“智慧政务”中的文本挖掘应用

摘要

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。

问题 1：通过过采样方法对数据进行预处理，得到较平衡的相关群众留言分类数据。利用 jieba 中文分词工具对留言主题+留言详情进行分词，Scikit-Learn 库中的 Bunch 数据结构来表示分词后的训练集语料库和测试集语料库。用 TF-IDF 算法计算权重向量，并利用 SVM 实现群众留言的分类。最后，使用 F-Score 对分类方法进行评价。

问题 2：对热点问题留言表进行数据预处理后，利用 LDA 主题分类模型对热点问题进行聚类。设置相关的热度算法，其中结合了偏差率和威尔逊区间算法进行相应的辅助计算。再进行数据的归类、汇总等工作，将热点问题的挖掘结果最终以 excel 形式输出呈现。

问题 3：分析所给数据，剖析评价方案的挖掘维度。先对问答进行无效处理，保留有效的数据进入评价系统。主要是从相关性、完整性、及时性、专业性四方面对部门回复进行评价打分。该问题从短文本语义相关性、语义的匹配度等方面进行挖掘处理。

关键词：过采样；jieba 中文分词；TF-IDF 算法；SVM；LDA；威尔逊区间；语义相关性

Application of text mining in "smart government"

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that mainly rely on manual work to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

Problem one: through the over sampling method to preprocess the data, we can get a more balanced classification data of the relevant public comments. Using the Chinese word segmentation tool of Jieba to segment the message subject + message details, bunch data structure in scikit learn database represents the training corpus and test corpus after word segmentation. TF-IDF algorithm is used to calculate the weight vector, and SVM is used to classify the public comments. Finally, F-score is used to evaluate the classification method.

Problem two: after the data preprocessing of the hot issues message table, the hot issues are clustered using LDA topic classification model. Set up the related heat algorithm, which combines the deviation rate and Wilson interval algorithm for the corresponding auxiliary calculation. Then we classify and summarize the data, and present the mining results of hot issues in the form of Excel.

Problem three: analyze the given data and the mining dimension of the evaluation scheme. First of all, the questions and answers are invalid, and the effective data are retained to enter the evaluation system. Mainly from the relevance,

integrity, timeliness, professionalism of the four sides of the Department's response to the evaluation and scoring. This problem is dealt with from the aspects of semantic relevance and semantic matching.

Keywords: Oversampling; Jieba Chinese word segmentation; TF IDF algorithm; SVM; LDA; Wilson interval; semantic relevance

目录

1、挖掘目标

本次挖掘是将互联网平台收集到的群众问政留言记录、相关部门对部分群众留言的答复意见数据，通过 jieba 分词、SVM 分类器、LDA 主题模型、语义相关性分析等方法进行处理，实现以下三个目标：

（1）建立关于群众留言内容的一级标签分类模型，协助工作人员在处理网络问政平台群众留言时，能够更加智能、高效、准确。

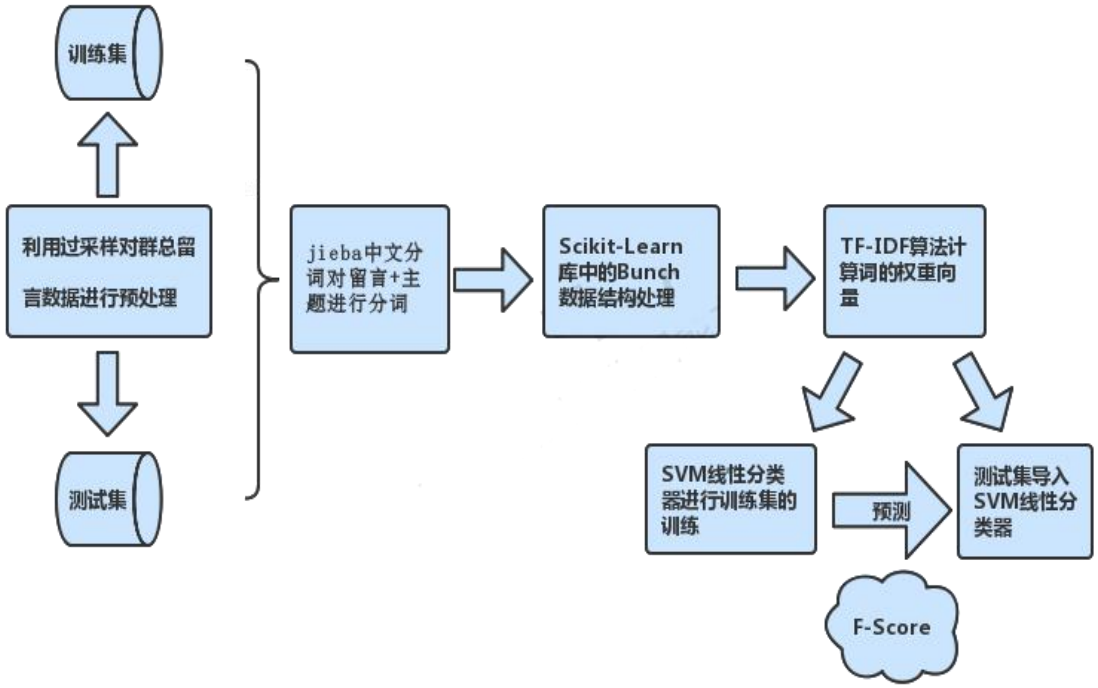
（2）利用群众在问政平台留言的相关数据，进行热点问题的挖掘，有利于相关部门进行有针对性地处理，从而提升服务效率。

（3）建立政府相关部门答复意见的评价方案，让政府的工作效率达到量化，从而达到督促政务处理的作用。并通过评价机制不断改善、提高相关部门的工作质量，让群众的民生问题能够得到积极的落实、响应。

2、问题 1 实现

2.1 分析方法与过程

2.1.1 流程图



2.1.2 数据预处理

2.1.2.1 数据不平衡问题

分析群众留言表，发现源数据存在数据不平衡的问题。

解决数据不平衡的方法有过采样和欠采样两种方法，因为本题数据量不多，所以利用过采样方法中的直接复制法，进行数据不平衡的处理。直接复制法，通过随机地直接复制类别样本数较少的类别样本，进而扩充某类数据少的样本数目。方法实现：将各个类别数量中进行汇总，获取类别数最大值。其他未达到类别数目最大值的样本数据就采用随机直接复制，从而达到每个类别的数量基本一致。对数据不平衡问题处理的 python 程序见附件 DataBalance.py。处理后的数据保存在附件 T1_ph.xlsx 中。

2.1.2.2 数据清洗

留言主题的文本数据中存在许多没有意义的符号，要将这些无意义的符号（例如首行缩进符、空格、回车行等）删去。为了避免无意义符号干扰后面的中文分词以及分类结果，需对数据进行清洗。采取直接滤过方法，将这些无关符号从文本中删除。

同时，要对数据集进行切分，将留言集切分为训练集和测试集，切分比例为 7：3。切分的比例不宜相同，否则导致训练集过少，分类模型的精度将下降。数据集切分后分别存放在 T1_Train 和 T1_Test 两个文件夹中，里面存放类别文件夹，每个文件夹的名称为类别名称，每个类别文件夹里存放这个类别的数据。由于 python 存放数据时候，名称相同的会直接覆盖，于是将行号也加入数据文本命名中，防止数据平衡时相同数据被覆盖。对数据切分的 python 程序见附件 DataConversion.py。

2.1.2.3 中文分词

在对留言信息进行分类之前，要将连续的字序列按照一定的规范重新组合成词序列的过程，即把非结构化的文本信息转换为计算机能够识别的结构化信息。原始留言数据都是中文文本数据，为了后面分类模型的建立，我们要对这些数据进行分词。采用的是 python 的中文分词包 jieba 进行分词。它的算法是基于前缀

词语实现高效的此图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图，采用了动态规划查找最大概率路径，找出基于词频的最大切合组合，对于未登录词，它采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法，使得中文分词效果大大提升。

由于方便，在进行分词的同时，也对数据进行数据清洗，保证了处理完的数据的有意义性。处理完的训练集和测试集分别存放在附件 train_corpus_seg 和附件 test_corpus_seg 两个文件夹中。

2.1.2.4 结构化表示-构建词向量空间

将分词后得到的训练集语料库和测试集语料库进行结构化处理，采用 Scikit-Learn 库中的 Bunch 数据结构来表示这两个数据集。Bunch 模式相当灵活，其属性可以动态设置，它具有 Dict 类的所有属性，比如对键/值对的遍历，以及判断某个属性是否存在。Bunch 相当于 python 的字典。用 Bunch 可以便于后面权重矩阵的计算，在其中创建 4 个成员，分别传入 train_corpus_seg 和 test_corpus_seg 数据集中的留言类别集合、留言文本的标签、留言文本文件的名字以及分词后的文本文件。将结构化的文本数据存放在附件 train_word_bag 和附件 test_word_bag 文件夹中。

2.1.2.5 生成 TF-IDF 向量

数据结构化后，将这些词语数据转换为向量，以供建立分类模型的使用。采用 TF-IDF 算法，将训练集和测试集所有文本文件统一到同一个词向量空间中，得到词典和权重矩阵。词典是单词和单词对应的序号。权重矩阵是一个二维矩阵，tdm[i][j]表示，第 j 个词（即词典中的序号）在第 i 个类别中的 TF-IDF 值。

TF-IDF 算法的原理如下：

第一步：计算词频 TF 权重。

$$词频（TF）= \frac{某关键词出现次数}{文章中关键词总数} \tag{2-1}$$

或者

$$词频（TF）= \frac{某关键词出现次数}{文章中出现最多次数关键词的出现次数} \tag{2-2}$$

这是考虑到留言文本的长短之分，为了便于不同的留言文本之间的比较，进行“词频”标准化。

第二步：计算逆文档频率 IDF 权重。

$$\text{逆文档频率（}IDF\text{）} = \log \frac{\text{语料库文档总数}}{\text{包含该词的文档数} + 1} \tag{2-3}$$

IDF 越大，表示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内容属性能力越强。

第三步：计算 TF-IDF。

$$TF - IDF = \text{词频（}TF\text{）} \times \text{逆文档频率（}IDF\text{）} \tag{2-4}$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

2.1.3 分类模型的建立

为了提高分类模型的 F1 分数，我们采用 python 的 sklearn 库中的各个分类模型进行对比，其中有基于多项式贝叶斯算法的分类模型、基于逻辑回归的分类模型、基于线性 SVM 的分类模型、基于决策树的分类模型、基于随机森林的分类模型、基于 KNN 最近邻的分类模型、基于多层神经网络的分类模型。在建立分类模型的同时，计算输出分类模型的准确率、召回率以及 F1 分数。通过几个模型之间的对比（在结果 2.2.1 处表 1），最终选择基于线性 SVM 的分类模型对留言进行分类。

线性 SVM 算法原理如下 ：

（1）硬边距（hard margin）

给定输入数据和学习目标：$X = \{X_1,...,X_N\}, y = \{y_1,...,y_N\}$，硬边界 SVM 是在线性可分问题中求解最大边距超平面（maximum-margin hyperplane）的算法，约束条件是样本点到决策边界的距离大于等于 1。硬边界 SVM 可以转化为一个等价的二次凸优化（quadratic convex optimization）问题进行求解[1][2]：

$$\max_{\omega,b}\frac{2}{\|\omega\|} \quad \text{s.t.} \quad y_i\left(\omega^T X_i+b\right)\geq 1 \iff \min_{\omega,b}\frac{1}{2}\|\omega\|^2 \quad \text{s.t.} \quad y_i\left(\omega^T X_i+b\right)\geq 1 \tag{2-5}$$

由(2-5)式得到的决策边界可以对任意样本进行分类： $sign\left[y_i\left(\omega^T X_i+b\right)\right]$。注意到虽然超平面法向量是唯一优化目标，但学习数据和超平面的截距通过约束条件影响了该优化问题的求解[2]。硬边距 SVM 是正则化系数取 0 时的软边距 SVM，其对偶问题和求解参见软边距 SVM，这里不额外列出。

（2）软边距（soft margin）

在线性不可分问题中使用硬边距 SVM 将产生分类误差，因此可在最大化边距的基础上引入损失函数构造新的优化问题。SVM 使用铰链损失函数，沿用硬边界 SVM 的优化问题形式，软边距 SVM 的优化问题有如下表示：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{N}L_i,\ \ L_i=\max\left[0,1-y_i\left(\omega^T X_i+b\right)\right] \quad \text{s.t.} \quad y_i\left(\omega^T X_i+b\right)\geq 1-L_i,\ \ L_i\geq 0 \tag{2-6}$$

(2-6)式表明可知，软边距 SVM 是一个 L2 正则化分类器，式中 Li 表示铰链损失函数。使用松弛变量，处理铰链损失函数的分段取值后，(2-6)式可化为：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{N}\xi_i \quad \text{s.t.} \quad y_i\left(\omega^T X_i+b\right)\geq 1-\xi_i,\ \ \xi_i\geq 0 \tag{2-7}$$

求解上述软边距 SVM 通常利用其优化问题的对偶性（duality），这里给出推导：定义软边距SVM的优化问题为原问题(primal problem)，通过拉格朗日乘子(Lagrange multiplier)： $\alpha=\{\alpha_1,\ldots,\alpha_N\},\ \mu=\{\mu_1,\ldots,\mu_N\}$ 可得到其拉格朗日函数[2][3]：

$$\wp(\omega,b,\xi,\alpha,\mu)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{N}\xi_i+\sum_{i=1}^{N}\alpha_i\left[1-\xi_i-y_i\left(\omega^T X_i+b\right)\right]-\sum_{i=1}^{N}\mu_i\xi_i \tag{2-8}$$

令拉格朗日函数对优化目标的偏导数为0，可得到一系列包含拉格朗日乘子的达式：

$$\frac{\partial \wp}{\partial \omega}=0 \Rightarrow \omega=\sum_{i=1}^{N}\alpha_i y_i X_i,\ \ \frac{\partial \wp}{\partial b}=0 \Rightarrow \sum_{i=1}^{N}\alpha_i y_i=0,\ \ \frac{\partial \wp}{\partial \xi}=0 \Rightarrow C=\alpha_i+\mu_i \tag{2-9}$$

将其带入拉格朗日函数后可得原问题的对偶问题（dual problem）[2][3]：

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\alpha_i y_i (X_i)^T (X_j) y_j \alpha_j]\\ &s.t.\ \sum_{i=1}^{N}\alpha_i y_i=0,\ 0\le\alpha_i\le C\end{aligned} \tag{2-10}$$

对偶问题的约束条件中包含不等关系，因此其存在局部最优的条件是拉格朗日乘子满足 Karush-Kuhn-Tucker 条件（Karush-Kuhn-Tucker condition, KKT）[2]：

$$\begin{cases}\alpha_i\ge 0,\ \mu_i\ge 0\\ \xi_i\ge 0,\ \mu_i\xi_i=0\\ y_i(\omega^T X_i+b)-1+L_i\ge 0\\ \alpha_i[y_i(\omega^T X_i+b)-1+L_i]=0\end{cases} \tag{2-11}$$

由上述 KKT 条件可知，对任意样本(X_i, y_i)，总有$\alpha_i=0$或 $y_i(\omega^T X_i+b)=1-\xi_i$，对前者，该样本不会对决策边界 $\omega^T X_i+b=0$产生影响，对后者，该样本满足 $y_i(\omega^T X_i+b)=1-\xi_i$意味其处于间隔边界上（$\alpha_i<C$）、间隔内部（$\alpha_i=C$）或被错误分类（$\alpha_i>C$），即该样本是支持向量。由此可见, 软边距 SVM 决策边界的确定仅与支持向量有关，使用铰链损失函数使得 SVM 具有稀疏性[2]。

2.2 执行结果与结论分析

2.2.1 F-Score 评价

F-score 是衡量一个分类模型好坏的重要算法，它综合了分类模型的查准率和查全率。它是查准率和查全率的调和平均值，当其中一个值非常高，另一个值非常低时，F-score 也会非常低；只有两者都非常高时，F-score 分数才会高，符合对查准率和查全率的权衡标准。

F-score 公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{2-12}$$

以下为各个分类模型的查准率、查全率以及 F1 分数：

表 1　分类模型的 F1 分数对比

	多项贝叶斯	SVC 分类	**线性 SVM**	逻辑回归	决策树	随机森林	KNN 最近邻	多层神经网络
查准率	0.918	0.951	**0.954**	0.936	0.858	0.928	0.874	0.952
查全率	0.918	0.951	**0.954**	0.936	0.861	0.929	0.874	0.952
F1	0.918	0.951	**0.954**	0.936	0.858	0.929	0.873	0.952
花费时间	0.03 秒	36.90 秒	**0.49 秒**	5.91 秒	3.96 秒	13.89 秒	0.01 秒	330.22 秒

通过表 1 可以清楚地看出，线性 SVM 的模型是 F1 分数最高，同时，花费的时间算较少的，于是采用的线性分类器是基于线性 SVM 的分类模型。

2.2.2 结论分析

通过对比几个分类模型，寻找 F-Score 分数最高的模型。从表 1 可以看出，线性 SVM 分类器所建立的分类模型评价分数达 0.954。这说明，在同等数据处理下，采用线性 SVM 分类模型进行留言分类更精确。在协助网络问政平台的工作人员对留言进行分类时，也更加高效、精确。不断优化算法、不断改进模型，建立更合适的分类模型，从而提高模型的 F-Score 分数，让问政平台能更加高效办公、服务群众。

3、问题 2 实现

3.1 分析方法与过程

3.1.1 数据预处理

将数据中无意义的符号清洗掉。jieba 分词词典采用搜狗输入法的词库而且 jieba 分词还具有新词识别能力，但是，自行添加新词可以保证更高的正确率，因此在原有的词典上添加了新词，新词即根据原始文本数据中的特有地名和聚类结果得到的一些专有词汇，从而提高 jieba 分词的准确率，部分新词表如下图 1，

详见附件 user.dict.utf8。

同时，为节省存储空间和提高搜索效率，利用合成停词表将文本数据中的停用词过滤掉，我们经过整理处理形成自己的停用词表。停用词表是用中文停用词表、百度停用词表、哈工大停用词表、四川大学机器智能实验室停用词库及互联网上前人经验总结出的停用词表进行合成，将重复的词语删除得到的停用词表。停用词表见附件 stop_words.txt。

A市 n	A3区 n	A8县 n	K10县 n
A9市 n	A4区 n	B7县 n	L6县 n
B市 n	A5区 n	C3县 n	M9县 n
C市 n	A6区 n	D7县 n	M14县 n
C4市 n	B4区 n	E7县 n	M6州 n
C5市 n	C2区 n	F9县 n	KTV n
D市 n	D区 n	J4县 n	kv n
E市 n	E4区 n	J5县 n	app n
F市 n	E5区 n	J9县 n	P2P n
L市 n	F区 n	K3县 n	ETC n
M市 n	G2区 n	K4县 n	5G网络 n
M5市 n	K1区 n	K5县 n	HPV九价 n
A1区 n	R区 n	K8县 n	长株潭 n
A2区 n	A7县 n	K9县 n	丽发新城 n

图 1　部分新词表

3.1.2 聚类簇数

3.1.2.1 确定簇类的意义

文本聚类模型搭建后，每次的迭代效果，呈现的结果都有所不同。此外，聚类簇数 K 值是一个自定义值。在面对一个庞大的留言详情数据集，确定一个最优 K 值，并不停运行迭代，是得到较科学、较精准的聚类模型的基石[4]。因此，引入结构风险，对模型的复杂度进行惩罚：

$$K = min_k[RSS_{min}(k) + \lambda k] \tag{3-1}$$

λ是平衡训练误差与簇的个数的参数，但是如何选取λ就变成了新的问题，有研究[4]指出，在数据集满足高斯分布时，λ=2m，其中 m 是向量的维度。

3.1.2.2 实现

将分好词的数据，使用 TF-IDF 进行特征词的选取，得到权重矩阵，之后利

用 K-means 算法通过选取不同 K 值然后进行循环，同时画出其对应的误差值，得到手肘图，通过寻求拐点来找到一个比较好的 K 值。

以下是结果手肘图，横坐标是 k 的数量，纵坐标是损失，根据手肘的拐点可确定，最优的簇聚类数目为 150。

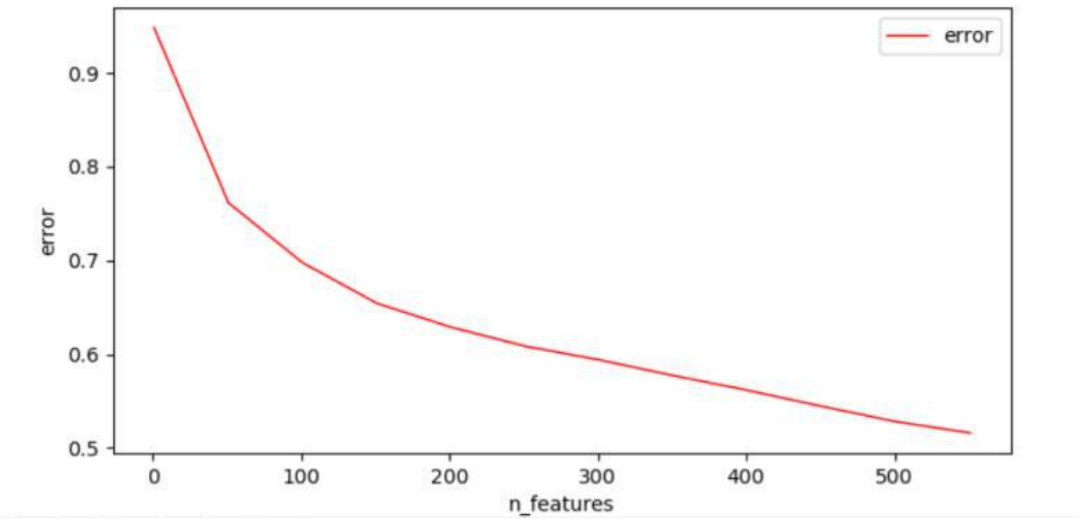


图 2　K 值手肘图

3.1.3 LDA 主题聚类

LDA 具体实现过程如下：

LDA 算法开始时，先随机地给 θ_d 和 ϕ_t 赋值（对所有的 d 和 t）。然后上述过程不断重复，最终收敛到的结果就是 LDA 的输出。

（1）针对一个特定的留言文本 d_s 中的第 i 单词 w_i，令该单词对应的 topic 为 t_j，则公式为：

$$p_j(w_i|d_s) = p(w_i|t_j)*p(t_j|d_s) \qquad (3\text{-}2)$$

（2）枚举 T 中的 topic，得到所有的 $p_j(w_i|d_s)$，其中 j 取值 1~k 根据这些概率值结果为 d_s 中的第 i 个单词 w_i 选择一个 topic。然后取令 $p_j(w_i|d_s)$ 最大的 t_j，即 argmax[j] $p_j(w_i|d_s)$；

（3）如果 d_s 中的第 i 个单词 w_i 在这里选择了一个与原先不同的 topic，就会

对θ_d和ϕ_t有影响了。它们的影响又会反过来影响对上面提到的 p(w|d)的计算。对 D 中所有的 d 中的所有 w 进行一次 p(w|d)的计算并重新选择 topic 看作一次迭代。这样进行 n 次循环迭代之后，就会收敛到 LDA 所需要的结果了[5]。

3.1.4 热度算法

3.1.4.1 热度指标

针对热度的计算，共设有 4 个指标，分别是爆发性、延长性、新鲜度和支持度。采用的策略思想是，对热度的总评分设置为 100 分，将四个指标按照所设置的比例分配相应分数，这里我们将四个指标权重都设置为 0.25，每个指标最低分为 0 分，最高分为 25 分。之后将四个指标单项分相加，总分就是话题热度排序的最终结果，并且按照总分进行从高到低的排序，详见附件 python 程序 redu.py。

（1）爆发性，使用[留言总数÷时间差]计算得分（时间差是指该类留言中最近的留言与时间最久远的留言之间的时间差，此处差值取天数，第（2）中的时间差与此处时间差相同），得分数值越大的爆发性越大，排名越靠前；

（2）延长性，使用[时间差]计算得分，得分数值越大的延长性越长，排名越靠前；

（3）新鲜度，使用[最新时间-当前时间]计算得分（最新时间指该类留言中时间最新的留言的时间，此处差值取天数），得分数值越小的新鲜度越高，排名越靠前；

（4）支持度, 使用威尔逊区间进行排序（排序的具体算法在 3.1.3.2 详述），排名即为得分数值，排名越靠前，数值越小，支持度越大；

热度指标分数的具体实现如下：

（1）各类留言各项指标的计算方式如下：

$$\text{diff} = \text{该指标第一名的数值} - \text{该指标最后一名的数值} \tag{3-3}$$

$$\text{rate} = \frac{\text{某类留言指标的数值} - \text{该指标最后一名的数值}}{\text{diff}} \tag{3-4}$$

$$\text{某类留言该指标数值的得分} = \text{该指标权重} \times 100 \times \text{rate} \tag{3-5}$$

（2）某类留言总分＝该类留言各指标得分之和

3.1.4.2 排名算法及威尔逊区间[6]

支持度的排名算法步骤：

第一步，计算每类留言项目的“好评率”（即赞成票的比例）。

第二步，计算每个“好评率”的置信区间（以 95%的概率）。

第三步，根据置信区间的下限值，进行排名。这个值越大，排名就越靠前。

根据本题需求，我们自行添加第四步，根据置信区间的上限值，在步骤三的排名基础上进行排名，下限相同的类别比较其上限值，值越大，排名就靠前。

其中第三步和第四步的排序均采用 python 中的 rank 函数，其 method 参数取值为 min, 意为对整个组使用最小排名，即当两个元素的值相同时，二者排名均取它们排名中的最小值。

威尔逊区间，基于 Z 检验，置信区间上下限如下：

$$\frac{p+\frac{1}{2n}z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}+\frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{1}{n}z^2_{1-\alpha/2}} \tag{3-6}$$

3.1.5 结果输出

将计算好的各个指标进行汇总，然后按照热度值总分进行排序，取热度值排名前五的留言类别，在 python 中利用第三方库 panda 将转换成 DataFrame, 用 excel 表格输出热点问题表。根据最高热点问题的类别 ID 升序排序，进一步整合输出热点问题留言明细表的 excel 形式。

3.2 执行结果及结论

3.2.1 执行结果

表 2 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	71.89	2019/01/06至2020/01/06	A市小区	居民生活环境脏乱差，垃圾污染严重，噪音油烟扰民
2	2	68.52	2019/01/04至2019/12/31	A市孩子、学生（尤其是小学生）	入学、升学困难，学校教育教学的相关问题
3	3	67.02	2019/01/18至2019/12/20	A市伊景园滨河苑等开发商	列出购房霸王规定，售房捆绑销售车位，欺诈消费者等其他购房问题
4	4	66.41	2019/01/08至2020/01/02	A市、西地省各公司	出现的虚假宣传、欺骗消费者、拖延退费、非法集资、偷税漏税和涉嫌传销等问题，且政府相关部门办事拖拉或坐视不理
5	5	65.82	2018/11/15至2020/01/07	A市国家政策、劳动和社会保障	住房公积金、医保社保、人才新政购房补贴等相关问题、工资拖欠、劳动合同纠纷及国家政策、当地政府单位的相关问题

表 3 热点问题留言明细表

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	191068	A00010562	A4区华章路路灯光源污染严重影响新领地小区居民生活！	2019-02-12 17:49:08	从昨天2月11号晚上7点开始，华	0	0
1	191579	A00017743	A市新奥燃气无法通过网络付费？	2019-09-18 09:16:06	我家住A8县和泰家园，使用的是	1	0
1	191699	A000106800	A6区润和紫郡鲲鹏物业太差	2019-09-03 09:05:47	尊敬的领导：我是A市A6区润和紫	0	0
1	194022	A00042107	坚决反对在A7县诺亚山林小区门口设置医院	2019-07-08 10:39:54	我们是诺亚山林小区的业主，已	1	0
1	195089	A00072618	举报A市博长山水香颐小区违规建设医疗机构	2019-09-04 12:56:14	尊敬的A市政府领导：我们系A市	0	0
1	199926	A000109613	恳请处理A8县市龙泉美墅旁可可槟榔的环保问题	2019-07-07 14:30:36	尊敬的领导　　这里主要就	2	0
1	200243	A0005101	A7县星沙螺丝塘路9号华中集团每天都排出大量白色气体，气味难闻	2019-07-08 15:03:21	现有星沙螺丝塘路9号华中集团每	1	0
1	201854	A00042351	对A8县西部修建高速公路的看法和建议	2019-01-06 14:14:21	西地省A8县西部为什么迟迟没有	0	0
1	202928	A00063359	请拆除A2区金线街40号的违章建筑	2019-02-27 23:50:26	我是A市A2区金线街40号东头201	0	0
1	207199	A00057942	A1区马王堆沁园小区b区停车位被个人占用了！	2019-06-06 22:31:28	嘉雨路沁园小区b3栋大量公用停	0	1
1	207467	A00014150	A4区金鹰小学附近一工地夜间施工扰民	2019-05-06 09:26:39	位于金鹰小学、恒大雅苑居民楼、	0	0
1	207521	A00050281	A市第一垃圾中转站散发恶臭	2019-09-06 12:49:02	我家位于A4区洪山路原山苑，A市	0	0
1	209203	A00063984	A3区永青东路环境差，到处是白色垃圾	2019-11-07 12:19:03	位于青山新村的永青东路上，环	0	0
1	209713	A00073319	新规划的A市第二大商圈，请给红星片区留一片绿心	2019-01-24 12:19:40	我是居住于A市红星商圈融城花园	6	1
1	210107	A00042107	坚决反对在A7县诺亚山林小区门口设置医院	2019-07-08 10:38:38	我们是诺亚山林小区的业主，已	14	0
1	210292	A00037700	指出A3区溁湾镇通程商业广场处红绿灯问题	2019-03-26 09:10:27	溁湾镇通往一桥中间(通程商业广	0	0
1	211492	A00085500	A5区昇明壹城沿街商铺商业空调的噪音、废气全排小区里了	2019-07-11 10:03:40	西地省A市A5区劳动东路昇明壹城	0	0
1	214913	A00079199	A6区茜茜美甲店欺骗消费者	2019-12-19 17:44:42	有关领导，您好我在18年2月6日	0	0
1	215563	A909231	A2区丽发新城小区旁边的搅拌厂是否合法经营	2019-12-06 12:21:32	领导，您好！我相咨询A2区丽发	0	0
1	216824	A909214	搅拌站大量加工砂石料噪音污水影响丽发新城小区	2019-12-25 12:15:57	最近一段时间以来，A2区丽发新	0	0
1	217064	A000112283	A2区金石蓉园小区居民被工地上的噪音给弄得午饭正常休息	2019-11-08 01:13:21	今日起，我所在小区A2区金石蓉	0	0
1	217244	A00076035	A3区中海国际社区有人违章搭建假山致房屋结构破	2019-04-02 17:40:09	2017年从中海地产购置23栋104商	0	0
1	219174	A00081998	A2区丽发新城小区内垃圾站散发严重臭味	2019-07-03 23:27:02	A2区丽发新城二期45栋，在离住	3	0
1	219675	A00088746	A市梅溪湖长郡中学附近交通乱象	2019-04-18 10:03:52	我是居住在梅溪湖长郡中学附近	0	0

3.2.2 结论

通过数据预处理、聚类簇数确定、LDA 主题聚类、热度计算、结果输出等挖掘步骤，总结出居民生活环境脏乱差、学校教育教学的相关问题、房地产商欺诈消费者、相关部门对偷税漏税及涉嫌传销等问题坐视不理、新政购房补贴等相关问题为排名前五的热点问题。热点问题的挖掘，让政府部门明确了近一段时间群众反响热烈的民生问题，给其提供了解决民生问题的目标和方向。摆脱效率比较低下的传统人工统计方法，减轻政府部门的人力负担和资源负担。该热点问题的挖掘，让政府部门能及时响应民生，及时处理民事，及时了解民情，使其积极落实为人民服务的宗旨和政策。

4、问题 3 实现

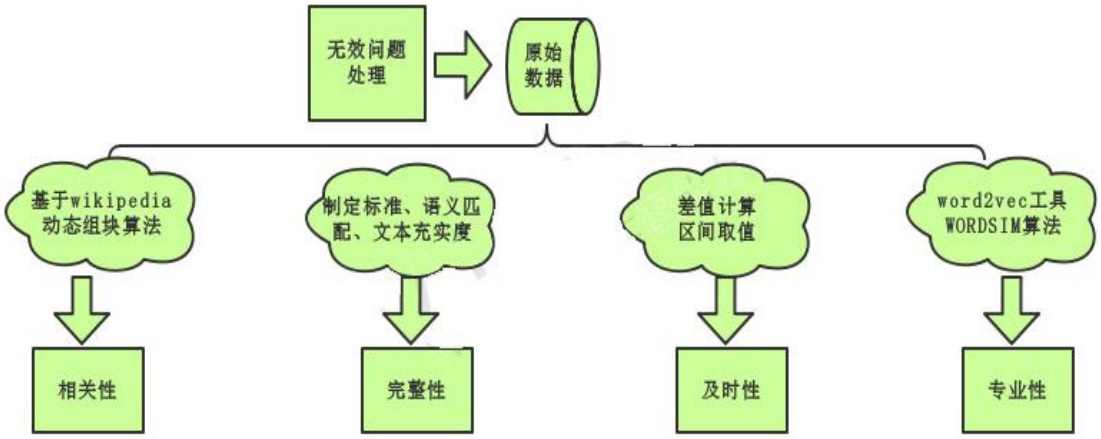
4.1 目的及意义

（1）建立政府意见答复的质量评价方案，是政府工作体系更加完善；

（2）量化政府的办事效率，使办公水平数字化、简洁化；

（3）精确定位政府的办事水准，从而提高或者改善政府部门的能力和效率。

4.2 方案框架图



4.3 框架搭建

4.3.1 指标设定

（1）相关性：根据分析政府部门的回复以及群众的额问题留言，计算政府

部门对问题回复的相关度值。

（2）完整性：针对政府部门的回复制定一套回复的标准或者模板，即判断该回复是否满足某种规范。

（3）及时性：根据分析群众留言时间和政府部门回复时间，来判断相关部门办事是否及时，是否积极。

（4）专业性：考虑政府部门回复是否专业也是一个重要性指标，体现相关部门员工的基本职业素养和办公能力。

4.3.2 无效问题处理

由于相关的群众留言描述不清或存在过于简短的问题，政府部门无法给出相应的合理的解答。因此，此类问题留言应做无效处理，贴上无效问题的标签，不列入评价系统。

先通过挖掘群众问题留言的文本，数据经过预处理之后，去掉无用词，判断文本长度的区间范围。挖掘过于简短或者描述不清问题的文本长度，设置问题留言的最少阈值。当留言文本的字数小于阈值，贴上无效问题标签，并不纳入评价系统进行评价打分。

4.3.3 指标实现

4.3.3.1 相关性

本题采用 Wikipedia 作为外部知识库，在计算词语之间相关度的基础上来计算短文本之间的语义相关度。通过动态组块算法，来计算问题留言和政府部门回复两个文本之间的相关度。本题将分析文本中相关度较高的词语聚集在一起，称之为动态组块。动态组块可看作是构成文本的语义单位[7]。

获取动态组块的步骤如下：

（1）对 A（即需要分析的文本）进行分词和词性过滤，获取其特征向量 A’=$\{a_1,a_2,\ldots,a_n\}$，令组块集合 *Chunk*（A）为空，即 *Chunk*（A）={}；

（2）取 A′中元素 a_i，如果 *Chunk*（A）为空，将（a_i）作为一个组块加入 *Chunk*（A）中，否则计算 a_i与 *Chunk*（A）每一组块中词语的平均相关度；

（3）如果 a_i 与 c_j（$c_j \in Chunk$（A））的平均相关度（记为 $\overline{r_{ij}}$）最大，且 $\overline{r_{ij}}$ ≥θ（θ为阈值），则将 a_i 加入 c_j，否则将（a_i）作为一个组块加入 *Chunk*（A）；

（4）反复执行步骤二和步骤三，直到 A′为空。

相关度计算如下：

设问题留言文本 A 和政府答复文本 B 的动态组块集合为：

$$Chunk（A）=\{CA_1, CA_2, \cdots, CA_n\}$$

$$Chunk（B）=\{CB_1, CB_2, \cdots, CB_n\}$$

两个组块之间的相关度，定义为组块包含词语之间相关度的均值，设：

$$CA_i=\{a_1, a_2, \cdots, a_n\}$$

$$CB_j=\{b_1, b_2, \cdots, b_n\}$$

则组块 CA_i和 CB_j 的相关度定义为：（形成组块特征相关矩阵）

$$SimC_{rel}(CA_i, CB_j) = \frac{\sum_{i=1}^{n}\sum_{i=1}^{n} \mathrm{Sim}(a_i, b_j)}{s \times t} \tag{4-1}$$

其中，$\mathrm{Sim}(a_i, b_j)$为词语 a_i和 b_j 之间的语义相关度值。

获取最大组块关联序列：

$$\mathrm{Max}L=\{S_{1,x1}, S_{2,x2}, \cdots, S_{n,xn}\}$$

计算 A 和 B 之间的组块相关度：

$$Sim_{sem}(A, B) = \frac{\sum_{i=1}^{n}\left(s_{i,\,xi} \times \sqrt{w_i \times w_{xi}}\right)}{n} \tag{4-2}$$

其中，W_i 和 W_{xi} 分别是 a_i和 b_j 在文本 A 和 B 中的权重，权重可用其频度表示。

4.3.3.2 完整性

在量化政府部门回复文本的完整性时，制定相应的回复模板。模板是通过挖掘众多回复文本中形成的，外加自定义的文本规范。回复模板除了规定了开头内

容、结尾内容的结构标准，还定义了一些必须出现的相关带有解释与艺术的关键词和礼貌用用语词。模板形成后与政府部门回复文本进行与语义匹配，匹配度越高分数越高。此外，用对数计算政府部门回复文本的长度，来量化其内容的文本内容充实度。将匹配度分数与文本充实度分数按照权重算总分。

4.3.3.3 及时性

利用及时性来衡量政府部门处理民生问题的积极性和响应度。采用群众问题留言时间和政府部门回复时间的差值，以天数来计算。制定评分的天数区间值，将所求得到差值进行对照，即可求得及时性的分数值。

4.3.3.4 专业性

利用第一题的一级标签、二级标签、三级标签等词语统计绘制成政府部门回复的专业领域词汇表。根据专业领域的词汇表可以判断政府部门工作人员回复内容的专业度和回复质量，可以认为政府部门工作人员在回复语句中所用的词语和相关专业领域的词语的语义相似度越高，用到的领域词越多，政府部门工作人员在该领域的专业水平越高。

引入函数 *WORDSIM*（Wi,Wpj）来描述问答中词语 Wi 与问政平台工作人员回复的专业领域 p 中的第 j 个词语 Wpj 的相似度，将回复信息的关键词与政府部门回复的专业领域的词汇库进行对比分析。本题中的 *WORDSIM*（Wi,Wpj）的计算方法采用 word2vec 工具[8]。

计算过程如下：

$$WORDSIM(w_i, w_j) = \frac{\sum_{i=1}^{n}(x_{i1} \times x_{i2})}{\sqrt{\sum_{i=1}^{n} x_{i1}^2} \times \sqrt{\sum_{i=1}^{n} x_{i2}^2}} \tag{4-3}$$

其中，两个词语 Wi 和 Wpj 的词向量表示为：Wi=$\{x_{11}, x_{21}, \cdots, x_{n1}\}$，Wpj=$\{x_{12}, x_{22}, \cdots, x_{n2}\}$；n 表示用 word2vec 训练词向量时设定的维度。

当 Wi=Wpj 时，可以通过系数 $1+\lambda'$ 增加相似度。因此，建立如下会专业度

的评价项 F1：

$$F_1 = \begin{cases} WORDSIM(w_i, w_{pj}), & w_i \neq w_{pj} \\ (1+\lambda')WORDSIM(w_i, w_{pj}), & w_i = w_{pj} \end{cases} \tag{4-4}$$

5、结论

走进民生、响应民生，打造“智慧政务”是必然，也是不可逆转的趋势。面对庞大的数据，建立基于自然语言处理技术，从而提升政府的管理水平和施政效率也是政府应该面对的挑战和新发展需求。通过这次“智慧政务”的文本挖掘，政府部门相关方面的工作可以得到数字化、信息化、科学化发展。建立良好的群众留言分类模型，从而减轻政府部门工作压力，同时也提高政府部门办事的效率。分析群众留言问题，挖掘热点问题，让政府部门能及时地响应民生，及时地处理民事，及时地了解民情。此外，制定政府部门回复的评价方案，能在一定程度上将政府部门的办事效率量化。并深度发现政府部门办公的优缺点，从而进一步不断完善和提高部门自身能力。

高速发展的 21 世纪，信息大爆炸的现代，如何利用更科学、合理的程序算法解决政府、机关、企业等，是值得我们大学生思考的问题。总之，学习算法、运用算法，用算法处理大数据，用算法解决棘手、复杂的问题，是我们迈向社会的必经之路。

6、参考文献

[1] 李航. 统计学习方法. 北京：清华大学出版社，2012：第七章，pp.95-135

[2] 周志华. 机器学习. 北京：清华大学出版社，2016：pp.121-139，298-300

[3] Friedman,J.,Hastie, T. and Tibshirani, R.. The elements of statisticallearning (Vol. 1, No. 10). New York, NY: Springer, 2001: Chapter 12, pp.417-438

[4] 王斌.信息检索导论.https://www.open-open.com/pdf/9354955662774d26b75e0edd16f8b294.html

[5] ZoeChen. LDA(Latent Dirichlet Allocation)主题模型算法 .http://blog.sina.com.cn/s/blog_8eee7fb60101cwzj.html

[6] 阮一峰.基于用户投票的排名算法（五）：威尔逊区间.阮一峰的网络日志.http://www.ruanyifeng.com/blog/2012/03/ranking_algorithm_wilson_score_interval.html

[7] 王荣波.Wikipedia 的短文本语义相关度计算方法. ComputerApplicationsandSoftwareVol.32NO. 1Jan. 2015

[8] 杨开平. 基于网络回复的律师评价方法. 计算机科学. 2018 年 09 期