': 1114, '人民法院': 1115, '元月': 1116, '家人': 1117, '侵犯': 1118, 'f7': 1119, '流程': 1120, '回答': 1121, '教室': 1122, 'l1': 1123, 'l6': 1124, 'l10': 1125, '指定': 1126, '征地': 1127, '校方': 1128
, '绿化': 1197, '阳光': 1198, '洒水车': 1199, '不好': 1200, '车站': 1201, '候选人': 1202, '确认': 1203, '黑': 1204, '福利': 1205, '询问': 1206, '县市': 1207, '世纪': 1208, '高速': 1209, '补': 1210, '3': 1211, '挂靠': 1212
'g8': 1282, '工商': 1283, '关停': 1284, '废水': 1285, 'd8': 1286, '取件': 1287, '中考': 1288, '课程': 1289, '照顾': 1290, '周岁': 1291, '保障局': 1292, '看病': 1293, '直销': 1294, '垃圾处理': 1295, '记录': 1296, '表': 12
366, '西': 1367, '早': 1368, '理解': 1369, '基础': 1370, '燃气公司': 1371, '肯定': 1372, '入住': 1373, '电': 1374, '处于': 1375, '拆': 1376, '天气': 1377, '打表': 1378, '高新区': 1379, '麻烦': 1380, '两': 1381, '责令': 13
1451, '答应': 1452, '项目部': 1453, '力度': 1454, '店': 1455, '试问': 1456, '几次': 1457, '水表': 1458, '工业园': 1459, '教育部门': 1460, '精神': 1461, '公道': 1462, '刘': 1463, '白天': 1464, '未果': 1465, '地段': 1466
5, '修': 1536, '专家': 1537, '院': 1538, '象': 1539, '县一中': 1540, '投标人': 1541, '周末': 1542, '两次': 1543, '出口': 1544, '频繁': 1545, '求': 1546, '招投标': 1547, '认证': 1548, 'm1': 1549, 'm13': 1550, '妈妈': 155
: 1619, '市里': 1620, '转到': 1621, '十年': 1622, '账户': 1623, '实际上': 1624, '网': 1625, '为啥': 1626, 'j4': 1627, '电影': 1628, '下达': 1629, '一句': 1630, '法律法规': 1631, '本村': 1632, '当地政府': 1633, '第一次':
03, '岭': 1704, '收房': 1705, '容积率': 1706, '明白': 1707, '发出': 1708, '选址': 1709, '侵害': 1710, '适合': 1711, '场': 1712, '三级': 1713, '认可': 1714, '质监局': 1715, '小城镇': 1716, '传统': 1717, '东山': 1718, '供
88, '发工资': 1789, '老工人': 1790, '新生儿': 1791, '生二胎': 1792, '501': 1793, '磁悬浮': 1794, '心声': 1795, '实验': 1796, '堆放': 1797, '地址': 1798, '不行': 1799, '地面': 1800, '不见': 1801, '商铺': 1802, '所属': 18
72, '驾驶员': 1873, '经营权': 1874, '包裹': 1875, '日起': 1876, '混乱': 1877, '支行': 1878, '拉': 1879, '不通': 1880, '行政部门': 1881, '担任': 1882, '培训班': 1883, '监考': 1884, '学员': 1885, '人事厅': 1886, '爸爸': 1

图 10 文本向量化

第四步，填充文本数据为定长序列，在之前的统计中，我们发现分词后的数据词语长度最大为 303，因此设置每条留言中词语的最长长度为 300，不足的词语在尾部补 0，多余的截去。

第五步，以 4：1 的比例划分训练集和测试集，分别用于模型训练和测试。如图所示：

```
(8146, 300) (8146, 7)
(906, 300) (906, 7)
```

图 11 划分训练集和测试集

第六步，进行 word2vec 词向量的初始化，我们加载之前训练好的 word2vec 模型，将词向量初始化为字典，用于 LSTM 的嵌入层。

### 3.7.3 利用 Keras 搭建 LSTM 模型

设置嵌入层：我们设置嵌入层的输入维度为词向量矩阵的长度，设置输出维度为 100，输入长度为文本数据的个数。

设置 SpatialDropout1D 层：将在每次训练中断开一定比例的输入神经元连接，作用与 Dropout 类似，但不同的是，其断开的是整个 1D 特征图，而不是单个神经元，能够提高特征图之间的独立性，我们这里设置断开连接的比例是 0.2。

设置 LSTM 层：我们设置 LSTM 层包含 100 个神经元，文本数据和隐藏层之间的 dropout 的比例为 0.2，隐藏层之间的 dropout 为 0.2。

设置全连接层：用于全连接层的激活函数主要有 tanh、softmax、relu 等，实验发现，使用 tanh 函数的效果要优于其他两种，因此我们选择 tanh 为激活函数。

LSTM 模型的具体参数如下：

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 300, 100)          2483500
_________________________________________________________________
spatial_dropout1d_3 (Spatial (None, 300, 100)          0
_________________________________________________________________
lstm_3 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_3 (Dense)              (None, 7)                 707
=================================================================
Total params: 2,564,607
Trainable params: 2,564,607
Non-trainable params: 0
_________________________________________________________________
None
```

图 12LSTM 模型参数表

LSTM 模型流程图如下：

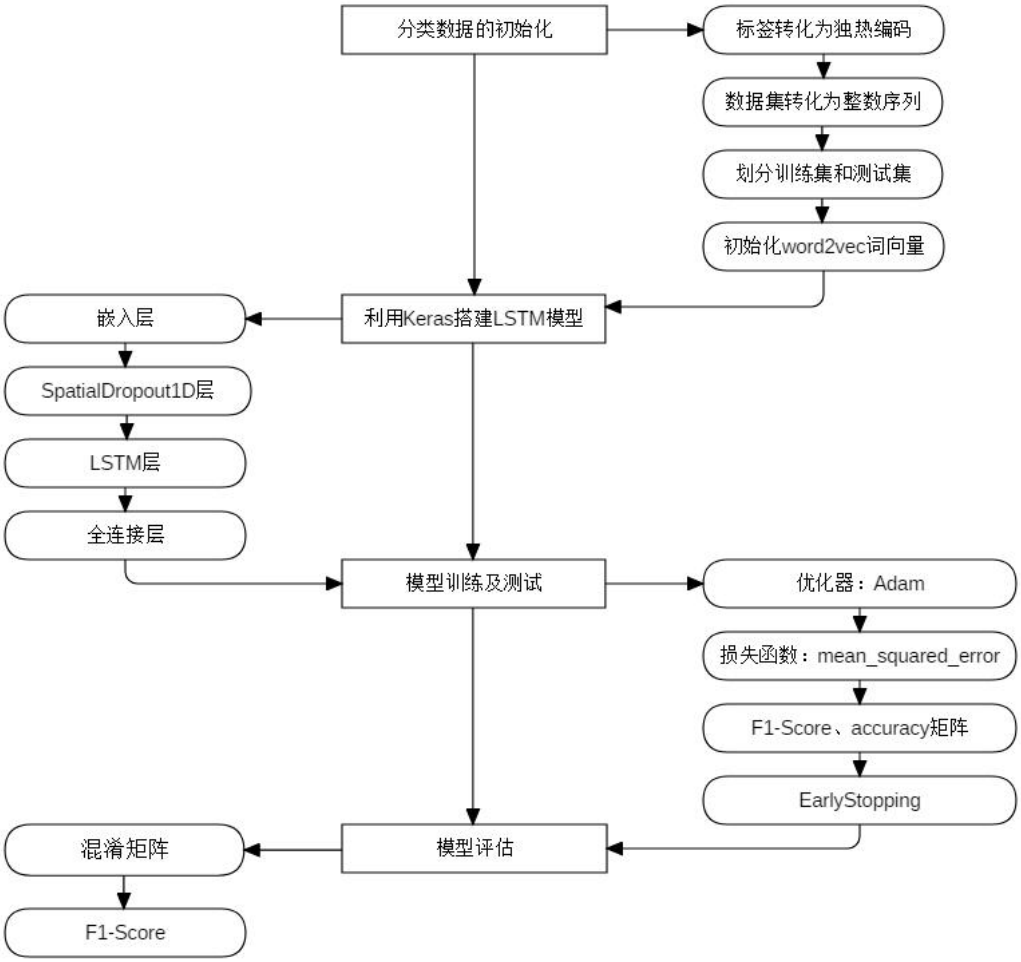


图 13LSTM 流程图

### 3.7.4 模型训练及预测

在模型训练中，优化器主要有 rmsprop，adam 等，实验发现采用 adam 作为优化器的效果较好，损失函数主要有 mean_squared_error、binary_crossentropy、categorical_crossentropy 等，实验发现采用 mean_squared_error 效果较好。因此我们采用 adam 作为优化器，mean_squared_error 作为损失函数。

接着我们引入评估方法对模型进行性能评估。

对模型进行评估的方法有很多种，常见的有使用准确率（accuracy）和 F1 分数等方法来评估模型的质量。

F1 分数（F1 Score）又称平衡 F 分数（balanced F score），他被定义为精确率

（ *precision* ）和召回率（ *recall* ）的调和平均数。被广泛应用在自然语言处理领域，用来衡量算法或系统的性能，即：

$$F_1=2\bullet\frac{precision\cdot recall}{precision+recall} \tag{2-4}$$

更一般的，我们定义 $F_\beta$分数为：

$$F_\beta=(1+\beta^2)\bullet\frac{precision\cdot recall}{(\beta^2\cdot precision)+recall} \tag{2-5}$$

而对于多分类问题，通常不使用准确率（accuracy）来评估模型的质量，因为当训练数据分布不平衡时，准确度（accuracy）就不能反映出模型的实际预测精度，因此我们采用 F1 分数和 accuracy 相结合的形式作为模型的 metric 来评估建立的模型。我们设置最大训练轮数 epochs 为 30，每次抽取 100 个文本数据进行训练，设置 validation_data,将之前划分好的测试集用于每次的测试，为了防止 epochs 过少，网络发生欠拟合及 epochs 过多网络发生过拟合，保证训练的高效性，在 callbacks 中我们指定 EarlyStopping，用于提早停止训练，监控的数据接口 monitor 为测试集 loss，patience 设置为 0.0001，即当测试集的 loss 三个 epoch 内提升的程度小于 0.0001 时，默认达到最优训练次数，停止训练。
代码如下：

```
history = model.fit(x_train, y_train, epochs=epochs, batch_size=batch_size,validation_data=(x_test,y_test),
                    callbacks=[EarlyStopping(monitor='val_loss', patience=3, min_delta=0.0001)])
```

图 14 最优训练次数

部分训练代码如下：

```
Train on 8146 samples, validate on 906 samples
Epoch 1/30
8146/8146 [==============================] - 135s 17ms/step - loss: 0.1073 - acc: 0.4131 - f1_m: 0.1251 - val_loss: 0.0793 - val_acc: 0.59
49 - val_f1_m: 0.4308
Epoch 2/30
8146/8146 [==============================] - 133s 16ms/step - loss: 0.0607 - acc: 0.7557 - f1_m: 0.6630 - val_loss: 0.0566 - val_acc: 0.74
83 - val_f1_m: 0.6906
Epoch 3/30
8146/8146 [==============================] - 131s 16ms/step - loss: 0.0369 - acc: 0.8957 - f1_m: 0.8530 - val_loss: 0.0560 - val_acc: 0.74
72 - val_f1_m: 0.7174
Epoch 4/30
8146/8146 [==============================] - 128s 16ms/step - loss: 0.0296 - acc: 0.9256 - f1_m: 0.8954 - val_loss: 0.0512 - val_acc: 0.76
71 - val_f1_m: 0.7554
Epoch 5/30
8146/8146 [==============================] - 130s 16ms/step - loss: 0.0187 - acc: 0.9541 - f1_m: 0.9468 - val_loss: 0.0495 - val_acc: 0.76
49 - val_f1_m: 0.7807
Epoch 6/30
8146/8146 [==============================] - 142s 17ms/step - loss: 0.0256 - acc: 0.9207 - f1_m: 0.9109 - val_loss: 0.0589 - val_acc: 0.72
96 - val_f1_m: 0.6957
Epoch 7/30
8146/8146 [==============================] - 154s 19ms/step - loss: 0.0152 - acc: 0.9678 - f1_m: 0.9593 - val_loss: 0.0501 - val_acc: 0.75
61 - val_f1_m: 0.7759
Epoch 8/30
8146/8146 [==============================] - 166s 20ms/step - loss: 0.0106 - acc: 0.9751 - f1_m: 0.9711 - val_loss: 0.0501 - val_acc: 0.76
05 - val_f1_m: 0.7591
```

图 15LSTM 部分训练代码

可以看到，在第八轮训练时，模型达到了最佳训练次数，训练集精度达到0.9711.

本题采用的 LSTM 模型准确度（accuracy）如下图所示：

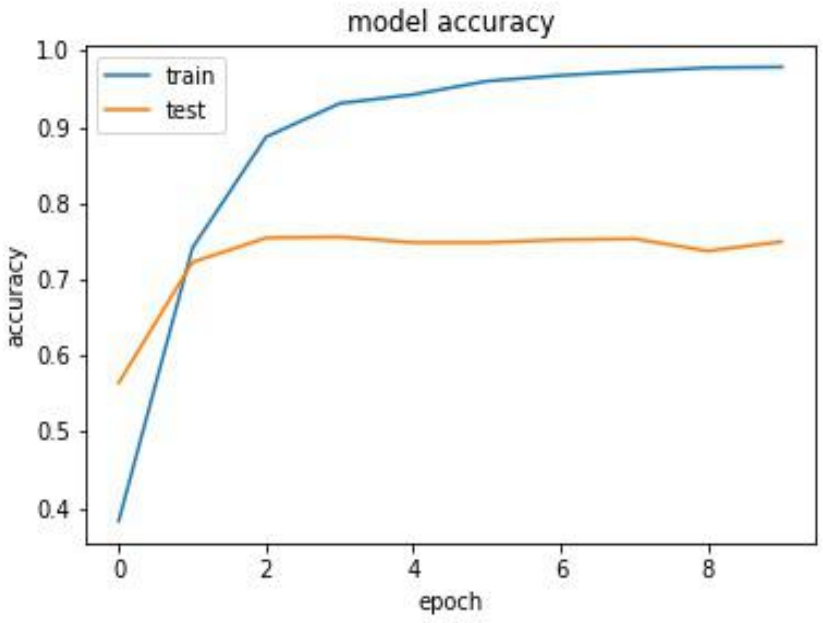


图 16 模型准确率

再计算模型的 F1 分数如图所示：

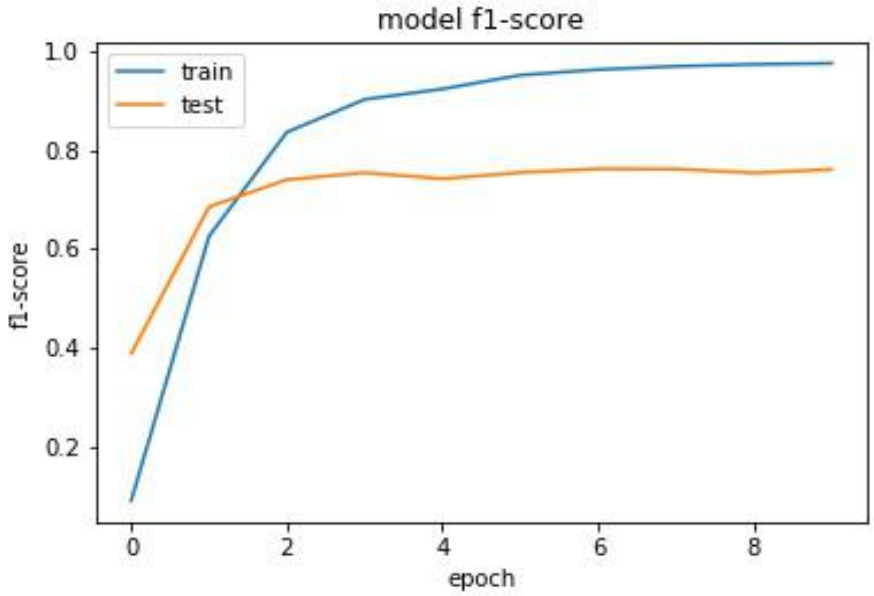


图 17 模型 F1 Score

另外，还需计算模型的丢失率（loss）,如图所示：

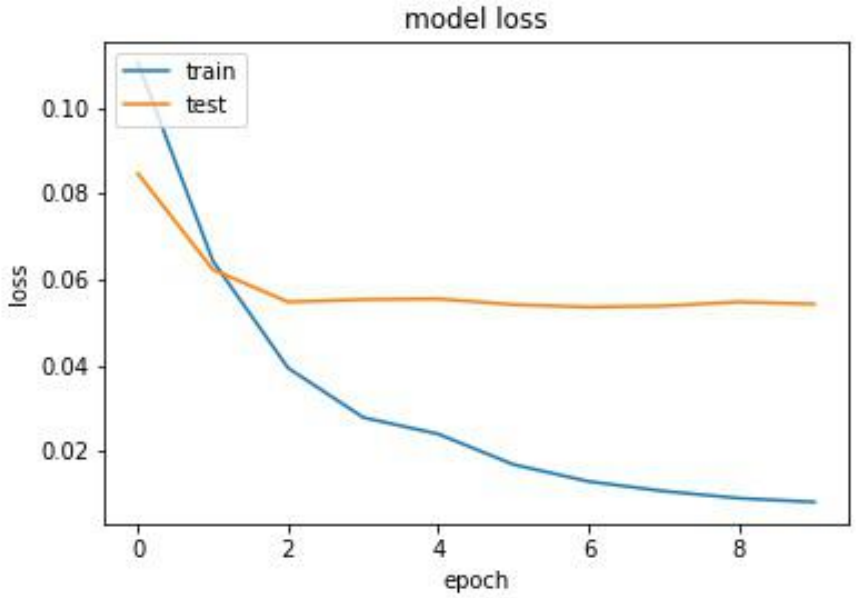


图 18 模型丢失率

### 3.7.5 模型评估

我们用划分好的测试集进行最终测试，并将预测值与真实值进行比较，得出测试集训练精度画出混淆矩阵：

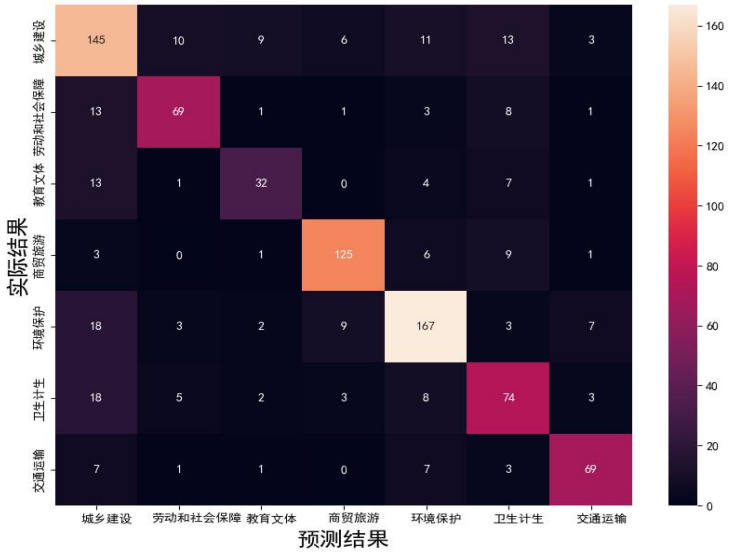


图 19 训练精度混淆矩阵

各类别 F1 Score 如下：

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 城乡建设 | 0.67 | 0.74 | 0.70 | 197 |
| 劳动和社会保障 | 0.76 | 0.75 | 0.75 | 96 |
| 教育文体 | 0.64 | 0.52 | 0.57 | 58 |
| 商贸旅游 | 0.89 | 0.92 | 0.90 | 145 |
| 环境保护 | 0.85 | 0.85 | 0.85 | 209 |
| 卫生计生 | 0.61 | 0.60 | 0.61 | 113 |
| 交通运输 | 0.79 | 0.73 | 0.76 | 88 |
| accuracy | | | 0.76 | 906 |
| macro avg | 0.75 | 0.73 | 0.74 | 906 |
| weighted avg | 0.76 | 0.76 | 0.76 | 906 |

图 20 分类别 F1 Score

从以上 F1 Score 来看，“商贸旅游”类 F1 Score 最高（0.90），而“教育文

体”类 F1 Score 最差只有 0.57，究其原因应是因为“教育文体”类训练数据最少只有 58 条，导致模型学习不够充分，产生预测失误较多。总体上，通过混淆矩阵，可以看出预测效果良好，只有少部分真实值与预测值不匹配。[4]

# 4 问题二分析

## 4.1 问题二流程图

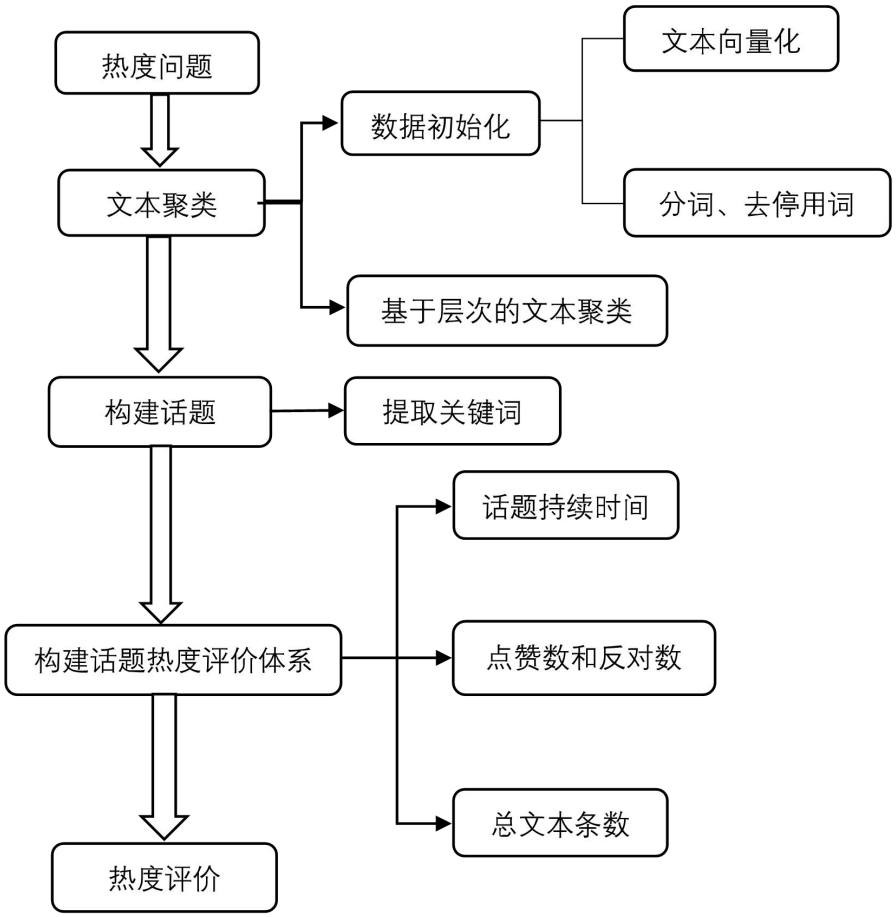


图 21 问题二流程图

## 4.2 基于层次的文本聚类

### 4.2.1 层次聚类介绍

层次聚类算法使用将数据组织分成若干组并形成一个相应的树状图的方法来进行聚类，其中又可以细分为两类，即自底向上的聚合层级聚类和自顶向下的分解层次聚类。聚合聚类的策略是先将每个对象各自作为一个原子聚类，然后对这些原子聚类逐层进行聚合，直至满足一定的终止条件为止。分解聚类则与其相反，它首先把所有的对象都看成一个聚类，然后将其不断分解直至满足终止条件。[5]

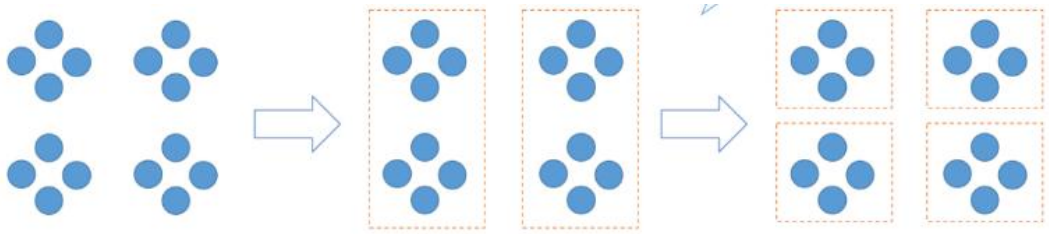

图 22 自顶向下算法

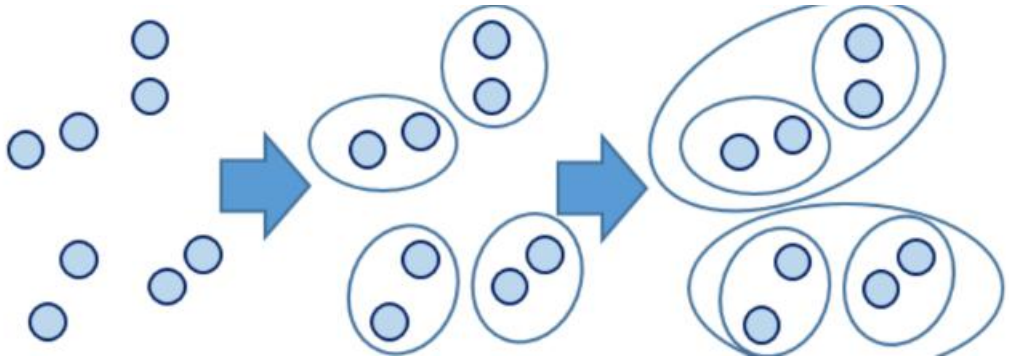

图 23 自底向上算法

对于层次聚类算法来说，根据度量两个子类相似度时依据的距离不同，又可分为 Single_Link, Complete_Link, Average_Link 三种层次聚类。其中 Single_Link 应用最为广泛。Single_Link 采用根据两个聚类中相隔最近两个点之间的距离来评价这两类之间的相似程度，另外两种分别根据最远距离和平均距离来进行不同类相似度评价。[6]

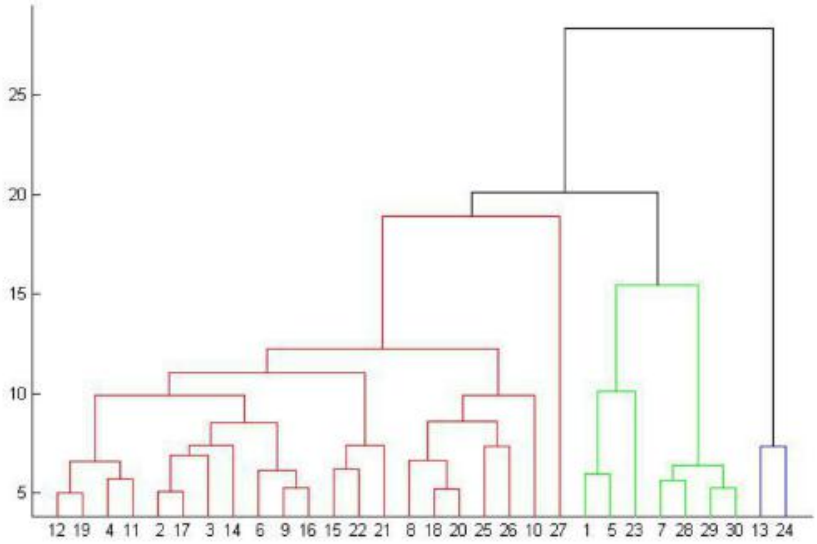


图 24 层次聚类算法

### 4.2.2Birch 聚类实现步骤

第一步，读入数据：读取附件 3 中“留言主题”，“留言详情”，“留言时间”，“点赞数”，“反对数”等数据。

第二步，取留言主题列进行分词，读取停用词表用来去除停用词。部分结果如下：

A3 区 一米阳光 婚纱 艺术摄影 合法 纳税
咨询 A6 区 道路 命名 规划 初步 成果 公示 城乡 门牌
A7 县 春华 镇金鼎村 水泥路 自来水 到户
A2 区 黄兴路 步行街 古道 巷 住户 卫生间 粪便 外排
市 A3 区 中海 国际 社区 三期 四期 空地 夜间 施工 噪音 扰民
A3 区麓 泉 社区 单方面 改变 麓 谷 明珠 小区 栋 架空层 性质
A2 区富 绿 新村 房产 性质
市 地铁 违规 用工 质疑
市 路 公交车 随意 变道 通行
A3 区 保利 麓 谷林语 桐梓 坡路 与麓 松路 交汇处 地铁 凌晨 点 施工 扰民
A7 县 特立 路 东四 路口 晚 高峰 太堵 建议 调整 信号灯 配时
A3 区 青青 家园 小区 乐果 果 零食 炒货 公共 通道 摆放 空调 扰民
拆除 聚美龙楚 西地省 商学院 宿舍 旁 安装 变压器 请求
市利保 壹号 公馆 项目 夜间 噪声 扰民
市 地铁 号线 星沙 大道 站 地铁 出入口 设置 不合理
A4 区 北辰 小区 非法 住 改商 何时能 解决
请 K3 县 乡村 医生 发 卫生室 执业 许可证
A7 县 春华 镇 石塘 铺村 党员 家开 麻将馆
咨询 异地 办理 出国 签证
投诉 市 温斯顿 英语 培训 学校 拖延 退费
A6 区乾源 国际 广场 停车场 违章 乱建 现象
A7 县 时代 星城 幢 非法经营 家庭旅馆

图 25 留言主题预处理结果

第三步，文本向量化计算

1 训练 word2vec 词向量：

采取之前第一问的方法，对本问的文本进行 word2vec 向量化处理。

2 得到 TF-IDF 权值矩阵：

TFIDF（Term Frequency-inverse Document Frequency,词频与逆文档频率）是一种统计方法，可以用于评估一个字词在整个语料库中的重要程度。它的主要思想是：如果某个词或短语在一个类别中出现频率较高，而在其他类别中出现频率较低，则认为该词或者短语具有很好的类别区分能力，适合用来分类。[7]TFIDF 的计算公式如下：

$$f(t,d)=\frac{f(t,d)}{\sum_k f(w_k,d)} \tag{4-1}$$

$$idf_t=\log(\frac{N}{1+df_t}) \tag{4-2}$$

其中，$f(t,d)$表示词条 t 在文档 d 中出现的次数，$df_t$表示语料库中包含词条 t 的文档数量，N 表示语料库中全部的文档数量。词条 t 的 TFIDF 权重即为：$tfidf_t=tf(t,d)\times idf_t$。

第四步，基于手肘法判断最佳聚类数，手肘法的核心指标是：SSE（误差平方和），公式如下：

$$SSE=\times\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right|^2 \tag{4-3}$$

其中$C_i$是第 i 个簇，$p$是$C_i$中的样本点，$m_i$是$C_i$的质心（$C_i$中所有样本的均值），$SSE$是所有样本的聚类误差，代表了聚类效果的好坏。

聚类数一般为 1 到数据总个数的平方根，我们根据不同聚类数计算其误差平方和 SSE，将 SSE 的变化趋势画成折线图，如图所示：

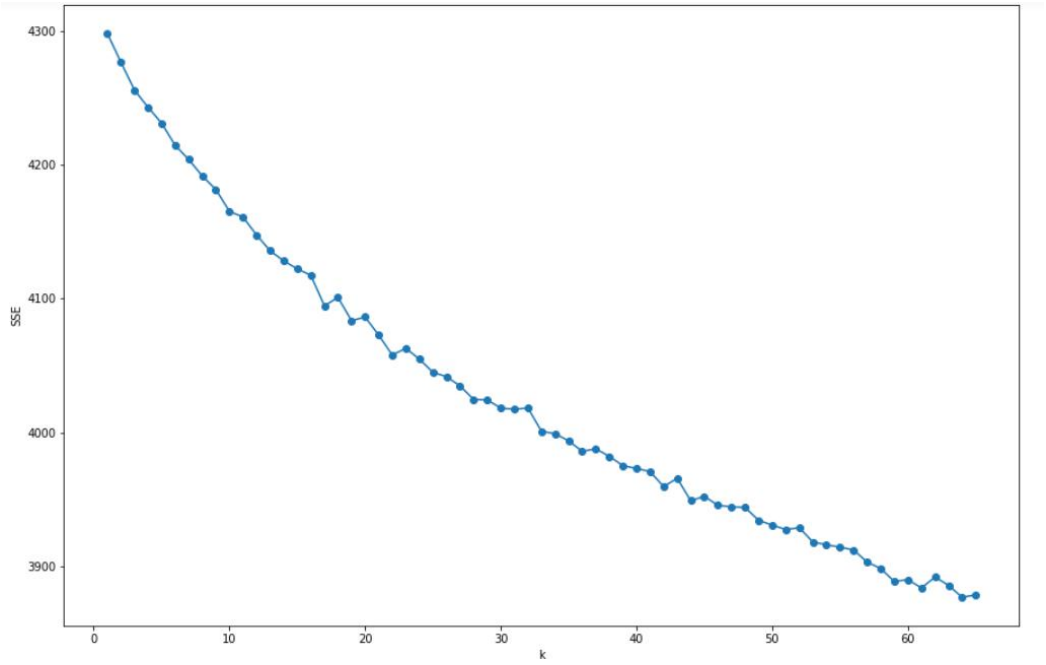


图 26 SSE 变化趋势

据此折线图判断，在聚类数 20-30 之前，SSE 变化趋势大，聚类数 20-30 之后，变化较为平缓，因此我们判断最佳聚类数在 20-30 之间，之后做出 20-30 内的 SSE 变化趋势图，做进一步判断。[8]

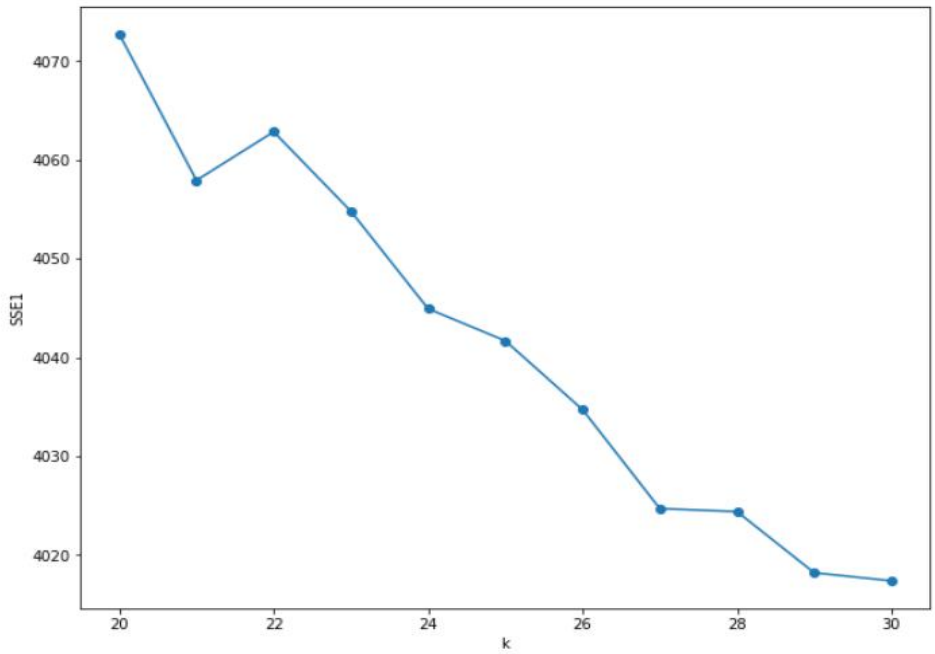


图 27 20-30SSE 变化趋势图

通过读图不难发现，26-27 变化趋势最大，因此判断最佳聚类数为 27.然后再

利用 TF-IDF 加权词向量获取句向量。

第四步，加权 word2vec 得到句向量，训练 Word2vec 模型后，累加词向量得到文档的向量化表示。这里我们使用 Scikit-learn 提供的 TfidfTransformer 模块计算词汇的 TFIDF 权重，将词向量和对应词汇的 TFIDF 权重相乘得到加权 Word2vec 词向量，累加加权词向量得到加权文本向量化表示。

第五步，BIRCH 的全称是利用层次方法的平衡迭代规约和聚类（Balanced Iterative Reducing and Clustering Using Hierarchies），BIRCH 算法利用了一个树结构来帮助我们快速的聚类，这个数结构类似于平衡 B+树，我们一般将它称之为聚类特征树(Clustering Feature Tree，简称 CF Tree)。树的每一个节点是由若干个聚类特征(Clustering Feature，简称 CF)组成。聚类特征树大致包含以下结构：每个节点包括叶子节点都有若干个 CF，而内部节点的 CF 有指向孩子节点的指针，所有的叶子节点用一个双向链表链接起来[9]。我们采用 BIRCH 进行聚类，并计算轮廓系数，聚类结果如图所示：

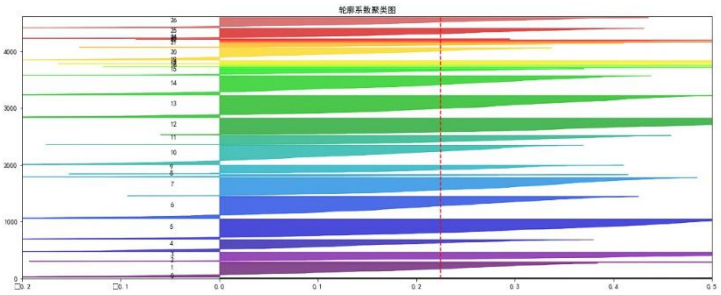

图 28 轮廓系数聚类图[10]

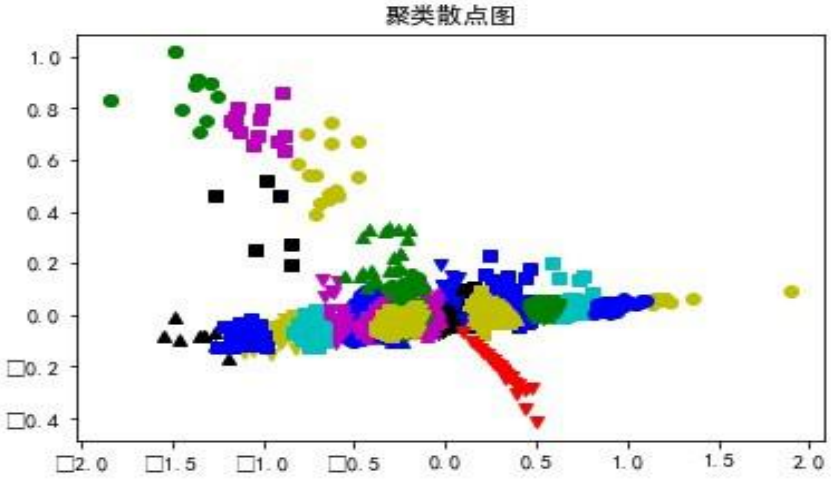


图 29 聚类散点图

第六步，针对每一个聚类构建话题，在每一组聚类数中提取关键词，每组提取 50 个。

第七步，将聚类结果分配到原数据，如图所示：

| | 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 | category |
|---|---|---|---|---|---|---|---|---|
| 0 | 188006 | A000102948 | A3区一米阳光婚纱艺术摄影是否合法纳税了? | 2019/2/28 11:25:05 | \n\t\t\t\t\t\n\t\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋320，... | 0 | 0 | 3 |
| 1 | 188007 | A00074795 | 咨询A6区道路命名规划初步成果公示和城乡门牌问题 | 2019/2/14 20:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t\tA市A6区道路命名规划已经初步成果公示文件，... | 0 | 1 | 5 |
| 2 | 188031 | A00040066 | 反映A7县春华镇金鼎村水泥路、自来水到户的问题 | 2019/7/19 18:19:54 | \n\t\t\t\t\t\n\t\t\t\t\t\t本人系春华镇金鼎村七里组村民，不知是否有相关... | 0 | 1 | 13 |
| 3 | 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 | \n\t\t\t\t\t\n\t\t\t\t\t\t靠近黄兴路步行街，城南路街道、大古道巷、一步... | 0 | 1 | 19 |
| 4 | 188059 | A00028571 | A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民 | 2019/11/22 16:54:42 | \n\t\t\t\t\t\n\t\t\t\t\t\tA市A3区中海国际社区三期四期中间，即蓝天璞... | 0 | 0 | 1 |
| 5 | 188073 | A909164 | A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质 | 2019/3/11 11:40:42 | \n\t\t\t\t\t\n\t\t\t\t\t\t作为麓泉社区麓谷明珠小区6栋居民，我们近期感... | 0 | 0 | 5 |
| 6 | 188074 | A909092 | A2区富绿新村房产的性质是什么? | 2019/1/31 20:17:32 | \n\t\t\t\t\t\n\t\t\t\t\t\t“二高一部”发出关于针对非法集资的打击的通知... | 0 | 0 | 26 |
| 7 | 188119 | A00035029 | 对A市地铁违规用工问题的质疑 | 2019/5/27 16:04:44 | \n\t\t\t\t\t\n\t\t\t\t\t\t我是一名在A市某地铁站上班的安检员，我是由中... | 0 | 0 | 25 |
| 8 | 188170 | A88011323 | A市6路公交车随意变道通行 | 2019/12/23 8:50:24 | \n\t\t\t\t\t\n\t\t\t\t\t\t12月21日下午17时52分许，6路公交车(... | 0 | 0 | 25 |
| 9 | 188249 | A00084085 | A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民 | 2019/9/17 4:25:00 | \n\t\t\t\t\t\n\t\t\t\t\t\t保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2... | 0 | 0 | 6 |

图 30 分配聚类结果到原数据

## 4.3 话题的基本概念

本题的目的是挖掘热点话题，以及对热点话题进行热度排序，因此先介绍话题的的基本概念。

话题，通常是指谈论的目的，谈话的主题。它可以是发生于某一时间段和某一地点的事件，以及与此事件引发的直接相关的事件。所以，话题可以被描述为某一事件及其引发的事件的集合。大多数情况下，话题必须包含一个核心事件，其他时间都是由核心事件所引发的。如果某一条留言与某个事件具有直接联系，那么可以推断该留言与这个时间所形成的话题相关。由此，我们就可以建立留言与话题热度之间的联系。

话题相比某一条留言而言还要具有概括性，它使用最精简的话语概括出事件，有时，甚至可以用一个词语、短语就可以代表一个事件。因此，话题一般由能够概括文章主题的主题词，具有关键特征的词语，和某领域内的术语这三种词组中的一种或多种组合而成。

## 4.4 指标体系的建立

### 4.4.1 评价话题热度的指标选择

话题热度反映了话题本身的重要程度和公众的话题参与度，这是一个综合性的指标，其影响因素多种多样，我们主要根据以下几个方面选择本题的影响因素：

1 话题持续时间，一般情况下，话题是对于某一段时间所发生的某件事情的即时报道和讨论，具有即时性，而随着时间的推移，人们对于某个话题的关注自然也会下降，从而旧的话题的热度会被新的话题所取代。时间久远的话题受到的关注比刚刚发生的话题要少很多，因此越新产生的话题具有越高的热度，所以选择话题持续时间作为评价话题热度的一项重要指标。

2 点赞数和反对数，我们注意到本题涉及到公众留言的话题设置了点赞和反对的操作，浏览网站的每一位网民都可以对每条留言内容进行点赞或反对，即认可留言内容可以选择点赞，不认可留言内容可以选择反对。这在一定程度上反映了某话题的公众关注度，某一话题的点赞数和反对数越多，表明公众对这一话题的重视程度越高，而公众的重视必然会引起社会各界的广泛关注，话题热度自然也就越高，因此我们选择话题的点赞数和反对数作为评价话题热度的另一重要指标。

3 文本总条数，大多数情况下，某一话题通常不止包含一两条文本，即可能有很多群众针对同一话题发表自己的留言，而文本总条数是指从首次发布时间到当前时间，涉及到某个话题的所有文本数量的总和。显然，文本总条数这一指标越高，说明对应的事件受到了更多公众的关注，话题热度也会随之升高。也就是，某话题的文本总条数越高，这一话题的热度越高。因此，我们选择文本总条数作为评价话题热度的一项十分关键的指标。[11]

#### 4.4.1.1 话题持续时间

在时间维度通过发布时间的信息挖掘内容热度，我们采用牛顿冷却定律作为时间衰减函数，即：

$$H(t) = H_a \bullet \exp[-\gamma \bullet (t - t_{last}) / 86400] \tag{4-3}$$

其中 $H_a$ 是内容的原始热度值，即通过互动量算出的热度值。采用以 e 为底的指数函数，$t_{last}$ 即该内容的最后回复时间。时间的单位都是秒，所以除去一天的秒数 86400，从而实现以天为单位做衰减。$\gamma$ 是所谓“重力因子”（gravity factor），

即该值越大，帖子的热度会更快地下落。

#### 4.4.1.2 点赞数和反对数

在时间维度对点赞量和反对量进行处理，这里可以使用基于票数的 Reddit 的热度排名算法，Reddit 算法是基于用户投票的热度排序算法，其数学表达式如下：

$$\int(t_s, y, z) = \lg z + \frac{y^{t_s}}{45000} \tag{4-4}$$

其中 $x$ 为点赞数与反对数之差（辅助变量），$t_s$ 为发布时间到现在的时间间隔，单位秒。$z$ 为 $x$ 和 1 的绝对值的最大值。$y$ 在当 $x$>0 时取 1；$x$<0 时取-1；$x$=0 时取 0。

#### 4.4.1.3 合计条数

合计条数即每一类话题的总文本条数，直接计算每一类话题包含的总文本条数和得出。

### 4.4.2 基于熵权法的热度问题分析

#### 4.4.2.1 熵权法介绍

熵权法是计算指标权重的经典算法之一，它是指用来判断某个指标的离散程度的数学方法。离散程度越大，即信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也越大。根据熵的特性，我们可以通过计算熵值来判断一个事件的随机性及无序程度，也可以用熵值来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的影响越大。具体步骤如下：
第一步，假设数据有 n 行记录，m 行变量，数据就可以用一个 n*m 行的矩阵表示。

$$A = [x_1, \ldots, x_m] \tag{4-5}$$

第二步，数据归一化处理 $x_{ij}$ 表示矩阵 A 的第 i 行第 j 列的元素。

$$x_{ij} = \frac{x_{ij} - \min(xj)}{\max(x_j) - \min(x_j)} \tag{4-6}$$

第三步，计算第 j 项指标下第 i 个记录所占比重。

$$P_{ij}=\frac{x_{ij}}{\sum_{1}^{n}x_{ij}}(j=1,2,...,m) \tag{4-7}$$

第四步，计算第 j 项的熵值。

$$e_j=-k*\sum_{1}^{n}P_{ij}*\log(P_{ij}),k=\frac{1}{\ln(n)} \tag{4-8}$$

第五步，计算第 j 项的差异系数。

$$g_j=1-e_j \tag{4-9}$$

第六步，计算第 j 项指标的权重。

$$w_j=\frac{g_j}{\sum_{1}^{m}g_j} \tag{4-10}$$

# 5 问题三分析

## 5.1 问题三流程图

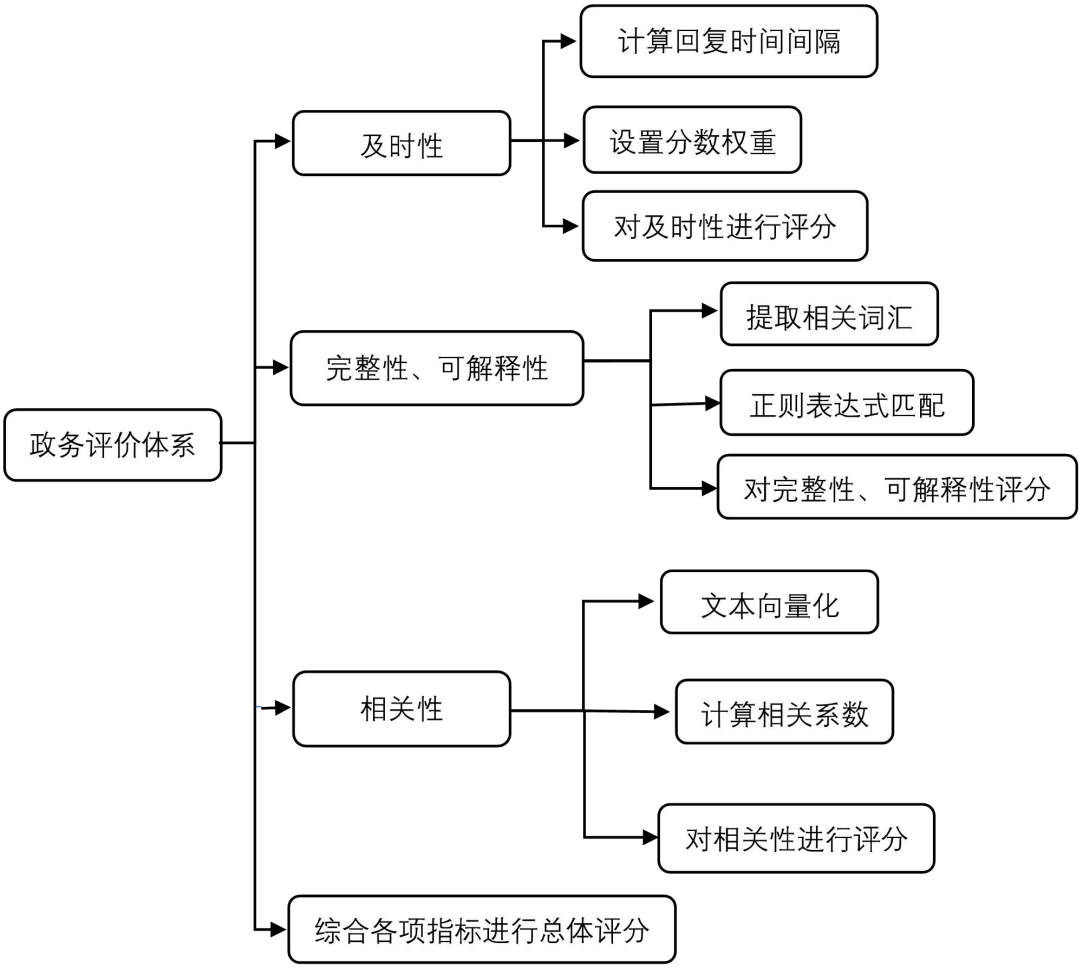


图 31 问题三流程图

## 5.2 政务水平的影响因素

1 及时性：及时性是衡量政府工作人员答复群众留言是否及时的指标。它是一种可以直接反映政府工作人员工作效率的良好手段，回复及时性强则意味着政府工作人员效率高，自然也就反映了政府良好的政务水平。而政府回复及时性强弱也会在群众心里留下自己对于政府政务水平的印象，如果政府回复及时性强，群

众在心里便会产生政府政务水平高的感受。综合以上两点，我们选择及时性作为政务水平的影响因素之一。

2 完整性：完整性是指政府工作人员对于群众留言的意见答复是否具有一定的规范以及是否包含了必要的称呼、敬语、感谢等句子成分。意见答复完整性高意味着政府工作人员执行政务是依据着某种行为规范的，所以完整性高对应这政府工作人员办事流程规范性强，即政务水平高。同时，政府工作人员回复的完整性强，留言的群众的心情便会感到不同程度愉悦，群众对于政府的认可度也有一定的提升，这也是政府政务水平高的体现。因此，我们选择完整性作为政务水平的又一影响因素。

3 可解释性：可解释性是指政府工作人员在对群众留言涉及到的问题进行回复时是否有针对反映的问题进行具体的调查，是否是根据当前社会形势回答，是否依据某些法律，是否按照有关部门依据，是否说明原因等。这项因素反映了政府工作人员在回复留言时有无针对留言内容思考、调研、考察后再进行回复，这充分反映了政府政务能力高低，如可解释性强则意味着政府工作人员追求实事求是，而不是应付差事。可解释性强弱能充分反映政务水平高低，因此，我们选择可解释性作为政务水平的第三个影响因素。

## 5.3 对于答复及时性的评价

### 5.3.1 关于答复及时性的核心概念

5.3.1.1 有效时间差

有效时间差指的是群众留言时间与政府工作人员答复群众留言的时间之间的时间差。这个时间差值能够充分反应政府工作人员答复留言的及时性，我们在统计时具体到天。这个差值与留言反映事件的性质、重要性、解决起来的难易程度并无相关性，它仅仅记录工作人员有无回复和回复速度，虽然并非直接涉及回复的质量，但可以作为评价回复质量的重要指标之一。计算公式为：

$$\text{有效时间差}=\text{回复时间}-\text{发布时间} \tag{5-1}$$

计算时间的方式以天为单位，间隔为 1 天。

5.3.1.2 反馈即时性

反馈即时性是指政府网站对于网民提出的建议、意见、反馈、监督等信息进行回复的有效时间差的相对大小值。这项指标没有精确定义的大小值，而是根据

回复时间的分布划分出的范围，因此是一个相对的概念，主要通过对比数据产生评价。根据时间差的数据，划分出不同的得分。[12]

### 5.3.2 针对回复及时性的处理步骤

第一步，首先读取附件 4 中的留言时间和回复时间的间隔。由于原数据的留言时间和答复时间中时间格式不同，为了方便处理，将留言时间和回复时间转化为 datetime 格式，即 yyyy-mm-dd 格式，方便后续的处理。将答复时间和留言时间相减，计算出时间间隔。部分代码如下：

| | 留言编号 | 留言用户 | 留言主题 | 留言时间 | 留言详情 | 答复意见 | 答复时间 | interval |
|---|---|---|---|---|---|---|---|---|
| 0 | 2549 | A00045581 | A2区景蓉华苑物业管理有问题 | 2019-04-25 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道... | 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... | 2019-05-10 00:00:00 | 15 days |
| 1 | 2554 | A00023583 | A3区潇楚南路洋湖段怎么还没修好? | 2019-04-24 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了... | 网友"A00023583"：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋... | 2019-05-09 00:00:00 | 15 days |
| 2 | 2555 | A00031618 | 请加快提高A市民营幼儿园老师的待遇 | 2019-04-24 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来... | 市民同志：你好！您反映的"请加快提高民营幼儿园教师的待遇"的来信已收悉。现回复如下：为了改善... | 2019-05-09 00:00:00 | 15 days |
| 3 | 2557 | A000110735 | 在A市买公寓能享受人才新政购房补贴吗? | 2019-04-24 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政... | 网友"A000110735"：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反... | 2019-05-09 00:00:00 | 15 days |
| 4 | 2574 | A0009233 | 关于A市公交站点名称变更的建议 | 2019-04-23 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t建议将"白竹坡路口"更名为"马坡岭小学"，原... | 网友"A0009233"，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议"白竹坡... | 2019-05-09 00:00:00 | 16 days |
| 5 | 2759 | A00077538 | A3区含浦镇马路卫生很差 | 2019-04-08 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t欢迎领导来A市泥泞不堪的小含浦镇滚泥巴，这个... | 网友"A00077538"：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、... | 2019-05-09 00:00:00 | 31 days |
| 6 | 2849 | A000100804 | A3区教师村小区盼望早日安装电梯 | 2019-03-29 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：您好！过去在小区买房是为了自己... | 网友"A000100804"：您好！针对您反映A3区教师村小区盼望早日安装电梯的问题,A3区... | 2019-05-09 00:00:00 | 41 days |
| 7 | 3681 | UU00812 | 反映A5区东澜湾社区居民的集体民生诉求 | 2018-12-31 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t我做为一东澜湾社区居民，我替社区几千户居民为... | 网友"UU00812"您好！您的留言已收悉。现将有关情况回复如下：一、关于小区附近幼儿园的问... | 2019-01-29 00:00:00 | 29 days |
| 8 | 3683 | UU008792 | 反映A市美麓阳光住宅楼无故停工以及质量问题 | 2018-12-31 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t我是美麓阳光a栋803业主，现工地已停工二个... | 网友"UU008792"您好！您的留言已收悉。现将有关情况回复如下：据查，美麓阳光项目位于A... | 2019-01-16 00:00:00 | 16 days |
| 9 | 3684 | UU008687 | 反映A市洋湖新城和顺路洋湖壹号小区路段公共绿化带的问题 | 2018-12-31 00:00:00 | \n\t\t\t\t\t\n\t\t\t\t\t胡书记好！根据规划，洋湖新城和顺路两侧都有1... | 网友"UU008687"您好！您的留言已收悉。现将有关情况回复如下：您所反映的地点为洋湖新城... | 2019-01-16 00:00:00 | 16 days |

图 32 留言明细表

第二步，我们统计政府答复时间间隔分布，结果如下图所示：

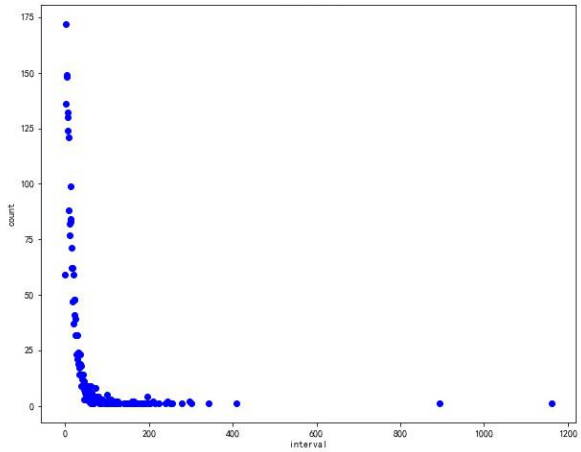


图 33 时间间隔分布图

由图可得知，答复时间间隔大多数聚集在 40 天之内，因此不同的时间间隔在评分中应该赋于不同的权重，随后我们计算不同的时间间隔在评分中所占的权重，我们设置每个时间间隔所占的权重为此时间间隔在总数据中出现的频率，计算出来的部分权重如下：

| | interval | count | count_weights |
|---|---|---|---|
| 0 | 0 days | 59 | 0.020952 |
| 1 | 1 days | 172 | 0.061080 |
| 2 | 2 days | 136 | 0.048295 |
| 3 | 3 days | 148 | 0.052557 |
| 4 | 4 days | 149 | 0.052912 |
| 5 | 5 days | 132 | 0.046875 |
| 6 | 6 days | 130 | 0.046165 |
| 7 | 7 days | 124 | 0.044034 |
| 8 | 8 days | 121 | 0.042969 |
| 9 | 9 days | 88 | 0.031250 |

图 34 时间间隔频率权重表

第三步，结合频率和十分制给分：当天回复的为 10 分，回复时间间隔在前 10%的为 9 分，回复时间间隔前 20%的给 8 分，依次类推对后面的时间间隔给分，在按照频率给分时，我们把 40 天之后回复的设置为 0 分，根据统计发现在 2816 条留言数据中，在 40 天之后答复的有 289 条，因此该方法可行[13]。各时间间隔对应

的得分如图所示：

| | interval | count | count_weights | score |
|---|---|---|---|---|
| 0 | 0 days | 59 | 0.020952 | 10 |
| 1 | 1 days | 172 | 0.061080 | 9 |
| 2 | 2 days | 136 | 0.048295 | 8 |
| 3 | 3 days | 148 | 0.052557 | 8 |
| 4 | 4 days | 149 | 0.052912 | 7 |
| 5 | 5 days | 132 | 0.046875 | 7 |
| 6 | 6 days | 130 | 0.046165 | 6 |
| 7 | 7 days | 124 | 0.044034 | 6 |
| 8 | 8 days | 121 | 0.042969 | 5 |
| 9 | 9 days | 88 | 0.031250 | 5 |
| 10 | 10 days | 82 | 0.029119 | 5 |
| 11 | 11 days | 77 | 0.027344 | 4 |
| 12 | 12 days | 83 | 0.029474 | 4 |
| 13 | 13 days | 99 | 0.035156 | 4 |
| 14 | 14 days | 84 | 0.029830 | 4 |
| 15 | 15 days | 71 | 0.025213 | 3 |
| 16 | 16 days | 62 | 0.022017 | 3 |
| 17 | 17 days | 62 | 0.022017 | 3 |
| 18 | 18 days | 47 | 0.016690 | 3 |
| 19 | 19 days | 59 | 0.020952 | 2 |
| 20 | 20 days | 37 | 0.013139 | 2 |

图 35 时间差得分表

最后保留及时性评分准则，用于后续评价体系的建立

## 5.4 对于答复可解释性和完整性的评价

### 5.4.1 可解释和完整性基本概念

对于政务服务而言，在答复意见时合理的解释和完整的解答是十分有必要的。可解释性是根据意见答复是否采用了某些依据，具有很强的客观性，受人为主观因素影响不大，可以作为评价政务水平的重要指标。完整性是语言表达上是否合

乎礼貌和某种行业行为规范，我们在查看数据时，发现比较完整的意见答复内容大多是形如“市民某某您好，关于你反映的某问题，现答复如下……感谢您”，诸如此类完整性强的留言可以使群众对政府留下积极的印象，更容易使群众认可政府的政务水平，因此也是评价政务水平的重要指标。

### 5.4.2 针对可解释性和完整性的处理步骤

第一步，总结出完整性和可解释性的代表词汇，以满分十分制赋予分数，每类词汇出现一次为 2 分。选取以下词汇的原因：完整性中我们选取的词汇分别在语句中起到称呼、问候语、回答问题的常用状语、用于回答问题的动词短语、礼貌用语等作用。可解释性中我们选取出现较多的依据用语、引用法律法规的状语、涉及回答依据的名词短语等词汇。具体如下图所示：

| 完整性 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 词汇1 | 分数 | 词汇2 | 分数 | 词汇3 | 分数 | 词汇4 | 分数 | 词汇5 | 分数 |
| 市民<br>网友 | 2 | 您好 | 2 | 关于<br>对于<br>有关<br>针对<br>反映 | 2 | 回复、<br>答复 | 2 | 感谢您 | 2 |
| 可解释性 | | | | | | | | | |
| 词汇1 | 分数 | 词汇2 | 分数 | 词汇3 | 分数 | 词汇4 | 分数 | 词汇5 | 分数 |
| 核实<br>经调查<br>经全面摸底<br>经咨询<br>了解<br>据了解<br>进行<br>经核实<br>经查<br>据查 | 2 | 目前<br>按照 | 2 | 根据<br>政策<br>法律 | 2 | 规定<br>下发<br>通知 | 2 | 由于<br>因为<br>因此<br>鉴于<br>所以 | 2 |

**图 36 完整性与可解释性赋分表**

第二步，采用正则表达式对答复内容中的文本进行匹配，首先将文本通过 compile 函数预编译为正则表达式对象，然后采用 search 方法进行匹配，每类词汇只匹配出一个即可给 2 分。部分答复意见的及时性、完整性、可解释性分数如下：

| | 留言主题 | 答复意见 | 及时性 | 完整性 | 可解释性 |
|---|---|---|---|---|---|
| 0 | A2区景春华苑物业管理有问题 | 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区景春花苑物业管理有问题"的调查核... | 3 | 10 | 4 |
| 1 | A3区潇楚南路洋湖段怎么还没修好? | 网友"A00023583"：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋... | 3 | 10 | 4 |
| 2 | 请加快提高A市民营幼儿园老师的待遇 | 市民同志：你好！您反映的"请加快提高民营幼儿园教师的待遇"的来信已收悉。现回复如下：为了改善... | 3 | 6 | 2 |
| 3 | 在A市买公寓能享受人才新政购房补贴吗? | 网友"A000110735"：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反... | 3 | 8 | 6 |
| 4 | 关于A市公交站点名称变更的建议 | 网友"A0009233"，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议"白竹坡... | 3 | 8 | 2 |
| 5 | A3区含浦镇马路卫生很差 | 网友"A00077538"：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、... | 1 | 10 | 2 |
| 6 | A3区教师村小区盼望早日安装电梯 | 网友"A000100804"：您好！针对您反映A3区教师村小区盼望早日安装电梯的问题,A3区... | 0 | 10 | 2 |
| 7 | 反映A5区东澜湾社区居民的集体民生诉求 | 网友"UU00812"您好！您的留言已收悉。现将有关情况回复如下：一、关于小区附近幼儿园的问... | 1 | 10 | 10 |
| 8 | 反映A市美麓阳光住宅楼无故停工以及质量问题 | 网友"UU008792"您好！您的留言已收悉。现将有关情况回复如下：据查，美麓阳光项目位于A... | 3 | 10 | 6 |
| 9 | 反映A市洋湖新城和顺路洋湖壹号小区路段公共绿化带的问题 | 网友"UU008687"您好！您的留言已收悉。现将有关情况回复如下：您所反映的地点为洋湖新城... | 3 | 10 | 4 |

图 37 部分答复意见得分

# 结语

本文主要运用数据挖掘领域中的文本分析包含的多种常见算法和模型，对政府政务中群众留言建立了分类模型和政务评价系统等一系列工作。

首先，我们依据原始数据的数据特点进行分析，针对每个属性制定了数据的预处理方案，尽可能使数据变得易于处理，减少工作量和提高计算效率。

进一步，在筛选了多种文本分类模型后，经过综合考虑，我们选择搭建了 LSTM 模型对文本进行分类处理。该模型与 RNN 模型相比，继承了其大部分特性，是一个优秀的变种模型，从结果来看，LSTM 非常适合解决文本分析的相关问题。针对模型处理结果，采用 F1-Score 对模型精度进行了评价,得分比较高，取得了比较良好的效果。不足之处在于没有做数据增强，略微影响模型精度。

再进一步，在第二问中，我们对数据进行基于层次的文本聚类，然后形成话题，选择评价指标后用熵权法进行评分从而对热点问题进行排序。其中，采用了相对客观的权重计算方法——熵权法使结果的客观性得到了保证，不足之处在于聚类效果仍有待提高。

最后，针对第三问，我们依然选择几个评价指标，然后建立用熵权法对政务水平进行打分的系统。综合来看，可行性较高但美中不足的是完整性和可解释性的词汇是人工选择的，没有指定一个好的自动评价体系，可能有略微误差。

# 参考文献


[1]蒲梅, 周枫, 周晶晶, et al. 基于加权 TextRank 的新闻关键事件主题句提取[J]. 计算机工程, 2017(8).

[2]张谦,高章敏,刘嘉勇. 基于 Word2vec 的微博短文本分类研究[J].信息网络安全,2017(01):57-62.

[3]赵明,杜会芳,董翠翠,陈长松.基于 word2vec 和 LSTM 的饮食健康文本分类研究[J]. 农业机械学报,2017,48(10):202-208.

[4]CSDN. 基于 LSTM 的中文文本多分类实战[EB/OL].

https://blog.csdn.net/weixin_42608414/article/details/89856566,2019.

[5]刘宁,陈凌云,熊文涛.基于文本挖掘的网络热点舆情分析——以问题疫苗事件为例[J].湖北工程学院学报,2019,39(06):60-64.

[6]贺玲, 吴玲达, 蔡益朝. Survey of Clustering Algorithms in Data Mining%数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007, 024(001):10-13.

[7]张谦,高章敏,刘嘉勇.基于 Word2vec 的微博短文本分类研究[J].信息网络全,2017(01):57-62.

[8]吴广建, 章剑林, 袁丁. 基于 K-means 的手肘法自动获取 K 值方法研究[J]. 软件, 2019, 040(005):167-170.

[9]刘建平. BIRCH 聚类算法原理[EB/OL].

https://www.cnblogs.com/pinard/p/6179132.html,2016.

[10]简书. 中文 NLP 笔记：7. 如何做中文短文本聚类[EB/OL].

https://www.jianshu.com/p/05634ae00c39，2019.02.02

[11]张国锋. 在文章聚类中话题热度排序的研究与实现[D].东华大学,2019.

[12]张祖宇. 政府网站反馈即时性调查研究[J]. 科技视界, 2013, 000(016):66.

[13]韩利, 梅强, 陆玉梅,等. AHP-模糊综合评价方法的分析与研究[J]. 中国安全科学学报, 2004(07):89-92+3.