C 题：“智慧政务”中的文本挖掘应用

目录

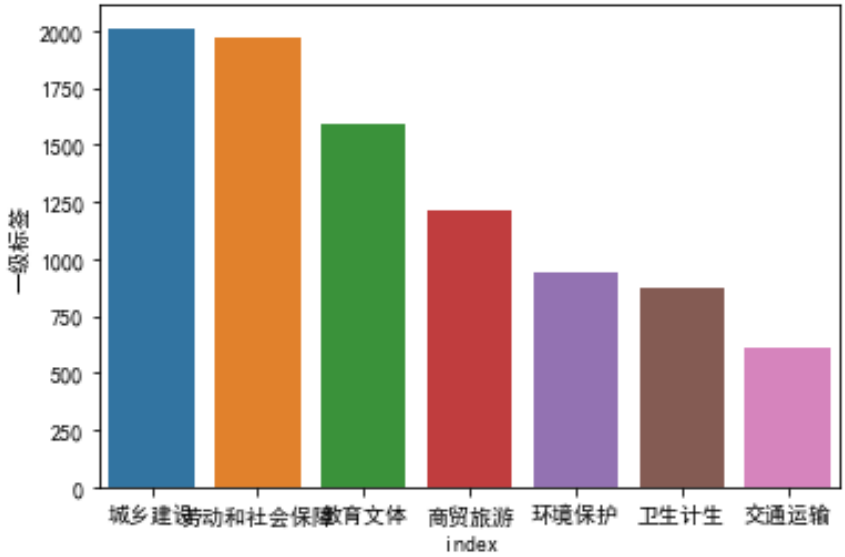
一、 基于改进 TF-IDF 的卷积神经网络留言分类

在第一部分留言分类中，我们根据 TF-ID 算法存在的问题进行改进，并使用改进的 TF-IDF 算法提取各个类别特征词，并以这些特征词来对数据进行增强，以保证数据的均衡性。此外，我们根据 TF-IDF 计算出的每一留言中每一 TF-IDF 的值，对每一个句子所有的词加和取平均，并排序，选择最有代表性，最重要的句子输入到 TextCNN 网络中进行拟合。最后得到，分类的 F1-score 为 0.918。

1.数据预处理

1.1 数据描述

首先对留言数据类别的分布进行观察，绘制柱形图如下图，其中，柱形图的高度代表着对应类别的个例数目。



由上图我们不难看出，留言数据类别的分布不均衡，即城乡建设和劳动和社会保障占大多数，而属于交通运输和卫生计生的、环境保护的个例数目较少，所以，在进行数据处理和模型拟合的过程中，要考虑数据不均衡的问题。

1.2 分词

中文分词指的是将一个汉字字序切分成一个个独立的词。我们知道哎英文中单词之间是以空格作为自然分解符的，而中文只是字，句段能通过明星的分解符来简单划界，唯独词没有一个形式上的分界符。中文分词不仅是中文文本分类的一大问题，也是中文自然语言处理的核心问题之一。

在这里我们使用 jieba 分词，jieba 小巧而且高效，是专门使用 Python 语言开发的分词系统，占用资源较小。分词后，对分词后的数据去除中文常见的停用词，在这里我们使用的是百度公开的停用词表。

2.数据向量化

目标：传统的机器学习文本分类算法把所有的分词后的单词作为特征，统计词频或 TF-IDF 作数据特征，进行分类。而我们认为，使用全部词语可能会因为某些词语出现频次较低或训练数据和实际数据词汇的分布不同，而易导致模型的过拟合、鲁棒性差。

进而，我们提出一种，基于类别关键词的特征构建方法。首先，我们使用改进后的 TF-IDF 提取出每一留言分类中的关键词，并以这些关键词为特征，构造数据，进行模型拟合。另一方面提取类别的关键词减少词表，可以进一步减小嵌入矩阵的大小

2.1 基于改进 TF-IDF 的关键词提取

2.1.1 TF-IDF

在信息检索中，TF-IDF（词频-逆文档频率）是一种统计方法，用以评估一个单词在一个文档集合或语料库中的重要程度。经常被用作信息检索、文本挖掘以及用户模型的权重因素。TF-IDF 的值会随着单词在文档中出现的次数的增加而增大，也会随着单词在语料库中出现的次数的增多而减小。TF-IDF 是如今最流行的词频加权方案之一。

TF-IDF 的各种改进版本经常被搜索引擎用作在给定用户查询时对文档的相关性进行评分和排序的主要工具。TF-IDF 可以成功地用于各种主题字段的停用词过滤，包括文本摘要和分类。

词频（TF）表示为一个给定词语 t 在一篇给定文档 d 中出现的频率，即对于在某一文档 d_j 里的词语 t_i 来说，t_i 的词频可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中，$n_{i,j}$ 为词语 t_i 在文档 d_j 中的出现次数，分母为该文档中所有的词语数量。

逆向文件频率（IDF）是词语对全部文档的重要性的指标，可以由总文件数除以包含该词语的文件数，再将得到的商取对数得到：

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} + 1$$

其中 |D| 是语料库中所有文档总数，分母是包含词语 t_i 的所有文档数。

根据 TF 和 IDF 值，便可以得到，每一文档中每一词语的 TF-IDF 值：

$$value_{i,j} = tf_{i,j} * idf_i$$

2.1.2 特征词提取

我们的目标是提取出每一类别中的关键词，而 TF-IDF 是针对每一文档计算的。因此，我们把所有的用户留言数据按类别分组，共得到 7 篇文档，然后，就每一文档分词，去除停用词，并计算 TF-IDF；然后对每一文档中的 TF-IDF 值排序，取出前 15%的词语作为该类别的关键词，最后得到的结果如下（以交通运输类别为例）：

根据上图词云，我们可以看出，TF-IDF 可以把代表该类别的一些关键词提取出来，比如“出租车”，“司机”，“公交”等，但是同时也存在一些问题，比如关键词中同时出现了一些在其他类别同样广泛存在的词语，比如“我们”，“一个”，“城市”等，不仅在交通运输类中，在其他类别中也存在由类似的情况。

我们分析原因后认为，可能是由于这些词语在文本中的词频过大，导致计算的 TF 值较大，且远大于其他词语的 TF 值，且由于我们的文档数量只有 7 篇，IDF 对词汇的重要性计算的修正值可能存在不足。最大的 IDF 的权重为 log(7/1)+1，约等于 2.9。最小为 1，故而，如果，词语的 TF 值较大，最后得出的 TF-IDF 仍较大。

此外，如果某些词语能鲜明地使得某个样本标记为某类别，但是这些词语仅仅存在于比较少的样本中，那么这些词语计算出的 TF-IDF

值可能较小，以至于不背加入到词表中。

因此，针对以上问题，我们需要对 TF-IDF 改进。

·2.1.3 TF-IDF 改进

针对上诉问题，我们设计了一个修正系数：

$$w_{i,j} = e^{k(n_{i,j}-\bar{n})}$$

其中：$w_{i,j}$为 d_j 篇文档中词语 t_i 的修正系数，k 为一个超参数系数，$n_{i,j}$为词语 t_i 在文档 d_j 中的出现次数，$\bar{n}$为该词语在所有文档中的平均频数。且 $n_{i,j}$和$\bar{n}$ 均是经过便准化后的次数于平均频数。标准化过程如下：

$$n_{i,j} = \frac{r\_n_{i,j}}{l_j}$$

其中，$r\_n_{i,j}$，为文档为文档中真实的权重，l_j为文档 j 的词语数量。$\bar{n}$可由经过标准化的$n_{i,j}$ 得到。此外，经过标准化后的数据可减少因数据不均衡或文档数量、长度不同对词频的影响。

在原始 TF-IDF 的基础上，乘以我们的权重，便可得到改进后的 TF-IDF 值：

即

$$value_{i,j} = w_{ij} * tf_{i,j} * idf_i$$

得到结果如下图

由上图我们可以看到，改进后的 TF-IDF 不仅去掉了类似于“我们”，“一个”之类的词汇，而且提取出了“的士”“线路”等比较重要，但可能词频较小的词汇。综上，我们改进的 TF-IDF 可以增强其提取关键词的能力

2.2 数据向量化

2.2.1 计算子句权重，并排序

然后使用我们改进后的 TF-IDF 值，以分词、取出停用词后的留言详情为数据，重新计算每一留言中所有词的 TF-IDF 值。并根据算出 TF-IDF 值，计算留言中每一分句的平均 TF-IDF，得到留言中每一分句的相对重要程度，并以此降序排序。经过这样操作后，留言中含有较

多重要词汇的句子，将会排在前面。

此外，我们认为，留言主题中包含较多的重要信息，因此，我们将每一条记录中的留言主题放置到所有排序后留言详情数据的最前面。

2.2.2 数据均衡化

根据我们最开始对原始数据类别分布的观察发现数据存在较严重的数据不均衡情况,因而,在这一部分,主要进行数据的均衡化处理。

根据我们在 2.1.3 中提取出的每一类别的关键词，还以交通运输为例，我们随机挑选交通运输中的 80-200 个关键词，作为一条扩充数据，依次类推，我们分别扩充交通运输，卫生计生，商贸旅游，教育文体，环境保护中的数据到 1500 例。并对

城乡建设类别数据欠采样，保存 1500 例个例。这样，每一类别数据个例就就都在 1500 左右。

2.2.3 数据向量化

在本节,将对按重要程度排序和均衡化后的留言数据进行向量化。

a. 留言数据截取

首先，我们需要查看每一留言的大致长度，结果如下表，发现，90%的留言数据大于 259。因此我们选择留言的长度为 260，即只截取留言数据的前 260 个数据，不足的以“PAD”填充。

指标	长度

mean	110
std	142
min	3
25%	33
50%	60
75%	128
90%	259
91%	276
93%	317
95%	376
97%	491
99%	729
max	3600

文本长度

b. 转换为数字特征

然后，根据截取后数据，统计词频，选取词频较多的词语，构造词汇字典，给每一词语进行编号，由文本数据转换为数字特征。

c. 转换为词向量

此外，在本样例中，我们使用预训练词向量来将数字特征转换为向量特征。预训练词向量使用由 Sogou News 搜狗新闻训练的词向量。

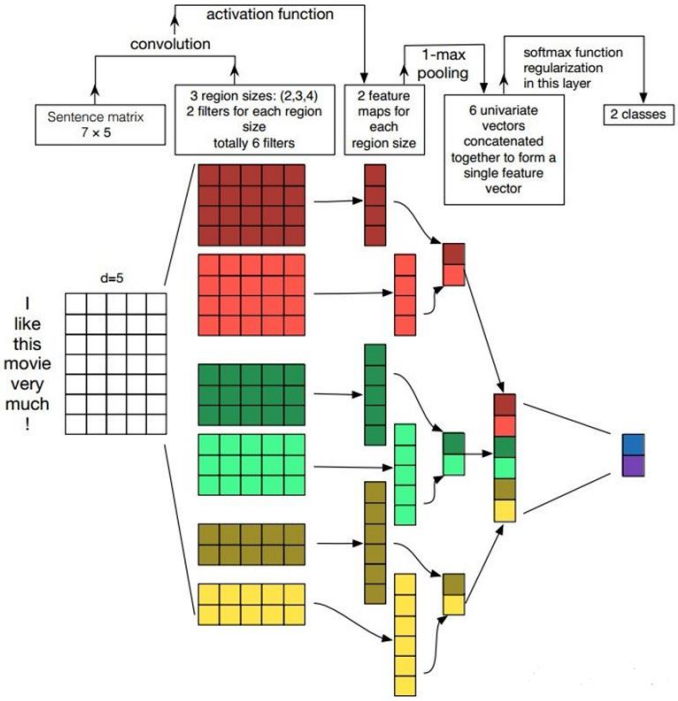
d. 数据集划分

最后将数据按 7：3 划分训练集与测试集。

3 模型拟合

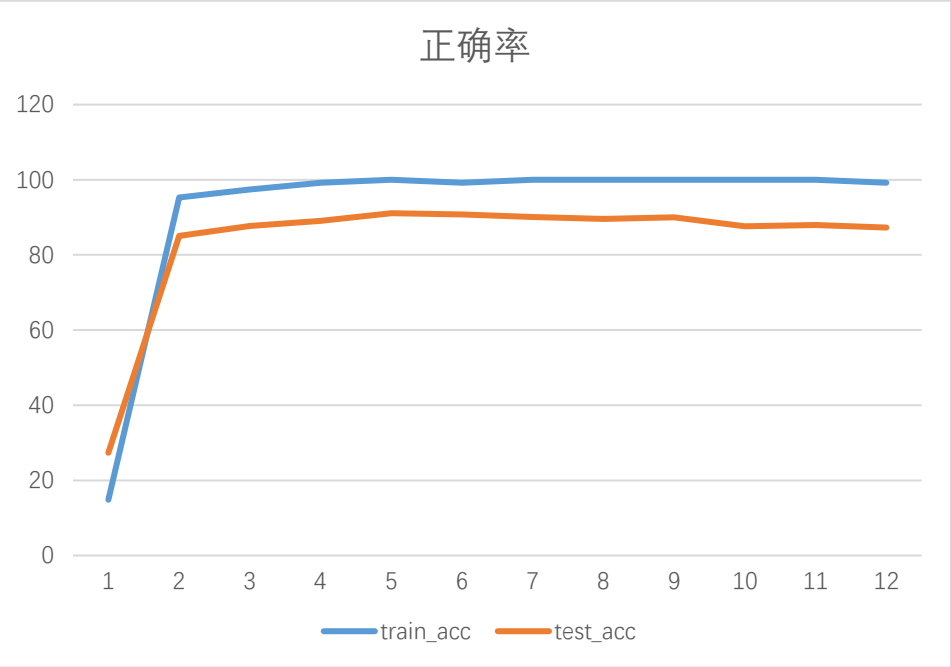
3.1 模型结构:

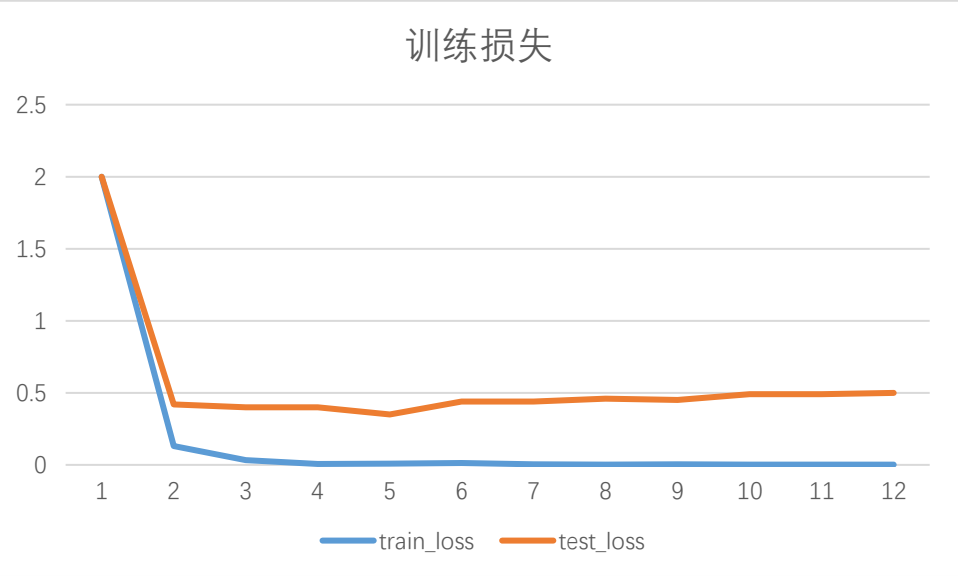
在本题中，我们使用 TextCNN 作为主要的分类模型，来对数据进行拟合、分类。TextCNN 的结构如下：



3.2 拟合过程

TextCNN 网络训练过程中的损失和正确率变化情况变化如下表。由于数据量相对来所比较小，故而模型在第 500batch 时就已经拟合，且随着训练的继续，开始出现过拟合的情况。





训练过程损失及正确率变化情况

batch	train_loss	train_acc	test_loss	test_acc
100	2	14.84	2	27.36
200	0.13	95.31	0.42	85.05
300	0.033	97.42	0.4	87.68
400	0.0067	99.2	0.4	89.1

500	0.0096	100	0.35	92.1
600	0.013	99.2	0.44	90.8
700	0.004	100	0.44	90.1
800	0.0028	100	0.46	89.61
900	0.0032	100	0.45	89.98
1000	0.0018	100	0.49	87.63
1100	0.0023	100	0.49	88
1200	0.0017	99.22	0.5	87.31

3.3 结果分析

我们除了使用 TextCNN 对数据进行拟合外，还使用 2.2.3.b 中的数字特征分别输入到 SVM，逻辑回归、朴素贝叶斯、随机森林中进行拟合，得到最终的结果如下。

模型	F1-score
TextCNN	0.918
SVM	0.887
逻辑回归	0.896
朴素贝叶斯	0.864
随机森林	0.857

根据上表的结果我们可知，数据在 TextCNN 模型上拥有更好的结果，最终的 F1 值为 0.918

4.模型不足和改进

（1）可以多尝试其他神经网络如 TextRCNN, TextRNN, FastText, TextRNN_Att, DPCNN, Transformer 等，并进行模型融合。

（2）可以尝试就留言主题进行以字为单位神经网络模型训练

二、热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题,如“XXX 小区多位业主多次反映 入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关 部门进行有针对性地处理，提升服务效率。请根据附件 3 将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出 排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题 对应的留言信息，并保存为“热点问题留言明细表.xls”

1.数据描述

在附件三中，共有 4327 条数据，每一条数据包括的字段有：留言编号、留言用户、留言主题、留言时间、留言详情、反对数和点赞数。数据的具体内容如下：

留言编号	留言用户	留言主题	留言时间
188006	A000102948	A3 区一米阳光婚纱艺术摄影是否合法纳税了？	2019/2/28 11:25:05

留言详情	反对数	点赞数
座落在 A 市 A3 区联丰路米兰春天 G2 栋 320，一家名叫一米阳光婚纱艺术摄影的影楼,据说年单这一个工作室营业额就上百万,因为地处居民楼内部，而且有蛮长的时间了，请税务局和工商局查一下,看看这个一米阳光有没有正常纳税！如果没有，应该会	0	0

怎么操作！

留言编号：

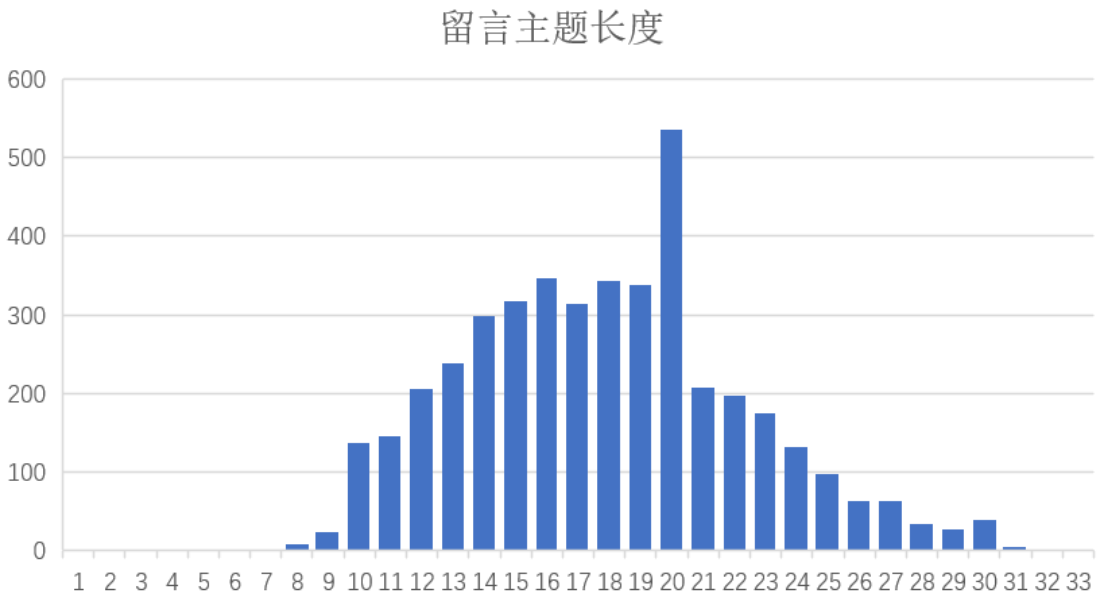
在所给数据中，留言编号由六位数字给出，作为每条数据的主键，起到辨识和唯一区分的作用，在对 NLP 任务的处理结束之后，可以通过留言编号对数据进行辨识。

留言用户：

留言用户通过“字母+数字”的形式给出，用于唯一辨识用户。我们认为留言用户的编号对文本聚类的帮助不大，而在构建热度指标时考虑将重复发言的留言文本进行处理，可以对热度指标构建起到作用。

留言主题：

留言主题是一句话，根据用户都不同，可能有陈述句和疑问句两种形式，留言主题的长度如下图所示：



上图中，横坐标为留言主题长度数值，纵坐标为留言主题长度

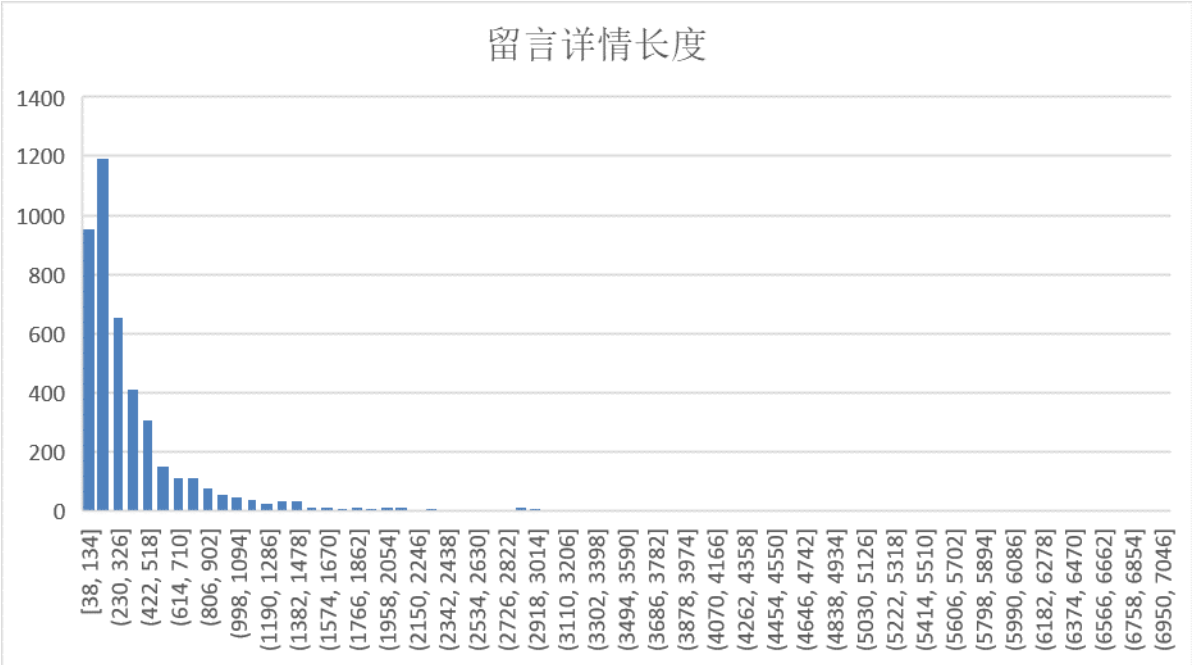
计数。由图可知，留言主题长度大致分布在(10，30)的区间上，有极少部分在长度区间之外，长度以(15，20)最为集中。由此可见，留言主题具有长度较小的特点，同时考虑到留言主题的表达较为凝练，主题性更加明确，因此我们认为留言主题可以作为文本聚类的重要依据。

留言时间：

留言时间精确到秒，格式为“年/月/日+小时/分钟/秒”，留言时间对于文本的聚类分析的作用不大，但可以作为热度指标的组成部分。

留言详情：

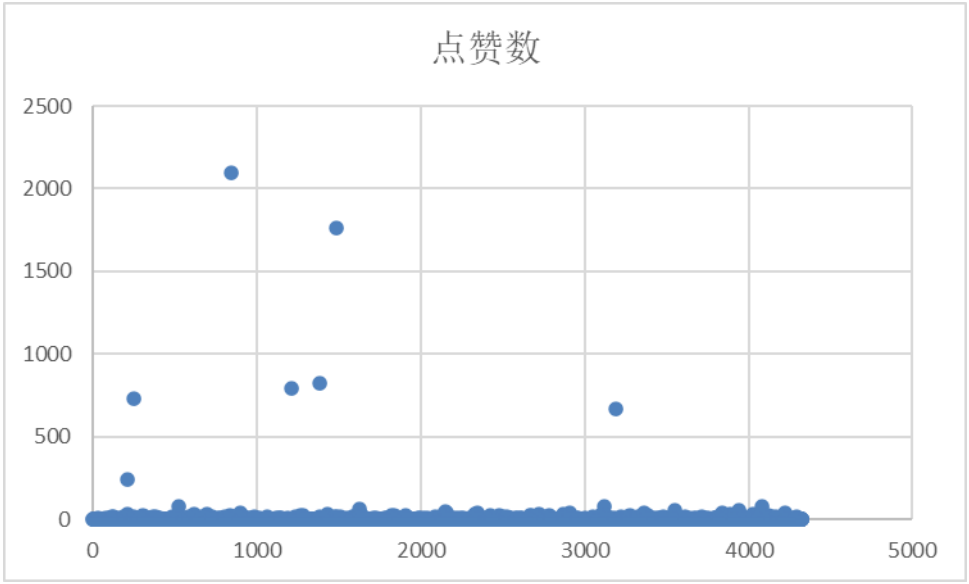
相较于留言主题而言，留言详情属于长文本，长度集中在(40，500)区间，过长的文本会带来维度和算法效率上的问题，因此文本聚类主要使用留言主题进行。

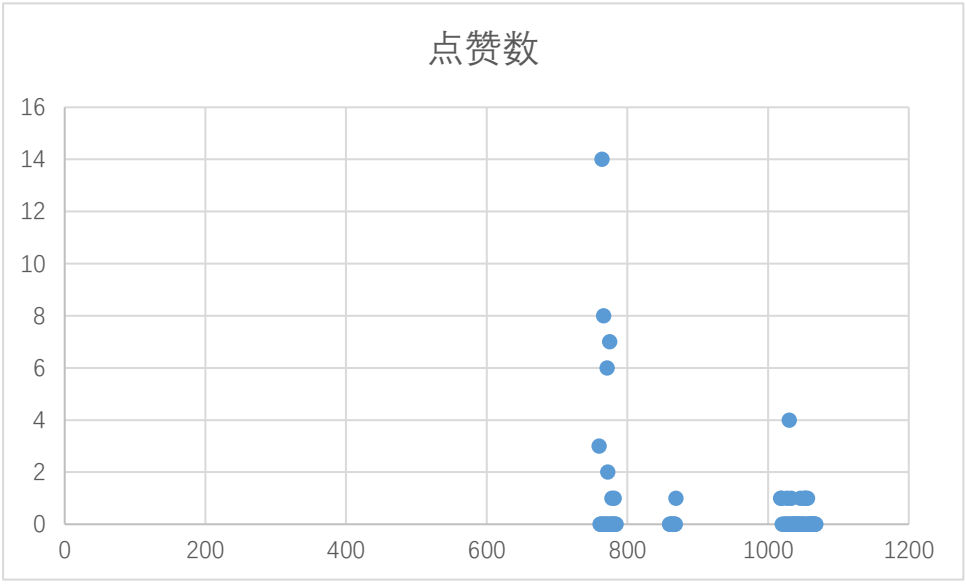


反对数和点赞数：

总体而言，这一指标数值都很小，但如下图所示，也存在少量

评论获得大量的点赞数和反对数的情况，因此在构建热度指标时也必须将这个因素考虑进去。





2. 文本的聚类分析

数据预处理

对于附件三的留言主题这一类的短文本，我们的预处理思路包括清洗、分词、去除停用词和赋权四个步骤。

文本清洗：

清洗操作的主要思路是删去无意义、重复的字段并且删除或补充含有空值的字段。对于每个字段来说，含有留言主题表达不明确的字段可以使用留言内容进行补充。

例如留言编号 189856 的字段，留言者 A00073717 输入的主题为“)”，留言内容为“A 市保利麓谷林语小区，是 A 市的大型小区，居住人口有一万多人。地下一层墙体严重开裂，并且地下水和泥沙大量溢出，特别是逢下雨天 严重，小区还算比较新，才建几年，楼高 33 楼，我们让专业的人员看过，不仅负一楼停车场经常被“淹”，并且大量泥沙漫出，而且严重影响威胁居民的生命生产安全。我们像保利物业和保利地产投诉和反应过很多次，未果。而且投诉时长可以用年来计算，保利的拖延理由是“等天晴出太阳再说”这个太阳一等就不知道是多久，最近保利更是出息了， 视居民的安全不顾，选择将裂缝“盲视”，在楼体裂缝处的表面全部“镀”上了钢板来装饰。（划重点）小区扶手全部锈掉去年发生了重大事故小孩靠着扶手休息就坠楼死亡，路灯被风一吹就，就把路过的老人砸倒，诸多问题多次引起了西地省台的新闻报道，但是小区毫无半点整改，和保利总部多次反应也未果，地方各部门也并且到处踢皮球。小区想成立业委会，A 市保利物业多翻阻拦，简直成了纵横的地霸。”我们将内容概括，给出他的留言主题“A 市保利麓谷林语小区地下一层墙体开裂”。

中文分词

对问题字段进行处理后，我们对短文本进行分词，通过

python3.0 采用中文分词工具 jieba 分词进行处理，具体的代码在“get_in.py”中给出。

如下表所示，分词后，短文本被转换成了多个词语。

分词前	分词后
A3 区一米阳光婚纱艺术摄影是否合法纳税了？	['A3', ‘区’, '一米阳光', '婚纱', '艺术摄影', ‘是否’, '合法', '纳税', ‘了’, ‘?’]
咨询A6区道路命名规划初步成果公示和城乡门牌问题	[‘咨询’, 'A6', ‘区’, '道路', '命名', '规划', '初步', '成果', '公示', ‘和’, '城乡', '门牌', ‘问题’]

去除停用词

在文本中，某些词对文本的主题没有贡献，只起到修饰性作用或辅助性作用，包括一些口语词、量词、形容词、副词和名词。因此，我们在分词后，使用“cn_stopwords.txt”中的停用词表，对分词结果做进一步清洗，即去除停用词。此外，我们还添加了一些与政务留言类型更相关但“cn_stopwords.txt”没有的词语，

包括“加快”、“请加快”、“导致”等。分词代码同样在“get_in.py”中，去除停用词后，分词结果如下所示

停用前	停用后
['A3', ‘区’, '一米阳光', '婚纱', '艺术摄影', ‘是否’, '合法', '纳税', ‘了’, ‘?’]	['A3', '一米阳光', '婚纱', '艺术摄影', '合法', '纳税']
[‘咨询’, 'A6', ‘区’, '道路', '命名', '规划', '初步', '成果', '公示', '城乡', '和', '门牌', '问题’]	['A6', '道路', '命名', '规划', '成果', '公示', '城乡', '门牌']

赋权：

根据词频，使用 TF-IDF 赋权并生成词向量，代码同样在“get_in.py”中，输出结果是词向量的权值矩阵。

聚类分析

经过向量化的文本具有维度 7745 维，维度过高对于不同聚类算法的表现会产生影响。我们选用 BIRCH 层次聚类算法，结合 PCA 降维，通过调整 BIRCH 中阈值和类别数两个超参数，进行聚类操作，具体代码在“first_birch.py”中给出。经过调试，得出 k=600，

阈值为 0.5046875 时，效果最好。

3. 热度指标构建

3.1.讨论密度

讨论密度指的是在第一条相关评论到最后一条相关评论的时间跨度内,考虑讨论次数的影响，得出的评价指标：

$$P = 1 - \frac{T\_sig}{C}$$

其中，P 是讨论密度，T_sig 是归一化时间跨度，C 是讨论次数。该指标的大小在（-1，1）之间。

3.2.反响程度

反向程度由赞同数和反对数形成，记为 F，表示为：

$$F = \frac{赞同数 - 反对数}{赞同数 + 反对数 + 0.0000001}/C$$

我们将赞同数看作公众对该留言的正面态度，将反对数看作负面态度，C 是讨论次数，0.0000001 是为了防止赞同数和反对数同时为零的情况。

该指标的大小在（-1，1）之间。

3.3.活跃性

活跃性 A 由时间分布的峰度给出。

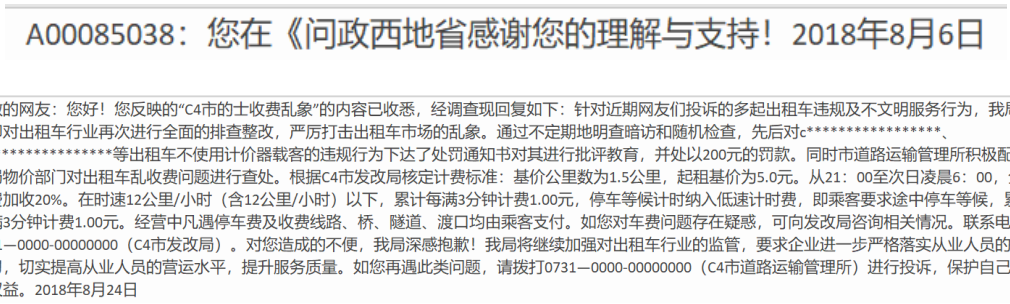
热度指标

综上所述，热度指标 R=A*P+F，结果如下

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	156	0.123922	2019/07/07-2019/09/01	伊景园滨河苑居民	伊景园滨河苑捆绑销售车位
2	548	0.089062	2018/11/15-2019/12/02	A 市引进人才	A 市人才购房补贴相关问题

3	40	0.063095	2019/01/23-2019/11/21	A 市居民	A 市地铁建造及使用问题
4	37	0.038542	2019/6/19-2020/01/09	丽发新城小区居民	A 市丽发新城社区搅拌站灰尘，噪音污染严重
5	180	0.02474	2019/01/06-2019/10/22	A 市居民	A 市社保卡等证件办理困难

三、答复意见的评价



A00085038：您在《问政西地省感谢您的理解与支持！2018年8月6日

尊敬的网友：您好！您反映的“C4市的士收费乱象”的内容已收悉，经调查现回复如下：针对近期网友们投诉的多起出租车违规及不文明服务行为，我局，立即对出租车行业再次进行全面的排查整改，严厉打击出租车市场的乱象。通过不定期地明查暗访和随机检查，先后对c**********、c**********等出租车不使用计价器载客的违规行为下达了处罚通知书对其进行批评教育，并处以200元的罚款。同时市道路运输管理所积极配合发改局物价部门对出租车乱收费问题进行查处。根据C4市发改局核定计费标准：基价公里数为1.5公里，起租基价为5.0元。从21：00至次日凌晨6：00，全程运费加收20%。在时速12公里/小时（含12公里/小时）以下，累计每满3分钟计费1.00元，停车等候计时纳入低速计时费，即乘客要求途中停车等候，累计每满3分钟计费1.00元。经营中凡遇停车费及收费线路、桥、隧道、渡口均由乘客支付。如您对车费问题存在疑惑，可向发改局咨询相关情况。联系电话：0731—0000-00000000（C4市发改局）。对您造成的不便，我局深感抱歉！我局将继续加强对出租车行业的监管，要求企业进一步严格落实从业人员的教育学习，切实提高从业人员的营运水平，提升服务质量。如您再遇此类问题，请拨打0731—0000-00000000（C4市道路运输管理所）进行投诉，保护自己的合法权益。2018年8月24日

我们以示例中的两条答复意见作为案例。

以上两条答复意见，是在附件四中真实存在的两条答复。从中我们可以看出，在庞大的数据面前，有部分的答复意见是相当完整且详细的。另一方面，也有相当大的一部分答复意见是非常简略和笼统的。对于对他们的评价，我们可以通过完整性、相关性和可解释性来进行，从而为评价体系的建立提供理论支持。

1 答复意见的完整性

答复意见来源于群众问政的留言记录。若要对答复意见的完整性进行评价的话，我们需要对其特征进行一些提炼：

1. 对群众的简单问候及感谢

在大部分答复意见中，基本上都有对于群众的问候，如“尊敬的网友”、“网友您好”等字样。交流沟通是人类行为的基础，而问好是表示友好的最初方式。在人与人交往中，问好起着重要作用，加上这一句，能够让被回应的市民感受到更好的体验。

2. 对留言提出的问题进行现状分析

对于市民提出的问题，可能很大一部分是由于既定政策和已有事实造成的。如果是由于政府的政策和规章制度造成的原因，则需要对当下的规则进行解释和说明。如果市民提出的问题是与某个公司或社区相关的，则需要对事实情况进行分析和确定。

3. 对问题进行处理并说明处理过程

网络问政平台的产生，就是为了能够有一个更直接的渠道，来解决老百姓提出的问题。对于产生的问题，意见回复中最好能给出解决方案或是政府的处理结果等信息，这样市民就能了解自己提出的问题到底是怎么被解决的。

4. 回复时间说明

在意见回复中添加回复时间的说明，能够保证意见回复的时效性。

5. 联系方式

在意见回复中留下联系方式，能够更好地与市民进行交流，拉近与老百姓之间的距离。同时，给出的联系方式也能够更有具体性，对于单件投诉的回复更加针对化，能够更好地解决相应的问题。

2 答复意见的相关性

相关性的评价，在于回复意见与市民留言的相关度是否较高。这时我们需要对两个数据的文本进行对比，找寻相关的词汇是否在两个数据中都有所体现。

举例如下，留言与回复如下所示：

在老木峪隧道分岔处，路牌和路标指示不清楚，晚上没有 路灯，一片漆黑，要不是两个施工队住在那里，就一点灯光也没有，黑的可怕，在客人的心理，就是一条乡道。老木峪隧道口五颜六色的灯光太刺眼了，路牌和路标客人一眼就能看清楚H市森林公园就是G1区源，所以每一天G1区源爆满，这就是区别。在阳和下高速以后，路上也没有我们H市国家森林公园指示牌，只有G1区源指示牌和景点牌等……麻烦书记、帮我们张家国家第一个森林公园牌子解决一下、这是整个张家村的村民心声、谢谢现在国家第一个森林公园确变成了锣鼓塔。建议大家去实地可以观察，这个才是最重要。相信有了路牌以后，就可以分清H市国家森林公园和G1区源之分。

您好，您所反映的问题，已转交相关部门调查处置尊敬的网友：你好！就您在10月24日《问政西地省》上所反映的问题，我局高度重视，迅速安排专人调查了解情况，现回复如下。一是您反映的关于老木峪隧道岔路口无路灯的问题：您反映问题时间是在10月下旬，正值路灯和标牌安装施工期间，当时部分路段的路灯还未通电。截至11月1日，G1区山大道景区段（含您在信中所提及的老木峪岔路口）的路灯安装和接线通电已全部通电亮灯。 二是您反映的森林公园标志牌的问题：按照G1区山大道的设计图纸和道路交通标志和标线国家标准及施工规范《GB-5768-2009》,G1区山大道森林公园路口的交通标志设置是完全参照以上图纸和规范设置的，一般道路标牌只会设置地名及路名，而在路口还会设置一块棕色的H市国家森林公园旅游专用指示标志牌，是符合设计、国家标准及相关规范的。目前该路段的道路标牌和H市国家森林公园旅游专用指示标牌正在安装施工中。 感谢您对我部工作的支持。 2018年11月9日

对于以上的例子，我们可以看出，留言中提到了“灯”、“指示牌”、“路牌”等具体信息，从中我们可以知道这是一条与这些元素相关的留言意见。对于这些信息，我们可以在答复意见中，能够看到“路灯”、“指示牌”等相应的信息。那么我们可以判定，这两条信息匹配度较高。与此同时，我们也可以知道，这一条答复意见的相关性较强。

另一方面，我们也可以通过地名来进行相应的匹配。如果能够对相应地区的问题进行比较高的匹配，那么说明这一条答复意见的相关性是比较高的。

3 答复意见的可解释性

可解释性，在这里指的是答复的可解释性。对于市民来说，答复意见最好比较直接，不能拐弯抹角的回复。

另一方面，在意见回复中引用相应的文件或政策时，最好不要整段整段的直接黏贴复制，这样内容过多的文本，比较冗杂，相比于直接给出意见，没有太多的影响。

引用

[1]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[2]张琳,李朝辉.文本分类中一种改进的特征项权重计算方法[J].福建师范大学学报(自然科学版),2020,36(02):49-54.