“智慧政务”中的文本分析与挖掘

摘 要

由于网络问政的便捷性，网络问政平台已逐渐成为政府部门感知民意、凝聚民智的重要途径，这就导致了各类社情民意的文本数据大量增加，而其中部分社情民意往往具有时效性，属于热点问题，更需要及时处理。内容繁杂、数据量大，这给依靠人工来进行留言分类、热点挖掘、意见答复的相关部门人员的工作带来了极大的困难。但与此同时，随着大数据、云计算等科学技术的发展，建立于自然语言处理、运用网络文本分析的“智慧政务”系统已然渐渐成为了政府施政创新发展的新趋势。

对于问题一，将附件二中的留言主题和留言详情进行去重去空等数据预处理，得到有效不重复的留言信息。利用 jieba 中文分词工具对留言信息进行分词，再进行去停用词和过滤的数据处理。然后，基于 TF-IDF 权重法提取候选特征词，形成词袋，构造词汇-文本矩阵。根据 TF-IDF 算法得到每个职位描述的 TF-IDF 权重向量，进而利用奇异值分解算法进行语义空间降维。根据 Complement NB 算法构建模型，对留言信息进行分类，F-score 方法进行分类结果分析。

对于问题二将附件三中的留言主题和留言详情进行数据清洗去重去空等数据预处理,利用 K-means 聚类算法对附件 3 中某一时段内反映的问题留言进行聚类，根据 K-means 聚类算法的评价指标 silhouette 系数对聚类结果进行检验；然后再利用 HanLP 进行命名实体识别特定地点和特定人群；依据聚类结果通过每个类别的数量多少、点赞数、反对数以及时间衰减程度构建热度评价指标进行综合排序，综合排序结果，选取前五个问题为热点问题。

对于问题三，将答复意见从与留言问题的相关性、内容的完整性、答复的可解释性以及答复的时效性四个方面构建量化评价模型，实现对答复意见质量的综合评价。

关键词： TF-IDF、Complement NB、F-score、K-means、HanLP

Abstract

Because of the convenience of the network to ask politics, the network political platform has gradually become an important way for government departments to perceive public opinion and gather people's wisdom, which has led to a large increase in the text data of various social sentiments, and some of them tend to be time-sensitive, a hot issue, and more need to be dealt with in a timely manner. The content is complicated and the amount of data is large, which makes it very difficult to rely on manual personnel to classify messages, hot spot mining, opinion responses to the work of the relevant departments. But at the same time, with the development of big data, cloud computing and other science and technology, the "smart government" system, which is established in natural language processing and uses the analysis of network text, has gradually become a new trend of innovation and development of government governance.

For question one, the message subject and message details in Annex II to go to empty and other data pre-processing, to get effective non-repeating message information. Using the jieba Chinese word breaker, word-sharing information is word-breakaway, and then de-stop words and filtered data processing. Then, based on TF-IDF weighting method, the candidate feature words are extracted, the word bag is formed, and the vocabulary-text matrix is constructed. According to the TF-IDF algorithm, the TF-IDF weight vector described by each position is obtained, and then the semantic spatial dimension is reduced using the singular value decomposition algorithm. Based on the Complement NB algorithm, the model is constructed, the message information is classified, and the F-score method is analyzed.

For question two will be in Annex III message subject and message details of data cleaning to re-empty and other data pre-processing, using K-means clustering algorithm to cluster the problem message reflected in a certain period of time in Annex 3, according to the evaluation index of K-means clustering algorithm Silhouette Coefficient index clustering results to verify the clustering results, and then use HanLP named entities to identify specific locations and specific groups of people According to the clustering results, the first five questions are selected as hot issues by weighting the number of each category, the number of likes and the number of objections.

For question three, the response syllables are constructed from the four aspects of relevance to the message question, the integrity of the content, the interpretability of the reply and the timeliness of the response, so as to realize the comprehensive evaluation of the quality of the responses.

Keywords: TF-IDF、Complement NB、F-score、K-means、HanLP

目 录

一、简介

1.1 挖掘意义

随着时代发展，科技进步，网络离人们的生活也越来越近。越来越多的人们习惯于运用网络留言来表达自己的疑问和看法。由此而来的网络问政也就开始发挥着日益重要的作用，网络问政通过其成本低、用时少、操作易等优点，已经逐渐成为各级政府部门体察民情、收集民意、回复民疑的重要方式之一。比如，很多学校官网会有的“阳光信箱”，以及“智慧政务”系统等。

但是由于相关技术依旧存在一些不完善的地方，导致工作人员的工作难度增大，我们这次进行文本分析与挖掘的意义正在于此，帮助工作人员减轻工作量，提高工作效率，解决人民群众的问题。

1.2 问题重述

对于问题一，解决留言信息分类问题，我们需要做的是选择出最合适的算法流程，来将海量的留言信息进行分类。对于算法的选择，我们需要考虑到留言信息的繁杂、重复等特点，比如文本语义问题带来的词语交叉，长文本的无意义表达，数据的不平衡问题，甚至是留言信息的多分类也提高了我们解决问题的难度，还需要去注意算法的精确度和召回率，力求准确快速地将留言信息进行分类，减少政府相关工作人员的工作量，提高工作效率。

对于问题二，热点问题挖掘，我们首先需要在大量的留言信息中尽量准确地找到相似的留言，再将相似的留言信息归为同一类问题，最后构建合理热度指标体系，对热点问题进行评价。而在此过程中，我们应该注意到，相似度的计算复杂，特征过多，热度指标构建，特殊地点人群识别带来的困难等问题。

对于问题三，答复意见的评价，我们需要从答复的完整性、相关性、可解释性等角度对答复意见的质量给出一套评价方案，即，我们需要考虑答复意见的内容是否与问题相关，答复格式是否满足某种规范，答复意见中内容的相关解释是否可行。

1.3 挖掘目标

网络问政凭借其信息传递速度快、空间距离小、成本低廉等优势，已成为政府收集民意和群众提出意见的主要渠道。本次建模目标是根据网络问政平台收集的各项信息与数据，利用 jieba 中文分词工具进行分词、Complement NB 算法、K-means 聚类及潜在语义分析，达到以下三个目标：

（1）对大量群众留言数据（其中包含结构化和非结构化文本数据），进行基本的数据预处理、中文分词、停用词过滤后，根据附件中的留言主题、留言信息等描述的具体内容，结合留言问题的关键词、特点等，进行一级标签的分类；

（2）对热点问题的挖掘，先对海量留言文本数据进行同问题一的预处理，然后进行文本特征抽取，通过 TF-IDF 算法转化为 TF-IDF 权值向量后降维，再采用 K-means 聚类，识别相似留言。最后，利用 HanLP 进行命名实体识别热点问题中的特定地点和特定人群。根据分类结果，通过每个类别的留言信息的数量多少、点赞数、反对数、时间间隔来进行加权排序，综合排序结果，选取热点问题。

（3）从留言信息的相关性、完整性、可解释性等角度，构建出相应的评价指标对留言答复意见的质量给出一套趋于完整的评价方案，并根据此评价方案尝试操作，以实现对留言答复意见的精准度高的评价。

二、留言分类

2.1 流程图

步骤一：数据预处理，对附件 2 中的‘一级分类’文本信息数据进行数值化处理，对‘留言信息’（包括‘留言主题’和‘留言详情’）进行去除重复项及空行、中文文本分词、停用词过滤的操作，以便我们进行后续的结果分析；

步骤二：文本向量化，基于 TF-IDF 权重法提取留言主题及详情中的关键词，构造出词汇-文本矩阵，进而利用奇异值分解算法进行语义空间降维，去除同义词的影响，简化我们的后续的相关计算；

步骤三：文本分类，根据 Complement NB 算法构建数学模型，对留言信息进行分类处理，再采用 F-score 方法对分类结果进行分析。

2.2 数据预处理

2.2.1 留言信息的去重、去空

在我们浏览题目中给出的文本数据信息时，我们发现留言数据信息中存在着一些重复的留言数据，以及空格或无意义现象。重复的信息以及空格、无意义等数据在我们之后的操作过程中是没有多大帮助的，反而会增加我们的数据处理量，降低数据处理的效率，所以我们首先需要去掉原始数据中的无效、重复的信息。在这里，我们对留言信息数据进行去重、去空处理，来方便我们后面的操作。

2.2.2 将留言信息表中文分词

在把留言信息数据转化为词向量之前，我们首先要把非结构化的文本信息（即留言信息）转换为计算机能够识别的结构化的文本信息。而在附件 2 留言信息表中，题目以中文文本的方式给出了数据，我们为了使留言信息能够更加完整地体现出来，决定将留言主题和留言详情合并为留言信息，再对这些留言信息进行中文分词处理，在这里，我们采用的是 python 的中文分词包 jieba 来进行分词操作。

具体的分词示例结果，我们采用词云图的方式来展现，将词频和词云图相结合，在整体上做成爱心的形状，便于观看。其中，部分分词结果如图 2.1 所示

图 2.1 一级标签为劳动和社会保障的分词结果

从图 1 中，我们可以直观地看到，公司、单位、工作、领导、人员、职工、工资、社保、退休等词的出现频率显著大于其他词语，其中的部分词语恰好体现出这些留言信息的标签——劳动和社会保障，这即达到我们进行中文分词的目的，为后续的词向量化提供便捷。

2.3 文本特征抽取

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，提高我们对文本处理的速度和效率。常用的方法有词频-逆向文档频率(TF-IDF)、互信息、信息增益、χ^2统计等。

(一)互信息(Mutual Information，MI)

在统计语言模型中，互信息用于表示两变个量间(表征f和类别c之间)的相关性。其互信息记作$MI(f,c)$可由下式计算：

$$M(f,c) = \log\left(\frac{p(f,c)}{p(c)p(f)}\right)$$

互信息没有考虑单词发生的频度，这是互信息一个很大的缺点，它导致互信息评估函数经常倾向于选择稀有词。

（二）χ^2统计（CHI）

χ^2统计方法度量词和文条t档类别之间的相关程度，并假设t和c之间符合具有一阶自由度的χ^2分布。令N表示训练语料中的文本总数，c为某一特定类别，t表示特定的词条，A表示属于c类且包含t的文档频数，B表示不属于c类且包含t的文档频数，C表示属于c类但不包含t的文档频数，D是既不属于c类也不包含t的文档频数，则t对于c的χ^2值由下式计算：

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

词条对于某类的χ^2统计值越高，它与该类之间的相关性越大，携带的类别信息也较多。

（三）信息增益(Information Gain, IG)

IG是一种在机器学习领域应用较为广泛的特征选择方法。它从信息论角度出发，以各特征取值情况来划分学习样本空间，根据所获信息增益的多少来筛选有效的特征。IG可以用下式表示：

$$IG(t) = p(t)\sum_{i=1}^{m} p(C_i|t)1g\frac{p(C_i|t)}{p(C_i)} + p(\bar{t})\sum_{i=1}^{m} p(C_i|\bar{t})1g\frac{p(C_i|\bar{t})}{p(C_i)}$$

式中$p(C_i|t)$表示文本中出现词条t时文本属于C_i的概率，$p(C_i|\bar{t})$表示文本中不出现词条t时文本属于C_i的概率，$p(C_i)$表示类别出现的概率，$p(t)$表示语料中包含词条t的文本的频率。

（四）词频-逆向文档频率(Term Frequency-Inverse Documentation Frequency, TF-IDF)

传统的 TF-IDF：

词频(Term Frequency，TF)是词语在文本中出现的频率，如果某一个词在一个文本中出现的越多，它的权重就越高，基本公式：

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

以上式子$n_{i,j}$中分子是该词在文件d_j中的出现次数，而分母则是在文件d_j中所有字词的出现次数之和。

逆向文档频率（Inverse Documentation Frequency，IDF)是指在少数文本中出现的词的权重比在多数文本中出现的词的权重高，因为在聚类中这些词更具有区分能力。它的基本公式如下：

$$idf_i = log\frac{N}{|\{j:t_i \in d_j\}|}$$

其中，N:语料库中的文件总数，|{j:$t_i \in d_j$}|:包含词语t_i的文件数目（即$n_{i,j} \neq 0$的文件数目）如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用 1+|{j:$t_i \in d_j$}|。

在 Shannon 的信息论的解释中:如果特征项在所有文本中出现的频率越高，它所包含的信息熵越小；如果特征项集中在少数文本中，即在少数文本中出现频率较高，则它所具有的信息熵也较高。

最后可以得出：

$$w_{ij} = tf_{ij} \times idf_i$$

这就是词的权重，上述方法各有利弊，信息增益的计算量相对其它几种方法较大；而对于互信息方法，在相同的条件概率下，稀有名词会比一般词获得更高的得分；卡方方法基于卡方分布，如果这种分布被打破，则对低频词不可靠。因此在本文中，我们采用目前公认的比较有效的 TF-IDF 算法抽取特征词条，将权重按照从大到小的顺序排列，抽取权重最大的 60000 个特征词作为候选特征词。

2.4 留言文本词向量化

在我们对留言文本数据进行分词后发现，由于计算机不能够直接对文本信息进行处理，所以需要我们先对文本信息进行处理，将留言文本信息表示成为计算机能够直接处理的形式，即将文本数字化。文本表示也可以称为文本特征表达，它不仅要求能够真实准确的反映出我们的文档内容，而且要对不同的文档具有区分的能力，需要把留言信息中的这些词语都转换为词向量，以供我们在挖掘分析时使用。根据图 2.1 分析，我们知道了并不是所有词频结果都能体现出正确的留言信息标签，仍存在一大部分的词语会对我们的分类结果产生干扰，为解决问题，我们在选择算法时考虑到 one-hot 算法虽然能够解决离散值特征的问题，在一定程度上起到扩充特征的作用，但是没有考虑到词之间顺序的词袋模型，假设条件时词之间相互独立则得到的特征会过于离散稀疏，而 TF-IDF 算法简单快速，结果比较符合实际。所以在这里，我们最终决定采用的是 TF-IDF 算法，把留言信息中的那些词语转换为词向量，而没有采用 one-hot 算法。

TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）词频（TF）=某个词在文本中出现的次数

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数，即:

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

$$\text{或词频（TF）} = \frac{\text{某个词在文本中出现的次数}}{\text{该文本中出现次数最多的词出现的次数}}$$

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（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布就越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率（IDF）} = \log\left(\frac{\text{谱料库中的文本总数}}{\text{包含该词的文本总数}+1}\right)$$

第三步,计算 TF-IDF 值(term frequency－inverse document frequency)。

TF-IDF = 逆文档频率（IDF）×词频（TF）

实际分析得出 TF-IDF 值与一个词在留言信息表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的留言信息中文本的关键词。

生成 TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找出每条留言信息的十个关键词；

（2）将每条留言信息提取的关键词，合并成一个集合，计算每条留言信息对于这

个集合中词的词频，如果没有则记为 0；

（3）生成各条留言信息的 TF-IDF 权重向量。

2.5 文本分类

2.5.1 朴素贝叶斯算法模型构建

解决分类问题的算法很多 ，单一的分类方法主要包括：决策树、贝叶斯、人工神经网络、K-近邻（KNN）、支持向量机和基于关联规则的分类等；朴素贝叶斯算法（Naive Bayesian algorithm）属于监督学习生成模型，实现简单，没有迭代并有坚实的数学理论基础（即贝叶斯定理）作为支撑；第二就是对大数量训练和查询时具有较高的速度，即使使用超大规模的训练集，针对每个项目通常也只会有相对较少的特征数，并且对项目的训练和分类也仅仅是特征概率的数学运算而已；第三就是对小规模的数据表现很好，能个处理多分类任务，适合增量式训练（即可以实时的对新增的样本进行训练）；第四就是对缺失数据不太敏感，算法也比较简单，常用于文本分类；第五就是朴素贝叶斯对结果解释容易理解。所以我们选用朴素贝叶斯算法对文本进行分类。

贝叶斯定理阐述了以下关系:

$$P(y|x_1,x_2,\cdots\cdots,x_n)=\frac{P(y)P(x_1,\cdots\cdots,x_n|y)}{P(x_1,\cdots\cdots,x_n)}$$

使用简单(naive)的假设-每对特征之间都相互独立:

$$P(y,x_1,\cdots\cdots,x_{i-1},x_{i+1},\cdots\cdots,x_n)=P(x_i|y)$$

对于所有的i都成立，这个关系式可以简化为

$$P(y|x_1,\cdots\cdots,x_n)=\frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_1,\cdots\cdots,x_n)}$$

由于在给定的输入中$P(x_1,\cdots\cdots,x_n)$是一个常量，我们使用下面的分类规则:

$$P(y|x_1,\cdots\cdots,x_n)\propto P(y)\prod_{i=1}^{n}P(x_i|y)$$

$$\Downarrow$$

$$\hat{y}=arg\max_{y}P(y)\prod_{i=1}^{n}P(x_i|y)$$

贝叶斯算法是以贝叶斯原理为基础，使用概率统计的知识对样本数据集进行分类。贝叶斯分类算法在数据集较大的情况下表现出较高的准确率，同时算法本身也比较简单。朴素贝叶斯算法是在贝叶斯算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立。也就是说没有哪个属性变量对于决策结果来说占有着较大的比重，也没有哪个属性变量对于决策结果占有着较小的比重。虽然这个简化方式在一定程度上降低了贝叶斯分类算法的分类效果，但是在我们对留言信息分类的实际运用中，极大地简化了贝叶斯算法的复杂性。补充朴素贝叶斯(CNB)算法是标准多项式朴素贝叶斯(MNB)算法的一种改进，特别适用于不平衡数

据集。具体来说，CNB 使用来自每个类的补数的统计数据来计算模型的权重。CNB 的发明者的研究表明，CNB 的参数估计比 MNB 的参数估计更稳定。此外，我们根据 CNB 算法得到的精确度为 0.8920，而由 MNB 算法得到的精确度为 0.8795,所以我们最后决定选择 CNB 算法来进行留言信息的一级标签分类。
计算权重的步骤如下:

$$\widehat{\theta_{cl}} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}}$$

$$w_{ci} = log\widehat{\theta_{cl}}$$

$$w_{ci} = \frac{w_{ci}}{\sum_j |w_{cj}|}$$

其中对不在类 c 中的所有记录 j 求和，d_{ij}可以是文档 j 中词语 i 的计数，

a_i是像 MNB 中一样的平滑超参数，同时

$$\alpha = \sum_i \alpha_i$$

第二个归一化解决了长记录主导 MNB 参数估计的问题。分类规则为:

$$\hat{c} = arg\min_c \sum_i t_i w_{ci}$$

即将记录分配给补充匹配度最低的类。

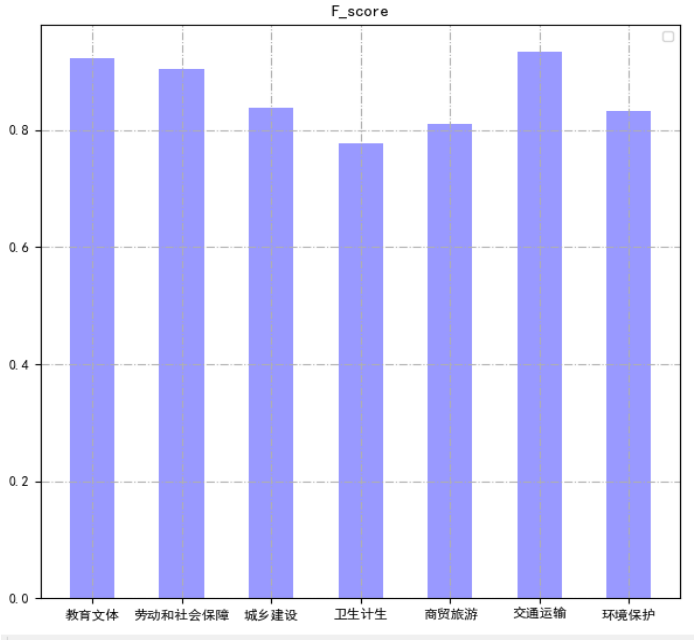
2.5.2 分类结果分析

在深度学习中，精确率(Precision)和召回率(Recall)是常用的评价模型性能的指标，从公式上看两者并没有太大的关系，但是实际中两者是相互制约的。我们都希望模型的精确率和召回率都很高，但是当精确率高的时候，召回率往往较低；召回率较高的时候，精确率往往较低。往往需要我们对模型的精确率和召回率做出取舍：比如在一般的搜索任务时，在保证召回率的同时，尽量提高精确率；在癌症检测、金融诈骗任务时，在保证精确率的同时，尽量提高召回率。但是在很多时候，我们需要综合权衡这 2 个指标，这就引出了一个新的指标 F-Score，这是综合考虑 Precision 和 Recall 的调和值.

$$F - Score = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

当$\beta = 1$时，成为$F_1 - Score$，这时召回率和精确率都很重要，权重相同。当有些情况下我们认为精确率更为重要，那就调整β的值小于 1，如果我们认为召回率更加重要，那就调整 β 的值大于 1，比如$F_2 - Score$。这里精确率更为重要，用到的是$F_1 - Score$。

分类结果可视化如表格 2.1：

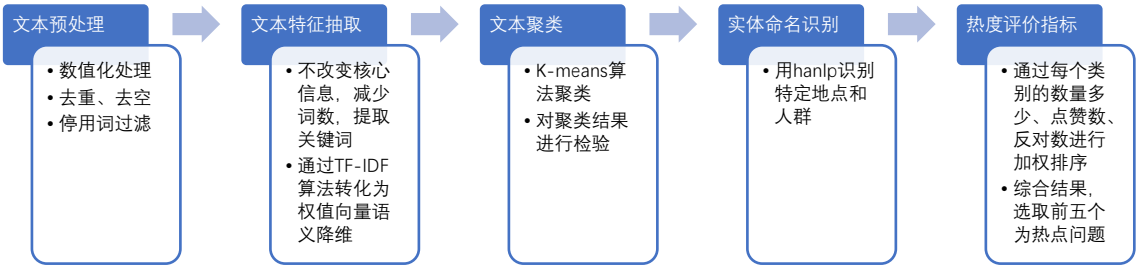


表格 2.1F-Score 精确率可视化

在表格 1 中，我们可以看到交通运输的精确率最高，其次就是教育文体、劳动和社会保障，卫生计生的精确率是最低的。对此，我们可以根据附件二的留言信息中的留言详情来进行分析，其中卫生计生部分受文本语义带来的词语交叉影响较大，如，附件二中属于卫生计生分类里的留言详情中，有留言问题描述问公共卫生服务补助经费问题的，这其实也可归属于劳动和社会保障分类里；有留言问题说医疗资源城市和农村分配不平均，这也可归属于城乡建设分类中，所以我们得到的结果中才会显示出它的精确率不高。

三、热点问题挖掘

3.1 流程图



步骤一，文本预处理，对附件三中的文本数据数值化处理，去除重复项及空行、

中文文本分词、停用词过滤，以便后续分析。由于留言信息中留言主题和留言详情中词语的权重不同，所以我们把它们分开来处理；

步骤二，文本特征抽取：在经过步骤一预处理后，为了简化计算，提高文本处理的效率，进行文本特征抽取，再通过 TF-IDF 算法转化为 TF-IDF 权值向量；

步骤三，文本聚类，利用 K-means 聚类算法对附件 3 中的某一时段内反映的问题留言进行聚类，然后再对聚类结果进行检验；

步骤四，实体命名识别:利用 HanLP 进行命名实体识别热点问题中的特定地点和特定人群；

步骤五，热度评价指标:根据得出的分类结果，通过每个类别的留言信息的数量多少、点赞数、反对数以及时间来进行加权排序，综合排序结果，选取前五个问题为热点问题。

3.2 数据预处理

对于留言详情，首先进行去重、去空、分词，然后再进行词频统计，由于文本数据量较大，我们将过滤掉出现的低频词，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；对于留言主题，我们直接采用去重、分词后的词，通过上述提取的新语料库建立词典。

3.3 文本特征抽取

3.3.1 提取留言信息特征词

由于留言主题和留言详情中包含的特征词的权重不同，主题包含的权重较大，详情包含的权重较小，所以我们分开提取留言主题和留言详情中的特征词。对于留言主题我们直接进行预处理，留言详情进行预处理和提取关键词，再把处理后的留言主题和留言详情拼接在一起，形成留言信息特征词。
留言信息部分特征词如下图 3.1：

```
留言编号
188006          A3区 一米阳光 婚纱 艺术摄影 合法 纳税 一米阳光 影楼 艺术摄影 居民楼 婚纱
188007        咨询 A6区 道路 命名 规划 初步 成果 公示 城乡 门牌 路名 更换 名牌 成果 时候
188031            A7县 春华 镇金鼎村 水泥路 自来水 到户 到户 水泥路 形象工程 村组 部分
188039        A2区 黄兴路 步行街 古道 巷 住户 卫生间 粪便 外排 外排 解决 卫生间 住户 粪便
188059    A市 A3区 中海 国际 社区 三期 四期 空地 夜间 施工 噪音 扰民 城管 噪音 施工 ...
                                      ...
360110         A市 经济学院 寒假 过年 期间 组织 学生 工厂 工作 家长 学子 求学 学生 西地省
360111               A市 经济学院 组织 学生 外出 打工 学校 孩子 流水线 外省 实习
360112                  A市 经济学院 强制 学生 实习 实习 学生 学校 安排 选择权
360113               A市 经济学院 强制 学生 外出 实习 实习 物流 寝室 回来 强制
360114            A市 经济学院 体育 学院 变相 强制 实习 实训 公司 儿童 老师 学期结束
Name: sp, Length: 4326, dtype: object
```

图 3.1 留言信息部分特征词

3.3.2 特征信息向量化

上述文本特征全部抽取构建一个词袋，根据留言信息的特征项对应词袋中的位置，组成统一维数的向量：

$$C = (t_1, t_2, \ldots t_n)$$

其中 C 为词袋集合，t_n是每个词在向量中对应的位置。

这样留言信息根据词袋组成了同一维数的词向量，再通过 TF-IDF 将它们向量化得到一个词汇-文本矩阵：

$$\begin{array}{c} \\ t_1 \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{ccc} d_1 & \ldots & d_n \end{array} \\ \begin{pmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{n1} & \cdots & \omega_{nn} \end{pmatrix} \end{array}$$

将留言信息的特征词转化为词频矩阵，统计每个词语的 TF-TDF 权值，即转化为 TF-IDF 权值向量。

3.3.3 奇异值分解降维

我们为了获得更好的留言信息的分类精度，会选择增加一些特征词。但是当特征词达到一定的数量之后，再继续增加特征词，可能存在一些特征词与最终精度的关系很小，加入新的特征词会导致增加冗余，却不能提高精度，甚至是不仅不能提高聚类的质量反而导致聚类质量的下降。我们选择使用的 PCA 就是一种典型的无监督线性降维方法。

PCA 降维理论介绍：

PCA 将数据投射到一个低维子空间实现降维。例如，二维数据集降维就是把点投射成一条线，数据集的每个样本都可以用一个值表示，不需要两个值。三维数据集可以降成二维，就是把变量映射成一个平面。一般情况下，n 维数据集可以通过映射降成 k 维子空间，其中 k 是选取的主成分数目。 主成分是原始数据集的属性集合乘以一个投影矩阵得到的，所以主成分不一定只有一个，而是由原始数据集的属性集合通过矩阵变换形成的新的属性集合。

PCA 降维具体步骤：

（1）数据预处理后求协方差矩阵：

a. 协方差（Covariance)：

协方差是度量两个变量的变动的同步程度，也就是度量两个变量线性相关性程度。

如果两个变量的协方差为 0，则统计学上认为二者线性无关。注意两个无关的变量并非完全独立，只是没有线性相关性而已；

如果协方差不为 0，如果大于 0 表示正相关，小于 0 表示负相关；

当协方差大于 0 时，一个变量增大是另一个变量也会增大；

当协方差小于 0 时，一个变量增大是另一个变量会减小；

协方差计算公式如下：

$$cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

b.协方差矩阵：协方差矩阵就是一个数据集里的变量两两之间协方差所组成的，它肯定是一个方阵，行是属性节点集合，列和行节点顺序一样。矩阵对角线上的，即属性自身的方差，C 矩阵的第(i,j)个元素就是数据集中第 i 和第 j 个元素的协方差。（i≠j）

（2）求协方差矩阵的特征值和特征向量，得到投影矩阵：

特征向量和特征值只能由方阵得出，且且并非所有方阵都有特征向量和特征值。

如果一个矩阵有特征向量和特征值，那么它的每个维度（特征）都有一对特征向量和特征值。

特征向量特征值求解如下式：

$$A\vec{v}=\lambda\vec{v}$$

（υ是特征向量，A 是协方差矩阵，它是一个方阵；λ是特征值。）

将协方差矩阵按照特征值大小排序特征向量，得到特征向量为每列组成的矩阵，从大到小取前 K 个特征值对应的特征向量作为列组成投影矩阵 P。

（3）得出 PCA 降维后的新样本数据：

如果我们要把 5 维的数据降成 3 维，那么我们就要用一个 3 维矩阵做投影矩阵。我们用数据矩阵点乘投影矩阵。将原先 m*n 的样本矩阵 A（不是上面的协方差矩阵）变换成一个 m*r 的样本矩阵 B，这样就会使得本来有 n 个 feature 的 A，变成了有 r 个 feature 的 B（r < n)，这 r 个其实就是对 n 个 feature 的 A 的一种提炼，我们就把这个称为 feature 的压缩。用数学语言表示就是：

$$A_{m\times n}P_{n\times r}=\widetilde{A_{m\times r}}$$

样本矩阵 B 就是 PCA 降维后的新矩阵。

对于我们的奇异值分解降维，PCA 特征值分解是一个提取矩阵特征很不错的方法，但是它只是针对方阵而言的。在本题中，文本数据信息形成的一个 m*n 的矩阵就不是方阵，而是一个普通的矩阵。我们怎样才能描述这样普通的矩阵的重要特征呢？

我们可以再看一下 SVD 公式（m 为样本数，n 为属性数）：

$$B=U\sum V^T$$

其中，B 是$m\times n$阶矩阵，则 U 是$m\times m$阶酉矩阵，$\sum$是$m\times n$阶非负实数对角

矩阵，而V^T，即 V 的共轭转置，是$m \times n$阶酉矩阵。

U，V 为正交矩阵，即乘以自己的转置得到 E；

U 称为数据矩阵的左奇异值向量；

V 称为数据矩阵的右奇异值向量；

Σ只有对角元素，且对角元素称为奇异值。

注意到我们的 SVD 也可以得到协方差矩阵$X^T X$最大的 d 个特征向量张成的矩阵V_B；但是 SVD 有一个好处，有一些 SVD 的实现算法可以不用先求出协方差矩阵$X^T X$，也能求出我们的右奇异矩阵V_B。也就是说，我们的 PCA 算法可以不用做特征分解，而是做 SVD 来完成。这个方法在样本量很大的时候很有效。实际上，在 scikit-learn 的 PCA 算法的背后真正的实现用的就是 SVD，而不是我们所认为的暴力特征分解。但除此之外，由样本矩阵 SVD 分解的左奇异矩阵结合奇异值矩阵，也可以对样本矩阵进行列降维。降维之前的数据为 4326×10715 的巨大矩阵，计算起来较困难，降维之后的数据为 4326×3438，维数有了一定程度的降低，降维之后提升了数据处理的效率，大大减少了程序运行的时间。

3.4 文本聚类

利用 K-means 聚类算法对附件 3 中的某一时段内反映的问题留言进行聚类，然后再根据 K-means 聚类算法的评价指标轮廓系数对聚类结果进行检验；生成留言信息的 TF-IDF 权重向量后，根据每个留言信息的 TF-IDF 权重向量，对留言信息进行分类。这里采用 K-means 算法把留言信息分类。

K-mean 聚类的原理如下：

假如有一个包含 n 个 d 维数据点的数据集

$$X=\{x_1, x_2, \cdots, x_i, \cdots, x_n\}$$

其中$x_i \in R^d$,K-means 聚类讲数据集 X 组织为 K 个划分 $C=\{c_k, i=1,2,\cdots,K\}$。每个划分代表一个类$c_k$有一个类别中心$\mu_i$，选取欧式距离作为相似性和距离判断准则，计算该类内个点到聚类中心μ_i的距离平方和

$$J(c_k)=\sum_{x_{i\in C_k}} \| x_i-\mu_k \|^2$$

聚类目标是使各类总的距离平方和 $J(C)=\sum_{k-1}^{K} J(c_k)$ 最小，$J(C)=\sum_{k-1}^{K} J(c_k)=\sum_{k-1}^{K}\sum_{x_{i\in C_i}} \| x_i-\mu_k \|^2=\sum_{k-1}^{K}\sum_{i-1}^{n} d_{ki} \| x_i-\mu_i \|^2$

其中，$d_{ki}=\begin{cases}1, 若 x_i \in c_i \\ 0, 若 x_i \notin c_i\end{cases}$,所以根据最小二乘法和拉格朗日原理，聚类中心$\mu_k$应该取为类别$c_k$类各数据点的平均值。

K-mean 聚类的算法步骤如下：

1、从 X 中随机取 K 个元素，作为 K 个簇的各自的中心。

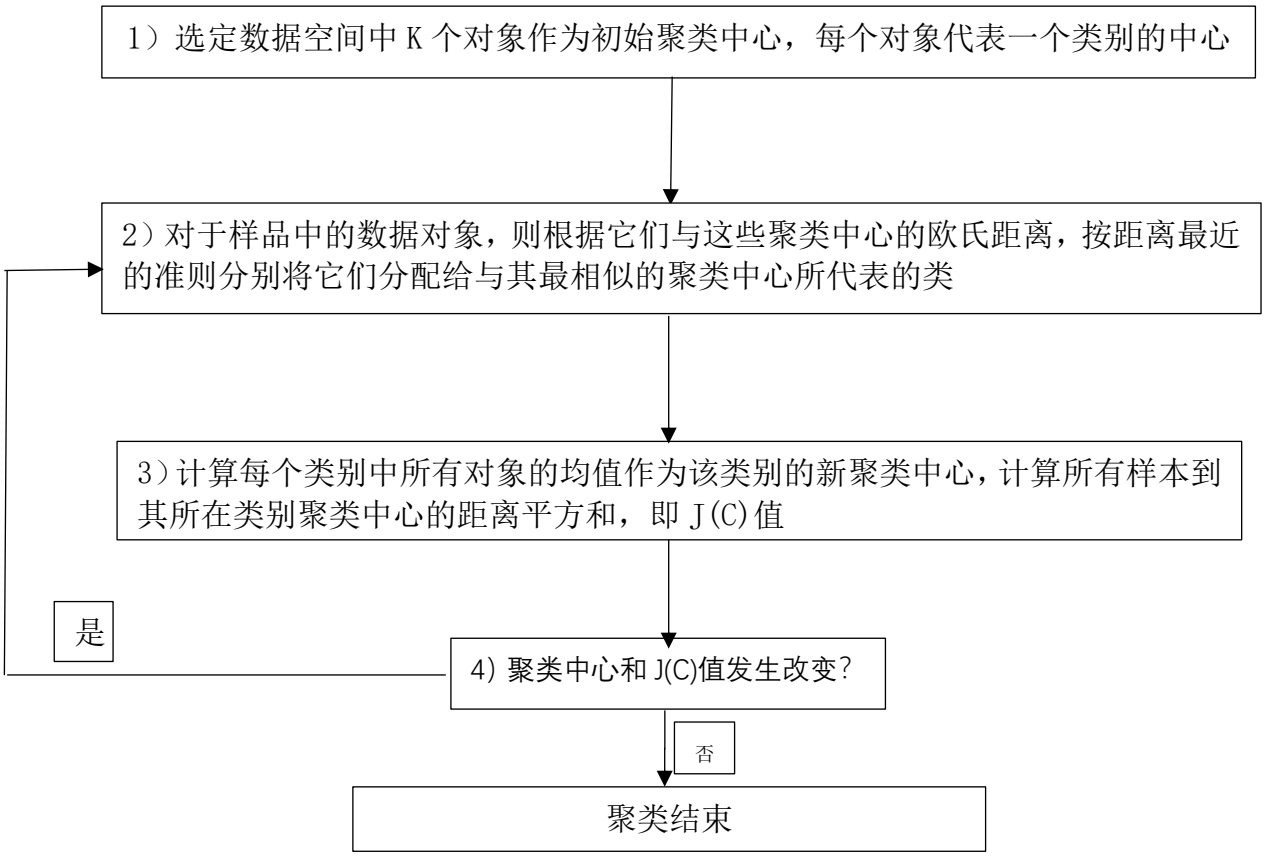
2、分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

4、将 X 中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化，将结果输出。

K-means 算法流程图如下



轮廓系数（Silhouette Coefficient），是聚类效果好坏的一种评价方式。它结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。对于其中的一个点 i 来说：

计算 a(i)=average(i 向量到所有它属于的簇中其它点的距离)

计算 b(i)=min(i 向量到与它相邻最近的一簇内的所有点的平均距离)

a(i)：i 向量到同一簇内其他点不相似程度的平均值

b(i)：i 向量到其他簇的平均不相似程度的最小值

那么 i 向量轮廓系数就为：

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

可见轮廓系数的值是介于[-1,1]，越趋近于 1 代表内聚度和分离度都相对较优。将所有点的轮廓系数求平均，就是该聚类结果总的轮廓系数。根据该轮廓系数，取 K=720。

部分聚类结果生成如图 3.2:

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	label_1
188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019/2/28 11:25:05	\n\t\t\t\t\t\n\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋320，...	0	0	679
188007	A00074795	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	\n\t\t\t\t\t\n\t\t\t\t\tA市A6区道路命名规划已经初步成果公示文件，...	0	1	295
188031	A00040066	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	\n\t\t\t\t\t\n\t\t\t\t\t本人系春华镇金鼎村七里组村民，不知是否有相关...	0	1	67
188039	A00081379	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	\n\t\t\t\t\t\n\t\t\t\t\t靠近黄兴路步行街，城南路街道、大古道巷、一步...	0	1	403
188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	\n\t\t\t\t\t\n\t\t\t\t\tA市A3区中海国际社区三期四期中间，即蓝天某...	0	0	289
...	...	...	...	...	...	...	...
360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	\n\t\t\t\t\t\t\t关于西地省A市经济学院寒假过年期间组织学生去工厂工作...	0	0	180
360111	A1204455	A市经济学院组织学生外出打工合理吗?	2019-11-05 10:31:38	\n\t\t\t\t\t\t\t一名中职院校的学生,学校组织我们学生在外边打工,...	1	0	180
360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	\n\t\t\t\t\t\t\t各位领导干部大家好，我是A市经济学院的一名学生，...	0	0	660
360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	\n\t\t\t\t\t\t\tA市经济学院强制16届电子商务跟企业物流专业实习。其中我...	3	0	660
360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	\n\t\t\t\t\t\n\t\t\t\t\t 书记您好，我是来自西地省经济学院体育学院...	9	0	660

图 3.2 部分聚类结果

3.5 热度评价指标

网络的便捷性使问政平台已逐渐成为政府部门和人民群众沟通交流的重要方式之一，这就导致了各种社情民意的文本数据急剧增加，而其中的部分社情民意往往具有时效性，这就属于热点问题，需要相关工作人员的及时处理。内容繁杂，数据量大，时效性有限，这给依靠人工来进行热点挖掘的相关部门人员工作带来了极大的不便。在热点留言信息中，我们可以通过各个类别问题的数量多少看出问题热度大小，根据点赞数和反对数也可以看出人们对该热点问题所持有的态度，人们对该信息进行点赞说明他们有同样的问题，或者较为认同此条留言，反对则说明他们可能存在同样的问题但不认同此留言，或者该留言与实际存在偏差，点赞数和反对数均可作为此留言的热度评价指标。所以我们依据聚类结果，通过每个类别的数量多少、点赞数、反对数进行加权排序，综合排序结果，选取前五个问题为热点问题。

在构建热度评价指标时，我们参考了关于新闻的热度算法基本原理，原理如下，

新闻热度分=初始热度分+用户交互产生的热度分 - 随时间衰减的热度分，

即，Score=S_0+S(Users) - S(Time)

对比我们的热点问题，思考发现，其中的用户留言占到的权重相对比较大，而由用户关注产生的热度分，即点赞数和反对数的权重相对较小，同时，热点问题是有较强时效性的内容，因此热点问题留言之后，热度必须随着留言间隔时间变大、变得陈旧而衰减。由于热点问题的强时效性，已经留言的问题热度值随着时间间隔变大而衰减，并且衰减趋势直至趋近于零热度。换句话说，如果一条留言要一直处于很靠前的位置，随着时间的推移它必须要有越来越多的点赞数或者反对数来维持。结合上述分析，考虑到底数为 e 的指数函数图像在负无穷大到零区域内处于比较平缓的趋势，相对其他函数来说底数为 e 的指数函数在负无穷大到零区域，比较适合构建关于时间的衰减。最后综合考虑，我们构建的热度评价指标如下，

热点问题分=（初始热度分+用户关注产生的热度分×0.5）e^{-kt}

其中初始热度值对应留言的总条数，用户关注产生的热度分对应点赞数和反对数之和，k 为冷却系数（我们这里加入 100 分的文章第 60 天希望下降 92%，反推得到 k=720），t 为问题初次留言到最后一次留言之间的时间间隔。

问题留言进入系统后，系统为之赋予一个初始热度值（即留言的总条数），该问题留言就进入了热度列表进行排序；随着问题不断被用户提出、表明态度（点赞或者反对）、留言时间间隔变化等，这些留言指标被视作帮助问题提升热度（即用户关注产生的热度）。

通过上面处理的结果，结合所构建的热度指标，对留言信息分类结果进行排序，并选出排名前五的类别。

3.6 命名实体识别

在将热点问题的前五名选出后，需要将留言信息中反映问题特定的地点或者人群识别出来。这里需要用到命名实体识别，常用的命名实体识别方法主要有 HanLP，NLTK 等。HanLP 是一系列模型与算法组成的 NLP 工具包，由大快搜索主导并完全开源，目标是普及自然语言处理在生产环境中的应用。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。NLTK 其本身自带的语料大部分是英文的，考虑到处理的对象为中文信息，NLTK 对中文识别并不是那么友好，这里我们选择 HanLP 中文自然语言处理进行命名实体识别热点问题中的特定地点和特定人群。

部分识别结果如图 3.3：

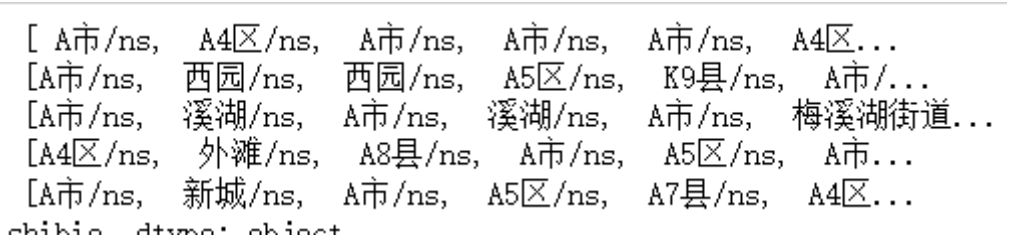


图 3.3 命名实体识别部分结果

3.7 提取摘要

考虑到目标表格有问题描述的需要，我们利用 HanLP 对排名前五的热点问题提取摘要。

3.8 热点问题表

3.1 热点问题表

热度排名	问题 ID	热度指标	时间范围	地点/人群	问题描述
1	1	943	2019-1-11 至 2019-7-8	西地省 A 市 A4 区	58 车贷案无进展
2	2	849	2019-4-13 至 2019-9-19	A 市五矿万境 K9 县	房屋质量存在问题
3	3	683	2019-3-12 至 2019-9-18	A 市梅溪湖	入学条件不明显
4	4	263	2019-1-18 至 2019-9-6	A 市绿地海外滩	高铁影响周边小区
5	5	114	2019-12-10 至 2019-9-2	A 市	小区水费出现问题

根据生成的热点问题留言明细表，我们可以看出有些留言信息的排名较为靠前的但条数不一定多，但是其用户点赞数和反对数较多或者时间间隔相对比较短，正是用户留言数点赞数、反对数以及时间根据热度指标的综合权重比较的结果。在“智慧政务”中根据热点问题表可以很直观的看出哪些问题是当下人们所关注和急需解决的问题，以便及时的反应给政府服务中心，尽快引起相关部门重视，从而使问题得到解决。

四、意见答复评价

4.1 相关性

潜在语义分析 LSA（Latent Semantic Analysis ）也叫作潜在语义索引 LSI（ Latent Semantic Indexing ） 顾名思义是通过分析文章（documents ）来挖掘文章的潜在意思或语义（concepts ）。LSA 被广泛用于文献检索，文本分类，垃圾邮件过滤，语言识别，模式检索以及文章评估自动化等场景。这里我们通过 LSA 解决留言信息和意见答复间相关性的问题。

潜在语义分析的基本原理是将文章和单词懂映射到语义空间（ “concept” space ）上，并在该空间进行对比分析。LSA 是一种自然语言处理中用到的方法，其通过“矢量语义空间”来提取文档与词中的“概念”，进而分析文档与词之间的关系。LSA 的基本假设是，如果两个词多次出现在同一文档中，则这两个词在语义上具有相似性。LSA 使用大量的文本上构建一个矩阵，这个矩阵的一行代表一个词，一列代表一个文档，矩阵元素代表该词在该文档中出现的次数，然后再此矩阵上使用奇异值分解（SVD）来保留列信息的情况下减少矩阵行数，之后每两个词语的相似性则可以通过其行向量的 cos 值（或者归一化之后使用向量点乘）来进行标示，此值越接近于 1 则说明两个词语越相似，越接近于 0 则说明越不相似。

潜在语义分析（LSA）的理论方法如下：

LSA 的关键思想是将文档和词汇映射到一个低维的向量空间，即潜在语义空间。LSA 利用奇异值分解的(Singular Value Decomposition, 缩写为 SVD)方法实现这种降维。下面介绍奇异值分解定理。

定理 1:任何一个矩阵X_{min}, X的秩记为 r，均可分解为两个正交矩阵和一个对角矩阵的乘积:

$$X = TSD^T$$

$T_{min} = (t_1, t_2, \cdots, t_r)$为正交矩阵，其中$t_1, t_2, \cdots, t_r$为 X 的左奇异向量，并且是$XX^T$的特征向量;$S_\infty = diag(\sigma_1, \sigma_2, \cdots, \sigma_r)$为对角阵，$\sigma_1, \sigma_2, \cdots, \sigma_r$为 A 的所有奇异值，同时也是$XX^T$或$X^TX$所有特征值的平方根,并且满足关系$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0; D_\infty = (d_1, d_2, \cdots, d_r)$为正交矩阵,其中$d_1, d_2, \cdots, d_r$为 X 的右奇异向量,并且是$X^TX$的特征向量。因此,矩阵 X 可以用下式表达:

$$X = \sigma_1 t_1 d_1^\tau + \sigma_2 t_2 d_2^\tau + \cdots + \sigma_r t_r d_r^\tau$$

SVD 的优势是通过一种简单的方法，就可以使原矩阵塌陷、找到一个规模大大减小的近似矩阵。LSA 在 SVD 的基础上保留最大的 k 个奇异值，而忽略其他较小的奇异值，k 就是低维空间的维数。然后进行奇异值分解反运算，得到原始矩阵的近似阵。k 应当足够小，去除掉不该保留的噪声，又要足够大以保留语义空间中的主要框架，通常根据经验来说，当 X 的秩为几千的时候，k 的取值为几百。

由于矩阵的奇异值正好对应由 EL=$(X_y: \| y \|_2 = 1)$所定义的超椭球 EL 的各半轴之长，所以 LSA 中降秩的过程可以视为去除了语义空间中代表低信息量(即噪声)的自由度，而保留了代表语义空间中主要信息的自由度。

令,$S_k = diag(\sigma_1, \sigma_2, \cdots, \sigma_k), T_k = (t_1, t_2, \cdots, t_k), D_k = (d_1, d_2, \cdots, d_k)$,则

$$\widehat{x_k} = T_k S_k D_k^{\tau}$$

为 X 的秩为 k 的近似阵，那么$\widehat{x_k}$与 X 到底有多近似呢?可以用下面的定理来说明:

定理 2：对于任意一个秩不大于 k 的矩阵，下式始终成立：

$$\| X - C \|_F \geq \| X - \widehat{x_k} \|_F$$

其中$\|\cdot\|_F$是矩阵的 Frobenius 范数，x 为用上述方法计算得来的 X 的近似阵。从数学分析的角度来看，与其他降维方法相比，塌陷的 SVD 式是原矩阵 X 在 k 维子空间上最佳的近似。

经过降秩得到相似矩阵：

$$\widehat{x_k} = (doc_1^*, doc_2^*, \cdots, doc_n^*) = (term_1^*, term_2^*, \cdots, term_m^*)^T$$

可以比较任意两个文档、任意两个词汇间的相似度。常用的方法是求两文档向量间(或两词汇向量间)的点乘、夹角余弦值或者是相关系数。根据上式，也可以把$\widehat{x_k}$写成：

$$\widehat{x_k} = \sigma_1 t_1 d_1^{\tau} + \sigma_2 t_2 d_2^{\tau} + \cdots + \sigma_k t_k d_k^{\tau}$$

对于其中的每一个文档向量，可以用下式表示：

$$doc_i^* = \sigma_1 d_{1,i} \cdot t_1 + \sigma_2 d_{2,i} \cdot t_2 + \cdots + \sigma_k d_{k,i} \cdot t_k$$

其中$d_{k,i}$表示向量d_k的第 i 个分量。可以将视$doc_i^* = (\sigma_1 d_{1,i} \cdot \sigma_2 d_{2,i}, \cdots, \sigma_k d_{k,i})^{\tau}$为文档 i 在 k 维向量空间中的表示，因为$t_1, t_2, \cdots, t_k$是正交向量，这样就得到了文档在低维空间中的向量表示。可以证明向量doc_i^*和doc_j^*的点乘或夹角余弦值与向量doc_i^t和doc_j^t的点乘或夹角余弦值是相同的。doc_i^t实际上是$D_k S_k$中第 i 个行向量的转置。因此可以得到如下结论:LSA 处理后，可以把矩阵$D_k S_k$中的行视为代表文档的向量，换句话说原 X 中的列向量被投影在T_k中的列向量所张成的低维空间中。同样，可以把矩阵$T_k S_k$中的行视为在低维空间中代表词汇的向量,X 中的行向量被投影在D_k中的列向量所张成的低维空间中。这里把这两个低维空间合称为潜在语义空间。如果有一查询文档 qdoc，这个文档不包含在 X 当中，如果想比较这个查询文档与 X 中任一文档的相似度，那么也要将查询文档投影到这个空间中去，而D_k中没有代表查询文档的行，可以用下述方法增加这一代表 qdoc 的行：

$$D_q = qdoc^{\tau} T_k S_k^{-1}$$

4.2 完整性

对于答复意见完整性的问题，考虑到相关部门对用户留言信息回复时大多数都会规定回复内容要满足某种规范，如开头要对留言用户进行问候加您好/你好、字数不能少于二十个子（考虑到有转交给相关部门的情况，这里简单对字数做规范，对答复意见过于敷衍的内容做一个筛查）、在对问题进行相关的解答时有回复如下/答复如下、在回复内容末尾加回复的时间。下面从以上几个方面构建答复意见完整性评价指标，将以上几个指标量化统一，进行对答复意见完整性的评价。

4.3 可解释性

首先我们给大家解读一下“解释”的含义，其实解释的意思就是在观察的基础上进行思考,合理地说明事物变化的原因，事物之间的联系，或者是事物发展的规律。在对答复意见可解释性评价时，通过对大篇幅答复意见的观察，我们构建以下几个指标对答复意见的可解释性进行评价。

根据……、调查……、研究……、符合……要求或者文章中有“《》”出现（此时说明该条答复对留言用户的问题解答时引用到相关部门的文件，此时说明该答复比较官方可解释性较强）。我们将以上几个指标量化，在对相关部门的答复意见的可解释性进行评价时，若以上几个指标该回复都满足，说明该答复意见的可解释性较强，反之则该条答复意见的可解释性不高。

4.4 时效性

考虑到答复是否及时的问题，我们可以比较用户留言时间和答复时间，通过观察附件四中的时间我们这里以时间间隔 20 天为标准，若用户留言之后 20 天以内有答复说明此条留言答复的时效性较高，反之，说明该条留言的时效性不高。在此基础上，将附件四中的留言时间和回复时间从表中抽离出来，比较两者之差Δt，若Δt小于等于 20（此时单位为天），则该条答复时效性高，若Δt大于 20，则该条答复时效性低。

4.5 评价分析

在对答复意见的质量进行评价时，综合以上相关性、完整性、可解释性、时效性等四大方面，若该条答复意见在上述四大指标中表现得都很好，则说明该条答复意见的质量较高。在对答复意见进行评价时，由于问题多样化，表述多样化，可能会出现一定的误差，相关部门在进行答复留言用户的问题时，可以提前制定相关制度，对答复意见做好规范，这样以来处理过程中可以减少其他问题带来的干扰，提升模型的准确率。

五、总结

总结本次比赛，我们分析问题，基于 TFIDF 权重法提取特征词，对文本留言信息进行分类；分析了带有不同特征词的留言文本信息中的留言主题和留言详情，利用 HanLP 进行命名实体识别热点问题中的特定地点和特定人群，得到分类结果；根据得出的分类结果，通过每个类别的留言信息的数量多少、点赞数、反对数以及时间间隔来进行加权排序，综合排序结果，选取前五个问题为热点问题。最后根据相关性、完整性、可解释性、时效性四个方面的分析判断，构建相应的模型，来帮助我们的相关工作人员对答复意见进行评价。

致谢，十分感谢泰迪杯官方给出的关于“智慧政务”模型的构建赛题，我们都

知道，在当今网络时代，网络无处不渗透进入人类的生活，政府如何更好地利用“智慧政务”平台来提升办公效率，继续为人民群众服务是一个重要而深刻的课题，我们希望我们的初步尝试能够让政府部门更好地工作，能解决民众生活中的问题。

六、参考文献

[1]王美方,刘培玉,朱振方.基于 TFIDF 的特征选择方法[J].计算机工程与设计,2007,28(23):5795-5796.

[2]周昭涛.文本聚类分析效果评价及文本表示研究[D].中国科学院研究生院(计算技术研究所), 2005.

[3]张跃,李葆青,胡玲,等.基于 K-Mean 文本聚类的研究[J].中国教育技术装备,2014(18):50-52.

[4]王礼礼.基于潜在语义索引的文本聚类算法研究[D].西南交通大学,2008.

[5]曹卫峰.中文分词关键技术研究.南京理工大学.硕士学位论文,2009.

[6]杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007.

[7]杜选.基于加权补集的朴素贝叶斯文本分类算法研究.《计算机应用与软件》,2014.2-3.

[8]刘云峰,齐欢,代建民.潜在语义分析在中文信息处理中的应用.华中科技大学系统工程研究所,2005.