# 基于“智慧政务”中的文本数据挖掘分析

**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。

本文针对“智慧政务”中的居民投诉建议文本评论数据，基于向量空间模型算法提取了文本关键词并我们采用了多种机器学习分类模型进行测试，从最终得到线性支持向量回归算法相对较优的结果，F1-Score 评价指标达 0.86。

在挖掘热点问题的前期处理上，使用了余弦相似度计算整理出文本相似的同类主题并加以筛选，通过在 SPSS 中建立基于因子分析法的热度评价指标模型，给出得分前五的主题样本作为 Top5 热点问题，分析比较了相关类问题的热度体现在各个指标上的具体表现。

为建立留言的答复意见的评价体系，我们定义了相关性、完整性、可解释性和及时性四个指标。答复意见和留言详情相关性的计算是基于 LDA 主题模型的中文编辑距离得到的，另外答复意见的可解释性使用了哈工大中文篇章关系的关联词表以及自定义的可解释性词典来判别。通过将这四项指标的得分相加得到某条答复意见的综合评分，分数越高，该答复的质量就越高，从而为决策者提供一个较为清晰完善的参考意见。

**关键词：TF-IDF 算法，因子分析法，LDA 主题模型，关联词表**

# 1 挖掘目标

本次建模针对“智慧政务”中的居民投诉建议文本评论数据，依次进行了数据预处理、关键词特征提取等基本操作后，通过线性支持向量回归模型测试，因子分析法以及 LDA 主题模型等多种数据挖掘模型，实现对政务文本的分类构建，热点问题深层次信息挖掘以及建立基于相关性、完整性、可解释性的答复问题评价体系。

# 2 总体流程

本次建模针对“智慧政务”中的居民投诉建议文本评论数据，首先进行了文本提取、数据预处理（包括数据清洗、中文分词、去停用词）等的基本操作，后续根据文本的不同类标签，通过基于向量空间模型算法提取了文本关键词。为确保分类效果，我们采用了多种分类模型进行测试，比较分析模型的优缺点。

在挖掘热点问题的前期处理上，使用了余弦相似度计算整理出文本相似的同类主题并加以筛选，通过在 SPSS 中建立基于因子分析法的热度评价指标模型，给出得分前五的主题样本作为 Top5 热点问题，分析比较相关类问题的热度体现在各个指标上的具体表现。

关于留言的答复意见的评价体系建立，我们定义相关性为构建的答复文本主题与留言主题之间的文本相似度，其计算是基于 LDA 主题模型的中文编辑距离得到的；定义完整性为有标准的开头句、结尾句，并且解决了留言中的问题；在可解释性方面，我们使用哈工大中文篇章关系的关联词表以及自定义的可解释性词典来判别答复意见是否具有可解释性。另外还加入了及时性的定义，即留言与答复时间间隔。设置以上四个指标的满分都为 1，通过将这四项指标的得分相加得到某条答复意见的综合评分，分数越接近 4，该答复的质量就越高。

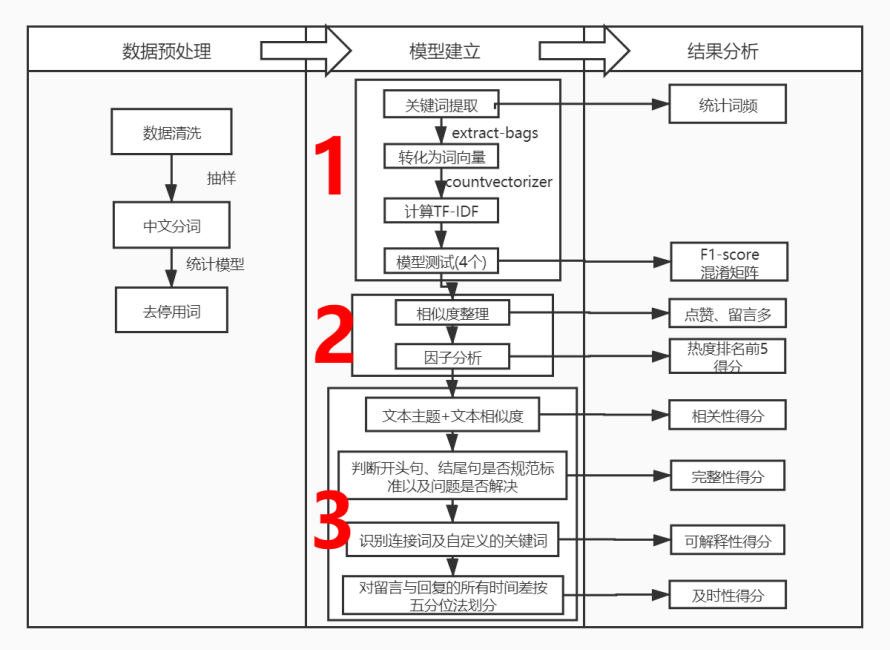


图 1 总流程图

# 3 数据预处理

## 3.1 数据介绍

附件 1 的数据为三级标签的详细分类，附件 2 和 3 给出了收集自互联网公开来源的群

众问政留言记录，其中附件 3 多包含了每条文本的点赞数和反对数。附件 4 则为相关部门对部分群众留言的答复意见。

通过浏览附件 2 中的文本数据，发现存在一些用户多次留言的情况，这对后面的热点问题挖掘有参考价值。在一级标签列中有七个标签，但数量不等，最少只有 600 多个样本，因此为了达到较好的分类效果就需要每类样本数都采用这个最小值。在留言详情列中，每条文本的中文字符数跨度较大，少至几十，多至几千，最多有 6212 个中文字符，在各类标签中的分布较为零散。

## 3.2 数据清洗

首先将附件 2 导入到 python 中，进行文本数据的预处理操作。因为文本中没有大量存在重复性的句段或词，所以这里没有进行压缩去词的操作，而是使用 drop_duplicates 语句进行简单去重。另外观察到留言详情列中存在大量包含字母与数字字符的时间，地点，车牌号等文本数据，为了精简文本信息，防止后续的关键词特征提取操作中出现大量的无关字符，需要对其进行正则化处理，去掉数字 0-9 和字母并存储在新的一列。

鉴于每一级类标签下还有不同的二级与三级标签，为增加样本的随机性，得到更好的分类效果，对七类标签分别使用 sample 函数均等抽取 500 例样本用于后续分类操作。

## 3.3 中文分词

在自然语言处理中，必不可少的是对一个段落或一个文章进行分词处理，对简化文本，提取特征起非常重要的作用。不同于西文文本，由于在中文中词与词之间没有明显的切分标记，这就需要选择一个单位来对文本其进行分割、词性标注等处理，需要运用一些算法来进行，即中文分词。中文分词由最开始的基于词典的方法发展到了基于统计语言模型的方法，并在逐渐地完善之中，取得了较好的研究效果。

### 3.3.1 正向/逆向最大匹配法（字典模型）

正向/逆向最大匹配算法是一种机械的分词方法，主要通过维护词典，在切分语句时，将语句的每个字符串与词表中的词逐一进行匹配，找到则切分，否则不予切分。[1]

这里需要有两个语料，一个是分词词典（也即是已经分词过的词典），另一个是需要被分词的文档。假定分词词典中的最长词有 x 个汉字符串，则用被处理文档的当前字符串中的前 x 个字作为匹配字段，查找字典。若此时分词词典中存在这样一个字符串，则匹配成功，而此时被匹配的字段切分出来。如果匹配失败，将匹配字段中的最后一个字去掉，对此时剩下的字串重新与分词词典进行匹配，如此下去直到匹配成功。也即是切分出一个词或剩余字串的长度为零为止，这个时候才是匹配了一轮，接着进行下一个 x 字字串的匹配，方法同上，直到文档被扫描完为止。以下给出最大逆向匹配的实例：

输入例句：S1="A 市公交线路的建议"；

定义：最大词长 MaxLen = 5；S2= ″ ″；分隔符 = “/”；

假设存在词表：…，A 市，公交线路，建议，…；

最大逆向匹配分词算法过程如下：

（1）S2=""；S1 不为空，从 S1 右边取出候选子串 W="线路的建议"；

（2）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W="路的建议"；

（3）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W="的建议"；

（4）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W="建议"；

（5）查词表，“建议”在词表中，将 W 加入到 S2 中，S2=″ 建议/″，并将 W 从 S1 中去掉，此时 S1="A 市公交线路的"；

（6）S1 不为空，于是从 S1 左边取出候选子串 W="公交线路的"；

（7）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W=″交线路的″；

（8）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W=″线路的″；

（9）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W=″路的″；

（10）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W=″的″，这 W 是单字，将 W 加入到 S2 中，S2=“ /的 /建议”，并将 W 从 S1 中去掉，此时 S1=″A 市公交线路″；

（11）S1 不为空，于是从 S1 左边取出候选子串 W=″市公交线路″；

（12）查词表，W 不在词表中，将 W 最左边一个字去掉，得到 W=″公交线路″；

（13）查词表，“公交线路”在词表中，将 W 加入到 S2 中，S2=“ 公交线路/ 的/ 建议/”，并将 W 从 S1 中去掉，此时 S1=″A 市″；

（14）S1 不为空，于是从 S1 左边取出候选子串 W=″A 市″；

（15）查词表，“A 市”在词表中，将 W 加入到 S2 中，S2=“A 市/ 公交线路/ 的/ 建议/”，并将 W 从 S1 中去掉，此时 S1=″″；

（16）S1 为空，输出 S2 作为分词结果，分词过程结束。

因为中文比较复杂以及中文的特殊性，逆向最大匹配大多时候往往会比正向要准确。但是匹配法也有一些不足之处，存在一些文本具有切分歧义的情况，比如：“不按核定营运线路及时间的客运车”可能被错误划分为“不按/核定/营运/线路/及时/间/的/客运/车”，从而导致分词效果很差。

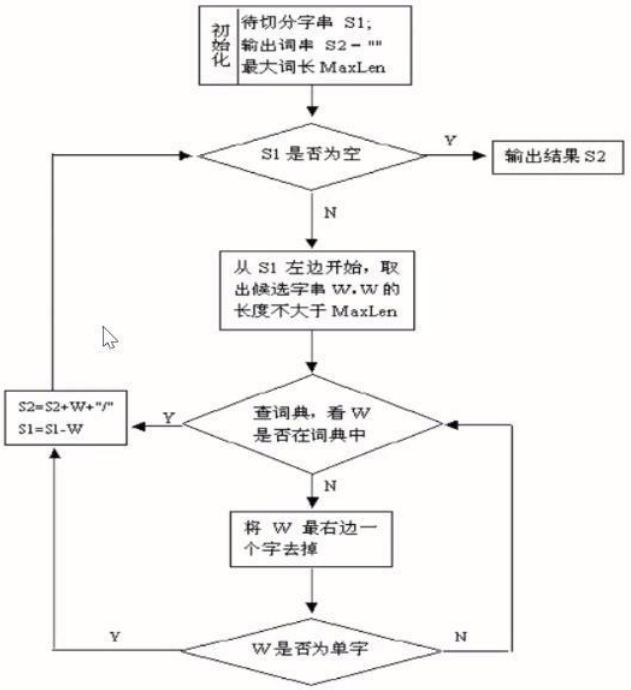


图 2 最大匹配法算法流程

### 3.3.2 最大概率法（统计模型）

#### 3.2.2.1 基本思想

对于分词来说，经过大量的文本训练的统计模型的效率比字典模型的方法高，同时效果也要好，不仅考虑了文字词语出现的频率信息，同时考虑了上下文的语境，对效率的提升是十分显著的。

这种统计模型所采用的方法是最大概率法，其基本思想是一个待切分的汉字串可能包含多种分词结果，将其中概率最大的那个作为该字符串的分词结果。[2]

$$r = \underbrace{\arg\max}_{i} P(W_{i1}, W_{i2}, \ldots W_{in_i})$$

由于涉及到$n_i$个分词的联合分布，故令分词结果及其概率表示为

$$S = W_1, W_2, \ldots, W_k$$

$$P(S) = P(W_1, W_2, \ldots, W_k) = p(W_1)P(W_2|W_1)\ldots P(W_k|W_1, W_2, \ldots, W_{k-1})$$

如果一个词的出现依赖于前面 N-1 个词，称这种模型为 N 元语言模型（N-gram）。在实践中用的最多的就是二元语言模型和三元语言模型，高于三元用的非常少，因为这样导致计算效率很低，时间复杂度高，精度提升很有限。

在 NLP 中，通常我们用到齐次马尔科夫假设，即每一个分词出现的概率只与前面一个分词相关，与其他分词无关。上式的概率公式可表示为

$$P(S) = P(W_1, W_2, \ldots, W_k) = p(W_1)P(W_2|W_1)\ldots P(W_k|W_{k-1})$$

通过标准语料库可以计算所有分词的二元条件概率：

$$P(\omega_2|\omega_1) = \frac{P(\omega_1, \omega_2)}{P(\omega_1)} \approx \frac{freq(\omega_1, \omega_2)}{freq(\omega_1)}$$

其中 $freq(\omega_1, \omega_2)$ 表示 $\omega_1, \omega_2$ 在语料库中相邻一起出现的频数，因此通过以上方法我们即可求出各种分词组合的联合分布概率，找到最大概率对应的分词即为最优分词。

#### 3.3.2.2 jieba 库分词

本文采用 python 中最常见的 jieba 库进行分词。

jieba 库分词采用的是基于统计的分词方法，首先给定大量已经分好词的文本，利用机器学习的方法，学习分词规律，然后保存训练好的模型，从而实现对新的文本的分词。主要的统计模型包括 N 元语言模型 N-gram，隐马尔可夫模型 HMM，最大熵模型 ME，条件随机场模型 CRF 等。这里列举一种 jieba 中文分词算法流程：

(1) 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）；

(2) 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；

(3) 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

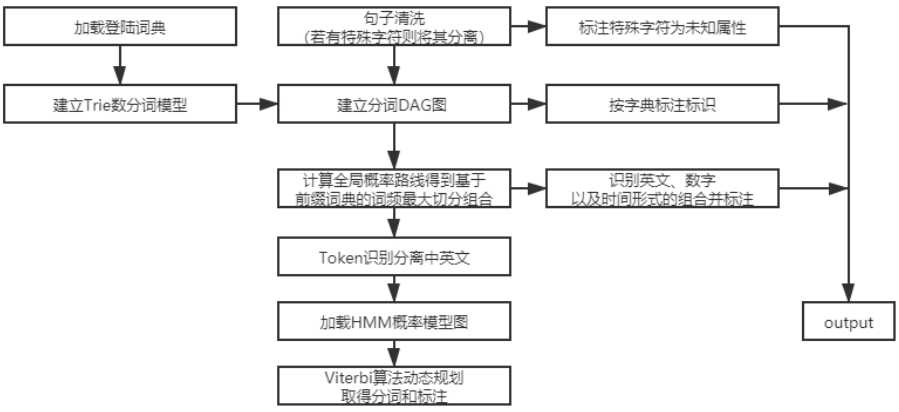


图 3 jieba 分词算法流程

### 3.4 去停用词

在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据之前或之后会自动过滤掉某些字或词，这些字或词被称为停用词。这些停用词都是非自动化生成的，生成后的停用词会形成一个停用词表。

停用词有两个特征，一是出现频率很高，大量文本中均含有该词；二是包含信息量较小，对文本标识与分类没有什么意义。实际上，并没有一个明确的停用词表能够适用于所有的工具，所以有时需要人工建立特定的针对不同文本数据特征的停用词表。

在附件 2 的留言详情列数据中，发现一是文本中职称和敬词较多，如“市长”，“处长”，“书记”，“尊敬的领导”，“同志”，“您好”，“谢谢”，“请求”，“感激不尽”等；二是时间较多，如“今晚”，“昨晚”，“昨天”，“凌晨”，“今天”，以及和年月日有关的名词；三是一些感官类的疑问词、形容词和语气副词等，如“全是”，“有种”，“本来”，“听说”，“想想”，“怎么回事”。这几类词在文本中出现的频率非常高，如不对其进行处理，在关键词特征提取中将大量重复出现，从而影响结果的分析。

表 1 部分停用词示例

| 职称和敬词类 | 时间类 | 感官类 |
| --- | --- | --- |
| 市长 处长 书记 | 年 月 日 | 全是 正好 好多 |
| 尊敬的 领导 同志 | 周一~周日 | 本来 迟迟 特别 |
| 您好 你好 你们好 | 今晚 昨晚 昨天 凌晨 今天 | 听说 样子 感觉 发现 |
| 谢谢 感谢 感激不尽 | 上午 下午 晚上 现在 | 希望 想想 想要 建议 |
| 请求 可不可以 | …… | …… |
| …… | | |

## 4 基于 TF-IDF 算法权重修正的关键词提取处理群众留言分类

### 4.1 词袋模型与向量化[3]

词袋模型假设我们不考虑文本中词与词之间的上下文关系，仅仅只考虑所有词的权重。而权重与词在文本中出现的频率有关。

词袋模型首先会进行分词，在分词之后，通过统计每个词在文本中出现的次数，我们就可以得到该文本基于词的特征，如果将各个文本样本的这些词与对应的词频放在一起，即向量化。向量化完毕后一般也会使用 TF-IDF 进行特征的权重修正，再将特征进行标准化，就可以将数据带入机器学习模型中计算。

向量化的 fitting 过程中采用 countvectorizer 函数，将根据语料库中的词频排序选出前 n 个词。一个可选的参数 minDF 也影响 fitting 过程，它指定词汇表中的词语在文档中最少出现的次数。另一个可选的二值参数控制输出向量，如果设置为真那么所有非零的计数为 1。这对于二值型离散概率模型非常有用。由于大部分文本都只会用词汇表中很少一部分的词，因此词向量中有大量的 0，也就是说词向量是稀疏的。因此在实际应用中一般使用稀疏矩阵来存储。

### 4.2 TF-IDF 算法关键词权重修正

TF-IDF 是一种统计方法，用以评估一字词对于一个语料库中的其中一份文档的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降，按词语重要性从中提取特定数量的词语组成这篇文章的关键词集合。一般来说，文档中的高频词应具有表征此文档较高的权重，或者是该词在全语料库中也是较高

的文档频率词。

TF（Term frequency）即关键词词频，是指一篇文档中关键词出现的频率

$$TF = N/M$$

其中 $N$ 代表词在某文档中的频次，$M$ 代表该文档的词数。

另外定义 IDF（Inverse document frequency）逆向文本频率，用于衡量关键词的指数

$$IDF = \log(D/D_w)$$

$D$ 为总文档数，$D_w$ 为出现了该词的文档数。最终将这两个指标相乘得到

$$TF - IDF = TF \times IDF$$

## 4.3 jieba 库提取关键词

首先导入先前处理好的分词列表，划分训练集测试集比例 7:3，再调用 extract_tags 函数端口进行关键词提取操作初始化。

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer,
TfidfVectorizer
keywords=jieba.analyse.extract_tags(i,topK=20)#提取关键词
……
countVectorizer = CountVectorizer()#调用
data_tr = countVectorizer.fit_transform(data_tr)#先生成词向量
X_tr = TfidfTransformer().fit_transform(data_tr.toarray()).toarray()#再计算 TF-IDF
```

图 4 jieba 库部分函数端口调用代码

在提取出关键词后对七个标签类的分别进行词频统计，在所有的样本中整理出每类前 20 个关键词如表 2 所示。对比七类标签关键词可以发现已经对样本提取出了明显的特征，第一类标签是城乡建设，关于“房子”，“物业”的词居多，体现出了与住户息息相关的留言反馈；第二类为环境保护，用户多反映在噪音、污水、周边环境问题上；交通运输上主要是用户反映的士司机乱收费现象，集中在路面问题；教育文体主要是说学校以及中小学学生的教育问题，可能涉及到了补课考试等担忧；劳动和社会保障提到了关于社保、退休，缴纳养老金等话题；商贸旅游中多数是投诉小区景区乱收费，开发商垄断的现象；卫生计生则是医患关系及独生子女的计生问题。

并且还可以发现七个标签的关键词里均出现了 A 市，说明用户所反映的 A 市存在的问题最多，其次为 K 市，另外在教育文体类中还出现了 B 市和 C 市，这是值得当权者注意的地方。

```
for i in data_key['key']:
    for j in i:
        if j not in word_fre.keys():
            word_fre[j] = 1
        else:
            word_fre[j] += 1
```

图 5 关键词词频统计代码

表 2 七个标签类 Top20 关键词

| 1 | 词频 | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 小区 | 85 | 污染 | 123 | 出租车 | 85 | 学校 | 180 | 社保 | 89 | A 市 | 70 | 医院 | 145 |
| 业主 | 81 | 环保局 | 110 | 快递 | 64 | 教育局 | 140 | 工资 | 71 | 电梯 | 68 | 医生 | 93 |
| 开发商 | 49 | 居民 | 86 | 司机 | 58 | 学生 | 123 | 职工 | 64 | 小区 | 59 | 政策 | 60 |
| A 市 | 46 | 排放 | 54 | 的士 | 54 | 教师 | 118 | 退休 | 63 | 传销 | 56 | 生育 | 54 |
| 房屋 | 43 | 噪音 | 51 | A 市 | 52 | 老师 | 94 | 工作 | 62 | 业主 | 45 | 小孩 | 53 |
| 居民 | 43 | 污水 | 51 | 收费 | 49 | 家长 | 89 | 单位 | 60 | 旅游 | 37 | 办理 | 43 |
| K 市 | 34 | 环保 | 51 | 车辆 | 38 | 教育 | 87 | A 市 | 51 | 景区 | 34 | 患者 | 40 |
| 房子 | 33 | 周边 | 42 | 公司 | 36 | 小学 | 62 | 人员 | 49 | 收费 | 31 | 计生办 | 37 |
| 住房 | 33 | 投诉 | 41 | 邮政 | 35 | 孩子 | 58 | 医保 | 49 | 垄断 | 31 | 独生子女 | 36 |
| 规划 | 31 | 环保部门 | 41 | 交通 | 31 | A 市 | 49 | 办理 | 46 | 投诉 | 29 | 准生证 | 36 |
| 房产证 | 27 | 生产 | 41 | 收取 | 30 | 补课 | 46 | 缴纳 | 43 | 物业 | 28 | 二胎 | 34 |
| 住户 | 27 | 环境 | 37 | 出行 | 28 | 中学 | 41 | 社局 | 41 | 收取 | 26 | 家庭 | 34 |
| 办理 | 24 | 环评 | 36 | 客运 | 27 | 小孩 | 40 | 政策 | 40 | 价格 | 24 | 户口 | 33 |
| 城管 | 23 | K 市 | 35 | 打表 | 25 | C 市 | 37 | 养老保险 | 40 | 故障 | 22 | 检查 | 33 |
| 交房 | 22 | A 市 | 33 | 费用 | 24 | 收费 | 31 | 公司 | 38 | 公司 | 21 | 卫生院 | 32 |
| 公积金 | 20 | 小区 | 30 | 乱收费 | 23 | 幼儿园 | 29 | 企业 | 36 | 开发商 | 21 | K 市 | 32 |
| 改造 | 18 | 影响 | 27 | 营运 | 22 | 工作 | 28 | 劳动 | 33 | 人员 | 19 | 计划生育 | 32 |
| 物业 | 18 | 刺鼻 | 26 | 路面 | 21 | 考试 | 28 | 享受 | 32 | 交房 | 19 | 计生 | 32 |
| 公园 | 17 | 排污 | 26 | 投诉 | 21 | 培训 | 27 | 劳动法 | 31 | 乱收费 | 19 | 卫生局 | 31 |
| 廉租房 | 17 | 养猪场 | 24 | K 市 | 20 | B 市 | 26 | 员工 | 30 | 检疫 | 18 | A 市 | 31 |

## 4.4 模型测试

在得到关键词向量矩阵后，需要使用机器学习分类模型对其进行分类效果的评价与可视化。此处调用了四个模型进行分类测试：

- 随机森林，ensemble.forest.RandomForestClassifier；
- 线性支持向量回归，svm.classes.LinearSVC；
- 朴素贝叶斯（多项式），naive_bayes.MultinomialNB；
- 线性回归（逻辑回归），linear_model.logistic.LogisticRegression。

首先可以画出箱体图比较出不同模型的准确率，从图 6 上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器，有若干个子分类器组合而成。一般来说集成分类器不适合处理高维文本数据，因为词向量矩阵有太多的特征值，使得集成分类器难以应付。另外三个分类器的平均准确率都在 82.5%以上，其中线性 SVC 的准确率最高。

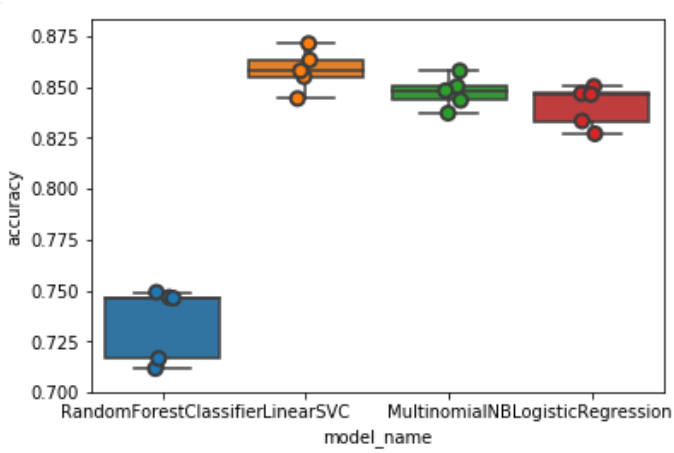


图 6 四模型准确率箱线图

进一步将表现最佳的模型的分类效果可视化，通过计算并绘制模型混淆矩阵（Confusion Matrix），可以直观地了解模型在哪一类样本里面表现得不是很好。这是一种在深度学习中常用的辅助工具，越接近对角矩阵分类效果越好，从图 7 中的数字块也可以看到，对角线颜色越浅且非对角线颜色越深分类效果越好。

横向比较七个标签的分类结果，明显观察到商贸旅游的分类结果最好，仅有 9 个错判；劳动和社会保障的分类结果最差，共错判了 51 个样本，其中最多错判了 16 个样本标签为卫生计生。

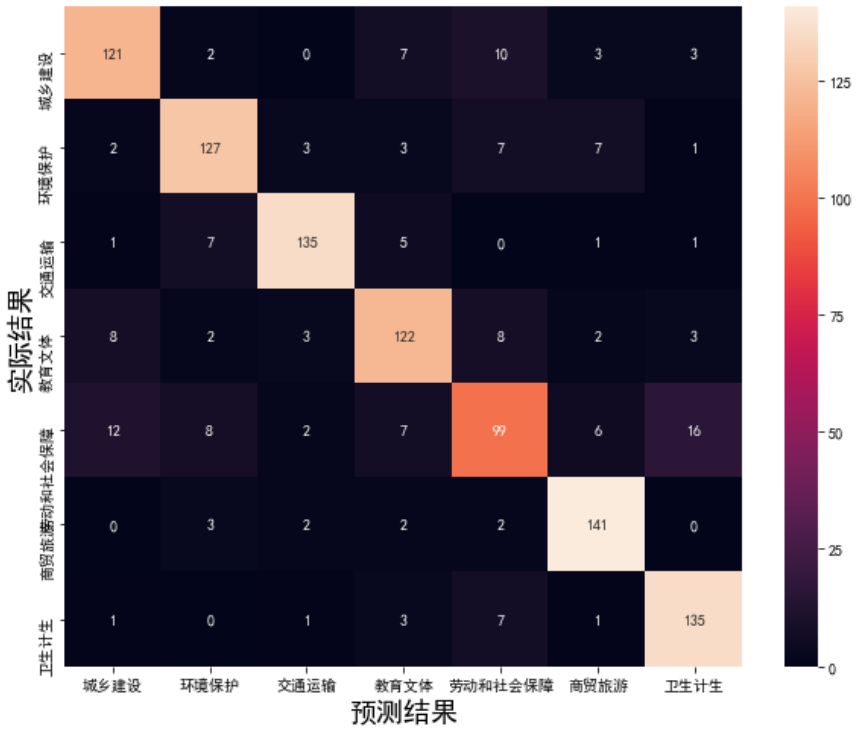


图 7 线性 SVC 混淆矩阵

表 3 线性 SVC 分类效果分析报告

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 城乡建设 | 0.85 | 0.88 | 0.86 | 161 |
| 环境保护 | 0.85 | 0.90 | 0.87 | 164 |
| 交通运输 | 0.91 | 0.91 | 0.91 | 164 |

| | | | | |
|---|---|---|---|---|
| 教育文体 | 0.84 | 0.80 | 0.82 | 163 |
| 劳动和社会保障 | 0.76 | 0.73 | 0.75 | 165 |
| 商贸旅游 | 0.89 | 0.85 | 0.87 | 165 |
| 卫生计生 | 0.91 | 0.95 | 0.93 | 164 |
| avg/total | 0.86 | 0.86 | 0.86 | 1146 |

# 5 基于相似度分类与因子分析法的热点问题挖掘

## 5.1 相似度计算

### 5.1.1 余弦距离的基本思想

在分析两个向量之间的相似性时，通常会采用余弦相似度来表示。对于两个文本之间的相似性，可以将其向量化，再计算他的余弦距离。余弦距离可以避免文本长度的不同而导致距离过大，它考虑的是两个文本的特征向量之间的夹角。

对于一个向量空间中的二维向量 a$(x_1, y_1)$和 b（$x_2, y_2$），夹角为$\theta$，如图 8 所示

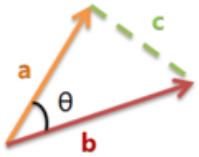


图 8

它的余弦相似度定义为

$$cos\theta = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

余弦相似度的范围是[-1,1]，余弦距离的定义公式为

$$dist(a,b) = 1 - cos\theta$$

这样余弦距离的范围为[0,2]

对于两个多维的向量 a$(a_1, a_2, \dots, a_n)$和 b$(b_1, b_2, \dots, b_n)$，其余弦距离为

$$dist(a,b) = 1 - \frac{\sum_{i=1}^{n}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

### 5.1.2 制作语料库

在文本相似度计算之前，对附件 3 先按照地点和人群进行分类，由于留言详情的冗余内容较多，因此对留言详情的文本内容调用 jieba 库的 extract_tags 函数进行关键词的提炼，同时对留言主进行分词和去停用词处理。

基于经过预处理分词文本，调用 gensim 库中的 corpora.Dictionary 方法获取目标文本的词袋，建立词典，词袋中用数字对所有词进行了编号。对于词典，使用 dictionary 中的 doc2bow 制作语料库，语料库是一组向量，向量中的元素是一个二元组(x,y)，x 表示对应分词的编号，y 表示对应分词出现的频次。并且把测试文本也转化为二元组向量。

### 5.1.3 相似度分析

使用 TF-IDF 模型处理语料库，得到每个词的索引以及 TF-IDF 值，gensim 库中的 similarities.SparseMatrixSimilarity 是利用余弦距离来计算文本相似度的函数，调用该函数计算每个目标文本与测试文本的相似度，并根据相似度进行排序。在这里选择文本相似

度大于 0.3 的文本作为同类留言，并记录留言的索引，由于可能会有疏漏，对于每条留言，当它与其它留言的同类留言中有两条以上相同的索引，就取他们的并集，以此来对不同的地点和人群的留言进行归类。

基于挖掘排名前 5 的热点问题的要求，本文筛选出了留言数量大于 3 的问题以及点赞数量大于 25 的问题，如表 4，用以后续通过计算相关的量化指标来建立热度评价体系，并给热点问题排名。

表 4 留言数和点赞数较多的 60 个问题主题

| 留言多（36） | 点赞多（24） |
|---|---|
| 巴罗比幼儿园无证办学问题 | 丽发新城断水问题 |
| 白云路断头路拉通时间问题 | 绿地城际空间站建筑质量差 |
| 财富名园空地建设问题 | 假期开放泉星小学田径场 |
| 楚行一卡通使用问题 | 泉星物流园油烟排放管道问题 |
| 楚仪路车辆停放与卫生问题 | 泉星物流园摆放游戏赌博机 |
| 西地省富惠天下诈骗案 | 金毛湾配套入学的问题 |
| 恒基凯旋门万婴格林幼儿园的普惠问题 | 加大东六线榔梨段拆迁力度 |
| 辉煌国际城商铺违法处置做餐饮 | 泉星生鲜市场装修工程未公开招标 |
| 经济学院强制学生外出实习 | 长房云时代存在房屋质量问题 |
| 漓楚路街道灯光问题 | “和包”支付问题 |
| 丽发新城搅拌站扰民 | 三一大道全线快速化改造启动时间问题 |
| A 市魅力之城油烟扰民 | 泉塘昌和商业中心以南的规划问题 |
| A 市魅力之城油烟扰民墙板开裂 | 星沙大道摩的飙行问题 |
| 人才住房补贴问题 | 建议外迁京港澳高速城区段至远郊 |
| 反对 A7 县诺亚山林小区门口设置医院 | A6 区事业单位工作人员的工龄计算问题 |
| 星沙镇星沙派小区违法违建 | 山千制药机械股份有限公司拖欠薪水 |
| A7 县星沙中贸城拒绝退还购房资金 | 建议将地铁 7 号线南延至 A 市生态动物园 |
| A7 县楚郡未来实验学校调学费 | 万润滨江天著毛坯房存在质量问题 |
| 扬帆小区夜市噪音及安全问题 | A 市地铁四号线北延线同心路站位置问题 |
| 伊景园滨河院捆绑销售车位 | 星沙滨湖路以南特立路以北土地出让问题 |
| A 市地铁 1 号线北延线开工时间询问 | 加快修建 A 市南横线的建议 |
| A 市地铁 3 号线松雅西地省站安全问题 | 建议 A 市收回东六路恒天九五工厂地块 |
| A 市地铁 6 号线施工扰民 | A 市东四线以西新安路以南的规划问题 |
| A 市地铁扫码问题 | 郝家坪小学扩建问题 |
| 建议增设 A2 区南塘城轨公交站 | |
| 对 A 市公交线路的建议 | |
| 建议完善 A 市公交站房设置工作 | |
| 长赣高铁噪音影响绿地海外滩小区居民 | |
| 泉星公园规划问题 | |
| 58 车贷诈骗案件 | |
| 月亮岛路沿线架设 110kv 高压线杆问题 | |
| 五矿万境 K9 县房屋质量问题 | |
| 长永高速工程进展询问 | |
| A 市市直学校教师招聘考试不公平 | |
| 督促政府落实退休教师补贴 | |
| 咨询反映临聘教师待遇问题 | |

## 5.2 因子分析法构建热度评价体系

### 5.2.1 指标的选取与计算

留言热点问题本质上是指某一条或某一类留言引起的用户关注和讨论的热烈程度[4]，如何定量地来描述热点问题是一个核心难点，这需要根据已知数据和相关文献来定义相关指标进行计算和建立评价体系，如表 5 所示。

首先直观地来看，问题的热度受留言数量的影响是最直接的，所以第一个维度指标定义为数量特征热度影响力。除留言数量外还包括了每类热点问题的不同用户数，从流量的角度看，用户数量的多少也体现出了问题的热度的大小。关于第一维度具体的指标量化，我们定义了留言率 1 和留言率 2，$n$ 为留言条数

$$L_1 = n/T$$

留言 1 代表该类问题的留言数和留言平台开发天数（即附件 3 留言最早与最晚时间差）的比值，计算得$T = 934$。

$$L_2 = n/L$$

留言 2 代表该类问题的留言数和平台总留言量的比值，$L = 4326$。

第二个维度是时空特征热度影响力，包括留言信息充实度、留言出现的时频和时长。从空间上看是内容所占留言平台的比重，定义为充实度

$$F = m_1 + m_2 + \ldots + m_n / Mn$$

其中$m_i$为第$i$条留言的字数，$M$为本类留言中的一条最多字数。时间上定义两个指标，时长和时频，时长即为该类问题最晚留言时间和最早留言时间的时间差$t$，以天计，时频为时长与留言数的比值

$$t = t_{晚} - t_{早}$$

$$P = t/n$$

第三个维度是受众情绪特征热度影响力，表现为点赞与反对相关的指标，定义点赞反对率为点赞数$up$加反对数$down$与每类留言数量的比值大小

$$E = up + down/n$$

表 5 热度评价体系指标

| | 一级指标 | 二级指标 | 指标内涵 |
|---|---|---|---|
| 留言热度评价指标 | 数量特征影响力 | 留言率 1 | 该类问题的留言数和留言平台开发天数（附件三留言最早与最晚时间差）的比率 |
| | | 留言率 2 | 该类问题的留言数和留言用户数的比率 |
| | | 用户数 | 该类问题的留言用户数 |
| | 内容特征影响力 | 留言信息充实度 | $\frac{m_1 + m_2 + \ldots + m_n}{Mn}$，其中$n$表示某类问题的留言数 |

| | | | |
|---|---|---|---|
| | | | $M$ 为本类留言中的一条最多字数<br>$m_i$ 表示第 $i$ 条留言的字数 |
| | | 留言出现时频 | 该类问题最早留言时间和最晚留言的时间差与留言数的比率 |
| | | 时长 | 该类问题最早留言时间和最晚留言的时间差 |
| | 受众特征影响力 | 点赞反对率 | 该类问题点赞反对数（累加）与该类问题留言用户数的比率 |

### 5.2.2 因子分析法原理及实现

#### 5.2.2.1 原理[5]

因子分析是一种降维、简化数据的技术，通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并抽取少数几个变量来表示基本的数据结构。抽取出的变量被称作“因子”，能反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而因子一般是不可观测的潜在变量。

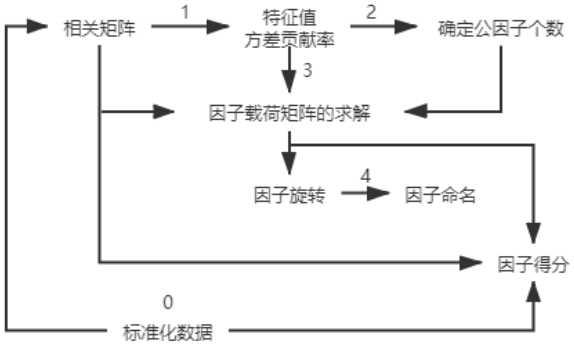


图 9 因子分析流程图

分析变量之间往往存在相关性，因子分析法可通过对多个变量的相关系数矩阵的研究，找出同时影响或支配所有变量的共性因子。每个变量都可表示成公共因子的线性函数与特殊因子之和，即

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots a_{im}F_m + \varepsilon_i, \quad (i = 1,2,\cdots,p)$$

式中的F1,F2,...,F_m为公共因子，e_i为X_i的特殊因子，该模型可用矩阵表示为X=AF+e，满足以下条件

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \quad X = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

- $M \le p$
- 公共因子与特殊因子不相关
- 公共因子互不相关且方差 $D(F) = 1$
- 特殊因子互不相关（不要求相等）

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}=(A_1,A_2,\cdots,A_m)$$

$a_{ij}$ 称为因子载荷，是第 $i$ 个变量在第 $j$ 个因子上的负荷，若把变量 $X_i$ 看成 $m$ 维空间中的一个点，则 $a_{ij}$ 表示它在坐标轴 $F_j$ 上的投影，矩阵 $A$ 称为因子载荷矩阵。

求公共因子的核心是求因子载荷矩阵，可通过变量 $X$ 的协方差矩阵 $\Sigma$ 分解式

$$\begin{aligned} D(X)&=D(AF+\varepsilon)=E\left[(AF+\varepsilon)(AF+\varepsilon)'\right] \\ &=AE(FF')A'+AE(F\varepsilon')+E(\varepsilon F')A'+E(\varepsilon\varepsilon') \\ &=AD(F)A'+D(\varepsilon) \end{aligned}$$

由上述满足的条件知，$\Sigma=AA'+D_\varepsilon$，如果 $X$ 经过了标准化，则 $\Sigma$ 为相关矩阵 $R=(\rho_{ij})$，即 $R=AA'+D_\varepsilon$.

对 $A$ 的估计方法较多，常用的是主轴因子法。令 $R^*=R-D_\varepsilon=AA'$，则称 $R^*$ 为 $X$ 的约相关阵。$R^*$ 中的主对角线的元素是 $h_i^{\,2}$，其中 $h_i^{\,2}=\sum_{i=1}^{m}a_{ij}^{\,2}$，非对角线的元素和 $R$ 中完全一样。记 $R^*=(\rho_{ij}^{\,*})_{p*p}$，则有

$$r_{ij}^{\,*}=\sum_{k=1}^{m}a_{ik}a_{jk}=\begin{cases}\sigma_{ij} & i\neq j \\ \sigma_{ii}-\sigma_i^{\,2} & i=j\end{cases}\quad i,j=1,2,\cdots,p$$

这里 $A$ 的解不唯一，这使得第一公共因子 $F_1$ 对 $X$ 的贡献 $g_1^{\,2}=\sum_{i=1}^{m}a_{i1}^{\,2}$ 达到最大，F2 的贡献次之，$F_m$ 最小，即相应的贡献按大小顺序排列。

$$g_t^{\,2}=\lambda_t^{\,*},\ \ A_t=\sqrt{\lambda_t^{\,*}}t_t^{\,*},\ \ t=1,2,\cdots,m$$

其中，$\lambda_t^{\,*}$ 为约相关阵 $R^*$ 的第 $t$ 大特征根，$t_t^{\,*}$ 为相应的单位特征向量，故求得载荷矩阵

$$\begin{aligned} A&=\left(\sqrt{\lambda_1^{\,*}}t_1^{\,*}\quad \sqrt{\lambda_2^{\,*}}t_2^{\,*}\quad \cdots\quad \sqrt{\lambda_m^{\,*}}t_m^{\,*}\right) \\ &=\left(t_1^{\,*}\quad t_2^{\,*}\quad \cdots\quad t_m^{\,*}\right)\begin{bmatrix}\sqrt{\lambda_1^{\,*}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_m^{\,*}}\end{bmatrix} \end{aligned}$$

### 5.2.2.2 检验及实现

因子分析的前提是观测变量之间具有相关性。如果相关性较低，则它们可能不会共享因子，只有相关性较高时，才适合做因子分析。

通过计算 KMO 和显著性得到 KMO=0.644>0.6，sig=0<0.01，如图 10 所示。说明模型参数通过了检验，变量之间存在一定的相关性，且在 0%的显著性水平下拒绝原假设。

**KMO 和巴特利特检验**

| | | |
|---|---|---|
| KMO 取样适切性量数。 | | .644 |
| 巴特利特球形度检验 | 近似卡方 | 36.766 |
| | 自由度 | 10 |
| | 显著性 | .000 |

图 10　KMO 与显著性检验结果

对公共因子进行提取，得到三个主成分，分别贡献方差的 40.075%，27.082%和 14.340%，使得总荷载达到 81.496%，如图 11 所示。

| | 初始特征值 | | | 提取载荷平方和 | | | 旋转载荷平方和 | | |
|---|---|---|---|---|---|---|---|---|---|
| 成分 | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % |
| 1 | 2.805 | 40.075 | 40.075 | 2.805 | 40.075 | 40.075 | 2.798 | 39.978 | 39.978 |
| 2 | 1.896 | 27.082 | 67.156 | 1.896 | 27.082 | 67.156 | 1.900 | 27.149 | 67.128 |
| 3 | 1.004 | 14.340 | 81.496 | 1.004 | 14.340 | 81.496 | 1.006 | 14.369 | 81.496 |
| 4 | .669 | 9.554 | 91.050 | | | | | | |
| 5 | .373 | 5.335 | 96.385 | | | | | | |
| 6 | .235 | 3.351 | 99.737 | | | | | | |
| 7 | .018 | .263 | 100.000 | | | | | | |

提取方法：主成分分析法。

图 11 公共因子提取及总方差解释

通过计算成分得分矩阵，如图 12，选取出每个因子中系数绝对值前三的变量作为主要因子，即对评价体系起重要作用的因子，分级与前面所述三个维度指标相同，即与预期指标分级相符：

- 数量特征热度影响力：留言率 1、留言率 2、用户数；
- 时空特征热度影响力：充实度、时长、时频；
- 受众情绪特征热度影响力：点赞反对率。

| | 成分 | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Zscore(时频) | -.118 | .440 | -.019 |
| Zscore(时长) | .013 | .487 | -.046 |
| Zscore(充实度) | -.177 | -.294 | -.080 |
| Zscore(留言率1) | .346 | .006 | .018 |
| Zscore(留言率2) | .276 | -.099 | -.071 |
| Zscore(用户数) | .342 | .014 | .024 |
| Zscore(点赞反对准) | .001 | -.019 | .990 |

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
组件得分。

图 12 各指标成分得分系数矩阵

## 5.3 热点问题具体留言分析

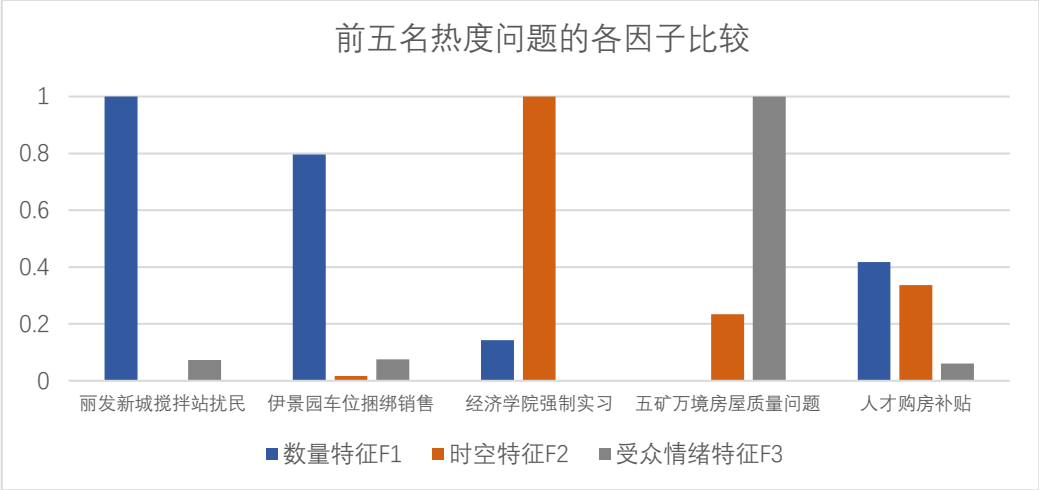


图 13 热度前 5 名问题各因子比较结果

对热度前 5 名的问题的三个标准化后的因子进行比较，如图 13，可以看到丽发新城搅拌站扰民问题和伊景园车位捆绑销售问题的热度来源主要是数量特征，说明这两类问题的留言数量以及留言用户数量众多，而时空特征的热度贡献最小；而对经济学院强制学生实习问题的热度贡献最大的是时空特征影响力，说明该问题留言内容充实度高，留言时间跨度很长，留言频率较高，反映出该问题的热度持续时间长，可能是问题迟迟得不到解决。

对于五矿万境 K9 县的房屋质量问题，热度来源主要是受众情绪特征热度影响力，而数量特征影响力的贡献率很小，说明该问题的点赞数量大，反映出用户对于该问题的关注度很高，问题的严重性比较大；与前面四个问题不同，人才购房补贴问题的热度来源相对比较平均，数量特征影响力和时空特征影响力的热度贡献相对比较大，体现了人才购房补贴在社会上的广泛关注度。

# 6 答复意见的评价

针对每条留言的答复意见，其质量效果可以由以下几个方面进行评价。一是相关性。该答复意见是准确地回复了对应的留言问题，还是答非所问。我们对答复意见进行文本主题的提取，并与留言主题进行相似度计算，得到相关性得分。二是完整性。通过判断该答复意见的开头句，结尾句是否符合某种规范标准，比如开头有问候语，结尾有感谢语和答复日期，以及答复是否解决了留言问题，从而得到完整性得分。三是可解释性。通过判断该答复中是否有背景调查、所做决定是否有法规等依据，以及识别连接词，特别是因果类的连接词，可以发现该答复是否有逻辑依据，从而得到可解释性得分。四是及时性。留言与答复时间间隔越小，说明答复越及时，及时性得分越高。

我们设置以上四个指标的满分都为 1，通过将这四项得分相加可以得到某条答复意见的综合评分，分数越接近 4，该答复的质量就越高。

## 6.1 答复意见的相关性

### 6.1.1 LDA 主题模型的应用

在评价系统中相关性满分为 1 分。由于使用 doc2vec 对留言详情和答复意见这两个文本直接做计算文本相似度的时间复杂度较大，并且为了弥补文本相似度计算对完全相同的词语敏感，对表达相同领域但不相同的词语不敏感的特征，我们首先使用 LDA 主题模型对每条

答复提取了 2 个主题，如图 14 是某条答复的 2 个主题。其中数字表示该词对主题的贡献率。

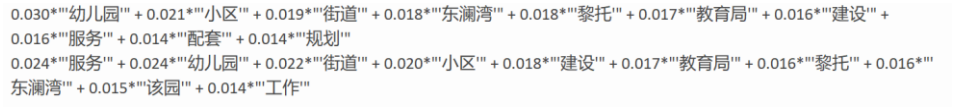


图 14　某条答复的文本主题

### 6.1.2 留言与答复的相似度计算

基于 LDA 主题模型提炼的答复主题关键词，对答复主题中的词与留言主题进行相似度计算，这里采用余弦距离和中文编辑距离进行加权运算，即

主题相似度 = 0.3*编辑距离 + 0.7*余弦距离

编辑距离是指两个字串之间，由一个转成另一个所需的最少编辑操作次数，如果它们的距离越大，说明它们越是不同。编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。[6]

例如：要将字符串“ABCD”转化为“DBDF”，那么最小需要经过 3 步操作：

1、将 A 替换为 D

2、删去 C

3、在最后面增加 F

因此它们的编辑距离为 3。

假设使用 d[i, j]表示将字符串 s[ : i ]转换为字符串 t[ : j ]所需要的最少编辑操作次数，当 i = 0，即 s 为空时，那么对应的 d[0, j] 就是增加 j 个字符，使得 s 转化为 t；反之，在 j = 0，即 t 为空时，对应的 d[i, 0]就是减少 i 个字符，使得 s 转化为 t。当 i 和 j 都不等于 0 时，要想得到将 s[ : i ]经过最少次数的编辑操作就转变为 t[ : j ]，就必须在最后一步操作之前以最少次数的编辑操作，使得现在 s 和 t 只需要再做一次操作或者不做就可以完全相等。

这里的“最后一步操作之前”指在 k 个操作内将 s[ : i ]转换为 t[ : j-1 ]或在 k 个操作内将 s[ : i-1 ]转换为 t[ : j ]或在 k 个操作内将 s[ : i-1 ]转换为 t[ : j -1]。与之对应的最后一步操作为将 t[j]加上 s[ : i ]或将 s[i]移除或者将 s[i]替换为 t[j]这样总共需要 k+1 个操作。在最后一种情况下，如果 s[i]刚好等于 t[j]，那么就经过 k 个操作将 s 转化为 t。

### 6.1.3 得分标准化

由于相关性的得分范围是[0,1]，而对于留言以及回复由于语义的不同，在利用相似度来度量相关性会使分数整体偏低，集中在 0.5 以下，因此，对上述得到的相似度做标准化处理，即相关性 $=\frac{\text{主题相似度}-\min(\text{主题相似度})}{\text{主题相似度极差}}$，使得相关性的分数区间不再集中在 0.5 以下。

## 6.2 答复意见的完整性

在评价系统中完整性满分为 1 分，其中开头句的完整性、结尾句的完整性和答复是否解决对应问题所占分数权重之比为 3:3:4。将每条答复意见以标点符号为切割符进行切割，形成多个短句。选取前 2 个短句和后 4 个短句。

1) 若问候语在前 2 个短句中有出现，判定开头句完整，得分为 0.3 分；若没有出现，判定开头句不完整，得分为 0 分。

2) 若感谢语以及答复日期在后 4 个短句中有出现，判定结尾句完整，得分为 0.3 分；若二

者都没有出现或其中一者没有出现，判定结尾句不完整，得分为 0 分。

3) 若转移问题的词语没有出现在后 4 个短语中，判定该答复解决了问题，得分为 0.4 分；

若转移问题的词语出现在后 4 个短句中，判定该答复没有解决问题，得分为 0 分。

问候语，感谢语，转移问题的词语具体如下：

表 6 问候语、感谢语及转移问题的词语

| | |
|---|---|
| **问候语** | ’你好’、 ’您好’、’你好’、’您好’、’你们好’、’您们好’、’收悉’ |
| **感谢语** | 感谢’、’谢谢’、 ’理解’、 ’我们非常乐意听取您的意见和建议’ |
| **转移问题的词语** | ’已转’、’已收悉’、’交办’、’转交’、’待核实后给您答复’、’办复’、<br>’敬请关注后续回复’ |

## 6.3 答复意见的可解释性

在评价系统中可解释性满分为 1 分。我们使用哈工大中文篇章关系语料库中的关联词表[7]，对答复意见进行连接词的显性识别[8]。连接词可表示因果、并列、转折、解说的关系，如图 15 所示，其中因果类连接词对于判定答复意见是否可解释性极具重要性，因为中文中习惯使用因果类连接词来叙述某人或某物基于某种原因[9]，引出某种结果。

此外，我们自定义可解释性的词典，如表 7 所示。我们将每条答复意见中的连接词和可解释性词语识别出来并制成列表，每条答复分别对应一个列表。由于答复意见中的背景描述经常是隐性表达，即没有明显的连接词或者关键词，且背景所占经常较多，故我们不能根据列表中的词数或者列表中的词数与答复意见文本长度的比率来判断可解释性的强弱。鉴于没有更好的办法，我们假设若列表中为空，则答复的可解释性很差，得分为 0 分。若列表不为空，则答复具有可解释性，得分为 1 分。

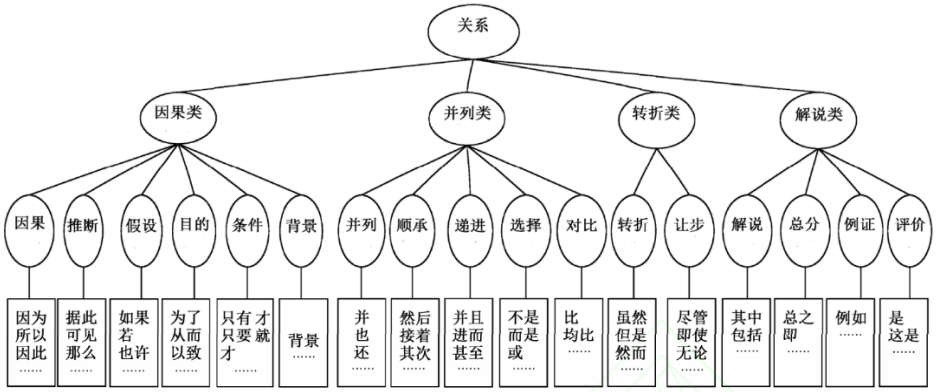


图 15 基于连接词的关系分别

表 7 自定义可解释性的词典

| **自定义可解释性的词典** |
|---|
| 依、按、根据、按照、调查、核实、经查、据查、针对、依据、依法、规定、流程、依规、方案、法律、规划、规定、审查、协商、部署、协议、政策、条件、符合、审批、认定、证、研究、通知、要求、核查、询问、查看、鉴定、记录、通过、了解、调阅、批准、核定、公示、属实、巡查、指示、查明、依照、查处、合规、合法、为了、确保、满足、查询 |

## 6.4 答复意见的及时性

在评价系统中及时性满分为 1 分。我们对留言时间与回复时间的所有时间差按五分位法划分，即取 1/5，2/5，3/5，4/5 分位数。若 3 天内答复，得分为 1 分；若在一个星期内答复，得分为 0.75 分；若在两个星期内答复，得分为 0.5 分；若在一个月内答复，得分为 0.25 分；若一个月以上才答复，得分为 0 分。

## 6.5 结果分析

最后，将相关性、完整性、可解释性和及时性四项得分相加，得到最终的答复意见总得分。我们从附件四中随机抽取某 10 条答复意见，对其质量进行评价分析，如图 16 所示，对其他答复意见的评价分析类似。其中的完整性得分由开头句完整性、结尾句完整性、问题解决这三个指标组成，如图 17 所示。

我们发现总得分较高的答复意见，其相关性、完整性、可解释性和及时性一般都比较高，若这四项指标中存在得分为 0 的情况，则该条答复意见的质量得分就比较低了，比如留言编号为 75000 的答复意见“您的留言已收悉。关于您反映的问题，已转 F7 县调查处理。”，其相关性得分和及时性得分都为 0，说明答复和留言并不那么相关且答复时间距离留言时间在一个月以上。并且结尾句没有感谢语和日期，也转移了问题，留言问题没有得到解决。但是因为“转至 F7 县调查处理”说明了问题没有解决的原因，所以该答复具有可解释性。总的来说，该答复质量得分只有 1.3 分，在这 10 条答复意见中质量最差。

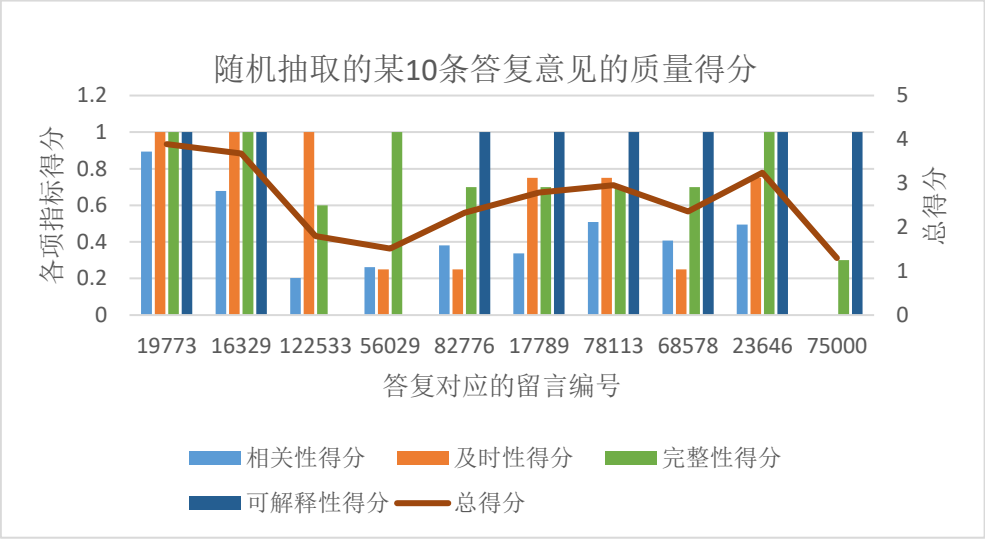


图 16 随机抽取的某 10 条答复意见的质量得分

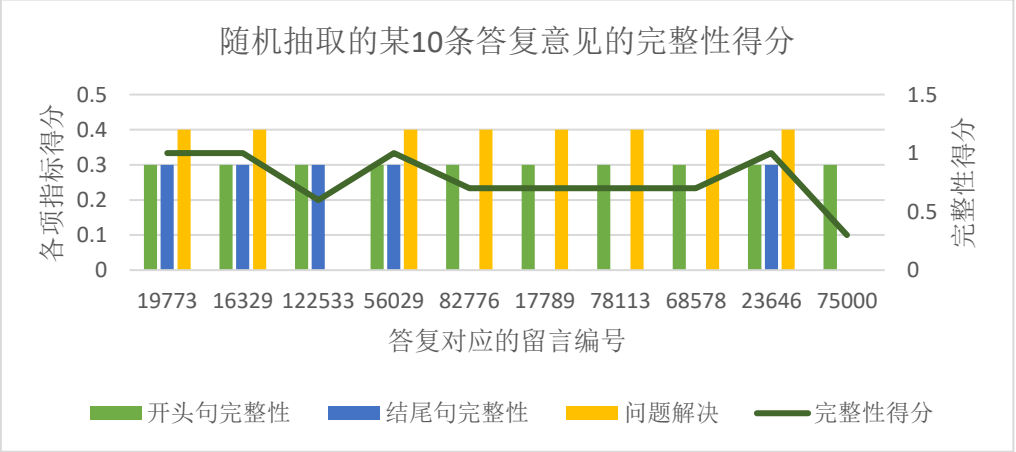


图 17 随机抽取的某 10 条答复意见的完整性得分

# 7. 总结

## 7.1 结论

本文首先对数据进行预处理，包括数据清洗，中文分词，去停用词，再根据文本的不同类标签，通过基于向量空间模型算法提取了文本关键词，并使用四个模型对留言内容进行一级标签分类，在得到的结果中，**线性支持向量回归算法相对最优的结果，F1-Score 评价指标达 0.86。**

在挖掘热点问题中，本文首先对不同的地点人群进行分类，再通过余弦距离对问题归类并筛选出留言数和点赞数量较多的留言问题。基于筛选的留言问题计算热度评价体系指标，对指标利用 spss 软件进行因子分析，得分前五的主题样本作为 Top5 热点问题，最终呈现出的**附件：热点问题表.xls** 和**附件：热点问题留言明细表.xls。**

在留言的答复意见的评价体系建立问题中，我们对答复留言多相关性，完整性，可解释性和及时性进行评分。定义相关性为构建的答复文本主题与留言主题之间的文本相似度；定义完整性为是否有标准的开头句、结尾句，是否解决留言中的问题；在可解释性方面，本文使用哈工大中文篇章关系的关联词表以及自定义的可解释性词典来判别答复意见是否具有可解释性；定义及时性为留言与答复时间的间隔；根据上述定义对其进行各项评分以及综合评分，结果由**附件：总得分（最终版）.xlsx** 呈现。

## 7.2 回顾与展望

在进行本次实验中，我们对文本挖掘有了更深的体会，同时也存在未解决的问题，在留言的答复意见的评价体系建立问题中，我们根据关联词表以及自定义的可解释性词典来判别答复意见是否具有可解释性，这是比较片面的，忽略了文本语义中没有用关联词表达出的关联关系。

本小组在能力有限，在时间有限的情况下，没有实现所有的想法，可解释性和相关性的衡量并没有想象中简单，因此我们只能将其简化。在实践过程中，我们通过看视频，查文献，查资料学习到了许多文本挖掘的知识，能力也得到了有效地锻炼，同时小组成员之间互帮互助，分工协作完成了这次活动。

# 参考文献


[1] https://www.cnblogs.com/csudanli/p/5409164.html

[2] https://zhuanlan.zhihu.com/p/66904318

[3] 姜霖,王东波.采用连续词袋模型(CBOW)的领域术语自动抽取研究[J].现代图书情报技术,2016(02):9-15.

[4] 何跃,蔡博驰.基于因子分析法的微博热度评价模型[J].统计与决策,2016(18):52-54.

[5] https://www.docin.com/p-1550177716.html

[6] 于长永,李淼淼,赵楚,马海涛.一种新颖的编辑距离限制下的相似性确认算法[J]2019,40(11): 1005 — 3026

[7] http://ir.hit.edu.cn/hit-cdtb/index.html

[8] 李艳翠,孙静,周国栋.汉语篇章连接词识别与分类[J].北京大学学报(自然科学版),2015,51(02):307-314.

[9] 李文翔,晏蒲柳,张滨,夏德麟.基于语料库的关联词识别方法[J].计算机工程与应用,2004(07):50-52.