# “智慧政务”中的文本挖掘与分析

# 摘　要

在政府政务中，收集民众的反应和意见，是一项极其重要的工作。近年来，各类民意反应构成的文本数据量不断增多，如果仅靠人工来进行群众留言的划分和热点问题的挖掘与分析，会给政务工作带来极大的不便。存在着效率低、耗时长、差错率高等问题。

本文旨在基于自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见数据，通过文本分析、文本特征提取、建立模型和文本挖掘与评估，对留言内容进行分类，挖掘出群众反应的热点问题并对有关部门给予的回复，从相关性、完整性、可解释性等角度进行质量评估。

针对问题一，我们根据所给的标签类型，对附件 2 表结构数据进行了各级标签数据的抽取再整合和文本预处理。进行文本特征提取后，通过多项式朴素贝叶斯、K 近邻、线性分类支持向量机和随机森林四种分类模型 F-Score 的评价和综合对比，最终选取了 LinearSVC 分类模型。通过 classification_report 模型评估和模型应用得到了较为准确的分类效果和较高的 F1 值。

针对问题二，首先，我们对预处理后的文本进行 TF-IDF 特征计算和选择，利用 gensim 计算文本相似度，通过调节参数来设置最有效的阈值进行文档归类。我们根据每类问题的归类数量、出现频率、点赞数和反对数构造了热度评价指标，根据指标的评分对热点问题进行排序。其次利用 LDA 模型提取，通过 LDAvis 得到了比较清晰的主题。再提取每个主题的时间范围，利用 Hanlp 识别具体的地点或人群之后，得到了最后的热点问题和热点问题留言明细表。

针对问题三，我们根据给出的数据构建了回复字数、回复时间间隔、礼貌用语数、回复等级和回复与留言的文本相似度五个评价指标。使用 kmeans 聚类方法聚成合适的类后，利用雷达图对聚类后的簇分析，并用分类模型对新的答复文本质量进行预测。

**关键词：**LinearSVC；LDA；Hanlp 命名实体识别；kmeans

# Abstract

In government affairs, collecting people's reactions and opinions is an extremely important task. In recent years, the volume of text data constituted by various public opinions has been constantly increasing. If people only rely on manual work to divide the comments of the masses and dig and analyze hot issues, it will bring great inconvenience to the work of government affairs. There are problems of low efficiency, long time consuming and high error rate.

This paper based on the public Internet source mass message records, and the relevant departments for some replies the opinion data, through text analysis, text feature extraction, model and evaluation and text mining, classifying the message content, digging out the hot issues and the relevant departments of the public to give reply, from the perspective of relevance, integrity and interpretability for quality evaluation.

Aiming at the first problem, according to the label type given, we carried out extraction and re-integration of label data at all levels and text preprocessing for the structure data of annex 2 table. After text feature extraction, the SVM classification model was finally selected through the evaluation and comparison of f-score of four classification models: polynomial naive bayes, k-nearest neighbor, support vector machine and random forest. A more accurate classification effect and a higher F1 value are obtained through the evaluation and application of the classification_report model.

Aiming at the second problem, first, we carried out tf-idf feature calculation and selection for the pre-processed text, used gensim to calculate the text similarity, and set the most effective threshold value for document classification by adjusting parameters. We construct heat evaluation indexes according to the number of categories, frequency of occurrence, number of thumb up and number of opposition of each type of problems, and rank hot issues according to the score of the indexes. Secondly, the LDA model was used to extract, and a clear topic was obtained through LDAvis. After extracting the time range of each topic and using Hanlp to identify specific places or groups, the final list of hot issues and comments on hot issues was obtained.

Aiming at the third problem, we constructed five evaluation indexes of response number, response time interval, number of polite expressions, response level and text similarity between reply and message according to the data given. After clustering into appropriate classes by using kmeans clustering method, radar map is used to analyze the cluster after clustering, and classification model is used to predict the quality of new reply text.

**Key words:** LinearSVC;LDA;Hanlp named entity recognition;kmeans

# 目 录

# 第一章：问题描述

## 1.1 问题描述

近年来，随着互联网的飞速发展，政务中社情民意类型的文本数据库迅速增长，为了能更好的、更快速的了解民意，更准确地回复各个类型的问题，留言划分和挖掘热点问题是比较有效的方法。但是采取人工根据经验处理是无法完成如此庞大的工作量的，还会出现效率低、耗时长、差错率高等问题。

大数据、云计算、人工智能等新型技术的出现，让政务服务变得智能化、电子化。较好地解决了上述问题，通过自然语言处理和文本挖掘则可以做到对群众反应的问题进行精准回复、精准解决。极大的提高了政务服务的效率。

对于所要解决的三个问题，我们对所给的文本数据进行了文本清洗和文本表示，提出了文本中比较有特征的信息。对于文本分类问题，通过四种分类器 F-Score 的评价和综合对比选择效果最好的进行分类和应用。对于热点问题的挖掘，我们采用了 gensim+LDA 的方式对文本进行归类并提取主题，再定义具体的热度指标对热点问题排序。最后，对于回复文本的评价，我们定义了五个评价指标并通过 kmeans 聚类和 TOPSIS 分析得到了评价结果。

## 1.2 论文结构安排

第一章：对论文需要解决的问题进行描述，并介绍整篇论文的结构。

第二章：根据“一级标签”对群众的留言进行分类，通过比较多项式朴素贝叶斯、K 近邻、线性分类支持向量机和随机森林四种分类模型各类一级标签的 F1-score 和精确度来选择最优模型，最后进行模型的应用。

第三章：采用 gensim+LDA 的方式对文本进行归类并提取主题，根据文本获取、文本表示、文本归类的流程实现热点问题的识别，定义具体的热度评价指标来实现热点问题的排序。

第四章：具体定义了回复文本字数、回复时间间隔、礼貌用语数、回复等级和回复相似度五个指标。通过 kmeans 聚类得到答复文本类型，并通过分类模型对新的答复文本进行预测。

# 第二章　群众留言分类及评估

对于提取到的群众的类型较多而杂乱的留言文本数据，较人工划分来说，通过机器学习分类模型对留言文本进行分类使留言划分体系更加清晰和精准。本章利用分类模型对比及评估来选取效果最好的分类器来进行分类及应用。

## 2.1 留言文本数据抽取

我们观察了留言文本数据，根据给出的“一级标签”对每一类标签的数据量进行探索，发现数据类别存在一定程度上的失衡，“城乡建设”、“劳动和社会保障”和“教育文体”标签型数据所占比例较大，如图 1。

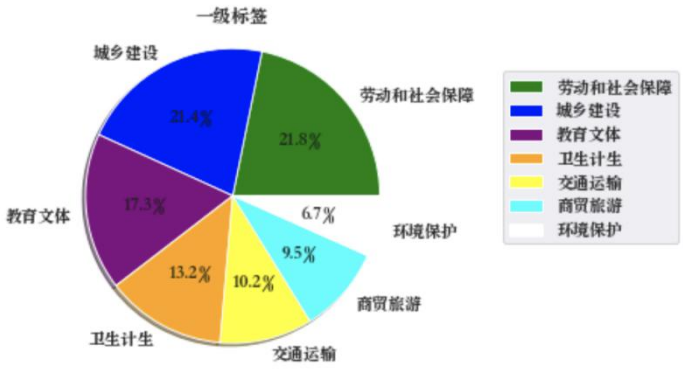


图 1　“一级分类”各标签所占比例图

如果直接采用不均衡数据进行建模预测，预测后的结果更容易偏向所占数据集比例大的标签，对分类器本身产生了“主观”的影响，所以，为了缓解类别失衡对分类器的影响，我们采取了分层抽取数据的方式，探索最小数量的类“交通运输”，以其文本数量 613 为标准，对其它类随机抽取与其相同数量的文本量，最终对不均衡的数据重新组合，得到均衡数据(data_new)。处理后的数据如图 2。

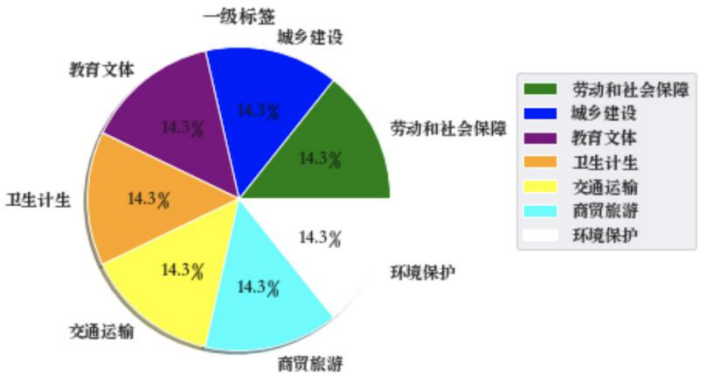


图 2　数据抽取再整合后效果图

## 2.2 分类文本预处理

在缓解类别失衡的问题后，我们接着进行文本的预处理，对必要文本进行筛选和清洗。

### 2.2.1 文本信息筛选

在群众留言文本中，大都是一些中文短文本类型数据，与长文本相比，短文本篇幅较短，一般不超过 200 个字的文本形式，通常只包含几个到几十个有实际意义的词语。所以我们将“留言主题”与“留言详情”合并作为分类依据，用以提高文本的丰富度和特征。

### 2.2.2 文本清洗

在得到合并后的留言内容文本(data_con)后，可以看出文本中存在一些没有意义的字符，如图 3。去除特殊符号以及多余空白，是让短文本的特征表示尽可能地只关注短文本自身词汇的特征和语义本身，降低其他符号对分类准确率的影响。采用正则表达式匹配将这些字符清洗掉，得到数据(data_clean)。如图 4。

```
6150    咨询养老金改革问题\n      \n          养老金改革进行了5年，可，  没有真正实现。...
6513    请求领导解决I5县落水洞煤矿下岗职工实际困难的报告\n      \n      尊敬的贺局长：...
6058    E6县工商局36年工龄的科员为何只享受副主任科员待遇\n      \n          我于15...
5962    关于新生儿办理新农合的问题咨询\n      \n          新生儿父亲是西地省农村户口办理...
5384    咨询外地人口在B市办理失业保险金的问题\n\t\t\t\t\t\n\t\t\t\t\t本人现...
6262    A市生育津贴政策为何迟迟未出台？\n      \n      贺厅长：     请问新的生育政策已经...
6629    请还省电业局全民合同制工人一个公道\n      \n      我1993年元月，被省电业局以...
6628    请I市接收见义勇为英雄人员\n      \n      本人系I市中级人民法院退休干部、三级警...
5743    城镇职工医保与城镇居民医保的报销差额该找哪里补回来？\n      \n      尊敬的市委书...
5828    国企老工人养老金太低！\n      \n          国企老工人他们为国家的建设事业奉献了自...
5638    要求L市企业社会保险处恢复退休养老金标准\n      \n          你好，我是1948 ...
5358    E9县农机公司国营职工的医药费谁负责？\n      \n          我叫马凯平，男，现年4...
5933    J6县肝移植后门诊抗排异药的报销是怎么样的？\n      \n      尊敬的郭县长：     您...
6960    G7县医保处为何终止我的退休职工医保\n      \n      我葛春芳，是西地省G7县华夏...
6051    D市社保局与交通局事故科勾结推诿工伤认定\n      \n          本人父亲于2012年...
7000    请求解决D市工业系统各主管局退休人员的待遇问题\n      \n      致D市人民政府和市...
6510    I市下岗工人的困惑\n      \n          我爱人正退2017-7月但到https:/...
```

图 3　合并后的留言内容文本（部分）

```
6150    咨询养老金改革问题养老金改革进行了5年可没有真正实现国企2005改制时期退休的老工人不仅养老...
6513    请求领导解决I5县落水洞煤矿下岗职工实际困难的报告尊敬的贺局长您好我是I市I5县落水洞煤矿的...
6058    E6县工商局36年工龄的科员为何只享受副主任科员待遇我于15年11月30日向省人社厅反映的问...
5962    关于新生儿办理新农合的问题咨询新生儿父亲是西地省农村户口办理了新农合医保母亲是A市城镇户口也...
5384    咨询外地人口在B市办理失业保险金的问题本人现在处于失业状态符合申领失业保险金条件本人来B市已...
6262    A市生育津贴政策为何迟迟未出台贺厅长请问新的生育政策已经公布这么久产假政策非常明朗为何配套的...
6629    请还省电业局全民合同制工人一个公道我1993年元月被省电业局以D6县劳动局名义招为"全民合同...
6628    请I市接收见义勇为英雄人员本人系I市中级人民法院退休干部三级警司温毅斌因为2016年农历中月...
5743    城镇职工医保与城镇居民医保的报销差额该找哪里补回来尊敬的市委书记您好我叫谢建华系D12市亲仁...
5828    国企老工人养老金太低国企老工人他们为国家的建设事业奉献了自己的一生是他们奠定了国家财富的基础...
5638    要求L市企业社会保险处恢复退休养老金标准你好我是1948年11月出生1965年10月参加工作...
5358    E9县农机公司国营职工的医药费谁负责我叫马凯平男现年48岁系西地省E9县农机公司国营职工于1...
5933    J6县肝移植后门诊抗排异药的报销是怎么样的尊敬的郭县长您好我是盘江乡小风村人去年六月在武汉同...
6960    G7县医保处为何终止我的退休职工医保我葛春芳是西地省G7县华夏有限责任公司职工现年76岁身份...
6051    D市社保局与交通局事故科勾结推诿工伤认定本人父亲于2012年11月16日下班途中因车祸去世D...
7000    请求解决D市工业系统各主管局退休人员的待遇问题致D市人民政府和市委领导请求政府有关部门公平合...
```

图 4　清洗后的留言内容（部分）

## 2.3 jieba 文本分词

中文分词是文本挖掘的基础，对于输入的一段中文，成功的进行中文分词，可以达到电脑自动识别语句含义的效果。现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

分词最常用的工作包是 jieba 分词包，jieba 分词是 python 写成的一个分词开源库，专门用于中文分词。jieba 库分词的三种模式：

精准模式：把文本精准地分开，不存在冗余

全模式：把文中所有可能的词语都扫描出来，存在冗余

搜索引擎模式：在精准模式的基础上，再次对长词进行切分

在本文我们用 jieba 库的精准模式 jieba.lcut 把文本精准地分开，直接生成一个 list (data_cut)。如图 5。

```
6150    [咨询，养老金，改革，问题，养老金，改革，进行，了，5，年，可，没有,...
6513    [请求，领导，解决，I5，县，落水洞，煤矿，下岗职工，实际困难，的，报告...
6058    [E6，县，工商局，36，年，工龄，的，科员，为何，只，享受，副，主...
5962    [关于，新生儿，办理，新农，合，的，问题，咨询，新生儿，父亲，是，西地...
5384    [咨询，外地，人口，在，B，市，办理，失业，保险金，的，问题，本人，...
6262    [A，市，生育，津贴，政策，为何，迟迟，未，出台，贺，厅长，请问，新...
6629    [请，还，省，电业局，全民，合同制，工人，一个，公道，我，1993，年...
6628    [请，I，市，接收，见义勇为，英雄，人员，本人，系，I，市，中级，人...
5743    [城镇职工，医保，与，城镇居民，医保，的，报销，差额，该，找，哪里，补...
5828    [国企，老工人，养老金，太低，国企，老工人，他们，为，国家，的，建设，...
5638    [要求，L，市，企业，社会保险，处，恢复，退休，养老金，标准，你好，我...
5358    [E9，县，农机，公司，国营，职工，的，医药费，谁，负责，我，叫，马...
5933    [J6，县，肝移植，后，门诊，抗，排异，药，的，报销，是，怎么样，的...
```

图 5　分词后的留言内容（部分）

## 2.3 去除停用词

我们看到分词后的文本中，仍有一些像“有”，“的”，“吗”这样的停用词出现，停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words。将这些词扔掉减少了索引量，增加了检索效率，并且通常都会提高检索的效果。于是我们导入停用词表 stoplist.txt 文件，对分词后的留言内容去除停用词，得到 data_after 变量。效果如图 6。

```
6150    [咨询, 养老金, 改革, 养老金, 改革, 年, 国企, 2005, 改制, 时期, 退休...
6513    [领导, 解决, I5, 县, 落水洞, 煤矿, 下岗职工, 实际困难, 报告, 尊敬, 贺...
6058    [E6, 县, 工商局, 36, 年, 工龄, 科员, 享受, 副, 主任科员, 待遇, 我...
5962    [新生儿, 办理, 新农, 合, 咨询, 新生儿, 父亲, 西地省, 农村户口, 办理, 新...
5384    [咨询, 外地, 人口, B, 市, 办理, 失业, 保险金, 处于, 失业, 状态, 符合...
6262    [市, 生育, 津贴, 政策, 迟迟, 未, 出台, 贺, 厅长, 新, 生育, 政策, 公...
6629    [省, 电业局, 全民, 合同制, 工人, 公道, 1993, 年, 元月, 省, 电业局,...
6628    [I, 市, 接收, 见义勇为, 英雄, 人员, 系, I, 市, 中级, 人民法院, 退休...
5743    [城镇职工, 医保, 城镇居民, 医保, 报销, 差额, 找, 补, 回来, 尊敬, 市委书...
5828    [国企, 老工人, 养老金, 太低, 国企, 老工人, 国家, 建设, 事业, 奉献, 一生...
5638    [L, 市, 企业, 社会保险, 处, 恢复, 退休, 养老金, 标准, 1948, 年, ...
5358    [E9, 县, 农机, 公司, 国营, 职工, 医药费, 负责, 马凯平, 男, 现年, 4...
5933    [J6, 县, 肝移植, 门诊, 抗, 排异, 药, 报销, 尊敬, 郭, 县长, 盘江, ...
6960    [G7, 县, 医保, 处, 终止, 退休职工, 医保, 我葛春芳, 西地省, G7, 县,...
```

图 6　去除停用词后的留言内容（部分）

## 2.4 文本特征提取

运用词袋法（Bag of Words）进行文本特征提取。文本分析是机器学习算法的主要应用领域。但是，文本分析的原始数据无法直接丢给算法，这些原始数据是一组符号，因为大多数算法期望的输入是固定长度的数值特征向量而不是不同长度的文本文件。

文本特征提取是将文本数据转换为特征向量的过程，词袋法不考虑词语出现的顺序，每个出现过的词汇单独作为一列特征，这些不重复的特征词汇集合为词表，每一个文本都可以在很长的词表上统计出一个很多列的特征向量，如果每个文本都出现的词汇，一般被标记为“停用词”不计入特征向量。

运用到的 sklearn CountVectorizer。CountVectorizer 是属于常见的特征数值计算类，是一个文本特征提取方法。对于每一个训练文本，它只考虑每种词汇在该训练文本中出现的频率。CountVectorizer 会将文本中的词语转换为词频矩阵，它通过 fit_transform 函数计算各个词语出现的次数。

在 sklearn TfidfVectorizer 中，TfidfVectorizer 基于 TF-IDF 算法。此算法包括两部分 TF 和 IDF，两者相乘得到 TF-IDF 算法。TF-IDF=TF*IDF。TF 为某个训练文本中，某个词的出现次数，即词频(Term Frequency)；IDF 为逆文档频率（Inverse Document Frequency）：

$$\text{词频(TF)} = \frac{\text{某个词在文章中出现的次数}}{\text{(文章的总词数)或(该文出现次数最多的的词的出现次数)}}$$

$$\text{逆文档频率(IDF)} = \log\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}$$

$$\text{TF}-\text{IDF} = \text{词频(TF)} \times \text{逆文档频率(IDF)}$$

特征提取后，我们得到了特征权重矩阵 feature_weight。

## 2.5 分类模型分析

### 2.5.1 多项式朴素贝叶斯（MultinomialNB）

朴素贝叶斯分类器是一种有监督学习，对于一个文档 d，多项式模型中，只有在 d 中出现过的单词，才会参与后验概率计算。

设某文档 d=(t1,t2,…,tk)，tk 是该文档中出现过的单词，允许重复，则：

先验概率 P(c)= 类 c 下单词总数/整个训练样本的单词总数

类条件概率 P(tk|c)=(类 c 下单词 tk 在各个文档中出现过的次数之和+1)/(类 c 下单词总数+|V|)

V 是训练样本的单词表（即抽取单词，单词出现多次，只算一个），|V|则表示训练样本包含多少种单词。P(tk|c)可以看作是单词 tk 在证明 d 属于类 c 上提供了多大的证据，而 P(c)则可以认为是类别 c 在整体上占多大比例(有多大可能性)。

### 2.5.2 K 近邻算法（KNeighborsClassifier）

K 最近邻(k-Nearest Neighbor, KNN)分类算法, 是一个理论上比较成熟的方法, 也是最简单的机器学习算法之一。该方法的思路是：在特征空间中，如果一个样本附近的 k 个最近(即特征空间中最邻近)样本的大多数属于某一个类别，则该样本也属于这个类别。

算法计算步骤：

1、算距离：给定测试对象，计算它与训练集中的每个对象的距离；

2、找邻居：圈定距离最近的 k 个训练对象，作为测试对象的近邻；

3、做分类：根据这 k 个近邻归属的主要类别，来对测试对象分类；

优点：

1、简单，易于理解，易于实现，无需估计参数，无需训练；

2、适合对稀有事件进行分类；

缺点：

1、样本容量较小的类域采用这种算法比较容易产生误分。

2、该方法的另一个不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的 K 个最近邻点。

3、可理解性差，无法给出像决策树那样的规则。

### 2.5.3 线性分类支持向量机（LinearSVC）

LinearSVC 实现了线性分类支持向量机，它是给根据 liblinear 实现的，可以用于二类分类，也可以用于多类分类。罚函数是对截矩进行惩罚。

主要思想是寻找一个超平面，使得两类实例点位于该超平面的两侧。由于这样的超平面不唯一，考虑到模型的鲁棒性，认为超平面应当使得所有实例点到超平面的距离最大。

如若超平面为 wx+b=0，则所有实例点应该满足 yi(wxi+b)-1>=0，此不等式即为约束条件。距离最大就是目标函数 min ||w||^2/2。

说明：

1）LinearSVC 是对 liblinear 的封装

2）liblinear 中使用的是损失函数形式来定义求解最优超平面的，因此类初始化参数都是损失函数形式需要的参数。

3）原始形式、对偶形式、损失函数形式是等价的

### 2.5.4 随机森林（RandomForestClassifier）

随机森林可以看成是 Bagging 和随机子空间的结合。随机森林是由一系列的分类器组合在一起进行决策，期望得到一个最“公平”的学习方法。构造每一个分类器需要从原数据集中随机抽取出一部分样本作为样本子空间，然后再从样本子空间中随机的选取一个新的特征子空间，在这个新空间中建立决策树作为分类器，最后通过投票的方法得到最终决果。

构建单棵决策树：

对于训练集 S，如果训练集数据都属于一个类标签 C,或 S 足够纯净 （85%以上的数据都属于类标签 C 时），否则创建叶子节点，表明类标签 C。否则

・选择“最具有信息”的属性 A

・依据 A 来划分训练集 S

・递归的划分训练集来构造子树

构建随机森林:

・从原始数据中产生 n 个随机抽样

・对于每一个抽样，训练一个未剪枝的决策树，对于每个节点，不是在所有属性中挑选分割最好的决策树，而是在 m 个抽样出来的属性中挑选最好的那个

・对数据集进行预测，并搜集各个树的预测结果，以众数（出现最多的值）给出最后的预测结果

## 2.6 模型评测及对比分析

我们将数据划分成两份，测试集占比 31%，训练集占比 79%。对于四种分类器对训练集的拟合，我们通过将文本特征权重与“一级标签”的预测，得到了各个分类器预测的正确率和耗时时长。如表 1。

表 1　不同模型正确率对比表

| | 正确率 | 耗时（s） |
|---|---|---|
| 多项式朴素贝叶斯 | 0.943 | 2.47 |
| K 近邻 | 0.863 | 1035.67 |
| **线性分类支持向量机** | **0.967** | **0.82** |
| 随机森林 | 0.924 | 6.56 |

暂时从分类器得出的 score 和耗时来说，我们可以看出线性分类支持向量机在正确率和耗时上都是显著更优的，K 近邻模型的表现较差，特别是耗时上。但是单以此表还无法充足说明分类效果。下面我们进行模型的评测:

常见的模型评测指标有准确率( Accuracy) 、精确率( Preci-sion)、召回率(Recall)和 F1 值。下面通过 classification_report 和混淆矩阵进行综合分析 。

四种模型的混淆矩阵如图 7、8、9、10 所示:

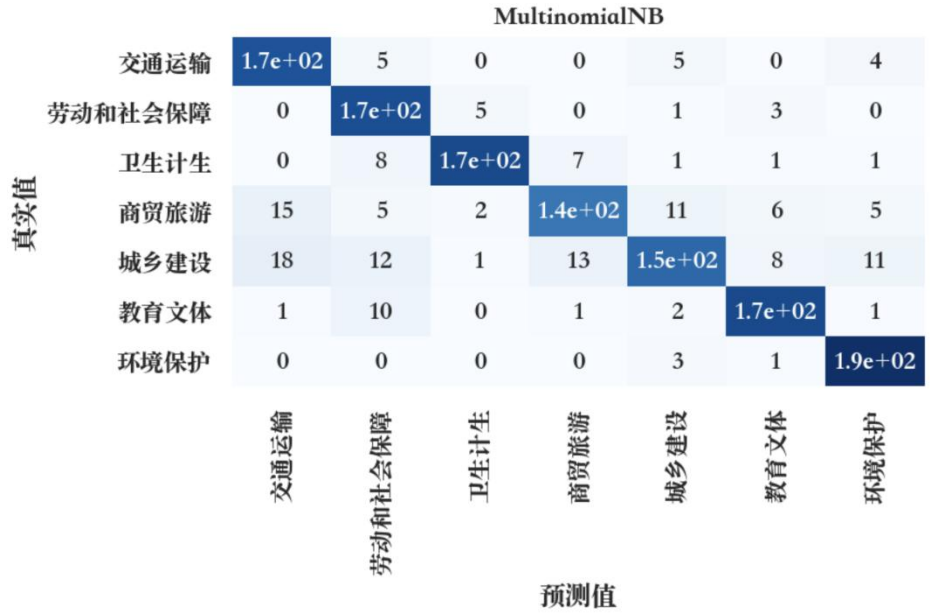


图 7　多项式朴素贝叶斯混淆矩阵

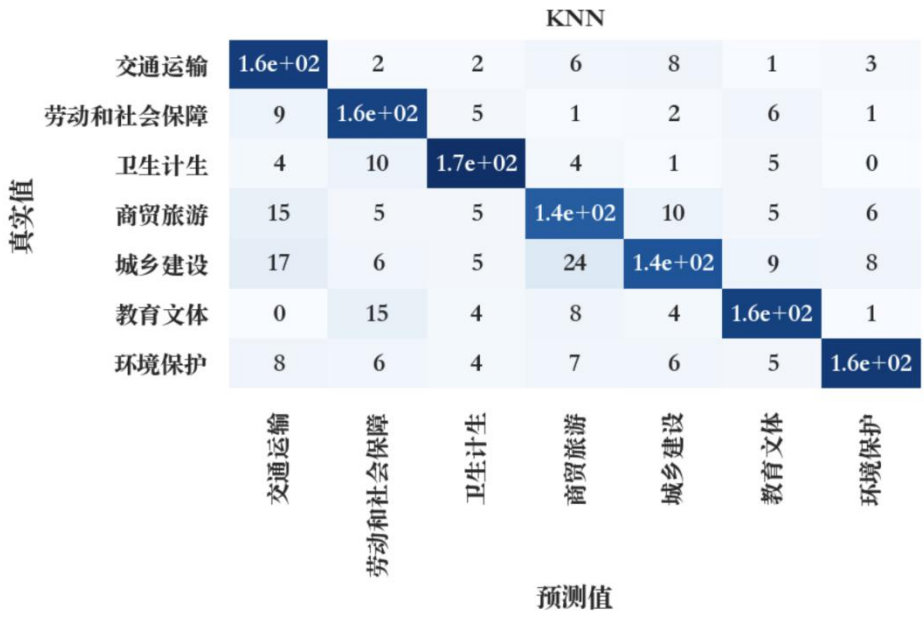


图 8　K 近邻混淆矩阵

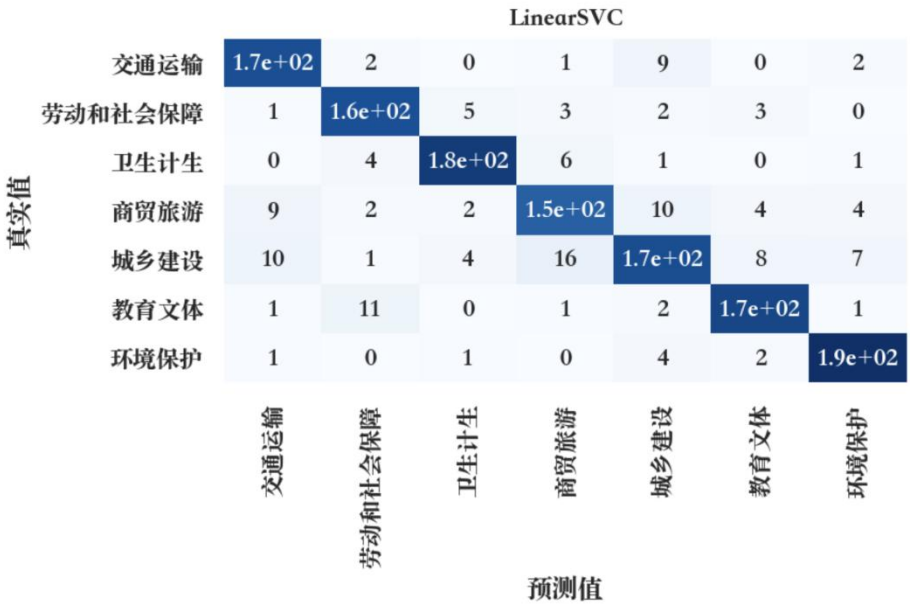


图 9　线性分类支持向量机混淆矩阵

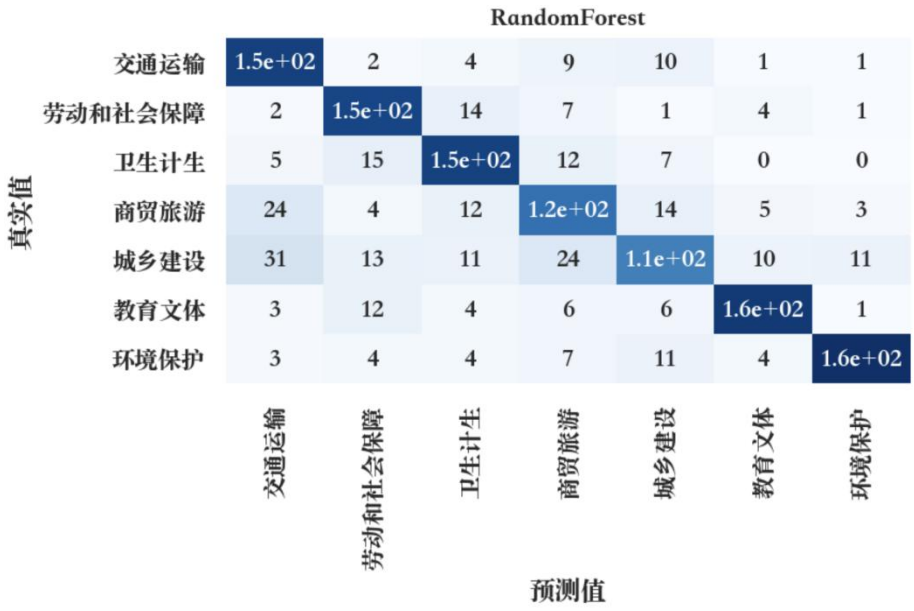


图 10　随机森林混淆矩阵

与二分类混淆矩阵一样，矩阵行数据相加是真实值类别数，列数据相加是分类后的类别数。

我们观察单个分类器内部各个标签的预测的效果，可以看到“环境保护”预测得较为准确，而“商贸旅游”预测的效果较差。而通过多个分类器的比较，可以看出，线性分类支持向量机混淆矩阵的正对角线的数，较其它分类器来说较大，说明预测得较为准确，效果更佳。

接下来我们对比各个标签不同分类器得出的 F1-score。

F1 分数认为召回率和精确率同等重要，当 F1 值较高时则能说明模型比较有效。F1 值( F1 Score) 的计算公式 (其中 P 代表 Precision，R 代表 Recall) ：

$$F1 = 2* P* R P+R$$

不同分类器得出的 F1-score 如表 2。

**表 2　不同分类器的准确率表**

| | F1-score |
|---|---|
| **多项式朴素贝叶斯** | 0.87 |
| **k 近邻** | 0.81 |
| **线性分类支持向量机** | 0.89 |
| **随机森林** | 0.80 |

综合表 2，可以看出，在各个标签中线性分类支持向量机与其它模型相比普遍有较高的 F1 值。同时，我们从 classification_report 可以看出，“环境保护”与“卫生计生”类别预测得效果较好。

## 2.7 结果分析

通过对群众留言的数据抽取、文本预处理、文本表示、模型应用和模型对比评估。我们最终选取了线性可分支持向量机作为分类模型，得到了 96%的正确率，较 KNN、MultinomialNB 和 RandomForest 分类器而言，F1 值分别有 2.4%、10.4%和 4.3%的提高，训练时间也更优，验证了线性可分支持向量机在此类中文短文本多分类问题上的适用性。

# 第三章　热点问题挖掘

为了能让处理的相关部门提升服务效率，及时关注群众反应问题热点，使群众获得有效服务，反应的问题得到及时解决。本章利用 gensim 通过设置阈值对文本进行归类，采用 LDA 的方式提取热点问题主题。提取问题的时间窗口，通过留言数、时间窗口内留言出现的频数、点赞数和反对数来定义具体的热度评价指标。

## 3.1 具体流程


图 11　热点问题挖掘流程图

## 3.2 文本预处理

对于附件 3 样式的留言文本数据的文本预处理，由于和第一章对群众留言数据的处理步骤相同，所以不再做过多阐述。我们同样经过处理可以得到清洗后文本(data_clean)，分词后文本(data_cut)以及去除停用词后文本(data_after)。

## 3.3 Gensim 文本表示与计算文本相似度

Gensim 是一个用于从文档中自动提取语义主题的 Python 库，拥有非常强大的功能。

它致力于处理原生，非结构化的数值化文本(纯文本)。Gensim 中用到的算法，比如 Latent Semantic Analysis(潜在语义分析 LSA)，Latent Dirichlet Allocation(隐含狄利克雷分配)或 Random Projections(随机预测)等等，通过单词在训练语料库的同一文档中的统计共生模式(statistical co-occurrence patterns)来发现文档的语义结构。这些算法都是非监督的算法，无需人工输入。

一旦得到这些统计模式，所有的普通文本就可以被用一个新的、语义代号简介的表示并用其查询某一文本与其它文本的相似性。

Gensim 中元素：

Corpus：数字化文档的集合，被用于自动推断文档的结构和主题等。由此，corpus 也称作 training corpus，被推断的这些潜在结构用于给新的文档分配主题，无需人为介入，也不需要给文档打标签。

Vector：在向量空间模型中，每个文档被表示成了一组特征，比如，一个单一的特征可能被视为一个问答对。

Sparse Vector：通常，大部分问题的答案都是 0，为了节约空间，我们会从文档表示中省略他们。Gensim 不会指定任何特定的 Corpus 格式，不管 Corpus 是怎样的格式，迭代时回一次产生这些 Sparse Vector。

### 3.3.1 训练文本数据

我们可以利用得到的语料(data_after)来构建词典(dictionary)，通过 Doc2bow 稀疏向量来得到语料库(corpus)。

Doc2Bow 是 Gensim 中封装的一个方法，主要用于实现 Bow 模型，该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个

单词的出现都是独立的。BoW 使用一组无序的单词(words)来表达一段文字或一个文档。

接着，通过 TF-IDF 模型算法，得出所有留言的 TF-IDF 值，如图 12。

```
array([[0.00000000e+00, 9.63458584e-02],
       [1.00000000e+00, 6.52254228e-02],
       [2.00000000e+00, 9.63458584e-02],
       ...,
       [1.24800000e+03, 6.67717979e-02],
       [1.24900000e+03, 6.67717979e-02],
       [1.25000000e+03, 2.00315394e-01]])
```

图 12　所有留言的 TF-IDF 二阶矩阵

我们通过 token2id 来查看所得到的特征数(也就是字典里面键的个数)为 1251 个,同时依据稀疏矩阵相似度来建立一个索引(index)。

### 3.3.2 计算各留言相互之间的相似度

我们遍历所有的留言文本，通过对测试数据进行分词和文本向量化，我们可以得到新的稀疏向量(new_vec),再就可以依次计算测试数据和训练数据集中每个文本的文本相似度了,得到了每条测试文本的相似度列表 sims。

## 3.4 文本归类

在得到文本的相似度后，要得到与测试文本相关的训练文本集，就必须将相关度高的文本进行归类。我们通过设置阈值，将多个文本判断为一个文本，来实现反应相似问题的留言文本的归类，并通过调节来找到效果最好的阈值。效果如图 13。例如留言文本 19 中，表示留言编号 188455，195183，208185，219575，229228，240253，276639 和 283655 反应的都是办理异地事宜问题。

```
<<---留言文本19--->>
['咨询异地办理出国签证的问题', '咨询A市潇楚卡异地优惠问题', '咨询A市转业士官异地安置问题', '请问A市何时能落实出入境政策？', '妻子公派国外访学，我能否申请出国探亲假？', '关于外籍人员办理签证事宜过程中的困难', '咨询A市士官转业异地安置时间计算问题', 'A7县黄兴镇居民如何开具全国范围内的无犯罪记录证明？']
归类文本数量:8
2019-05-27 16:04:44


<<---留言文本20--->>
['投诉A市温斯顿英语培训学校拖延退费！', 'A市温斯顿英语梅溪新天地校区何时退费', 'A市梅溪湖温斯顿恶意拖欠退款', 'A市楚通驾校套路学员，学费只能交不能退', 'A市民办培训机构乱象丛生，温斯顿英语培训强设霸王条款', 'A市温斯顿楚府英语恶意拖欠退款', '在A市金茂悦二期艺棵树培训学校交的学费要不回', 'A市温斯顿英语世纪金源校区还没退费给我', 'A市国安驾校退费遭遇霸王条款', '投诉A市温斯顿英语培训机构拖延退费', '投诉A市德睿国际英语拖延退款', 'A市温斯顿英语（梅溪新天地校区）店大服务差，退款难', 'A市温斯顿英语培训机构退费怎么这么难？']
归类文本数量:13
2019-04-15 16:23:09
```

图 13　测试数据的归类效果图(部分)

### 3.5 热度指数的定义

日常生活中我们所指的热点问题是指在某一时期或某一地点出现频率较高的问题，具有引起群众广泛关注、参与讨论、激起民众情绪、引发强烈反响的特点。在此问题中要得到排名前 5 的热点问题，首先要做的就是定义具体的热点指标，再进行热点问题的排序。

在此过程中，我们发现，如果仅从每个问题的时间间隔通过问题出现的次数来定义，存在着不同问题时间间隔不同的影响，不能很好的控制单一变量。对此，我们引入不同问题在不同时间段内出现的频率来定义热点指数。

每条群众留言所对应的点赞数和反对数是其它群众赞同或不赞同的重要依据，它在一定程度上反应了这条留言的真实程度和重要程度。可以给有关部门在处理问题优先级上提供重要依据。

综合以上方面，我们赋给每个指标一定的权重：

引如变量 heat_num[]，赋予权重 0.3，存储每类问题的数量；time_lag[]，存储每类问题的时间间隔(以天为单位)；Freq[]，赋予权重 0.5，存储每类问题出现的频率；赋予权重 0.5；Like[]，Dislike[]，分别赋予权重 0.1 与（-0.1），存储每类问题的总点赞数和总反对数。Heat_value[]，存储每类问题的热点指数，则：

$$Freq[i] = \frac{heat\_num[i]}{time\_lag[i]}$$

$$\text{Heat\_value[i]} =$$

$$\text{heat\_num[i]} \times 0.3 + \text{Freq[i]} \times 0.5 + \text{Like[i]} \times 0.1 - \text{Dislike[i]} \times (-0.1)$$

对得到的热点问题指数，我们保留两位小数并排序，前五热点问题及热度指数如表 3。

表 3　前五热点问题及热度指数表

| 热点问题 | Heat_num | Freq | Like | Dislike | Heat_value |
|---|---|---|---|---|---|
| 1 | 46 | 0.12 | 164 | 1 | 30.16 |
| 2 | 66 | 0.27 | 36 | 3 | 23.24 |
| 3 | 18 | 0.05 | 172 | 2 | 22.43 |
| 4 | 49 | 0.23 | 52 | 2 | 19.82 |
| 5 | 32 | 0.11 | 66 | 3 | 15.96 |

## 3.6 提取时间窗口

我们将前五类热点问题的时间窗口存储到 time_range[]变量中，以天为单位，将时间间隔存储到 time_lag[]变量中。如表 4。

表 4　时间窗口表

| 问题序号 | 时间窗口 |
|---|---|
| 1 | 2019-01-03 至 2020-01-06 |
| 2 | 2019-03-18 至 2019-11-20 |
| 3 | 2019-01-02 至 2020-01-07 |
| 4 | 2019-06-25 至 2020-01-26 |
| 5 | 2019-01-02 至 2019-11-01 |

## 3.7 LDA 提取主题

LDA 是一种挖掘文本潜在主题的概率生成模型，它依据此常识性假设:隐含主题集是由一系列相关特征词组成，文档集合中所有文档均按照一定比例共享隐含主题集合。

具体生成文档的过程如下：

1. 按照先验概率 $p(d_i)$选择一篇文档$d_i$

2. 从 Dirichlet 分布 $\alpha$ 中取样生成文档$d_i$的主题分布 $\theta_i$，主题分布由超参数为 $\alpha$ 的 Dirichlet 分布生成

3. 从主题的多项式分布 $\theta_i$中取样生成文档$d_i$第 j 个词的主题$z_{i,j}$

4. 从 Dirichlet 分布 $\beta$ 中取样生成主题$z_{i,j}$对应的词语分布$\emptyset_{z_{i,j}}$，词语分布$\emptyset_{z_{i,j}}$由参数为 $\beta$ 的 Dirichlet 分布生成

5. 从词语的多项式分布$\emptyset_{z_{i,j}}$中采样最终生成词语$w_{i,j}$

对于归类好的热点问题集，我们将它们放入 LDA 模型中去拟合，并输出前 20 个权重最高的主题词语。如图 14。

```
Topic #6:
车辆 道路 大道 地铁 路口 建议 交通 出行 停车 设置 车道 通行 行人 路段 高速 马路 a7 人行道 红绿灯 城市
Topic #7:
新城 搅拌站 噪音 区丽发 搅拌 万博汇 建发泱玺 污染 喇叭 声音 镇上 粉尘 国道 区亭 发泱玺 百米 拍摄 短缺 女生宿舍 泥土
Topic #8:
变频器 机器 森蒂 维保 垃圾池 报价 印刷 认可 评审 之悦 杂志 紫郡润 不亮 海尔 盛凯 城润 星城润 美郡润 三润 城星澜
Topic #9:
公交 公交车 号线 出行 线路 南路 公交线路 公交站 建议 上班 一趟 希望 木莲 运营 时间 增加 开通 巴士 中路 地铁站
Topic #10:
公司 工作 西地省 户口 平台 有限公司 组织 政策 教师 领导 考试 员工 驾校 村民 2018 工资 安置 资金 投资 诈骗
```

图 14　热点问题主题词语

同时我们利用 LDAvis 来做可视化，来查看每个主题中最相关的前 30 个词语，而某个词语主题的相关性，由**λ**参数来调节，以确定最相关的 30 个术语是出现频率最高的，还是该主题最独特的。

如果**λ**接近 1，那么在该主题下更频繁出现的词，跟主题更相关；

如果**λ**接近 0，那么在该主题下更特殊、更独有的词，跟主题更相关；

我们选择将**λ**调节为 0.8，主要来查看那些出现频率较高的词，例如关于“A5 区劳动东路魅力之城小区油烟扰民”这类问题，红色横条代表该术语在选定主题中出现的频次，而浅蓝色横条代表该术语在语料库中出现的频次。可以看出，代表主题的主要关键词有“小区、油烟、烧烤、影响、居民” 等等。如图 15。

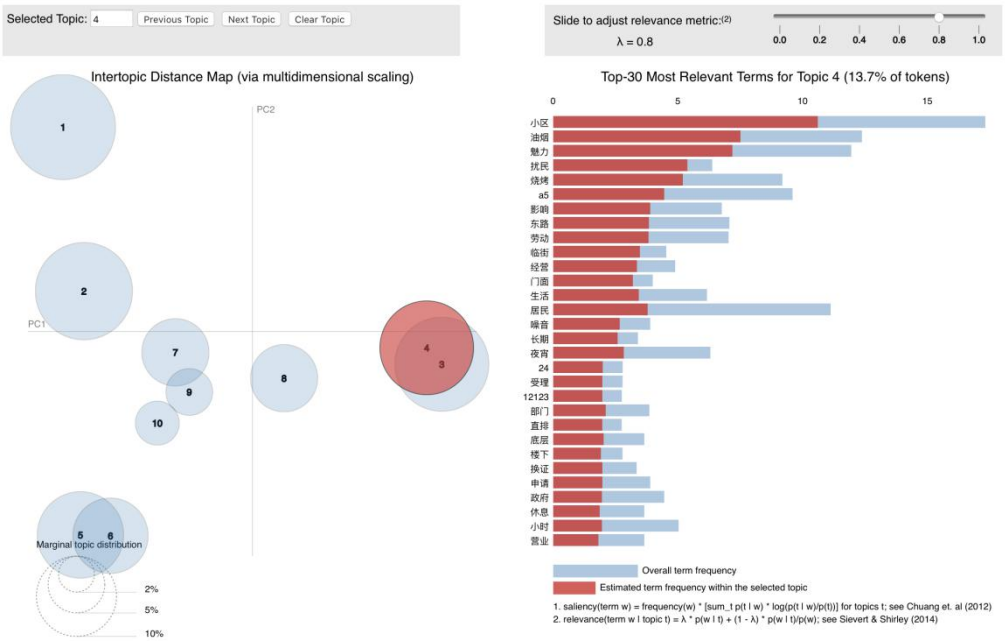


图 15　LDAvis 主题可视化

当点击右边面板的术语时，左边面板代表主题的气泡也会随之发生变化，每个气泡的位置不变，但面积变成由该术语在这些主题上的分布比例决定，图 16

为点击“经营”术语后的可视化结果，可以看出该术语主要出现在主题 4 中，在主题 8 中也占少许。

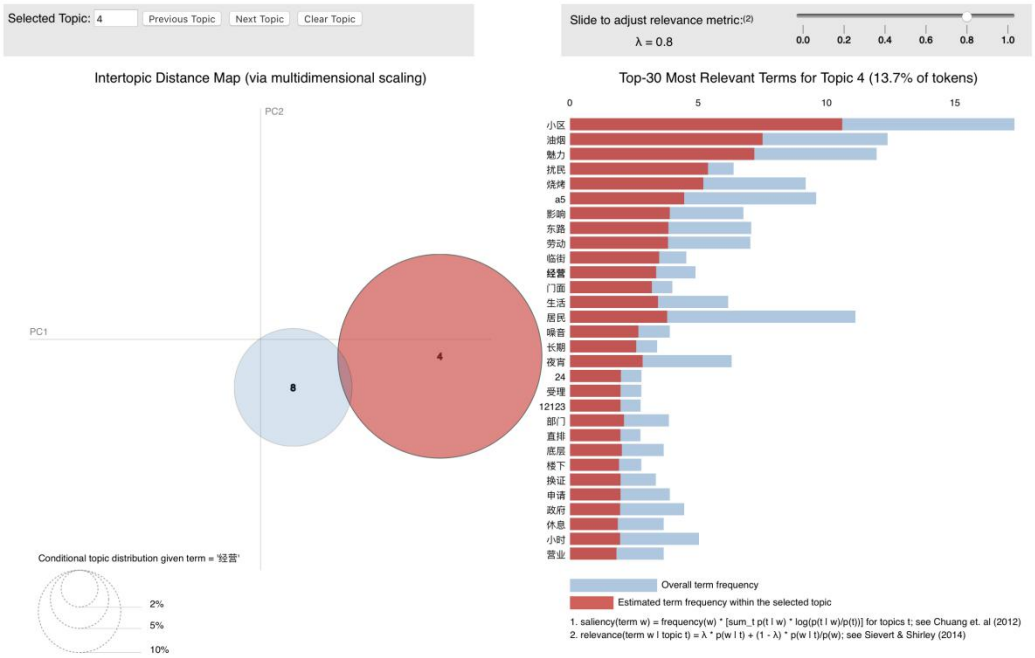


图 16　点击“经营”术语后的 LDAvis 主题可视化

## 3.8 Hanlp 命名实体识别

为了得到每类问题特定的地点或人群，我们选择 Hanlp 这个自然语言处理工具包，因为句法和语义分析依赖于词性标注，而词性标注又依赖于分词，所以对每类问题的问题集，创建流水线(pipeline)，灵活地将多个组件（统计模型或规则系统）组装起来，如图 17。


图 17　命名实体识别流程图

在创建 pipeline 后，我们将流水线序列化存储为 loc.json 文件，如图 18。

```
{
  "id": 29,
  "form": "西地省",
  "cpos": "NR",
  "pos": null,
  "head": 33,
  "deprel": "nsubj",
  "lemma": null,
  "feats": null,
  "phead": null,
  "pdeprel": null
},
```


图 18　pipeline 序列化后的 loc.json 文件

对于每一类识别出来的地点，我们发现存在着地点不具体、不单一的问题，例如“A 市广铁集团武广新城伊景园滨河苑”这类问题，识别的地点可能会不完整，如图 19。所以我们将出现频率最多的地点和人群提取出来汇总，再与其它词性的词合并成完整的地点或人群，其它词性的词主要包括：

“nr”：人名　　“ns”：地名　　“nt”：机构团体


```
[('滨河苑', 'NS', 2, 5), ('广铁', 'NT', 7, 9), ('伊景园', 'NS', 2, 5), ('滨河苑', 'NS', 5, 8), ('
广铁集团', 'NT', 5, 9), ('广铁集团', 'NT', 2, 6), ('伊景园', 'NS', 15, 18), ('滨河苑', 'NS', 18, 21
), ('伊景园', 'NS', 4, 7), ('滨河苑', 'NS', 7, 10), ('武广新城', 'NS', 4, 8), ('A市市政建设开发有限公
司', 'NT', 0, 12), ('广铁', 'NT', 13, 15), ('伊景园', 'NS', 0, 3), ('滨河苑', 'NS', 3, 6), ('广铁
集团', 'NT', 0, 4), ('新城坑', 'NS', 4, 7), ('房伊景园', 'NS', 2, 6), ('滨河苑', 'NS', 6, 9), ('武
广新城', 'NS', 2, 6), ('滨河苑', 'NS', 9, 12), ('广铁集团', 'NT', 13, 17), ('广铁', 'NT', 2, 4), (
'伊景园', 'NS', 10, 13), ('滨河苑', 'NS', 13, 16), ('伊景园', 'NS', 11, 14), ('滨河苑', 'NS', 14,
17), ('A7', 'NS', 5, 7), ('金科天悦', 'NT', 8, 12), ('A7县', 'NS', 0, 3), ('金科天悦小区', 'NS', 3
, 9), ('伊景园', 'NS', 7, 10), ('滨河苑', 'NS', 10, 13), ('城建开发公司', 'NT', 4, 10), ('伊景园',
'NS', 9, 12), ('滨河苑', 'NS', 12, 15), ('伊景园滨', 'NT', 4, 8), ('河苑', 'NS', 8, 10), ('开', 'N
T', 10, 11), ('伊景园', 'NS', 5, 8), ('滨河苑', 'NS', 8, 11), ('武广新城', 'NS', 0, 4), ('A7县黄兴
镇卫生院', 'NT', 0, 9), ('伊景园', 'NS', 14, 17), ('滨河苑', 'NS', 17, 20), ('伊景园滨河院', 'NT', 4
, 10)]
```


图 19　命名实体识别地点(合并前)

## 3.9 热点问题表及留言明细表的归并

在得到前五热点问题的每个问题的热度指数、时间范围、地点或人群以及 LDA 提取出的主题描述后，我们根据热点指数对它们进行降序排序，得到了热点问题表，并存入 csv 文件中，如图 20。之后，对每类问题完整地依次将它们的留言明细信息放入留言明细表，同样存入 csv 文件中。

| | 热度排名 | 问题ID | 热度指数 | 时间范围 | 地点/人群 | 问题描述 |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 30.160 | 2019-01-03至2020-01-06 | A市地铁线延线 | A市地铁施工建设扰民 |
| 1 | 2 | 2 | 23.235 | 2019-03-18至2019-11-20 | A市广铁集团武广新城伊景园滨河苑 | 强行捆绑车位销售给业主 |
| 2 | 3 | 3 | 22.425 | 2019-01-02至2020-01-07 | A7县泉星公园 | 闲置土地公园建设 |
| 3 | 4 | 4 | 19.815 | 2019-06-25至2020-01-26 | A市暮云街道丽发新城社区搅拌厂 | 小区附近搅拌站扰民 |
| 4 | 5 | 5 | 15.955 | 2019-01-02至2019-11-01 | A市 | 加快建设力度 |

图 20　热点问题表 csv 文件

可以看出，热度指数最高的是“A 市地铁施工建设扰民”，相对热点问题二，问题的数量和出现频率虽然不及，但是群众的点赞数尤其高，反对数较少，反映了此类问题的“流量”较大，是这段时间以来群众一直关注、亟待解决的问题。同样，热点问题三相对与问题四和问题五，虽然问题频率较低，但是群众的点赞数相对很高。

如表 5, 对于像热点问题一和三这样出现时间不集中但是点赞数较高的类型, 主要体现了群众关注的长期性, 一直以来都是群众十分关心确没有得到及时解决的问题，我们称这类问题为“持续关注型”问题。而对于出现时间较为集中，留言数量较多的热点问题二、四、五。我们称这类问题为“突发热点型”问题。这两种类型的问题都是有关部门亟需解决的问题，而总的来说，有关部门可以根据问题的热度指数来决定处理问题的缓急程度。

表 5　前五热点问题及热度指数表

| 热点问题 | Heat_num | Freq | Like | Dislike | Heat_value |
|---|---|---|---|---|---|
| 1 | 46 | 0.12 | 164 | 1 | 30.16 |
| 2 | 66 | 0.27 | 36 | 3 | 23.24 |
| 3 | 18 | 0.05 | 172 | 2 | 22.43 |
| 4 | 49 | 0.23 | 52 | 2 | 19.82 |
| 5 | 32 | 0.11 | 66 | 3 | 15.96 |

# 第四章：答复意见评价

对于群众反应的问题，有关部门给出的答复的质量是值得去评价的，答复内容是否相关，是否完整；答复是否及时，是否礼貌等等都是非常重要的评价指标。政务服务是一种无形的劳动，服务能直接满足服务对象的某种需求，没有中间转换环节。因此，每一项服务工作、每一个服务过程都必须达到服务质量要求，才能实现服务对象满意。

在本章中，我们对建立的答复意见评价指标进行聚类分析。通过 kmeans++ 的聚类方法对不同的答复进行聚类并分析各类特点。最后以聚类的结果作为标签，使用分类模型对新的答复意见文本进行预测。

## 4.1 答复意见评价指标定义

### 4.1.1 回复字数

我们首先对每条答复意见进行预处理，去掉多余的不必要的空白符和标点符号，统计有效的答复内容字数，通过回复字数的多少，可以在一定程度上反应答复的完整性。引入变量 Len[]，用来存储每条回复意见的字数。

### 4.1.2 回复时间间隔

回复时间间隔是反映政务答复是否及时的一个重要指标，如果是突发热点事件或话题，及时的处理与回复是对群众问题的一个及时参考和重要解决方案。如果回复过慢，群众等待时间长，很容易引起更大范围的影响。我们以天为单位，计算每条留言得到的答复的时间间隔，并存储到 Time_lag 变量中。

### 4.1.3 礼貌度

我们观察了附件 4 给出的答复意见数据，发现完整的礼貌文本格式如图 21 所示。对此，我们对文本分词后，导入礼貌用语词.txt 文件，遍历每一条答复意见文本，统计每条文本中礼貌词语的数量，并存入 Polite_num 变量中。

网友***，**您好**/现将网友***在***反映******的问题向该网友**答复如下**：**您好**，首先**感谢您**对我们工作的信任与支持，关于**您******的留言，反映******的情况**已收悉**，现*****答复如下**：

**……**

**……**

再次**感谢您**的理解和关心。

****年**月**日

图 21　礼貌文本格式

### 4.1.4 答复等级

我们创建了答复等级标签，对每条答复意见文本进行了人工标注，标注分为1、2、3，分别对应好、中、差三个等级，具体的定义方法如下：

**好**：答复文本对群众反应的问题作出了详细的解释，在解释后根据完整具体的法律条文或相应的政策给出详细的解释方案，承诺对此进行及时处理或已处理完毕，并说明了这样解决的原因与改善。

**中**：答复文本对群众反应的问题作出了详细的解释，在解释后根据法律条文或相应的政策给出了解释方案和原因。但未作出具体的解决方案。

**差**：答复文本未对群众反应的问题进行解释，没有提出有效的解决方案，仅仅只是“已收悉”、“已转移相关部门处理”或答复文本极其不完整。

标注后的附件 4 如图 22 所示：

| 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 | 答复等级 |
|---|---|---|---|---|---|---|---|
| 2549 | A00045581 | 区景蓉华苑物业管理有问 | 2019/4/25 9:32:09 | 公司却以交20万保证金，不 | 停车管理费，在业主大会结束后业委会也 | 2019/5/10 14:56:53 | 3 |
| 2554 | A00023583 | 篇楚南路洋湖段怎么还没 | 2019/4/24 16:03:40 | 面的生意带来很大影响，里 | 整体换填，且换填后还有三趟雨污水管 | 2019/5/9 9:49:10 | 2 |
| 2555 | A00031618 | 提高A市民营幼儿园老师 | 2019/4/24 15:40:04 | 司时更是加大了教师的工作 | 幼儿园聘任教职工要依法签订劳动合同， | 2019/5/9 9:49:14 | 2 |
| 2557 | A000110735 | 公寓能享受人才新政购房 | 2019/4/24 15:07:30 | 户A市，想买套公寓，请问 | 年龄35周岁以下（含），首次购房后，可 | 2019/5/9 9:49:42 | 2 |
| 2574 | A0009233 | A市公交站点名称变更的 | 2019/4/23 17:03:19 | 为“马坡岭小学”，原“马坡岭 | 马坡岭”的问题。公交站点的设置需要方便 | 2019/5/9 9:51:30 | 1 |
| 2759 | A00077538 | A3区含浦镇马路卫生很差 | 2019/4/8 8:37 | 再把泥巴冲到右边，越是上 | 您问题中没有说明卫生较差的具体路段， | 2019/5/9 10:02:08 | 2 |
| 2849 | A000100804 | 教师村小区盼望早日安装 | 2019/3/29 11:53:23 | 为老社区惠民装电梯的规范 | A3区人民政府办公室下发了《关于A市A | 2019/5/9 10:18:58 | 2 |
| 3681 | UU00812 | 区东澜湾社区居民的集体 | 2018/12/31 22:21:59 | 好远，天寒地冻的跑好远， | 修前期准备及设施设备采购等工作。下一 | 2019/1/29 10:53:00 | 3 |
| 3683 | UU008792 | 篇阳光住宅楼无故停工以 | 2018/12/31 9:55:00 | 没得到相关准确开工信息。 | 落实分户检查后，西地省楚江新区建设工 | 2019/1/16 15:29:43 | 3 |
| 3684 | UU008687 | 和顺路洋湖壹号小区路段 | 2018/12/31 9:45:59 | 交桥等地方做立体绿化，取 | 分也按规划要求完成了建设，其中西边绿 | 2019/1/16 15:31:05 | 2 |
| 3685 | UU0082204 | 2区大托街道大托新村违 | 2018/12/30 22:30:30 | 规划局审批通过《温室养殖 | 付一笔耕地征收补偿款给原大托村，但截 | 2019/3/11 16:06:33 | 3 |
| 3692 | UU008829 | 阳村D区安置房人防工程 | 2018/12/29 23:27:51 | 区安置房地下室近两万平方 | 续，按长人防发[2014]7号文件要求，鄱阳 | 2019/1/29 10:52:01 | 3 |
| 3700 | UU00877 | 段请求修建一座人行天 | 2018/12/29 11:55:34 | ，大量从小区开车出去的业 | 局配合进行具体选址，招标（邀标）进 | 2019/1/14 14:34:58 | 1 |
| 3704 | UU0081480 | 市芒果金融平台涉嫌 | 2018/12/28 17:18:45 | 省相关政府部门的大力支持 | 的相关警情，已由银盆岭派出所立刑事案 | 2019/1/3 14:03:07 | 1 |
| 3713 | UU0081227 | 建议增开A市261路公交车 | 2018/12/28 7:53:25 | 时以上！天寒地冻，其他公 | 。由于驾驶员工作时间长，劳动强度大， | 2019/1/14 14:33:17 | 2 |
| 3720 | UU008444 | 路与披塘路交叉路口通行 | 2018/12/27 15:18:07 | ：https://baidu.com/。但其 | 的“披塘路路口两端各拆除20米中间花坛， | 2019/3/6 10:26:14 | 3 |
| 3727 | UU0081194 | 区桐梓坡路益丰大药房以 | 2018/12/27 1:55:21 | 更以各种理由拒绝退货，并 | 据您提供的信息进行投诉信息的登记分 | 2019/1/3 14:02:47 | 2 |
| 3733 | UU008706 | 在A市梅溪湖开办一个图 | 2018/12/26 16:51:40 | 称。建议在艺术中心先期借 | 业。梅溪湖二期金菊路与雪松路东南角 | 2019/1/14 14:32:40 | 1 |

图 22　标注后的文本数据（部分）

### 4.1.5 文本相似度

回复和提问的文本相似度是衡量答复内容的相关性的重要指标，在本章中我们选择了利用余弦相似度算法来计算文本相似度。

余弦相似度算法：如何计算回复和提问的文本相似度。我们可以把它们想象成空间中的两条线段，都是从原点（[0, 0, ...]）出发，指向不同的方向。两条线段之间形成一个夹角，如果夹角为 0 度，意味着方向相同、线段重合,这是表示两个向量代表的文本完全相等；如果夹角为 90 度，意味着形成直角，方向完全不相似；如果夹角为 180 度，意味着方向正好相反。因此，我们可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。

计算公式为：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(xi \times yi)}{\sqrt{\sum_{i=1}^{n}(xi)^2} \times \sqrt{\sum_{i=1}^{n}(yi)^2}}$$

具体流程：

分词->列出所有词->分词编码->词频向量化->套用余弦函数计量两个句子的相似度

我们引入 word_vector1 与 word_vector2，来用给定形状和类型的用 0 填充的矩阵存储向量，并套入统计词频函数中并输出。最后套用余弦函数计量两个句子的相似度。

我们引入 Score 变量，来存储每条留言文本与答复文本的相似度，以答复文本第一条为例，提问与答复文本向量如图 23，带入函数后得到相似度 0.615。

```
[ 0.  1.  1.  0.  0.  1.  1.  9. 10.  1.  2.  1.  5.  1.  1.  1.  1.  0.
  0.  0.  0.  1.  1.  0.  0.  1.  0.  2.  1.  0.  0.  3.  0.  2.  0.  1.
  0.  0.  0.  1.  8.  1.  4.  1.  1.  3.  2.  1.  1.  3.  1.  0.  0.  1.
  2.  1.  1.  0.  3.  0.  0.  1.  1.  1.  1.  1.  1.  0.  1.  0.  0.  1.
  0.  1.  0.  1.  1.  5.  6.  1.  1.  1.  3.  1.  1.  0.  1.  0.  1.  1.
  1.  1.  0.  0.  1.  1.  1.  0.  1.  0.  3.  1.  0.  1.  3.  0.  0.  0.
  1.  1.  0.  0.  3.  2.  1.  5.  0.  2.  2.  0.  0.  0.  1.  1.  0.  1.
  1.  1.  1.  0.  1.  1.  0.  1.  5.  1.  1.  0.  0.  0.  1.  0.  0.  0.
  1.  1.  0.  1.  1.  1.  0.  2.  0.  1.  0.  0.  1.  1.  0.  0.  1.  0.
  0.  1.  1.  1.  1.  1.  0.  1.  4.  0.  0.  6.  0.  2.  1.  1.  1.  0.
  0.  0.  3.  1.  1.  1.  1.  7.  2.  1.  1.  1.  1.  1.  0.  1.  1.  0.
  0.  4.  2.  0.  1.  3.  0.  1.  0.  0.  0.  2.  1.  2.  1.  0.  0.  0.
  1.  0.  3.  1.  1.  0.  0.  1.  0.  0.  1.  0.  0.  0.  1.  1.  1.  0.
  0.  0.  1.  2.  0.  1.  1.  1.  1.  2.  0.  1.  1.  8.  5.  2.  1.  1.
  1.  1.  0.  6.  1.  0.  2.  0.  6. 13.  0.  4.  1.  0.  0.  1.  1.  0.
  0.  0.  4.  3.  1.]
[ 4.  0.  0.  1.  1.  3.  0. 10.  2.  2.  1.  0.  5.  0.  0.  3.  0.  2.
  1.  1.  1.  0.  0.  4.  2.  1.  1.  0.  0.  2.  1.  4.  1.  0.  2.  0.
  1.  4.  2.  3.  0.  0.  0.  0.  0.  1.  0.  2.  1.  2.  0.  3.  2.  2.
  1.  1.  1.  2. 12.  5.  2.  1.  1.  3.  0.  0.  1.  1.  0.  2.  1.  2.
  2.  0.  1.  2.  1.  5.  4.  2.  0.  2.  2.  0.  0.  2.  0.  1.  3.  0.
  0.  1.  1.  3.  0.  3.  0.  1.  4.  1.  5.  0.  1.  0.  0.  1.  1.  1.
  0.  0.  3.  1.  2.  0.  0.  8.  1.  3.  0.  2.  1.  1.  0.  3.  5.  0.
  0.  4.  0.  2.  5.  0.  3.  0.  2.  0.  0.  2.  1.  2.  0.  3.  1.  3.
  2.  2.  2.  0.  0.  1.  1.  6.  1.  1.  1.  2.  2.  0.  2.  2.  0.  1.
  1.  2.  2.  1.  1.  3.  2.  0.  0.  2.  2.  5.  2.  1.  0.  0.  2.  1.
  1.  1.  0.  1.  0.  0.  1.  0.  1.  0.  1.  0.  0.  2.  3.  0.  1.  3.
  1.  1.  2.  2.  0.  0.  1.  4.  3.  1.  1.  0.  2.  1.  3.  4.  2.  1.
  2.  2.  3.  0.  0.  1.  3.  0.  2.  1.  0.  1.  1.  1.  3.  1.  3.  2.
  1.  2.  0.  0.  1.  3.  0.  0.  0.  2.  1.  0.  6. 13.  1.  0.  4.  0.
  0.  0.  2.  0.  2.  1.  0.  1.  2. 17.  1.  3.  0.  1.  1.  0.  0.  2.
  1.  2.  1.  1.  0.]
```

图 23　提问与答复文本向量

## 4.2 指标聚合

对于定义好的五类指标，我们对它们进行聚合得到 Data1，如图 24。统计标准化后的数据，如图 25。

| | 回复字数 | 回复时长 | 答复等级 | 礼貌用语数 | 相似度 |
|---|---|---|---|---|---|
| **0** | 418 | 15 | 3 | 6 | 0.615115 |
| **1** | 278 | 14 | 2 | 5 | 0.255307 |
| **2** | 327 | 14 | 2 | 7 | 0.588239 |
| **3** | 271 | 14 | 2 | 12 | 0.362846 |
| **4** | 139 | 15 | 1 | 5 | 0.492625 |
| **5** | 216 | 31 | 2 | 10 | 0.129164 |
| **6** | 226 | 40 | 2 | 4 | 0.399964 |
| **7** | 578 | 28 | 3 | 6 | 0.394717 |
| **8** | 468 | 16 | 3 | 6 | 0.425796 |

图 24　聚合后的指标

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **count** | 2816.000000 | 2816.000000 | 2816.000000 | 2816.000000 | 2816.000000 |
| **mean** | 0.045047 | 0.156379 | 0.602450 | 0.173485 | 0.438229 |
| **std** | 0.054731 | 0.166885 | 0.360182 | 0.105102 | 0.177259 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 0.017720 | 0.040404 | 0.500000 | 0.100000 | 0.319598 |
| **50%** | 0.034165 | 0.111111 | 0.500000 | 0.166667 | 0.447406 |
| **75%** | 0.055430 | 0.212121 | 1.000000 | 0.233333 | 0.563690 |
| **max** | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

图 25　标准化后的指标数据

## 4.3 Kmeans++答复文本聚类

K-Means 算法的思想比较简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。

用数据表达式表示，假设簇划分为(1,2,...)(C1,C2,...Ck)，则我们的目标是最小化平方误差 E：

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} ||x - \mu i||_2^2$$

其中μi 是簇 Ci 的均值向量，有时也称为质心，表达式为：

$$\mu i = \frac{1}{|Ci|} \sum_{x \in Ci} x$$

k 个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的 k 个质心。如果仅仅是完全随机的选择，有可能导致算法收敛很慢。K-Means++算法就是对 K-Means 随机初始化质心的方法的优化。

K-Means++的对于初始化质心的优化策略如下：

1. 从输入的数据点集合中随机选择一个点作为第一个聚类中心μ1

2. 对于数据集中的每一个点 xi, 计算它与已选择的聚类中心中最近聚类中心的距离 $D(x) = \arg\min||x - \mu r||_2^2$， $r = 1,2...k_{selected}$

3. 选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大

4. 重复 b 和 c 直到选择出 k 个聚类质心

5. 利用这 k 个质心来作为初始化质心去运行标准的 K-Means 算法

我们对原始数据进行归一化到[0,1]区间后，将转化后的数据进行 PCA 降维，再代入 kmeans++中拟合。

其次选取不同的聚类数，查看不同聚类个数得到的聚类效果。我们发现，当聚类数为 4 时，聚类的效果最好。轮廓系数最高，为 0.70。如图 26。

聚类后，我们得到四类聚合的数量分别为 1174、1000、500 和 142 个。引入 text0，text1，text2 和 text3 分别存储四类答复文本的留言编号。

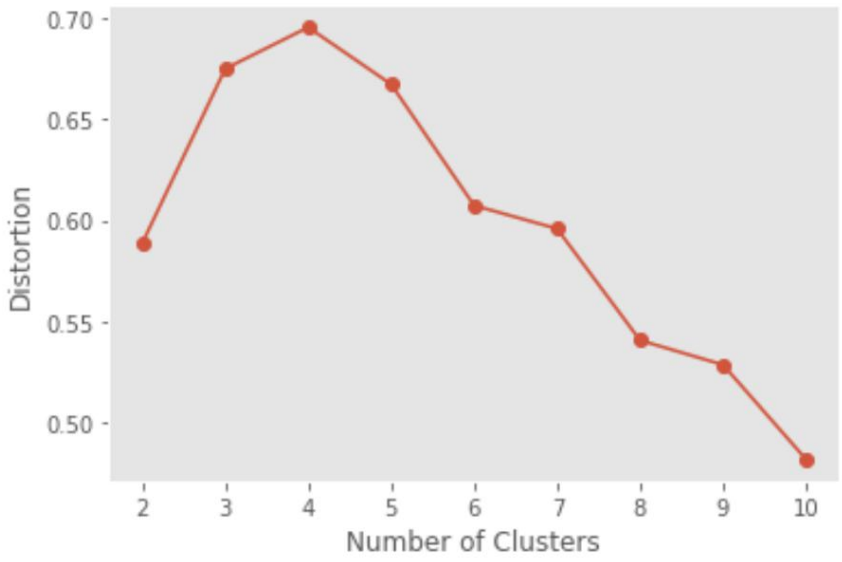


图 26　标准化后的指标数据

## 4.4 聚类结果分析

对于得到的聚类数据，我们将每类的聚类中心点放入雷达图中进行绘制。通过对比不同类的回复字数、回复时长、礼貌用语数、答复等级和相似度五类指标，来分析不同类的特点。聚类雷达图如图 27。

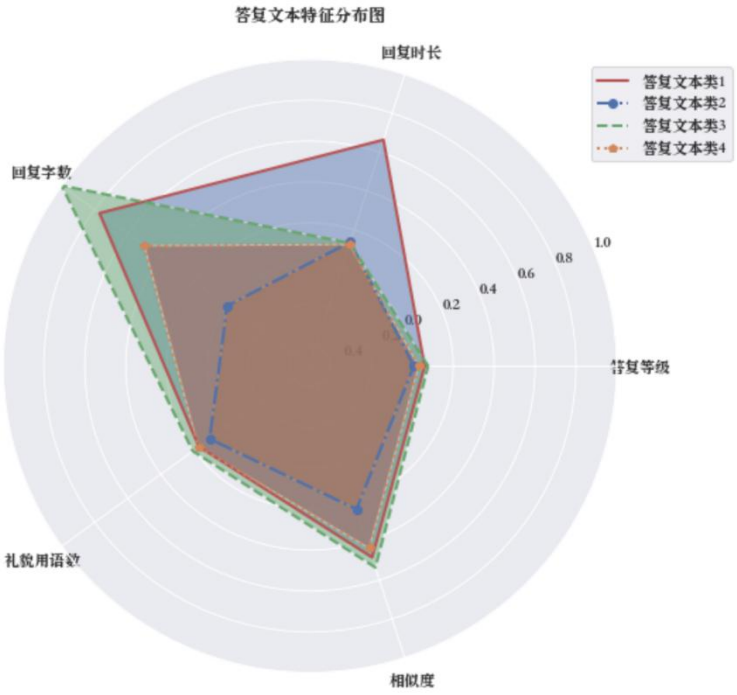


图 27　四类答复文本特征分布图

**我们将留言文本分为 4 类：**

答复文本类一：回复字数、礼貌用语、相似度、答复等级表现均不错，回复时长表现差。

答复文本类二：回复字数、礼貌用语、相似度、答复等级表现均差，回复时长表现佳。

答复文本类三：回复字数、礼貌用语、相似度、答复等级和回复时长表现均佳。

答复文本类四：礼貌用语、相似度、答复等级和回复时长表现均不错，回复字数表现较差。

**我们将上述 4 类留言文本进行具体描述：**

第一类留言文本定义为：延误型较高质量回复

我们发现，虽然此类文本回复的完整性、可解释性和与提问问题的相关性较为不错，但是存在着回复时间过长的问题，对于一些突发热点性问题来说，可能存在这没有得到及时的回复而错过了最佳解决时间的问题。

第二类留言文本定义为：低质量回复

此类文本回复虽然时间较短，但是存在着回复不完整（例：具体的解决方案链接没有给出）；回复相似度不高（例：回复的只是时间）的问题。回复的内容没有针对性，没有给出具体的原因，没有给群众带来有效的解决方案。

第三类留言文本定义为：及时型高质量回复

此类文本回复在五类指标都有较佳的表现。符合完整的规范格式，答复内容与提问问题高度相关。对于群众提出的问题礼貌得作出了解释，并提出了问题的原因，较为及时地给出了有效的解决方案。

第四类留言文本定义为：及时型中质量回复

此类文本在回复时间上是最为及时的，但在内容上，此类回复只是提出了一些建议和法律政策条文依据。并没有针对群众反应的问题提出有效的解决方案或承诺。所以此类文本虽然有一定的参考性，但是对于群众问题的解决还需要一些实质的方案。

## 4.5 答复意见评价类型预测

在得到聚类结果后，我们将结果添加到附件 4 中，形成文本类型标签，接着划分附件 4 答复意见文本数据，将 30%的文本作为测试集，将 70%的文本作为训练集。

作为分类模型，线性可分支持向量机在此类问题上具有不俗的效果，通过间隔最大化或者等价的求出相应的凸二次规划问题得到的分离超平面$w^* \cdot x + b^* = 0$,以及决策函数：$f(x) = sign(w^* \cdot x + b^*)$

$|w \cdot x + b|$描述了点 x 距离超平面的远近，对于正确分类的点来说，这个式子与函数间隔是相等的。因此，函数间隔可以表示分类预测的正确性和确信度。

对于多分类的问题，可将带求解的多分类的问题转化为二分类问题的延伸；即将多分类任务拆分为若干个二分类任务的求解，具体方法如下：

OVO(One-Versus-One)：一对一

OVA/OVR(One-Versus-All/One-Versus-the-Rest)：一对多

Error Correcting Output Codes(纠错码机制)：多对多

将二分类算法推广到多分类算法的一种常见的方法是“一对多余”，在“一对多余”中对每个类别都对应二分类模型，这样每个类别都有一个系数(w)和一个截距(b)，将这个类别与所有其它类别尽量分开，这样就生成了类别个数同样多的二分类。

我们将训练集放入线性可分支持向量机中拟合，进行模型的评估。对于新的答复文本，我们对其进行分词和特征提取后，就可以通过训练后的模型进行预测，得到属于的答复文本类型，并且根据类型的特点可以知道此条新的答复文本存在的问题以便进行改善。

# 参考文献

[1]王传廷. 基于自然语言处理与非负矩阵分解的中文文本分类研究[D].武汉理工大学,2009.

[2]刘金岭,王新功.基于中文短信文本聚类的热点事件发现[J].情报杂志,2013,32(02):30-33.

[3]段丹丹,唐加山,温勇,袁克海.基于 BERT 的中文短文本分类算法的研究[J/OL].计算机工程: 1-12[2020-04-15]

[4]李霄野,李春生,李龙,张可佳.基于 LDA 模型的文本聚类检索[J].计算机与现代化,2018(06):7-11.

[5]刘春磊,梁瑞斯,邸元浩.基于 TFIDF 和梯度提升决策树的短文本分类研究[J].科技风,2019(24):231.

[6]冯勇,屈渤浩,徐红艳,王嵘冰,张永刚.融合 TF-IDF 和 LDA 的中文 FastText 短文本分类方法[J].应用科学学报,2019,37(03):378-388.

[7]陈莉萍,杜军平.突发事件热点话题识别系统及关键问题研究[J].计算机工程与应用,2011,47(32):19-22.

[8]陈海利,孙志伟,庞龙.基于随机森林的文本分类研究[J].科技创新与应用,2014(02):55.

[9]陈俊宇,郑列.基于 R 语言的商品评论情感可视化分析[J].湖北工业大学学报,2020,35(01):110-113.

[10] gensim 文档[DB/OL]. https://radimrehurek.com/gensim/auto_examples/index.html#