“智慧政务”中的文本挖掘应用

摘要

为了解决网络问政平台中的群众留言的分类与热点问题提取，本文由机器学习相关理论入手，分析群众留言文本的特点，基于现有的TF-IDF算法，提取出留言文本的特征向量。

针对问题1，本文根据留言文本的实际特征，对模型进行了适当的假设简化，结合现有线性支持向量机算法，使用TF-IDF算法提取出的留言文本的特征向量训练出LinearSVC分类器，在留言文本分类上取得了较好的结果。

针对问题2，本文分析了留言文本中反应问题的特征和地点与人群相关词语在文本中的分布特征，提出了基于DBSCAN算法的聚类模型和热度相关评价指标，对不同地点不同人群的留言问题进行聚类。并且，根据信息论中的信息熵相关理论，我们采用熵权法计算出热度评价指标的权重，最后通过加权求和的方式计算出留言问题的热度并进行排序，挑选出排名前五的热点问题。

针对问题3，本文参考了现有的机器翻译系统和对话系统的质量评价标准，从留言答复的及时性，内容相关性，答复有效性等几个方面提出了4条留言质量评价准则，并且仍然通过计算交叉熵的方式确定各指标权重，建立了留言答复质量评价模型

关键词：机器学习 TF-IDF算法 线性支持向量机 DBSCAN算法 熵权法

一、挖掘目标

1.1 问题背景

网络时代下市长信箱、阳光热线等各种网络问政平台成为解民意、汇聚民智、凝聚民气的重要渠道。为了对大量的社情民意相关的文本数据进行及时的处理与分析，建立基于自然语言处理技术的智慧政务系统成为社会治理创新发展的新趋势。

1.2 相关数据

给定数据附件一是详情分类的三级标签体系，其中的一级分类标签是第一问中进行文本多分类的依据；附件二给出了带有一级分类标签的留言数据；附件三给出了一些详情上有重合的、带有群众的点赞与反对意见的留言数据；附件四的数据不仅给出了留言详情，还给出了答复意见和时间。

在分析给定数据的基础上，我们根据附件一的标签体系在互联网上搜集了大量公开来源的群众问政留言记录，将在附件文件 data 中给出。

1.3 挖掘目标

（1）参考附件 1 提供的详情分类三级标签体系，建立关于留言详情的文本多分类模型，对留言数据进行一级分类，并使用 F-Score 方法对模型进行评价。

（2）根据附件 3 将某一时间段内反应特定地点或特定人群问题的留言进行归类，定义和合理的热度评价指标，并给出评价结果。按指定格式给出排名前 5 的热点问题和相应热点问题对应的留言信息。

（3）针对附件 4 中相应留言的答复意见，制定一套方案对答复意见的质量给出评价，并尝试实现。

二、总体流程分析

2.1 总体流程图

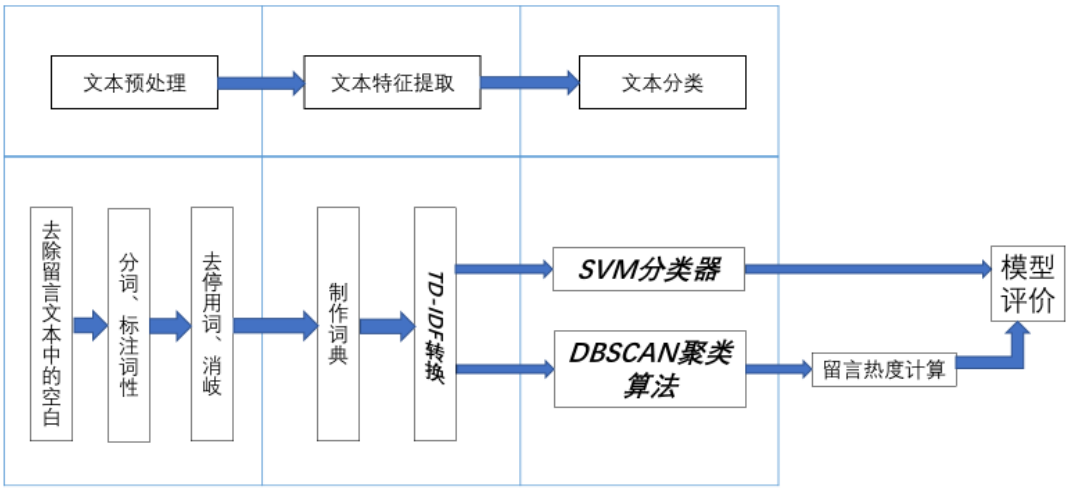

图 1：总体流程

针对问题1群众留言分类，建立文本分类模型，首先通过文献查阅和学习我们知道文本分类的方向主要有二分类、多分类和多标签分类。根据对文本分类概念的学习和对题意的分析，问题1我们应当建立的是文本多分类模型。通过文献资料的学习，我们了解到当前文本分类主要采用传统机器学习方法和深度学习方法。深度学习方法的分类效果较传统机器学习有所提高，但考虑到其模型训练的时间花费相对较大，理论上的分类效果在实际中有时难以达到，且训练需要的数据集要求更大，因此本问题我们仍然采用传统的机器学习方法进行分类。我们将问题的解决分为三个部分：文本预处理，文本特征提取和分类模型构建。首先我们对留言的主题和详情进行预处理，得到分词和去停用词结果，然后通过TF-IDF方法计算文本的特征值，最后采用多项式朴素贝斯模型和线性支持向量机模型对提取好的文本特征进行分类，比较分类结果。

针对问题2，在文本预处理和特征提取上的步骤和第一问类似。考虑到事先并不能确定需要聚类的留言究竟分属于多少不同的事件，这里采用了基于密度的DBSCAN聚类算法，对留言特征抽象为区域中的点，把具有足够高密度的区域划分为簇，得到同一时间的留言簇。对与问题的热度进行计算和排名，应当主要考虑留言出现的频度以及点赞和反对的数量，并且利用交叉熵对各指标所占权重进行计算，最后得到关于问题的热度并进行排名，输出结果。

针对问题3，我们在阅读了大量文献，并且参考学习了现有的对话系统以及机器翻译系统的评价准则后，提出了关于回复的及时性，问题与答复的相关性，答复与问题的长度对比等相关的度量手段。

三、问题 1 基于 SVM 算法的留言分类模型

3.1 基本假设

根据实际问题，为了简便对数据的处理以及优化模型，我们对模型提出如下假设：

（1） 假设每条数据能且只能对应 15 个一级分类中的一类；

（2） 假设留言数据中给出的一级分类标签一定正确；

3.2 数据处理流程

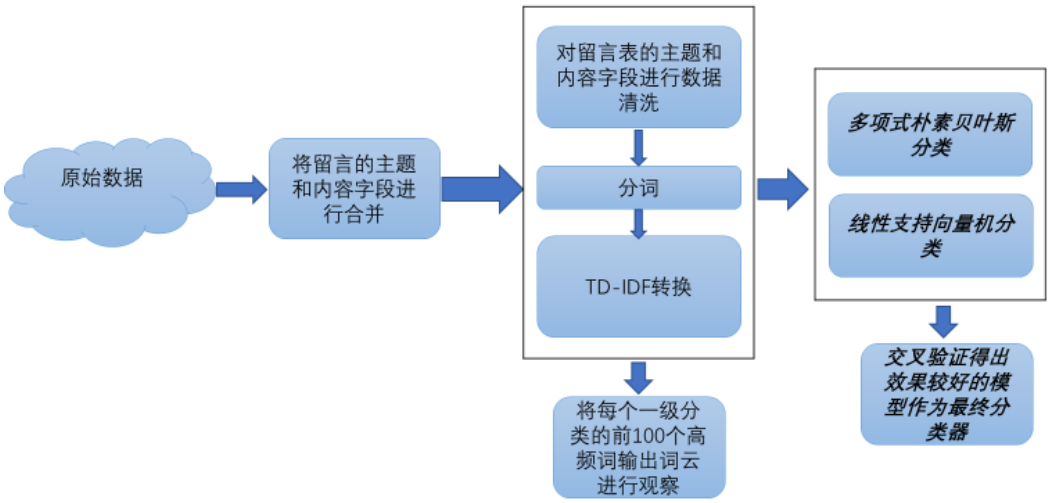

图 2：问题 1 流程

3.3 文本预处理

将附件 2 中给出的留言主题和详情字段进行合并，然后去除空白，将文本中出现的英文统一转换为小写，并且利用正则表达式的方式清理掉与详情无关的字符。

随后，我们在对文本进行聚类之前，先要把非结构化的详情文本信息转换为计算机可以识别的结构化信息。因此，为了便于转换，首先我们需要对这些详情文本进行中文分词。这里采用了 python 的中文分词包 jieba 进行分词，使用的是默认精确方式和 jieba 自带的词典。jieba 采用了基于前缀词典进行的词图扫描，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），同时采用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，采用 Viterbi 算法进行计算，这样能够更好的实现中文分词效果。

3.4 TF-IDF 转换 [1]

3.2.1 TF-IDF 算法

在对文本进行分词处理后，需要把这些词语转换为向量，以供分类使用。这里采用 TF-IDF 算法，把文本信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF（Term Frequency）权重。

$$\text{词频（TF）}=\frac{\text{某个词在文本中出现的次数}}{\text{文本的总词数}} \tag{1}$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库，即语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本详情属性能力越强。

$$\text{逆文档频率（IDF）}=\log\frac{\text{语料库的文本总数}}{\text{包含该词的文本书}+1} \tag{2}$$

第三步，计算 TF-IDF 值。

$$\text{TF-IDF}=\text{词频（TF）}\times\ \text{逆文档频率（IDF）} \tag{3}$$

3.2.1 TF-IDF 转换

这里采用 python 的 sklearn 库中的 feature_extraction.text.TfidfVectorizer 函数进行转换处理，得到关于留言文本的特征矩阵。

在使用函数时，max_df 取 0.95 表明当词频大于 0.95 的时候我们就忽略该词条；其余参数取默认值即可。

3.5 训练分类器

在对留言文本进行特征提取后，我们使用了四种分类算法，尝试对文本进行一级分类，挑选出效果最好的作为最终的结果给出。

3.5.1 训练项式朴素贝叶斯分类器[2]

在进行分类预测之前我们首先要拿处理好的特征向量训练分类器。朴素贝叶斯的算法过程如下：

我们假设训练集为 m 个样本 n 个维度，如下：

$$\left(x_1^{(1)},x_2^{(1)},\cdots,x_n^{(1)},y_1\right),\left(x_1^{(2)},x_2^{(2)},\cdots,x_n^{(2)},y_2\right),\cdots,\left(x_1^{(m)},x_2^{(m)},\cdots,x_n^{(m)},y_m\right)$$

共有 k 个特征输出类别，分别为$C_1, C_2,\cdots, C_k$，每个特征输出类别的样本个数为$m_1, m_2,\cdots, m_k$，在第 k 个类别中，如果是离散特征，则特征值为m_{jl}，其中 l 取值为 $1,2,\cdots, S_j$为特征 j 不同的取值数。输出为$X^{(test)}$的分类。

算法流程如下：

（1） 如果没有 Y 的先验概率，则计算 Y 的 K 个先验概率：

$$\mathrm{P}(Y=C_k)=\frac{m_k+\lambda}{m+K\lambda}, \tag{4}$$

否则$\mathrm{P}(Y=C_k)$为输入的先验概率。

（2） 分别计算第 k 个类别的第 j 维特征的第 l 个取值的条件概率：

$$\mathrm{P}(X_j = x_{jl} | Y = C_k) \tag{5}$$

（3） 对于实例$X^{(test)}$，分别计算：

$$\mathrm{P}(Y = C_k)\prod_{j=1}^{n} P(X_j = x_j^{(test)} | Y = C_k) \tag{6}$$

（4） 确定实例$X^{(test)}$的分类$C_{results}$：

$$\mathrm{C}_{result} = \underbrace{\arg\max}_{C_k} P(Y = C_k)\prod_{j=1}^{n} P(X_j = X_j^{(test)} | Y = C_k) \tag{7}$$

这里我们调用 python 第三方包 sklearn 库的 naive_bayes. MultinomialNB 方法对贝叶斯分类的算法进行实现。函数参数取默认值。

3.5.2 训练线性支持向量机分类器[3]

在使用多项式朴素贝叶斯分类器进行分类的同时，我们也用线性支持向量机模型训练了分类器并进行预测。支持向量机(Support Vector Machine)是 Cortes 和 Vapnik 于 1995 年首先提出的[4]，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力（或称泛化能力）。

在这里我们利用了 python 第三方包 sklearn 库的 svm.LinearSVC 方法将预处理的到的特征向量进行训练得到分类器。函数参数取默认值。

3.5.3 模型比较

在模型训练完成后，分别对两个模型进行交叉验证，比较分类效果的好坏，将效果更好的分类器作为最终选择的模型。

3.6 结果预测与模型评价

使用挑选出的分类器模型对附件 2 的数据进行预测，并计算其每一类分类结果的 F-score 分数作为评价。

四、问题 2 基于 DBSCAN 算法的分层聚类模型

4.1 数据处理流程

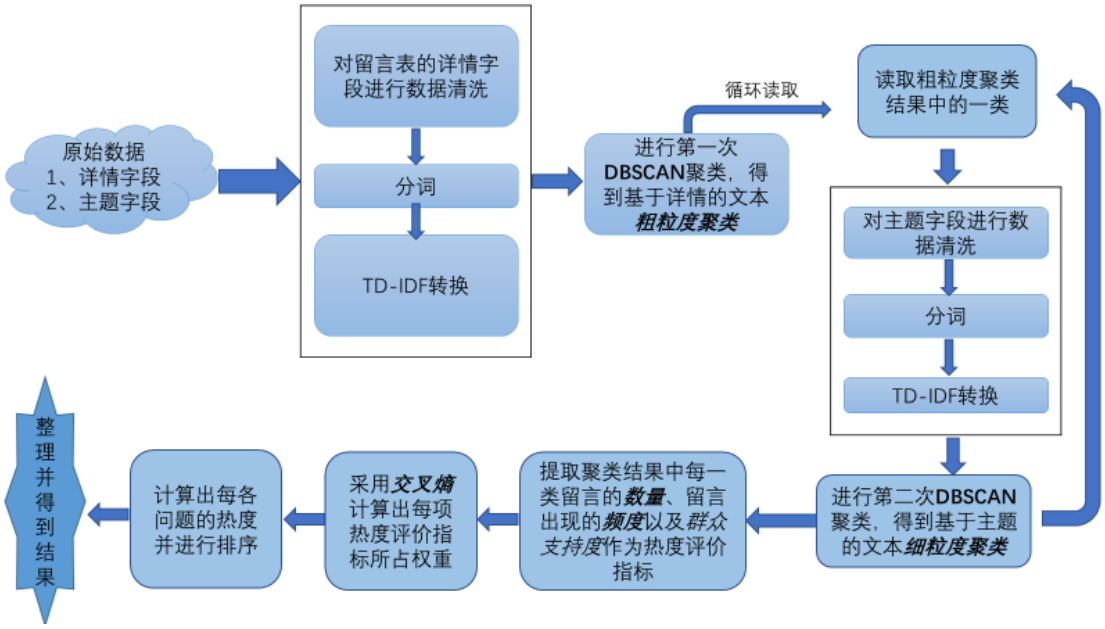

图 3：问题 2 流程

由于进行 TF-IDF 转换的时候，如果语料词库过大，会导致矩阵过于稀疏并且消耗过多内存的问题，但如果我们将转换的 max_features 参数设置为排名靠前的高频词，则会导致文本特征矩阵中的地点及人群的特征权重过低而被忽视，导致聚类结果不能很好的区分不同地点人群的相同问题。因此，我们将聚类分两个层级进行，第一层我们使用详情文本，对留言问题进行一个粗粒度的聚类，在这个过程中，由于详情的文本对于问题描述的信息较为详细，所以能粗粒度的将不同的话题进行区分。

得到对详情文本的粗粒度聚类结果后，我们分别针对每一个聚类，将其主题字段重新进行清洗和 TF-IDF 转换，此时，由于已经进行了一次聚类，所以文本的语料范围不会太大，则可以将全部语料用于特征提取。这样得到的文本主题特征向量能够在细粒度上反映文本之间的差别，尤其是地点和人群这两个特征上的差别，从而取得一个较为精确的聚类结果，把不同地点不同人群的不同问题进行区分。

4.2 文本预处理一

4.2.1 留言详情字段的数据清洗与分词

将附件 3 中给出的留言详情字段进行空白去除，将文本中出现的英文统一转换为小写，并且利用正则表达式的方式清理掉与详情无关的字符。

随后，我们在对详情（title）字段进行粗粒度聚类之前，先要把非结构化的详情文本信息转换为计算机可以识别的结构化信息。因此，为了便于转换，首先我们需要对这些详情文本进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。

4.2.2 生成 TF-IDF 向量

在对 title 文本进行分词后，需要把这些词语转换为向量，以供分类使用。这里采用 TF-IDF 算法，把 title 信息转换为权重向量。这里采用 python 的 sklearn 库中的 feature_extraction.text.TfidfVectorizer 函数进行转换处理，得到关于留言详情的特征矩阵。

4.3 基于留言详情的文本粗粒度聚类

生成留言详情的 TF-IDF 特征矩阵后，根据每条留言详情的 TF-IDF 权重向量，对文本进行聚类。这里采用 DBSCAN 算法对主题文本进行聚类。[5]

DBSCAN 聚类的原理如下：

假设样本集 D=(x1,x2,...,xm)(x1,x2,...,xm)，DBSCAN 通过在数据集中寻找被低密度区域分离的高密度区域，将分离出的高密度区域作为一个独立的类别。在 DBSCAN 算法中数据点被分类三类：

核心点（Core Point）：若样本x_i的ε邻域内至少包含了 MinPts 个样本，即$N_\varepsilon(x_i) \geq MinP$ts,则称样本点$x_i$为核心点。

边界点（Border Point）：若样本x_i的ε邻域内包含的样本数小于 MinPts，但是它在其他核心点的邻域内，则称样本点x_i为边界点。

噪音点（Noise）：既不是核心点也不是边界点的点。

DBSCAN 聚类的算法步骤如下：

输入：样本集 D=(x1,x2,...,xm)(x1,x2,...,xm)，邻域参数(ε,MinPts)(ε,MinPts)，样本距离度量方式

输出： 簇划分 C.

（1） 初始化核心对象集合 Ω=ϕΩ=ϕ, 初始化聚类簇数 k=0，初始化未访问样本集合 ΓΓ = D， 簇划分 C = ϕ ϕ。

（2） 对于 j=1,2,...m，按下面的步骤找出所有的核心对象：

I： 通过距离度量方式，找到样本x_jx_j的εε−邻域子样本集 Nϵ(x_j)Nϵ(x_j)

II ：如果子样本集样本个数满足|Nϵ(x_j)|≥MinPts|Nϵ(x_j)|≥MinPts，将样本x_jx_j加入核心对象样本集合：Ω=Ω∪{x_j} Ω=Ω∪{x_j}

（3） 如果核心对象集合 Ω=ϕ Ω=ϕ，则算法结束，否则转入步骤 4。

（4） 在核心对象集合 ΩΩ 中，随机选择一个核心对象 oo，初始化当前簇核心对象队列 Ωcur={o}Ωcur={o}，初始化类别序号 k=k+1，初始化当前簇样本集合 Ck={o}Ck={o}，更新未访问样本集合 Γ=Γ−{o} Γ=Γ−{o}。

（5） 如果当前簇核心对象队列 Ωcur=ϕ Ωcur=ϕ，则当前聚类簇 Ck Ck 生成完毕，更新簇划分 C={C1,C2,...,Ck}{C1,C2,...,Ck}，更新核心对象集合 Ω=Ω−Ck Ω=Ω−Ck， 转入步骤 3。

（6） 在当前簇核心对象队列 ΩcurΩcur 中取出一个核心对象 o′ o′，通过邻域距离阈值εε找出所有的 ϵϵ−邻域子样本集 Nϵ(o′) Nϵ(o′)，令 Δ =Nϵ(o′)∩Γ Δ =Nϵ(o′)∩Γ，更新当前簇样本集合 Ck=Ck∪Δ Ck=Ck ∪Δ，更新未访问样本集合 Γ=Γ−Δ Γ=Γ−Δ，转入步骤 5。

输出结果为： 簇划分 C={C1,C2,...,Ck}{C1,C2,...,Ck}。

这里我们调用 python 第三方包 sklearn 库的 sklearn.cluster.DBSCAN 方法对贝叶斯分类的算法进行实现，每次聚类迭代 300 次。其中粗粒度聚类时 eps 参数取 0.5（范围时 0 到 1 之间的一位小数），min_samples 取 3。参数的选取基于程序调试时对聚类结果的分析。eps 参数越大聚类结果中出选的混淆越多，而 eps 参数小于等于 0.5 时聚类个数基本收敛。

用全部数据附件 3 中给出的 4327 条数据进行粗粒度聚类，得到的聚类结果为 194 条。为了使得计算简便，我们将条目小于 7 的聚类认为是噪声数据，即没有机会成为热点问题，不参与接下来的细粒度分类。这样剩下来的就是 40 个粗粒度分类结果。

4.4 文本预处理二

对第一层的详情粗粒度聚类结果中的每一个分类，重复 4.2 文本预处理中的步骤，得到每一个粗分类的文本关于主题的特征矩阵。

4.5 对文本主题特征矩阵进行 DBSCAN 聚类

与步骤 4.3 的处理方法相同，我们对每一个粗粒度分类的结果的主题字段再进行一次 DBSCAN 聚类，得到较为准确的不同地点人群的问题的最终聚类结果。其中 eps 参数取 0.5（范围时 0 到 1 之间的一位小数），min_samples 取 2。参数的选取基于程序调试时对聚类结果的分析。eps 参数越大聚类结果中出选的混淆越多，而 eps 参数小于等于 0.5 时聚类个数基本收敛。聚类结果见热点问题明细表.excel。

4.6 对聚类的结果进行热度计算并排序

4.6.1 计算热度评价指标

为了评价问题的热度，我们选取了问题文本出现的数量W_c、问题文本在预料中出现的频度W_f、以及群众支持度W_s作为每个问题的热度度量指标。这些指标均认为数值越大越好。下面给出这些热度度量指标的计算方法：

$$\text{数量}W_c = \text{同一问题出现的文本次数} \tag{8}$$

$$\text{频度}W_f = \frac{\text{某一问题出现的文本次数}}{\text{问题出现的时间跨度天数}(\Delta t)} \tag{9}$$

$$\text{问题出现的时间跨度 } \Delta t = \text{问题最近一次出现的时间} - \text{最早出现的时间} \tag{10}$$

$$\text{群众支持度}W_s = \text{某一问题点赞总数} - \text{某一问题反对总数} \tag{11}$$

对 4.5 中得到的最终聚类结果，分别计算它们的热度评价指标，形成问题

热度指标水平矩阵 R′ 。

$$\mathrm{R}' = \begin{bmatrix} r'_{11}r'_{12}\cdots r'_{1n} \\ r'_{21}r'_{21}\cdots r'_{2n} \\ r'_{m1}r'_{m1}\cdots r'_{mn} \end{bmatrix}_{m\times n} \tag{12}$$

m 表示有 m 个样本（m 个问题），n 表示有 n 个评价参数（热度度量指标），这里的 n=3。

4.6.2 熵权法计算指标权重和问题热度

然后我们采用熵权法客观的计算出每个热度度量指标的权重。熵权法的计算步骤如下：

第一步，将问题热度指标水平矩阵 R′ 进行归一化。对矩阵的每一列进行归一化处理，得到归一后的矩阵 R。计算公式如下：

$$r_{ij} = \frac{r'_{ij} - \min\limits_{i}(r'_{ij})}{\max\limits_{i}(r'_{ij}) - \min\limits_{i}(r'_{ij})} \tag{13}$$

第二步，熵权处理，计算每个指标的熵H_j。

$$H_j = -k\sum_{i=1}^{m} f_{ij}\cdot \ln f_{ij} \tag{14}$$

$$f_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}} \tag{15}$$

$$\mathrm{k} = \frac{1}{\ln m} \tag{16}$$

其中f_{ij}为第 j 个指标下第 i 个问题的指标值的比重。

第三步，计算每个指标的熵权λ_j。

$$\lambda_j = \frac{1-H_j}{\sum_{j=1}^{n}(1-H_j)} = \frac{1-H_j}{n-\sum_{j=1}^{n}H_j} \tag{17}$$

第四步，对指标进行加权求和,求出每个问题的热度。

$$w_i = \sum_{j=1}^{n} \lambda_j\, r_{ij} \tag{18}$$

4.6.3 进行热度排序得出热点问题

将 4.6.2 中计算出的问题热度进行排序，挑选出前五名作为热点问题，得出热点问题表和热点问题留言明细表。

五、问题答复质量评价指标模型

5.1 评级指标计算流程

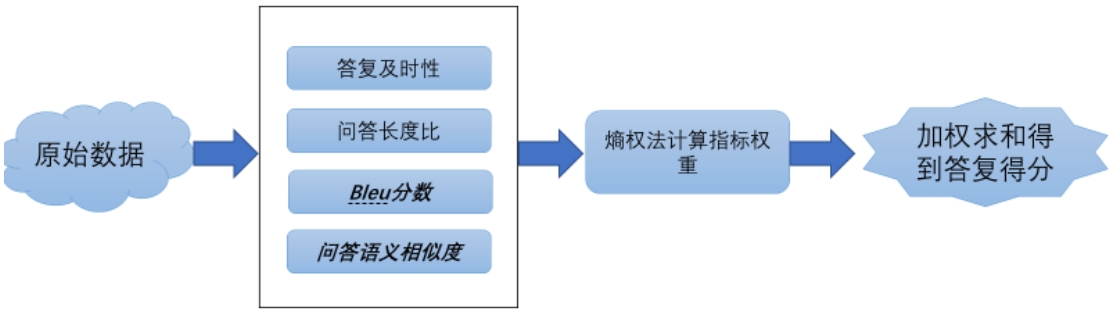

图 4：问题 3 流程

5.2 评价指标的提出

通过查阅文献，和参考现有的对话系统评价方法[6]，我们提出了以下几个评价指标：

（1） 答复及时性W_t[7]，通过对每条留言的留言时间与答复时间做差值，求出答复距离留言的时间长短，我们可以评价答复意见的及时性。

（2） 问答长度比W_l，为了比较答复的完整性和认真程度，我们假设问题文本越长，那么答复应当越详细，即长度越长。因此，将答复的长度与留言详情的长度的比值作为一个指标，问答长度比越大，说明答复越详细，完整性越高。

（3） 答复的 Bleu 分数W_b，计算答复意见关于留言详情与 Bleu 分数，也就是答复意见与留言详情的词重叠程度，能够在一定程度上反应答复与留言详情的相关性，答复意见的 Bleu 分数越高，代表量大幅与留言的相关性越高[8]。

（4） 问答语义相似度W_s，由于基于统计词频的 Bleu 分数有时并不能很好的衡量留言详情与答复意见在语义上的相关性，因为表达意思相同的情况下，留言详情和答复意见可能会使用不同的词汇，这就会导致在问答相关性的评价指标上出现误差。因此我们采用向量极值法[9]（Vector Extrema）计算出的留言详情与答复意见之间的语义相似度作为问答相关性度量的一个指标。

5.3 评价指标的计算方法

下面详细给出各项指标的计算方法：

（1） 答复及时性W_t：

$$W_t = MAX(\text{某一留言的答复时间} - \text{某一留言的提问时间}) - (\text{某一留言的答复时间} - \text{某一留言的提问时间}) \quad (19)$$

其中W_t单位为天，且将MAX（某一留言的答复时间 − 某一留言的提问时间）取基本值 0。

（2） 问答长度比W_l：

$$W_l = \frac{\text{留言答复的文本长度}}{\text{留言详情的文本长度}} \quad (20)$$

（3） 答复的 Bleu 分数W_b：

I：Bleu 分数的原理[8]：

BLEU(bilingual evaluation understudy) 中文名称为双语互译质量辅助工具，它被广泛应用于机器翻译领域，用来衡量衡量机器翻译的结果与专业人工翻译结果的差异。本质上讲 BLEU 就是用来衡量机器翻译文本与参考文本之间的相似程度的指标,取值范围在 0-1, 取值越靠近 1 表示机器翻译结果越好。这里我们借鉴它的思想，将其用于衡量留言详情与留言答复的相关性。

II：计算方法：

使用一个累加器表示留言答复中的词在留言详情中出现的次数, 从留言答复文本中的第一个词开始比较, 如果在参考文本中出现过, 那么计数加 1。 最后使用这个累加值除以留言答复中的词数目即可计算得到文本的 BLEU 取值。

$$W_b = \frac{\text{留言答复中的词在留言详情中出现的累计次数}}{\text{留言答复中的词数}} \quad (21)$$

（4） 问答语义相似度W_s：

Vector Extrema 是一种在句子级向量上计算相似度的方法。通过筛选词向量的每一维来选择整句话中极值最大的一维作为这个句子的向量表示，然后通过计算留言详情与答复之间的余弦距离来表示问答之间的语义相似度。

I：采用 cw2vec 模型计算出基于语义的文本词向量[10]。

II：筛选词向量的每一维来选择整句话中极值最大的一维作为这个句子的向量表示：

$$\mathrm{e}_{\gamma d} = \begin{cases} \max_{we\gamma} e_{wd} & if\ e_{wd} > |\min_{w'e\gamma} e_{w'd}| \\ \min_{we\gamma} e_{wd} & otherwise \end{cases} \quad (22)$$

III：计算留言详情与回复之间的余弦距离。

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (23)$$

5.3 熵权法确定指标权重

与问题 2 中的处理相同，采用熵权法客观的计算出每个答复质量度量指标

的权重。基本步骤如下：

（1） 根据答复质量评价指标，计算出每条答复文本的的指标度量值，形成留言答复质量水平矩阵 R′ 。

（2） 将留言答复质量水平矩阵 R′ 进行归一化。对矩阵的每一列进行归一化处理，得到归一后的矩阵 R。

（3） 熵权处理，计算每个指标的熵H_j。

（4） 计算每个指标的熵权λ_j。

（5） 对指标进行加权求和,求出每条答复文本的评价分数

过程我们已经尝试实现，源代码保存在文件“Q3 评价指标.ipynb”中

六、结果分析

6.1 问题1结果分析

6.1.1 数据预处理

首先我们对附件2中的数据进行分析，得到文本关于一级分类的样本分布，统计结果如下：

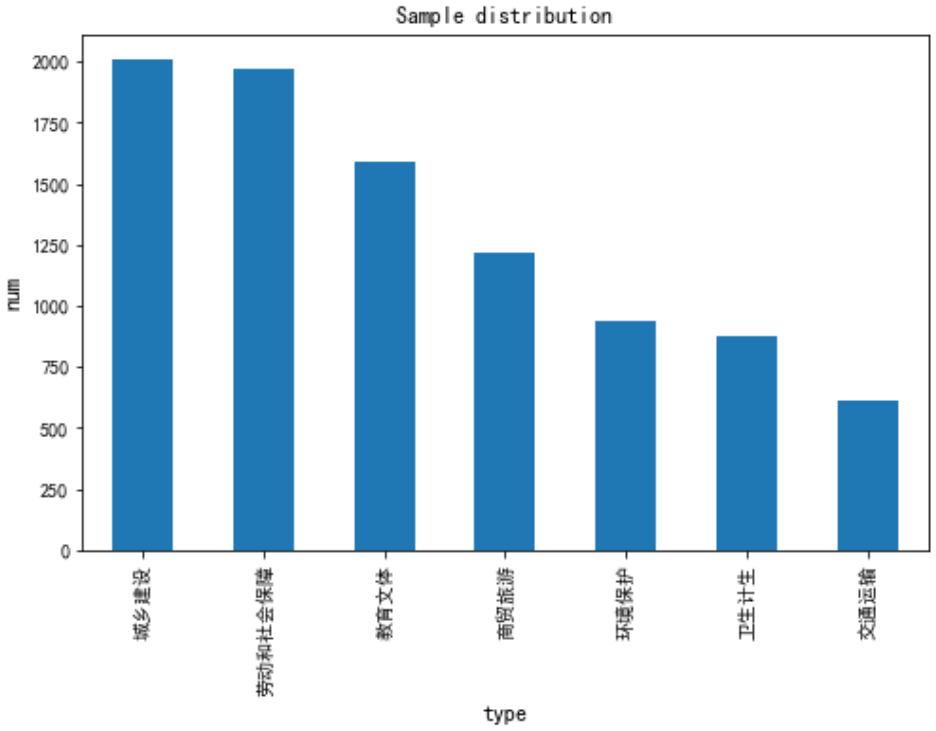

图5：附件2样本分布条形统计图

在对文本进行清洗分词后，我们计算出每一类文本的数量排名前100的高频词，转换成wordcloud（词云.png）。我们发现每个一级分类都有代表其特点的

高频词，因此可以考虑基于统计词频的概率预测方式进行分类。同时，我们也观察到不同分类的高频词存在部分重叠，这可能会对分类的准确率产生影响。这里给出城乡建设和环境保护两个分类的词云：

图6：一级分类词云示例

随后我们对分词结果进行 TF-IDF 转换。

6.1.3 分类器结果

通过对清洗后的文本进行分词、TF-IDF转换，然后将文本特征向量和标签ID作为参数分别对贝叶斯分类器和LinearSVC分类器进行训练，随后分别对两个模型进行交叉验证，画出ACC箱体图如下：

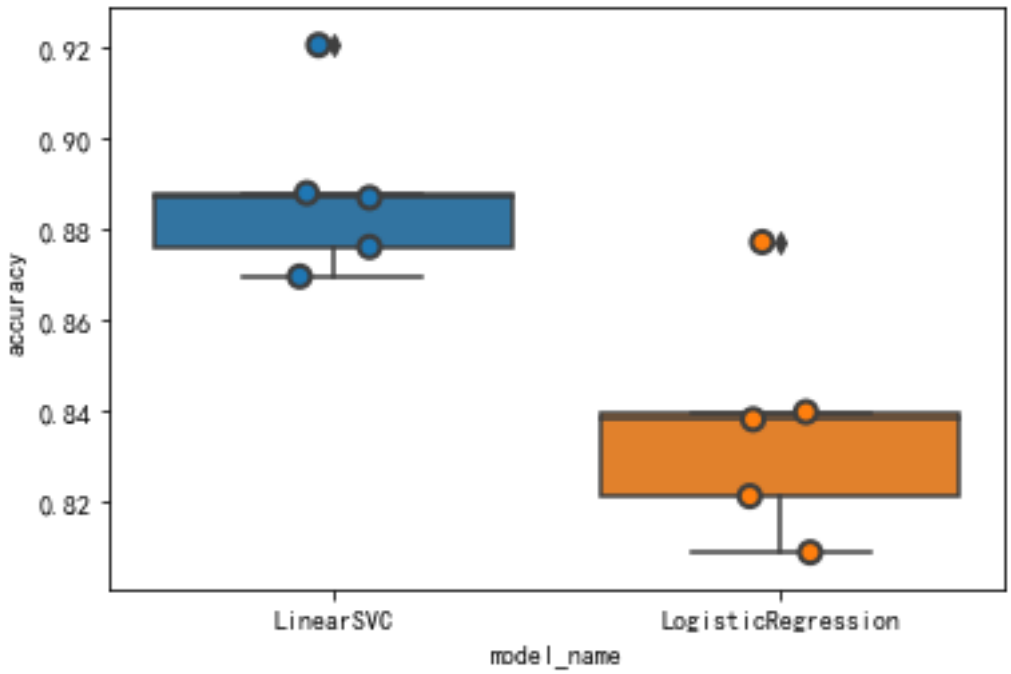

图7：ACC箱体图

根据对二者准确率的比较，线性支持向量机的分类结果更适用于对留言文本进行分类。

6.1.3 对支持向量机分类结果的分析

根据上文比较得出，线性支持向量机在留言分类上的效果更好。用训练好的LinearSVC模型对附件2数据的33%进行预测，得到的准确率（accuracy）为0.914445。并输出预测结果的混淆矩阵如下：

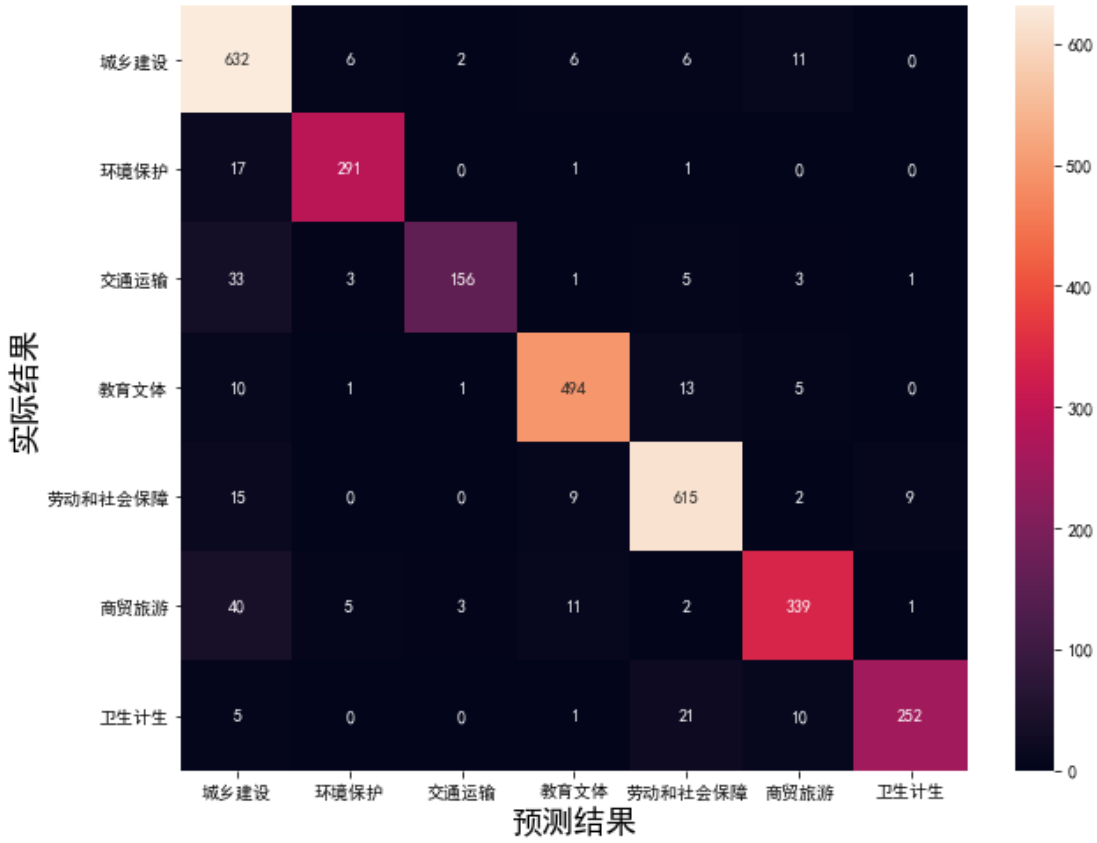

图8：混淆矩阵

由图可知，LinearSVC分类器对交通运输的分类结果最好，只有6例错误，对城乡建设的分类结果最差。分析结果较差原因，可能是由于城乡建设分类的高频词范围较广，与其他分类的高频词存在重叠。

6.1.4 线性支持向量机模型评价

对LinearSVC模型的分类结果，分别计算每一类的精确率、召回率、F-score分数、预测数量，得出分类结果评价表如下：

表 1：问题 1 分类结果评价

	Precision	Recall	F1-scores	Support
城乡建设	0.84	0.95	0.89	663
环境保护	0.95	0.94	0.94	310
交通运输	0.96	0.77	0.96	202
教育文体	0.94	0.94	0.94	524
劳动和社会保障	0.93	0.95	0.94	650
商贸旅游	0.92	0.85	0.88	401
卫生计生	0.96	0.87	0.91	289

为了进一步比较训练样本数量对分类效果的影响，我们将附件 2 中的卫生计生和交通运输两类数据与其他五类从政府网站收集到的公开留言数据共 28463 条，作为改进的大样本的训练数据进行清洗和 TF-IDF 转换。样本分布图如下：

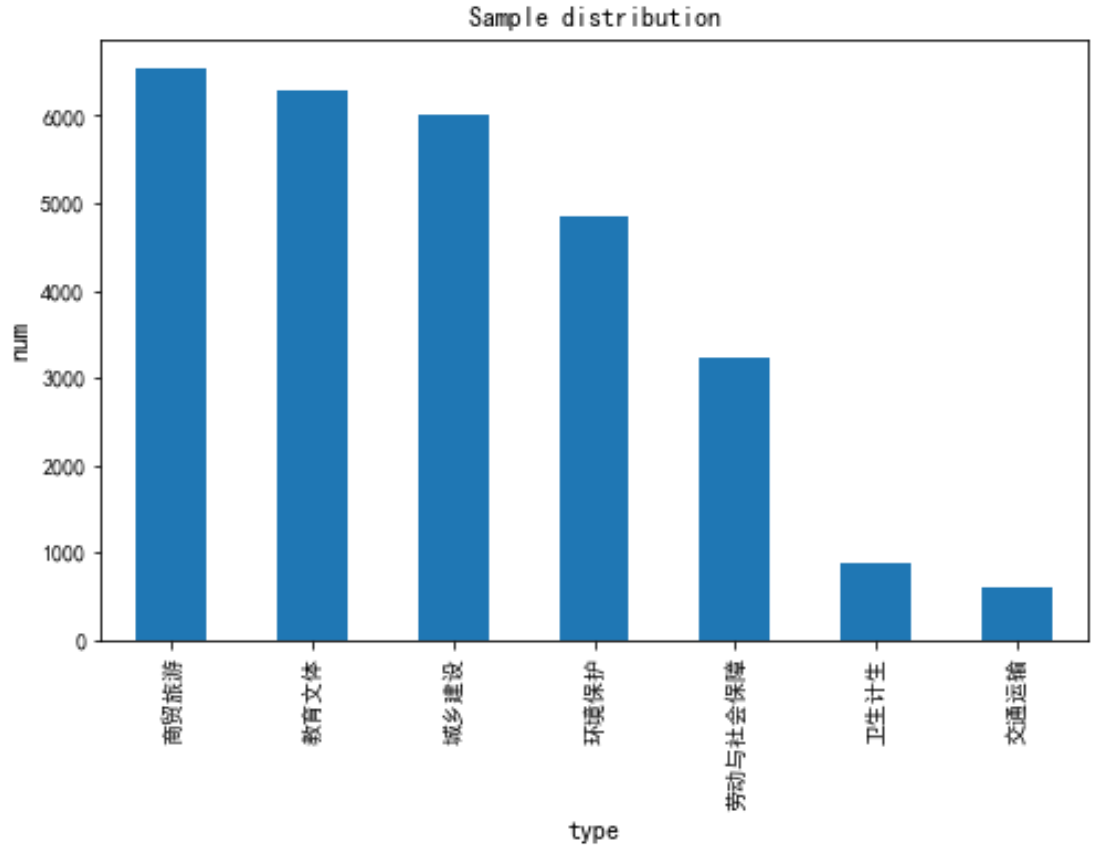

图 9：大样本分布条形统计图

再次对 LinearSVC 模型进行训练，并取训练数据的 33%作为测试集，对模型进行检验，得到的准确率为 0.940681。并画出混淆矩阵图如下：

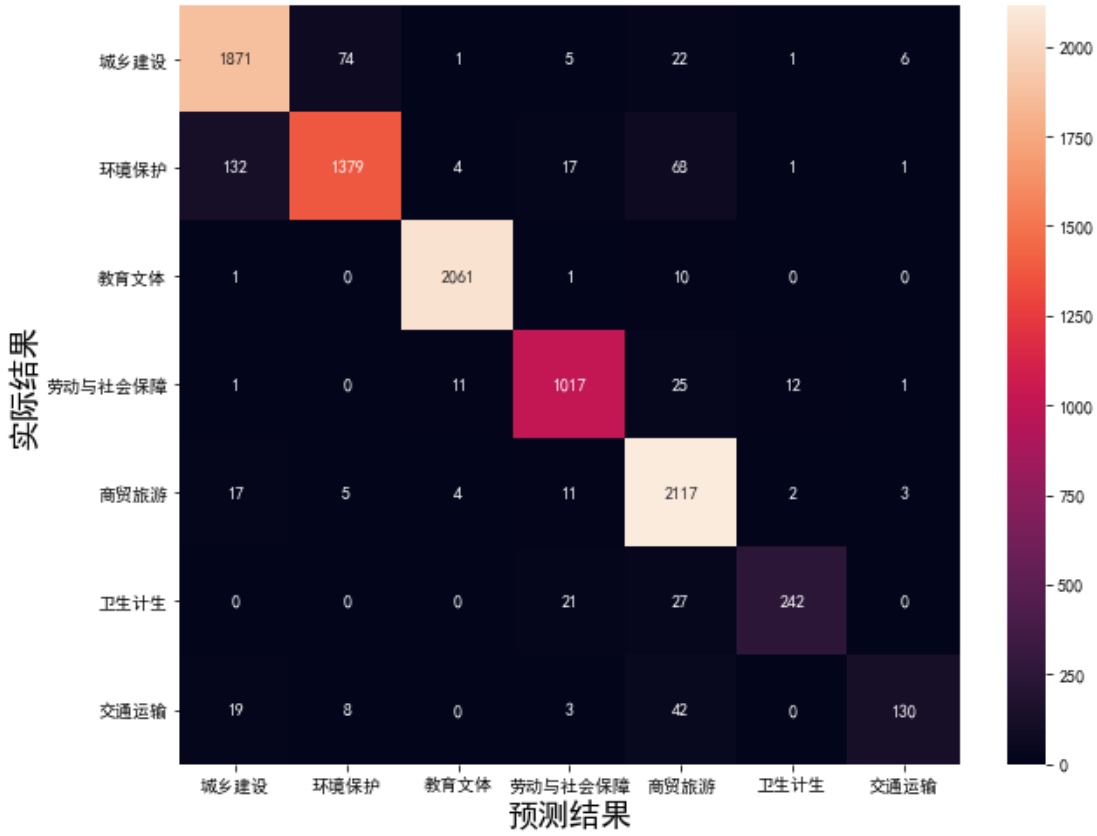

图 10：大样本预测混淆矩阵

并且，我们仍然计算出每一类的精确率、召回率、F-score 分数、预测数量，得出分类结果评价表如下：

表 2：问题 1 大样本分类结果评价

	Precision	Recall	F1-scores	Support
城乡建设	0.92	0.94	0.93	1980
环境保护	0.94	0.86	0.90	1602
教育文体	0.99	0.99	0.99	2073
劳动和社会保障	0.95	0.95	0.95	1067
商贸旅游	0.92	0.98	0.95	2159
卫生计生	0.94	0.83	0.88	290
交通运输	0.92	0.64	0.76	202

从图表中不难看出，随着样本规模的增大，分类的准确率和 F-score 分数等指标都在上升，但也可以发现，由于卫生计生和交通运输的数据规模没有变化，导致在大样本中的占比减少，使得分类的 F-score 分数出现下降。因此，数据的规模与均衡性对模型的分类结果产生重要影响。

6.2 问题2结果分析

6.2.1 数据预处理

对附件3 留言的详情字段进行数据清洗和分词，示例结果如下：

	title	time	content	yes	no	review	review_	review_cut	content_	content_cut
0	A3区—米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	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...	0	0	A3区一米阳光婚纱艺术摄影是否合法纳税了?\n座落在A市A3区联丰路米兰春天G2栋320，一...	A3区 一米阳光 婚纱 艺术摄影 合法 纳税 座落在 A市 A3区 联丰路 米兰 春天 G2...	[A3区, 一米阳光, 婚纱, 艺术摄影, 合法, 纳税, 座落在, A市, A3区, 联丰...	座落在 联丰路 米兰 春天 G2 名叫 一米阳光 婚纱 艺术摄影 影楼 工作室 营业额 地处...	[座落在, 联丰路, 米兰, 春天, G2, 名叫, 一米阳光, 婚纱, 艺术摄影, 影楼,...
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	A市A6区道路命名规划已经初步成果公示文件，什么时候能转化成为正式的成果，希望能加快完成的路...	0	1	咨询A6区道路命名规划初步成果公示和城乡门牌问题\nA市A6区道路命名规划已经初步成果公示文...	咨询 A6区 道路 命名 规划 成果 公示 城乡 门牌 A市 A6区 道路 命名 规划 成果...	[咨询, A6区, 道路, 命名, 规划, 成果, 公示, 城乡, 门牌, A市, A6区,...	道路 命名 规划 成果 公示 文件 转化 成果 希望 加快 路名 规范 道路 安装 名牌 变...	[道路, 命名, 规划, 成果, 公示, 文件, 转化, 成果, 希望, 加快, 路名, 规...
2	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	本人系春华镇金鼎村七里组村民，不知是否有相关水泥路到户政策和自来水到户政策，如政府主导投资村...	0	1	反映A7县春华镇金鼎村水泥路、自来水到户的问题\n本人系春华镇金鼎村七里组村民，不知是否有相...	A7县 春华 金鼎村 水泥路 自来水 到户 春华 金鼎村 村民 不知 相关 水泥路 到户 政...	[A7县, 春华, 金鼎村, 水泥路, 自来水, 到户, 春华, 金鼎村, 村民, 不知, ...	春华 金鼎村 村民 不知 相关 水泥路 到户 政策 自来水 到户 政策 政府 投资 村民 集...	[春华, 金鼎村, 村民, 不知, 相关, 水泥路, 到户, 政策, 自来水, 到户, 政策...
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	靠近黄兴路步行街，城南路街道、大古道巷、一步两搭桥小区（停车场东面围墙外），第一单元一住户卫...	0	1	A2区黄兴路步行街大古道巷住户卫生间粪便外排\n靠近黄兴路步行街，城南路街道、大古道巷、一步...	A2区 黄兴路 步行街 古道 住户 卫生间 粪便 外排 靠近 黄兴路 步行街 南路 街道 古...	[A2区, 黄兴路, 步行街, 古道, 住户, 卫生间, 粪便, 外排, 靠近, 黄兴路, ...	靠近 黄兴路 步行街 南路 街道 古道 搭桥 小区 停车场 围墙 单元 住户 卫生间 粪便 ...	[靠近, 黄兴路, 步行街, 南路, 街道, 古道, 搭桥, 小区, 停车场, 围墙, 单元...

图 11：问题 3 数据预处理

6.2.2 粗粒度DBSCAN聚类结果分析

对预处理完成的结果进行TF-IDF转换，随后将转换生成的特征矩阵进行DBSCAN聚类，搜索边距eps取0.5，min_samples取3，得到的聚类结果如下图：

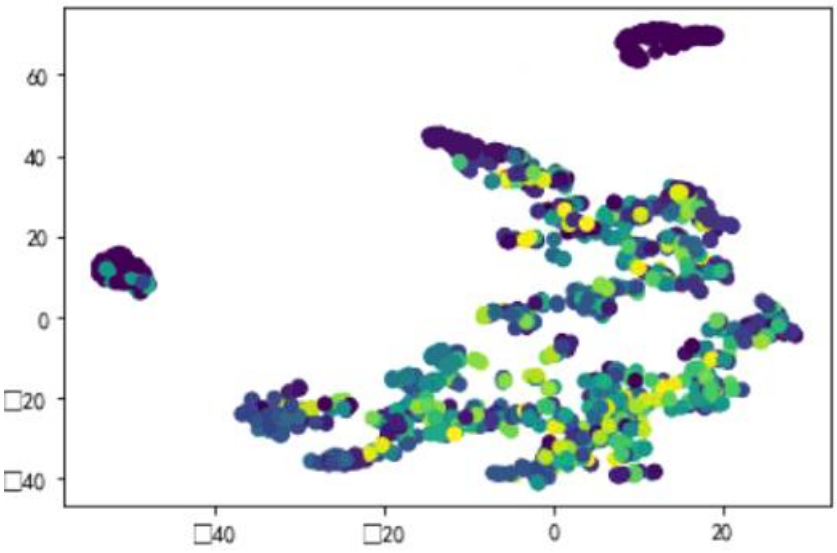

图12：粗粒度DBSCAN结果

可以看到由于样本数量众多，许多杨笨在提取特征之后相互之间的区别仍然不是特别明显，所以一次聚类可能并不能顺利将不同事件区分开，查看第一次聚类的结果我们也发现其中由部分数据出现了事件混淆。

并且输出聚类结果的直方图（保存在粗粒度直方图.png），下图为部分聚类类结果直方图如下：

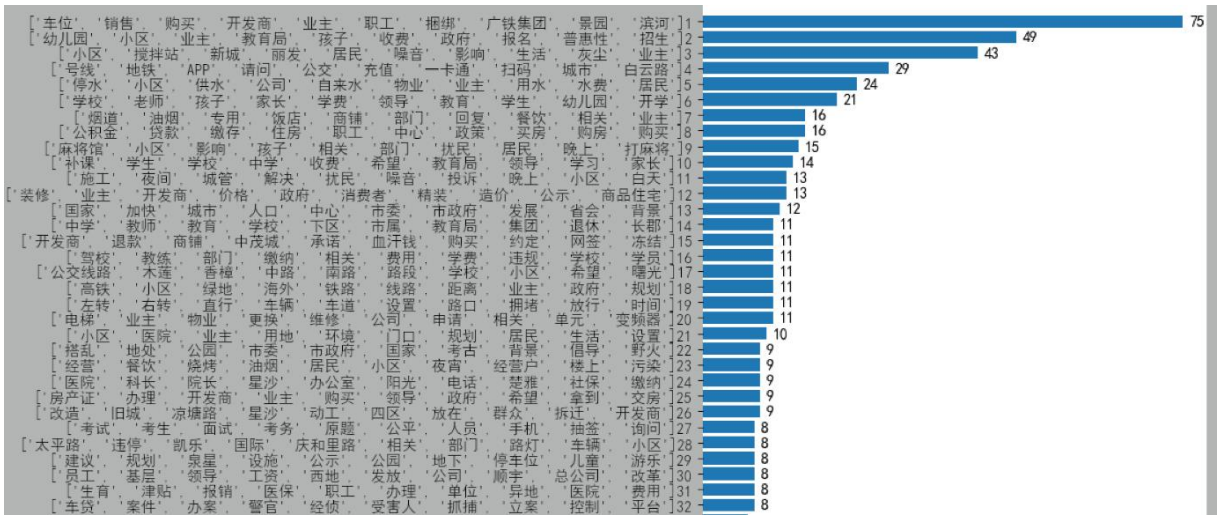

图13：粗粒度聚类部分结果直方图

对聚类结果进行观察，发现能够较好的区分不同的事件，但部分样本存在未将问题同类但地点与人群不同的条目分开。因此我们进行了第二轮基于主题细粒度的聚类。

6.2.3 细粒度聚类结果与分析

依次将粗粒度聚类的结果中主题字段进行数据清洗、TF-IDF转换，然后将特征矩阵进行细粒度聚类其中eps参数取0.5（范围时0到1之间的一位小数），min_samples取2。参数的选取基于程序调试时对聚类结果的分析。eps参数越大聚类结果中出选的混淆越多，而eps参数小于等于0.5时聚类个数基本收敛。画出其中两个粗粒度聚类的聚类结果如下图：

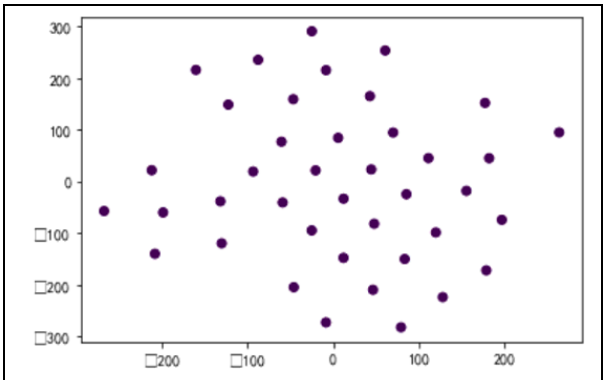

图14：细粒度聚类1

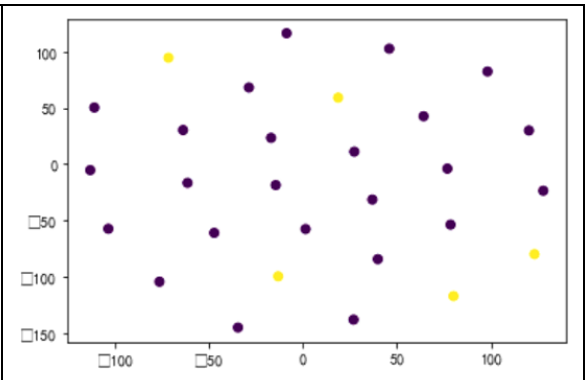

图15：细粒度聚类2

据图观察可知，进行细粒度分类的时候各样本点之间的距离比较均匀，能够更好的区分出不同地点不同人群的不同话题。

6.2.4 热度计算

计算细粒度聚类结果中每个问题的各项热度评价指标，并对形成的热度评价指标矩阵用熵权法进行权重计算。随后将各问题的指标进行加权求和，计算得出每个问题的热度并且进行排序，输出前五名作为热点问题。热点问题结果如下表：

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	21.49563167	2019-09-01 至 2019-01-08	广铁集团伊景园	A 市伊景园滨河苑向广铁集团职工捆绑销售车位
2	5	20.61069866	2020-01-15 至 2019-11-02	A2 区暮云街道丽发新城	丽发新城小区附近水泥搅拌站噪音粉尘污染扰民
3	6	12.10400210	2020-01-03 至 2019-01-23	A 市魅力之城	A 市魅力之城小区夜宵摊产生油烟和噪音污染扰民
4	15	5.66462511	2019-10-30 至 2019-05-09	A 市人才补贴	A 市人才补贴政策实施
5	12	5.30344468	2019-11-29 至 2019-06-03	A 市公交线路	A 市公交线路整改

通过热度计算，综合各个指标可以得到样本中热度最高的问题是“A市伊景园滨河苑向广铁集团职工捆绑销售车位”，而分析其他几个问题也可发现，数量多的不一定热

度也高，结果是多方因素综合的结果

6.3 问题3模型评价

对问题3中提出的答复质量评价模型进行分析，发现模型存在以下不足：

一是计算问题与答复之间的相似度和答复时间上的及时性只能在一定程度上反应答复的质量，并不能反映全局上留言问题有无及时得到解决。因此，我们可以引入一个指标，答复有效性W_e来进行评价。下面给出了W_e的计算思路：

答复意见越有效，越能很好的解决留言问题则当答复意见给出后相同的问题提出的留言就会越少。因此我们对留言进行一次聚类，找出相同问题，然后再计算出每个问题的答复有效性W_e，即可评价出答复的有效性。计算思路如下：

I：将留言进行聚类：

采用问题 2 类似的处理方法，将留言详情进行一次细粒度 DBSCAN 聚类。

II：计算每一类留言的答复有效性W_e：

①读取每一类留言的最早答复时间$t_{earlist}$；

②将该论文留言的留言时t_i间与最早答复时间$t_{earlist}$进行比较，计算在t_i在$t_{earlist}$之后的文本数量 k；

③计算$W_s = \frac{1}{k}$；

W_e的默认值取 1，即答复给出后没有同类问题留言的文本$W_e = 1$。

二是交叉熵计算各项指标权重体现的是一个客观的指标权重，而留言答复的质量好坏可能需要专业人士对各指标的而权重进行衡量，因此仅采用交叉熵的方式计算权重也许不能很好的衡量答复意见的质量。在进一步对评价模型进行改进的时候可以考虑加入专业人士的意见对权重的调整。并且，加权求和计算出的答复意见分数，也需要通过专业人士指定的量化标准来对其好坏进行判定，仅仅考虑分数在样本中的排名只能得出相对优秀答复而不是客观优秀答复。

七、参考文献

[1] 金镇晟．基于改进的 TF-IDF 算法的中文微博话题检测与研究[D]．北京理工大学．2015．

[2] 贺 鸣 孙建军 成 颖．基于朴素贝叶斯的文本分类研究综述[J]．情报科学，2016，34（7）：147-154．

[3] 平 源．基于支持向量机的聚类及文本分类研究[D]．北京邮电大学．2012．

[4] 杜圣东．基于多类支持向量机的文本分类研究[D]．重庆大学．2007．

[5] Domenica ArliaMassimo Coppola．Experiments in Parallel Clustering with DBSCAN [C]．European Conference on Parallel Processing，2001．

[6] 张杨子．面向对话系统回复质量的自动评价研究[D]．哈尔滨工业大学．2018．

[7] 张继勋 韩冬梅．网络互动平台沟通中管理层回复的及时性、 明确性与投资者投资决策[J]．管理评论，2015，27（10）：70-83．

[8] 汪菊琴 高俊涛．基于实例的 BLEU 翻译评价方法[J]．电脑知识与技术，2016，34（7）：147-154．

[9] Xiaoming Xue Jianzhong Zhou Yongchuan Zhang Xiao Jian Xuemin Wang．An Extrema Extension Method Based on Support Vector Regression for Restraining the End Effects in Empirical Mode Decomposition [C]．Proceedings of the 2013 2nd International Symposium on Manufacturing Systems Engineering，2013．

[10] 高统超 张云华．基于 cw2vec 和 BiLSTM 的中文商品评论情感分类[J]．软件导刊，2019．