“智慧政务”中的文本挖掘应用

摘要：

本文以某政府近几年来的群众问政留言和政府回复意见为研究对象，进行文本挖掘，研究的主要问题包括：对群众留言的分类，热点问题的聚类分析热度评价和对答复意见的综合评价评价。目的在于通过python实现机器学习的方法，实现处理群众留言的自动化，节省了人工成本，有利于更好的了解民意，适应大数据时代的科学施政。

在数据预处理部分，我们通过pandans库处理xlsx文件的大量数据，应用正则表达式，去停用词等方法进行数据清洗，基于词袋模型与tf-idf建立词向量并针对中文文本的高维度特点进行了降维处理。

对于群众留言分类，我们训练XGBoost模型进行多分类，并通过混淆矩阵计算F-Score对来对模型进行评价，调优。

对于热点问题的挖掘，我们采用的是无监督聚类算法-DBSCAN，主要利用其无需提前设定聚类类别数量特点，并对结果进行了二次处理，包括设定热度评价，进行命名实体识别，创建专用字典，以提高结果质量，并进行热度排名。

对于答复意见的评价，综合相关性、可解释性、规范性、时效性等多个角度角度对答复意见进行了全方面分析并量化实现，并计算出各个指标融合的总评分。

关键词： python 群众留言处理 XGBoost DBSCAN

目录

摘要：.....1
关键词：python 群众留言处理 XGBoost DBSCAN.....1
一、 问题重述.....3
1、 问题背景.....3
2、任务.....3
二、 问题分析.....3
三、 问题求解.....4
1、群众留言分类.....4
1.1 数据预处理.....4
1.2 建立模型.....6
1.3 模型评价.....10
2、热点问题提取.....11
2.1 数据预处理.....11
2.2 建立模型.....12
2.3 结果处理.....15
3、答复意见的评价.....18
3.1 对评价指标的选取.....18
3.2 相关性的量化实现.....18
3.3 规范性的量化实现.....20
3.4 可解释性的量化实现.....21
3.5 时效性的量化实现.....21
3.6 加权求总分.....22
3.7 处理结果.....22
四、 参考文献.....23

一、问题重述

1、 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

2、任务

任务一：群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

任务二：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”。

任务三：答复意见的评价

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、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

二、问题分析

针对任务一：对于本题，关键在于对留言数据的提取与分析处理，通过 python 实现文件的读取、数据的提取和处理以及数据的保存，并删除了附件2中无用数据。建立关于留言内容的一级标签分类模型。

针对任务三：本题的关键在于对于答复意见的评价指标的选取与量化，以及实现

的方法。在本题中我选取了相关性、可解释性、规范性和实效性，并对其取不同的权重来得到一个回复意见的最终评分。

三、 问题求解

1、群众留言分类

1.1数据预处理

首先，对附件2中的数据进行提取和处理。利用pandas库将xlsx文件提取为dataframe格式，以便对文本进行预处理。题目所给数据噪声较小，常规数据清洗步骤即可完成预处理，具体思路如下：

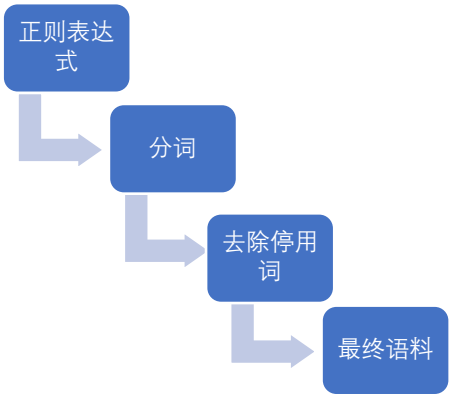


图 1、数据预处理步骤

对于第二与第三题，数据预处理过程与第一题大同小异，在介绍后两题时将不再进行展开。

1.1.1正则表达式

相对于利用停用词库，使用replace()方法去除空格，换行符等更加方便，而且不影响分词结果。值得一提的是，部分编码为/u3000等的空格无法在utf8格式读取时正常显示为空格，因此采用内置库unicodedatak库的normalize()函数扩充编码，使其正常显示并被去除。

1.1.2分词

接下来，为了筛选其中的无用关键词，我们先把数据内容进行分词操作，这里我们定义了代码中的jieba.lcut函数，用精确分词的方式将内容切成列表。

1.1.3去除停用词

由于其中无失意词汇较多且存在大量标点符号，我们下一步决定使用去除停用词的方法，对分词结果进行了清洗。我们首先在网站上查找了常用停用词表，直接进行去停用词处理。通过分析清洗结果，我们又在停用词表中添加了英文字母，年份，并且发现了“http”，“baidu”等网址字符带来的噪音，使得语料库得到了大量缩减，从而得到我们建立模型所用的最终语料库。

1.1.4数字化标签

对于数据中的已知标签类别，由于在代码中难以表示，这里我们将已知的7种分类进行数字化标签，即分别用数字0-6中的一个数字表示。

1.1.5计算tf-idf

为了进一步用数据的方式呈现关键字的词频，我们决定计算其tf-idf值，使词频量化。接下来进行tf-idf值的计算，导入sklearn.feature_extraction.text，并引用其中的TfidfVectorizer（相对词频向量生成器）。然而语料库数据庞大，而一些词的词频很低。为了更加清晰地展示其中一些关键字的低频特点，这里我们建立了一个关于tf-idf值的柱状图，如下：

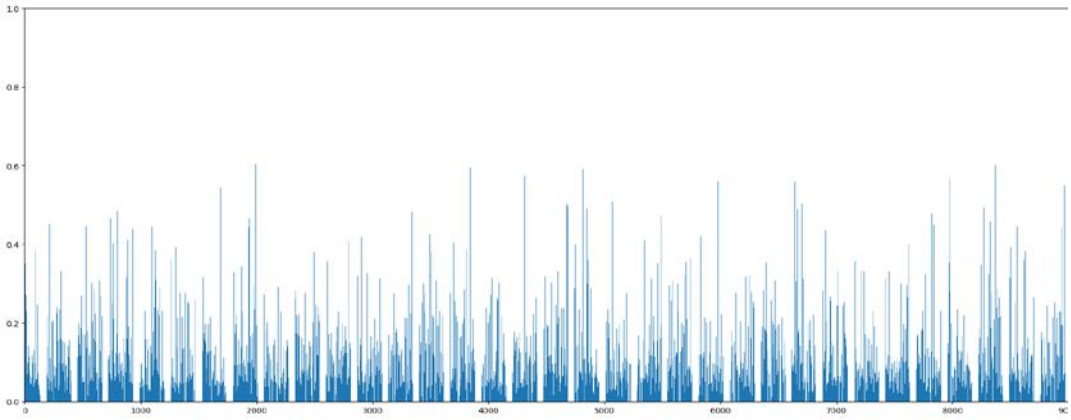


图 2、tf-idf值柱状图

显然在抽取的9000个词组中，存在大量tf-idf值极小的低频词，通过更改y轴范围我们观察到下图：

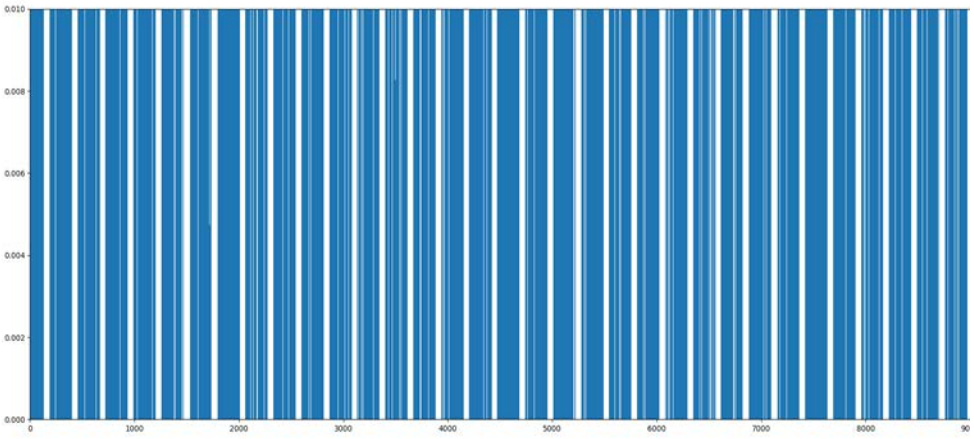


图 3、y轴范围小时tf-idf柱状图

即使Ymax取到了0.01，依然存在大量tf-idf值几乎为零的词汇，它们在算法分类过程中，不但不能提高分类效果，反而会拉低高频词汇的权重，降低分类效果这里我们写了一个函数filter，用来降低维度，在第一题中只取每一个样本的tf-idf值前250的词汇，就降低低词频词汇的影响。而为了保持其向量维度一致，所以在训练模型时用向量矩阵，而保存过程，我们则用稀疏矩阵的方法来节省空间。如下图

```
(0, 40265)    0.13496168699920294
(0, 40122)    0.12439171697601462
(0, 40270)    0.15635676763316497
(0, 65306)    0.086355198516O7133
(0, 27169)    0.23141705772044147
(0, 67306)    0.1848255942711494
(0, 7558)     0.18120388117990946
(0, 66999)    0.13672252258584547
(0, 43234)    0.10323832838633101
(0, 2364)     0.17184878252125435
(0, 22828)    0.16056819260718572
(0, 40764)    0.23141705772044147
(0, 73357)    0.10005058844521562
(0, 27243)    0.12389336617984692
(0, 51150)    0.21602467459889338
(0, 72670)    0.2622051444069502
```

图 4、稀疏矩阵

左侧为每一个非零量的索引，右侧为对应的tf-idf值。经过以上处理，我们基本完成了数据的预处理。

1.2建立模型

对于分类模型的选择，我们倾向于采用解释性强、简单、速度快的决策树类的模型。经过对比调查，分析印证最终我们选择了近些年在数据挖掘竞赛上比较热门的XGBoost模型[1]，它既保有决策树模型优点的同时，具备精度高、速度快、可扩展性高、防止过拟合特点的xgboost模型。

1.2.1模型原理

XGBoost是一个树集成模型，直白的讲它将K个树的结果进行加权求和作为最终的预测值。f表示棵cart树。

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathrm{F}$$

假设给定的样本集有n个样本，m个特征，则

$$\mathrm{D} = \{(\mathbf{x}_i, y_i)\} \left(|\mathrm{D}| = n, \mathbf{x}_i \in \square^m, y_i \in \square\right)$$

其中 xi 表示第i个样本，yi 表示第i个类别标签，回归树（CART树）的空间F为

$$\mathrm{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\} \left(q : \square^m \to T, w \in \square^T\right)$$

其中q代表每棵树的结构，他将样本映射到对应的叶节点；T是对应树的叶节点个数；f(x)对应树的结构q和叶节点权重w。所以XGBoost的预测值是每棵树对应的叶节点的值的和。

我们的目标是学习这k个树，所以我们最小化下面这个带正则项的目标函数：

$$\mathrm{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

$$\text{where}\,\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2$$

最后将关于树模型的迭代转化为了关于叶子节点的迭代，并求出最优的叶节点分数。

将叶节点的最优值带入目标函数，最终目标函数的形式为：

$$\square^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T$$

Xgboost的一个重要特点就是监督学习，监督学习就是训练数据有标签的学习。虽然以上的公式推导过程比较复杂，但xgboost对应的模型实际上就是一堆CART树。

下面我们举一个用来判断一个人是否喜欢玩电脑的例子：

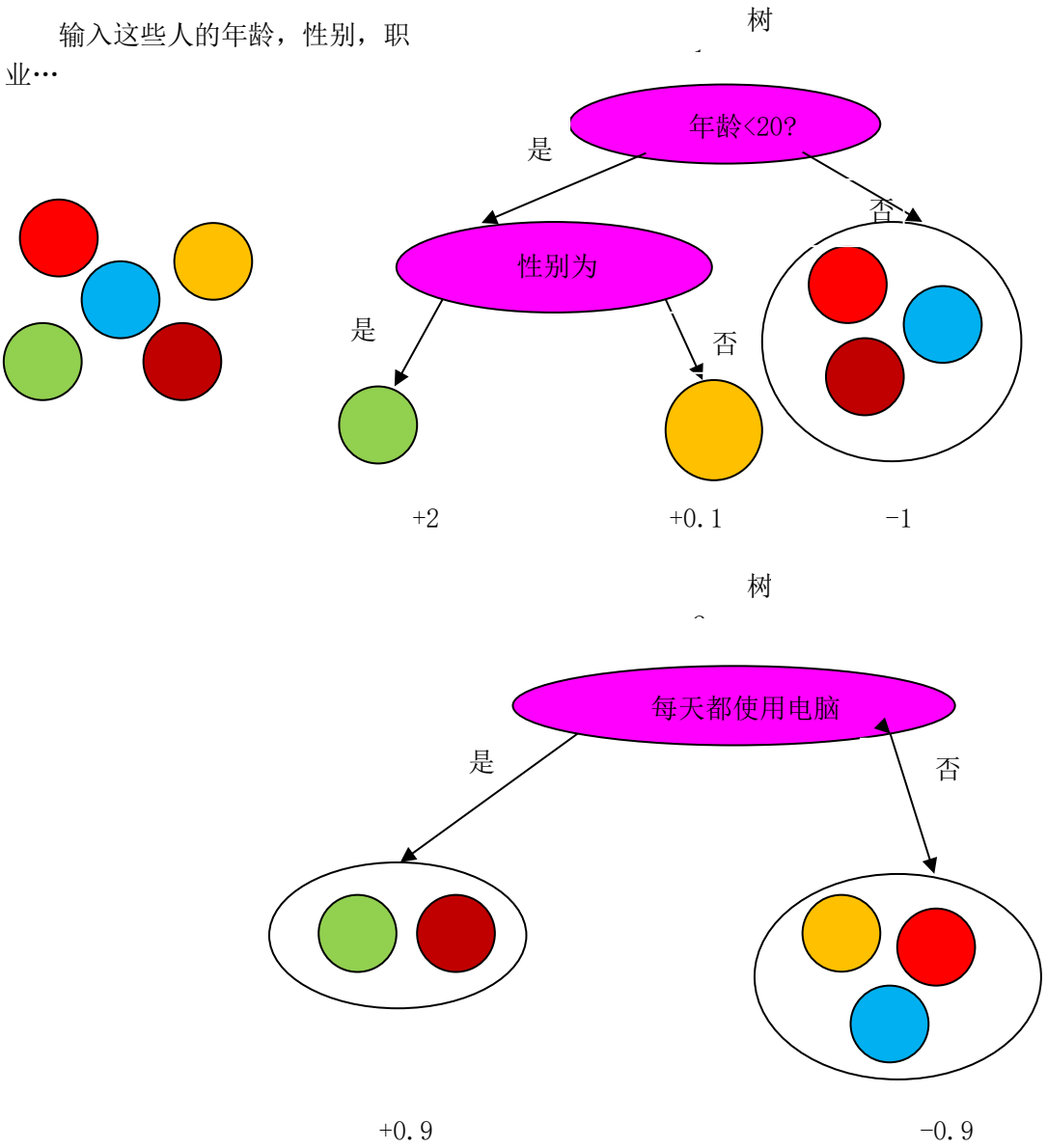


图 5

根据决策树可得f(绿)=2+0.9=2.9；而f(蓝)=-1-0.9=-1.9。当然实际xgboost算法中存在防止过拟合的参数限制树的分裂长度以及惩罚项。

1.2.2模型调参

直接调用xgboost库，用train()方法进行建模。其中有大量参数可以对模型进行调优。

XGBoost的参数可简单分为三类：

一是通用参数：宏观控制包括弱学习器类型，运行信息打印，线程数等。

二是Booster运算参数：主要是通过修改Booster的数学参数控制每一步的

booster(tree/regression)，可以调控模型的效果和计算代价。是调参主要对象。

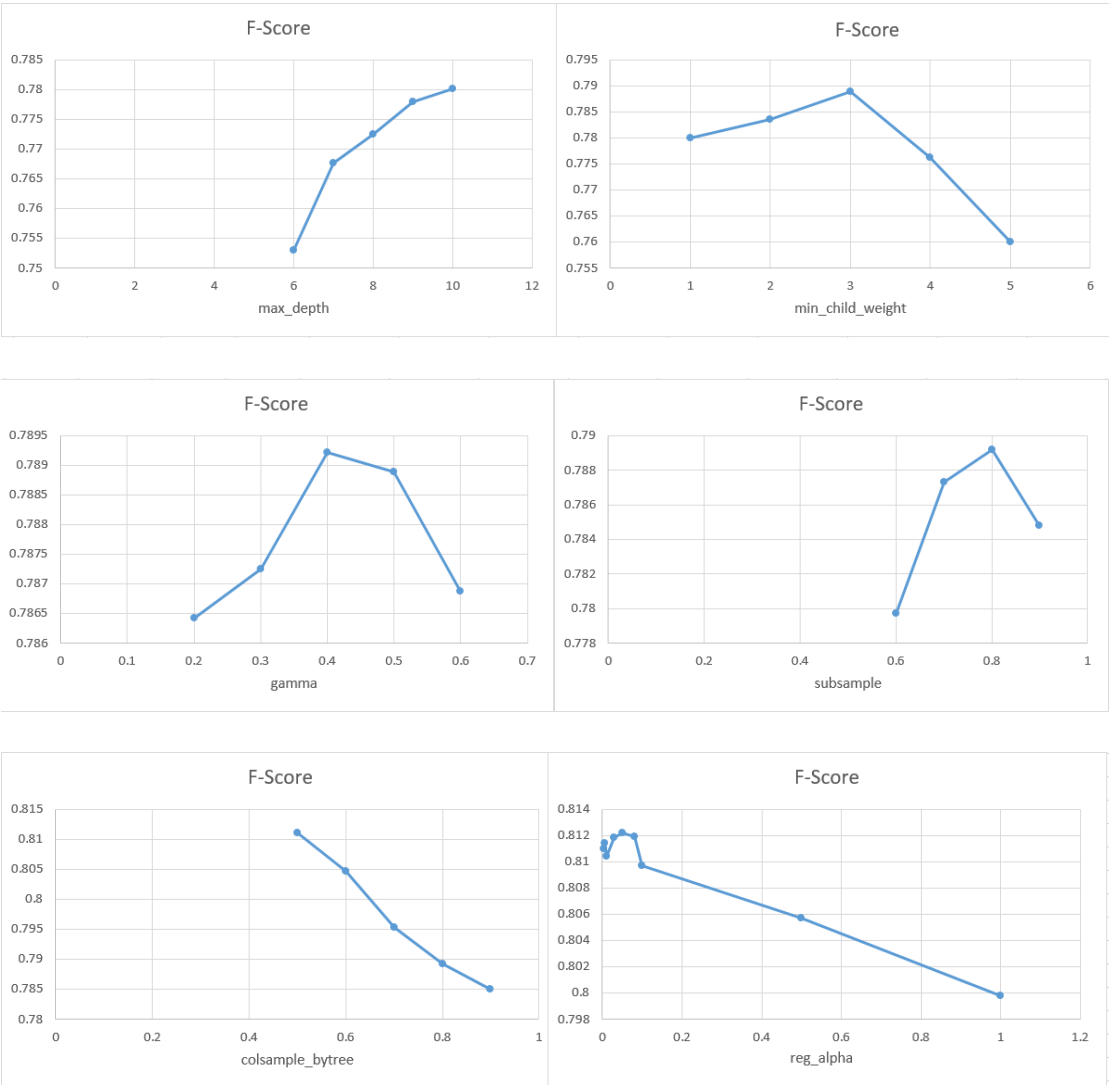
三是学习目标参数：控制函数类型，比如二分类，多分类等。

具体调参思想是基于‘贪婪’的。即先一个参数进行调参，获得最优解后，该参数保持不变。对下一个参数重复上一过程直至获得所有参数最优解。进行调参的参数顺序为下表从左到右：

表 1、参数顺序

max_depth	min_child_weight	gamma	subsample	colsample_bytree	reg_alpha	reg_lambda	learning_rate	num_boost_round
10	3	0.4	0.8	0.5	0.05	0.1	0.4	40

F-Score随参数变换折线图依次为：



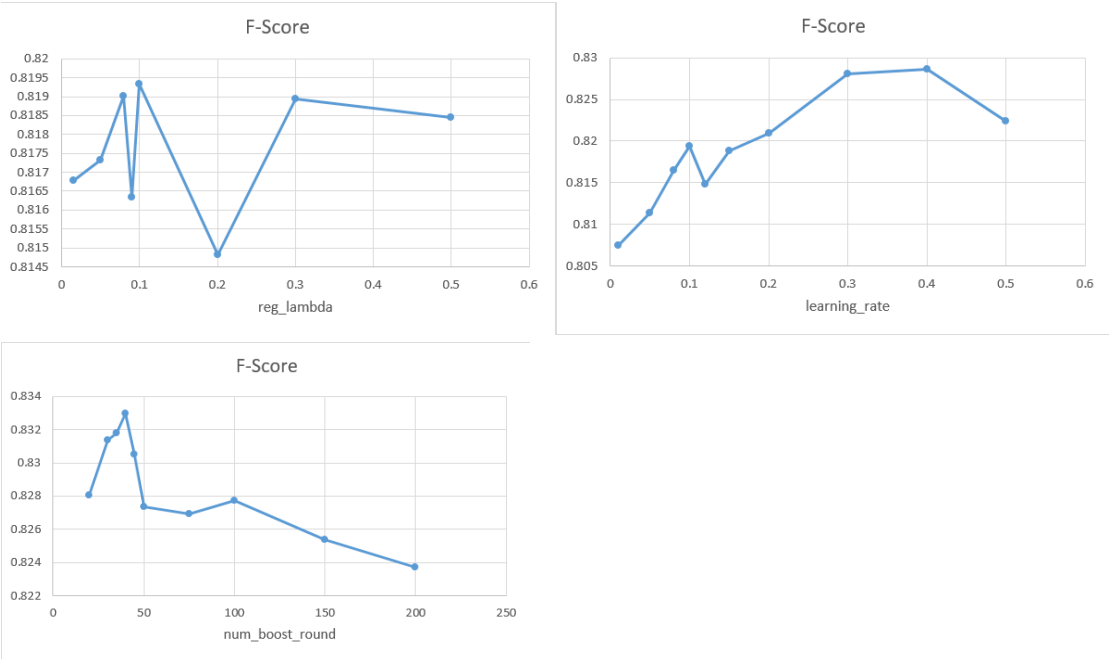


图 6、F-Score随参数变化折线图

可以看出reg_alpha与reg_lambda参数在无规则波动，调参意义不大，而max_depthcol与sample_bytree虽然为单调趋势，考虑到样本过拟合问题，限定在了常用范围内。

1.3模型评价

1.3.1建立混淆矩阵

为了评价精度，我们需要一种可视化工具。基于之前采用了xgboost模型，这里建立特别用于监督学习的混淆矩阵。混淆矩阵的每一列代表了预测类别 ，每一列的总数表示预测为该类别的数据的数目;每一行代表了数据的真实归属类别 ，每一行的数据总数表示。该类别的数据实例的数目。每一列中的数值表示真实数据被预测为该类的数目。

1.3.2计算查准率和查全率

由于题干中已经给出F-score的方法对分类方法进行评价，公式为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

我们下一步只需要根据混淆矩阵得出其查全率和查准率，即可得到其F-score，从而对我们的分类方法进行准确的评价。下图为调优过后的结果：

```
[02:55:16] 4604x76001 matrix with 313183 entries loaded from train.txt
[02:55:16] 4606x76001 matrix with 291183 entries loaded from exam.txt
<xgboost.core.DMatrix object at 0x000001C91A3DE908>
[[815.  18.  19.  27.  53.  63.   9.]
 [ 34. 415.   2.   2.   8.   6.   2.]
 [ 45.   5. 209.   6.  20.  18.   4.]
 [ 17.   2.   0. 758.  11.   6.   0.]
 [ 24.   0.   1.  21. 894.   8.  37.]
 [ 99.   6.  11.  20.  35. 419.  18.]
 [ 25.   0.   1.   8.  25.  16. 364.]]
查全率:
[0.8117529880478087, 0.8848614072494669, 0.6807817589576547, 0.9546599496221663,
 0.9076142131979695, 0.6891447368421053, 0.8291571753986332]
查准率:
[0.7695939565627951, 0.9304932735426009, 0.8600823045267489, 0.9002375296912114,
 0.8546845124282982, 0.7817164179104478, 0.8387096774193549]
F-Score:0.832949
```

图 7、调优结果

2、热点问题提取

热点问题的提取，没有可供参照的分类数据集，且结合问题背景可知，每隔一段时间，热点问题都会发生难以预测的变化。如果使用监督类算法，显然需要能够囊括未来可能发生的所有问题的已分类数据集作为训练集，这显然是难以实现的，因此我们不得不选择无监督的聚类算法来解决问题。而常见的聚类算法可简单分为三大类：一是基于划分（以K-means类算法为代表）、二是基于层次（BIRCH算法、CURE算法等）、三是基于密度（DBSCAN算法、OPTICS算法等）。K-means算法的改进型算法速度精度都不错，但聚类中心数量K在我们的问题中未知，只能采用交叉验证的方法求取，如果定期进行热点问题提取，调参代价过大。基于层次的方法计算量极大，只适合小样本聚类。为了应对文本样本的高维度，大样本数量的特点，只能选择基于密度的聚类算法，本文中采用的是DBSCAN算法[2]。

2.1数据预处理

数据预处理的过程在1.1中已有详细介绍，而第二题中我们的思路基本一致，做法大同小异，值得一提的是，我们所选用的sklearn库的sklearn.cluster.DBSCAN()的数据接口与第一题中的xgboost算法不同，不能采用稀疏矩阵与内置的DMarix格式，只能采用numpy.array格式的高维矩阵，因此过程量巨大。将该tf-idf高维矩阵保存为csv格式属性如下图：

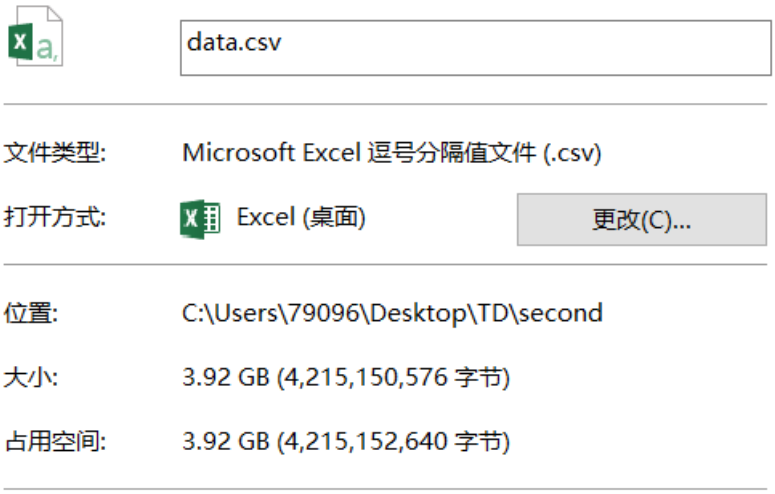


图 8、csv格式属性

该过程量无法打包进上交附件中，但可以通过运行second.py获得。由于体积原因无法用通常的文本处理软件打开，不过其格式就是简单的numpy.array数组形式。

2.2建立模型

DBSCAN算法的模型，实质是根据向量空间中的样本密度进行聚类，因此得到了一个显著优于传统K-means算法的特性，可以获得非凸聚类结果，即其聚类的簇的形状不一定为类圆形，而是密度分布决定簇的形状，例如二维空间中可以得到环形，月牙形的聚类簇，这是K-means算法无法做到的。

2.2.1模型原理

首先在样本空间中随机选择一个点，以这个点为圆心画一个半径为esp的圆，如果这个圆中包含的样本点大于等于我们所规定的最小样本点数量min_sample，那么就在包含的样本点中随机选择一个点作为下一个圆心，不断重复，直至新生成的圆中包含的样本点数量少于min_sample，那么这一个圆心就是边界点，最初的圆心为核心点，如果核心点没有满足min_sample的数量，那么称其为离群点，也就是噪声样本。二维空间向量原理示意图如下：

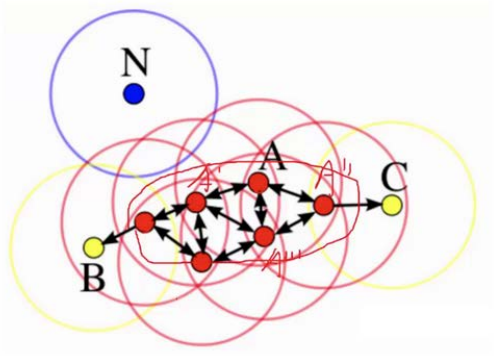


图 9、二维空间向量原理示意图[2]

下面我们利用naftaliharris网站[3]提供的DBSCAN可视化模拟结果展示DBSCAN算法的优越之处，如下图

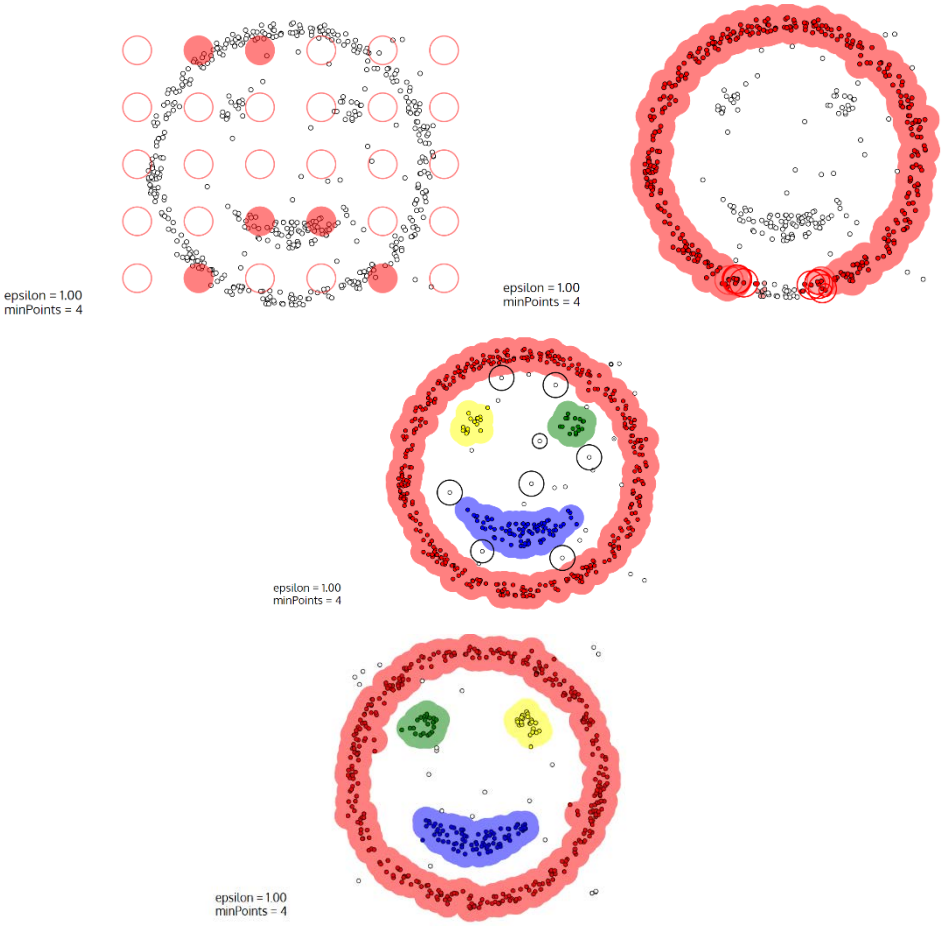


图 10、可视化模拟结果展示[3]

对于这个笑脸图形，只有基于密度聚类算法能够很好拟合。

2.2.2模型实现

通过调用sklearn库的sklearn.cluster.DBSCAN()函数[4]，进行建模。由于数据量巨大，每次建模花费的时间巨大。在本人i7-7700HQ+8GB RAM的平台上每次建模花费时间都在40分钟左右。

```
In [2]: runfile('C:/Users/79096/Desktop/TD/second/DBSCAN_model.py', wdir='C:/Users/
79096/Desktop/TD/second')
开始运行：
2020-05-06 18:21:25
向量加载完毕
2020-05-06 18:25:03
模型训练完毕
2020-05-06 18:59:43
```

图 11

而最终生成的model文件也很大。如下图所示：

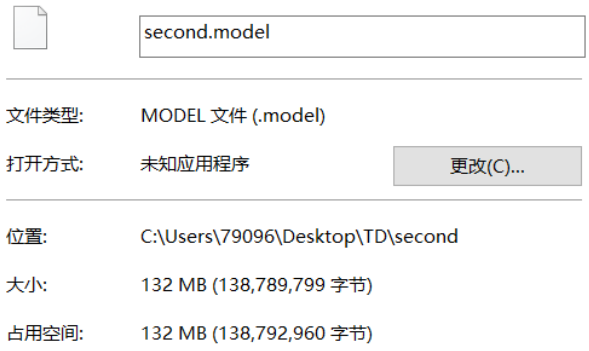


图 12

2.2.3模型结果

将聚类结果的留言主题按照每类一列保存为csv文件。

	A	B	C	D	E
1	0	1	2	3	4
2	关于拆除聚美龙楚在西地省商学院宿舍旁安装	投诉滨河苑针对广铁职工购房的霸王规定	A市万家丽南路丽发新城居民区附近搅拌站扰	关于A7县恒基凯旋门万婴格林幼儿园办普惠园	A3区西湖街道茶场村五组什么时候能启动征地
3	关于拆除西地省商学院宿舍、A3区一小旁变压	关于伊景园滨河苑捆绑销售车位的维权投诉	投诉A2区丽发新城附近建搅拌站噪音扰民	A3区中海国际社区幼儿园无法满足实龄幼儿就	请问A3区西湖街道茶场村五组是如何规划的
4	A3区一小学门口违规建设的日咀鹤停车场违	车位捆绑违规销售	丽发新城小区旁边建搅拌站	A3区高心麓城小区配套幼儿园无证办学且拒绝	A3区西湖街道茶场村五组何时才能启动拆迁?
5	关于拆除西地省商学院宿舍、A3区一小旁变压	关于广铁集团铁路职工定向商品房伊景园滨河	A市丽发新城违建搅拌站，彻夜施工扰民污染	咨询A7县金科时代中心小区配套幼儿园的问题	A3区西湖街道茶场村五组是如何规划的?
6	日咀鹤停车场美聚龙楚公司充电桩项目存在隐	投诉A市伊景园滨河苑捆绑车位销售	A市丽发小区建搅拌站，噪音污染严重	A7县深业睿城幼儿园迟迟不开园	A3区西湖街道茶场村五组何时启动拆迁?
7		关于A市武广新城违法捆绑销售车位的投诉	A2区丽发新城附近修建搅拌厂，严重污染环境	反映A7县恒基凯旋门小区配套幼儿园公办或普	A3区西湖街道茶场村五组不属于拆迁部分的村
8		咨询人防车位产权的问题	A市丽发新城小区侧面建设混泥土搅拌站，粉	A4区君悦幼儿园为什么一直不改为普惠或公	A3区西湖街道茶场村五组什么时候能拆迁
9		家里本来就困难,还要捆绑买卖车位	投诉丽发新城小区附近违建搅拌站噪音扰民	A7县松雅安置区的配套公办幼儿园在哪里?	请问A市西湖街道茶场村五组这边有什么规划
10		伊景园滨河苑捆绑车位销售合法吗？！	A2区丽发新城附近违规乱建混凝土搅拌站谁来	咨询A3区保利西海岸小学的开学问题	咨询A3区西湖街道茶场村五组的拆迁规划
11		坚决反对伊景园滨河苑强制捆绑销售车位	A2区丽发新城小区旁边的搅拌厂是否合法经营	A7县星沙恒大翡翠华庭小区内幼儿园收费是否	反映A3区西湖街道茶场村拆迁问题
12		伊景园滨河苑强行捆绑车位销售给业主	搅拌站大量加工砂石料噪音污水影响丽发新城	咨询A7县普惠性幼儿园招生问题	
13		A市武广新城坑客户购房金额并且捆绑销售车	丽发新城小区旁的搅拌站严重影响生活	反映A7县恒基凯旋门小区配套幼儿园公办或者	
14		伊景园滨河苑项目绑定车位出售是否合法合规	噪音、灰尘污染的A2区丽发新城附近环保部门	咨询A7县山水湾小区幼儿园性质问题	
15		询问A7县圣力华苑车位的问题	A2区丽发新城附近修建搅拌厂严重影响睡眠	举报A3区长房云时代小区幼儿园无招标公告，	
16		广铁集团要求员工购房时必须同时购买车位	投诉A2区丽发新城附近建搅拌站噪音扰民	A3区高心麓城芭比罗幼儿园还在办学	
17		投诉A市伊景园滨河苑定向限价商品房违规涨	暮云街道丽发新城小区附近水泥搅拌站非法经	坚决反对A3区巴罗比幼儿园继续留在高心麓城	
		关于房伊景园滨河苑销	丽发新城小区附近建搅	A7县海德公园楼盘没有	

图 13、csv文件展示图

当参数设定理想时，聚类效果不错。经历多次尝试，我们选择esp=1.11，min_sample=5 。

2.3结果处理

只用DBSCAN处理的结果是很粗糙的。由于DBSCAN算法原理的非凸性易知，两个样本之间的相似度不高但是具有某部分强烈关联的特点时可能被归为一类，如下图：

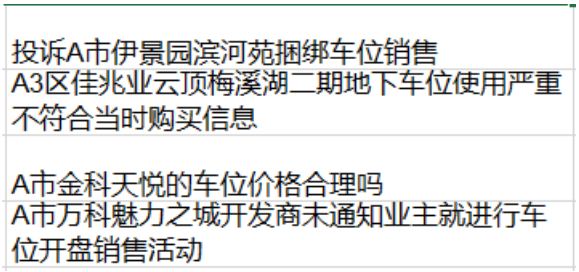


图 14

这三个样本都因为问题与车位相关而被分为一类，所以我们必须对结果进行二次处理。具体思路如下：

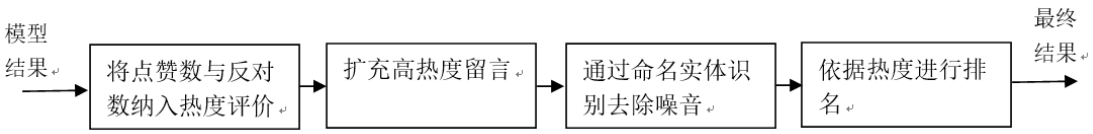


图 15

2.3.1热度评价

首先综合结果与题目背景，我们对热度评价量化如下：

1）每条留言基础分为1分

2）每一个点赞为留言加0.02分

3）每一个反对为留言减0.02分

然后考虑到部分高赞留言可能没有类似问题，或类似问题不满足min_sample数量，我们将未进入模型结果的留言单独记为一类问题加入模型结果中，这个过程中我们只对点赞数超过250的留言进行了操作，因为它的热度刚好满足模型分类结果的最低热度即250*0.02=5。

2.3.2命名实体识别

命名实体识别的实现是极其复杂的，因此我们选择了java开源项目HanLP[5]的封装python库pyhanlp进行实现，可结果差强人意。对于样本中的重要命名实体识别目标：小区名的正确识别率极低。对于这种情况开源作者的建议是针对业务要求添加自定义词典，结合问题背景，我们有理由任务，业务甲方能够为我们提供A市小区的注册名单，基于这份名单我们能够通过自定义词典来修正分词结果。而在实际问题中，我们未能得到相关数据，但是我们对数据进行了缩减，只按照顺序提取目前热度的前10 的样本，手动创建里面包含的命名实体，保存为词典txt文件，添加到pyhanlp库的内置词典中并将优先级设置为除核心词典外的最优集。

图 16

该词典格式中，nsp为pyhanlp库未定义的词性类型，数字为统计词频，此处设为1不影响结果。

具体的筛选过程为提取每类问题中的nsp词性词汇，只保留其中出现次数最多的一种词汇，将不包含这一词汇的问题删除，得到最后的分类结果。以下是前五项问题的筛选留存结果：

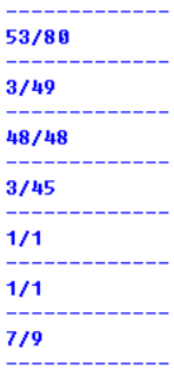


图 17、筛选留存结果

其中原热度评价第二与第四的问题留存率极低。通过观察我们发现出现这一结果的原因是，这两类问题的主要人群/地点差异很大，但面临的是相同问题，以原热度第二问题样本列表展示如下：

B
1
关于A7县恒基凯旋门万婴格林幼儿园办普惠园的咨询
A3区中海国际社区幼儿园无法满足实龄幼儿就读需求
咨询A4区万国三期幼儿园的学位问题
A2区中建A1区嘉苑业主有房有户却不能就读仰天湖小学
咨询A4区万国城三期幼儿园学位问题
A3区高心麓城小区配套幼儿园无证办学且拒绝转成普惠制幼儿园
A市美联幼儿园什么时候能招生
咨询A7县金科时代中心小区配套幼儿园的问题
希望A市在仰天湖中建小学校门口修建人行天桥或地下通道
A7县深业睿城幼儿园迟迟不开园
反映A7县恒基凯旋门小区配套幼儿园公办或普惠问题
反映A4区万国城三期幼儿园学位问题
A4区君悦幼儿园为什么一直不改为普惠或公办?
A7县松雅安置区的配套公办幼儿园在哪里?
咨询A3区保利西海岸小学的开学问题
A7县星沙恒大翡翠华庭小区内幼儿园收费是否合理?
A3区中海业主的小孩都没办法上中海幼儿园
咨询A7县普惠性幼儿园招生问题
请解决A2区中建A1区嘉苑的业主小孩入学问题
反映A7县恒基凯旋门小区配套幼儿园公办或者普惠问题
反映2019年六艺天骄A7县松雅湖幼儿园招生问题
咨询A7县山水湾小区幼儿园性质问题
举报A3区长房云时代小区幼儿园无招标公告，无教学历史
A3区高心麓城芭比罗幼儿园还在办学
反映西地省农业大学的孩子入读幼儿园一事
坚决反对A3区巴巴比幼儿园继续留在高心麓城小区内办学

图 18

可见其问题特点高度相似，但问题人群与地域差异很大，属于DBSCAN算法特点

造成，在预料之中。显然经过此次筛选原第一、第三、第五、第六、第七的问题成为了最终结果的热度前五名。最终结果按照题目要求写入相应文件。

3、答复意见的评价

3.1对评价指标的选取

对于一个回复意见，群众最关心的便是自己的问题是否得到正面的的答复，而没有跳到其他的问题上去，即回复意见是否于留言相关，因此相关性便是一项重要的指标。

作为政府的官方答复，需要规范专业，因此规范性也是评价回复的不可或缺的指标。政府的答复意见不能直接答复“您的问题已经解决”或“您的问题不能解决”，需要有法规文件、会议精神等权威支持，这里我们用可解释性表示。而群众得到回复的时间也是影响群众满意度的一项的指标，这里我们用时效性来表示。

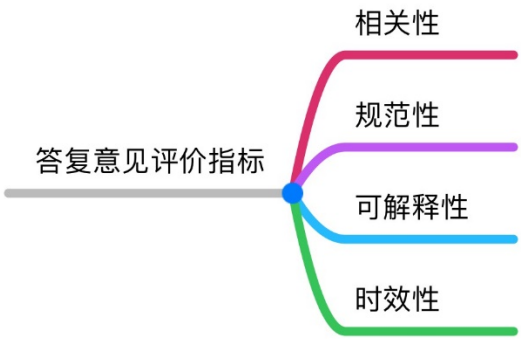


图 19、评价指标

3.2相关性的量化实现

评判回复意见的相关性，需要对留言和答复的相似度进行比对，流程如下：

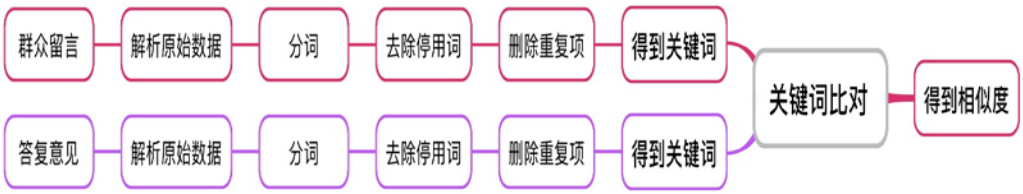


图 20、相似度流程图

其中前几步的原理已在问题1（群众留言分类）中解释过，这里我们重点分析一下“关键词比对”这一步。其原理大致如下：

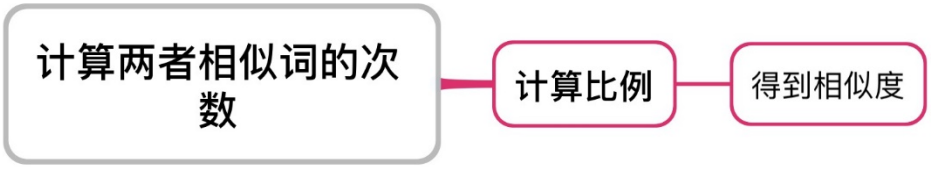


图 21

若回复意见和群众留言相关，则经过处理后的关键词必然符合相同词、近义词、反义词、相似词中的一样，故可以通过统计词频来得到相似度。得到相似度的情况如下图所示：

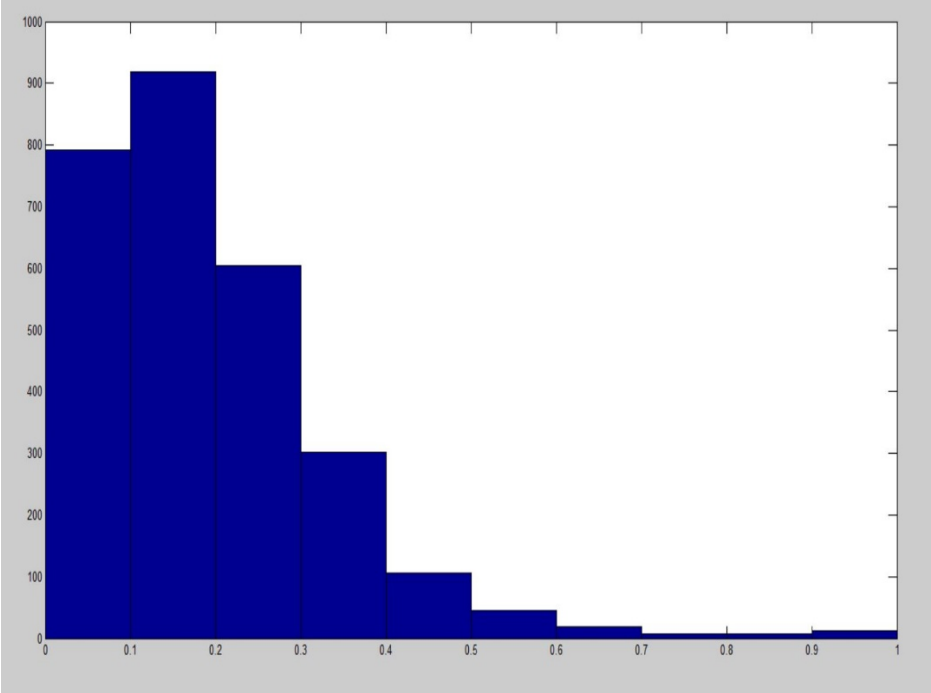


图 22、相似度直方图

可以看出相似度集中在0.5以后，因此对于相似度0.5以上的答复，我们认为是必然相关的，即给予其相关性的满分。而对其他的相似度低于0.5的答复，对其相关性给予1-10分的详细分级，分数越高的说明其相关程度越高。

3.3规范性的量化实现

对于规范性，我们取两种量化等级，即规范和不规范。下面是对规范性的评判标准：

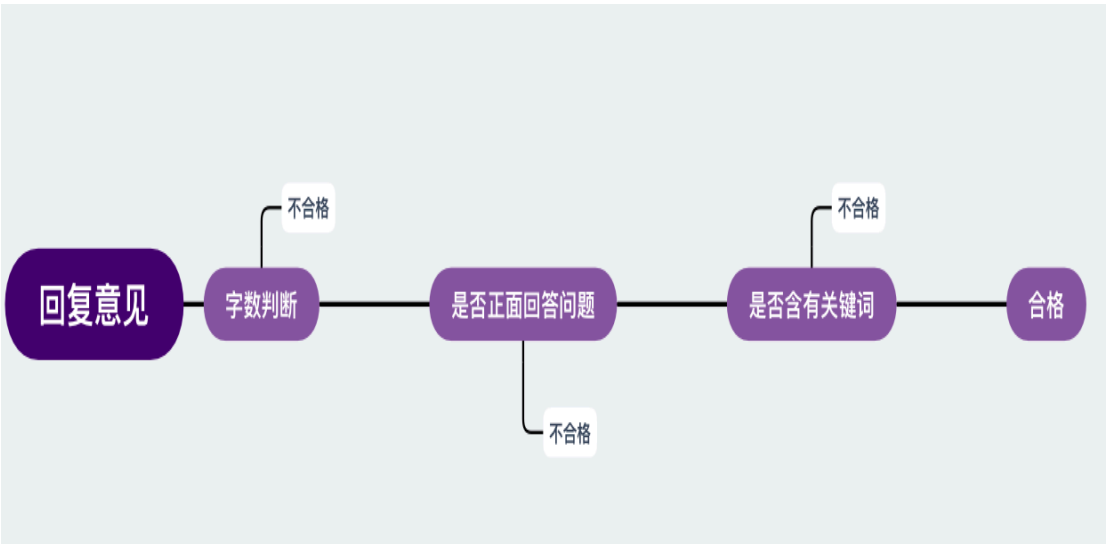


图 23、规范性的评判标准

因为政府的回复是对群众生活问题中的解释，故字数过少的不合格。未正面回答问题，包含“转交”等类似字样的，也不合格。没有“您好”等类似礼貌用语关键词的也不合格。

3.4可解释性的量化实现

对于答复的可解释性，我们取两个量化等级,即“是”和“否”。“是”表示解释清楚，“否”表示未解释清楚。对于一个政府性质的留言，其说服力在于是否有已有的规章制度、法律文件、或会议精神对其进行解释。因此我们可以通过检索关键词的方法来判断其是否引用了相关的文件与法规，从而得到文件的可解释性。

3.5时效性的量化实现

对于政府答复的时效性，我们取五个量化等级，从快到慢为“迅速”、“及时”、“一般”、“迟缓”。首先我们需要计算回复时间和留言时间的时间差。这里为了计算简便，我们进行了简化。只计算了天数，忽视了小时和分钟的影响。可以得到下图：

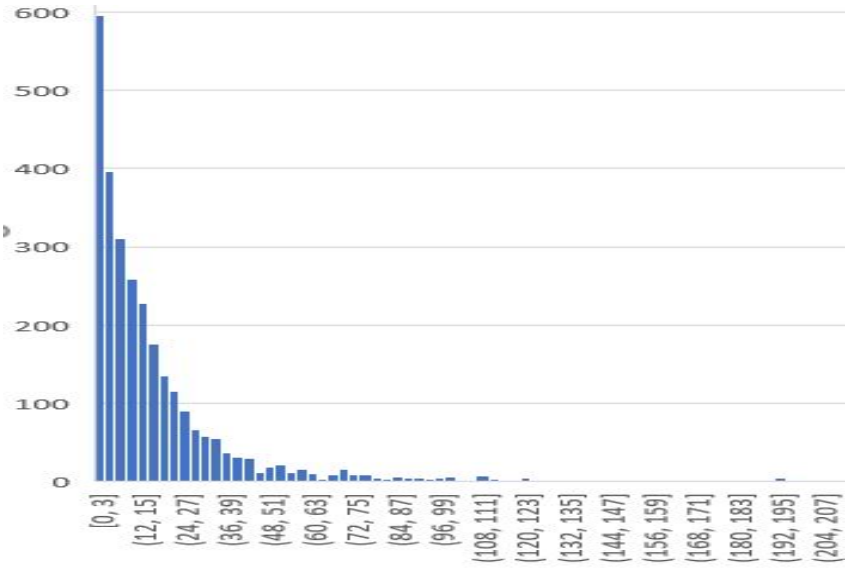


图 24、回复时间直方图

可以看出大部分集中在一个月之内，考虑到可能处理不同的指标，以及群众希望得到较快回复的心理。因此我们取0-10%为迅速，10%-30%为及时，30%-60%为一般，60%之后为迟缓。从而对回复时间一一量化评价，得到回复的时效性。

3.6加权求总分

对于一个留言的回复，留言者最关心就是其问题是否得到正面的回复和有法理支撑的解释。

及时的得到解答也很重要，因此时效性排名第三。最后便是规范性的问题。因此对于评价指标取下图的各个权重，来得到一个总分。

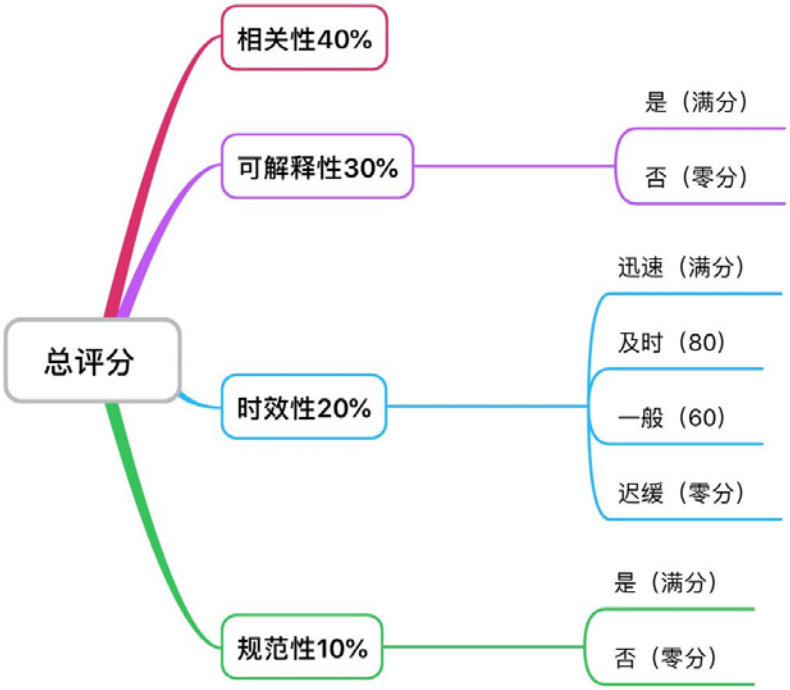


图 25、评价指标权重

3.7处理结果

经过对回复意见的处理量化，得到下图所示的Excel表格：

答复意见	答复时间	次回复时	文本相似度	相关性	复是否规	可解释性	时效性	总评分
现将网友	########	15	0.238095	6	是	是	迟缓	64
网友“AO(	########	14	0.068182	2	是	否	迟缓	18
市民同志:	########	14	0.244898	6	是	是	迟缓	64
网友“AO(	########	14	0.5	10	是	是	迟缓	80
网友“AO(	########	15	1	10	是	否	迟缓	50
网友“AO(	########	31	0.078947	2	是	否	迟缓	18
网友“AO(	########	40	0.120879	4	是	是	迟缓	56
网友“UU(	########	28	0.226415	6	是	是	迟缓	64
网友“UU(	########	16	0.246154	6	是	是	迟缓	64
网友“UU(	########	16	0.157303	4	是	是	迟缓	56
网友“UU(	########	70	0.311475	8	是	是	迟缓	72
网友“UU(	########	30	0.538462	10	是	是	迟缓	80
网友“UU(	########	16	0.078431	2	是	是	迟缓	48
网友“UU(	########	5	0.041096	2	是	否	一般	30

图 26、处理结果

四、参考文献

[1]: XGBoost 重要参数(调参使用)

[2]: DBSCAN聚类算法——机器学习（理论+图解+python代码）

[3]: X Visualizing DBSCAN Clustering

[4]: 用scikit-learn学习DBSCAN聚类

[5]: Hanlp自然语言处理工具作者hankcs的Github