关于政务文本信息的挖掘应用

摘要：近年来，随着社会信息化、智能化水平的不断提高，社会的管理水平和工作效率有着极大地提高。政府作为大量数据文本接收载体，需要对与日俱增的文本数据进行划分、整理，为减少人工和提高效率，建立智能政务系统成了当务之急。在这里需要解决下列问题。

针对问题一：本文首先将附件 1、附件 2 中的非结构化数据进行数值化处理、中文分词及停用词过滤等数据预处理，然后根据 TF-IDF 算法提取分类特征词，进行向量化，再通过 KNN 算法和 SVM 算法分别对留言信息进行划分并通过召回率、准确率、F-Score 进行分类评估。

针对问题二：为从群众留言信息中挖掘出热点问题，将处理过的信息进行 DBSCAN 密度聚类，通过参数枚举，以调整兰德指数和轮廓系数为指标得出最优聚类参数，再根据点赞数、发表时间、记录数构建热度指标模型，得到 5 大热点问题，再根据结果进行关键信息提取，输出要求结果。

针对问题三：首先对于问题进行分类，选取每类优秀回复，构建词汇库，利用 word2vec 算法对未评价回复进行训练, 形成特征词汇库，将回复信息的关键词与优质回复的词汇库进行对比分析。相似度高的，相较于回复评价也相对高。

关键词：文本挖掘；TF-IDF；jieba；DBSCAN 算法；轮廓系数；ARI；TextRank；

Abstract:In recent years, with the continuous improvement of the level of social informatization and intelligence, the level of social management and work efficiency has been greatly improved.As the receiving carrier of a large number of data texts, the government needs to divide and sort out the increasing amount of text data. In order to reduce labor and improve efficiency, it is imperative to establish an intelligent government affairs system.The following problems need to be solved here.

In view of the problem a: at first, this paper will be unstructured data in attachment 1, attachment 2 numerical processing, Chinese word segmentation and stop words filtering data preprocessing, and then according to the TF - IDF algorithm to extract key words, vectorization, again through the KNN algorithm and SVM algorithm for message information respectively divided and through the recall ratio and accuracy, the F - classified Score evaluation.

For problem 2: for the excavated from the message information hot problems, the processed information to the density of DBSCAN clustering, enumerated by parameters, in order to adjust the rand index and coefficient of contour for index, it is concluded that the optimal clustering parameters according to the number of thumb up, again published time heat index model is established, the number of records to get five hot issues, and then according to the results of key information extraction, and result output requirements.

For problem 3: first, classify the problem, select each type of excellent reply, build a vocabulary, use word2vec algorithm to train the unevaluated reply, form a characteristic vocabulary, and compare and analyze the key words of reply information with the vocabulary of high-quality reply.Those with a high degree of similarity were also rated higher than those with a high degree of similarity.

Key words: text mining;TF-IDF;Jieba;DBSCAN;Contour coefficient;ARI;TextRank

目录

1、挖掘目标

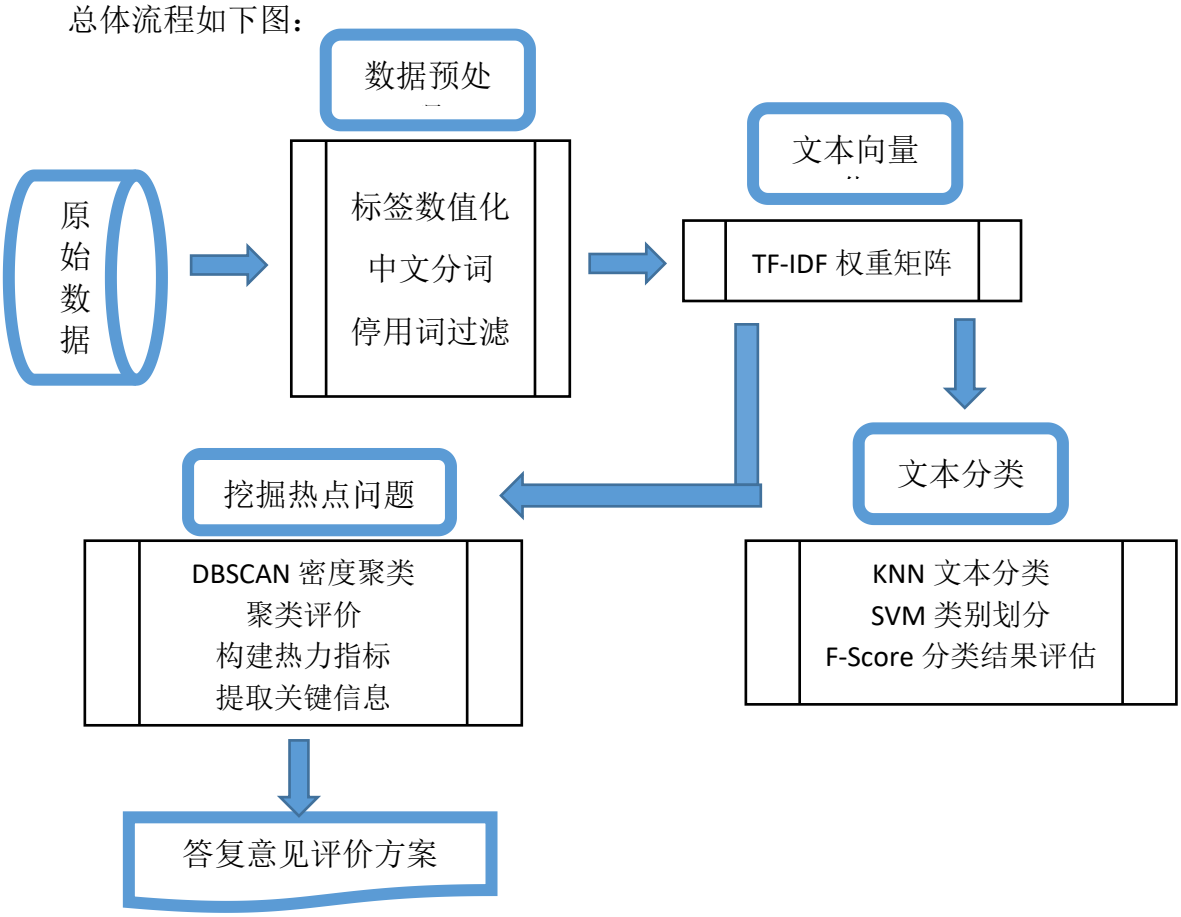
近年来，随着社会信息化、智能化水平的不断提高，社会的管理水平和工作效率有着极大地提高。政府作为大量数据文本接收载体，需要对与日俱增的文本数据进行划分、整理，为减少人工和提高效率，建立智能政务系统成为当务之急。

本次建模目的是利用互联网公开来源的群众问政留言系统，以及相关部门对部分群众留言的答复意见，包含了非结构和结构化文本书籍，对相关数据进行数据预处理后，对群众留言的基本信息进行自动分类体系划分，分别采取 TF-IDF 加权法和 KNN、SVM 算法实现对群众留言的自动标记，并利用 F-score 对分类结果进行评估；另一方面，为从群众留言信息中挖掘出热点问题，将信息进行 DBSCAN 密度聚类，根据点赞数、发表时间、记录数进行热度排名，并导出相关信息；其次是根据群众留言与回复信息进行答复意见评估，给出实现方案。

2、分析方法和过程

2.1 总体流程、步骤

总体流程如下图：



步骤一：进行数据预处理，对分类体系进行结构化处理，对附件 2 的数据进行中文库分词、停用词过滤，以便建立模型。

步骤二：进行文本向量化，利用 TF-IDF 算法对处理好的数据进行加权，构建数据矩阵。

步骤三：文本分类，根据 KNN、SVM 算法对未分类的数据信息进行类别划分，使用 F-Score 对于分类结果进行评估。

步骤四：利用文本聚类挖掘热点问题，利用数据矩阵，使用 DBSCAN 算法进行密度聚类，通过聚类评估，得到最优参数，构建热力指数模型，进行热力排名，得到五大热点问题，并提取关键信息，输出结果。

步骤五：答复意见评价分案。

2.2 群众留言分类

2.2.1 数据预处理

（1）数据描述

分析所得数据，基本上的字段都为文本格式，附件一为内容分类标签体系，需要对分类标签进行结构化处理；附件二为用户留言分类信息，包含留言时间、人工划分的一级类别，对于文本分类的主要依据是留言主题和留言详情；附件三是留言详情，在附件二的留言信息中加入了点赞数、反对数；附件四是相关部门对留言的答复意见，主要根据留言主题、留言详情和答复意见的文本数据以及答复响应时间进行答复评估。

（2）标签数值化

对文本进行分类，需要对分类标签数值化，也就是对附表 1 的分类体系进行数值化处理，总计 15 个一级分类标签、103 个二级分类、517 个三级分类标签，对它们进行数值化：

分类数值化对照表

一级分类	数值化
城乡建设	1
党内政务	2

二级分类（部分）	数值化
安全生产	101

国土资源	3
环境保护	4
经济管理	5
纪检监察	6
交通运输	7
科技与信息产业	8
民政	9
农村农业	10
贸易旅游	11
卫生计生	12
政法	13
教育文体	14
劳动和社会保证	15

城市建设和市政管理	102
城乡规划	103
党的建设	201
港澳台侨	202

三级分类（部分）	数值化
事故处理	10101
违章建筑	10302
港澳事务	20203
病退及提前退休人员待遇	150806
辞职辞退	150606

完成对分类属性的数值化，再根据得到的数值化标签，对附件二中的人工分类结果进行标记。

（3）中文分词

相对于英文文本依靠空格等进行分词，中文文本分词比较复杂，需要借助于外部资源，jieba 结巴分词 (Jieba) 作为一个强大的分词库，能够很好进行中文分词处理。jieba 分词支持三种分词模式:

1) 精确模式, 此模式试图以最高精度来对句子进行划分, 适用于文本分析;

2) 全模式, 此模式可以扫描出句中全部可成词的词语, 并且速度很快, 但它并不可以解决歧义问题;

3) 搜索引擎模式, 此模式基于精确模式对长词在进行切分, 可以将此模式用于搜索引擎分词[1]。

针对分析的文本数据，采用精确模式，但是直接对文本进行分词会存在大量的标点符号和无意义词，所以应该对分词结果进行进一步过滤。

（4）停用词过滤

停用词表有自动抽取和人工构造两种方式，由于文本数据专业化程度不高，含新词不多，在这里直接引入了标准的经典停用词表和 df 计算过滤，对于高出或低于阈值的文本频率的词进行进一步过滤，在这里过滤掉在文档中出现率大于

0.75 和小于 0.002 的词。同时基于留言主题和留言详情的重要程度不一，对这两列分开处理，并统计其词频。

留言主题: 495 [['A', '市', '西湖', '建筑', '集团', '占', '道', '施工', '安全隐患'], ['A', '市', '在水一方', '大厦', '人为', '烂尾', '多年', '安全隐患'], ['A3', '区', '杜鹃', '文苑', '小区', '外', '非法', '汽车', '检测站', '开业'], ['民工', 'A6', '区明发', '国际', '工地', '受伤', '工', '地方', '拒绝', '支付', '医疗费'], ['K8', '县', '丁字街', '商户', '乱', '摆摊'], ['K8', '县', '南门', '街', '干净', '整洁', '几天', '样子'], ['K8', '县冷', '江东', '路', '蓝波', '旺', '酒店', '外墙', '装修', '无人', '施工'], ['K8', '县', '九亿', '广场', '公厕', '安装', '照明灯'], ['K4', '县', '石期', '市镇', '农贸市场', '旁边', '公厕', '旱厕', '脏', '乱', '差'], ['K', '市域', '轨道交通', '规划', '建议'], ['K', '市域', '轨道交通', '规划', '建议'], ['请问', 'A', '市', '乘坐', '地跌', '爱心卡'], ['地铁', '号线', '施工', '导致', 'A', '市锦楚', '国际', '星城', '小区', '三期', '一个月', '停电', '10', '来次'], ['A6', '区涸', '紫', '郡', '用电', '解决'], ['A', '市锦楚', '国际', '新城', '月份', '停电'], ['咨询', 'A', '市', '楼盘', '供暖', '一事'], ['A', '市能', '北方', '居民小区', '统一', '建设', '供暖', '设备'], ['A', '市', '供暖'], ['K9', '县', '坐', '公交车', '元'], ['希望', 'K6', '县', '路', '路', '公交车', '延迟', '收班', '时间'], ['K6', '县', '公交车', '监控'], ['请求', '延后', 'L1', '区迎丰', '公园', '清晨', '路灯', '熄灯', '时间', '半小时'], ['L', '市中坡', '山', '公园', '内溜狗', '有损', '景区', '环境', '应', '严禁'], ['请', 'L', '市', '公园', '里', '门', '球场', '修好'], ['A5', '区雅礼洋湖', '实验', '中学', '垃圾箱', '长期', '未', '清理'], ['A4', '

部分结果显示

（5）TF-IDF 算法

TF-IDF (Term Frequency-inverse Document Frequency) 是一种加权统计算法, TF 指的是词出现的频率, IDF 是该词出现的文件在总文件里出现的反频率指数。该算法的主要思想为:如果某词在某类的文本中出现的频率很高, 但是这个词在其他类文本中出现的频率很低, 那么认为该词具有此类文本某些代表性的特征, 可用词对此类文本进行分类[2].

在某文件中词语 t_i 的 tf 公式如下，其中 N_{i} ， j 表示词 t_i 在文件中出现的总次数, 分母表示中文件中所有词出现的次数的总和。

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

对于某个词 t_i 的 idf 计算公式如下，其中,$|D|$ 是全部文件的总数, 分母表示中包含有 t_i 的文件的数目。

$$idf_t = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

某个词 t_i 的 $tfidf$ 就是该词的 tf 和 idf 进行相乘，计算公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

在这里重新利用已经进行初步过滤的结果进行加权计算，按照权重从大到小进行排序，抽取出每一类文本权重比较大的特征词进行进一步分类计算。

2.2.2KNN 算法

该算法的基本思路是:在给定文本后考虑在训练集文本中与该新文本距离最近的k篇文本数据,根据这k篇文本数据所属的类别判定新文本所属的类别[3],具体的算法步骤如:

1) 计算测试数据与各个训练数据之间的距离;
2) 按照距离的递增关系进行排序;
3) 选取距离最小的k个点;
4) 计算前k个点所在类别的出现频率;
5) 返回前k个点中出现频率最高的类别作为测试数据的预测分类。[4]

将已经进行人工分类的数据进行训练集划分，KNN 算法对训练集的依赖程度大，需要对训练集数据进行合理划分，训练集的分类结果也尽量全面，实现样本平衡。同时 KNN 算法计算量较大，需要计算未知文本到所有已知文本的距离。在全部数据里随机选择 100 条测试数据进行分类结果：

原标签:

[1 5 4 1 1 4 4 5 5 7 1 4 6 5 1 5 4 4 4 5 4 4 5 5 2 7 5 5 5 6 1 6 1 4 2 5 1
5 5 6 6 2 5 4 1 5 2 6 5 6 1 1 4 5 2 1 7 7 2 1 5 6 4 5 1 7 5 1 3 4 5 5 1 1
3 2 4 4 2 4 7 3 6 5 7 3 6 2 6 1 7 2 6 1 5 2 6 4 3 2]

KNN 预测标签结果:

[1 5 4 1 1 4 4 5 5 7 1 4 6 5 1 5 4 4 4 5 4 4 5 5 2 7 5 5 5 6 1 6 1 4 2 5 1
5 5 6 1 2 4 2 1 5 2 4 5 6 1 1 4 5 6 1 7 7 2 1 5 6 4 5 1 7 5 1 3 4 5 5 1 2
3 2 4 4 2 5 7 3 6 5 7 3 6 2 5 1 7 2 1 3 5 2 6 4 3 2]

2.2.3 SVM 算法

SVM(Support Vector Machines)——支持向量机和 KNN 算法一样是有监督的分类算法，能够很好解决非线性、高维度问题，是文本分类的重要算法。支持向量机大致可以分为三类：线性可分支持向量机、线性支持向量机以及非线性支持向量机这模型。

在这里，使用线性可分支持向量机（LinearSVC）。由于对于文本进行 15 个类型分类，需要在任意两类样本之间设计一个 SVM，因此 k 个类别的样本就需要设计 $\frac{k(k-1)}{2}$ 个 SVM。当对一个未知样本进行分类时，最后得票最多的类别即为该未知样本的类别。对于文本这类非线性数据，需要引入核函数，实现非线性数据的高维线性可分。在这里使用的是线性核，计算公式为：

$$K(x_i^T, x_j) = x_i^T x_j \text{ 、}$$

在 python 中，通过 sklearn 工具包可以是直接实现该算法，上文选择的 100 条测试数据进行 SVM 分类结果：

[1 5 4 7 1 4 4 5 5 7 1 4 6 5 1 5 4 4 4 5 4 4 5 5 2 7 5 5 5 1 1 6 1 4 2 5 1
5 5 6 1 2 4 1 1 5 2 6 5 6 1 1 4 5 2 1 7 7 2 1 5 6 4 5 1 7 5 1 3 4 5 5 1 1
3 1 4 4 2 5 7 3 6 5 7 3 6 2 6 1 7 2 1 4 5 2 6 4 3 2]

使用全部数据进行测试，显示结果如下：

```
true:
[5 6 3 ... 4 3 6]
pred:
[5 6 3 ... 4 3 6]
score: 0.9747722603310166
```

2.2.4 分类评估

通过两种算法的分类结果，分别从召回率、准确率、F-score 三个指标来评估两种分类算法效果。

召回率：是检索出的相关文本数和全部数据中所有的相关文本数的比率，衡量的是分类模型的查全率。

$$召回率 = \frac{检索的相关文本数}{系统所有文本总数}$$

准确率：是检索出的相关文本数与检索到的文本总数的比率，衡量的是分类模型的查准率。

$$准确率 = \frac{检索的相关文本数}{检索的所有文本总数}$$

F-Score：分类常用的评价标准，公式如下，P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

$$F\text{ - score} = \frac{1}{n}\sum \frac{2P_i R_i}{P_i + R_i}$$

准确率和召回率两者都是与分类效果呈现正相关的关系，但是两者往往相互影响，不能达到很高的值，相对而言，F-score 可以简单综合评价这两个指标。两种算法分类（随机选取 100 条测试数据）评估如下表：

KNN、SVM 评估指标表

	精确率	召回率	F-score
KNN	0.89802711907975	0.90751429856693	0.8995002580465
SVM	0.93007215007215	0.91850330955594	0.92077437266883

从评估数据看，SVM 算法相对于 KNN 算法分类结果更加精确。

2.3 热点问题挖掘

2.3.1 密度聚类 DBSCAN 算法

聚类算法属于常见的无监督分类算法，常见的聚类算法可以分成基于分区和基于层次划分的两类算法，但是存在以下缺点：

1) 需要事先确定聚类的个数
2) 只适用于具有凸形状的簇
3) 对内存的占用资源比较大

4) 需要确定停止分裂的条件

5) 计算速度慢

DBSCAN(Density Based Spatial Clustering of Applications with Noise)是一种基于密度的聚类算法，能够有效解决上述问题，它将满足确定密度规则的区域划分为簇，能够有效减少噪声值对于簇的影响，中心思想是对于集群中的每一个点，在给定半径范围内，相邻的点数必须超过最小阈值。DBSCAN 算法需要确定两个重要的参数:

1) Eps:领域半径

2) MinPts:在簇内领域中点的最少个数。

参数的设置直接影响了聚类的效果，所以对于参数的确定需要进行进一步的计算。

将附件三的数据进行预处理，过滤了大于 0.75、小于 0.001 的词条和单字词，按语料词频排序取了前 7000 个词进行聚类训练，考虑到自适应参数代码实现比较困难，并且显示的结果需要进行热力指标排序过滤排序，筛选出前五大的热点问题，显示结果较少，可以进行参数枚举分析，参数以每次 0.02 的增幅显示聚类结果，根据枚举聚类结果选择最优。

2.3.2 聚类评价

由于参数枚举，需要选择最优的聚类结果，需要进行聚类评价，在这选取了调整兰德指数 ARI 和轮廓系数作为聚类评价指标，通过枚举对比选取最佳聚类参数。

（1）兰德指数 RI 与调整兰德指数 ARI

兰德系数(Rand Index)指的是样本数据的测试值与实际值之间的相似度，RI 取值范围是[0,1]，值越大代表着聚类结果效果越好。给定 n 个对象集合 $S=\{O_1,O_2,\ldots,O_n\}$，假设 $U=\{u_i,\ldots,u_R\}$和 $V=\{v_1,\ldots,v_C\}$表示 S 的两个不同划分并且满足 $\bigcup_{i=1}^{R} u_i = S = \bigcup_{j=1}^{C} v_j$， $u_i \cap u_{i*} = \varnothing = v_j \cap v_{j*}$其中 $1 \leqslant i \neq i* \leqslant R$，$1 \leqslant j \neq j* \leqslant C$。

假设 U 是外部评价标准即 true_label，而 V 是聚类结果。设定四个统计量：

a 为在 U 中为同一类且在 V 中也为同一类别的数据点对数

b 为在 U 中为同一类但在 V 中却隶属于不同类别的数据点对数

c 为在 U 中不在同一类但在 V 中为同一类别的数据点对数

d 为在 U 中不在同一类且在 V 中也不属于同一类别的数据点对数

此时，兰德系数为：

$$\mathrm{RI}=\frac{a+b}{a+b+c+d}$$

兰德系数的值在[0,1]之间，当聚类结果理想时为 1。但是兰德系数存在对于两个随机的划分,数值不是一个接近于 0 的常数的问题，需要调整兰德系数假设模型的超分布为随机模型，U 和 V 集合进行随机划分，各类别和各簇的数据点数目固定，与兰德指数相比在处理随机结果时，保证了分数更加趋近与零，从而拥有更好的区分度[5]。调整的兰德系数计算公式为：

$$\mathrm{ARI}=\frac{RI-E(RI)}{\max(RI)-E(RI)}$$

其中 E(RI)以及 max(RI)计算公式如下：

$$E(RI)=E(\sum_{i,j}\binom{n_{ij}}{2})=\frac{\left[\sum_{i}\binom{n_i}{2}\sum_{j}\binom{n_j}{2}\right]}{\binom{n}{2}}$$

$$\max(RI)=\frac{1}{2}\left[\sum_{i}\binom{n_i}{2}+\sum_{j}\binom{n_j}{2}\right]$$

（2）轮廓系数

轮廓系数(silhouette coefficient) 结合了凝聚度和分离度两种指标来对聚类结果进行评估，其计算步骤如下：

1) 对于第 i 个对象，计算它到所属簇中所有其他对象的平均距离为 a_i

2) 对于第 i 个对象和不包含该对象的任意簇，计算该对象到给定簇中所有对象的平均距离为 b_i

3) 第 i 个对象的轮廓系数为：

$$\mathrm{S_i}=\frac{\mathrm{b_i}-\mathrm{a_i}}{\max(\mathrm{a_i},\mathrm{b_i})}$$

（其中S_i为所有点的轮廓系数的均值，a_i为i到同一簇内其他点不相似程度的均值，b_i为i到其他簇的平均不相似程度的最小值）

从计算公式可以看出，轮廓系数的取值范围为[-1, 1]，结果接近于 0 表示聚类结果有重叠的情况，越趋近于 1 代表内聚度和分离度都相对较优。

两种聚类指标都可以通过调用 sklearn 包实现，兰德系数和轮廓系数都是越接近于 1，结果越优。经过反复实验，可以得出在簇内领域中点的最少个数取 2 时，结果比较稳定。

2.3.3 构建热度指标

通过 DBSCAN 密度聚类算法可以得到完整聚类结果，需要排序过滤出前五大热点问题，根据已知信息因素：时间范围、聚类记录数、赞同数、反对数，构建热度指标，对热点问题进行热力排名。

首先，对于热点问题的时间范围进行确认，对于每一个热点问题的细则留言时间进行排序，得出时间范围。由于问题的时间跨度比较大，设置月为时间间隔，假设距离最新时间的越久，热度也会随之递减，热力计算公式可以表示为：

$$\text{Hot\_score} = \sum((\text{count} + \text{agree\_count} * 0.1 - \text{dis\_count} * 0.05) * a\text{^}i)$$

其中：

a 为热度系数

i 为距离该热点问题时间范围内最新月份的月份差（数据中的最新时间为 2020 年 1 月，将此设为最新时间）

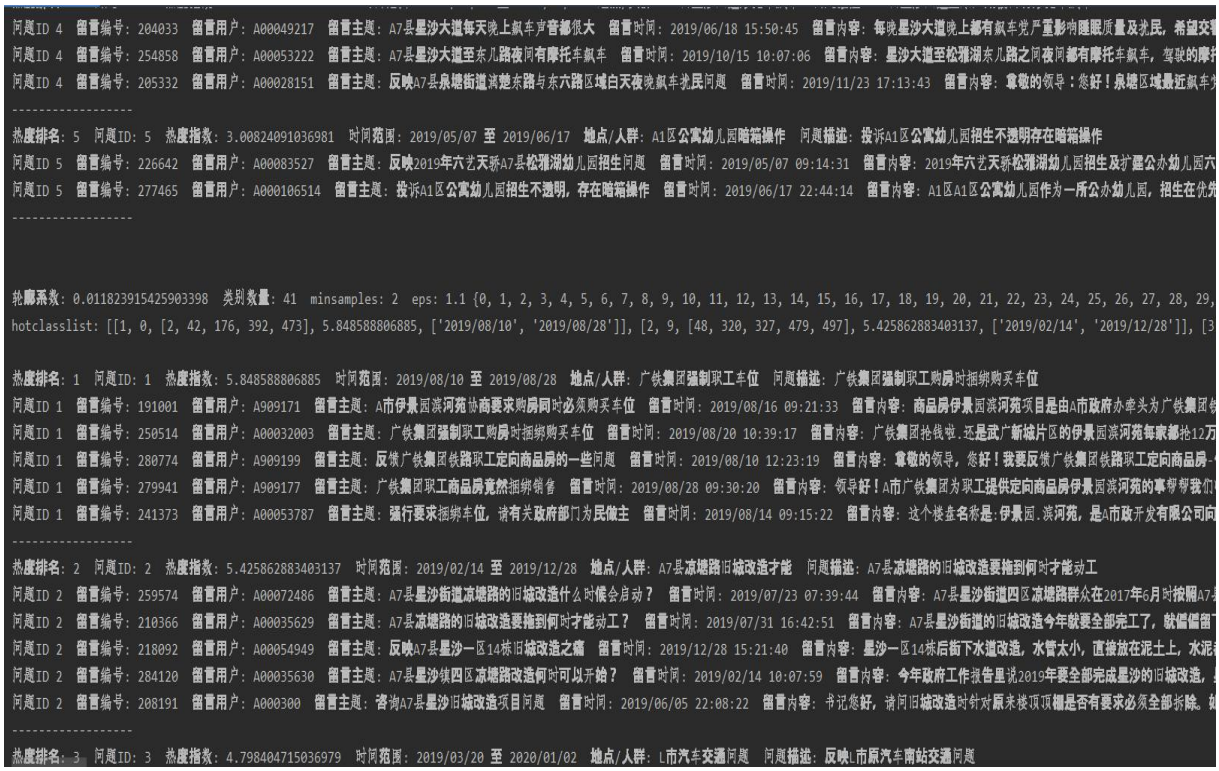
count 为以月为时间间隔内的留言次数

agree_count 为这一个月内留言记录的点赞数

dis_count 为这一个月内留言记录的反对数

根据热力计算公式，需要对各个聚类结果细则按照月份进行汇总，得出每类每个月的留言数量、反对数、赞成数以及和最新日期的月份差，进行赋权累加。在这里对于记录数、点赞数和反对数设置固定加权值，分别为 1、0.1 和 0.05，a 为热力系数，为小于 1 的数，为了适应数据时间间隔与最新时间差异较大，把

a 设定固定值 0.99，间隔月份越大，a^i 越小，每个聚类的热度指标进行排序，筛选出前五大热点问题。



热度排序显示结果（部分）

2.3.4 获取关键信息

经过上述步骤，得到前五大热点问题的详细信息，从中需要提取出人群、地点、时间范围等关键信息。对于时间范围的获取只需要对热点问题的细则进行时间排序就可能得到，主要的是对地点和人群这类关键词的提取。

关键词提取主要有关键词分配和关键词抽取两个方法。对于地点来说，大部分的地点信息都含有省、市、区、号、街道等后缀词，根据附件 3 的文本信息，留言详情与留言主题匹配度较高、概括度高，可以对于留言主题进行关键句自动提取，通过命名实体识别进行词性区分，只保留关键信息所需要的词性，组合成关键信息。

TextRank 是根据 PageRank 迭代思想而来的，可以自动用来抽取关键句，具体公式如下：

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

（$WS(V_i)$表示某一个语句的权重，$(1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$表示每个相邻句子对该语句的贡献程度）

命名实体识别（Named Entity Recognition）是自然语言处理的一个基础性任务，其目的是识别语言文本中人名、地名、组织机构名等实体。命名实体识别的主要技术方法分为基于规则、基于统计、二者混合这三种方法。在这里运用的是基于统计的层叠 HMM-Viterbi 方法。

主要的操作步骤是：

1) 进行角色标注
2) 统计词频和转移矩阵
3) 进行识别，再一次角色标注
4) 模式匹配，进行第二层隐马模型细分

在 Python 中通过调用 HanLP 工具包[1]中的 NLPTokenizer 类来进行 NLP 分词，NLPTokenizer 会进行文本中全部命名实体的识别和词性标注，将带有词性的词语进行集合，将单词标记和词性标注应用于句子，实现名词短语分块，以使用正则表达式来识别命名实体，正则表达式指示句子的分块规则，所需的信息应该形成名词短语。提取结果如下：

地点/人群	问题描述
A 市公交车	A 市 205 路公交车经常不按时发车
A 市伊景园滨河院车位	投诉 A 市伊景园滨河院捆绑销售车位
A 市高心巴比仑幼儿园	A 市高心巴比仑幼儿园无证办学十年了有没有人管
A 市力度	请 A 市加快一圈二场三道建设力度
夫妻房 A 市人才	夫妻共同买的房为何申请 A 市人才购房补贴不通过

[1] HanLp 工具包来自 www.github.com/hankcs/HanLP

2.3.5 结果输出分析

根据上述聚类结果和输出的五大热点问题信息表及问题细则表如下（部分）：

五大热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群
1	1	72.51	2019/01/19 至 2019/02/25	月 A4 区公安分局案情
2	2	72.13	2019/08/23 至 2019/09/06	A 市赣州高铁绿地海外小区
3	3	45.42	2019/07/07 至 2019/09/01	伊景园车位
4	4	28.24	2019/11/13 至 2020/01/09	新城小区搅拌站问题
5	5	17.66	2019/01/22 至 2019/12/18	A7 县街区幼儿园普惠幼儿园

热点问题细则表（部分）

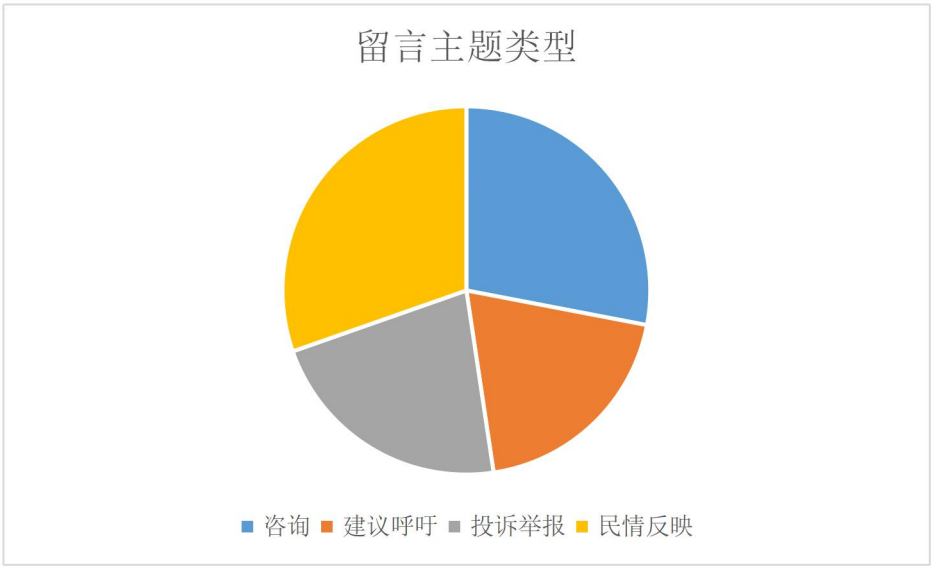
问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	217032	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞	2019/02/25 09:58:37	胡市长：您好！西地省展星投资有限雄…	790	0
1	264119	A00084445	58 车贷立案五个月过去，A4 区公安分局未公布过任何案情	2019/01/19 09:47:23	58 车贷立案五个月过去，A4 区公安分局…	0	0
2	191951	A00041448	A4 区绿地海外滩小区距渝长厦高铁太近了	2019/08/23 14:21:38	尊敬的领导：你好，近日看到了渝长厦高铁最…	1	0
2	202575	A00092007	咨询 A 市绿地海外滩二期与长懿高铁问题	2019/09/04 18:32:42	我们是 A 市 A4 区绿地海外滩二期 5 栋居民…	17	0
2	216316	A00097196	A4 区绿地海外滩二期业主被噪音扰得快烦死了	2019/09/06 10:16:27	我们是 A 市 A4 区绿地海外滩二期居民…	2	0

从中可以看出在上文中构建的热力指标受点赞数和记录数的影响比较大，热力排名第一、第二的问题虽然问题细则的记录数比较少，但是点赞数较多，因此导致了热力排名比较高。综合来说，热力指标受点赞数和记录数的影响比较大，由于数据基本上为 2019 年，时间影响程度不大。

2.4 答复意见评价建立

对于答复意见评价，要先从问题入手，对于回复问题类型进行区分,问题分类模块需要完成两个基本功能: ①提供可用于准确定位和验证答案的约束; ②指示下一步工作中需要针对不同答案类型作不同的处理[6]。因此问题分类体系的定义需要依据实际问题情况。根据附件 4 的文本信息，所有问题基本上可以分为咨询、建议呼吁、投诉举报、民情反映这四大类。具体情况如下图：

留言主题词云



留言主题分类数据饼图

在 4 大类基础下按照附件一的分类体系再进行 15 个类别细分。随机对附件 4 中 600 条记录根据留言主题进行四大类人工分类标识。主要步骤分为：

1) 根据留言问题和留言详情分别进行部分人工标识和所有数据分类

2) 人工标识优质回复

3) 进行特征提取

4) 相似度计算

将附件四的留言详情进行分词、去停用词等预处理后，利用附件二中的训练模型对各条记录进行问题二级分类。同时根据文本分类的步骤利用人工标识的数据进行一级分类模型训练，提取四大分类中信息特征，根据模型对所有留言问题进行分类。

在数据分别进行了一级分类和二级分类后，从附件四随机抽取不同分类的优质回复，再分别结合问题类型和分类领域知识对优秀回复进行特征提取，形成特征词汇库，需要引入函数 WORDSIM（w_i，w_{pj}）来描述问答中词语 w_i 与同类别的词袋中的第 j 个词语 w_{pj} 的相似度，将回复信息的关键词与优质回复的词汇库进行对比分析。相似度高的，相较于回复评价也相对高。

利用 word2vec 可以方便地计算出回复中的词语和优质回复的词汇库的相似度 WORDSIM（w_i，w_{pj}）。

$$\text{WORDSIM}（w_i，w_{pj}）=\frac{\sum_{i=1}^{n}（x_{i1}\times x_{i2}）}{\sqrt{\sum_{i=1}^{n}x_{i1}^2}\times\sqrt{\sum_{i=1}^{n}x_{i2}^2}}$$

其中，两个词语 w_i 和 w_{pj} 的词向量表示为: w_i = (x11, x21, x31, …, xi1…, xn1) , w_{pj} = (x12, x22, x32, …, xi2…, xn2) ;n 表示用 word2vec 训练词向量时设定的词向量的维度[7]。

3、结论

本文通过对互联网上的群众问政留言记录进行分析挖掘，实现了对群众留言的整合分类以及热点问题的提取，这对政府相关部门工作效率的提升有着重要意

义。随着网络的普及以及发展，传统依靠分工进行留言整理以及热点提取的方式已经跟不上数据的大量增长。本次对群众问政留言的建模，运用了 KNN、SVM 分类算法、DBSACAN 密度聚类算法，以及分类、聚类评价指标，熟悉运用 jieba、sklearn、hanlp 等工具包，实现群众留言的分类管理，并统计了群众集中反映的问题，初步构建回复评价模型。

分析挖掘模型实现了对群众留言记录的分类，对应部门可及时对留言进行处理，群众的问题也能得到及时解决，这大大提升了政府的工作效率，实现了“智慧政务”的服务化。将某一时间段群众集中反映的问题进行提取，并给出了对应问题的合理的热度指数，提取了时间范围、地点、人群，简单描述了问题要点，政府额能够实现在重点业务领域更加智能便捷，多数群众关注的问题能够得到针对性解决，实现了“智慧政务”的智能化，但是实现“智慧政务”在还需要进一步的努力来完善。

4、参考文献

[1]王志超, 孙建斌, 秦瑞丽.基于分词的关联规则预测系统研究[J].计算机应用与软件, 2018, 35 (12) :140-143.WANG Zhi-chao, SUN Jian-bin, Qin Rui-li.Association Rule Prediction System Based on Word Segmentation[J].Computer Applications and SoftWare, 2018, 35 (12) :140-143.81.
[2]唐明, 朱磊, 邹显春.基于 Word2Vec 的一种文档向量表示[J].计算机科学, 2016, 43 (06) :214-217, 269.TANG Ming, ZHU LEi, ZOU Xian-chun.Document Vector Representation Based on Word2Vec[J].Computer Scien ce, 2016, 43 (06) :214-217, 269.
[3]马建斌,李滢,滕桂法,等.KNN 和 SVM 算法在中文文本自动分类技术上的比较研究[J].河北农业大学学报,2008(03):120-123.
[4]祁小军,兰海翔,卢涵宇,丁蕾锭,薛安琪.贝叶斯、KNN 和 SVM 算法在新闻文本分类中的对比研究[J].电脑知识与技术,2019,15(25):220-222.
[5]张霄. 基于数据特征的标签传播聚类算法研究[D].兰州大学,2019.
[6]于娟. 面向政府公文领域的中文问题分类[C]. 中国通信学会青年工作委员会.2008 年中国高校通信类院系学术研讨会论文集（下册）.中国通信学会青年工作委员会:中国通信学会青年工作委员会,2009:80-84.
[7]杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学,2018,45(09):237-242.