第八届泰迪杯数据挖掘挑战赛

“智慧政务”中的文本挖掘应用

摘要

随着信息技术、信息产业的飞速发展，信息传播和更新的速度日新月异，这也使得信息的增长呈现井喷式发展，增长速度异常迅猛。如何利用好这些信息，逐渐成为企业、政府关注的焦点。

智慧政务正在成为互联网时代政府管理的新形态，智慧政务运用信息和通信技术手段并以大数据为核心，为政府指挥决策，解决问题。在新的社会形势下，智能化政府是贯彻和落实国家信息化发展的重要战略。本文通过分析当前各类社情民意的留言数据，为政府职能部门发现当下群众主要意见和建议并做出合理的解决改善方案提供了一定的帮助。

针对问题一，我们首先对数据进行简单分析，利用 jieba 分词对数据处理并对留言信息计算 TD-IDF 值，同时采用 sklearn 中的 chi 进行卡方检验，找出每个分类中关联度最大的两个词语和两个词语对。通过比较 4 种分类器（Logistic Regression(逻辑回归)、(Multinomial) Naive Bayes(多项式朴素贝叶斯)、Linear Support Vector Machine(线性支持向量机)、Random Forest(随机森林)），得出最优模型对其分类。

针对问题二，每个留言详情做分词处理，计算每个词汇 tfidf 值，构建词典分数矩阵，对词典进行分类聚类，并可视化输出，再构建情感词典分析模型，对模型输入情感词典进行分类训练，利用模型的计算方式和机器学习的计算方式来计算每个留言的情感分数，根据情感分析模型，分析留言详情，将留言详情按照热度分成五大类型，利用层次聚类以及 LDA 主题模型，找到每个类中最热点的问题，提取出关键字，通过对数据可视化分析与统计得出最热门的五个问题，利用 Excel 筛选出对应的留言详情。

针对问题三，根据相关部门对留言的答复意见，采用 NLPIR 大数据语义智能分析系统针对大留言内容进行处理，从相关性、可解释性角度对留言回复进行分析特征词提取、情感分析，对答复意见的质量给出一套评价方案，并尝试实现。

关键词：智慧政务；Python；Snowmlp；LinearSVC；TF-IDF 值；LDA；nlpir；层次聚类

Text Mining Application in "Smart Government Affairs"

Abstract

With the rapid development of information technology and information industry, the speed of information dissemination and updating is changing with each passing day, which also makes the growth of information show a blowout development, the growth rate is extremely rapid. How to make good use of this information has gradually become the focus of attention of enterprises and governments.

Smart government affairs is becoming a new form of government management in the Internet era. Smart government affairs use information and communication technology and take big data as the core to guide government decisions and solve problems. Under the new social situation, intelligent government is an important strategy to implement and implement the development of national informatization. By analyzing the current message data of various social conditions and public opinions, this article provides some help for the government functional departments to discover the current main opinions and suggestions of the masses and make reasonable solutions to improvement.

For problem one, we first analyze the data briefly, use jieba word segmentation to process the data and calculate the TD-IDF value of the message information, and use chi in sklearn to perform chi-square test to find the two most relevant in each category Words and two word pairs. By comparing 4 classifiers (Logistic Regression, Logistic Regression, Multinomial Naive Bayes, Linear Support Vector Machine, Random Forest), the optimal the model classifies it.

For problem two, word segmentation is processed for each message detail, the TF-IDF value of each vocabulary is calculated, the dictionary score matrix is constructed, the dictionary is classified and clustered, and then the sentiment dictionary analysis model is constructed, and the model is input to the sentiment dictionary for classification training Calculate the sentiment score of each message using the calculation method of the model and the calculation method of machine learning, and get the five most popular questions through visual analysis and statistics of the data.

In response to question three, according to the response comments of relevant departments on the message, the NLPIR big data semantic intelligent analysis system is used to process the content of the large message. The message reply is analyzed for feature word extraction, sentiment analysis and response The quality of the opinions gives a set of evaluation plans and attempts to achieve them.

Keywords: smart government, Pyhton, Snowmlp, LinearSVC, TF-IDF value,

LDA, NLPIR, hierarchical clustering

目录

1. 挖掘目标.....1
2. 分析过程与方法.....1
总体流程图.....1
2.1 问题 1 分析方法与过程.....2
2.1.1 流程图.....2
2.1.2 数据分析.....2
2.1.3 数据预处理.....4
2.1.4 分类器选择.....6
2.1.5 模型评估.....7
2.2 问题 2 分析方法与过程.....7
2.3 问题 3 分析方法与过程.....9
3. 结果分析.....10
3.1 问题 1 结果分析.....10
3.1.1 分类器选择结果.....10
3.1.2 线性支持向量机模型评估.....10
3.2 问题 2 结果分析.....12
3.2.1 对数据分析可视化结果.....12
3.2.2 热点问题表与对应留言信息.....14
3.3 问题 3 结果分析.....15
3.3.1 从相关性分析.....15
3.3.2 从可解释性分析.....16
4 结论.....17
5 参考文献.....18

第八届泰迪杯数据挖掘挑战赛

1. 挖掘目标

随着时代的高速发展，各类有关于社情民意的文本数据不断上升。相对于政府部门来说人工分类和人工划分热点是具有极大挑战性的。同时，随着大数据、云计算等技术的发展，建立自然语言处理技术的智慧政务系统已经是社会管理创新的新趋势，对于提升政府工作效率有着极大的推动作用。

本次建模目标是互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。利用 jieba 中文分词工具对留言内容进行分词、Linear Support Vector Machine(线性支持向量机)模型进行一级标签分类等方法达到以下三个目标：

1） 利用文本多分类的方法对对留言进行分类，建立关于留言内容的一级标签分类模型。

2） 根据群众的留言信息，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，并给出评价结果挖掘出热点问题，即某一时段内群众集中反映的某一问题，有助于相关部门进行有针对性地处理，提升服务效率。

3） 根据相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2. 分析过程与方法

总体流程图

这里，我们对问题进行详细分析，得到如下总体流程。

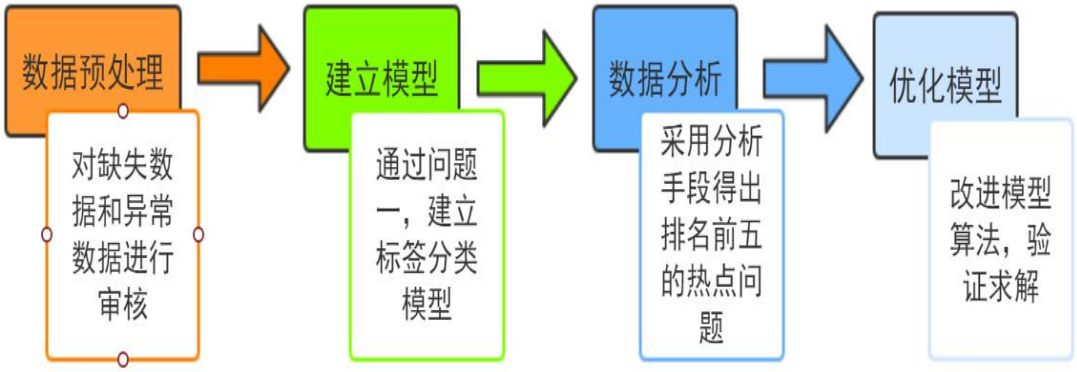


图 1：总体流程图

本用例主要包括如下步骤：

步骤一：数据预处理，在题目给出的数据中，对数据进行中文分词分类，在对留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法计算词频

步骤二：建立标签一级分类模型，并进行模型评估

步骤三：数据分析附件 3 留言信息，通过情感分数得出热点问题

2.1 问题 1 分析方法与过程

2.1.1 流程图

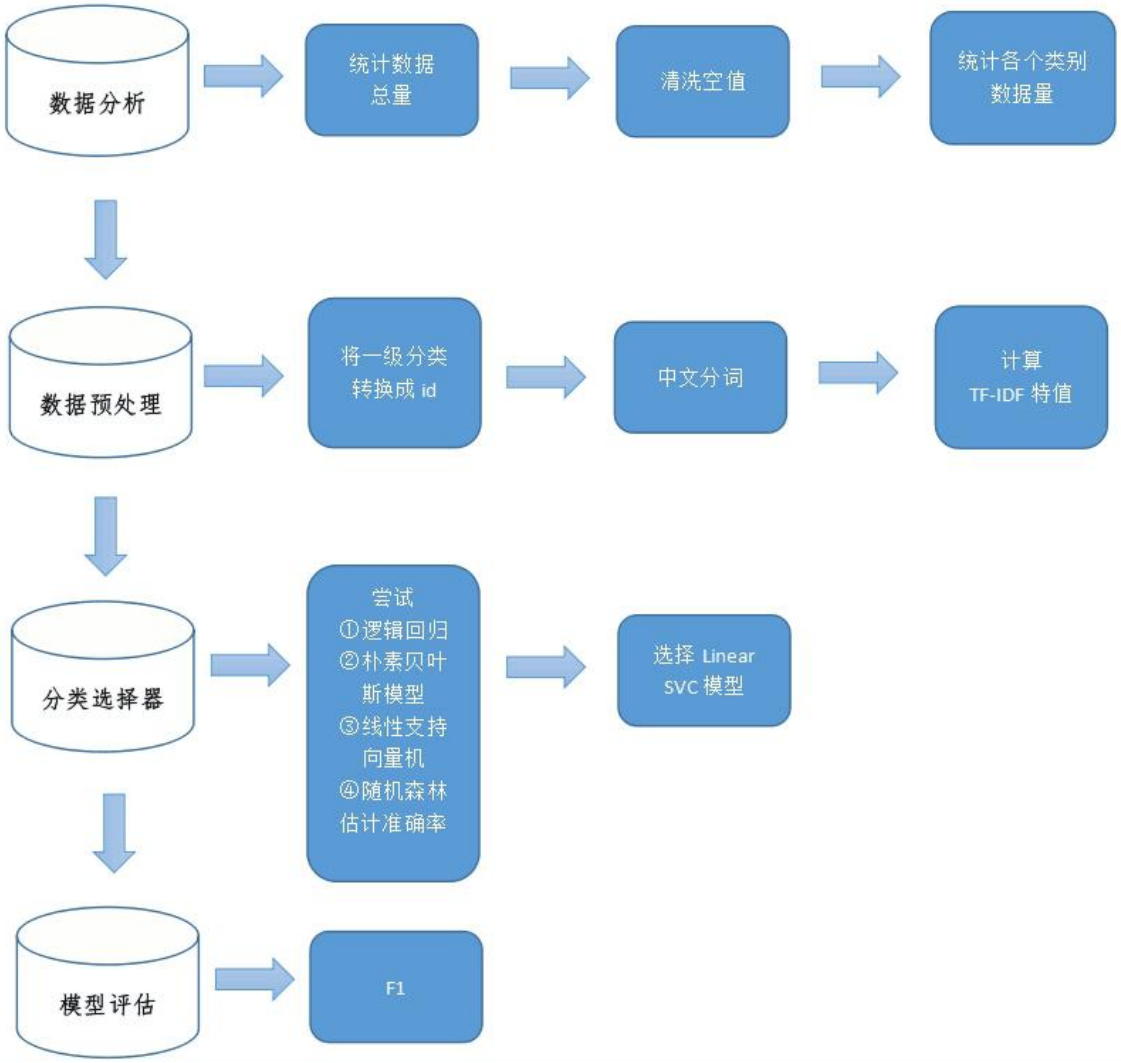


图 2：问题 1 流程图

2.1.2 数据分析

2.1.2.1 数据统计

步骤如下：

（1）在附件 2 数据中，数据总量为: 9210。查看数据中的“留言详情”以及“一级分类”，其次把不同留言信息数据分到不同的分类中去,且每条数据只能对应 7 个类中的一个类。

	留言详情	一级分类
7228	\n\t\t\t\t\t\n\t\t\t\t\t在A4区月湖安置小区，传销骗了我20万说投资...	商贸旅游
7127	\n\t\t\t\t\t\n\t\t\t\t\t我经优客租公司介绍租下了A市四方坪的房子，刚...	商贸旅游
3123	\n\t\t\t\t\t\n\t\t\t\t\t 从M3县，M4市至E5县，E市装黑土的后...	交通运输
8032	\n\t\t\t\t\t\n\t\t\t\t\t涟水名城电梯房的八部电梯多年未搞年检，但作为...	商贸旅游
451	\n\t\t\t\t\t\n\t\t\t\t\t B6县建筑市场专门没有按照有关法律法规执...	城乡建设

图:3：数据示例

（2）清洗掉数据中的空值

```
print("在 一级分类 列中总共有 %d 个空值." % df['一级分类'].isnull().sum())
print("在 留言详情 列中总共有 %d 个空值." % df['留言详情'].isnull().sum())
df[df.isnull().values==True]
df = df[pd.notnull(df['留言详情'])]
```

```
在 一级分类 列中总共有 0 个空值.
在 留言详情 列中总共有 0 个空值.
```

图 4：查看是否有空值

可以看出数据完整，在一级分类对应的列以及留言详情对应的列中没有出现数据缺漏的情况。

（3）统计一下各个类别的数据量如图是对留言的各个一级标签进行排序计数。

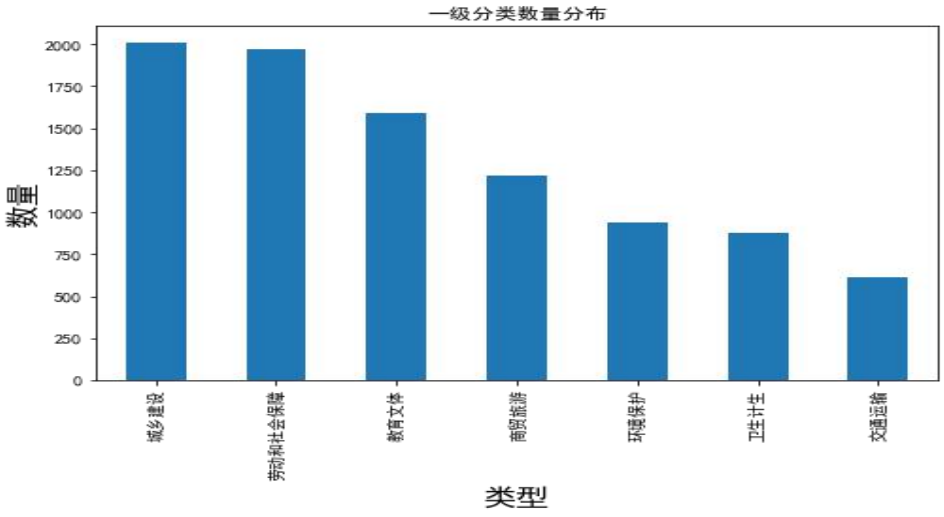


图 5：一级分类数量条形图

总体来看小区居民在各个分类都有意见，可以看出“城乡建设”与“劳动社会保障”类数量较多，说明小区居民对城乡建设和劳动社会保障分类意见较多，

城乡建设和劳动社会保障对小区居民影响较大。

“交通运输”相对于“城乡建设”与“劳动社会保障”分类，意见较大幅度减少。

“教育文体”、“商贸旅游”、“环境保护”和“卫生计生”分类的数量虽然没有城乡建设和劳动社会保障分类那么多，但是居民对其也存在不少意见，群众留言热度也挺高。

2.1.3 数据预处理

2.1.3.1 一级分类标签转换 id

将一级分类标签转换成 id（0-6），这样便于以后的分类模型的训练。

	一级分类	一级分类_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图 6：标签转换图

2.1.3.2 对留言信息进行中文分词

在对留言信息进行挖掘分析之前，先要把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。在附件 2 中，以中文文本的方式给出了数据。 为了便于转换，先要对这些留言信息进行中文分词。这里采用 python 的 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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，使得能更好的实现中文分词效果。这包括删除文本中的标点符号，特殊符号，还要删除一些无意义的常用词(stopword)。因为这些词和符号对系统分析预测文本的内容没有任何帮助，反而会增加计算的复杂度和增加系统开销，所有在使用这些文本数据之前必须要将它们清理干净。

2.1.3.3 TF-IDF 算法

TF-IDF（term frequency－inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF-IDF 是在单词计数的基础上，降低了常用

高频词的权重,增加罕见词的权重。TF-IDF 算法的具体原理如下：
第一步，计算词频，即 TF 权重（TermFrequency）。

词频(TF)=某个词在文本中出现的次数。 （1）

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$词频（TF）=\frac{某个词个词在文本中出次数}{文本的总词数} \quad （2）$$

或

$$词频（TF）=\frac{某个词个词在文本中出次数}{该文本出现次数最多的词的出现次数} \quad （3）$$

第二步，计算 IDF 权重，即逆文档频率（InverseDocument Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本 中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$逆文档频率（IDF）=\log\left(\frac{语料库的文本总数}{包含该词的文本数+1}\right) \quad （4）$$

第三步，计算 TF-IDF 值（TermFrequencyDocumentFrequency）。

$$TF-IDF=词频（TF）*逆文档频率（IDF） \quad （5）$$

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词。

2.1.3.4 计算 TF-IDF 特征值

生成 TF-IDF 特征值的说明：

（1）使用 sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取文本的 TF-IDF 的特征值；

（2）使用了参数 ngram_range=(1,2),表示我们除了抽取留言详情中的每个词语外,还要抽取每个词相邻的词并组成一个“词语对”,如： 词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2,词 3)，(词 3，词 4)。从而扩展特征集的数量,因为有了丰富的特征集才有可能提高我们分类文本的准确度；

（3）参数 norm='l2',是一种数据标准划处理的方式,可以将数据限制在一点的范围内比如说(-1,1)。

2.1.3.4 卡方检验（chi2）

卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度,实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为 0，表明理论值完全符合。

我们采用卡方检验的方法来找出每个分类中关联度最大的两个词语和两个

词语对。卡方检验是一种统计学的工具,用来检验数据的拟合度和关联度。在这里我们使用 sklearn 中的 chi2 方法。找出了每个分类中关联度最强的两个词和两个词语对。这些词和词语对能很好的反映出分类的主题。

交通运输	Most correlated unigrams:	快递
		出租车
	Most correlated bigrams:	的士司机
		出租车司机
劳动和社会保障	Most correlated unigrams:	退休
		社保
	Most correlated bigrams:	劳动关系
		退休人员
卫生计生	Most correlated unigrams:	医生
		医院
	Most correlated bigrams:	社会抚养费
		乡村医生
商贸旅游	Most correlated unigrams:	传销
		电梯
	Most correlated bigrams:	小区电梯
		传销组织
城乡建设	Most correlated unigrams:	小区
		业主
	Most correlated bigrams:	住房公积金
		公积金贷款
教育文体	Most correlated unigrams:	学生
		学校
	Most correlated bigrams:	教育局领导
		培训机构
环境保护	Most correlated unigrams:	环保局
		污染
	Most correlated bigrams:	附近居民
		严重污染

图 7：最大关联词图

2.1.4 分类器选择

根据训练集的不同，若训练集很小，那么高偏差分类器要优于低偏差分类器；然而，随着训练集的增大，低偏差分类器将胜出高偏差分类器，因为高偏差分类器不足以提供准确的模型。本文通过训练集的大小判断筛选出较适用于本次的分类的模型。为了更好评估它们的准确率，使用如下四种模型进行比较：

1.Logistic Regression(逻辑回归)

2.(Multinomial) Naive Bayes(多项式朴素贝叶斯)

3.Linear Support Vector Machine(线性支持向量机)

4.Random Forest(随机森林)

选择出准确率最高的模型进行分类

2.1.5 模型评估

（1）针对平均准确率最高模型，我们将查看混淆矩阵，并显示预测标签和实际标签之间的差异。

（2）使用 F-Score 对分类方法进行评价。

2.2 问题 2 分析方法与过程

2.2.1 流程图

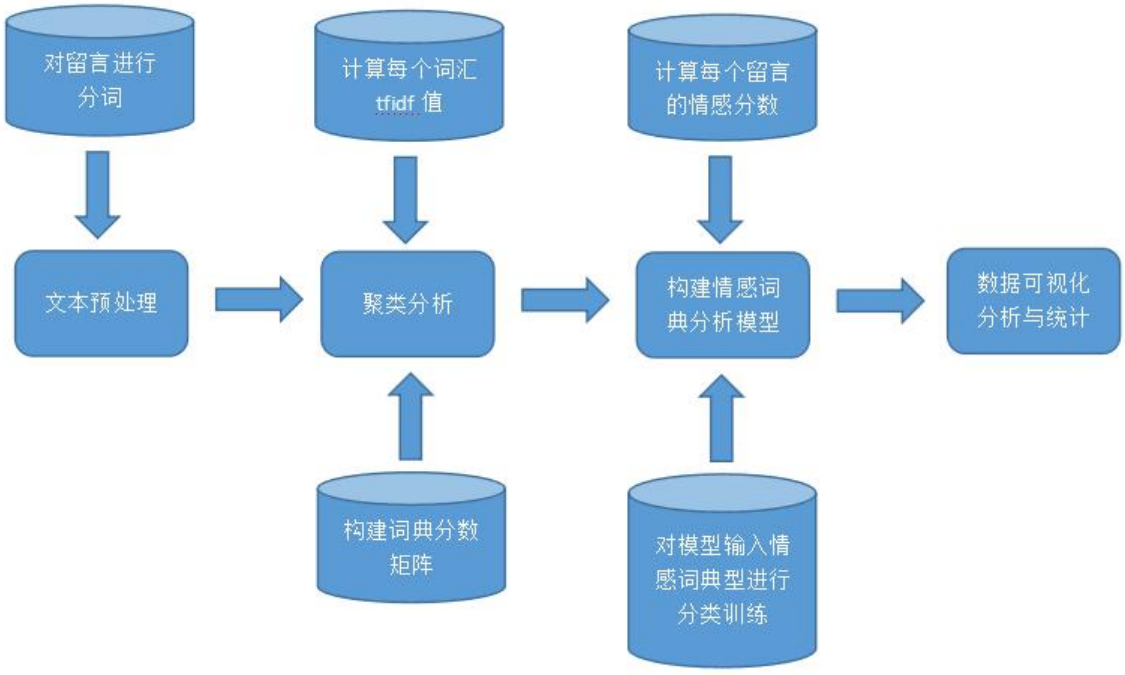


图 8：问题 2 流程图

2.2.1 文本预处理

采用 jieba 分词，对留言信息进行分词，分词出来的效果是能使得提取文章中所有的关键词，以方便进行话题聚类和情感分析，jieba 库是一个简单实用的中文自然语言处理库，在文本分词方面，文章采用 jieba 分词对采集到的的文本进行分词工作，jieba 基于前缀词典实现高效的词图扫描，生成句子中所有可能词构成的有向无环图，jieba 本身采用动态规划查找最大概率路径的形式，找到

该词的频率，jieba 本身也附带高效的语句训练，这里使用 snownlp 自带的语句文本以及制定好的语句文本训练来提高准确率。

2.2.2 计算 TF-IDF

有了关键词的文本矩阵之后，需要对每个关键词每条文本数据计算向量矩阵，对每个关键词的重要性占比作一个准确的数学计算，这样就能提高聚类后的话题准确性。

（1）计算每个词汇 TF-IDF 值

（2）构建词典分数矩阵

每个关键词需要计算 TF-IDF 值，TF 是词频，表示关键字在文本中出现的频率，IDF 是逆向文件频率，如果包含词条 t 的的文档越少，IDF 越大，则说明词条具有很好的类别区分能力，此时，需要将文档相似问题转换为数学向量矩阵问题，可以通过 VSM 向量空间模型来存储每个文档的词频和权重，特征抽取完后，因为每个词语对实体的贡献度不同，所以需要对这些词语赋予不同的权重。
它表示 TF（词频）和 IDF（倒文档频率）的乘积：
其中 TF 表示某个关键词出现的频率，IDF 为所有文档的数目除以包含该词语的文档数目的对数值。

TF-IDF 计算权重越大表示该词条对这个文本的重要性越大，它的目的是去除一些"的、了、等"出现频率较高的常用词。

2.2.3 层次聚类

层次法（Hierarchicalmethods）先计算样本之间的距离。每次将距离最近的点合并到同一个类。然后，再计算类与类之间的距离，将距离最近的类合并为一个大类。不停的合并，直到合成了一个类。其中类与类的距离的计算方法有：最短距离法，最长距离法，中间距离法，类平均法等。比如最短距离法，将类与类的距离定义为类与类之间样本的最短距离。

层次聚类算法根据层次分解的顺序分为：自下底向上和自上向下，即凝聚的层次聚类算法和分裂的层次聚类算法。

2.2.4 构建情感分析模型

方法所采用的的情感词典来自于 BOSON 中文语义开放平台的情感词典，BosonNLP 情感词典是从微博、新闻、论坛等数据来源的上百万篇情感标注数据当中自动构建的情感极性词典。因为标注包括微博数据，该词典囊括了很多网络用语及非正式简称，对非规范文本也有较高的覆盖率。该情感词典可以用于构建社交媒体情感分析引擎，停用词，程度副词均来自该词典并且在添加了医学方面的部分词典，提高情感词典的准确率。

在获得了情感词、程度副词、停用词的打分标准之后，我们需要建立模型，计算句子情感得分，我们认为情感分数计算逻辑识所有情感词语分数之和，以下是得分公式。得分公式：

score=（-1）^（否定词数量）*(程度副词数量)*(句子得分)。

例如：不是很喜欢看书，不是为否定词，很为程度副词，喜欢和看书为情感词。

那么情感分数 score=(-1)^1*1.75*2.35*(-0.74)=3.04325。
其中 1 是指一个否定词，1.75 是程度副词的数值，2.35 和-0.74 分别是喜欢和看书的数值。

2.2.5 数据可视化与统计

2.2.5.1LDA 主题模型

文档主题生成模型（Latent Dirichlet Allocation，简称 LDA）通常由包含词、主题和文档三层结构组成。LDA 模型属于无监督学习技术，它是将一篇文档的每个词都以一定概率分布在某个主题上，并从这个主题中选择某个词语。文档到主题的过程是服从多项分布的，主题到词的过程也是服从多项分布的。

文档主题生成模型（Latent Dirichlet Allocation，简称 LDA）又称为盘子表示法（Plate Notation），下图是模型的标示图，其中双圆圈表示可测变量，单圆圈表示潜在变量，箭头表示两个变量之间的依赖关系，矩形框表示重复抽样，对应的重复次数在矩形框的右下角显示。LDA 模型的具体实现步骤如下：

（1）从每篇网页 D 对应的多项分布 θ 中抽取每个单词对应的一个主题 z。
（2）从主题 z 对应的多项分布 φ 中抽取一个单词 w。
（3）重复步骤 1 和 2，共计 Nd 次，直至遍历网页中每一个单词。

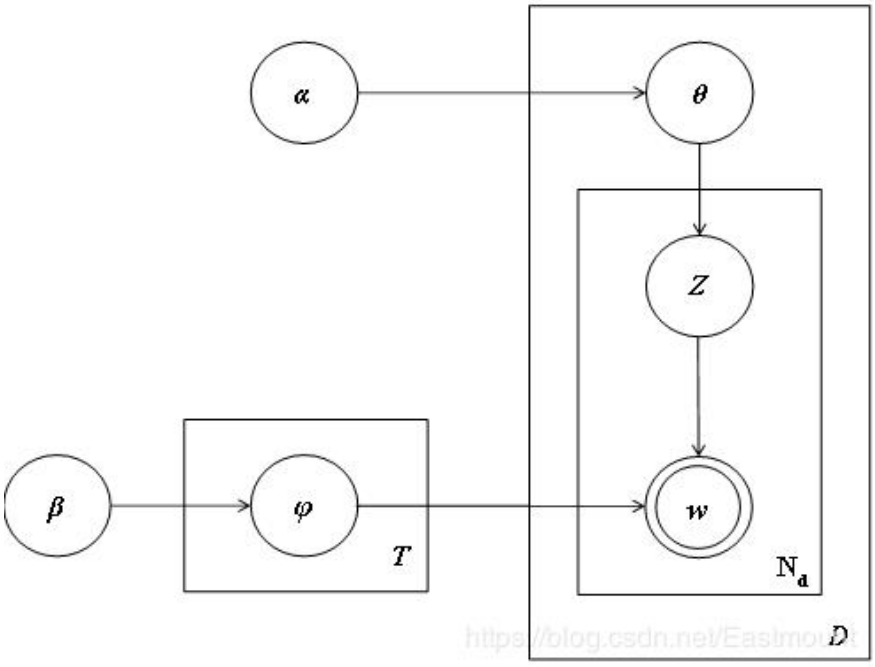


图 9：LDA 模型

根据情感分析模型，分析留言详情，将留言详情按照热度分成五大类型，利用层次聚类以及 LDA 主题模型，找到每个类中最热点的问题，提取出关键字，利用 Excel 筛选出对应的留言详情。

2.3 问题 3 分析方法与过程

2.3.1 语义分析系统

NLPIR 大数据语义智能分析平台是针对大数据内容处理的需要，融合了网络精准采集、自然语言理解、文本挖掘和网络搜索技术的十三项功能，提供客户端工具、云服务、二次开发接口。

利用语义分析系统，对留言详情与留言回复进行分析，如下三个步骤：

（1）将留言详情与对应的留言回复抽取出来，每五条为一个文本；

（2）从相关性进行分析，分析每五条的留言详情与留言回复；

（3）从可解释性进行分析，主要分析每五条的留言详情与留言回复的情感是否呈现上升趋势。

3. 结果分析

3.1 问题 1 结果分析

3.1.1 分类器选择结果

比较 4 种不同的机器学习模型,并评估它们的准确率，使用如下四种模型进行比较:

1.Logistic Regression(逻辑回归)

2.(Multinomial) Naive Bayes(多项式朴素贝叶斯)

3.Linear Support Vector Machine(线性支持向量机)

4.Random Forest(随机森林)

```
model_name
LinearSVC                0.869932
LogisticRegression       0.800547
MultinomialNB            0.628336
RandomForestClassifier   0.404880
Name: accuracy, dtype: float64
```

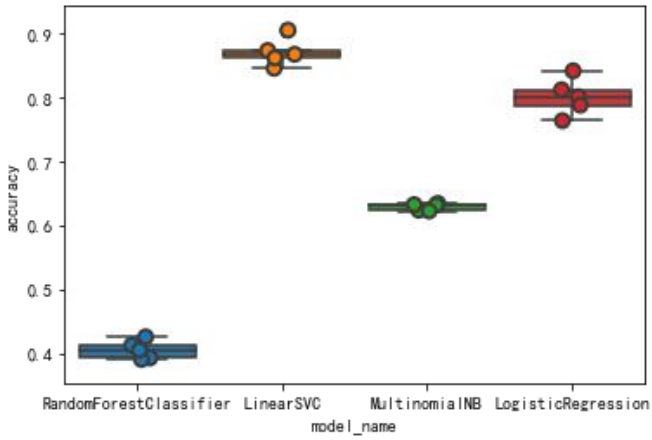


图 10：模型精准度图（1）　　图 11：模型精准度图（2）

可以看出 Linear Support Vector Machine(线性支持向量机)模型准确率最高。

3.1.2 线性支持向量机模型评估

Linear Support Vector Machine(线性支持向量机)模型评估结果：

混淆矩阵的主对角线表示预测正确的数量,除主对角线外其余都是预测错误的数量.从上面的混淆矩阵可以看出“环境保护”，“交通运输”类预测最准确,

“教育文体”，“卫生生计”类准确度其次。“城乡建设”类预测的错误数量教多。

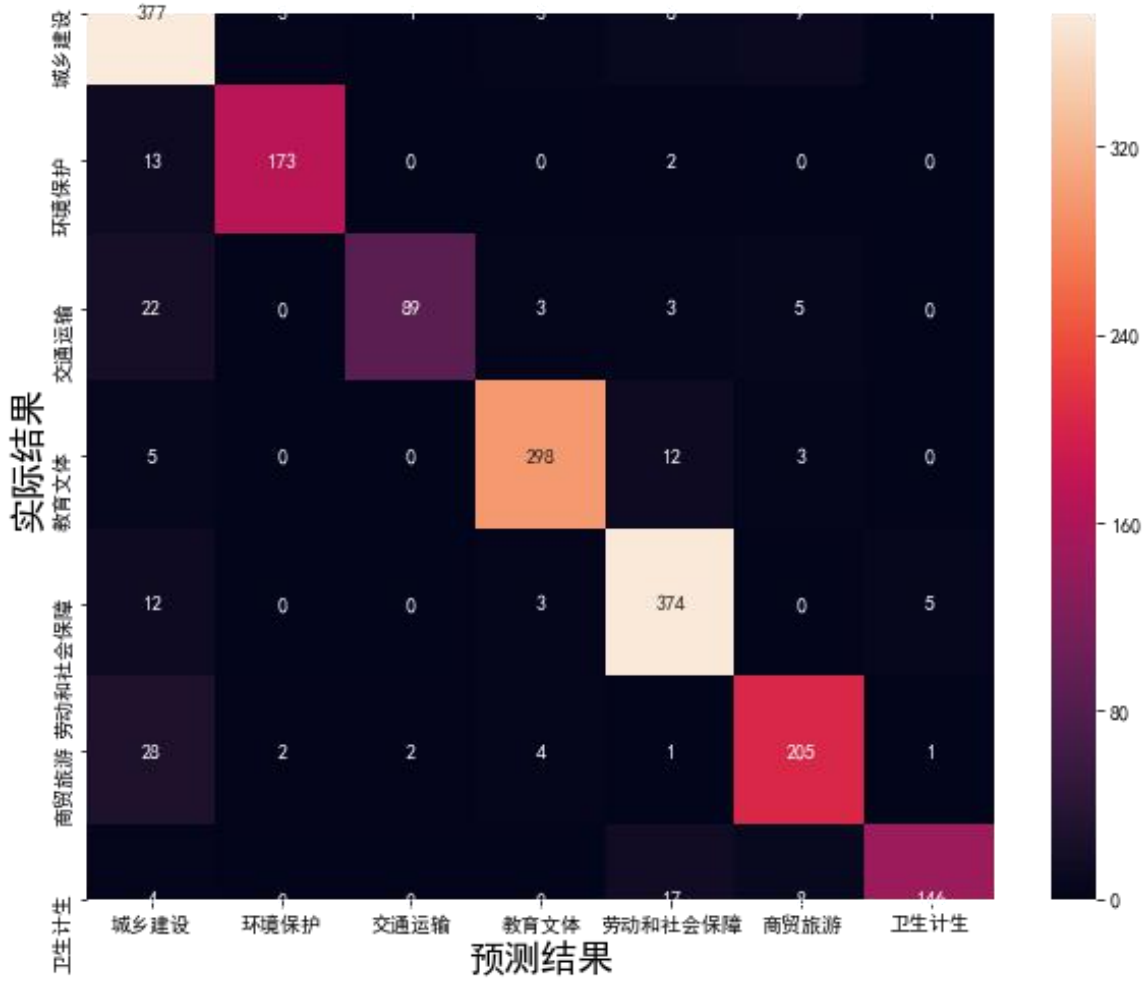


图 12：预测结果混淆矩阵

多分类模型一般不使用准确率(accuracy)来评估模型的质量，因为accuracy 不能反应出每一个分类的准确性，因为当训练数据不平衡(有的类数据很多，有的类数据很少)时，accuracy 不能反映出模型的实际预测精度，这时候我们就需要借助于 F1 分数、ROC 等指标来评估模型。

下面我们将查看各个类的 F1 分数.

	precision	recall	f1-score	support
城乡建设	0.82	0.94	0.87	402
环境保护	0.97	0.92	0.95	188
交通运输	0.97	0.73	0.83	122
教育文体	0.96	0.94	0.95	318
劳动和社会保障	0.90	0.95	0.92	394
商贸旅游	0.89	0.84	0.87	243
卫生计生	0.95	0.83	0.89	175
accuracy			0.90	1842
macro avg	0.92	0.88	0.90	1842
weighted avg	0.91	0.90	0.90	1842

accuracy 0.9022801302931596

图 13：F1 分数

通过二分类方法分析可以看出环境保护和交通运输的 precision 更接近 1，也就是更接近理想状态，预测更准确，劳动和社会保障的 recall 更接近理想状态，f1-score 是精确率与召回率的调和平均值，可以看出环境保护和教育文体类预测最准确，accuracy 为 0.90228 被预测正确的比例接近 1，说明预测是对的。

3.2 问题 2 结果分析

3.2.1 对数据分析可视化结果

（1）对留言信息进行情感分析

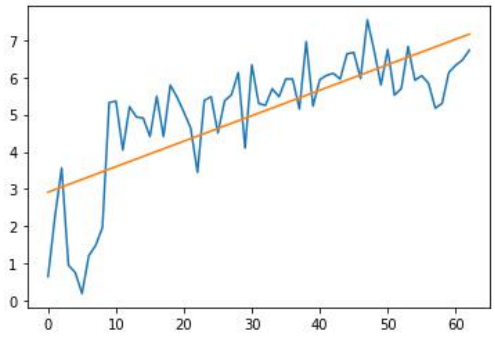


图 14-1：机遇词典情感分析

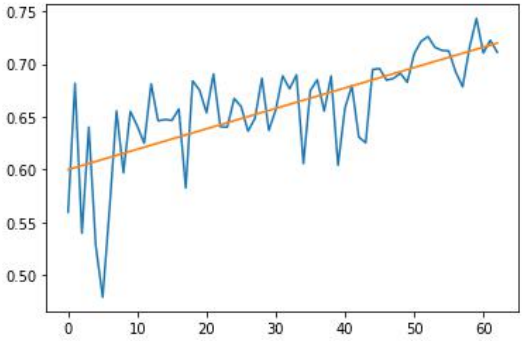


图 14-2：机器学习情感分析

这两个图反映出的是不同的方法计算出来的分数会有差别，但是总体都是趋于向上的，随时间的推移，平均情感分数逐渐上升，说明群众的问题随着时间的推移在一点点的被解决，积极情感才会越来越多。

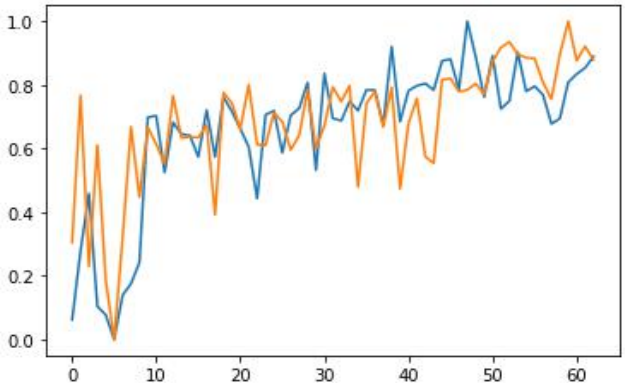


图 15：情感分析模型对比

两个方法的对比，可以看出这两条线基本重合，说明机遇词典的方法与机器学习方法差距不大。

（1）热点问题生成模型

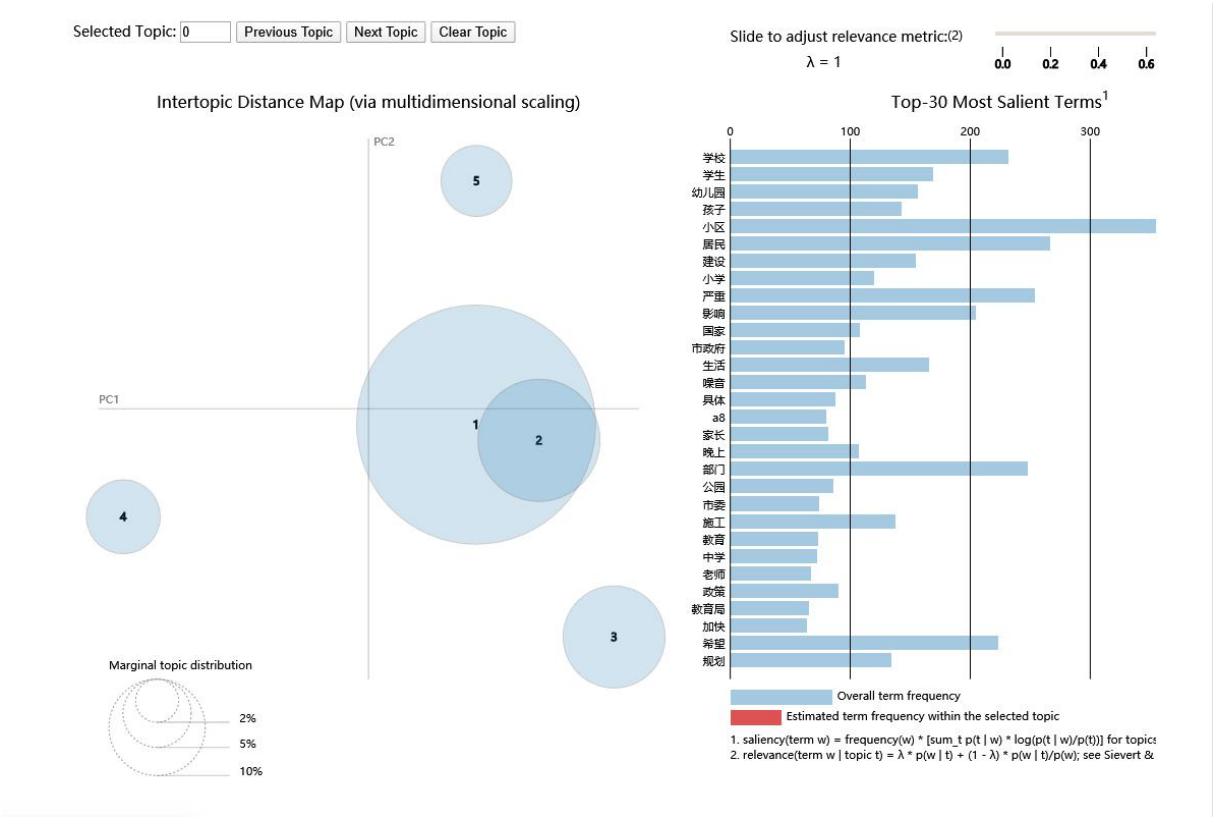


图 16：LDA 热点问题模型

（2）根据情感分数分析得出五大类热点留言

问题 id	热度指数	时间	问题
1	11793	2019/01/01-2019/12-31	小区生活建设问题
2	7888	2019/01/01-2019/12-31	公司发展问题
3	5385	2019/01/01-2019/12-31	领导督察社会问题
4	1781	2019/01/01-2019/12-31	生活质量水平问题
5	1684	2019/01/01-2019/12-31	学校学习问题

表 1：五大类热点留言

（4）层次聚类，将五大类热点留言层次聚类可视化得出每个类中热点最高的问题

图 17：热点问题的层次聚类

3.2.2 热点问题表与对应留言信息

热度排名	问题 id	热度指数	时间	地点/人群	问题
1	1	11793	2019/07/21-2019/09-25	A 市 A5 区魅力之城小区	小区临街餐饮店油烟扰民
2	2	7888	2019/01/11-2019/10-25	西地省九九富达实业发展集	A 市 58 车贷诈骗
3	3	5385	2019/01/28-2020/1-3	A7 县星沙地区	A7 县星沙地区社会问题
4	4	1781	2019/01/14-2019/12-15	A 市居民区	噪音扰民影响生活质量
5	5	1684	2018/5/17-2019/11-05	A 市经济学院学生	学校强制学生去定点企业实习

表 2：热点问题表

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	284147	A909113	A5区劳动东路魅力之城小区一	2019/07/21 10:29:36	局长：	0	3
1	360107	A0283523	A5区劳动东路魅力之城小区一	2019-07-21 10:29:36	局长：	3	0
1	360108	A0283523	A5区劳动东路魅力之城小区一	2019-08-01 16:20:02	局长：	6	0
1	272122	A909113	A5区劳动东路魅力之城小区一	2019/08/01 16:20:02	局长：	0	6
1	282172	A0004013	A6区高塘岭街道望福佳园附近	2019/8/13 11:17:45	位于A市A	5	1
1	287386	A909116	A市万科魅力之城商铺无排烟管	2019/08/18 14:44:00	A市万科魅	0	0
1	360104	A012417	A市魅力之城商铺无排烟管道，	2019-08-18 14:44:00	A市魅力之	0	0
1	360106	A235367	A市魅力之城小区底层商铺营业	2019-08-26 01:50:38	2019年5月	0	0
1	246362	A909114	A市魅力之城小区底层商铺营业	2019/08/26 01:50:38	2019年5月	0	0
1	195095	A0003908	魅力之城小区临街门面油烟直	2019/09/05 12:29:01	魅力之城	0	3
1	192337	A0004231	A3区梅溪湖看云路一师润芳园	2019/9/5 12:23:38	《张傅师	0	1
1	360100	A324156	魅力之城小区临街门面油烟直	2019-09-05 12:29:01	魅力之城	3	0
1	360103	A0012425	A5区劳动东路魅力之城小区临	2019-09-25 00:31:33	A5区劳动	1	0
1	246598	A0005484	A5区劳动东路魅力之城小区临	2019/09/25 00:31:33	A5区劳动	0	1

图 18：热点问题留言明细表

从热点问题表可以看出热度指数最高的问题为 A 市 A5 区临街餐饮店油烟扰民问题，A 市 58 车贷诈骗和 A7 县星沙地区社会问题也有挺多人提出，热度仅次于餐饮油烟扰民问题，A 市噪音扰民和 A 市经济学院学校强制学生去定点企业实习的问题热度指数排到了前五，虽然热度只为排名第一的十分之一，但反应的问题热度相对较高，需要得到重视。

由热点问题留言明细表可以看出居民们对热点问题的留言恢复也存在一些不满意的情况，居民们对热度第一的 A5 区劳动东路魅力之城小区餐饮店油烟扰民问题的留言详情不满意的人数多于满意的人数。

3.3 问题 3 结果分析

3.3.1 从相关性分析

相关系分析是通过比较两个或多个具备相关元素的变量进行分析，从而衡量变量之间的关系密切程度。文本的关联性计算作为信息检索处理中的一项基础性技术，直接影响结果的好坏。而传统的基于词语字符串匹配方法已经不适用于解决复杂的语言关联问题。

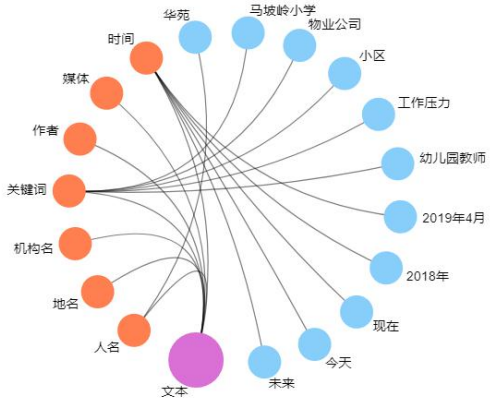


图 19-1：问题 1-5 相关性分析

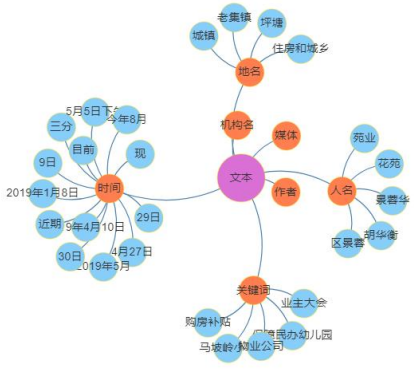


图 19-2：问题 1-5 留言回复相关性分析

由图 19 可以得出，在 1-5 的留言问题中，通过语义分析系统 NLPIR 可知文本主要内容有人名、地名、机构名以及关键词等，留言问题最多的分别是与物业公司、马坡岭小学、小区、工作压力和幼儿园教师的关键词。留言的作者、媒体与机构等文本信息是未明确的状态。在对问题的回复中，可看出领导对于时间段和关键词部分的回复较多，对地名和人名的回复较少。

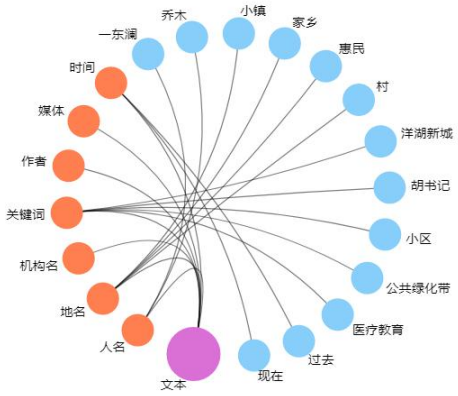


图 20-1：问题 6-10 相关性分析

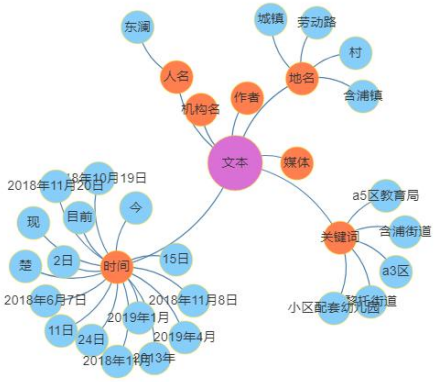


图 20-2：问题 6-10 留言回复相关性分析

由图 20 可以得出，在 6-10 的留言问题中，关键词主要与洋湖新城、胡书记、小区、公共绿化带以及医疗教育相关。媒体、作者和机构未进行问题留言。在 6-10 问题的回复中相较于问题 1-5 而言，问题 6-10 针对时间段的回复更加频繁，对关键词和地名的回复次之，对于媒体、作者和机构的回复为零。

3.3.2 从可解释性分析

NLPIR 情感分析技术对留言进行意见挖掘，文本情感分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程，通过对居民们留言的情绪状态可以从中挖掘出有用情感信息，从而对分析结果的可解释性进行说明。

（1）第一条到第五条的留言及回复分析：

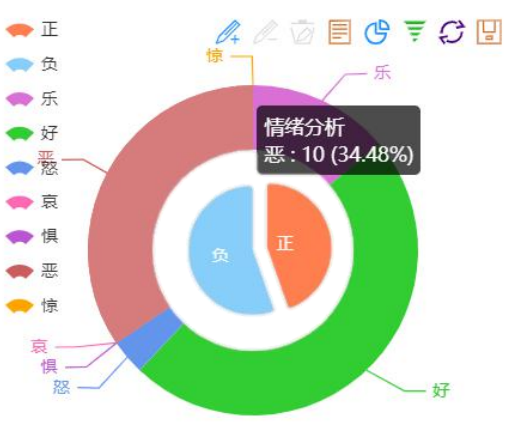


图 21-1：用户留言问题 1-5 情感分析

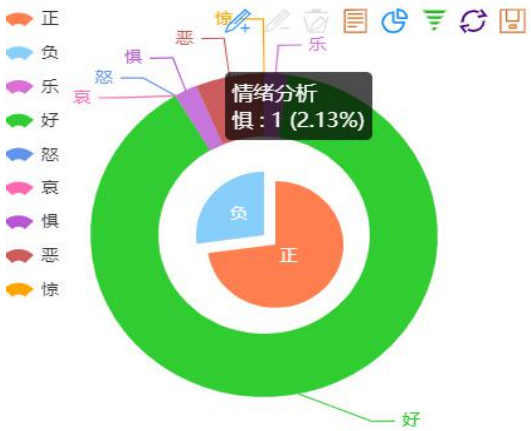
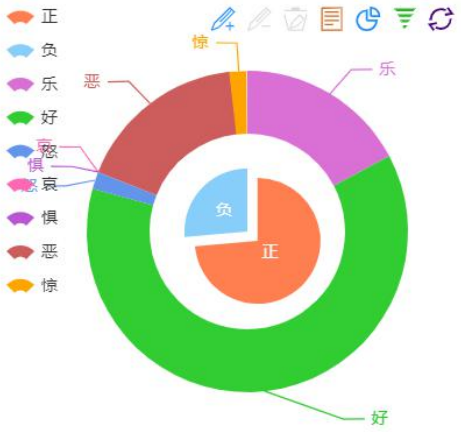


图 21-2：领导回复留言问题 1-5 情感分析

如图 21-1 是对留言详情（第一条到第五条的留言详情）进行情感分析：可以看出，用户在留言时大概是处于“恶”的情感状态，用户在留言时是带着负面情绪留言的，渴望得到满意的回复。

如图 21-2 是留言回复（第一条到第五条）情感分析图：领导在答复意见时大部分情况态度还是好的，只有少部分情况带着较差的情绪。

（2）第六条到第十条的留言及回复分析：



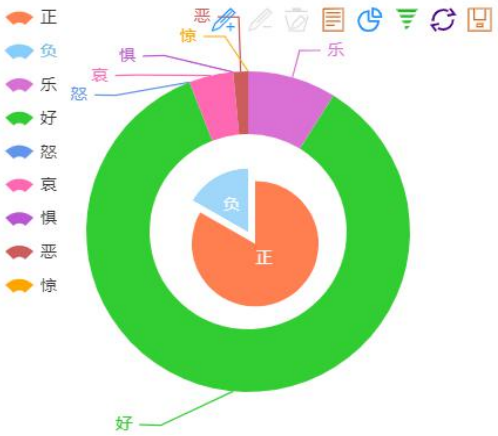


图 22-1:用户留言问题 6-10 情感分析 图 22-2:领导回复留言问题 6-10 情感分析

由图 22-1 留言情感分析可以看出，超过一半居民在留言时还是保持较好的情绪，和 1-5 留言问题对比，可以得出居民们对待问题的情感越负面它的热度就越高，由此可得出可解释性也越高。

由图 22-2 情感分析可以看出领导在回复居民们的留言时情感还是大部分处于“好”的情况，与热度 1-5 留言回复情感分析对比很容易可以看出在回复热度 6-10 留言问题时情绪为“恶”的状态明显减少。

4 结论

对社情民意的留言数据进行分析研究，了解群众主要意见和建议和如何更好地解决问题的方法，对社区居民们有重大意义，云留言产生的大量数据同时也是数据分析、数据挖掘的一个课题、一个难题。随着科技发展，传统的社区纸质留言已经落后，新型云留言、云回复以及对回复满意度进行点赞或反对的智慧政务已经慢慢走近人们的生活。

我们利用数据挖掘、数据分析以及模型的建立，从大量数据分析出有价值的信息最终，实现智慧政务。根据 TF-IDF 算法、jieba 分词和 Linear Support Vector Machine(线性支持向量机)模型，由 jieba 分词计算出关键词，由算法和模型统计小区留言中热度最高的问题，由分析结果可以看出，居民留言的问题分为交通运输、城乡建设、劳动社会保障、教育文体、商贸旅游、环境保护和卫生计生七大问题分类，并评估分类结果的准确性为 90%。

对留言信息进行情感分析，从而定义热度较高的问题，可以看出有关城乡建设和劳动社会保障类问题的留言较多，通过进一步数据挖掘处理得出热点问题，可以发现热点问题为：A5 区劳动东路魅力之城小区餐饮店油烟扰民问题；A 市 58 车贷诈骗问题；A7 县星沙地区社会问题噪音扰民影响生活质量；学校强制学

生去定点企业实习五个热点问题。

对留言回复从相关性和可解释性分析得出没留言回复是根据实际情况合理地进行回复，工作者认真对待每一条留言；回复的情感应是正面的，乐意亲切，并且具有可信度。

5 参考文献

[1] 周昭涛;文本聚类分析效果评价及文本表示研究[D];中国科学院研究生院(计算技术研究所);2005.
[2] 李金华;基于 SVM 的多类文本分类研究[D];山东科技大学;2010.
[3] 张磊;基于支持向量机的反垃圾电话技术研究[D];哈尔滨工程大学;2010.
[4] 徐明,高翔,许志刚;基于改进卡方统计的微博特征提取方法[J];计算机工程与应用;2014,50(19):113-117.
[5] 张华平;商建云;NLPIR-Parser:大数据语义智能分析平台[J];2019,《语料库语言学》（1）:87-104.
[6] 胡学钢,董学春,谢飞;基于词向量空间模型的中文文本分类方法[J].合肥工业大学学报:自然科学版;2007,30(10):1261-1264.
[7] 张保富,施化吉,马素琴;基于 TFIDF 文本特征加权方法的改进研究[J];计算机应用与软件;2011,28(2):17-20.
[8] 付永陈;基于博客搜索的博文情感倾向性分析技术的研究[D];东北大学;2010.
[9] 赵志升,靳晓松,温童童,梁俊花;基于 Python-Snownlp 的新闻评论数据分析[J];科技传播;2018 年 18 期.
[10] 胡吉明,陈果;基于动态 LDA 主题模型的内容主题挖掘与演化[J];图书情报工作;2014 年 02 期.
[11] 李熙铭;基于主题模型的多标签文本分类和流文本数据建模若干问题研究[D];吉林大学;2015.