“智慧政务”中的文本挖掘和综合分析

摘 要：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。本文将利用自然语言处理和文本挖掘的方法解决留言分类，挖掘并整理出留言中的热点问题，最后从答复的相关性、完整性、可解释性等角度对有关部门答复意见的质量进行评价。

在本次数据挖掘过程中，针对问题一，本文先对文本进行剔除数字和英文字母、中文分词及停用词过滤等数据预处理，然后基于词频-文档频率（TF-IDF）、支持向量机（SVM）等数据挖掘模型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，最后使用 F-Score 对分类方法进行评价。

针对问题二，本文首先对文本进行剔除数字和英文字母、中文分词及停用词过滤等数据预处理，然后基于 TF-IDF 权重法提取特征词，形成词袋，再通过基于 LDA 的文本聚类算法，将附件 3 根据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，最后定义话题热度指标，给出热度评价结果，得出排名前 5 的热点问题。

针对问题三，本文通过表层语言特征、相似性特征和语言解释性特征对文本进行特征提取。基于给定的问答质量判定标准，对留言内容及答复意见先进行了人工标注，并通过提取特征集，利用算法设计和实现了基于特征集的问答质量分类器，最后用分类器给每个答复意见的质量等级进行分类，得出每个答复意见的质量等级。

关键词：TF-IDF；中文分词；分类器；LDA；特征词；文本聚类

Text mining and comprehensive analysis in "intelligent government affairs"

Abstract: WeChat, weibo, mayor's mailbox, sunshine hotline and other network political platform gradually become an important channel for the government to understand public opinion, gather people's wisdom, and pool people's spirit. The increasing amount of text data related to various social situations and public opinions has brought great challenges to the work of relevant departments which mainly relied on manual workers to divide messages and sort out hot topics. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government. This paper will use the methods of natural language processing and text mining to solve the message classification, mining and sorting out the hot issues in the message, and finally from the relevance, integrity, interpret-ability and other aspects of the response to the quality of the department.

In the data mining process, in view of the question one, this article first to eliminate text Numbers and English letters, Chinese word

segmentation and stop words filtering data preprossessing, and then based on word frequency, document frequency (TF-IDF), support vector machine (SVM), and other data mining model according to certain classification system (refer to annex 1 provides the content of the classification level 3 label system) for message classification, so that the follow-up message assigned to the corresponding functional departments to deal with the masses, finally using F - Score to evaluate classification method.

To solve the second question, this article first to eliminate text Numbers and English letters, Chinese word segmentation and stop words filtering data preprocessing, and then based on TF - IDF weight method to extract the key form the word bag, again through the text clustering algorithm based on the LDA, attachment 3 to a certain period reflect a specific location or classify the specific people problem message finally defined topic heat index, the evaluation results are given, are the top five.

As for question three, this paper extracts the features of text through the features of surface language, similarity and language interpretation. Based on the given q&a quality criterion, the q&a quality classifier based on the feature set is designed and realized by extracting the feature set.

Key words: TF-IDF; word segmentation; classifier. LDA. key words; text clustering

目录

1. 挖掘目标

本次建模的目标是针对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，在对文本进行相关的预处理、中文分词、停用词过滤后，一方面通过建立包括词频-文档频率（TF-IDF）、支持向量机（SVM）、VSM特征词建模和 LDA 主题建模等多种数据挖掘模型实现对群众留言进行分类以及对留言进行文本聚类，将相似的留言聚集在一起，方便发现热点话题；另一方面从答复的相关性、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。期望此次文本挖掘和综合分析能在一定程度上减少工作人员的工作量、提升工作效率，降低差错率高等问题。

2. 总体流程和步骤

本论文的分析流程大致可分为以下四步：

第一步：获取分析用的原始数据（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）；

第二步：针对每一问对获取的数据进行基本的处理操作，包括数据预处理、中文分词、停用词过滤等操作；

第三步：文本数据经过处理后，运用多种数据挖掘模型对文本数据进行多方面的分析；

第四步：从对应结果的分析中获取文本数据中有价值的内容。

3. 群众留言分类

3.1 具体流程和步骤

3.2 数据预处理

3.2.1 数据描述

通过观察所给数据，可以发现数据量比较大，且文本有大量英文字母和数字、连接词（“但是”、“并且”、“然而”等）等对文本信息贡献不大的词。如果不做处理会对后续分析造成影响，并且留言文本信息存在大量噪声特征，如果把这些数据也引入进行分词、词频统计乃至文本分类，则必然会对分类结果的质量造成很大的影响，于是针对群众留言分类，需要先对数据进行预处理。

3.2.2 文本预处理

（一）去除数字、英文字母

英文字母和数字对文本信息贡献不大。本文首先剔除文本中的数字和英文字母。去除数字、英文字母后的结果示例：

去除前：B3 区公园住宅 6 栋后面有人直排生活用厨房用水，导致排水沟堵塞，恶臭，空气污染严重，物业部作为，多次反应都不处理，只有请李局长您帮忙了。谢谢！

去除后：区公园住宅栋后面有人直排生活用厨房用水，导致排水沟堵塞，恶臭，空气污染严重，物业部作为，多次反应都不处理，只有请李局长您帮忙了。谢谢！

（二）中文分词

分词顾名思义是指将文本的语句或词序序列切断，分为一个个单独的词。对于中文而言，由于其表达上的特点，不仅存在一词多义的现象，而且存在词句的多变性和连贯性，是的很多情况夏，同一句子的不同断句存在着文字层面上的明显歧义。而词句作为中文基本的语义表达单位，文本预处理中的关键的一步是将词语分割开而尽可能地保证语义的准确性，分词技术即是针对解决这一问题而产生的。

本文采用的分词工具是 jiebaR 包。示例结果如下表，原始文本数据附件 2 的全部中文分词处理见作品附件：问题一文本预处理结果.xlsx

分词前	分词后
A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。	″位于″,″书院″,″路″,″主干道″,″的″,″在水一方″,″大厦″,″一楼″,″至″,″四楼″,″人为″,″拆除″,″水″,″电″,″等″,″设施″,″后″,″烂尾″,″多年″,″用″,″护栏″,″围着″,″不但″,″占用″,″人行道路″,″而且″,″护栏″,″锈迹斑斑″,″随时″,″可能″,″倒塌″,″危机″,″过往行人″,″和″,″车辆″,″安全″,″请求″,″有关″,″部门″,″牵头″,″处理″

上图的分词结果是没有停用词过滤的结果，可以看到，其中表达无意义的字词，对后续分析会造成很大影响，因此接下来需要进行停用词过滤。

（三）停用词处理

经过中文分词这一步骤，已将初始的文本处理成词的集合。停用词是指在文本中分布普遍且均匀的高频词语，一般是基于语法结构而加入的词，可以理解为古汉语里的“之”、“乎”、“者”、“也”等这类信息量低的高频词。因为这

类词在不同的文本里皆分布均匀，在特征选取的过程中，停用词的介入可能会造成选出的特征几乎都是停用词，从而影响结果的分析。

文本采用基于停用词表的文本停用词过滤方式，将分词结果与停用词表中的词语进行匹配。若匹配成功，则进行删除处理。示例结果如下表，原始文本数据附件 2 的全部停用词处理结果见作品附件：问题一文本预处理结果.xlsx

停用词处理前	停用词处理后
″位于″,″书院″,″路″,″主干道″,″的″,″在水一方″,″大厦″,″一楼″,″至″,″四楼″,″人为″,″拆除″,″水″,″电″,″等″,″设施″,″后″,″烂尾″,″多年″,″用″,″护栏″,″围着″,″不但″,″占用″,″人行道路″,″而且″,″护栏″,″锈迹斑斑″,″随时″,″可能″,″倒塌″,″危机″,″过往行人″,″和″,″车辆″,″安全″,″请求″,″有关″,″部门″,″牵头″,″处理″	″大道″,″西行″,″便″,″未管″,″路口″,″加油站″,″路段″,″人行道″,″包括″,″路灯杆″,″圈″,″西湖″,″建筑″,″集团″,″燕子″,″山″,″安置房″,″项目″,″施工″,″围墙″,″上下班″,″期间″,″这条″,″路上″,″人流″,″车流″,″极多″,″隐患″,″非常大″,″文明″,″整改″,″极″,″不文明″,″路段″

3.3 文本特征提取

文本特征提取的步骤为：

（1）统计每一类内文档所有出现的词语及其频率；

（2）统计每一类内出现词语的总词频，并取其中的若干个频率最高的词汇作为这一类别的特征词集。特征提取部分结果示例如下：

城乡建设的特征词集为：

业主　开发商　物业　规划　房屋　房产证　建设　项目　工程　城管　房子　施工　社区　建筑　住房　质量　拆迁　改造　违法　住户

征收环境保护的特征词集为：

排污　环保　环保局　环境　环评　空气　破坏　身体健康　生产　气味　投诉　危害　污染　污水　环境保护　噪音　环保部门　健康　检

测 排放 生活

3.4 文本特征表示

经过上述文本预处理后，虽然已经去掉部分停用词，但还是包含大量词语，给文本向量化过程带来困难，所以特征抽取的主要目的是在不改变文本原有核心信息的情况下尽量减少要处理的词数，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。

本文采用词频-文档频率（TF-IDF）方法进行降维来获得特征集。

在使用 TF-IDF 方法时，由于文档长短不易，一个词无论其重要与否可能会在较长的一个文档中多次出现，所以需要将词频进行归一化处理。对于一个文档中的特征词，其重要程度可以表示成标准化之后的“词频”，用下式表示

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

式中 $tf_{i,j}$ 表示特征词 i 在文档 j 中出现的频率，分子部分的 $n_{i,j}$ 表示该特征词 i 在文档 j 中出现的次数。分母表示在文档 j 中所有的特征词出现的次数之和。

IDF（inverse document frequency）是衡量词在文档中的重要性，可以用总的文档数量除以包含该特征词的文档数量之后再取对数，具体的表述为下式：

$$idf_{i,j} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

其中 $|D|$ 表示文档的总数，$|\{j : t_i \in d_j\}|$ 表示第 i 个词出现在第 j 个文档里，即文档频数。则 TF-IDF 表示为：

$$tf-idf_{i,j} = tf_{i,j} \times idf_j = \frac{n}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

从上式中可以看出，在文档中多次出现的应该赋予较大的权重，用式中的 $tf_{i,j}$ 来确保其较大的权重。但是，如果该词在每个文档中都出现，就无法准确地判别该词代表哪个或者哪种文档。因此需要适当地降低其权重，用 idf_j 来降低该

特征词权。df_j 值越高，$\log\frac{|D|}{|\{j:t_i\in d_j\}|}$ 值越低，也意味着 idf_j 值越低。$tf_{i,j}$ 与 idf_j 相互结合加权的结果不但保证了高频词的重要性，而且确保了其在整个文档集中分布的区分能力。

根据 TF-IDF 的原理，计算出 TF-IDF 值，将其按大到小依次排列，再根据分析文档的内容与目的确定一个阈值，选择大于阈值的特征词构成特征集进行文本的分析。

TF-IDF 方法部分结果如下：

```
> inspect(tdm1)
<<TermDocumentMatrix (terms: 77837, documents: 2009)>>
Non-/sparse entries: 128893/156245640
Sparsity           : 100%
Maximal term length: 17
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
        Docs
Terms    1057       1169 1957 417 51       745 756 758 801 948
           0 0.00000000    0   0  0 0.0000000   0   0   0   0
  房屋     0 0.00000000    0   0  0 0.0000000   0   0   0   0
  房子     0 0.00000000    0   0  0 0.0000000   0   0   0   0
  公积金   0 0.00000000    0   0  0 0.0000000   0   0   0   0
  公园     0 0.00000000    0   0  0 0.5094753   0   0   0   0
  规划     0 0.00000000    0   0  0 0.0000000   0   0   0   0
  居民     0 0.08151482    0   0  0 0.0000000   0   0   0   0
  开发商   0 0.00000000    0   0  0 0.0000000   0   0   0   0
  小区     0 0.00000000    0   0  0 0.0000000   0   0   0   0
  业主     0 0.00000000    0   0  0 0.0000000   0   0   0   0
```

图 1 城乡建设

```
> inspect(tdm2)
<<TermDocumentMatrix (terms: 77837, documents: 938)>>
Non-/sparse entries: 68134/72942972
Sparsity           : 100%
Maximal term length: 17
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
        Docs
Terms    2100      2101 2102       2115 2125      2308      2368 2410      2538      2643
           0 0.0000000    0 0.02303517    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  村民     0 0.0000000    0 0.07679167    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  环保     0 0.0000000    0 0.00000000    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  环保局   0 0.0000000    0 0.00000000    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  环境     0 0.0000000    0 0.00000000    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  居民     0 0.1086864    0 0.00000000    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  排放     0 0.0000000    0 0.00000000    0 0.1201115 0.0000000    0 0.4250099 0.0000000
  生产     0 0.0000000    0 0.00000000    0 0.0000000 0.0000000    0 0.0000000 0.0000000
  污染     0 0.0000000    0 0.00000000    0 0.0000000 0.0000000    0 0.0000000 0.0835084
  污水     0 0.0000000    0 0.10366145    0 0.3718291 0.3455382    0 0.4385677 0.0000000
```

图 2 环境保护

```
> inspect(tdm3)
<<TermDocumentMatrix (terms: 77837, documents: 613)>>
Non-/sparse entries: 29585/47684496
Sparsity           : 100%
Maximal term length: 17
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
        Docs
Terms    2973 3017       3124       3174 3258 3434 3493      3501      3502      3542
             0    0 0.04368738 0.00000000    0    0    0 0.0000000 0.0000000 0.0258558
  车辆       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.4683797 0.0000000
  出租车     0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
  的士       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
  滴滴       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
  客运       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
  快递       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
  收费       0    0 0.00000000 0.05379902    0    0    0 0.1931874 0.0000000 0.0000000
  司机       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
  邮政       0    0 0.00000000 0.00000000    0    0    0 0.0000000 0.0000000 0.0000000
```

图 3 交通运输

得到的特征集可视化后部分结果如下图所示，在词云图中，词频的高低决定词的大小。

图 4 城乡建设

图 5 环境保护

图 6 交通运输

3.5 分类器选择和训练

特征提取之后就获得了文档集合的特征子集，文本向量对应所有的文本，完成对文本的表示工作。下一步对文本的分类工作就转化为对文本对应的向量进行分类，即利用分类算法对分类器进行设计。最具有代表性的算法主要有 K-最近邻分类（KNN)）、决策树、朴素贝叶斯分类方法和支持向量机（SVM）等。

3.4.1 支持向量机（SVM）原理

本文采取支持向量机（SVM）方法。SVM 的基本思想是：对于特征空间中两类点不能靠超平面分开的非线性问题，SVM 采用映照方法将其映照到更高维的空间，并求得最佳区分二类样本点的超平面方程，作为判别为止样本的判据。

线性可分情形

SVM 算法是从线性可分情况下的最优分类面（Optimal Hyperplane）提出的。

所谓最优分类面就是要求分类面不但能将两类样本点无错误地分开，而且要求两类的分类空隙最大。D 维空间中线性判别函数的一般形式为 $g(x)=w^Tx+b$，分类面方程是 $w^Tx+b=0$，我们将判别函数进行归一化，使两类所有样本都满足 $|g(x)|\geq 1$，此时离分类面最近的样本的 $g(x)=1$，而要求分类面对所有样本都能正确分类，就是要求它满足

$$y_i(w^Tx_i+b)-1\geq 0, i=1,2,...,n \text{。}$$

式中使等号成立的那些样本叫做支持向量（Support Vectors）。两类样本的分类空隙（Margin）的间隔大小：

$$M\arg in=2/\|w\|$$

因此，最优分类面问题可以表示成如下的约束优化问题，即在约束条件下，求函数

$$\phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w^Tw)$$

的最小值。为此，可以定义如下的 Lagrange 函数：

$$L(w,b,\alpha)=\frac{1}{2}w^Tw-\sum_{i=1}^{n}\alpha_i[y_i(w^Tx_i+b)-1]$$

其中，$\alpha_i\geq 0$ 为 Lagrange 系数，我们的问题是对 w 和 b 求 Lagrange 函数的最小值。把上式分为对 w、b、α_i 求偏微分并令它们等于 0，得：

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$
$$\frac{\partial L}{\partial b}=0\Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0$$
$$\frac{\partial L}{\partial \alpha_i}=0\Rightarrow \alpha_i[y_i(w^Tx^i+b)-1]=0$$

以上三式加上原约束条件可以把原问题转化为如下凸二次规划的对偶问题：

$$\begin{cases}\max\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i^Tx_j)\\ s.t\quad a_i\geq 0, i=1,2,...,n\\ \quad\quad \sum_{i=1}^{n}a_i y_i=0\end{cases}$$

这是一个不等式约束下二次函数机制问题，存在唯一最优解。若α_i^*为最优解，则

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

α_i^*不为零的样本即为支持向量，因此，最优分类面的权系数向量是支持向量的线性组合。

b^*可由约束条件$\alpha_i\left[y_i\left(w^T x_i + b\right) - 1\right] = 0$求解，由此求得的最优分类函数是

$$f(x) = \mathrm{sgn}\left(\left(w^*\right)^T x + b^*\right) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i x_i^* x + b\right)$$

sgn( )为符号函数。

非线性可分情形

当用一个超平面不能把两类点完全分开时（只有少数点被错分），可以引入松弛变量$\xi_i\left(\xi_i \geq 0, i = \overline{1,n}\right)$，使超平面$w^T x + b = 0$满足：

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i$$

当$0 < \xi_i < 1$时样本点x_i仍旧被正确分类，而当$\xi_i \geq 1$时样本点x_i被错分。为此，引入目标函数：

$$\psi(w, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

其中 C 是一个正常数，称为惩罚因子，此时 SVM 可以通过二次规划（对偶规划）来实现：

$$\begin{cases} \max \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i^T x_j\right) \\ s.t \quad 0 \leq a_i \leq C, i = 1,2,...,n \\ \sum_{i=1}^{n} a_i y_i = 0 \end{cases}$$

3.4.2 SVM 参数选择

经分析对支持向量机有着重要影响的参数是：惩罚因子，核函数及其参数的选取。惩罚因子 C 用于控制模型复杂度和逼近误差的折中，C 越大则对数据的拟

合程度越高，学习机器的复杂度就越高，容易出现“过学习”的现象。而 C 取值过小，则对经验误差的惩罚小，学习机器的复杂度越低，就会出现“欠学习”的现象。当 C 的取值大到一定程度时，SVM 模型的复杂度将超过空间复杂度的最大范围，那么当 C 继续增大时将几乎不再对 SVM 的性能产生影响。SVM 的核函数包括线性核函数，RBF 核函数，多项式核函数，高斯核函数等，对于构建一个 SVM 模型来说首先需要做的就是选择核函数和核参数。根据 Vapnik 等人的研究表明，对于不同类型的核函数，SVM 模型所选择的支持向量的个数基本相同，但是其核函数参数和惩罚因子 C 的选择却对 SVM 模型的性能有着重要影响，如 RBF 核函数的参数 gamma 的取值就直接影响模型的分类精度，也就是说对于一个 RBF 核的 SVM 模型，要想提高其分类精度首先需要考虑的就是如何选取其核参数 gamma 和惩罚因子 C。

一级标签分类结果见作品附件：一级分类标签.xlsx

3.6 结果评价

本文使用 F-Score 对分类方法进行评价。

这种评价方法并不是围绕一条曲线进行的，而是针对一个点开展的。它从模式识别的角度出发，将文本分类作为一个待识别的问题进行评价。文本分类中常用的指标为查全率、查准率和 F_1 评估值。下面详细介绍国际上通用的查全率、查准率、准确率和 F_1 系统性能评价方法。

查全率（Recall）即正确预测为正的占全部实际为正的比例。表示为：

$$R=\frac{\text{正确预测为该类的文档数}}{\text{测试集中实际属于该类的文档数}}\times 100\%$$

查准率（Precision）即为正确预测为正的占全部预测为正的比例，表示为

$$P=\frac{\text{正确预测为该类的文档数}}{\text{测试集中预测属于该类的文档数}}\times 100\%$$

查准率和查全率反映了分类质量的两个不同方面，两者必须综合考虑。因此，列入了 F_1 测试值作为评价指标，它是查全率和查准率的调和平均数，表示为

$$F_1=\frac{2\times R\times P}{R+P}$$

计算结果如下：

	SVM_PRECISION	SVM_RECALL	SVM_FSCORE
1 城乡建设	0.5	0.73	0.59
2 环境保护	0.8	0.66	0.72
3 交通运输	0.65	0.28	0.39
4 教育问题	0.82	0.73	0.77
5 劳动与社会保障	0.78	0.85	0.81
6 商贸旅游	0.52	0.62	0.57
7 卫生计生	0.87	0.53	0.66

表 1 F-Score 法计算结果

4. 热点问题挖掘

4.1 具体流程和步骤

本文将对留言进行文本聚类，将相似的留言聚集在一起，方便发现热点话题。文本聚类主要是依据著名的聚类假设：同类的文档相似度较大，而不同类的文档相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程，以及不需要预先对文档手工标注类别，因此具有一定的灵活性和较高的自动化处理能力，已经成为对文本信息进行有效地组织、摘要和导航的重要手段。具体流程如下：

4.2 数据预处理

4.2.1 中文分词

分词顾名思义是指将文本的语句或词序序列切断，分为一个个单独的词。对于中文而言，由于其表达上的特点，不仅存在一词多义的现象，而且存在词句的多变性和连贯性，是的很多情况夏，同一句子的不同断句存在着文字层面上的明显歧义。而词句作为中文基本的语义表达单位，文本预处理中的关键的一步是将词语分割开而尽可能地保证语义的准确性，分词技术即是针对解决这一问题而产生的。本文采用的分词工具是 jiebaR 包。

4.2.2 停用词处理

经过中文分词这一步骤，已将初始的文本处理成词的集合。停用词是指在文本中分布普遍且均匀的高频词语，一般是基于语法结构而加入的词，可以理解为古汉语里的“之”、“乎”、“者”、“也”等这类信息量低的高频词。因为这类词在不同的文本里皆分布均匀，在特征选取的过程中，停用词的介入可能会造成选出的特征几乎都是停用词，从而影响结果的分析。

文本采用基于停用词表的文本停用词过滤方式，将分词结果与停用词表中的词语进行匹配。若匹配成功，则进行删除处理。

本题的文本预处理结果见作品附件：问题二文本预处理结果.xlsx

4.3 构建词袋空间

构建词袋空间的步骤如下：

（1）将所有留言读入到程序中，再将每个留言切词；

（2）去除每个留言中的停用词；

（3）统计所有留言的词集合；

（4）对每个留言，都构建一个向量，向量的值是对应词语在该留言中出现的次数。

4.4 TF-IDF 构建词权重

前面已经得到了文本的向量化表示，但是只用词频进行数值化明显是不够的。因此，这里使用 TF-IDF 来度量每个词的重要程度。

4.5 使用聚类算法进行聚类

所谓文本聚类就是将无类别标记的文本信息根据不同的特征，将有着各自特征的文本进行分类，使用相似度计算将具有相同属性或者相似属性的文本聚类在一起。这样就可以通过文本聚类的方法把相似的留言聚集在一起，方便发现热点话题。

4.5.1 传统的文本聚类算法

传统的聚类算法是根据文本的特征词进行聚类的，即利用文本中的词频统计信息来进行文本间相似度度量的评价，然后根据文本间的相似度进行聚类。但是，与其他的自然语言处理问题一样，文本聚类和文本所隐含的语义密切相关。因此，人们总是希望可以在文档的概念级别和语义级别来进行聚类算法。

传统的文本聚类算法中，文本的表示通常选择向量空间模型，把词作为特征项，把文档构成一个高维、稀疏的词语-文本矩阵。但是向量空间模型只是在词语层面对文档之间的关系进行分析，无法在语义层面对文档之间的关系进行分析。当前的一个新的研究方向就是通过主题模型来对文本进行建模，把文本表示为多个主题按照一定比例进行的混合，使得可以充分地挖掘文本集合的内在关系。

4.5.2 基于 LDA 和 VSM 相结合的文本聚类算法

传统的聚类分析一般在特征词层面对文本的相似度进行计算，这种方法往往会丢失大量的深层语义的信息。通过 LDA 主题模型，我们可以得到文档-主题-特征词三层模型结构。LDA 主题模型可以从文本内部挖掘出潜在的主题知识，但是因为模型的维度比较低，难以保持文本信息的完整性，使得区分文本能力不够强。通过将两种方法进行有机的结合，从特征词和主题两方面对文本进行聚类分析，结合了两种模型的优点，弥补两种方式的不足，提高聚类的准确率。

正如我们上文所讲，传统的聚类方法是采用基于 TF-IDF 权值策略的 VSM 特征词空间向量，来计算文本的相似度的。

一篇文档 d_i，采用 TF-IDF 权值策略的文本向量为 $d_{i(TF-IDF)}=(w_1,w_2,...,w_N)$，其中 N 为特征词的个数；基于 LDA 主题模型的文本向量为 $d_{i(LDA)}=(t_1,t_2,...,t_T)$，其中 T 为潜在主题的个数。

两个文本 d_i 和 d_j，基于 TF-IDF 权值策略来计算两个文本相似度的公式为：

$$S_{TD-IDF}(d_i,d_j)=\frac{d_{i(TF-IDF)}\times d_{j(TF-IDF)}}{\left|d_{i(TD-IDF)}\right|\times\left|d_{j(TF-IDF)}\right|}$$

基于 LDA 主题模型得到的主题向量来计算两个文本相似度的公式为：

$$S_{LDA}(d_i,d_j)=\frac{d_{i(LDA)}\times d_{j(LDA)}}{\left|d_{i(LDA)}\right|\times\left|d_{j(LDA)}\right|}$$

本文分别利用 LDA 主题模型和基于 TF-IDF 权值策略的 VSM 向量空间模型来计算文本的相似度，然后把两种方式计算的相似度进行线性结合，得到文本的相似度，然后使用 K-means 聚类算法来进行聚类分析。线性结合的公式如下所示：

$$S(d_i,d_j)=\lambda S_{TF-IDF}(d_i,d_j)+(1-\lambda)S_{LDA}(d_i,d_j)$$

其中，λ 为线性相关的系数。

LDA 主题建模的聚类的图形化结果如下图所示：

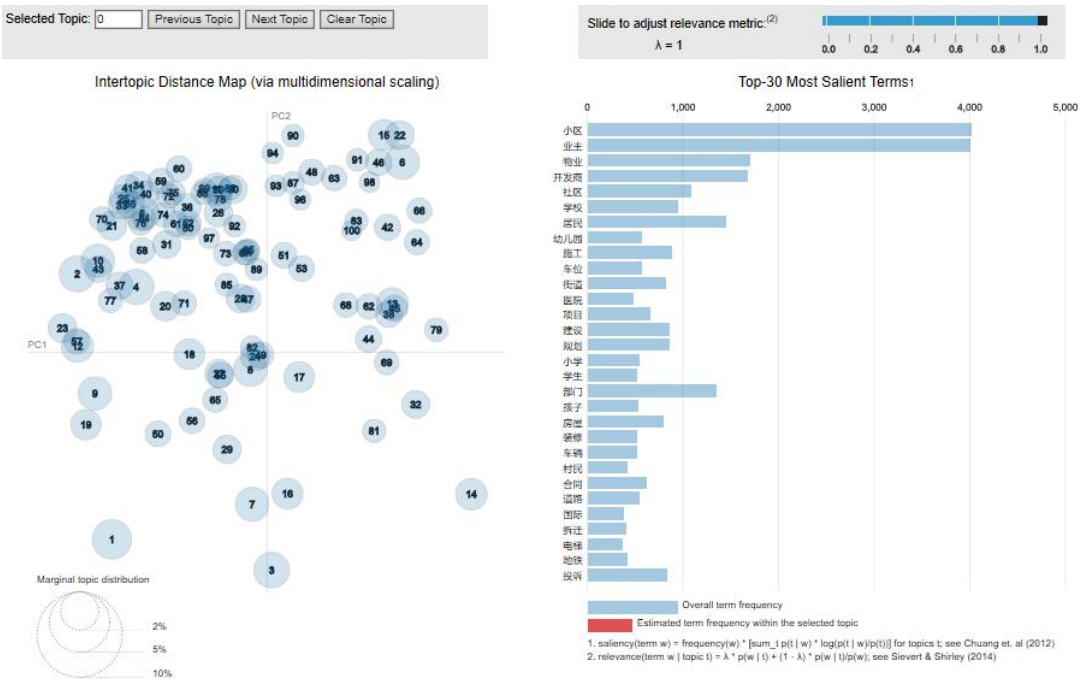


图 7 LDA 主题建模结果

4.6 热度评价

问题热度受很多因素的影响，根据留言的特点，主要考虑影响问题热度的三个因素：涉及该问题的留言量、留言时间、点赞量。

定义话题热度指标 Hot，该指标的计算公式为：

$$Hot(\mathrm{T})=\alpha\frac{S_n}{S_N}+\beta\log\ \mathrm{p}_n$$

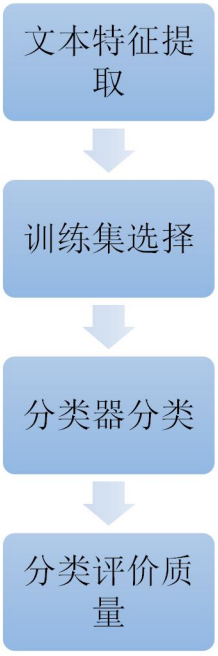
其中，S_N为单位时间内所有留言数量，S_n为问题 T 相关的留言数量，p_n为问题 T 的点赞数，α，β是调整因子，本文取α=0.8，β=0.1.

热点问题表及热点问题留言明细表见作品附件:热点问题表.xlsx 和热点问题留言明细表 xlsx

5. 答复意见的评价

5.1 具体流程和步骤

本文根据留言的答复意见的文本特征对答复的质量进行评价。为了方便评价，我们将质量等级分为高和低两类。先对部分答复意见进行初步人工标注，运用分类器对高低质量的答复进行分类，从而对全部答复意见评价质量等级。本题具体流程如下：



5.2 特征提取

5.2.1 表层语言特征

表层语言特征是指不需要经过复杂的分析就能从答案上下文中提取出来的特征。表层语言特征不需要复杂的算法就可以提取出来，且可以通过一定形式在一定程度上反映出答复者的语言风格。一个高质量的答案应包含合理的表层语言特征。表层语言特征包括答复的长度、动词数、名词数和非停用词的词数。

5.2.2 相似特征

根据对问答对文本进行分析，可以发现问答对里问句和答案之间存在一定的相似性，相似度越高一般表面答案质量会比较高。本文使用基于向量空间模型的方法对问题和答案的相似度进行计算。

向量空间模型是其他模型的基础，被广泛地应用于文本检索领域中，其在相似度计算方面占有重要地位。其计算公式如下：

$$sim\left(w,w^{1}\right)=\frac{\sum_{i=1}^{i=n}\left(w_1\times w_1^{'}\right)}{\sqrt{\sum_{i=1}^{i=n}w_1^{2}\times\sum_{i=1}^{i=n}w_1^{'2}}}$$

w 为一个问题句子，其对应的向量为 $w=\{w_1,w_2,.....,w_n\}$；对应的待比较的答案句子 $w^{'}$，其对应的向量为 $w^{'}=\{w_1^{'},w_2^{'},.....,w_n^{'}\}$。

相似度越高说明答复意见质量越好，相似性结果见作品附件中的相似性.csv。

5.2.3 语言解释性特征

语言翻译模型在自然语言处理领域占有重要的地位，其中，上下文环境和词序对于语言翻译模型而言是非常重要的因素。

假设任意一个问句 q 和答案句 a，定义 a 翻译成 q 的概率为 $P\left(q\mid a\right)$，于是将 a 翻译成 q 的问题则变成求解公式：

$$a^{1}=\arg\max P\left(q\mid a\right)$$

将句子中词语对其为止设置为 A,这有公式

$$P\left(q\mid a\right)=\sum_{A}P\left(q,A\mid a\right)$$

假设问句 q 长度为 m，记做 $b_1=b_1^{m}=b_1b_2...b_m$。答案句 a 的长度为 m，记做 f = $f_1^{m}=f_1f_2...f_m$,则最终计算公式为：

$$P\left(q,A\mid a\right)=p\left(m\mid a\right)\prod_{j=1}^{m}P\left(b_j\mid b_1^{j-1},f_1^{j-1},m,a\right)P\left(f_j\mid b_1^{j},f_1^{j-1},m,a\right)$$

通俗来讲，其模型原理是：给定两个对应的句子，其词语经常在句子中同时地出现，那么他们之间语言翻译性就越高。在本文中，将语言翻译行特征表示为问题和答案上下文之间语义关联的概率，通过分析发现，对于某些特定类别的问题，其高质量的答案往往期望与问题存在上下文间的语义关联。例如原因类问题，

答案中期望包含因为等词语。

5.3 训练集选择

人工标注了答复意见的好坏，1 代表高质量的答复，0 代表低质量的答复，具体标准如下：

（1）给出的答复意见和留言的内容是否相关；

（2）答复是否规范；

（3）关于留言的内容的相关解释是否到位。

人工标注的答复意见的质量评价见作品附件：答复质量人工评价.xlsx

5.4 分类器选择

选用逻辑回归算法，逻辑回归分析主要是研究自变量与因变量为二分类或多酚类观察结果之间的关系。逻辑回归分类主要是根据已有数据建立回归公式，以此来进行分类，由以下公式可知，将每个数据集上的因变量乘以回归系数，将其结果求和，最后输入到 $f(z)$ 逻辑函数公式中，根据其函数值进行分类。其中，$x_1,.....,x_k$ 是所有变量，θ_k 是回归系数，z 代表了所有因变量的总贡献的值。如下图所示，横轴为 z 值，纵轴为函数 $f(z)$ 的值，从图中可以看出 $f(z)$ 的结果是介于 0 和 1 之间，越接近于 1 表面发生该事物的可能性越高。也就是说，逻辑回归分类器最大的问题就是求一组权值系数。

$$f(z)=\frac{e^z}{e^z+1}=\frac{1}{1+e^{-z}}$$

$$Z=\theta_0+\theta_1x_1+\theta_2x_2+...+\theta_kx_k$$

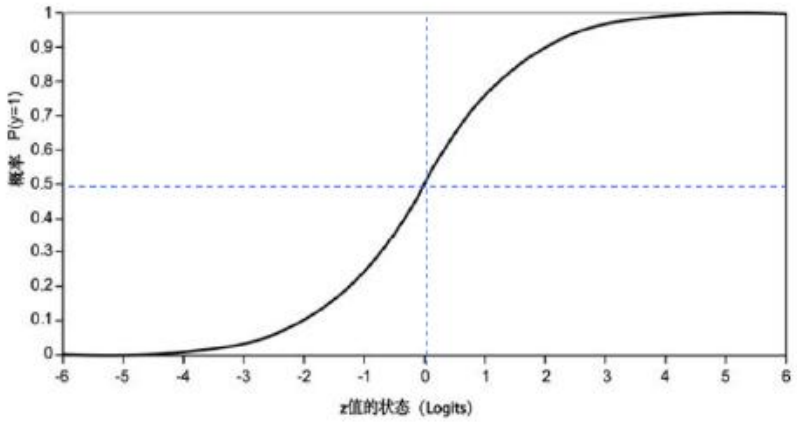


图 8 逻辑回归模型

5.5 质量等级分类结果

通过以上步骤对部分答复意见进行初步人工标注，运用分类器对高低质量的答复进行分类，达到对全部答复意见评价质量等级的目的，结果见作品附件：全部答复意见质量评价结果.xlsx

6. 结论

总结本次比赛，我们针对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，在对文本进行相关的预处理、中文分词、停用词过滤后，通过建立包括词频-文档频率（TF-IDF）、支持向量机（SVM）、VSM 特征词建模和 LDA 主题建模等多种数据挖掘模型实现对群众留言进行分类以及对留言进行文本聚类，将相似的留言聚集在一起，方便发现热点话题，最后从答复的相关性、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。

7. 参考文献

[1]姜博闻. 基于向量空间模型的文本分类及 R 语言实现[D].山东师范大学,2018.

[2]车蕾,杨小平.多特征融合文本聚类的新闻话题发现模型[J].国防科技大学学报,2017,39(03):85-90.

[3]高星. 面向新闻的话题发现和热度评估方法研究[D].东北师范大学,2017.

[4]李青. 高校网络舆情话题热度趋势预测研究[D].山东科技大学,2017.

[5]崔敏君,段利国,李爱萍.多特征层次化答案质量评价方法研究[J].计算机科学,2016,43(01):94-97+102.

[6]王少鹏,彭岩,王洁.基于 LDA 的文本聚类在网络舆情分析中的应用研究[J].山东大学学报(理学版),2014,49(09):129-134.

[7]刘海旭. 基于 PCA 和 LDA 的文本分类系统设计与实现[D].北京邮电大学,2013.

[8]李晨,巢文涵,陈小明,李舟军.中文社区问答中问题答案质量评价和预测[J].计算机科学,2011,38(06):230-236.

[9]方匡南. 基于数据挖掘的分类和聚类算法研究及 R 语言实现[D].暨南大学,2007.