

面向自然语言处理的“智慧政务”文本挖掘应用研究

摘 要

在当今的大数据时代里，伴随互联网和移动互联网的高速发展，层出不穷的网络留言、评论和点踩信息，已经成为政府机构迅速抓取热点话题、了解民心所向的重要途径。但各类社情民意相关的文本数据规模不断扩大，使得以往人工进行留言划分和热点整理的方法成为历史。因此，本文通过构建基于自然语言处理技术的智能模型，来解决留言分类和热点排名等问题。

针对问题一：在进行数据分析处理后，基于经典的 bert 模型进行改进，构建留言分类模型，再通过不断调制参数比较输出的 f1 值、accuracy 值和 loss 值结果得到较优解。

针对问题二：首先分析数据从而构建详细的热度综合评价指标体系，再处理数据以适应该体系。利用 jieba 分词和词频-逆文档（TF-IDF）模型进行文本分词和特征提取，构建词向量，然后利用 k-means 和 DBSCAN 结合的方式进行聚类。最后对长度前十类别进行热度排序和定义，再匹配原详细数据得到热点详情。

针对问题三：首先在数据处理阶段剔除掉答复意见的开场白话，形成新的数据集，再从答复的相关性、完整性、实时性、可解释性四个角度对答复意见给出评价方案。

关键词：自然语言处理；bert 模型；TF-IDF；k-means；DBSCAN.

Abstract:

In today's era of big data, with the rapid development of the Internet and mobile Internet, an endless stream of network messages, comments and information has become an important way for government agencies to quickly grasp hot topics and understand the aspirations of the people. However, the scale of all kinds of text data related to social conditions and public opinions is constantly expanding, which makes the previous methods of manual message division and hot spot sorting become a history. Therefore, this paper constructs an intelligent model based on natural language processing technology to solve the problems of message classification and hot spot ranking.

For problem 1: after data analysis, the classic Bert model was improved to build a message classification model, and then the better solution was obtained by continuously comparing the output f1 value, accuracy value and loss value through continuous modulation parameters.

For problem two: first, analyze the data to build a detailed heat comprehensive evaluation index system, and then process the data to adapt to the system. Jieba word segmentation and word frequency-inverse document (tf-idf) model are used to extract text word segmentation and features, construct word vectors, and then cluster by means of k-means and DBSCAN. Finally, the length of the top ten categories of heat sorting and definition, and then match the original detailed data to get the details of hot spots.

For question 3: firstly, the opening remarks of replies are removed in the data processing stage to form a new data set, and then the evaluation scheme is given from the four perspectives of relevance, integrity, real-time performance and interpretability of replies.

Keywords: natural language processing; Bert model; TF - IDF; K - means; DBSCAN.

目录

摘 要	- 1 -
Abstract:	- 2 -
一、引言（问题描述）	- 4 -
二、群众留言分类	- 4 -
2.1 模型框架	- 4 -
2.2 数据分析与预处理	- 5 -
2.2.1 数据分析	- 5 -
2.2.2 数据预处理	- 7 -
2.3 bert 模型改进	- 8 -
2.4 调制参数	- 10 -
三、热度排名	- 13 -
3.1 模型框架	- 13 -
3.2 构建热度评价指标	- 14 -
3.3 数据处理	- 14 -
3.4 构建词向量并聚类	- 15 -
3.4.1 分词、特征提取	- 15 -
3.4.2 聚类	- 16 -
3.5 热度定义	- 17 -
3.6 匹配留言	- 18 -
四、答复意见评价方案	- 19 -
4.1 数据的观察与处理	- 19 -
4.2 设计评价方案	- 20 -
五、总结与展望	- 21 -
5.1 总结	- 21 -
5.2 展望	- 21 -
5.2.1 基于 Bert 模型分类器的改进	- 21 -
5.2.2 热度定义的改进	- 22 -
5.2.3 对留言回复意见的改进	- 23 -
参考文献	- 25 -

一、引言（问题描述）

随着互联网和信息产业的快速发展，信息传递和更新的速度呈几何倍数增长，这也令信息总量爆发式提升。而由于产生数据的来源多样性以及数据本身的无序性、无规则性等原因，致使海量数据的价值密度不断降低。如何从海量数据中提取信息产生高价值，成为了当今信息处理研究领域的主流。这也为大数据、云计算、人工智能等技术的蓬勃发展提供了良好的前景。

近年来，微博、微信、QQ 等各大交流平台普及，逐步成为了政府了解民意、汇聚民智、凝聚民气的重要渠道，而文本数据量的高速提升，使得人工进行留言分类和热点问题整理几乎变得不可行，建立基于自然语言处理的智慧政务系统已经成为政府妥善治理社会的新走向，对提升政府的管理水平与施政效率当有极大的推动效果。

基于对自然语言处理技术以及对分类挖掘、热点问题的相关认识，本文将致力于研究公众的留言分类和热点排序问题，并建立一个自然语言处理模型，在 F1-score、准确率以及泛化能力上都表现出不错的效果。

二、群众留言分类

2.1 模型框架

群众留言分类问题显然属于自然语言处理（NLP）的范畴，经过查找资料，我们选择了 NLP 中处理无标注预料特别出色的 bert 模型作为模型的基础：

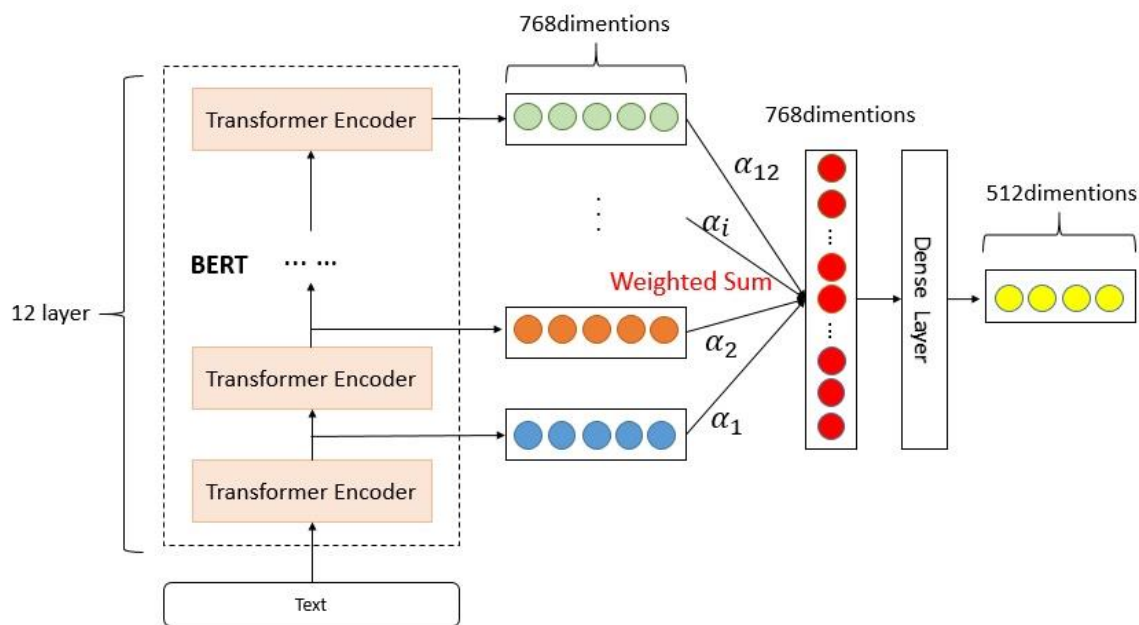


图 1. BERT 模型框架图

在此之上进行优化构建自己的群众留言分类模型。该模型主要包括三个部分：数据分析与预处理、bert 模型改进以及调置参数。

第一步：数据分析与预处理。我们根据对问题的理解，首先对原始数据进行分析观察，然后通过预处理将各种脏数据和无用数据清除，再得出适应于 bert 模型的数据集。

第二步：bert 模型改进。Bert 模型只是一个基础模型，还需要对其进行进一步的改进以适应留言分类问题，从而做到正确的输入输出。

第三步：调置参数。Bert 模型预训练需要输入相应参数，而参数的改变会影响输出结果，所以在后期需要对模型参数进行微调，多次训练，从而得到最优的参数。

2.2 数据分析与预处理

2.2.1 数据分析

高质量的数据集是模型优化和改进的基础，而对整个留言数据集进行分析处理可以令我们对数据集有一个的宏观的把控认知，从而更好地对数据做出妥善处理，使数据的使用率和准确率达到最大化。也更为后面调制参数做好了铺垫，减少了重复摸索的概率。

留言分类模型的构建过程只用到了附件 2.xlsx 中的数据集。

1	留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
2	24	A00074011	明建筑集团上海施工有安	20/1/6 12:09:	围墙门。每天尤其上	城乡建设
3	37	U0008473	市大厦人为烂尾多年，安	20/1/4 11:17:	看，不但占用人行道	城乡建设
4	83	A00063999	市A1区苑物业违规收停	9/12/30 17:06	已多次向物业和社区	城乡建设
5	303	U0007137	南路A2区华庭楼顶水箱	19/12/6 14:40:	品，霉是一种强致癌	城乡建设
6	319	U0007137	A2区华庭自来水好大一	19/12/5 11:17:	品，霉是一种强致癌	城乡建设
7	379	A00016773	市盛世耀凯小区物业无	19/11/28 9:08:	物业不是为业主服务	城乡建设
8	382	U0005806	咨询A市楼盘集中供暖一	9/11/27 17:14	月亮岛片区近年规划	城乡建设
9	445	A00019209	西路可可小城长期停水	9/11/19 22:39	求帮助至今没有找到	城乡建设
10	476	U0003167	收取城市垃圾处理费不	9/11/15 11:44	在的物业公司也未给	城乡建设
11	530	U0008488	A3区魏家坡小区脏乱差	9/11/10 18:59	人让人好好休息一下	城乡建设
12	532	U0008488	A市魏家坡小区脏乱差	9/11/10 12:30	人让人好好休息一下	城乡建设
13	673	A00080647	市四届非法业委会涉嫌	9/10/24 11:29	责令B4区有关部门	城乡建设
14	994	U0005196	梅溪湖壹号御湾业主用	19/9/18 22:43:	别的城市都已经一	城乡建设
15	1005	U0006509	翡翠湾强行对入住的业	19/9/18 13:36:	清地公司和金晖物	城乡建设
16	1110	A00099772	市锦楚国际星城小区三	19/9/9 11:07:	是无通知，突然断	城乡建设
17	1309	U0005083	和紫郡用电的问题能不	19/8/21 15:12:	起之后，我们的用	城乡建设
18	1440	A0003288	际新城从6月份开始停	19/8/6 10:28:	的生活，而且我们	城乡建设
19	1775	U0002150	城区南西片区城铁站	19/7/4 18:52:	达A市，并且规划有	城乡建设
20	1783	U0004763	政府加大对滨水新城	19/7/4 14:25:	的或者几个半大小	城乡建设
21	1827	U000613	区楚府线几个小区经	19/7/1 20:14:	已停电三次。说是	城乡建设
22	2603	A00099650	团及西地省辉东安建工	19/4/20 16:50:	说不出。去年8月，	城乡建设
23	3607	A00046529	山水嘉园1栋三单元群	19/1/8 10:08:	隐患，投诉给物业	城乡建设
24	3742	A00013884	小区外的非法汽车检测	18/12/26 10:13:	设备检定（省级）	城乡建设
25	3800	U0001518	修建中速磁悬浮（最高	18/12/20 1:23:	与京广高铁B市西	城乡建设
26	3874	U0007328	旁露天垃圾池子臭味熏	18/12/11 21:40:	无专人值守和管理	城乡建设
27	3980	U0001518	线西延二期暂缓修建，	18/12/1 1:05:	至I市的中速磁悬	城乡建设
28	3981	U0001518	站至A市火车站、A市	18/12/1 0:14:	A市南站、A市东	城乡建设
29	4042	U0007328	道脏乱差、露天垃圾池	18/11/25 12:23:	区（大桥二区安置	城乡建设
30	4055	A00025379	市一样给居民小区统一	18/11/23 9:08:	也跟北方一样，统	城乡建设

图 2. 附件 2 留言数据部分截图

留言数据主要分为六个部分：留言编号、留言用户、留言主题、留言时间、留言详情以及一级标签。基于留言文本分类的目的，我们可以很容易看出，留言编号、留言用户和留言时间三个部分的内容意义不大，可以考虑在数据预处理阶段剔除。而留言分类问题的关键显然就是留言主题和留言详情，所以在与处理时应当将两者和一级标签保留。我们建立模型的目的是要实现对数据集的学习训练，从而能够对完全陌生的测试集进行准确率较高的标注分类。通过资料搜集，我们发现在自然语言处理领域处理无标注语料具有强大功能的bert模型能够很好地胜任这一要求。基于bert模型创建留言分类模型，则需要语料和标注两个部分，所以我们有三个方向的数据集可以考虑：

- (1) 一级标签和留言主题。
- (2) 一级标签和留言详情。
- (3) 一级标签、留言主题与留言详情合并后的数据。

通过综合考量，我们用了第(2)个方向的数据集，这里我们默认数据预处理阶段的数据集就是以上的格式，剔出了留言编号、留言用户和留言时间和留言主题。

2.2.2 数据预处理

1.数据预处理阶段首先考虑的是处理脏数据的问题。所谓的脏，指数据可能存在于以下几种问题（主要问题）：

- （1）数据缺失（Incomplete）是属性值为空的情况。
- （2）数据噪声（Noisy）是数据值不合常理的情况。
- （3）数据不一致（Inconsistent）是数据前后存在矛盾的情况。
- （4）数据冗余（Redundant）是数据量或者属性数目超出数据分析需要的情况。
- （5）数据集不均衡（Imbalance）是各个类别的数据量相差悬殊的情况。
- （6）离群点/异常值（Outliers）是远离数据集中其余部分的数据。
- （7）数据重复（Duplicate）是在数据集中出现多次的数据。

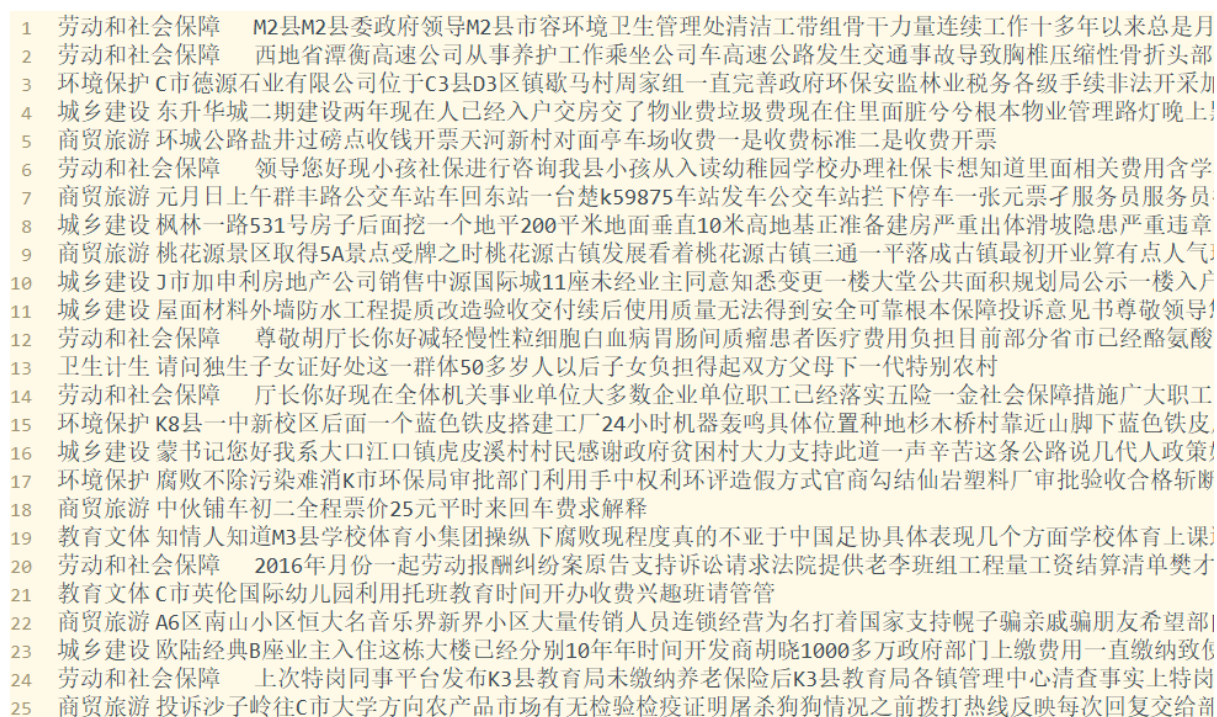
经过统计，原始数据主要有数据冗余、数据重复的问题。对于数据冗余情况，我们采取的解决措施是直接删除 bert 模型一次所能处理的最大长度后面的内容；对于数据重复问题，我们仍采用直接删除重复项的方法。

2.数据预处理第二阶段就是删除标点符号和停用词，对留言分类的自然语言处理模型来说，标点符号和停用词不仅没有意义，还会对分类的准确率造成一定影响，提升了训练数据集的难度，所以我们决定剔除标点和停用词。

3.第三阶段是给数据集打乱顺序。由于深度学习模型训练的 batch_size 是一定的，所以如果直接采用原数据，则每次训练的数据可能都是同一类型，就会产生过拟合，所以在训练模型前一定要打乱数据。

4.第四阶段是划分数据。有了模型后，训练集就是用来训练参数的，说准确点，一般是用来梯度下降的。而验证集基本是在每个 epoch 完成后，用来测试一下当前模型的准确率。因为验证集跟训练集没有交集，因此这个准确率是可靠的。也就是说，从狭义来讲，验证集没有参与梯度下降的过程，也就是说是没有经过训练的；但从广义上来看，验证集却参与了一个“人工调参”的过程，我们根据验证集的结果调节了迭代数、调节了

学习率等等,使得结果在验证集上最优。因此,我们也可以认为,验证集也参与了训练。那么就很明显了,我们还需要一个完全没有经过训练的集合,那就是测试集,我们既不用测试集梯度下降,也不用它来控制超参数,只是在模型最终训练完成后,用来测试一下最后准确率。由于我们只有一个大的标注数据集,想要完成一个有监督模型的测试,那么通常使用均匀随机抽样的方式,将数据集划分为训练集、验证集、测试集,这三个集合不能有交集,设置比例是 8:1:1,当然比例是人为可调的。从这个角度来看,三个集合都是同分布的。



1 劳动和社会保障 M2县M2县委政府领导M2县市容环境卫生管理处清洁工带组骨干力量连续工作十多年以来总是月
2 劳动和社会保障 西地省潭衡高速公司从事养护工作乘坐公司车高速公路发生交通事故导致胸椎压缩性骨折头部
3 环境保护 C市德源石业有限公司位于C3县D3区镇歇马村周家组一直完善政府环保安监林业税务各级手续非法开采加
4 城乡建设 东升华城二期建设两年现在人已经入户口交房交了物业费垃圾费现在住里面脏兮兮根本物业管理路灯晚上!
5 商贸旅游 环城公路盐井过磅点收钱开票天河新村对面亭场收费一是收费标准二是收费开票
6 劳动和社会保障 领导您好现小孩社保进行咨询我县小孩从入读幼稚园学校办理社保卡想知道里面相关费用(含学
7 商贸旅游 元月日上午群丰路公交车车站回东站一台楚k59875车站发车公交车站拦下停车一张元票子服务员服务员
8 城乡建设 枫林一路531号房子后面挖一个地平200平米地面垂直10米高地基正准备建房严重出体滑坡隐患严重违章
9 商贸旅游 桃花源景区取得5A景点受牌之时桃花源古镇发展看着桃花源古镇三通一平落成古镇最初开业算有点人气!
10 城乡建设 J市加申利房地产公司销售中源国际城11座未经业主同意知悉变更一楼大堂公共面积规划局公示一楼入户
11 城乡建设 屋面材料外墙防水工程提质改造验收交付续后使用质量无法得到安全可靠根本保障投诉意见书尊敬领导!
12 劳动和社会保障 尊敬胡厅长你好减轻慢性粒细胞白血病胃肠间质瘤患者医疗费用负担目前部分省市已经酪氨酸
13 卫生计生 请问独生子女证好处这一群体50多岁人以后子女负担得起双方父母下一代特别农村
14 劳动和社会保障 厅长你好现在全体机关事业单位大多数企业单位职工已经落实五险一金社会保障措施广大职工
15 环境保护 K8县一中新校区后面一个蓝色铁皮搭建工厂24小时机器轰鸣具体位置种地杉木桥村靠近山脚下蓝色铁皮
16 城乡建设 蒙书记您好我系大口江口镇虎皮溪村村民感谢政府贫困村大力支持此道一声辛苦这条公路说几代人政策!
17 环境保护 腐败不除污染难消K市环保局审批部门利用手中权利环评造假方式官商勾结仙岩塑料厂审批验收合格斩断
18 商贸旅游 中伙铺车初二全程票价25元平时来回车费求解释
19 教育文体 知情人知道M3县学校体育小集团操纵下腐败现程度真的不亚于中国足协具体表现几个方面学校体育上课
20 劳动和社会保障 2016年月份一起劳动报酬纠纷案原告支持诉讼请求法院提供老李班组工程量工资结算清单樊才
21 教育文体 C市英伦国际幼儿园利用托班教育时间开办收费兴趣班请管管
22 商贸旅游 A6区南山小区恒大名音乐界新界小区大量传销人员连锁经营为名打着国家支持幌子骗亲戚骗朋友希望部
23 城乡建设 欧陆经典B座业主入住这栋大楼已经分别10年年时间开发商胡晓1000多万政府部门上缴费用一直缴纳致停
24 劳动和社会保障 上次特岗同事平台发布K3县教育局未缴纳养老保险后K3县教育局各镇管理中心清查事实上特岗
25 商贸旅游 投诉沙子岭往C市大学方向农产品市场有无检验检疫证明屠杀狗狗情况之前拨打热线反映每次回复交给部

图 3. 数据预处理后的验证集部分数据截图

2.3 bert 模型改进

BERT 非常友好的一点就是对于 NLP 任务,我们只需要对最后一层进行微调便可以用于我们的项目需求。我们只需要将我们的数据输入处理成标准的结构进行输入就可以了。在 run_classifier.py 文件中有一个基类 DataProcessor 类,在这个基类中定义了一个读取文件的静态方法_read_tsv,四个分别获取训练集,验证集,测试集和标签的方法。接下来我们要定义自己的数据处理的类,我们将我们的类命名为 MyTaskProcessor,

MyTaskProcessor 继承 DataProcessor，用于定义我们自己的任务。

```
1 class MyTaskProcessor(DataProcessor):
2     """Processor for my task-news classification """
3     def __init__(self):
4         self.labels = ['劳动保障', '城乡建设', '教育文体', '卫生计生', '交通运输', '商贸旅游', '环境保护']
5     def get_train_examples(self, data_dir):
6         return self._create_examples(
7             self._read_tsv(os.path.join(data_dir, 'train.tsv')), 'train')
8     def get_dev_examples(self, data_dir):
9         return self._create_examples(
10            self._read_tsv(os.path.join(data_dir, 'dev.tsv')), 'val')
11    def get_test_examples(self, data_dir):
12        return self._create_examples(
13            self._read_tsv(os.path.join(data_dir, 'test.tsv')), 'test')
14    def get_labels(self):
15        return self.labels
16    def _create_examples(self, lines, set_type):
17        """create examples for the training and val sets"""
18        examples = []
19        for (i, line) in enumerate(lines):
20            guid = '%s-%s' % (set_type, i)
21            text_a = tokenization.convert_to_unicode(line[1])
22            label = tokenization.convert_to_unicode(line[0])
23            examples.append(InputExample(guid=guid, text_a=text_a, label=label))
24        return examples
25
```

图 4. MyTaskProcessor 类代码截图

接下来就可以训练数据集了，然而原生 BERT 代码中验证集的输出指标值只有 loss 和 accuracy，并没有题目中要求的测定指标 F1 值，所以需要另写程序算出 F1 值。F1 值计算公式为：

$$F1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i}$$

由于 bert 模型本身只支持二分类问题，对于该多分类问题并不适用，故需要另找方法，在查阅相关资料后，我们发现可以通过在 bert 模型基础上计算混淆矩阵的方法解决多分类问题，所以另编写了一个 metrics.py 程序算出查准率 P 值、查全率 R 值和 F1 值。

```

1 import numpy as np
2 import tensorflow as tf
3 from tensorflow.python.ops.metrics_impl import _streaming_confusion_matrix
4 def get_metrics_ops(labels, predictions, num_labels):
5     # 得到混淆矩阵和update_op, 在这里我们需要将生成的混淆矩阵转换成tensor
6     cm, op = _streaming_confusion_matrix(labels, predictions, num_labels)
7     tf.logging.info(type(cm))
8     tf.logging.info(type(op))
9     return (tf.convert_to_tensor(cm), op)
10
11 def get_metrics(conf_mat, num_labels):
12     # 得到numpy类型的混淆矩阵, 然后计算precision, recall, f1值。
13     precisions = []
14     recalls = []
15     for i in range(num_labels):
16         tp = conf_mat[i][i].sum()
17         col_sum = conf_mat[:, i].sum()
18         row_sum = conf_mat[i].sum()
19
20         precision = tp / col_sum if col_sum > 0 else 0
21         recall = tp / row_sum if row_sum > 0 else 0
22
23         precisions.append(precision)
24         recalls.append(recall)
25
26     pre = sum(precisions) / len(precisions)
27     rec = sum(recalls) / len(recalls)
28     f1 = 2 * pre * rec / (pre + rec)
29
30     return pre, rec, f1

```

图 5. Metrics 程序截图

至此，bert 模型改进基本完善，可以进行较好地训练，得到正确的输入输出。

2.4 调制参数

运行 bert 模型需要输入训练参数，不同参数会产生不同的训练结果，所以调制参数阶段的目的是要使训练结果达到最优。

```

1  --task_name=mytask
2  \
3  --do_train=true
4  \
5  --do_eval=true
6  \
7  --data_dir=../GLUE/glue_data/cnews-tab/
8  \
9  --vocab_file=../GLUE/BERT_BASE_DIRinese_L-12_H-768_A-12vocab.txt
10 \
11 --bert_config_file=../GLUE\BERT_BASE_DIRinese_L-12_H-768_A-12/bert_config.json
12 \
13 --init_checkpoint=../GLUE\BERT_BASE_DIRinese_L-12_H-768_A-12/bert_model.ckpt
14 \
15 --max_seq_length=128
16 \
17 --train_batch_size=16
18 \
19 --learning_rate=2e-5
20 \
21 --num_train_epochs=2.0
22 \
23 --output_dir=mytask_output

```

图 6. Bert 模型输入参数示例图

一共需要输入 12 个参数，其他参数都是一些名称和路径问题，对训练效果不会产生影响，我们主要关注的是第 8-11 个参数: max_seq_length 是 bert 模型一次能够处理数据的最大长度，我们选用大于 80% 以上留言详情长度的值作为最大长度，以下是留言详情长度统计：

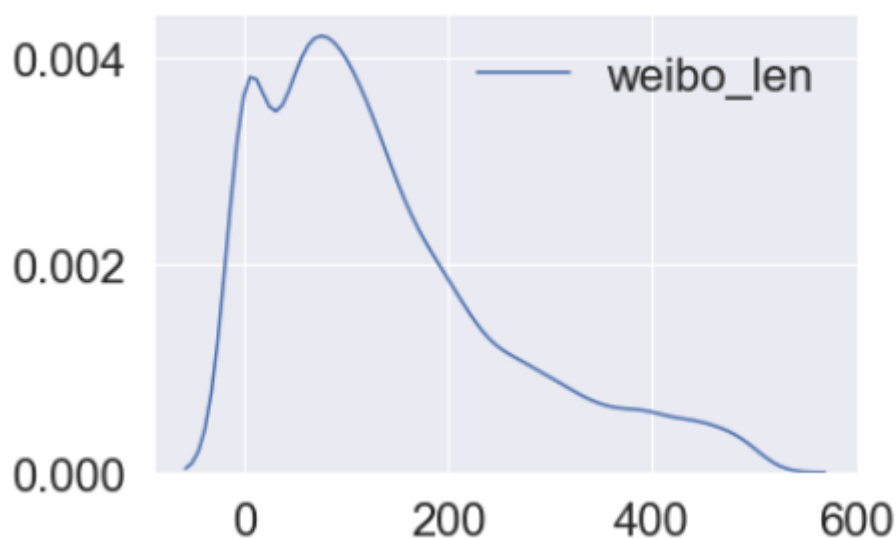


图 7. 留言详情长度统计图

分析统计图，我们采用估值 300 作为模型训练的最大长度 max_seq_length；train_batch_size 是我们用来更新梯度的批数据大小；learning_rate 是学习率；num_train_epochs 是训练模型的迭代次数。我们主要看损失是否收敛在一个稳定值，若收敛则当前设置的 epoch 为最佳。调制参数的核心就是利用控制变量法不断改变某一个参数的输入值，记录下来作比较，得出最优参数。

下面是我们经过试验得出的三个相对较优结果及相应参数：

```
INFO:tensorflow:evaluation_loop marked as finished
INFO:tensorflow:***** Eval results *****
INFO:tensorflow:eval_precision: 0.9209826858280975
INFO:tensorflow:eval_recall: 0.9229627566115174
INFO:tensorflow:eval_f1: 0.9219716580971636
INFO:tensorflow:eval_accuracy: 0.9262472987174988
INFO:tensorflow:eval_loss: 0.27129846811294556
root@fc63d29c9816:/opt/bert_test/bert-master# python run_classifier.py --task_name=mytask --do_train=true --do_eval=true --data_dir=./GLUE/glue_data/cnews-tab/ --vocab_file=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/vocab.txt --bert_config_file=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/bert_config.json --init_checkpoint=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/bert_model.ckpt --max_seq_length=300 --train_batch_size=40 --learning_rate=2e-5 --num_train_epochs=3.0 --output_dir=mytask_output
```

图 8. train_batch_size=40 输出结果图

```
INFO:tensorflow:eval_precision: 0.9205163506277294
INFO:tensorflow:eval_recall: 0.9196735255948056
INFO:tensorflow:eval_f1: 0.920094745100197
INFO:tensorflow:eval_accuracy: 0.9273318648338318
INFO:tensorflow:eval_loss: 0.2692359685897827
root@fc63d29c9816:/opt/bert_test/bert-master# python run_classifier.py --task_name=mytask --do_train=true --do_eval=true --data_dir=./GLUE/glue_data/cnews-tab/ --vocab_file=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/vocab.txt --bert_config_file=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/bert_config.json --init_checkpoint=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/bert_model.ckpt --max_seq_length=300 --train_batch_size=50 --learning_rate=2e-5 --num_train_epochs=3.0 --output_dir=mytask_output
```

图 9. train_batch_size=50 输出结果图

```
INFO:tensorflow:evaluation_loop marked as finished
INFO:tensorflow:***** Eval results *****
INFO:tensorflow:eval_precision: 0.9196971083804275
INFO:tensorflow:eval_recall: 0.923050971502824
INFO:tensorflow:eval_f1: 0.9213709878701118
INFO:tensorflow:eval_accuracy: 0.9262472987174988
INFO:tensorflow:eval_loss: 0.2530841529369354
root@fc63d29c9816:/opt/bert_test/bert-master# python run_classifier.py --task_name=mytask --do_train=true --do_eval=true --data_dir=./GLUE/glue_data/cnews-tab/ --vocab_file=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/vocab.txt --bert_config_file=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/bert_config.json --init_checkpoint=./GLUE/BERT_BASE_DIR/chinese_L-12_H-768_A-12/bert_model.ckpt --max_seq_length=512 --train_batch_size=64 --learning_rate=2e-5 --num_train_epochs=3.0 --output_dir=mytask_output
```

图 10. train_batch_size=64 输出结果图

主要是 train_batch_size 参数不同致使结果有些许差异，互有优劣，取 40 时 f1 值最高；取 50 时准确率 accuracy 最高；取 64 时损失值最低。

三、热度排名

3.1 模型框架

随着信息技术的高速发展，网络舆情总是在瞬息间发生变化，抓住一段时间的网络热点话题，可以抓住舆论走向，对于政府机关来说，是最真实也是最有效获取社会民生问题的方法，从而能做到及时作出答复和改变，使人民过的更好。热度排名模型就是基于这个原因建立的，能够更清晰的得出一段时间内人民最关心的实际问题，提升政府机关的工作效能。热度排名模型最主要包括五个部分：构建热度评价指标、数据处理、构建词向量并聚类、热度定义以及匹配留言。

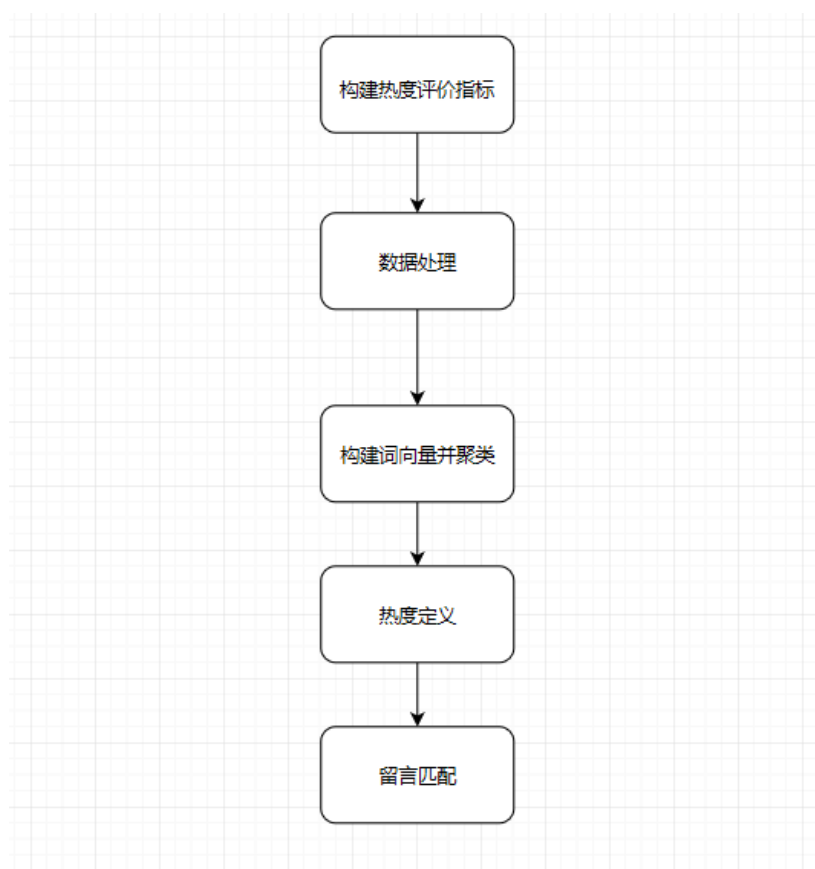


图 11. 热度排名模型框架图

3.2 构建热度评价指标

如何评价一个话题是否为热点话题是热度排名模型最关键一步，这里我们参考了微博热度评价指标的构建方法，再基于本体数据集，得出比较妥善的评价方法：考虑内容特征热度影响力（这里由于数据原因只考虑话题出现的及时性和突然性，即时间因素）和传受众特征热度影响力（即点赞数、反对数和同一个话题的留言数）。同一个话题若是有多人提及，毫无疑问应该还是比较重要的，而一个话题点赞数和反对数的多少也可以直观反映出该话题的影响程度。基于直接影响大于间接影响因素的考虑，我们人为设定点赞数或反对数与留言数的权重比为 0.6:1。而话题的及时性也至关重要，如果是时间跨度很长的话题，但是实际上已经解决或是每隔一段时间就发生，就显得没那么重要、所以我们对传受众特征热度影响力最高的 10 个话题做了如下处理：

$$\text{热度} = \frac{\text{传受众特征热度影响力}}{\text{话题持续时间}}$$

以算出的数值作为最终热度进行排名得出结果。据此构建出的热度评价指标体系如下表所示：

热度 综合 评价 指标 体系	一级指标	二级指标	指标内涵	权值
	传受众 特征热 度影响力	点赞数和反对 数之和	话题的所有点 赞数和反对数 之和	0.6
		留言数	属于该话题的 留言数	1
	内容特征热度 影响力	话题持续时间	话题结束时间 减去开始时间	

3.3 数据处理

基于给出的热度评价指标，我们考虑到很难在聚类过后找到某一类中各数据对应的准确点赞数和反对数。所以我们创新性的将某一留言的点赞数与反对数之和乘以 0.6 的权值得出的数转化为该留言的重复数据的数目，也就是人工加入重复数据进行聚类，这样就可以直接通过类的长度得出其传受众特征热度影响力。

189180,A000106515,A市人才购房补贴申请是否与单位注册地有关?,2019/6/18 9:51:36,"

胡书记,您好。我于19年三月在长购房,五月初申报了人才购房补贴,六月份审核未通过,理由是“工作单位非在长工商注册的企业”。我所在公司为集团化公司,注册地在北京,但A市公司有数千人。事实上,[政府发文](2017)112号文件的要求是“在长工作”,我的社保公积金可以体现在长工作,单位出具的在职证明也写明了工作地点在A市。个人认为,审核人员错误解读了政策文件。此外,最近今天反复拨打了政策咨询电话0000-00000000几十次,没有一次打通。A市人才app也没有留言咨询的功能。民众没有便利的渠道咨询了解相关政策。请书记安排相关部门解答相关问题,并完善政策解读交流渠道。

*,0,3

189180,A000106515,A市人才购房补贴申请是否与单位注册地有关?,2019/6/18 9:51:36,"

胡书记,您好。我于19年三月在长购房,五月初申报了人才购房补贴,六月份审核未通过,理由是“工作单位非在长工商注册的企业”。我所在公司为集团化公司,注册地在北京,但A市公司有数千人。事实上,[政府发文](2017)112号文件的要求是“在长工作”,我的社保公积金可以体现在长工作,单位出具的在职证明也写明了工作地点在A市。个人认为,审核人员错误解读了政策文件。此外,最近今天反复拨打了政策咨询电话0000-00000000几十次,没有一次打通。A市人才app也没有留言咨询的功能。民众没有便利的渠道咨询了解相关政策。请书记安排相关部门解答相关问题,并完善政策解读交流渠道。

*,0,0

图 12. 重复数据加入效果图

3.4 构建词向量并聚类

3.4.1 分词、特征提取

我们使用的是“留言主题”数据,文本聚类首先要进行分词和特征提取,并将一些停用词、字母、标点删去,以期让聚类效果提升。这里我们采用中文分词效果显著的 jieba 进行分词,而特征提取则采用 TF-IDF 算法,处理后数据如下:

```
1  市 的 尿 毒 症 患 者 太 苦 了
2  反 映 区 东 岸 乡 新 安 村 八 组 拆 迁 问 题
3  市 汽 车 南 站 何 时 能 建 好
4  县 江 背 镇 的 自 来 水 主 管 为 什 么 经 常 爆 裂
5  市 银 华 园 共 用 部 位 设 施 的 租 金 去 哪 了
6  市 中 级 人 民 法 院 指 鹿 为 马
7  区 暮 云 水 电 八 局 南 托 基 地 噪 音 扰 民
8  市 楚 江 路 辅 道 交 通 信 号 灯 设 置 不 合 规 吧
9  县 楚 龙 西 路 何 时 才 能 打 通
10 县 文 体 中 心 乒 羽 中 心 何 时 能 建 成 开 放
11 区 中 海 国 际 社 区 期 北 门 前 面 那 片 空 地 通 宵 施 工 扰 民
12 请 解 决 县 松 雅 湖 烂 尾 问 题
13 区 保 利 大 都 汇 违 规 调 规
14 市 国 道 东 八 线 货 车 噪 音 扰 民
15 市 城 际 铁 路 小 编 组 列 车 设 置 不 合 理
16 反 映 区 井 圭 路 农 贸 市 场 的 诸 多 问 题
17 请 政 府 接 管 湖 楚 财 富 实 质 性 启 动 善 后 清 退 工 作
18 区 学 而 思 培 优 华 盛 花 园 培 训 点 噪 音 扰 民 影 响 居 民 生 活
19 县 星 沙 电 建 星 湖 湾 高 层 房 子 为 什 么 一 直 不 开 盘
20 加 快 市 国 王 陵 考 古 公 园 建 设 刻 不 容 缓
```

图 13. 分词、特征提取后数据部分截图

3.4.2 聚类

文本聚类一般采用 k-means 聚类和 DBSCAN (密度聚类)。但两者均有不足之处，所以我们大胆采用两者合并使用的方法来提高词向量聚类的准确性：

第一步：先用 k-means 聚类的方法对整个数据集进行聚类。通过我们不断调参实验，发现当 $k=5$ 时聚类效果最好。

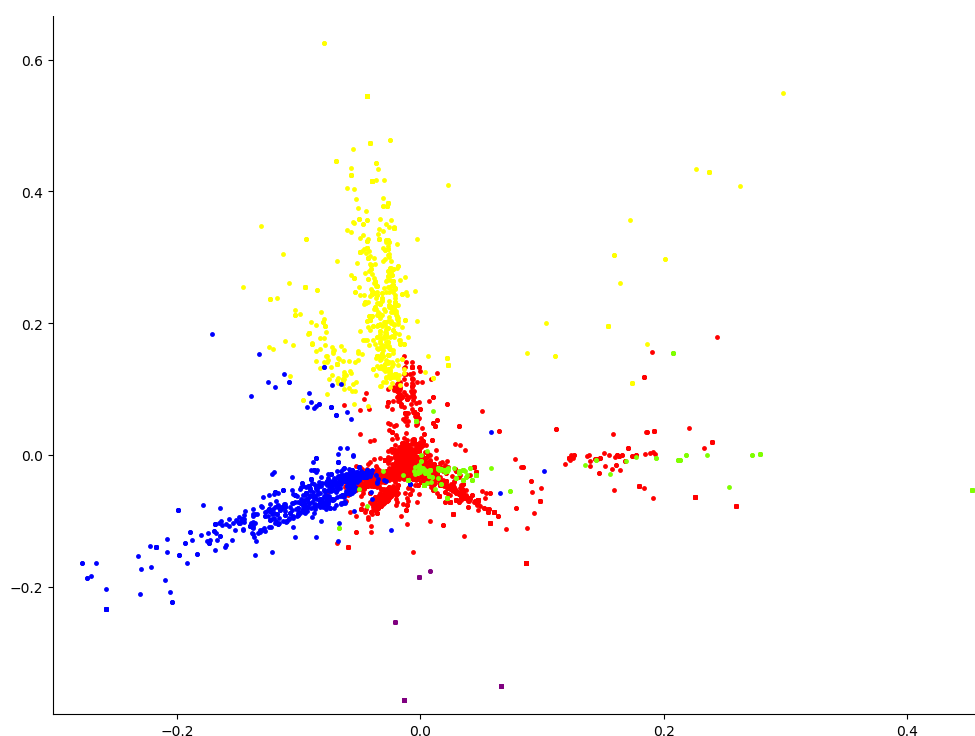


图 14. K-means 聚类效果图

但是也能明显感觉到有一些类有拖尾数据，所以更坚定了我们使用密度聚类来优化的想法。

第二步：将第一步得到的五个数据集分别进行密度聚类，但是考虑到 k-means 聚类可能将一个大的类分割成了两个以上的小类，基于热度排名只需要前五个热点话题的需求，我们取五个数据集密度聚类后的长度前 20 名类数据打乱，再次进行密度聚类，得出最终聚类结果。

1 ['市','拉','县改黑','市','县改黑','小','拉','拉','区','拉','拉','县改黑','县改黑',
2 ['请加快市国家中心建设','请加快市国家中心建设','请市加快地铁建设','请市加快轨道交通建设力度'],
3 ['县星改造','县星沙改造拖何年何月动工','县星沙改造拖何年何月动工','县星沙街道改造进行','县星早
4 ['市伊景园捆绑销售车位是否合理','投诉市伊景园捆绑车位销售','景园捆绑车位销售合法','投诉市伊景园
5 ['区长施工','区工地长期施工扰民','小区建搅拌站合理','区公寓工地施工扰民','区小区工地施工扰民
6 ['请问市包支付任务不让市场正竞争','请问市包支付任务不让市场正竞争','请问市包支付任务不让市场
7 ['问问市经规划','问问市经规划','问问市经规划','问问市经规划','问问市经规划','问问市经规划'],
8 ['市大道全线快速化改造启动','市大道全线快速化改造启动','市大道全线快速化改造启动','市大道全线快
9 ['建议市经收回恒天工厂地块打造商业综合体','建议市经收回恒天工厂地块打造商业综合体','建议市经收回
10 ['居住地铁号线县松地省站方向民众心声','居住地铁号线县松地省站方向民众心声','居住地铁号线县松地省
11 ['请敦促区教育局落实发放原市退休教师文明单位奖','请敦促区教育局落实发放原市退休教师文明单位奖'],
12 ['建议加大县东榔梨段拆迁力度','建议加大县东榔梨段拆迁力度','建议加大县东榔梨段拆迁力度','建议加
13 ['区小学','区小学','区小学','市小学停','区小学','区小学','区小学','区小学','区小学','区小学'],
14 ['市长时代房子裂缝质量堪忧','市长时代房子裂缝质量堪忧','市长时代房子裂缝质量堪忧','市长时代房子
15 ['建议外迁高速城区','建议外迁高速城区','建议外迁高速城区','建议外迁高速城区','建议外迁高速城区
16 ['建议出让星沙路路土地','建议出让星沙路路土地','建议出让星沙路路土地','建议出让星沙路路土地'],
17 ['县东高速能否通车','请问县东高速能否通车','县东高速能否通车','请问县东高速能否通车','县东高速
18 ['区月亮岛路高压线建议','区月亮岛路高压线建议','区月亮岛路架设高压线投诉','区月亮岛路架设高压线
19 ['反映市地铁号线站点地下通道问题','反映市地铁号线站点地下通道问题','反映市地铁号线站点地下通道问
20 ['咨询市高铁站选址问题','咨询市高铁站选址问题','咨询市高铁站选址问题','咨询市高铁站选址问题'],

图 15. 最终聚类结果部分截图

3.5 热度定义

对排名前十的热点话题进行定义和描述,得出热点问题表。

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述	
1	1	1259.2	2019/8/19至2019/8/19	A市A5区汇金路五矿万境K9县	A市A5区汇金路五矿万境K9县存在一系列问题	
2	2	147.6	2019/6/19至2019/6/19	A市富绿物业丽发新城	A市富绿物业丽发新城强行断业主家水	
3	3	49	2019/1/10至2019/1/10	西地省	建议西地省尽快外迁京港澳高速城区段至远郊	
4	4	34.06	2019/8/23至2019/9/5	A4区绿地海外滩小区	A4区绿地海外滩小区距长赣高铁太近	
5	5	13.522	2019/3/12至2019/5/29	A市辰北三角洲幼儿园, A市金毛湾, A2区中建A1区嘉苑, A5区万科魅力之, A3区协调解决旭辉御府/子女, 业主小孩	配套入学的问题	
6	6	9	2019/1/2至2019/1/11	A市经开区东六线以西泉塘	问问A市经开区东六线以西泉塘和商业中心以南的有关规划	
7	7	7.83	2019/1/11至2019/5/29	A市	A市58车贷特大集资诈骗案	
8	8	2.27	2019/3/6至2019/4/12	A6区月亮岛	A6区月亮岛路高压线建议	
9	9	0.683	2019/7/7至2019/9/1	伊景园滨河苑	伊景园滨河苑捆绑车位销售	
10	10	0.3	2019/3/13至2020/1/6	A3区保利麓谷林语桐梓坡路与麓松路交汇处, A市丽发新城, A市赤岗岭地铁站, 西地省人民医院, A5区万科金域华府别墅区, A4区天健盛世A1区工地等	小区工地施工扰民	

图 16. 热点问题表截图

考虑到热点问题和时间的关系，我们提取排名前五的热度问题 做了一个主题河流图，来更加直观的查看热点问题随时间的变化，图中宽度为热度，横坐标为时间：

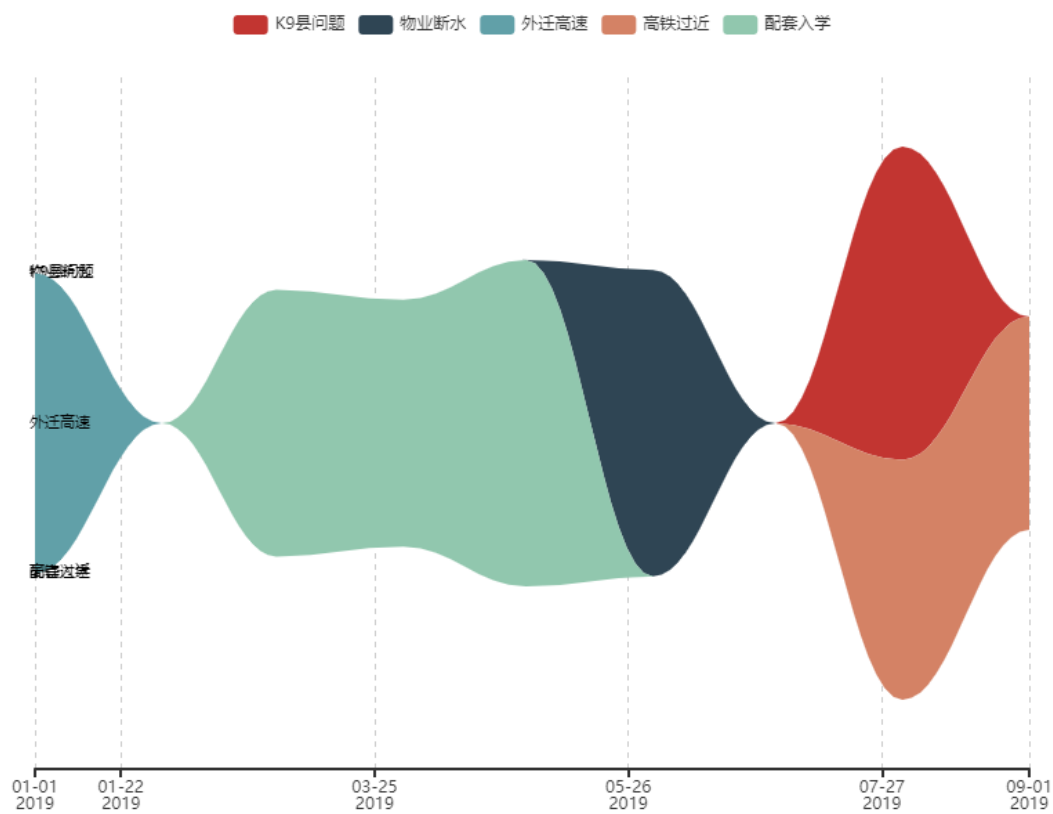


图 17. 热度前五话题河流图

3.6 匹配留言

用关键词匹配算法对排名前十的话题进行匹配，找到相关的留言数据，汇集成留言明细表。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	208636	A00077171	汇金路五矿万境K9县存在一	2019/8/19 11:34:04	曾发生过狗咬人，请问有人对养宠物	0	2097
2	193091	A00097965	绿物业丽发新城强行断业主	2019/6/19 23:28	物业只提供地摊上买的收据，对于	0	242
3	284571	A00074795	省尽快外迁京港澳高速城区	2019/1/10 15:01:26	连接着g4、长浏高速出口，进出车辆	0	80
4	191951	A00041448	地海外滩小区距渝长厦高铁	2019/8/23 14:21:38	是轻轨也不是火车，这是高铁，每天	0	1
4	202575	A00092007	市绿地海外滩二期与长慧高	2019/9/4 18:32:42	约定9月底交房入住。在城镇化高歌	0	17
4	246974	A00082652	长慧高铁征地路线对A6区周边	2019/8/26 17:40:23	也会按同样的路线导入A市高铁西站	0	0
4	263672	A00041448	小区距长慧高铁最近只有30	2019/9/5 13:06:55	希望能回复到我如下问题：1、关于	0	669
5	222919	A0001719	七三角洲幼儿园入园难、小学	2019/2/15 10:24:04	、海，安全隐患大。这都与配套规划	0	1
5	223297	A00087522	映A市金毛湾配套入学的问题	2019/4/11 21:02:44	月31日，A市教育局暂未将金毛湾楼	5	1762
5	224997	A00011154	区中建A1区嘉苑的业主小孩	2019/5/29 15:39:45	位完全不能满足《A市城市中小学幼	0	11
5	247839	A00051965	办调解决旭辉御府业主子女入	2019/3/12 11:19:22	旭辉御府业主配套入学资格是因为旭	0	0
5	253314	A00089954	A5区万科魅力之城小孩入学	2019/4/17 23:59:08	。17年底买房时告知我们，凭购房	0	0
5	255733	A00061749	晖御府业主子女的配套入学资	2019/3/12 3:18:35	大民生事宜必须公开公正！试问长郡	0	0
6	233542	A00080329	东六线以西泉塘昌和商业中心	2019/1/2 20:27:26	恒天九五重工原厂房和闲置的西地省	0	24
6	239670	A00080329	东六线以西泉塘昌和商业中心	2019/1/11 15:46:04	地省达源置业有限公司厂房，有什	0	41
6	256358	A00080329	东六线以西泉塘昌和商业中心	2019/1/2 20:27:07	恒天九五重工原厂房和闲置的西地省	0	29
7	214238	A0006178	安派出所对58车贷一案办案	2019/1/20 22:28:40	的。因此我们想问一下：第一，58车	1	2
7	217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	苏纳和小股东、苏纳弟弟苏吕是挂名	0	790
7	218132	A00010609	请求过问A市58车贷案件进展	2019/1/29 19:15:49	把此信息告知办案警官毛浚时，他说	0	0
7	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	切盼望有案情消息总是失望，四处诉	0	821
7	223787	A00034861	案件创造全国典型诈骗案，过	2019/1/11 21:12:34	出的相关资产4任由犯罪嫌疑人上蹿	0	0
7	226265	A00010644	正办理58车贷案件，还我们	2019/5/28 15:08:51	相信经侦，但这样的经侦我们怎么相	0	3
7	234320	A00010659	让A市因为58车贷案件而臭名	2019/7/8 17:16:57	从通报的内容看还是一如既往的敷衍	0	0
7	240554	A00029163	车贷老板跑路美国，经侦拖	2019/2/10 20:58:40	容嫌犯。这是涉嫌保护伞直观表现。	0	6
7	254532	A00010606	性退出立案近半年没有发过	2019/1/14 22:08:20	犯罪。于是，在58官网上挂出选举	0	3
7	264119	A00084445	个月过去，A4区公安分局未	2019/1/19 9:47:23	由于未控制人员，未查封资产，未查	0	0
7	268251	A00010609	立案近半年毫无进展，单位	2019/2/2 15:03:05	控制平台高管和资产，这种情况，	0	25
7	272413	A00010606	车贷恶性退出，A4区立案已	2019/1/14 20:23:57	借款人极为不满但我们出借人又没有任	0	2

图 18. 留言明细表部分截图

四、答复意见评价方案

4.1 数据的观察与处理

通过观察发现附件 4 的答复意见这一列中存在冗余内容，真正的答复都在“答复如下:”的后面，并且每一个数据项的末尾都会有类似“感谢您对我们工作的关心、监督与支持。”字样的话语，所以针对以上数据出现的问题并结合常用的数据处理方法，我们将数据进行分割，提取出真正的回复内容，具体方法如下：

设置两个数组 a 和 b，对答复意见的每一行数据进行遍历，如果出现“:”就将对应的 index 放入 a 中，如果出现“。”或者“!”，就将对应的 index 放入到数组 b 中，遍历完成之后对原始的答复意见进行分片处理，范围是 a 的最后一个值+1 到 b 倒数第二个值，即 $a[-1]+1 : b[-2]$

原始数据如下：

您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“42区景蓉苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉苑业委会于2019年4月10日至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真梳理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5月5日下午，辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019年5月9日

网友“A00023583”：您好！针对您反映A3区潇湘南路洋湖段怎么还没修好的问题，A3区洋湖街道高度重视，立即组织精干力量调查处理，现回复如下：您反映的为潇湘大道西线道路工程项目，该项目位于坪塘老集镇，目前正在进行土方及排水施工。因该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大，周期较长。加之坪塘集镇原有管线、排水渠道较多，需先处理管线和渠道才能进行道路施工，且因近期持续雨天，为保证道路施工质量，需在晴好天气才能正常施工。目前该项目已完成75土方及50排水，预计今年8月底将完工通车。感谢您对我们工作的关心、监督与支持。2019年4月29日

图 19. 答复意见原始数据部分截图

处理后数据如下：

经调查了解，针对来信所反映的“小区停车收费问题”，景蓉苑业委会于2019年4月10日至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真梳理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5月5日下午，辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序

您反映的为潇湘大道西线道路工程项目，该项目位于坪塘老集镇，目前正在进行土方及排水施工。因该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大，周期较长。加之坪塘集镇原有管线、排水渠道较多，需先处理管线和渠道才能进行道路施工，且因近期持续雨天，为保证道路施工质量，需在晴好天气才能正常施工。目前该项目已完成75土方及50排水，预计今年8月底将完工通车

图 20. 答复意见处理后数据部分截图

4.2 设计评价方案

对于这个问题，我们主要从四个方面来考虑：相关性，完整性，实时性，可解释性。

首先说相关性，简单来讲就是进行文本相似度计算，通过比较答复意见和留言详情的文本相似度，来判断答复的内容是否和留言详情相关；接着说完整性，完整性我们认为就是来判断对于留言详情中所提出的问题，在答复意见中是否都给出了回答，即反过来匹配相关问题是否都答复了；然后是实时性，所谓实时性就是看针对留言详情中提出的问题答复的是否及时，是否能够在其实就是合理性解决的是否恰当；最后是可解释性，我们认为可解释性就相当于合理性，来判断解决的问题是否恰当。

四个方面，我们认为不满足任何一条都认为答复意见是不合理的，针对这四个方面我们设计了一套评价方案，如下：

通过 TF-IDF 算法对留言详情和处理后的答复意见进行关键词提取，然后将提取出来的关键词使用余弦相似度的方法进行文本相似度匹配，设置一个阈值，如过余弦值大于这个阈值，就说明留言性情和答复意见是相关的，如果小于这个阈值就证明留言详情和答复意见是不相关的，同时也说明答复意见是不合理的。然后用答复时间-留言时间得

到间隔时间（只精确到月），将所有的间隔时间加和求平均，将这个均值设置为阈值（也可以通过其他方法设置，合理即可），如果间隔时间大于这个阈值，则认为答复意见不满足实时性；针对事件的合理性，我们采用的解决方法是将每个答复时间都作为一个节点，每一次都对留言时间大于这个答复时间的留言详情同对应这个答复时间的答复意见进行文本相似度匹配，如果匹配的结果都小于前面设定的阈值，则证明这个答复意见是合理的，因为这样保证了在这个时间节点之后并没有人在对相同的问题进行留言，说明人们对这个答复意见是满意的。

五、总结与展望

5.1 总结

为了帮助政府部门更好的了解民情，体会民意，同时减轻工作人员的非必要的作业负担。我们首先建立了基于 Bert 模型改进的中文文本分类器，对留言实现自动化分类别。

针对第二题的热点问题，我们通过自定义热度：

$$\text{热度} = \frac{\text{传受众特征热度影响力}}{\text{话题持续时间}}$$

按照先 k-means 聚类，再密度聚类优势互补实现了热点问题的挖掘，并且通过文本的相似匹配回溯整个事件的记录。最后，我们使用文本匹配，对文本的及时性、合理性等处理，提出了针对政府回复的建议，基本实现本赛题设计的目标。

5.2 展望

5.2.1 基于 Bert 模型分类器的改进

Bert 可以说是当前 NLP 方面的万能机器，而且在各大公司的实践当中，Bert 模型确实在各个任务都有所提升，但是 Bert 模型在文本分类的方面的提升程度并不高，不

过 Bert 模型在词性标注方面效果提升比较高，所以在第一题上我们更好的实现方式是使用 Bert 模型进行词性标注，因为文本分类的关键在于名词性词语的差距，我们希望通过以此来达到更好的提取特征，提高正确率的效果。

在数据处理的时候，我们还发现给出的分类数据有明显的分类错误的现象，对此，我们的做法是不做处理，因为如果人工的对所有的训练数据集进行纠错处理，违背了我们这次竞赛的初衷，其次，在实际的工作环境中，出现训练集标注错误是一个正常的现象。所以综合考虑我们还是不做处理，直接使用了训练集。但是如果可以定位到错误的记录，然后通过训练的模型对数据进行更正，得到新的训练数据集，重新加强对模型的训练，以达到更好的效果。

其次不同类别的数据数目不同，甚至差异明显，因而造成训练模型的时候，各种类型的训练不是均衡的，可能会影响最终结果的准确率等。因此应该想办法在保证差距的同时尽量的缩小因为数据的不均衡带来的影响。

5.2.2 热度定义的改进

针对第二题，我们组认为这是一个非常开放的题目，但是我们认为我们的热度定义有点简单，而且存在一定的问题。按照添加数据的方式来处理点赞和反对虽然实现了同时处理同一类文本的数量和文本的点赞与反对，但是可能会造成因为数值类型的限制，导致得到的数据不是那么精准，比如如果点赞是 1 反对是 0，则添加记录应为 0.6 条，在实际操作中视为不添加。在实际生活中，我们往往也无法忽略一个因素对于热点新闻的影响，这个因素就是关键人物和水军，在现在的网络环境，网络知名人物的行为，往往很容易煽动网民的行为，因此在实际生活中，我们必须对这种网络上的关键人物赋予更高的处理权重，相反水军应该被屏蔽，或者设置低权重。其次在本次的热度定义时，我们很难对自己的热度定义是否合理进行检验，所以如果这个题目可以给出一些已经发掘的热点问题及其排名，并且给出相应的数据的话，我们可以通过因子分析的方法，找到影响热度定义的真正的隐性因素，但是这个题目比较开放，所以我们的处理方式也是很个性化的，权重的赋值方面往往是题目与现实之间磨合，而且本次数据提供的维度方面相对而言还是比较少的，因此我们很难定义出来关键人物等因素，我们本来尝试通过已有的数据得到关键人物等因素，但是发现数据不齐全，很难得到预期的效果，而且很

容易适得其反，其次，本次的文本时间跨度长，因此相对而言文本在时间的跨度方面就显得密度很低，这对热度的定义也带来一定的问题，考虑时间很可能会掩盖一些时间跨度比较长但是确实是一直比较热点的信息，不考虑的话，与实际生活的常识又不符合，在考虑第一题的方面，也应该对不同类的事件赋予不同的权重来进行热度的比较，现实中对于政府部门往往民生的问题比娱乐的问题更加容易受到关注。

热点，是实时性非常强的一个名词，所以热点一定要与一定的时间点来绑定，站在今天说两三年之前的热门新闻，就不能称之为热点了，在这里我们认为还可以引进牛顿冷却定律来对热度定义，正如温度随时间改变一样，热度也会改变。

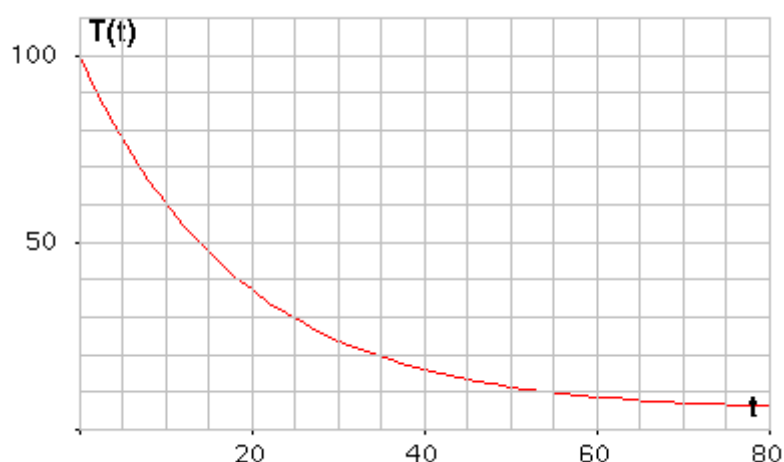


图 21. 热度衰减图

基于牛顿冷却定律可以得到：

$$\text{本期温度} = \text{上一期温度} \times \exp(-(\text{冷却系数}) \times \text{间隔的小时数})$$

冷却系数需要自己来决定，冷却系数的确定与热点新闻的时效性强弱有关，但是在此处要注意，不同类的热点事件的冷却系数应该是不同的，而不是统一的，从而达到更好的效果。

5.2.3 对留言回复意见的改进

留言的回复，我们主要从四个方面来考虑：相关性，完整性，实时性，可解释性。但是在实际的实现过程中，可解释性，完整性的判断结果可能并不如人为主观判断更好，而且可解释性等非常依赖参数的设置，而且针对实时性方面我们采用的是超过某一个时间间隔，即视为不及时，但是实际情况中，可能不同的问题类需要不同的时间阈值，甚

至不同的具体留言问题需要不同的阈值。

参考文献

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/pdf/1810.04805.pdf>
- [2] 何跃, 帅马恋, 冯韵. 中文微博热点话题挖掘研究. <https://wenku.baidu.com/view/8c94f12c6ad97f192279168884868762caae8b84.html?fr=search,2013-11-18/2020-05-01>
- [3] 梁昌明, 李东强. 基于新浪热门平台的微博热度评价指标体系实证研究. <https://www.docin.com/p-1687586794.html, 2015-08-18/2020-05-02>
- [4] <https://wenku.baidu.com/view/289168876137ee06eff918da.html>
- [5] 王玉洁. K-means 算法的改进及其在文本数据聚类中的应用[D]. 西安科技大学, 2016.
- [6] 于宽. 改进 K-Means 算法在文本聚类中的应用[D]. 大连交通大学, 2006.
- [7] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述. <https://wenku.baidu.com/view/7789d35710661ed9ac51f30c.html, 2009-04-03/2020-05-03>
- [8] 阮一峰. 基于用户投票的排名算法(四): 牛顿冷却定律. http://www.ruanyifeng.com/blog/2012/03/ranking_algorithm_newton_s_law_of_cooling.html