

基于自然语言处理的智慧政务文本挖掘应用

摘要

近年来,多种网络问政平台逐步发展,网络问政已成为了市民反映社情民意,政府聚民智、通民意、解民惑的重要渠道,密切了政府与人民群众的关系。建立基于自然语言处理的智慧政务系统是社会与科技发展的必然趋势,对提升政府的管理水平和施政效率具有极大的推动作用。

对于问题 1,原始数据预处理后,利用 TF-IDF 算法输出文本向量,再通过 LSA 进行降维后分别对文本分类中常见的机器学习算法进行训练与测试,对比八种算法的准确度、F1-score 和所需时间,选择出表现最优的 LinearSVC 算法,进行参数调优后运用未降维前的文本向量通过交叉验证的方法训练模型,得到最终 F1-score 为 0.88 的分类模型。

对于问题 2,对附件三的留言主题和留言详情进行数据预处理后进行特征提取,使用 LSA 对数据进行降维,再利用 K-距离曲线确认 DBSCAN 聚类参数进行聚类,结合聚类结果计算每类的 Reddit 和时间范围,最后根据定义的热度指标公式:话题热度值=该话题的 Reddit*该话题的分布率 得到每类话题的热度值,根据热点问题的两个条件筛选出某一时间段特定人群集中反映的热点问题。

对于问题 3,利用余弦相似度计算答复与留言的相关度,使用正则表达式提取关键词判断答复是否具有完整性、可解释性,并统计答复意见的字数和答复与留言的时间间隔,以此作为判断答复意见时效性和重视度的要素,最后利用层次分析法对五个性质进行加权计算得分,综合评价答复意见的质量。

关键词: TF-IDF 机器学习 LSA DBSCAN 聚类 Reddit 算法 余弦相似度 层次分析法

基于自然语言处理的智慧政务文本挖掘应用

摘要	1
1、挖掘目标.....	3
2、分析过程与结果.....	3
2.1 问题 1 分析过程与结果.....	3
2.1.1 数据预处理.....	4
2.1.1.1 地点标签提取.....	4
2.1.1.2 中文分词.....	4
2.1.1.3 去除停用词.....	5
2.1.2 TF-IDF 算法.....	6
2.1.3 LSA 降维	6
2.1.4 机器学习.....	8
2.1.5 最优模型的选择	9
2.1.6 模型参数最优化	10
2.1.7 分类结果与模型评价	11
2.2 问题 2 分析过程与结果.....	13
2.2.1 挖掘流程示意图与名词解释.....	13
2.2.1.1 挖掘流程框架.....	13
2.2.1.2 名词解释.....	13
2.2.2 数据预处理.....	14
2.2.3 DBSCAN 聚类.....	14
2.2.3.1 文本特征抽取.....	15
2.2.3.2 DBSCAN 聚类.....	16
2.2.4 计算 Reddit	18
2.2.5 计算分布率.....	20
2.2.6 话题热度值.....	22
2.2.7 热点问题.....	23
2.3 问题 3 分析过程与结果.....	24
2.3.1 答复的相关性	24
2.3.2 答复的完整性和可解释性	24
2.3.3 答复的时效性	24
2.3.4 答复的重视度	25
2.3.5 答复意见评级	25
2.3.6 评价方案的实现	27
3、参考文献.....	30

1、 挖掘目标

本次建模目标，是利用群众的留言的主题、留言详情以及一级分类标签，在对文本信息进行分词、停用词过滤等基本的预处理后，采用 TF-IDF 算法进行文本数据向量化以及 LSA 算法实现降维，通过机器学习模型建立关于留言内容的一级标签分类模型；通过 DBSCAD 聚类、Reddit 算法实现对某一时间段内某一被集中反映的热点话题的挖掘，以便有关部门及时处理，通过改进 Reddit 算法和时间范围定义热度指标，实现热点话题的排名。针对有关部门的回复信息，通过词频实现文本向量化，计算留言信息与回复信息的余弦值度量两者的相关度，通过关键词判断回复的完整性与可解释性，由留言与答复的时间间隔判断回复的时效性，再由回复字数判断回复部门对于留言的重视程度。运用层次分析法确定回复信息的五个性质的权重，运算可得答复信息的最终得分，给予每一个留言答复质量对应的可量化评价。

2、 分析过程与结果

2.1 问题 1 分析过程与结果

本题的挖掘流程框架如下：

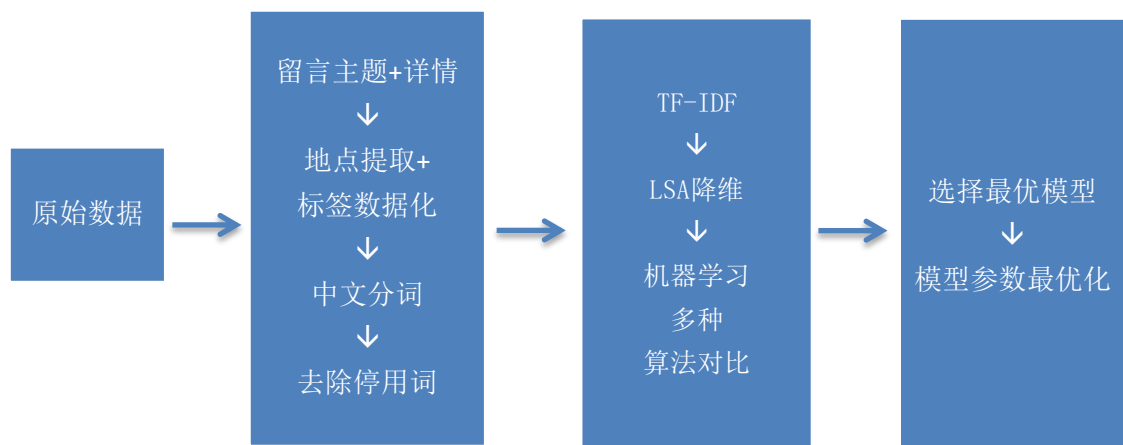


图 1：题 1 挖掘流程图

2.1.1 数据预处理

由于附件 2 的留言主题和留言详情均包含重要信息,所以第一步先将这两列合并创造新列 **message** 进行分析,然后对一级标签数值化。一级标签:城乡建设,党务政务,国土资源,环境保护,纪检监察,交通运输,经济管理,科技与信息产业,民政,农村农业,商贸旅游,卫生计生,政法,教育文体,劳动和社会保障十四大类,分别对应数字“1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14”。

对数据进行去重去空处理,结合哈工大的停用词表以及本文本特点,整理更为合适的停用词表进行停用词过滤,运用 **jieba** 分词^[1]进行精准模式下的分词操作。

2.1.1.1 地点标签提取

提取附件 1 二级分类、三级分类的内容以及留言内容中的 A 市 A 县 A 区等类似的地点名词作特殊词表,在后续中文分词中不进行拆分。

[‘安全生产’, ‘城市建设和市政管理’, ‘城乡规划’, ‘村镇建设’, ‘工程质量’, ‘国有土地上房屋征收’, ‘其他’, ‘住房保障与房地产’, ‘党的建设’, ‘港澳台侨’, ‘国防外交’, ‘民族宗教’, ‘群众团体’, ‘宣理’, ‘土地征收’, ‘土地资源管理’, ‘环保管理’, ‘环境污染’, ‘建设项目审批’, ‘党政处分’]
[‘J市’, ‘G市’, ‘E市’, ‘F市’, ‘D市’, ‘F8市’, ‘B市’, ‘L市’, ‘C4市’, ‘K市’, ‘C5市’, ‘M5市’, ‘M7市’, ‘G3县’, ‘F5县’, ‘A7县’, ‘B7县’, ‘L7县’, ‘I5县’, ‘I3县’, ‘L6县’, ‘F6县’, ‘J9县’, ‘K1县’, ‘K9县’, ‘M9县’, ‘D8县’, ‘J3县’, ‘G8县’, ‘E5县’, ‘M3县’, ‘I4县’, ‘J5县’, ‘F7县’, ‘H3县’, ‘E7县’, ‘K5县’, ‘A4区’, ‘M1区’, ‘E3区’, ‘E区’, ‘F区’, ‘A1区’, ‘B4区’, ‘D区’, ‘B区’, ‘K2区’, ‘区’, ‘E5区’, ‘D1区’, ‘C2区’, ‘R区’, ‘K1区’, ‘A3区’, ‘C3区’, ‘M区’, ‘A2区’, ‘A5区’]

图 2

2.1.1.2 中文分词

采用 **python** 中文分词模块 **jieba** 包对附件三的留言主题留言详情进行分词。**jieba** 算法^[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),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,使用了 **Viterbi** 算法

```

0 [A, 市, 西湖, 建筑, 集团, 占, 道, 施工, 有, 安全隐患, \n, \t, ...
1 [A, 市, 在水一方, 大厦, 人为, 烂尾, 多年, , , 安全隐患, 严重, \n, ...
2 [投诉, A, 市, A1, 区苑, 物业, 违规, 收, 停车费, \n, \t, \t,...
3 [A1, 区, 蔡锷, 南路, A2, 区华庭, 楼顶, 水箱, 长年, 不洗, \n, \...
4 [A1, 区, A2, 区华庭, 自来水, 好大, 一股, 霉味, \n, \t, \t, ...
5 [投诉, A, 市, 盛世, 耀凯, 小区, 物业, 无故, 停水, \n, \t, \t,...
6 [咨询, A, 市, 楼盘, 集中, 供暖, 一事, \n, \t, \t, \t, \t,...
7 [A3, 区, 桐梓, 坡, 西路, 可可, 小城, 长期, 停水, 得不到, 解决, \n...
8 [反映, C4, 市, 收取, 城市, 垃圾处理, 费, 不, 平等, 的, 问题, \n,...
9 [A3, 区, 魏家坡, 小区, 脏乱差, \n, \t, \t, \t, \t, \t, ...
10 [A, 市, 魏家坡, 小区, 脏乱差, \n, \t, \t, \t, \t, \t, \...
11 [A2, 区, 泰华, 一村, 小区, 第四届, 非法, 业委会, 涉嫌, 侵占, 小区业主...
12 [A3, 区梅, 溪湖, 壹号, 御湾, 业主, 用水, 难, \n, \t, \t, \t...
13 [A4, 区鸿涛, 翡翠, 湾, 强行, 对, 入住, 的, 业主, 关水, 限电, \n,...
14 [地铁, 5, 号线, 施工, 导致, A, 市锦楚, 国际, 星城, 小区, 三期, 一个...
15 [A6, 区润, 和, 紫, 郡, 用电, 的, 问题, 能, 不能, 解决, \n, \t...
16 [A, 市锦楚, 国际, 新城, 从, 6, 月份, 开始, 停电, 好, 多次, 了, \...
17 [给, A9, 市, 城区, 南, 西, 片区, 城铁, 站, 设立, 的, 建议, \n,...
18 [请, A6, 区政府, 加大, 对滨水, 新城, 的, 绿化, 建设, \n, \t, \...
19 [A5, 区楚府, 线, 几个, 小区, 经常, 停电, \n, \t, \t, \t, \...

```

图 3: jieba 分词部分结果

中文分词后，留言信息存在大量的标点和无意义的词（如：， 的 他等），这些在文档中大量存在但无意义的词称为停用词，停用词会对后续的分析处理产生较大的影响，所以接下来进行停用词的去除。

2.1.1.3 去除停用词

运用网上下载的哈工大的停用词表并结合实例数据中的一些高频但无意义的词、字符，整理形成新的停用词表，分词后根据该停用词表去除停用词。

由于热度话题的存在，同一地点名词往往在同一类型的留言中集中出现，地点名词对于题一中分类模型的训练会造成一定程度的偏斜，不利于模型的泛化性，故将其加入到停用词表中分词过程中进行过滤

```

0 西湖 建筑 集团 占 道 施工 大道 西行 便 道 未管 路口 加油站 路段 人行道
1 在水一方 大厦 人为 烂尾 多年 严重 位于 书院 路 主干道 在水一方 大厦 一楼
2 投诉 苑 物业 违规 收 停车费 尊敬 领导 苑 小区 位于 火炬 路 小区 物业 程
3 蔡锷 南路 华庭 楼顶 水箱 长年 不洗 华庭 小区 高层 二次 供水 楼顶 水箱 长
4 华庭 自来水 好大 一股 霉味 华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗
5 投诉 盛世 耀凯 小区 物业 无故 停水 2015 年 购买 盛世 耀凯 小区 17 栋 楼
6 咨询 楼盘 集中 供暖 一事 西地省 地区 常年 阴冷 潮湿 气候 近年 气候 逐渐
7 桐梓 坡 西路 可可 小城 长期 停水 得不到 解决 尊敬 胡书记 您好 家住 桐梓
8 反映 收取 城市 垃圾处理 费 平等 问题 梅家田 社区 辖区 小区 居民 每年 依
9 魏家坡 小区 脏乱差 尊敬 政府 领导 您们好 魏家坡巷 业主 多年 小区 脏 乱

```

图 4: 去除停用词部分结果

2.1.2 TF-IDF 算法

计算机无法直接处理文本信息，因此我们需要将文本表示为计算可处理的模式，本题选择了目前文本挖掘技术中成熟常用的向量空间模型。利用 TF-IDF^[2]算法输出文本向量。TF-IDF 是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。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。TF 指的是某词在文章中出现的总次数，IDF=则为 \log_e （语料库中文档总数/包含该词的文档数+1）。计算公式如下

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{j,k}}$$
$$idf_i = \lg \frac{|D|}{|\{j: t_i \in d_j\}|}$$

TF-IDF=TF*IDF，TF-IDF 值越大，表示该词对文本越重要。选择 TF-IDF 算法是基于政务文本的地点名词、专有名词多的特点，运用 TF-IDF 算法可以避免专有名词的权重被降低，影响后续的分类判断。

2.1.3 LSA 降维

由 TF-IDF 算法得（9093，,77708）的高稀疏巨大矩阵，该矩阵维数过多，考虑到文本信息中存在语意相近的情况，且下一步骤为多种模型的测试，计算量较大，故我们放弃原有文本直接向量化。我们重新运用 TF-IDF 算法，规定参数 (max_df=0.5（表示“忽略出现在 50% 以上文档中的词”）,min_df=5（表示“忽略出现少于 5 个文档中的词”）)进行特征抽取。

关于特征提取，我们有从分好类的语料中分别提取再整合形成总的特征或者从统一的语料库中统一提取特征两种方式。Nicholas Evangelopoulos 的研究发现，从统一语料库中提取的主题可以更有助于分类模型的训练。在他们的研究中，当从统一语料库中提取主题时，使用少于一半的主题就可以达到相同的分类精度。因此，我们选用从统一的语料库中提取主题的方式实现特征提取^[3]。

此时得到的新矩阵大小为（9093，17381）。然而，这个矩阵对于八个机器学

习模型的训练与测试来说的计算量还是太大了，所以我们决定通过 LSA 实现降维。

LSA^[4] (latent semantic analysis)潜在语义分析，也被称为 LSI(latent semantic index)，是 Scott Deerwester, Susan T. Dumais 等人在 1990 年提出来的一种新的索引和检索方法。类似于传统向量空间模型，通过向量来表示词和文档间的关系；而不同的是，LSA 基于高维的文档降到低维空间的思想，将词和文档映射到潜在语义空间，去除了原始向量空间中的部分无关信息，提高了信息检索的精确度。映射得到的空间被称为潜在语义空间。这个映射必须是严格线性的而且是基于共现表奇异值分解 (SVD)。通过在 SVD 分解，可以去掉一些 “信息噪声”，只保留最重要的维度信息。

SVD 原理如下：矩阵 A 是原矩阵， A 维度较大，且稀疏，直接运算计算量过大，我们希望能降低矩阵 A 的维度。SVD 将矩阵 A 分解为三个独立矩阵的乘积： $A=U*S*V$ ， S 是矩阵 M 奇异值的对角矩阵。 U 、 V 中的 t 是一个超参数，我们将选择奇异值中最大的 t 个数，且只保留矩阵 U 和 V 的前 t 列，实现将原始的向量转化成一个低维隐含语义空间中。超参数 t 我们可以根据想要查找的主题数量进行选择和调整。简单来说，截断 SVD 可以看作只保留我们变换空间中最重要的 t 维。

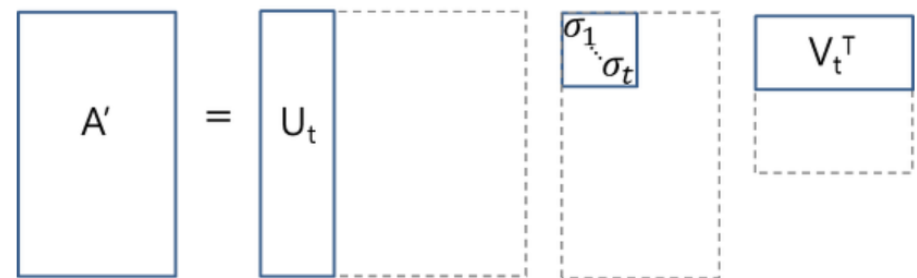


图 5

对于分类模型，我们期望机器学习结果具有更高的泛化能力，因此，我们选用了 sklearn 中的 TruncatedSVD (截断奇异值分解)。TruncatedSVD 与一般 SVD 不同的是，它可以生成指定维度的矩阵，即超参数 t 。

“message” 特征矩阵 X 规模为 (9093, 17381)，由于计算能力的限制，我们无法通过循环去寻找大矩阵的最优维度，而超参数 t 的选择，直接影响了降维效果因此，我们根据 “message” 每行文本的长度画图观察

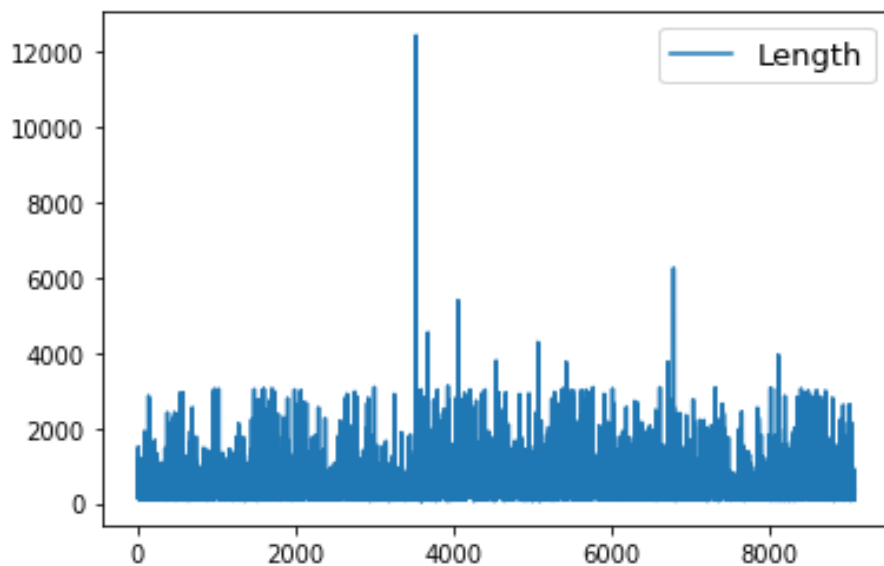


图 6

文本长度多接近 2000，因此，我们对 t 取值 2000，通过 TruncatedSVD，化为规模为(9093,2000)的新特征矩阵 X 。通过 SVD 解释方差来检验降维效果，当 $t=2000$ 时，解释方差为 73%，我们可以认为超参数取值 2000 时，降维后文本特征仍能较好地呈现文本信息。

LSA 得到的结果并未标准化，若没有标准化，绝对值大的特征将影响应有的分类效果，因此，我们将通过 `fit_transform` 函数对特征矩阵 X 进行归一化处理。

2.1.4 机器学习

查找资料确定主流的多分类文本常用的机器学习模型，分别是 GaussianNB，LinearSVC，LogisticRegression，SGDClassifier，SVC，KNeighborsClassifier，AdaBoostClassifier，DecisionTreeClassifier，运用处理好的数据进行模型训练及测试，得到各个模型分类结果的准确度，F1-score 以及训练和测试时间，形成模型对比。结果如下图。

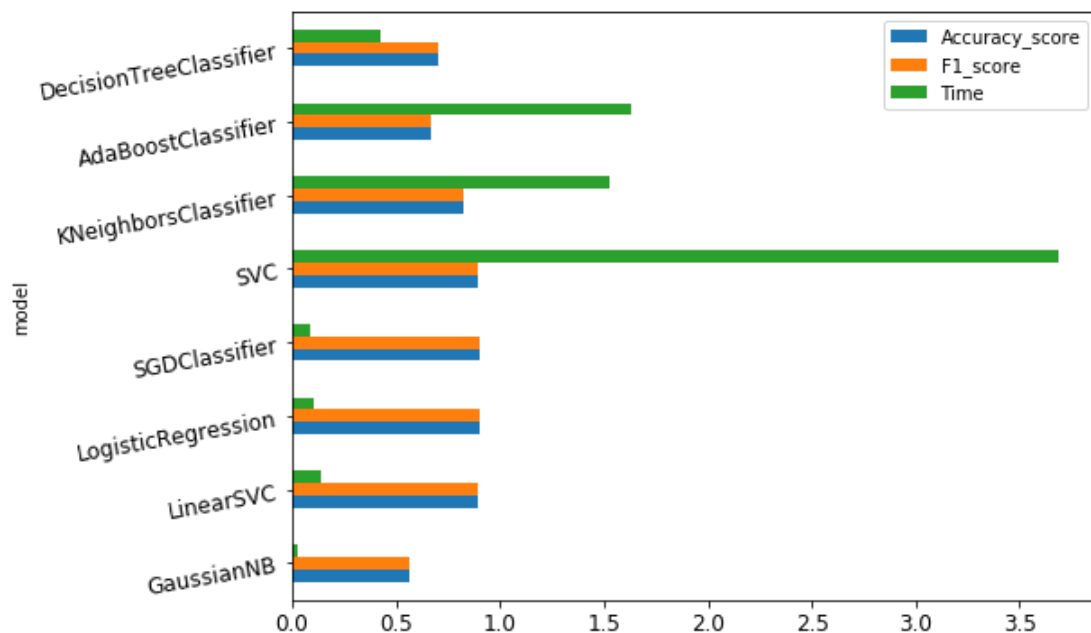


图 7 八种模型效果对比图

综合评价，SGDClassifier, LogisticRegression, LinearSVC 三个模型表现更佳。

2.1.5 最优模型的选择

结合政务文本数据特点考虑 SGDClassifier, LogistigRegression, LinearSVC 三个模型。

大数据情形下使用 LinearSVC 模型可能会有点慢，而 SGDClassifier 使用了随机梯度下降做训练，所以每次训练并没有使用全部的样本，因此收敛速度会快很多，所需时间也就少很多。但 SGDClassifier 对于特征的幅度非常敏感，我们难以掌握特征幅度，不便于训练前对特征做幅度调整，其次，本数据类别间具有明显的不平衡性，对此我们可以通过交叉验证降低这种不平衡性所带来的误差。而 SGDClassifier 每次只使用一部分(mini-batch)做训练，在这种情况下，我们使用交叉验证(cross-validation)并不是很合适。

属于 SVM 其中之一的 LinearSVC，分类中寻求的是一个超平面，通过超平面实现对多维数据的分类，因此并不是每个样本都对其平面有影响。同类样本间差距越小，不同类样本间差距越大，分类效果越好。

LogisticRegression 是每个样本点对超平面都有影响,因此，更适合通过因素变动预测结果变动方向及其概率，而非多分类问题。

综上所述，我们选取了 LinearSVC 模型，进行了更为详细的分析。

由于该政务数据的收集具有地域性以及热点问题多人多次留言的集中性，某类一级标签所对应的留言往往有某一个或某几个地点数据集中多次出现的现象，这对于模型的训练极为不利。因此，我们通过正则表达式提取该类文本信息，并形成列表加入停用词表，在分词的过程中进行删除。

其次本数据集是不均衡数据，这里在划分训练集与测试集时，采用与原有数据集一级标签同比例的方式抽样，即分层采样，以减少不均衡的数据集对模型训练的影响。

2.1.6 模型参数最优化

在机器学习中，超参数选择不恰当，就会出现欠拟合或者过拟合的问题。调参，一般是凭经验微调，或者选择不同大小的参数，带入模型中，挑选表现最好的参数。我们将采取单一变量调参的模式，寻找损失函数、损失函数的惩罚项、惩罚项系数 C 的最优值。

首先是考察损失函数对预测的影响，在 LinearSVC 中，损失函数有 hinge, squared_hinge 两种类型，分别运用这两个损失函数，训练模型再预测分类，计算分类的准确度，准确度均为 0.9，则使用 LinearSVC 中默认的损失函数。

其次，对于损失函数的惩罚项，使用 L1 正则化的模型即 Lasso 回归与使用 L2 正则化的模型即 Ridge 回归（岭回归）进行模型训练再预测分类，计算分类的准确度，L1 的得分为 0.89，L2 的得分为 0.90，且 L2 正则化可以防止模型过拟合，因此我们选择 L2 作为损失函数的惩罚项，我们创建一个从 0.01 到 100 的个数为 50 的等比数列，以此作为 C 的备选，逐一代入模型训练，分别计算测试集与预测集的分类预测的准确度。以 C 的取值为横轴，准确度得分为纵轴画图。由图可知，当 C 较小时，误分类点重要性较低，此时误分类点较多，分类器性能差，当 C 接近一时，测试集的得分接近最高点，而后测试集的得分不再增加，训练集的得分则不断增加，此时已是过拟合状态，因此， C 取值 1。

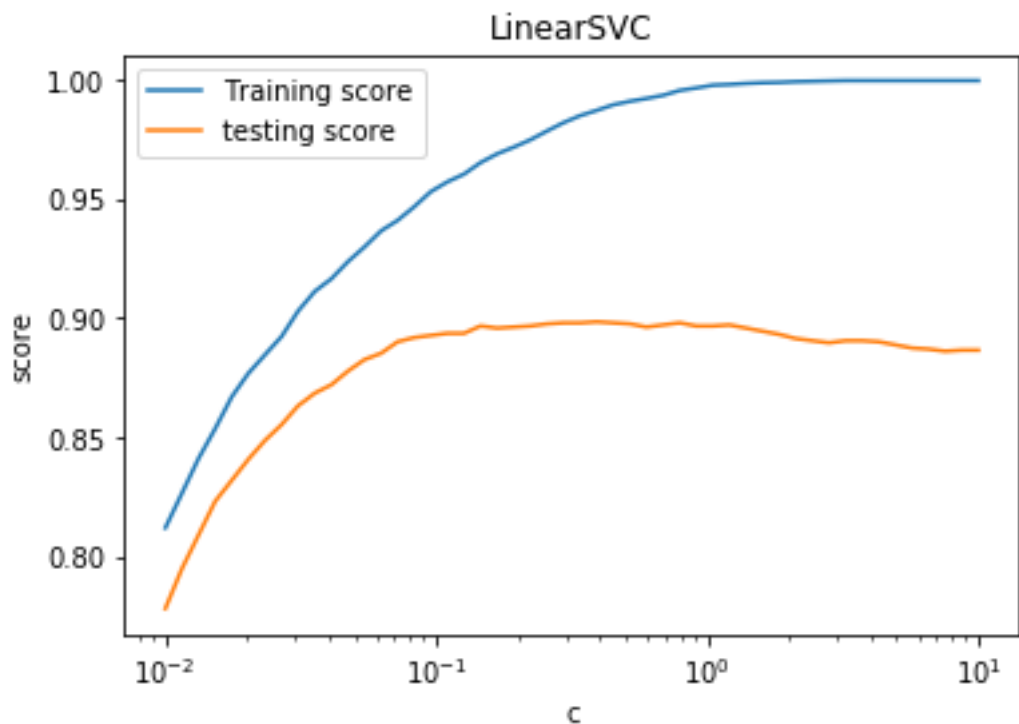


图 8 不同 C 值的 LinearSVC 模型训练集与测试集得分曲线

2.1.7 分类结果与模型评价

由上述测试，我们为 LinearSVC 选择了损失函数、损失函数的惩罚项、惩罚项系数 C 的最优值，将利用 TF-IDF 算法得经过去重去空去停用词的所有文本的向量值，进行模型训练，为降低数据的不平衡，我们将采用交叉验证的方法，取 CV=5 进行 LinearSVC 模型训练，得到最终的 F1 分数.为 0.88.

F1-Score 公式：

若预测结果和真实结果有如下关系：

	真实 1	真实 0
预测 1	True Positive(TP)真阳性	False Positive(FP)假阳性
预测 0	False Negative(FN)假阴性	True Negative(TN)真阴性

则有：

$$\text{查准率 (precision)} : p = \frac{TP}{TP + FP}$$

$$\text{召回率 (recall)} : r = \frac{TP}{TP + FN}$$

F1 分数（F1-Score）：
$$F1 = 2 * \frac{p * r}{p + r}$$

将分类的预测结果与实际分类用热力图进行直观展示如下。

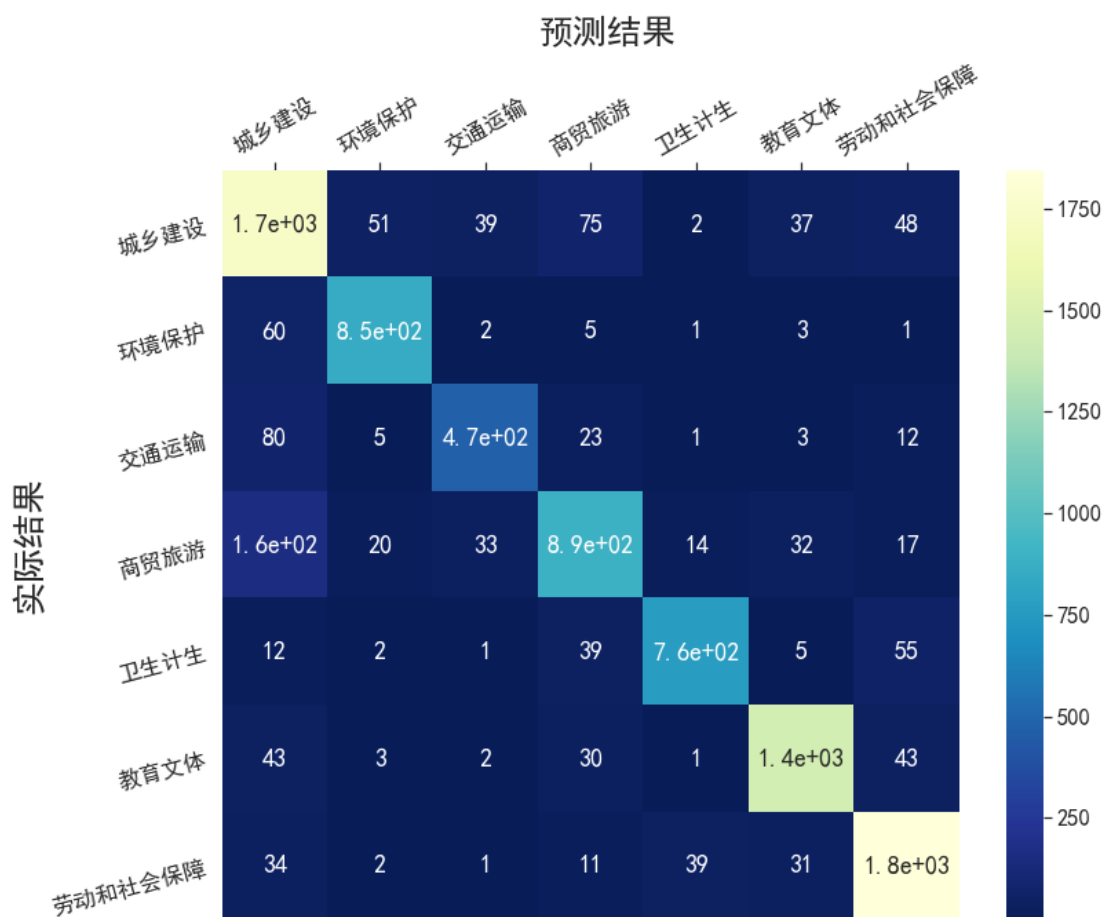


图 9：LinearsVC 模型的预测结果实际分类热力图

2.2 问题 2 分析过程与结果

2.2.1 挖掘流程示意图与名词解释

2.2.1.1 挖掘流程框架：

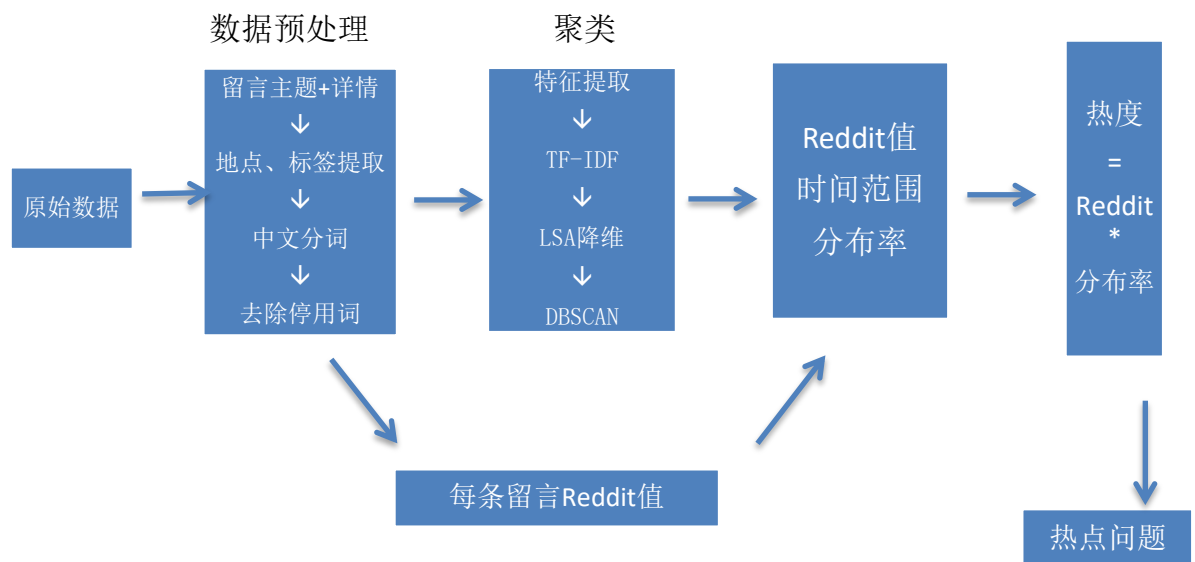


图 10：题 2 挖掘流程图

2.2.1.2 名词解释

定义热点问题属性^[6]：留言频率（数量）、留言分布率、时间、点赞反对人数

定义热度评价指标的：话题热度值=该话题的 Reddit*该话题的分布率

✧ Reddit: Reddit 是美国最大的网络社区，它根据一系列算法（称为 Reddit 算法）计算“热点文章排行榜”。本体先利用其数学公式计算每条留言的 Reddit 值，再对出现的极端情况进行公式改进，使其更符合本体留言信息的特性。修改后的 Reddit 公式更合理的体现留言频率、时间跨度以及点赞人群与反对人群对热度的贡献力度。

✧ 分布率: 在话题时间范围内，该话题的留言数量占该段时间留言总数的比重。

✧ 热点问题需满足的俩个条件：

1) 热度值>热度值的标准差 109.71

2) 分布率>分布率的平均值 0.009

2.2.2 数据预处理

与题一进行相似的预处理与分词过程。不同的是，分词后去除停用词时，由于地点名词有利于某一时间段内某一被集中反映的热点话题的挖掘，故保留地点名词，分词结果如图 10。

```
0    A3区 一米阳光 婚纱 艺术摄影 是否 合法 纳税 座落在 A市 A3区 联丰路 米
1    咨询 A6区 道路 命名 规划 初步 成果 公示 城乡 门牌 问题 A市 A6区 道路
2    反映 A7县 春华 镇金鼎村 水泥路 自来水 到户 问题 系 春华 镇金鼎村 七
3    A2区 黄兴路 步行街 古道 巷 住户 卫生间 粪便 外排 靠近 黄兴路 步行街
4    A市 A3区 中海 国际 社区 三期 四期 中间 空地 夜间 施工 噪音 扰民 A市
5    A3区 麓 泉 社区 单方面 改变 麓 谷 明珠 小区 栋 架空层 使用 性质 麓
6    A2区 富绿 新村 房产 性质 二高 一部 发出 非法 集资 打击 通知 中是
7    A市 地铁 违规 用工 问题 质疑 一名 A市 地铁站 上班 安检员 中介 公司
8    A市 路 公交车 随意 变道 通行 12 月 21 日 下午 17 时 52 分许 路 公交
9    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
10   A7县 特立 路 东四 路口 晚 高峰 太堵 建议 调整 信号灯 配时 近来 下午
```

图 11：去除停用词效果

2.2.3 DBSCAN 聚类

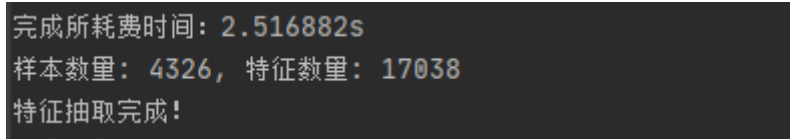
聚类算法是无监督学习，也就是无标签（label）数据输入，其原理是把相似的数据划分在一起。聚类算法中比较常用的有 K-means 和 DBSCAN 两种算法^[8]，K-means 算法需预先选择聚类的簇个数，由于分类前，话题个数无法确认，矩阵的维度较大，无法通过循环进行大量计算确定最优的簇个数，因此，K-means 算法不适用于此数据。考虑到留言中属于热度话题的留言是有限的，更多的是一个话题只有 1-2 个相关的留言，采用 DBSCAN 算法进行聚类，可以通过调整 min_samples 参数，即簇内最少个数将此类留言识别为噪声点，以达到更优的聚类效果。

聚类算法的普遍难点在于：如何选择合适的参数（调参），怎样评估聚类效果是好的（评估）。我们将通过 k-距离曲线来确认 DBSCAN 的半径参数，并用轮廓系数评价聚类的好坏。

2.2.3.1 文本特征抽取

预处理后的数据提取 TF-IDF 特征值、2-gram 特征类。2-Gram 将文本里面的内容按照字节进行大小为 2 的滑动窗口操作,形成了长度是 2 的字节片段序列,比如 “A 市 经济 体育 学院 强制 实习”,2-gram 切分就是 “A 市经济 经济体育 体育学院 学院强制 强制实习”。2-Gram 抽取到的特征,更能代表文本的特性,可以对后续的文本聚类产生良好的推动作用。

通过参数的设定: $\text{max_df} = 0.50$ (表示“忽略出现在 50% 以上文档中的词”)、 $\text{min_df} = 5$ (表示“忽略出现少于 5 个文档中的词”)完成了一次特征提取,此时的矩阵大小为 (4271, 16200)。



```
完成所耗费时间: 2.516882s
样本数里: 4326, 特征数里: 17038
特征抽取完成!
```

图 12

然而, DBSCAN 不能很好反映高维数据,所以对抽取的特征进行降维是很有必要的。

高维数据的降维,常用的方法有 LSA 与 PCA。LSA 将词和文档映射到潜在语义空间,从而去除了原始向量空间中的一些“噪音”,提高了信息检索的精确度。它是依赖于 SVD 分解,即奇异值分解的算法。PCA 的实现一般有两种,一种是用特征值分解去实现的,一种是用 SVD 去实现的。由于特征值分解要求输入的矩阵必须为方阵,所以一般 PCA 多用 SVD 实现。虽然 LSA 与 PCA 具有相似性,都是依靠于 SVD 算法实现,这里,我们将选用 LSA (潜在语义分析) 实现降维,原因是: 由于 PCA 必须计算协方差矩阵,需要在整个矩阵上操作,特征矩阵 X 规模为 (4326, 17073), 矩阵过大,计算资源可能会不够用。SVD 则对原始数据直接进行奇异值分解,计算量大大降低了。

更进一步考虑,我们选用了 sklearn 中的 TruncatedSVD (截断奇异值分解)。TruncatedSVD 与一般 SVD 不同的是,它可以生成指定维度的矩阵,它是一种正则化方法,它牺牲部分精度换取稳定性,使得结果具有更高的泛化能力。根据 “message” 每行文本的长度画图观察:

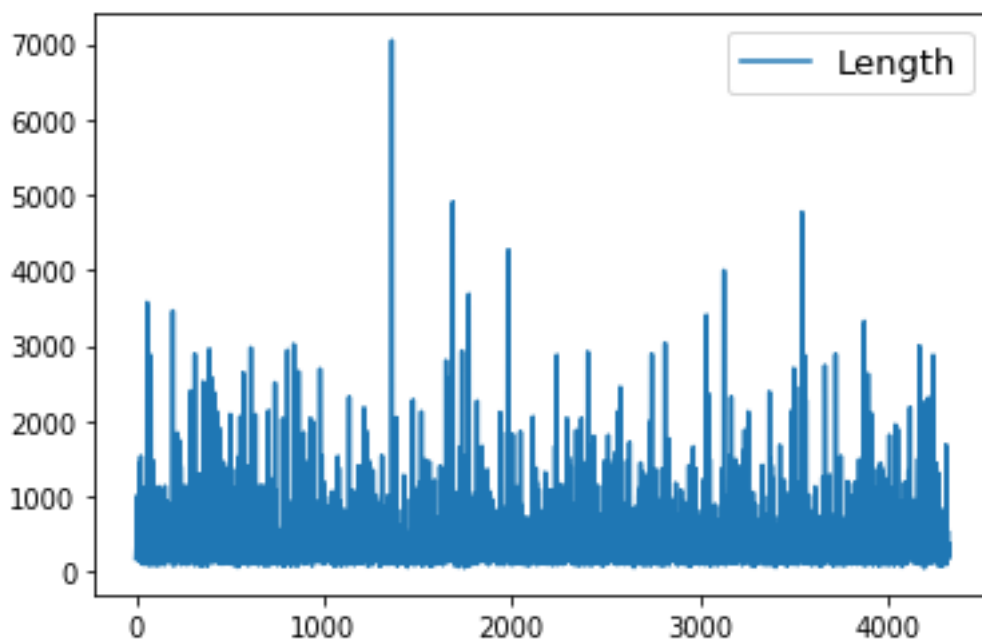


图 13

由图：文本长度多低于 1500，以期维度为 1500 时能尽可能的保留文本中必要的信息，过滤长文本中冗杂的信息，使得降维后文本特征仍能较好地呈现文本信息。通过 TruncatedSVD，把原先规模为 (4326, 17073) 的特征矩阵 X 化为规模为 (4326, 1500) 的新特征矩阵 X 。我们通过 SVD 解释方差来检验降维效果，对属性 `explained_variance_ratio` 进行求和，即查看降维后所有新特征中每个新特征向量所占的信息量占原始数据总信息量的百分比之和。

```
完成所耗费时间：108.499561s  
SVD解释方差的step：77%  
PCA文本特征抽取完成！
```

图 14

2.2.3.2 DBSCAN 聚类

DBSCAN^{[9][10]}是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，不需要事先划分簇的个数，可在存在噪声的空间数据库中发现任意形状的聚类。

为确定 DBSCAN 的 `eps` 半径参数，我们绘制 K-距离曲线来确定该参数。

k-距离[2]是指：给定数据集 $P=\{p(i); i=0,1,\cdots,n\}$ ，对于任意点 $P(i)$ ，计算点 $P(i)$ 到集合 D 的子集 $S=\{p(1), p(2), \cdots, p(i-1), p(i+1), \cdots, p(n)\}$ 中所有点之间的距离，距离按照从小到大的顺序排序，假设排序后的距离集合为 $D=\{d(1), d(2), \cdots, d(k-1), d(k), d(k+1), \cdots, d(n)\}$ ，则 $d(k)$ 就被称为 k -距离。也就是说， k -距离是点 $p(i)$ 到所有点（除了 $p(i)$ 点）之间距离第 k 近的距离。对待聚类集合中每个点 $p(i)$ 都计算 k -距离，最后得到所有点的 k -距离集合 $E=\{e(1), e(2), \cdots, e(n)\}$ 。

对集合 E 进行升序排序后得到 k -距离集合 E' ，需要拟合一条排序后的 E' 集合中 k -距离的变化曲线图，然后绘出曲线，通过观察，将急剧发生变化的位置所对应的 k -距离的值，确定为半径 Eps 的值。

采用欧氏距离计算 k -距离曲线，绘制结果如下：

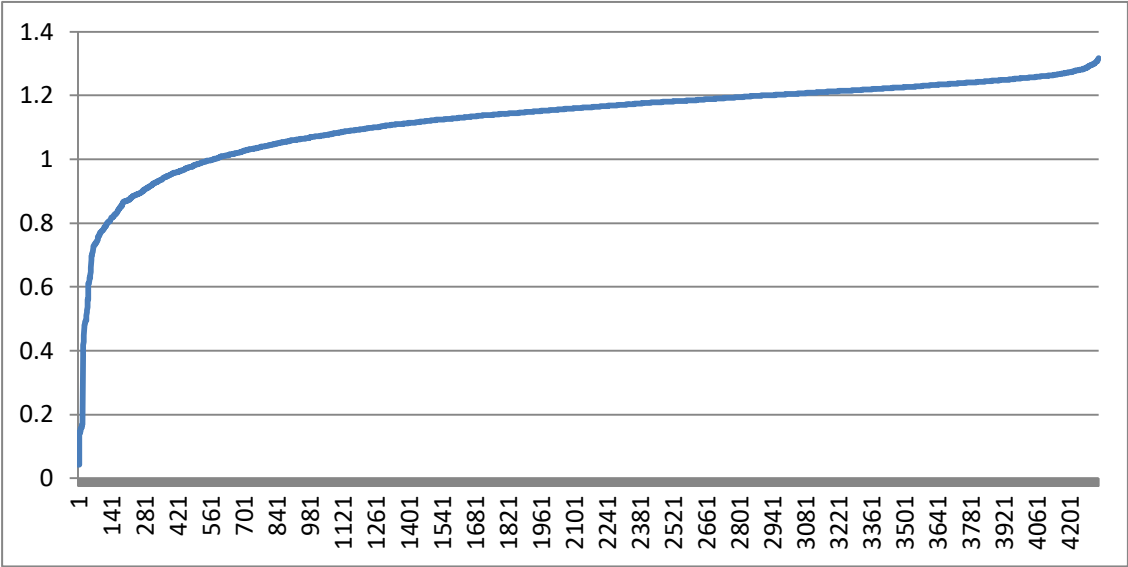


图 15：:K-距离曲线

由图得选取半径为 0.8，因挖掘的是热点问题，我们认为留言数应 ≥ 4 ，所以最终进行 DBSCAN（ $eps=0.8,min_samples=4$ ）聚类，并利用轮廓系数评价聚类效果。得结果如下：

表格 1	
参与聚类的留言条数	257
噪点数	3967
聚类数	65
轮廓系数	-0.015

轮廓系数<0，效果欠佳，为此我们重新选取半径为 1，进行聚类得：

表格 2	
参与聚类的留言条数	874
噪点数	3079
聚类数	148
轮廓系数	0.019

2.2.4 计算 Reddit

Reddit^[11] “热门排名算法”的数学表达式如下：其中 45000 是 12.5 个小时，是 Reddit 算法衡量发布时间对热度的影响参数。

Given the time the entry was posted A and the time of 7:46:43 a.m. December 8, 2005 B , we have t_s as their difference in seconds

$$t_s = A - B$$

and x as the difference between the number of up votes U and the number of down votes D

$$x = U - D$$

where $y \in \{-1, 0, 1\}$

$$y = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and z as the maximal value, of the absolute value of x and 1

$$z = \begin{cases} |x| & \text{if } |x| \geq 1 \\ 1 & \text{if } |x| < 1 \end{cases}$$

we have the rating as a function $f(t_s, y, z)$

$$f(t_s, y, z) = \log_{10} z + \frac{y t_s}{45000}$$

图 16: Reddit 的数学表达式

利用该数学表达式计算得 Reddit 散点图如下

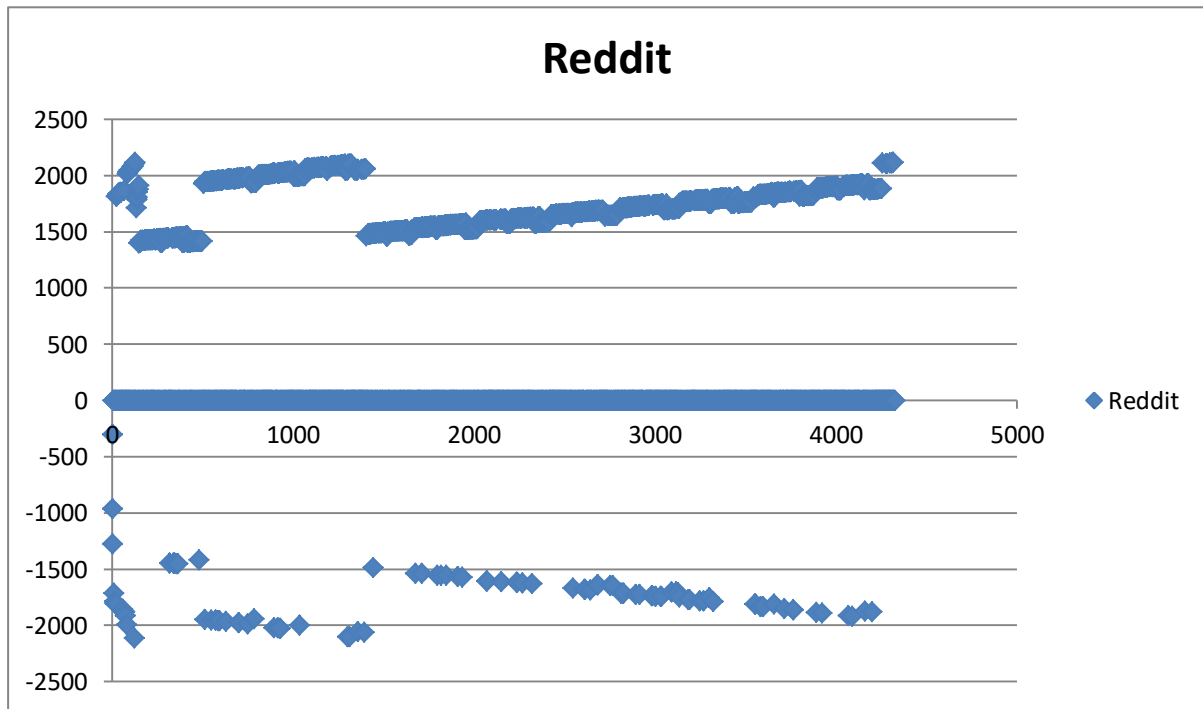


图 17: Reddit 散点图

由图可知 Reddit 值分布在 $(-2300, -300) \cup 0 \cup (1400, 2200)$ 之中，大部分取 0，非 0 部分又呈两极分化分布。这是因为：

- ✧ 本题留言数据中点赞数、反对数基本为 0，这会使得 $\frac{y t_s}{45000}$ 部分为 0， x 也为 0，进而 \log_{10}^z 也为 0，所以大部分 Reddit 为 0。
- ✧ 当点赞数与反对数的差值 x 即使不为 0，仍较小，这会使得公式中的 \log_{10}^z 非常小。

因此对 Reddit 进行改良，使其更符合本题中留言数据的特性。改良如下：

- ✦ 留言热度与留言自身的正负向无关，仅与大众关注度有关。故本体中 y 均取 1，即忽略掉改参数。
- ✦ 又考虑时间间隔秒数部分 $\frac{y t_s}{45000}$ 却非常大，导致 \log_{10}^z 对最后的值基本没影响。

故把公式改进为：
$$f(t_s, z) = \Theta(\log_{10}^z) * \frac{t_s}{45000}$$

其中 $\Theta(\log_{10}^z)$ 是对 \log_{10}^z 进行 $(0, 1)$ 区间标准化、弱化不同点赞数、反对数之间的差值距离；改良后得新的散点图：

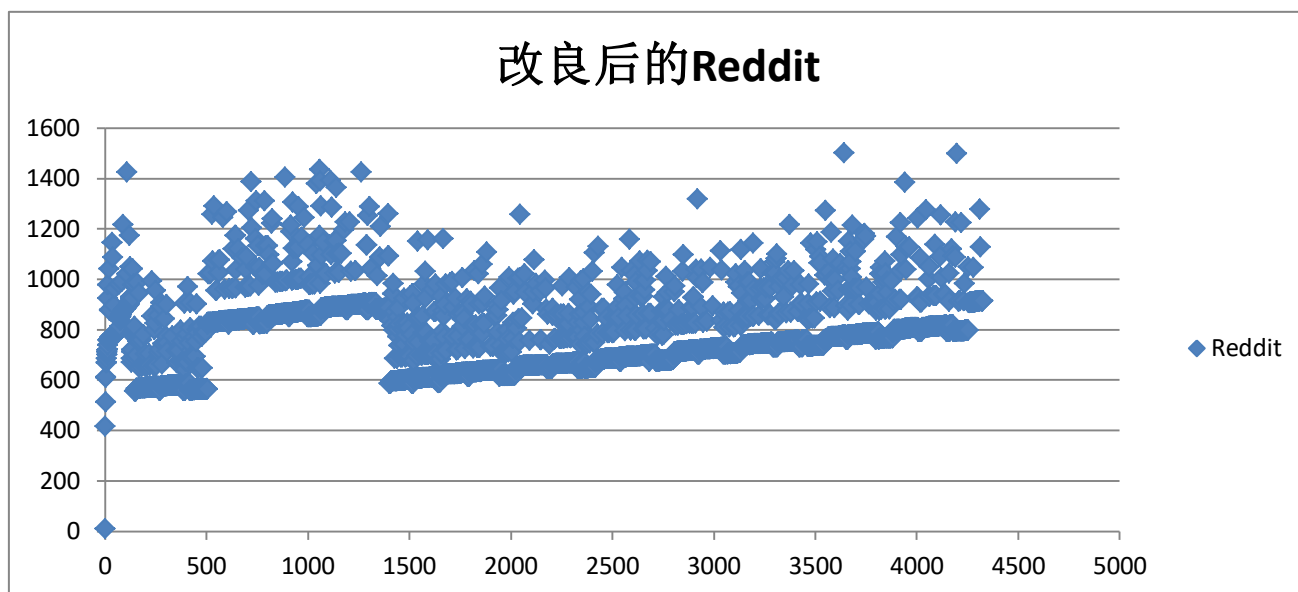


图 18: 改良后 Reddit 散点图

改良后单条留言的 Reddit 值分布较均匀，消除了大部分取 0，其余部分俩极分化的极端情况，能更好的比较每条留言之间的热度高低。

根据聚类结果计算每一类的 Reddit 值：

表格 3: 簇类以其 Reddit(注: 红色-1 类为噪点)			
簇类	reddit	簇类	reddit
-1	2353902	6	3391.772
0	17153.11	7	51945.15
1	2421.677	8	16720.9
2	3651.775	9	4718.073
3	57802.76	10	56900.94
4	3932.57	11	6017.784
5	11237.06	12	11608.44

2.2.5 计算分布率

分布率是指在话题时间范围内，该话题的留言数量占该段时间留言总数的比重。为此欲求出每类话题的分布率须求出每类话题的留言数和对应时间范围的留言总数。这里我们用 excel 函数来进行相关计算。

该话题的留言数就是该簇包含的个数，这在聚类完成后就可以得到。我们重点要求的是该簇的留言时间范围以及该时间范围的留言总数。

留言范围是指聚类完成好可以得到每一类包含的留言以及其时间，取每一类第一条留言的时间以及最后一条留言时间（这里是指时间上的第一条和最后一条）即可得到该类的留言时间。同时也可以得到每一类包含的留言数。

表格 4：部分簇的时间范围(注：红色-1 类为噪点)

簇类	个数	第一条留言时间	最后一条留言时间	时间范围
-1	3075	2017-06-08 17:31:20	2020-01-08 14:44:38	2017/06/08-2020/01/08
0	23	2018-11-15 16:07:12	2019-12-02 11:57:49	2018/11/15-2019/12/02
1	4	2018-05-17 08:32:04	2019-04-28 17:32:51	2018/05/17-2019/04/28
2	5	2019-03-19 19:37:45	2019-05-30 17:34:02	2019/03/19-2019/05/30
3	72	2019-03-11 14:52:12	2020-01-04 03:01:43	2019/03/11-2020/01/04
4	4	2019-07-21 10:29:36	2019-08-01 16:20:02	2019/07/21-2019/08/01
5	13	2019-07-28 12:49:18	2019-09-25 00:31:33	2019/07/28-2019/09/25
6	4	2019-10-31 21:17:22	2019-11-02 10:19:56	2019/10/31-2019/11/02
7	57	2019-11-02 10:18:00	2020-01-26 19:47:11	2019/11/02-2020/01/26
8	22	2019-01-10 00:20:55	2020-01-03 14:38:34	2019/01/10-2020/01/03
9	6	2019-01-11 10:10:46	2019-12-17 10:10:08	2019/01/11-2019/12/17
10	77	2019-01-02 17:25:52	2020-01-06 13:54:54	2019/01/02-2020/01/06
11	6	2019-01-02 20:27:07	2019-11-08 15:48:07	2019/01/02-2019/11/08
12	15	2019-01-08 21:34:48	2019-07-08 17:16:57	2019/01/08-2019/07/08
13	23	2019-01-02 11:03:59	2019-11-20 11:06:47	2019/01/02-2019/11/20

将留言时间升序排列，对每一类取第一条时间对应的行号以及最后一条时间对应的行号，俩者的差即是这段时间的留言总条数。

表格 5：部分簇留言时间范围内的留言总数

簇类	第一条留言时间的行号	最后一条留言时间的行号	该时间范围的留言总数
-1	1	4322	4321
0	4	3886	3882
1	2	1352	1350
2	831	1751	920
3	727	4277	3550
4	2399	2540	141
5	2489	3223	734
6	3596	3610	14
7	3609	4326	717
8	121	4272	4151
9	132	4075	3943
10	16	4295	4279
11	18	3673	3655
12	100	2221	2121
13	13	3788	3775

将该类留言数除以该时间范围的留言总数得到每一类的分布率

表格 6：分布率

簇类	个数	Reddit	留言时间范围	该时间范围内留言总数	分布率
-1	3075	2353901.661	2017/06/08-2020/01/08	4321	0.711640824
0	23	17153.10569	2018/11/15-2019/12/02	3882	0.005924781
1	4	2421.676616	2018/05/17-2019/04/28	1350	0.002962963
2	5	3651.775373	2019/03/19-2019/05/30	920	0.005434783
3	72	57802.76114	2019/03/11-2020/01/04	3550	0.02028169
4	4	3932.570161	2019/07/21-2019/08/01	141	0.028368794
5	13	11237.05599	2019/07/28-2019/09/25	734	0.017711172
6	4	3391.771511	2019/10/31-2019/11/02	14	0.285714286
7	57	51945.15465	2019/11/02-2020/01/26	717	0.079497908
8	22	16720.89745	2019/01/10-2020/01/03	4151	0.005299928
9	6	4718.073156	2019/01/11-2019/12/17	3943	0.001521684
10	77	56900.9378	2019/01/02-2020/01/06	4279	0.017994859
11	6	6017.783702	2019/01/02-2019/11/08	3655	0.001641587
12	15	11608.44466	2019/01/08-2019/07/08	2121	0.007072136
13	23	15523.23511	2019/01/02-2019/11/20	3775	0.006092715

热度评价指标为话题热度，可用话题分布率乘以话题 Reddit 得到

2.2.6 话题热度值

表格 7：部分话题的热度值结果

簇类	个数	Reddit	留言时间范围	该时间范围的留言总数	分布率	热度
-1	3075	2353901.661	2017/06/08 -2020/01/08	4321	0.711640824	1675132.517
7	57	51945.15465	2019/11/02 -2020/01/26	717	0.079497908	4129.531123
3	72	57802.76114	2019/03/11 -2020/01/04	3550	0.02028169	1172.337691
10	77	56900.9378	2019/01/02 -2020/01/06	4279	0.017994859	1023.924331
6	4	3391.771511	2019/10/31 -2019/11/02	14	0.285714286	969.0775746
26	57	44159.1176	2019/01/07 -2019/12/20	4055	0.01405672	620.7323559
78	10	8983.291767	2019/12/15 -2020/01/03	213	0.046948357	421.7507872
98	9	8400.516096	2019/03/26 -2019/04/15	228	0.039473684	331.5993196
103	4	2618.511389	2019/04/11 -2019/04/16	44	0.090909091	238.0464899
29	33	26365.08082	2019/01/08 -2020/01/03	4177	0.007900407	208.2948688
5	13	11237.05599	2019/07/28 -2019/09/25	734	0.017711172	199.0214276
128	4	3634.838788	2019/06/25 -2019/07/02	86	0.046511628	169.0622692
118	4	2812.045222	2019/05/31 -2019/06/06	78	0.051282051	144.2074473
53	23	16868.62293	2019/02/17 -2019/11/08	3191	0.007207772	121.5851856
0	23	17153.10569	2018/11/15 -2019/12/02	3882	0.005924781	101.6283954
13	23	15523.23511	2019/01/02 -2019/11/20	3775	0.006092715	94.578651
75	6	5503.022477	2019/12/06 -2020/01/07	363	0.016528926	90.9590492

由上表可得有些话题留言分布率低于 0.009，热度值低于 109.71，利用热点问题的两个条件筛选得特定地点/人群的热点问题（按热度将序排列）：

2.2.7 热点问题

表格 8：热点问题

簇类	个数	Reddit	留言时间范围	该时间范围的留言总数	分布率	热度
7	57	51945.15465	2019/11/02 -2020/01/26	717	0.079497908	4129.531123
3	72	57802.76114	2019/03/11 -2020/01/04	3550	0.02028169	1172.337691
10	77	56900.9378	2019/01/02 -2020/01/06	4279	0.017994859	1023.924331
6	4	3391.771511	2019/10/31 -2019/11/02	14	0.285714286	969.0775746
26	57	44159.1176	2019/01/07 -2019/12/20	4055	0.01405672	620.7323559
78	10	8983.291767	2019/12/15 -2020/01/03	213	0.046948357	421.7507872
98	9	8400.516096	2019/03/26 -2019/04/15	228	0.039473684	331.5993196
141	4	3313.388453	2019/08/23 -2019/08/26	52	0.076923077	254.8760348
103	4	2618.511389	2019/04/11 -2019/04/16	44	0.090909091	238.0464899
5	13	11237.05599	2019/07/28 -2019/09/25	734	0.017711172	199.0214276
128	4	3634.838788	2019/06/25 -2019/07/02	86	0.046511628	169.0622692
118	4	2812.045222	2019/05/31 -2019/06/06	78	0.051282051	144.2074473
4	4	3932.570161	2019/07/21 -2019/08/01	141	0.028368794	111.5622741

2.3 问题 3 分析过程与结果

本题针对附件 4 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、时效性以及重视度对答复意见的质量给出一套评价方案。

2.3.1 答复的相关性

相关性是指答复内容与留言的问题有关联，对此采用余弦相似度计算留言详情和答复意见的相关程度。余弦相似度，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估两者的相似度。具体步骤如下：

- 1) 分别对留言详情和答复意见进行分词；
- 2) 找出留言详情和答复意见两组文字的关键词；
- 3) 每组文字各取出若干个关键词，形成一个集合；
- 4) 计算每组文字集合中的词的词频，生成两组文字各自的词频向量；
- 5) 计算两组向量的余弦相似度，值越大就表示内容相关程度越高。

2.3.2 答复的完整性和可解释性

完整、可解释的答复意见需具有开头结尾和相应的理论支撑，答复意见满足完整性需满足问好、回复以及表达对群众支持的感谢三个要素，而满足可解释性则要引用法律条例对答复进行解释。

对此，判断答复意见是否具有完整性和可解释性，可以使用 python 中的正则表达式 `re.findall` 对答复意见进行关键词提取，查找答复意见中是否存在“您好/你好”、“答复/回复”、“感谢/谢谢”、“《》”的关键字符。

2.3.3 答复的时效性

相关部门收到留言后，需要及时对留言的问题提出解决方案，对此，可以将

答复的时间减去留言的时间，统计各组留言和答复的时间间隔，然后对时间间隔取四分位数，根据四分位数来判定每组答复的时效性的优良中差。

2.3.4 答复的重视度

相关部门对留言的重视程度拟有三个判断标准，分别为留言的字数、针对性的回复和问题的核实。

对于字数的统计，可以使用 python 中的 count 函数分别统计留言中的中文、英文以及数字的个数，三者的总和即为留言的字数，统计各组的答复意见的字数，利用 excel 的 quartile 函数算出留言字数的四分位数，依据四分位数把留言字数分为优良中差四个级别。

对于答复意见中是否存在针对性的回复和问题的核实，使用 python 中的正则表达式 re.findall 查找答复中是否存在“留言已收悉/留言反映/所反映的”和“调查/高度重视/核实/经查”来进行判断。

2.3.5 答复意见评级

赋予每个性质 10 分原始分，根据每个性质的要素个数，将每个性质 10 分原始分平均分配到各个要素。每个性质原始分的分配准则如表 9 所示

表格 9

性质	分配准则
相关性	余弦相关度*10
完整性	三个要素各占 $\frac{10}{3}$ 分
可解释性	有引用法律条例得 10 分，无则得 0 分
时效性	时间间隔的优良中差级别对应得分分别为 10、7.5、5、2.5
重视度	① 答复字数对应级别的得分权重乘以 $\frac{10}{3}$ 即为答复字数的得分 (答复的字数的优良中差得分权重分别为 1、0.75、0.5、0.25)

	② 针对性的回复、问题的核实各占 $\frac{10}{3}$ 分
--	-----------------------------------

根据层次分析法的标度方法（如表 10 所示）写出五个性质的判断矩阵（如表 11 所示），利用判断矩阵计算五个性质的得分权重。五个性质的权重与得分的乘积之和即为答复意见的评分，最后根据评分标准（如表 12 所示）对答复意见进行评级。

层次分析法^[12] 根据问题的性质和要达到的总目标，将问题分解为不同的组成因素，并按照因素间的相互关联影响以及隶属关系将因素按不同层次聚集组合，形成一个多层次的分析结构模型，从而最终使问题归结为最低层(供决策的方案、措施等)相对于最高层(总目标)的相对重要权值的确定或相对优劣次序的排定。

表格 10：标度及其含义

标度	含义
1	同样重要性
3	稍微重要
5	明显重要
7	强烈重要
9	极端重要
2, 4, 6, 8	上述两相邻判断的中值
倒数	A 和 B 相比如果标度为 3 那么 B 和 A 相比就是 1/3

表格 11：五个性质的判断矩阵

	相关性	可解释性	时效性	重视度	完整性
相关性	1	$\frac{1}{2}$	1	$\frac{1}{2}$	3
可解释性	2	1	2	1	6
时效性	1	$\frac{1}{2}$	1	$\frac{1}{2}$	3
重视度	2	1	2	1	6
完整性	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	1

表格 12：评级标准	
评分	等级
[0, 2.5]	不及格
(2.5, 5]	一般
(5, 7.5]	中等
(7.5, 10]	优秀

一致性指标用 CI 计算，CI 越小，说明一致性越大.一致性指标为： $CI=(\lambda -n)/(n-1)$ 。计算得判断矩阵的最大特征值 $\lambda =n=5$ ，所以 CI 为 0，该判断矩阵为一致性矩阵。可直接选取其中一列计算权重，权重的计算结果如下：

$$\text{相关性: } \frac{1}{1+2+1+2+\frac{1}{3}} = \frac{3}{19}$$

$$\text{可解释性: } \frac{2}{1+2+1+2+\frac{1}{3}} = \frac{6}{19}$$

$$\text{时效性: } \frac{1}{1+2+1+2+\frac{1}{3}} = \frac{3}{19}$$

$$\text{重视度: } \frac{2}{1+2+1+2+\frac{1}{3}} = \frac{6}{19}$$

$$\text{完整性: } \frac{\frac{1}{3}}{1+2+1+2+\frac{1}{3}} = \frac{1}{19}$$

2.3.6 评价方案的实现

将附件 4 中 2816 组答复意见的数据分别进行相关性、完整性、可解释性、时效性和重视度的计算判断，结果如下所示

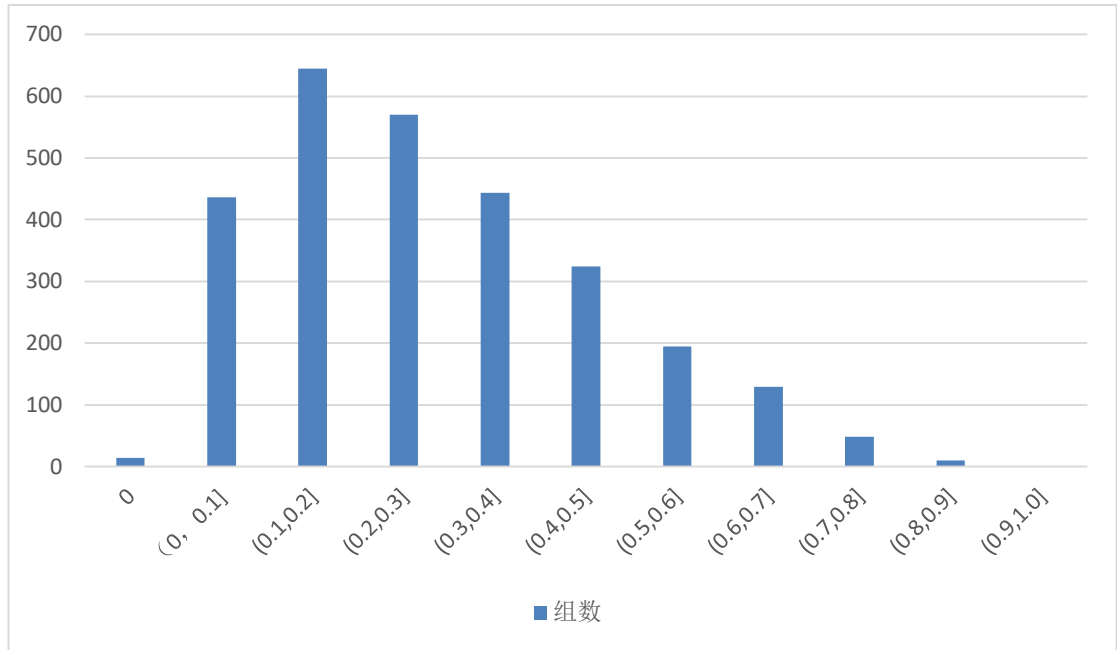


图 19：答复与留言相关度的区间统计

	A	B	C	D	E	F	G	H	I	J
1		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	law	value
2	0	2549	A00045581	A2区景蓉	2019/4/25	2019年4月 现将网友在2019/5/10	['《问政西地省》']			1
4	2	2555	A0003161	请加快提	2019/4/24	地处省会A 市民同志：2019/5/9	['《中共A市委A市人民政府关于学前教			1
5	3	2557	A0001107	在A市买公	2019/4/24	尊敬的书记网友 “A00 2019/5/9	['《问政西地省》’， ‘《A市人才购房			1
8	6	2849	A0001008	A3区教师	2019/3/25	尊敬的胡书记网友 “A00 2019/5/9	['《关于A市A3区既有多层住宅增设电			1
9	7	3681	UU00812	反映A5区	2018/12/5	我做为一方网友 “UU0 2019/1/25	['《A市A5区城镇小区配套幼儿园专项			1
13	11	3692	UU008829	A5区鄯阳	2018/12/2	胡书记：总网友 “UU0 2019/1/25	['《A市结合民用建筑修建防空地下室			1
17	15	3720	UU008444	关于A市新	2018/12/2	2016年下半年 网友 “UU0 2019/3/6	['《国有土地房屋征收补偿协议》’]			1
20	18	3747	UU008201	希望相关	2018/12/2	希望相关 网友 “UU0 2019/1/8	['《建筑工程夜间施工登记证明》’]			1
24	22	3762	UU008105	呼吁A5区	2018/12/2	尊敬的市领导网友 “UU0 2019/1/16	['《中华人民共和国道路交通安全法			1
25	23	3777	UU008162	关于A市地	2018/12/2	A市委市政 网友 “UU0 2019/1/25	['《A市轨道交通线网规划修编规划》’]			1
31	29	3871	UU008227	希望能在A	2018/12/1	怡海星城 网友 “UU0 2019/3/15	['《A市城市中小学校幼儿园规划建设			1
35	33	3910	UU008195	举报A市	2018/12/2	本人于201 网友 “UU0 2018/12/2	['《西地省物业服务收费管理办法》’]			1

图 20：可解释性（部分答复的法律条例的提取结果）

	A	B	C	D	E	F	G	H	I	J	K	L
1		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	result1	result2	result3	
2	0	2549	A00045581	A2区景蓉	2019/4/25	2019年4月 现将网友在2019/5/10	['您好']		['答复']	['感谢']	['感谢']	
3	1	2554	A0002358	A3区潘楚	2019/4/24	潘楚南路 网友 “A00 2019/5/9	['您好']		['回复']	['感谢']		
4	2	2555	A0003161	请加快提	2019/4/24	地处省会A 市民同志：2019/5/9	['你好']		['回复']	['感谢']		
5	3	2557	A0001107	在A市买公	2019/4/24	尊敬的书记网友 “A00 2019/5/9	['您好']		['回复']			
6	4	2574	A0009233	关于A市公	2019/4/25	建议将 “白 网友 “A00 2019/5/9	['您好']		['答复']	['感谢']		
7	5	2759	A0007753	A3区含浦	2019/3/25	欢迎领导来 网友 “A00 2019/5/9	['您好']		['回复']	['感谢']		
8	6	2849	A0001008	A3区教师	2019/3/25	尊敬的胡书记网友 “A00 2019/5/9	['您好']		['回复']	['感谢']		
9	7	3681	UU00812	反映A5区	2018/12/5	我做为一方网友 “UU0 2019/1/25	['您好']		['回复']	['感谢']		
10	8	3683	UU008792	反映A市	2018/12/5	我是美麓 网友 “UU0 2019/1/16	['您好']		['回复']	['感谢']		
11	9	3684	UU008687	反映A市	2018/12/5	胡书记好！ 网友 “UU0 2019/1/16	['您好']		['回复']	['感谢']		
12	10	3685	UU008220	反映A2区	2018/12/5	我家住在A 网友 “UU0 2019/3/11	['您好']		['回复']	['感谢']		
13	11	3692	UU008829	A5区鄯阳	2018/12/2	胡书记：总 网友 “UU0 2019/1/25	['您好']		['回复']	['感谢']		

图 21：完整性（部分答复的完整性关键词提取结果）

	A	B	C	D	E	F	G	H	I	J	K
1		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	result1	result2	字数统计
2	0	2549	A00045581A2区景蓉	2019/4/25	2019年4月现将网友在2019/5/10	['留言反映', '所反映的',	['调查', '核实', '调查				417
3	1	2554	A00023583A3区潇楚	2019/4/24	潇楚南路	网友 "A00 2019/5/9	['		['高度重视', '调查']		277
4	2	2555	A00031618	请加快提清	2019/4/24	地处省会A 市民同志: 2019/5/9	['		['		325
5	3	2557	A00011073	在A市买公	2019/4/24	尊敬的书记网友	"A00 2019/5/9	['	['留言已收悉']		270
6	4	2574	A0009233	关于A市公	2019/4/23	建议将 "白	网友 "A00 2019/5/9	['	['留言已收悉']		139
7	5	2759	A00077533A3区含浦	2019/4/23	欢迎领导	网友 "A00 2019/5/9	['		['高度重视']		211
8	6	2849	A00010080A3区教师	2019/3/25	尊敬的胡	网友 "A00 2019/5/9	['		['高度重视', '调查']		224
9	7	3681	UU00812	反映A5区	2018/12/3	我做为一	网友 "UU0 2019/1/29	['	['留言已收悉']	['核实', '核实']	576
10	8	3683	UU008792	反映A市美	2018/12/3	我是美麓	网友 "UU0 2019/1/16	['	['留言已收悉']	['	468
11	9	3684	UU008687	反映A市洋	2018/12/3	胡书记好!	网友 "UU0 2019/1/16	['	['留言已收悉', '所反映的']	['	203
12	10	3685	UU0082204	反映A2区	2018/12/3	我家住在A	网友 "UU0 2019/3/11	['	['留言已收悉']	['经查']	449
13	11	3692	UU008829	A5区鄱阳	2018/12/3	胡书记: 您	网友 "UU0 2019/1/29	['	['留言已收悉']	['调查', '核实']	374
14	12	3700	UU00877	A4区万国	2018/12/3	尊敬的书记	网友 "UU0 2019/1/14	['	['留言已收悉']	['经查']	123
15	13	3704	UU0081480	举报A市芒	2018/12/3	尊敬的领导	网友 "UU0 2019/1/3	['	['留言已收悉', '所反映的']	['经查']	90
16	14	3713	UU0081227	建议增开A	2018/12/3	建议增开A	网友 "UU0 2019/1/14	['	['留言已收悉']	['经查']	187
17	15	3720	UU008444	关于A市新	2018/12/3	2016年下	网友 "UU0 2019/3/6	['	['留言已收悉']	['经查']	592

图 22：重视度（部分答复的完整性关键词提取结果以及字数统计）

	A	B	C	D	E	F	G	H	I	J
1		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	day	second
2	0	2549	A00045581A2区景蓉	2019/4/25	2019年4月现将网友在2019/5/10				15	1315484
3	1	2554	A00023583A3区潇楚	2019/4/24	潇楚南路	网友 "A00 2019/5/9			14	1273530
4	2	2555	A00031618	请加快提清	2019/4/24	地处省会A 市民同志: 2019/5/9			14	1274950
5	3	2557	A00011073	在A市买公	2019/4/24	尊敬的书记网友	"A00 2019/5/9		14	1276932
6	4	2574	A0009233	关于A市公	2019/4/23	建议将 "白	网友 "A00 2019/5/9		15	1356491
7	5	2759	A00077533A3区含浦	2019/4/23	欢迎领导	网友 "A00 2019/5/9			31	2683488
8	6	2849	A00010080A3区教师	2019/3/25	尊敬的胡	网友 "A00 2019/5/9			40	3536735
9	7	3681	UU00812	反映A5区	2018/12/3	我做为一	网友 "UU0 2019/1/29		28	2464261
10	8	3683	UU008792	反映A市美	2018/12/3	我是美麓	网友 "UU0 2019/1/16		16	1402483
11	9	3684	UU008687	反映A市洋	2018/12/3	胡书记好!	网友 "UU0 2019/1/16		16	1403106
12	10	3685	UU0082204	反映A2区	2018/12/3	我家住在A	网友 "UU0 2019/3/11		70	6111363
13	11	3692	UU008829	A5区鄱阳	2018/12/3	胡书记: 您	网友 "UU0 2019/1/29		30	2633050

图 23：时效性（部分答复与留言的时间间隔）

对应的五个性质的得分乘以相应的权重，获得最终的评分，评分等级的分布如图 22 所示。其中等级为一般的答复所占比例最大，为 49%，其次是良好和不及格，分别占比 25%、19%，占比最低的等级为优秀，比例为 7%。

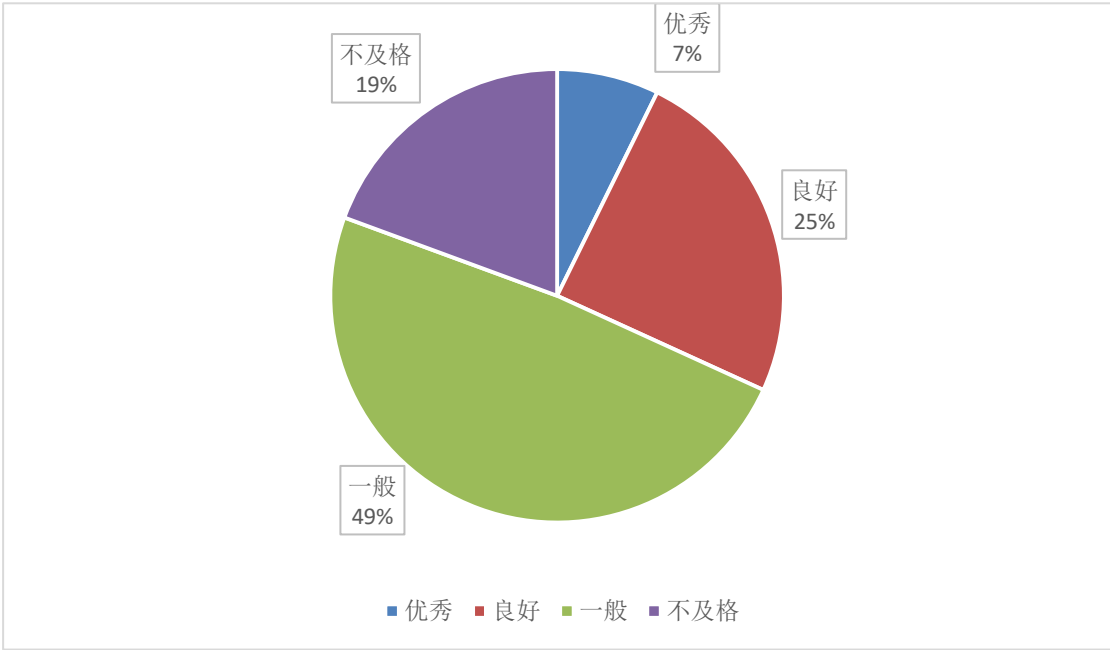


图 24 答复意见的等级分布

3、参考文献

- [1]: CSDN-Johnson07222 《中文分词的基本原理以及 jieba 分词的用法》
- [2]石凤贵. 《基于 TF-IDF 中文文本分类实现》[J]. 现代计算机, 2020(06):51-54+75.
- [3]Evangelopoulos, N., Amirkiaee, S.Y. Extracting LSA topics as features for text classifiers across different knowledge domains. Qual Quant 54, 249 - 261 (2020).
<http://h-s.doi.org.gduf.vpn358.com/10.1007/s11135-019-00954-x>
- [4]:搜狐《教程 | 一文读懂如何用 LSA、PSLA、LDA 和 lda2vec 进行主题建模》
选自 Medium 作者: Joyce Xu
- [5]:麦家健,朱凌峰,莫毅宇,陈志刚. 基于自然语言处理技术的警务情报文本挖掘分析[J]. 中国安防, 2019, (09):96-98.
- [6]: 龚海军. 网络热点话题自动发现技术研究[D]. 华中师范大学, 2008.
- [7]:CSDN-红豆和绿豆-《文本挖掘之文本聚类的介绍以及应用》
- [8]:CSDN-PanDawson-《K-Means & DBSCAN 聚类算法》
- [9]:CSDN-XianenZhou-《聚类方法: DBSCAN 算法研究(1)--DBSCAN 原理、流程、参数设置、优缺点以及算法》
- [10]: 简书-山的那边是什么_ 《DBSCAN 聚类原理》
- [11]:CSDN-博主-xiaomin_yan 《Reddit 算法分析》
- [12]许树柏. 实用决策方法:层次分析法原理[M]. 天津大学出版社, 1988.