“智慧政务”中的文本挖掘应用

摘要

网络问政平台凭其传递速度快、空间距离小、成本低廉、言论自由等优势，逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，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。

针对问题一：首先将附件 2 中的非结构化数据进行去重去空，本文选择 jieba 分词，收集了 4 个停用词表进行准确度对比后选择四川大学机器停用词库进行停用词过滤，运用 TextRank 算法进行关键词提取。数据预处理完成后基于 TFIDF 权重法提取特征词，形成词袋，构造词汇-文本矩阵。由于文本数据量庞大，矩阵维度高，并且具有高稀疏度、同义词影响因而可能产生语义歧义的缺点，因此利用基于潜在语义（LSA）分析的奇异值分解算法（SVD）对词汇-文本矩阵进行空间语义降维。 经过上述处理后，对比各类模型，初步选择了对中文处理较精准的三个模型进行进一步比较，使用 F1-Score 对分类方法进行评价后选择基于 SVM 模型训练的关于留言内容的一级标签分类模型。

针对问题二：本文对比 3 个命名实体识别方法，选择了 CRF 对附件 3 中的留言标题提取对应的实体内容，人群、地点，进行文本摘要提取出关键信息，用于提高聚类的准确性，并选择 K-means 算法进行聚类。由于聚类的类别数目未知，本文结合肘部法和轮廓系数法选择最优 K 值为 20，得出聚类后的各类标签用于分类。然后我们定义了合理的热度评价一级指标：内容特征热度、传受众特征热度影响力，及 3 个二级指标对一级指标进行量化，指标内涵便于理解，用层次分析法得出 3 个二级指标的比重，得到热度指数=字数充实度*0.1429+点赞数*0.4286-反对数*0.4286。最后对热度指数进行排序得到热点问题的具体数据，提取标题中的实体得到热点问题的人群、地点、时间，对同个标签的详情再次做文本摘要得到问题描述。

针对问题三：从答复的相关性、完整性、可解释性等角度定义二级指标：相似度、匹配度、回复规范性、结论性、字数充实度，指标内涵定义了对它们的具体量化过程。再次使用层次分析法得出 5 个二级指标的比重，得出评价方案：Y=0.3153* 相似度+0.1561*匹配度+0.1377*回复规范性+0.1211*结论性+0.2697*字数充实度。

关键词：SVM 支持向量机；jieba 分词；K-means 聚类；奇异值分解（SVD）；层次分析法；CRF 实体识别；文本摘要；自定义量化指标

Abstract

Network ask ZhengPing set with its transfer speed, small space distance, low cost, the advantages of the freedom of expression, gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, establishing the wisdom of the e-government system based on natural language processing technology has is the new trend of development of social management innovation, to enhance the management level of government and governance efficiency has a great role in promoting.

Aiming at problem 1: firstly, the unstructured data in attachment 2 was de-duplicated and de-nullified. In this paper, jieba word segmentation was selected, four stop word lists were collected for accuracy comparison, and then the machine stop word database of sichuan university was selected for stop word filtering, and TextRank algorithm was used for keyword extraction. After data preprocessing, feature words were extracted based on TFIDF weight method, word bag was formed, and lexicot-text matrix was constructed. Due to the large amount of text data, high dimension of matrix, high sparsity and synonym influence, semantic ambiguity may be generated. Therefore, the singular value decomposition algorithm (SVD) based on latent semantics (LSA) analysis is used to reduce the spatial semantic dimension of lexicographical matrix. After the above treatment, three models with more accurate Chinese processing were preliminarily selected for further comparison by comparing various models, and f1-score was used to evaluate the classification method, and then the first-level label classification model based on SVM model training on message content was selected.

For problem 2: this paper compares three named entity identification methods, selects CRF to extract the corresponding entity content, population and location from the comment title in annex 3, extracts the key information by text summary, for improving the accuracy of clustering, and selects k-means algorithm for clustering. Since the number of categories of clustering is unknown, the optimal K value of 20 is selected by the elbow method and the contour coefficient method in this paper, and various labels after clustering are obtained for classification. Then we define the heat of the reasonable evaluation indicators: heat, content characteristic heat influence audience characteristics, and three secondary index to quantify the primary index, index connotation understanding, using analytic hierarchy process (ahp) it is concluded that the proportion of three secondary indexes, heat index = words enriched degrees * 0.1429 + thumb up number - antilog * * 0.4286 0.4286. Finally, the heat index is sorted to get the specific data of hot issues, the entity in the title is extracted to get the population, place and time of hot issues, and the details of the same label are summarized again to get the problem description.

Aiming at question 3: from the perspective of relevance, completeness and interpretability of replies, the secondary indicators are defined: similarity, matching, reversibility, conclusion and word enrichment. The specific quantification process of them is defined by the index connotation. The analytic hierarchy process was used again to obtain the proportion of 5 secondary indicators, and the evaluation scheme was obtained: Y=0.3153* similarity +0.1561* matching +0.1377*

reversibility +0.1211* conclusion +0.2697* word enrichment.

Key words: SVM support vector machine; Jieba; K - means clustering; SVD; Analytic hierarchy process; CRF entity identification; Text abstract; Custom metrics

目录

1 问题重述

网络问政平台凭其传递速度快、空间距离小、成本低廉、言论自由等优势，逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，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。

本次建模目标是利用时间跨度为接近 10 年(共接近 1 万条记录)的实际群众问政留言记录信息数据，其中包含结构化和非结构化文本数据，在对文本数据进行基本的预处理、中文分词、停用词过滤后，使用数据抽取的方法调整数据不平衡带来得影响，根据附件 2 给出的数据，基于 TFIDF 权重法文本的向量化表示，使用 F-Score 对分类方法进行评价后选择基于 SVM（支持向量机）模型训练的关于留言内容的一级标签分类模型。

利用群众问政留言信息数据（附件 3），对留言标题进行实体识别、然后利用 K-means 聚类、文本摘要实现把反映同一问题的留言数据进行分类，定义热度评价指标并对分好类的数据排序得出热点问题及其特定的特点。

通过对答复详情的指标性定义，得出每个答复所能得到的分数，此分数在处理庞大的文本数据时，可以得到简洁明了的指标，因此可以更加完善地保证回答问题的准确性和易懂性。

2 技术方案

本文的总体架构及思路如下：

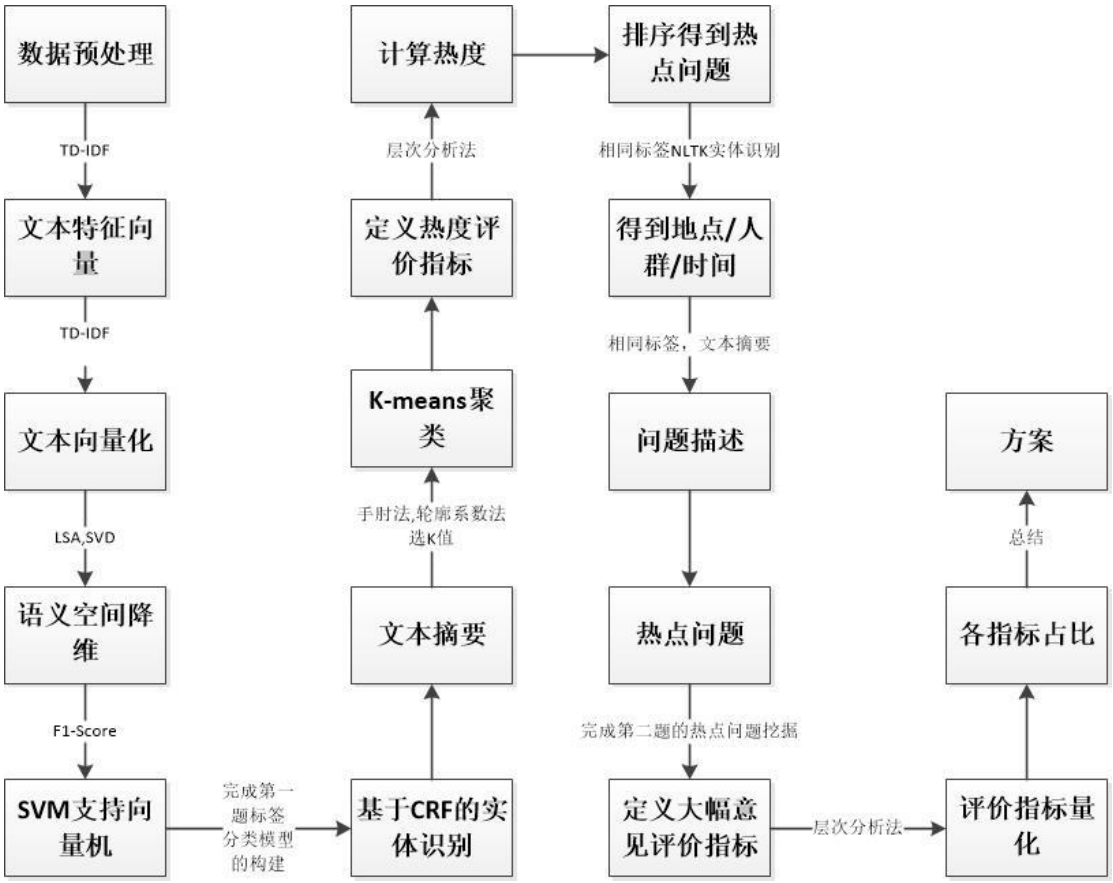


图 1：总体流程图

首先，我们对数据进行预处理，使用 TD-IDF 进行文本特征向量抽取以及文本向量化，为了解决语义问题使用 LSA、SVD 进行语义空间降维，完成以上步骤后进行模型选择与训练，通过 F1-Score 得到得分最高的模型 SVM 支持向量机，完成第一题标签分类模型的构建。

本文采用 NLTK 实体识别对留言标题进行特定地点、人群的提取，对留言详情进行文本摘要得出详情中较为重要的信息，结合手肘法和轮廓系数法得出最优 K 值使用 K-means 聚类得到分类结果。其次，自定义热度评价指标利用层次分析法得出各指标的比重，根据公式计算热度，对热度指数排序后得到前 5 个标签的热点问题，再对相同标签的留言标题实体识别、留言详情进行文本摘要得到问题描述，最终完成热点问题的挖掘。

本文从相关性、完整性、可解释性等角度定义答复意见评价指标，并将评价指标量化后利用层次分析法得出各指标比重，总结得出公式及具体的答复意见评价方案。

2.1 问题一的解决方案

2.1.1 处理流程图

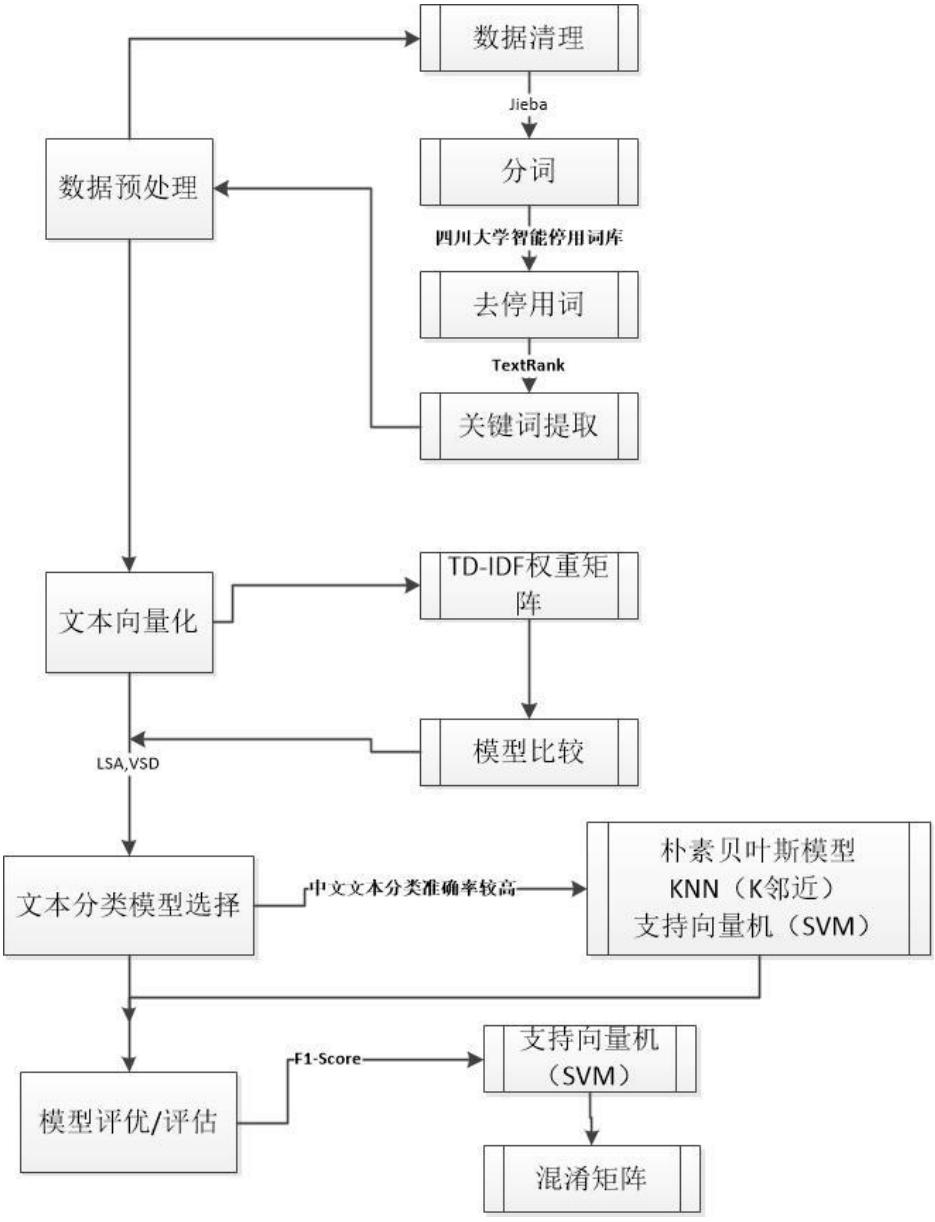


图 2：问题一流程图

首先进行数据清理：去重去空、数据抽取使得数据平衡、特殊字符处理等，利用 jieba 对留言详情进行分词，再用四川大学智能停用词库去除停用词，利用 TextRank 算法进行关键词提取。选取 TD-IDF 进行文本向量选择和文

本向量化表示，利用 LSA、SVD 进行语义空间降维使得中文歧义减少。然后初步通过筛选得出中文文本分类准确率较高的 3 个模型，计算和对比 F1-Score 后选择最优模型——支持向量机（SVM）模型，再利用混淆矩阵对模型进行评估。

2.1.2 数据预处理

结构化与非结构化的数据导致了有脏数据的产生，因为这些数据的存在我们无法直接进行数据挖掘，或挖掘结果差强人意。为了提高数据挖掘的质量产生了数据预处理技术。

数据预处理有多种方法：数据清理、数据集成、数据变换、数据归约等。这些数据处理技术在数据挖掘之前使用，大大提高了数据挖掘模式的质量，降低实际挖掘所需要的时间。[1]

（一）数据清理

首先本文对数据进行数据抽取解决数据不平衡的问题，然后去重、去空、对特殊字符进行处理、利用正则表达式清理数据。

（1）数据抽取

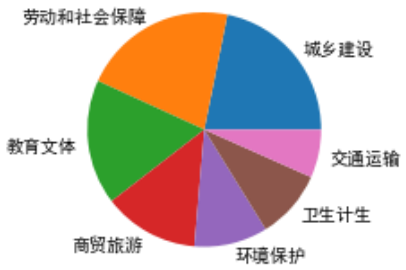


图 3：数据抽取前的数据分布图

```
城乡建设          2009
劳动和社会保障      1969
教育文体          1589
商贸旅游          1215
环境保护           938
卫生计生           877
交通运输           613
Name: 一级标签, dtype: int64
```

图 4：数据抽取前的具体数据

实际数据中样本比例不平衡的情况，也叫数据倾斜。本题的 7 个标签在附件 2 的占比也十分的不均衡，为了后续更好的建立模型，需要进行数据抽取使

得数据平衡。本文根据标签的数据最小值进行抽取，该值应<=613，为了不让某一个标签的全部值都被抽取，本文取该值为 500，为了使每次抽取的数据相对稳定，设置 random_state=100。

```
(3500, 6)
环境保护          500
商贸旅游          500
教育文体          500
卫生计生          500
交通运输          500
劳动和社会保障      500
城乡建设          500
Name: 一级标签, dtype: int64
```

图 5：抽取后的样本容量以及标签数量以及每个标签抽取的数据

(2) 去重、去空

附件 2 去重前有 9210 条数据，去重后有 9052 条数据，共清理了 158 条重复的留言详情信息，只有 1 条空消息，并对其删除处理。

(3) 特殊字符处理

我们观察文本可以发现留言详情中有许多特殊字符例如\xa0——不间断空白符、\u3000、\u2800、\t 等 Unicode 字符串。这些特殊字符不会出现在停用词表里，它的存在非常影响我们做数据的分析。本文使用 Python 标准库的 unicodedata 模块，它提供了 normalize 方法将 Unicode 字符转换为正常字符，它会让字符回归到我们期望看到的样子，同时不损害其它正常的空白字符，而且还能还原其它非空白字符，它们在处理某些字符的时候增加了额外的兼容特性。

(4) 正则表达式

\n——换行符、\t——横向跳到下一制表符位置等转义字符也是我们预处理的重点对象，因此我们采用 python 的 re 模块中的 apply 方法、lambda 正则表达式结合 re 模块中的 sub 方法进行删除。

（二）分词

分词是指将完整的一句话根据其语义分拣成一个词语项集[2]，该词语项集作为参与关联规则挖掘的基本单元[3]。中文分词是指以词作为基本单元，运用计

算机的数据库内容自动地对中文文本进行词语的切分，这样方便计算机识别出各语句中的重点内容。

(1) 分词方法的对比与选择

方法	优点	缺点
Jieba 分词	1、可以添加或管理自定义词典 2、通过计算分词后 TF/IDF 权重进行关键词抽取 3、分词后可以进行词性标注	1、自定义词典时，带空格的词不支持
HanLP 分词	1、自定义分词、词性方便 2、可分出多单词的英文名称(词典数据可带空格) 3、可动态增删词库 4、动态添加词典前五千万速度很快，5m 左右	1、动态添加词典前五千万快的很，越往后越慢 2、词典文件添加自定义词典速度略慢，添加 100w 需要 2m30s
Pyltp 分词	1、依存句法分析 2、语义角色标注	1、以非 UTF-8 编码的文本输入进行分析，结果可能为空

表 1：分词方法对比图

中文文本的特点是词与词之间没有明显的界限，比英文的难点在于语义分割后产生歧义。从文本中提取词语时需要分词，经过上述对比后，本文择优采 jieba 分词[4]，对附件 2 中每一条留言详情进行中文分词，jieba 分词用到的算法：

- 基于 Trie 树结构[5]实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）
- 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分组合
- 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

jieba 分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。

(2) 创建自定义词典

我们需要对特定的场景来进行分词，这时会有一些特定领域内的专用词汇，这些词汇往往是词库里没有的，解决这个问题的方法是创建自定义词典，自定义词典的有两个重要方法：载入词典 jieba.load_userdict(file_name)，往词库里添加单词：jieba.add_word(' word ')。我们混合使用两种方法对分词进行优化。

（三）去停用词

常用的停用词表有：哈工大停用词表、百度停用词表、中文分词表、四川大学机器只能实验室停用词库，我们将附件 2 的留言详情列分别用该四种停用词表分词后，放入 3 个对中文分类比较准确的分类器：朴素贝叶斯模型、KNN 模型、SVC 支持向量机模型中进行评价测试。

停用词表	**朴素贝叶斯模型**	**KNN 模型**	**SVC 支持向量机模型**
哈工大停用词表	0.8595238095238096	0.4531746031746032	0.8317460317460318
百度停用词表	0.8595238095238096	0.4785714285714286	0.8357142857142857
中文分词表[0]	0.8603174603174604	0.4396825396825397	0.8341269841269842
四川大学机器智能实验室停用词库	0.8611111111111112	0.4492063492063492	0.8365079365079365
最优停用词表	**四川大学机器智能实验室停用词库**	**百度停用词表**	**四川大学机器智能实验室停用词库**

表 2：停用词表使用后各模型的准确率对比

（四）关键词提取

对中文文本关键词提取的方法是采用不同方法 对文本分割后的分词进行计算权重，进行加权的方 法有 TF-IDF 算法和 TextRank 算法。

(1) TF-IDF 算法

TF-IDF（Term Frequency-inverse Document Frequency）是一种统计方法，其中 TF（Term Frequency）的意思是词频，IDF（Inverse Document Frequency） 的意思是逆文本频率指数，TF-IDF 算法所求实际 上就是这两者相乘所得的乘积。该算法的主要思想 为：若某词在一类指定的文本中出现的频率很高， 而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[6]。

上述引用总结就是,一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章，越能与其它文章区分开来。

TF 公式如下：

$$TF_w = \frac{N_w}{N} \qquad (2\text{-}1\text{-}1)$$

IDF 的计算公式：

$$IDF_w = \log\left(\frac{Y}{Y_w+1}\right) \quad (2\text{-}1\text{-}2)$$

其中Y是语料库的文档总数，Y_w是包含词条w文档数，分母加一是为了避免w未出现在任何文档中从而导致分母为 0 的情况。

TF-IDF 的就是将 TF 和 IDF 相乘

$$TF - IDF_w = TF_w \times IDF_w \quad (2\text{-}1\text{-}3)$$

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) TextRank 算法

TextRank 算法是一种基于图的算法，它是一种 排序算法，用于处理文本，可用于提取关键词[7]。 TextRank 可由一个有向有权图 G=(V,E) 表示，图中任两点 vi ，vj 之间的边的权重为 Wji，对于给顶 点 vi 的 TextRank 计算公式如下：

$$TR(v_i) = \frac{1-d}{n} + d\left(\sum_{v_j \in In(v_i)} \frac{w_{ji}}{\sum_{v_k \in Out(v_j)} w_{jk}} TR(v_j)\right), i = 1,2,\ldots n \quad (2\text{-}1\text{-}4)$$

其中，In(vi) 为指向该点的点集合，Out(vi) 为该 点所指向的集合，d 为阻尼系数，取值在 0 到 1 之间， 表示某点指向其他任意点的概率。TextRank 算法基于 PageRank 算法[8]，步骤为：

1. 分割文本，过滤。
2. 采用分割单位建立图模型。
3. 根据公式（2-1-4）在节点进行权重迭代，收敛时结束。
4. 根据权重的大小对节点采用的是倒序进行排序，排序后根据重要性假设得到了 T 个候选关键词。
5. 在原始文本中对候选词检测它们之间是 否相邻，相邻的时候将他们组合成多词关键词。

比较两种算法后，结果差异不大，本文采取 TextRank 算法提取关键词。

不 0.005235843173036554
市 0.004375776352905975
小区 0.0042002437011455515
没有 0.0039819714022969206
业主 0.0037307715811675333
人 0.0033799521032317455
后 0.0027857258323416064
问题 0.002541173843831651
上 0.0024358871309301624
公司 0.0024296726619382784
说 0.002339580886666192
现在 0.002319482498064392
进行 0.0022119793489803028
开发商 0.0020947585598755426
区 0.002056073087906972
中 0.0020013641187461985
西地省 0.0019970088999209192
领导 0.0019481295399696049
政府 0.001920185047857172
会 0.001882991855117252

图 6：TextRank 算法关键词提取及词频

（3）词云制作

词云是文本数据的视觉表示，由词汇组成类似云的彩色图形，用于展示大量文本数据。有了词云我们可以更直观的了解文本的内容和重点。

它可以：

- 快速感知最突出的文字
- 快速定位按字母顺序排列的文字中相对突出的部分

图 7：词云

可以从词云中看出大家对具体时间和地名，具体职位的意见较多，也提出了较多的解决方案。

2.1.3 文本特征向量选择

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。常用的方法有词频-逆向文档频率(TF-IDF)、互信息、信息增益、X2 统计等。

（1）信息增益[9](IG)

IG 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：

$$IG(t) = p(t)\sum_{i=1}^{m} p(C_i|t)lg(\frac{p(C_i|t)}{p(C_i)}) + p(\bar{t})\sum_{i=1}^{m} p(C_i|\bar{t})lg\frac{p(C_i|\bar{t})}{p(C_i)} \quad (2\text{-}1\text{-}7)$$

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。

（2）X2 统计[10]

X2 统计方法度量词条 t 和文档类别之间的相关程度，假设 t 和 c 之间符合具有一阶自由度的 X2 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 类但是包含 t 的文档频数，C 表示属于 c 类但是不包含 t 的文档频数，D 是既不属于 c 也不包含 t 的文档频数，则 t 对于 c 的 X2 值由下式计算：

$$x^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (2\text{-}1\text{-}6)$$

词条对于某类的 X2 的统计值越高，它与该类之间的相关性越大，携带的类别信息也较多。

（3）互信息[11](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），可由以下公式计算：

$$TR(v_i) = \frac{1-d}{n} + d\left(\sum_{v_j\in In(v_i)} \frac{w_{ji}}{\sum_{v_k\in Out(v_j)} w_{jk}} TR(v_j)\right), i = 1,2,\ldots n$$

(2-1-5)

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。

（4）词频-逆向文档频率[12](TF-IDF)

1.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：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2\text{-}1\text{-}8)$$

以上式子 ，$n_{i,j}$中 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。

2.逆向文档频率(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，因此这个方法对文本辨识度更高。

它的基本公式如下：

$$idf_i = log\frac{N}{|\{j:t_i \in d_j\}|} \quad (2\text{-}1\text{-}9)$$

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:

$$w_{ij} = tf_{ij} \times idf_i \quad (2\text{-}1\text{-}10)$$

算法	利弊
互信息 MI	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，有利于稀有名词多的专业性强的文本分析
x 2 统计	基于 x 2 分布，如果信息的特征分布不符合卡方分布则对低频词不可靠
信息增益 IG	计算量较大
词频-逆向文档频率 TF-IDF	有效且应用广泛，对文本辨识度更高

表 3：文本特征向量算法的利弊比较

因此本文采用有效且应用广泛的的 TF-IDF 算法抽取特征词条，将权重按照从大到小的顺序排列，抽取权重最大的前 1000 个特征词作为候选特征词。

2.1.4 文本的空间向量模型

计算机不能够直接处理文本信息，我们需要对文本进行处理，将文本表示成为计算机能够直接处理的形式，即文本数字化。文本表示[13](Test Expression)，即通过某种形式将文本字符串表示成计算机所能处理的数值向量,也称为文本特征表达，通过文本表示后的文章不仅具有准确性而且在对比运算中具有区分性。目前常用的文本表示模型有向量空间模型、布尔模型和概率模型等。本文采用向量空间模型[14](Vector Space Model, VSM)。

其主要思想：将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项(词条)对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重，这里的权重可以由 TF-IDF 等算法得到。向量空间模型就是将文本表示成为一个特征向量：

$$V(d) = (t_1, w_1(d), t_2, w_2(d), \ldots, t_n, w_n(d)) \quad (2\text{-}1\text{-}9)$$

其中，$t_i(i = 1,2,\ldots,n)$为文档 d 中的特征项，$w_i(i = 1,2,\ldots,n)$为特征项的权值，由 TF-IDF 算法可以计算得出。

（一） 文本的向量化表示

方法	优点	缺点
one-hot 词集模型	1. 统计各词在文本中是否出现 2. 直观	没有考虑词的频率
word2vec	1. Word2vec 会考虑上下文 2. 维度更少，所以速度更快 3. 通用性很强，可以用在各种 NLP 任务中 4. 分布式表示方法：保证了词的相似性 ，保证了词空间分布的相似性	1. 由于词和向量是一对一的关系，所以多义词的问题无法解决。 2. Word2vec 是一种静态的方式，虽然通用性强，但是无法针对特定任务做动态优化
LDA	以标签，类别衡量差异性的有监督降维方式，相对于 PCA 的模糊性，其目的更明确，更能反映样本间的差异。	局限性大，受样本种类限制，投影空间的维数最多为样本数量 N-1 维。
TF-IDF	1. 实现简单，相对容易理解 2. 离散化表示方法	1. 严重依赖语料库，需要选取质量较高且和所处理文本相符的语料库进行训练 2. 不能反应词的位置信息 3. 无法衡量词向量之间的关系

表 4：文本向量化模型利弊的比较

对比四种方法后，发现 word2vec 和 TF-IDF 得出的结果效果较好且两者结果差异度不大，因此选择实现简单、容易理解的 TD-IDF 生成文本向量化，其思想是，先根据所有训练文本，不考虑其出现顺序，只将训练文本中每个出现过

的词汇单独视为一列特征，构成一个词汇表(vocabulary list)，该方法又称为词袋法(Bag of Words)。

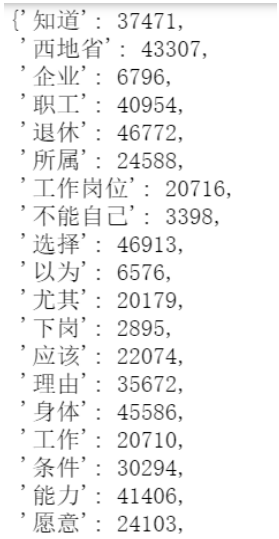


图 11：生成的词汇表

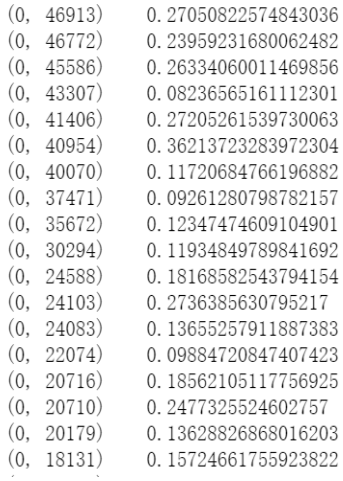


图 12：TD-IDF 文本向量化表示

（二） 语义空间降维

文本向量化表示后就可以直接比较两向量的夹角的余弦值进行相似度的计算。但是我们构造的词汇-文本矩阵是一个的巨大矩阵计算量十分大，而且中文文本分析存在语义上的歧义，留言详情中的信息文本中存在同义词和近义词等词语，即使通过特征抽取转化得到的文本向量可能达不到自然语言属性本质的要求。

因此，我们用潜在语义分析 (Latent Semantic Analysis，LSA)理论将留言详情中文本向量空间中非完全正交的多维特征投影到维数较少的潜在语义空间上。而LSA 对特征空间进行处理时用的关键技术就是奇异值分解 (Singular Value Decomposition，SVD)，在统计学上，它是针对矩阵中的特征向量进行分解和压缩的技术。

SVD 的计算公式如下：

$$Data_{m\times n} = U_{m\times m} \sum_{m\times n} V^{T}_{n\times n} \qquad (2\text{-}1\text{-}10)$$

下面的制图简单形象的展示了 SVD 的计算原理：

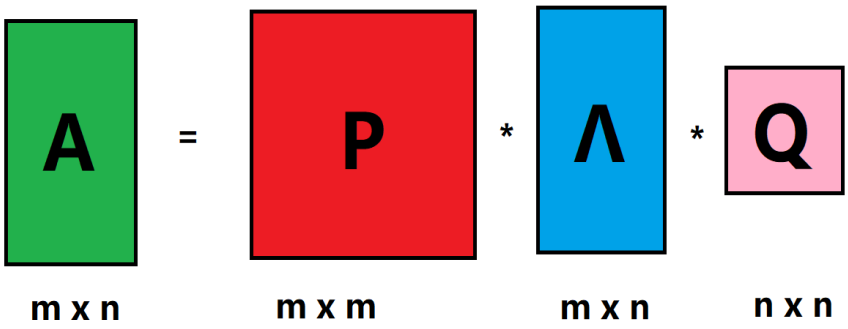


图 8：SVD 奇异值分解原理

2.1.5 文本分类模型

	方法	优点	缺点
传统方法	逻辑回归（LR）	1. 结果通俗易懂，自变量的系数直接与权重挂钩，可以解释为自变量对目标变量的预测价值大小 2. 速度快，效率高，并且容易线上算法实现	1.（特征空间）特征空间很大时，性能不好 2.（精度）容易欠拟合，精度不高 3. 目标变量中每个类别对应的样本数量要充足，才支持建模。
	隐马尔科夫方法（HMM）	解决了标注问题	做了齐次马尔科夫假设及观测股利性假设，可能出现标记偏置
	支持向量机方法（SVM）	1. 使用核函数可以向高维空间进行映射 2. 使用核函数可以解决非线性的分类 3. 分类思想很简单，就是将样本与决策面的间隔最大化 4. 分类效果较好 5. 模型占用小	1. SVM 算法对大规模训练样本难以实施 2. 用 SVM 解决多分类问题存在困难 3. 对缺失数据敏感，对参数和核函数的选择敏感，尤其是径向基函数
	KNN 方法	1. 非线性分类 2. 训练时间快 3. 可以进行增量训练 4. 异常数据不敏感	1. 不平衡数据效果差 2. 迷你占用内存大 3. 计算量大（相似度距离） 4. 容易过拟合 5. K 值选择没有理论指导
	Adaboost 算法	1. 很好的利用了弱分类器进行级联。 2. 可以将不同的分类算法作为弱分类器。 3．AdaBoost 具有很高的精度。 4．相对于 bagging 算法和 Random Forest 算法，AdaBoost 充分考虑的每个分类器的权重	1. AdaBoost 迭代次数也就是弱分类器数目不太好设定，可以使用交叉验证来进行确定。 2．数据不平衡导致分类精度下降。 3．训练比较耗时，每次重新选择当前分类器最好切分点。
	贝叶斯方法（NB）	1. 实施简单，非常高效（计算量小、存储占用低），可以在大数据场景中使用 2. 适用多类问题分类 3. 可以进行增量训练 4. 参数代表每个特征对输出的影响，可解释性强，便于分析误判	1. 特征之间有关联时效果受影响，由于使用了样本属性独立性的假设，所以如果样本属性有关联时其效果不好

	决策树方法	1．决策树易于理解和解释，可以可视化分析，容易提取出规则。 2．可以同时处理标称型和数值型数据。 3．测试数据集时，运行速度比较快。 4．决策树可以很好的扩展到大型数据库中，同时它的大小独立于数据库大小	1．对缺失数据处理比较困难。 2．容易出现过拟合问题。 3．忽略数据集中属性的相互关联。 4．ID3 算法计算信息增益时结果偏向数值比较多的特征。
深度学习方法	人工神经网络算法	1. 分类准确度高，学习能力极强。 2. 对噪声数据鲁棒性和容错性较强。 3. 有联想能力，能逼近任意非线性关系。	1. 神经网络参数较多，权值和阈值。 2. 黑盒过程，不能观察中间结果。 3. 学习过程比较长，有可能陷入局部极小值。
	K 均值聚类算法	1. 算法原理简单。需要调节的超参数就是一个 k。 2. 由具有出色的速度和良好的可扩展性。	1. 当数据量非常大时，算法的时间开销是非常大的。 2. 对离群点很敏感。

表 5：分类模型的利弊比较

由表对比可知，深度学习的方法虽然速度和可扩展性强，但是对于文本向量这一稀疏矩阵，不是特别精准。对比传统学习方法后，我们初步筛选出对处理中文较友好的模型——KNN，SVM，朴素贝叶斯模型。

2.2 问题二的解决方案

2.2.1 处理流程图

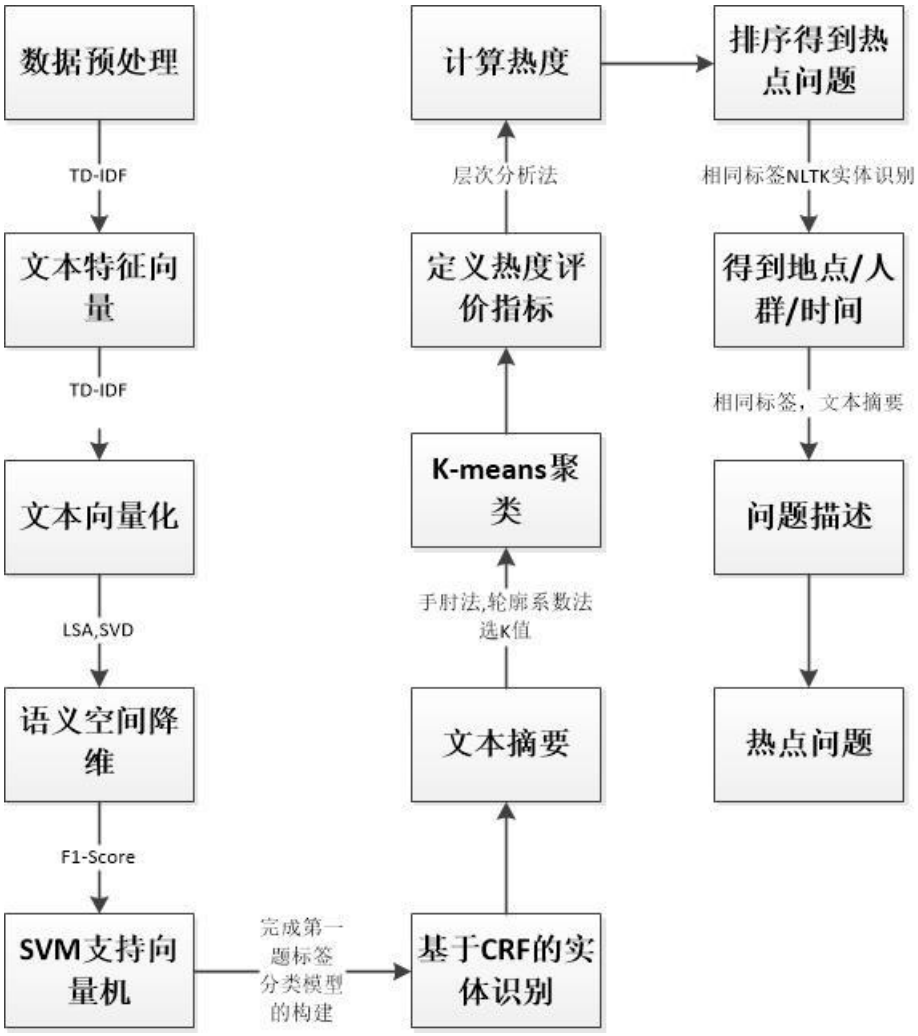


图 14：问题二流程图

该题先通过基于 CRF 的实体识别，得到地名，通过地名找到相关地点的的数据，把每个地点的数据中的留言详情进行文本摘要，之后将摘要结果进行 K-means 聚类，得到该地点的问题分类，其他地点重复该操作。问题分类后定义热度评价指标，用层次分析法找出各指标权重，之后计算热度值并进行排序，得到热度前 5 的问题。

2.2.2 实体识别

使用基于 CRF 的命名实体识别，把命名实体识别过程看作一个序列标注问题。基本思路为将给定的文本首先进行分词处理，然后对人名、简单地名和简单的组织机构名进行识别，最后识别复合地名和复合组织机构名，利用已标注的大规模语料对 CRF 模型的参数进行训练。

在训练阶段，先将分词语料的标记符号转化成用于命名实体序列标注的标记，如用 NA 表示人名的起始用字，ORG 表示名字的内政府组织。

确定特征模板。特征模板以当前位置的前后 n 个位置范围内的字串及其标记作为观察窗口：（…w－n/tag－n，…，w－1/tag－1w0/tag0，w1/tag1，…，wn/tagn，…）。防止窗口开得较大时，算法的执行效率会太低，而且模板的通用性较差，且窗口太小时，所涵盖的信息量又太少，不足以确定当前位置上字串的标记，将 n 值取为 2～3，即以当前位置上前后 2～3 个位置上的字串及其标记作为构成特征模型的符号。

由于不同的命名实体通常出现在不同的上下文中，因此通常使用不同的特征模板来识别不同的命名实体。同时,考虑到字符串出现在左右一个人的名字可以帮助确定边界的一个人的名字在某种程度上,比如一些名称,一些动词和标点符号,等等,因此,一些“引用边界词”(左参考边界词或右引用边界词)总结也可以作为特征。

$$
\begin{aligned}
&(WC^{+}+TC^{-})\\
&=argmax_{(WC,TC)}P(WC,TC|W,T)\\
&=argmax_{(WC,TC)}\frac{P(WC,TC,W,T)}{P(W,T)}\\
&\approx argmax_{(WC,TC)}P(WC,W)\times[P(TC,T)]^{\beta}\\
&\approx argmax_{(WC,TC)}P(WC)\times P(W|WC)\times[P(TC)\times P(T|TC)]^{\beta}
\end{aligned}
\qquad (2\text{-}2\text{-}1)
$$

特征函数确定以后，训练 CRF 模型参数 λ 。

方法	优点	缺点
Hanlp	1.自定义分词、词性方便 2.可简单分出歧义单词的名 3.词典数据可带空格 4.可动态增删词库,速度很快,	1. 动态添加词典越多往后越慢 2.词典文件添加自定义词典速度略慢 3.实体识别冗余度高
NLTK	1. 更加倾向于分词和词性标准 2. 可以使用 NLTK 下的 treebank 包将文本绘制为树形，使结果更加清晰易读。	输出文本的冗余性，不利于读者很好的识别命名实体，需要我们对文本做进一步处理。

CRF	它可以把七类实体很清晰的标注出来，而没有多余的词性。	开发环境受限制。NER 是基于 java 开发的，所以在用 python 实现时可能由于 jar 包或是路径问题出现很多问题

表 6：实体识别模型的利弊比较

常用的三个实体识别方法有 Hanlp、NLTK、Staford NER，经过对比后我们发现 NLTK 冗余度居中，识别性较强，开发环境不受限制，因此选择 NLTK 进行实体识别。

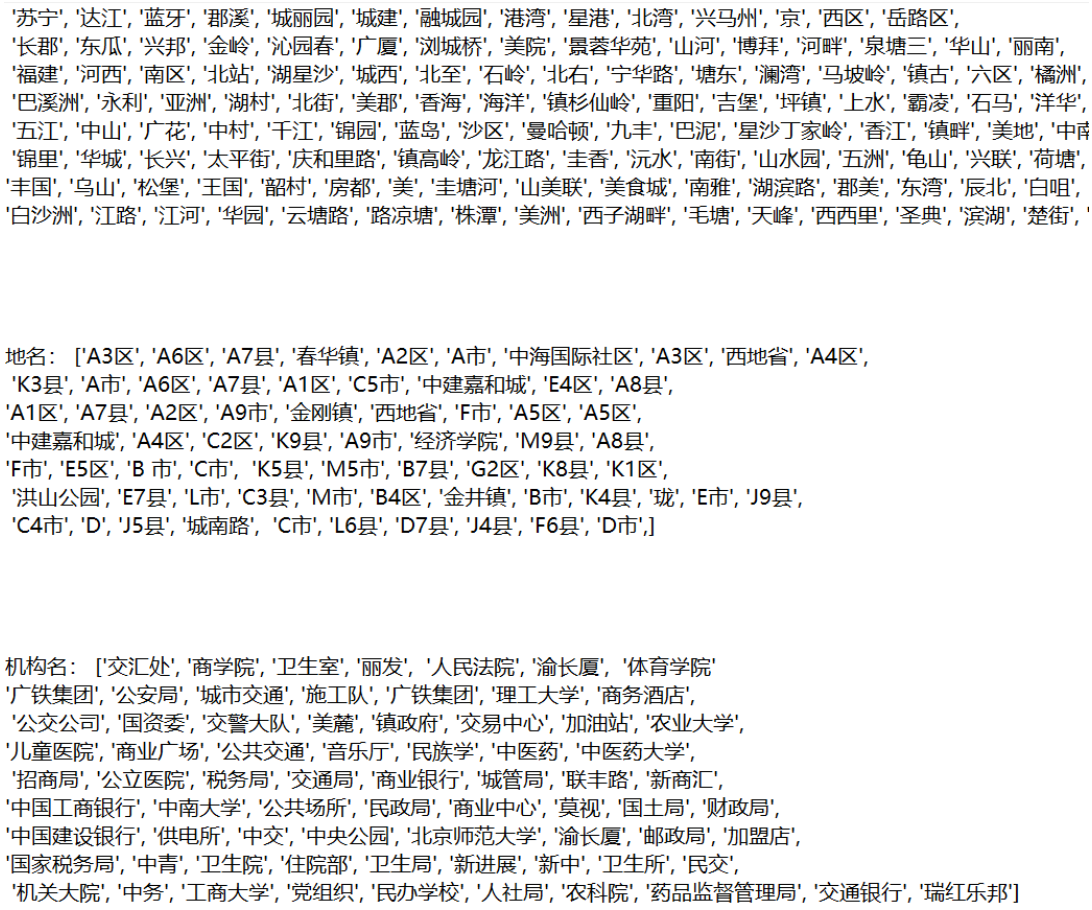


图 9：实体识别得到的地名和机构名

2.2.3 文本摘要

文本摘要就是将文本或文本集合转换为包含关键信息的简短摘要，利用词频统计进行自动摘要，其核心想法是：只考虑句子包含的关键词，句子包含的关键词越多，就说明这个句子越重要。[15]

实现的算法如下（伪码表示）：

```
Summarizer(originalText, maxSummarySize):

        // 计算原始文本的词频，生成一个数组，比如[(10,'the'), (3,'language'), (8,'code')...]
        wordFrequences = getWordCounts(originalText)

        // 过滤掉停用词，数组变成[(3, 'language'), (8, 'code')...]
        contentWordFrequences = filtStopWords(wordFrequences)

        // 按照词频进行排序，数组变成['code', 'language'...]
        contentWordsSortbyFreq = sortByFreqThenDropFreq(contentWordFrequences)

        // 将文章分成句子
        sentences = getSentences(originalText)

        // 选择关键词首先出现的句子
        setSummarySentences = {}
        foreach word in contentWordsSortbyFreq:
            firstMatchingSentence = search(sentences, word)
            setSummarySentences.add(firstMatchingSentence)
            if setSummarySentences.size() = maxSummarySize:
                break

        // 将选中的句子按照出现顺序，组成摘要
        summary = ""
        foreach sentence in sentences:
            if sentence in setSummarySentences:
                summary = summary + " " + sentence

        return summary
```

图 10：文本摘要伪代码

该算法首先计算文本词频生成数组，然后过滤停用词，按词频排序，之后把文章切分成句子，选择关键词首先出现的句子，将句子出现顺序组成摘要，最后返回摘要，算法结束。

下面是摘要后和摘要前的文本，摘要后只是保留了相关问题的描述，冗余部分被去除。

摘要后结果：

A 市 A6 区道路命名规划已经初步成果公示文件，给道路安装好路名牌，同时 A6 区农村的门牌号 10 年都未曾更换过，没有充分发挥路名和地名的作用。A6 区行政区划已经调整完毕，

摘要前：

A 市 A6 区道路命名规划已经初步成果公示文件，什么时候能转化成为正式的成果，希望能加快完成的路名规范，给道路安装好路名牌，对变更的路名牌及时更换。同时 A6 区农村的门牌号 10 年都未曾更换过，什么时候会统一更换，现在某些时候，我们找一个地方，都只能说是某某路口之类的，没有充分发挥路名和地名的作用。A6 区行政区划已经调整完毕，那么门牌的更新也应该同步开展。

2.2.4 文本聚类

（一）文本相似度计算

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体

将的差异大小，则需要对文本信息进行相似度计算，将根据相似特性的信息进行归类。目前相似度计算方法分为距离度量和相似度度量。本文选择基于距离度量的欧几里得距离计算留言详情进行文本摘要后的文本见差异。

计算文本相似度常见的方法有欧式距离、余弦相似度、Jaccard 距离、编辑距离。本文采用余弦相似度用向量空间中两个向量的夹角的余弦值来衡量两个文本间的相似度，相比距离度量，余弦相似度更加注重两个向量在方向上的差异，由于附件 2 数据接近 10000，数据量庞大，所以差异度大更有利于我们进行归类，因此选择了更注重差异性的余弦相似度计算。

计算公式如下：

$$cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=1}^{n} Y_i^2}} \qquad (2\text{-}2\text{-}2)$$

得出部分数据：

```
array([[ 0.00000000e+00,  7.69941264e-01,  8.01167620e-01, ...,
         7.65605710e-01,  7.96071619e-01,  7.99216680e-01],
       [ 7.69941264e-01,  0.00000000e+00,  6.96049170e-01, ...,
         6.37486804e-01,  8.18768177e-01,  6.68506634e-01],
       [ 8.01167620e-01,  6.96049170e-01, -2.22044605e-16, ...,
         7.98672021e-01,  7.68032582e-01,  7.20357118e-01],
       ...,
       [ 7.65605710e-01,  6.37486804e-01,  7.98672021e-01, ...,
         0.00000000e+00,  5.11385498e-01,  5.87134154e-01],
       [ 7.96071619e-01,  8.18768177e-01,  7.68032582e-01, ...,
         5.11385498e-01, -2.22044605e-16,  6.11512952e-01],
       [ 7.99216680e-01,  6.68506634e-01,  7.20357118e-01, ...,
         5.87134154e-01,  6.11512952e-01, -2.22044605e-16]])
```

图 11：文本相似矩阵

由图可观察得，因为矩阵中数据既有相差小的即文本相似的，也有相差大的负数与正数即辨识度高，该文本相似矩阵性能良好。

（二）聚类

(1) K-means 聚类原理

在聚类分析中，我们希望能有一种算法能够自动的将相同元素分为紧密关系的子集或簇，K 均值算法（K-means）为最广泛的一种算法。k-means 是硬分类，一个点只能分到一个类。K 即原始数据最终被聚为 K 类或分为 K 类，Means 即均值点。K-Means 的核心就是将一堆数据聚集为 K 个簇，每个簇中都有一个中心点称为均值点，簇中所有点到该簇的均值点的距离都较到其他簇的均值点更近。

(2) K 值选择

由于我们对最终的归类个数不明确，因此必须在训练模型前进行最优的 K 值选择。K 值选择的方法有手肘法和轮廓系数法，当情况简单时我们直接采用手肘法，当结果不明确或者模棱两可时我们再结合轮廓系数法对 K 值进行最优选择。

1.手肘法

手肘法的核心指标是 SSE(sum of the squared errors，误差平方和)，

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad (2\text{-}2\text{-}3)$$

Ci 是第 i 个簇，p 是 Ci 中的样本点，mi 是 Ci 的质心（Ci 中所有样本的均值），SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

其核心思想是：随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和 SSE 自然会逐渐变小。并且，当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，而当 k 到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，也就是说 SSE 和 k 的关系图是一个手肘的形状，而这个肘部，斜率变化最大的点对应的 k 值就是数据的真实聚类数。

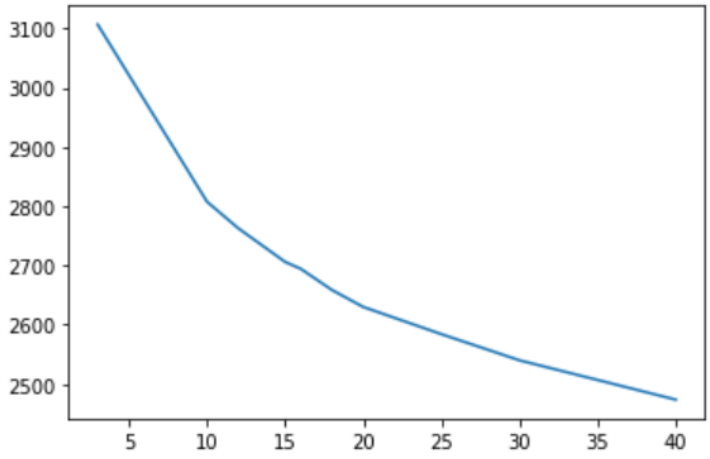


图 12：手肘法示例图（肘部明显）

当肘部即拐点十分明显如上图时，我们直接选择最优 K 值，上图的最优 K=10。

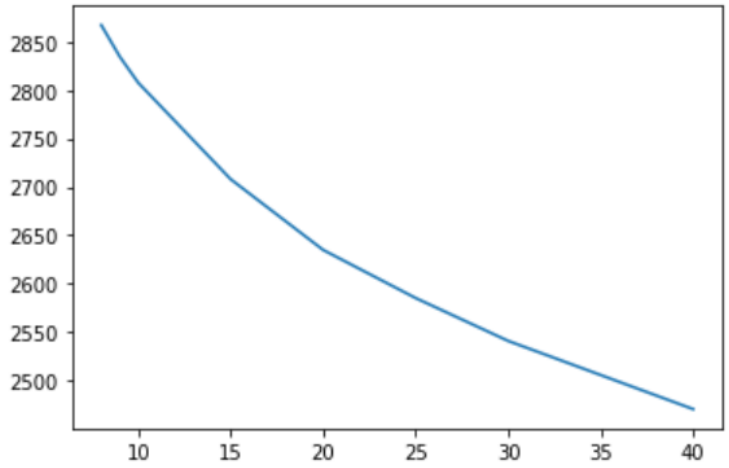


图 13：手肘法示例图（肘部不明显）

当肘部不是十分明显，曲线比较光滑或有歧义时我们对斜率变化较大的值选取出来（上图的 10、15、20）再对这个值附近的区间进行轮廓系数法比较，最终选择得出最优 K 值。

2.轮廓系数法

该方法的核心指标是轮廓系数（Silhouette Coefficient），定义如下：

$$S = \frac{b-a}{max(a,b)} \qquad (2\text{-}2\text{-}4)$$

其中，a 是某个数据与同簇的其他样本的平均距离，称为凝聚度，b 是某个数据（Xi）与最近簇中所有样本的平均距离，称为分离度。最近簇的定义是

$$C_j = agr\ min_{C_k} \frac{1}{n} \sum_{p \in C_k} |p - X_i|^2 \qquad (2\text{-}2\text{-}5)$$

其中 p 是某个簇 Ck 中的样本。用 Xi 到某个簇所有样本平均距离作为衡量该点到该簇的距离后，选择离 Xi 最近的一个簇作为最近簇。

求出所有样本的轮廓系数后再求平均值就得到了平均轮廓系数。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的 k 便是最佳聚类数。

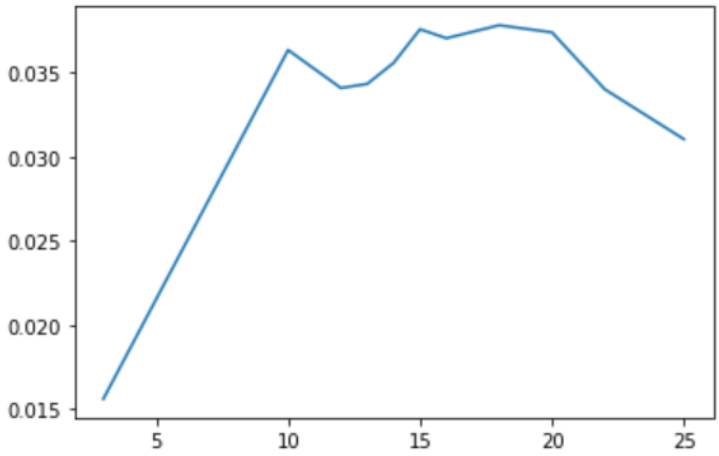


图 14：轮廓系数法示例图

由上图中的 K=18 时轮廓系数最高，图像中处于最高点，因此选为最优值。

（3）聚类流程

步骤：

1. 首先确定 K 值（即你想把数据聚为几类，K 值是 K-Means 算法中唯一的参数）；
2. 从原始数据集中随机选择 K 个点作为初始均值点（步骤 1 和 2 为准备工作）；
3. 依次从原始数据集中取出数据，每取出一个数据就和 K 个均值点分别计算距离（默认计算点间的欧氏距离），和谁更近就归为这个均值点所在的簇；
4. 当步骤 3 结束后，分别计算各簇当前的均值点（即求该簇中所有点的平均值）；
5. 比较当前的均值点和上一步得到的均值点是否相同，如果相同，则 K-Means 算法结束，否则，将当前的均值点替换掉之前的均值点，然后重复步骤 3。

输入: 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
聚类簇数 k.

过程:

```
从 D 中随机选择 k 个样本作为初始均值向量 {μ_1, μ_2, ..., μ_k}
repeat
    令 C_i = ∅ (1 ⩽ i ⩽ k)
    for j = 1, 2, ..., m do
        计算样本 x_j 与各均值向量 μ_i (1 ⩽ i ⩽ k) 的距离: d_ji = ||x_j − μ_i||_2;
        根据距离最近的均值向量确定 x_j 的簇标记: λ_j = arg min_{i∈{1,2,...,k}} d_ji;
        将样本 x_j 划入相应的簇: C_λj = C_λj ⋃ {x_j};
    end for
    for i = 1, 2, ..., k do
        计算新均值向量: μ'_i = (1/|C_i|) Σ_{x∈C_i} x;
        if μ'_i ≠ μ_i then
            将当前均值向量 μ_i 更新为 μ'_i
        else
            保持当前均值向量不变
        end if
    end for
until 当前均值向量均未更新
```

输出: 簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$

图 15：K-means 聚类算法 选自机器学习[16]

2.2.5 定义合理热度评价指标

热度综合评价指标体系	一级指标	二级指标	指标内涵
	内容特征热度影响力	字数充实度	文字数量和最多字数的比率
	传受众特征热度影响力	点赞率	点赞数与点赞数加反对数之和的比 即：点赞数/（点赞数+反对数）
		反对率	反对数与点赞数加反对数之和的比 即：反对数/（点赞数+反对数）

表 7：热度评价指标

2.2.6 层次分析法得出各指标比重

1、通过构建比较矩阵：

充实度 点赞数 反对数

$$\begin{matrix}1 & \frac{1}{3} & \frac{1}{3}\\ 3 & 1 & 1\\ 3 & 1 & 1\end{matrix}$$

2、层次的一致性检验

（1）一致性指标定义一致性指标为：

$$CI = \frac{\lambda - n}{n-1} \qquad (2\text{-}2\text{-}6)$$

(2) 检验系数 CR：

$$CR = \frac{CI}{RI} \qquad (2\text{-}2\text{-}7)$$

（3） 随机一致性指标 RI：

随机一致性指标 RI 和判断矩阵的阶数有关，一般情况下，矩阵阶数越大，则出现一致性随机偏离的可能性也越大，其对应关系下图：

矩阵阶数	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

表 8：随机一致性指标关系图

2.3 问题三的解决方案

根据题目要求，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评价方案，将从相关性、完整性、可解释性进行定量评价 。

（1）相关性。相关性可分为语义相似度和关键字匹配度，两个指标综合地影响着。均可以判断答复内容是否符合大众所提出的问题。

（2）完整性。答复都有回复规范性，有着回答的标准，越靠近标准则表示此回复越有说服力和可靠性。

（3）可解释性。从样本和特征两个维度考虑,包括理解问题的难度、找出各类的典型样本和识别重要特征,因此可制定我们合理答复意见评价指标。

2.3.1 处理流程图

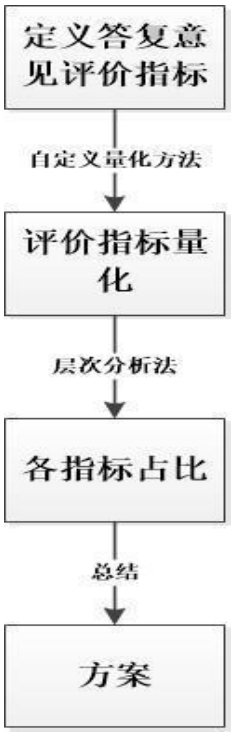


图 16：问题三流程图

通过自定义答复意见指标，将答复意见规范化，指定准确格式，并将该格式量化成可运算形式，得出回复在标准下的得分后，运用层次分析法，得出各项指标的占比，通过占比来给得出相应的评分方案。

2.3.2 定义答复意见评价指标

答复意见评价指标	一级指标	二级指标	指标内涵
	相关性	相似度	附件 4 中答复意见中文本信息与留言详情中文本信息的文本相似度
		匹配度	在留言详情中文本信息出现的关键字出现在答复意见中的频率。
	完整性	回复规范性	出现规范性语言的频率，规范性语言由答复意见中得出。
	可解释性	结论性	出现解释性语言的频率，解释性语言由答复意见中得出。
		字数充实度	文字数量和最多字数的比率

表 9：答复意见评价指标

我们对这三个方面进行量化，并采用层次分析法来进行计算和评价。对于答复的相关性我们采用文本相似度及关键字匹配度来对文本进行量化、可操作

化和可视化处理。完整性依据定义的回复标准格式，将回复文本与标准格式进行文本匹配。依照同样的方法对其进行解释性中结论性的量化与对比，且从字数充实度丰满解释性的叙述。

2.3.3 评价指标量化

（一）文本相似度

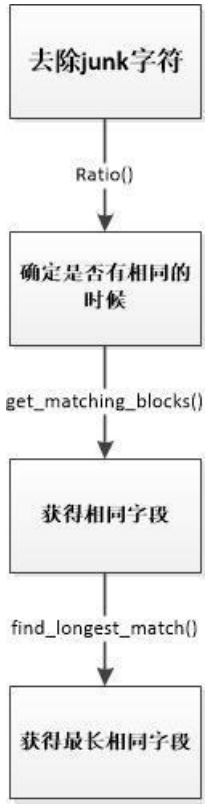


图 17：文本相似度流程图

文本相似度计算的具体流程：

- 去掉 junk 字符
- __chain_b()
  首先创建字典 self.b2j，以字符为 key，出现的位置放在一个 list 中当作元素，然后去掉 key 为 junk 的元素
- ratio():
  计算所有匹配片段的长度之和 T，然后 2*T/(len(a)+len(b))
  结果在[0,1]，相同的时候返回 1，没有相同片段返回 0
- get_matching_blocks():
  依次获取最长的相同段，然后分裂下去，最后返回所有相同的段。
- find_longest_match():
  上面方法中用来查找最长相同段的子方法。
  如果有多段最长的段长度一样，返回比较靠前的。

（二）关键字匹配度

使用 TextRank 提取关键字操作，将附录 4 中的留言详情进行关键字提取，将提出的关键字存入字符串数组中。此时将回复详情与关键字数组做匹配，查看是否进行了针对该关键字进行回复及处理。

其中 TextRank 算法的核心公式如下，其中ω_{ji}用于表示两个节点之间的边连接具有不同的重要程度：

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{\omega_{ji}}{\sum V_k \in Out(V_j) \omega_{jk}} WS(V_j) \qquad (2\text{-}3\text{-}1)$$

提取关键词和关键词组的具体步骤如下：

1）将给的文本进行分词分割，即$T = [S_1, S_2, \ldots, S_m]$；

2）对于每个单个句子$S_i \in T$，进行分词和词性标注，去除停用词，保留所需词性的词语，即$S_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,n}]$，$t_{i,j}$为第 i 句筛选保留后的词语；

3）建立词图

G=(V,E)，其中 V 为用保留的词语组成的节点集合，使用共现关系建立任意两个点所构成的边： 如果它们对应的词在长度为 K 的窗口中共现，即存在边。K 表示窗口大小

4）依据上文提到的公式，迭代求出每个节点的权重；

5）对将节点依据权重倒序排序，得到 top-t 关键词；

6）在文中标记 top-t 的位置，若在他们相邻的词组之间可以成为新的词组。则判定为关键词组

（三） 字数充实度

使用 excel 对各回复详情进行字数统计，并将每个回答的字数与最长回答字数的比例求出

字数充实度
0.02105862
0.02902675
0.03969835
0.04240182
0.07441662
0.04923164
0.14029596
0.03685259
0.02148549

图 18：部分字数充实度示例图

（四）规范性，结论性量化

规范性语言以及解释性语言是通过词频统计以及关键词提取得出，人工判断后进行规范化的。

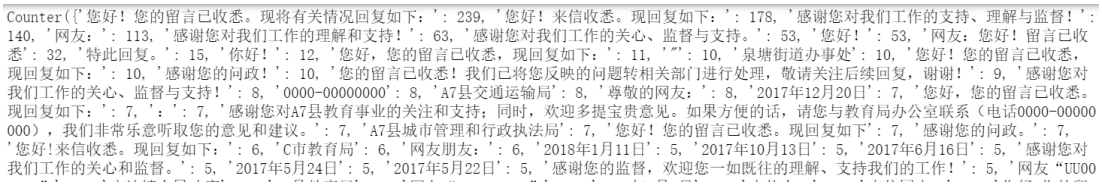


图 19：部分句频率统计图

项目	具体内涵
规范性语言	“尊敬的网友：你/您好！” "你/您的留言/来信已收悉 “经核查/调查/处理” "现将有关情况回复如下：” “感谢你/您对我们工作的支持、理解与监督！/谢谢/感谢您的问政” XXXX 年 XX 月 XX 日 “特此回复”
分词后	'尊敬','网友','你好','您好','!','留言','来信','收悉','核查','调查','处理','现','有关','情况','回复','如下',':','感谢','工作','支持','理解',','监督','谢谢','问政','特此','年','月','日'
解释性语言	“由于/因此” “如有需要请致电/联系电话：” “高度重视此事/项目” “根据《》条例/法律/[政府发文]” “市/县/镇/村” “（1）（2）（3）（4）（5）一 二 三 四 五 1 2 3 4 5” “您/你所反映的问题/已转交相关单位调查处置”
去除相同字后	'由于','因此','如有需要','请致电','联系','电话','高度重视','此事','项目','根据','《','》','条例','法律','政府发文','市','县','镇','村','（1）','（2）','（3）','（4）','（5）','一','二','三','四','五','1','2','3','4','5','反映','问题','转交','相关','单位','调查','处置'

表 10：答复意见评价指标的解释

在 sklearn 中使用 metrics.pairwise import cosine_similarity 步骤：
第一步：对数据使用 DataFrame 化，并进行数组化
第二步：对数据进行分词，并去除停用词，使用’’.join 连接列表
第三步：np.vectorizer 向量化函数，调用函数进行分词和停用词的去除
第四步：使用 TF-idf 词袋模型，对特征进行向量化数字映射
第五步：使用 from sklearn.metrics.pairwise import cosine_similarity，对两两样本之间做相关性矩阵，使用的是余弦相似度计算公式

2.3.4 基于层次分析法得出指标的占比

把问题分为两个层次，最上层为目标层——构建答复意见的评价指标。第二层为准则层，准则层分为一级和二级指标，一级指标影响目标层，分别为相关性、完整性、可解释性；二级指标影响一级指标，相关性的二级指标为相似度和匹配度，完整性的二级指标为回复规范性，可解释性的二级指标为结论性和字数充实度。构建二级指标的比较矩阵可求出五个指标的权重。
比较矩阵：

[1,2,3,3,1;

0.5,1,1,2,0.5;

1/3,1,1,2,1/3;

1/3,1/2,1/2,1,1;

1,2,3,1,1]

3 实验结果分析

3.1 问题一的实验结果

3.1.1 模型选择

（一）符号及公式说明

符号	说明
TP	实际正例；预测正例
FP	实际反例；预测正例
TN	实际反例；预测反例
FN	实际正例；预测反例

表 11：符号说明

查全率——实际为正例中，被预测为正例的比例。

$$recall = sensitivity = TRR = \frac{T_P}{(T_P+FN)} \quad (3\text{-}1\text{-}1)$$

查准率——预测为正例中，实际也为正例的比例。

$$precision = \frac{TP}{(TP+FP)} \quad (3\text{-}1\text{-}2)$$

F1-score———一个兼顾考虑了 Precision 和 Recall 的评估指标。

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \quad (3\text{-}1\text{-}3)$$

其中 P_i为第 i 类的查准率，R_i为第 i 类的查全率。

（二）对比选择

对比各类模型后，初步选择了对中文处理较精准的朴素贝叶斯模型、KNN（K 最邻近）模型、SVM（支持向量机）模型进行进一步比较。

将数据划分为测试集和训练集，将测试集设置为训练集的 30%,为了保证每次抽取的测试集变化不会太大设置 randomstate：

cv_train,cv_test,y_train,y_test=train_test_split(cv_data,data_new['一级标签'],test_size=0.3,random_state=123)

评价得分由公式算出：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \quad (3\text{-}1\text{-}4)$$

得出每个分类器的 score 进行比较：

分类器	评价得分
朴素贝叶斯模型	0.846031746031746
K 最邻近（Knn）	0.7968253968253968
支持向量机（Svm）	0.8690476190476191

表 12：各分类器的评价得分

由此可知 SVM（支持向量机）的 score 最高，因此选择基于 SVM（支持向量机）模型训练的关于留言内容的一级标签分类模型。

3.1.2 模型评估

macro avg——宏平均，就是对 7 个标签对应的类的 recall 值进行平均，权重是 1/7。weighted avg——就是用样本的比例充当权重来计算的。

我们运用 sklearn.metrics 中的 classification_report 模块对三个模型进行查全率、查准率、F1-score 的计算和比较。

	Precision	Recall	F1-score
	朴素贝叶斯模型		
macro avg	0.85	0.85	0.84
weighted avg	0.85	0.85	0.84
	K 最邻近（Knn）		
macro avg	0.80	0.80	0.80
weighted avg	0.80	0.80	0.80
	支持向量机（Svm）		
macro avg	0.87	0.87	0.87
weighted avg	0.87	0.87	0.87
最优模型	**支持向量机（Svm）**	**支持向量机（Svm）**	**支持向量机（Svm）**

表 13：各模型的查全率、查准率还是 F1-score 指标

由表格可知无论是查全率、查准率还是 F1-score 支持向量机 SVM 都是最优选择，证明了我们的选择是正确的。

接下来对支持向量机 SVM 的每个标签的查全率、查准率、F1-score 继续做评估：

	precision	recall	f1-score	support
交通运输	0.79	0.90	0.84	166
劳动和社会保障	0.88	0.87	0.88	168
卫生计生	0.92	0.91	0.91	182
商贸旅游	0.84	0.81	0.83	176
城乡建设	0.81	0.75	0.78	205
教育文体	0.91	0.93	0.92	167
环境保护	0.94	0.92	0.93	196

图 20：SVM 模型的 7 个标签的查全率、查准率、F1-score

混淆矩阵(confusion matrix) 是人工智能中的 一种可视化工具，特别用于监督学习，可以形象化的展示评价分类模型的好坏,通过构建混淆矩阵来 计算准确率。[17]

		预测值	
实际值		TP	FN
		FP	TN

表 14：性能度量标准

```
[[150   1   1   0  10   1   3]
 [  3 146  11   1   5   1   1]
 [  0   6 165   8   2   1   0]
 [ 16   1   1 143   8   6   1]
 [ 19   7   2  12 154   4   7]
 [  0   4   0   3   4 156   0]
 [  2   0   0   3   8   2 181]]
```

图 21：SVM 得到的混淆矩阵

3.2 问题二的实验结果

3.2.1 聚类结果分析

首先用肘部法得出该图

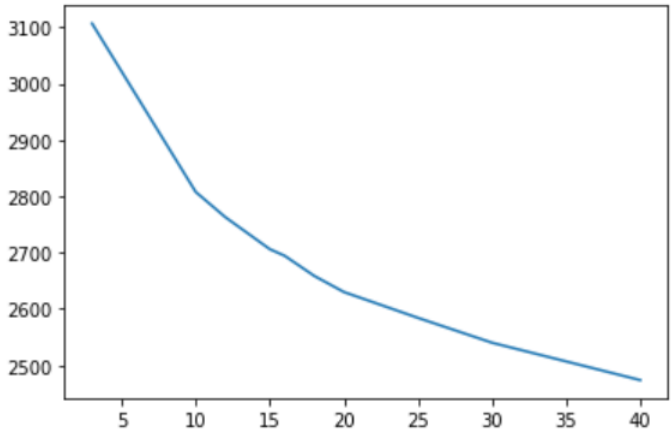


图 22：K-means 算法肘部法结果图

曲线较为平滑，但观察可得出肘部即斜率变化较大的值在 10、15、20 的邻近，因此再使用轮廓系数法进行进一步的分析 K 的最优值。

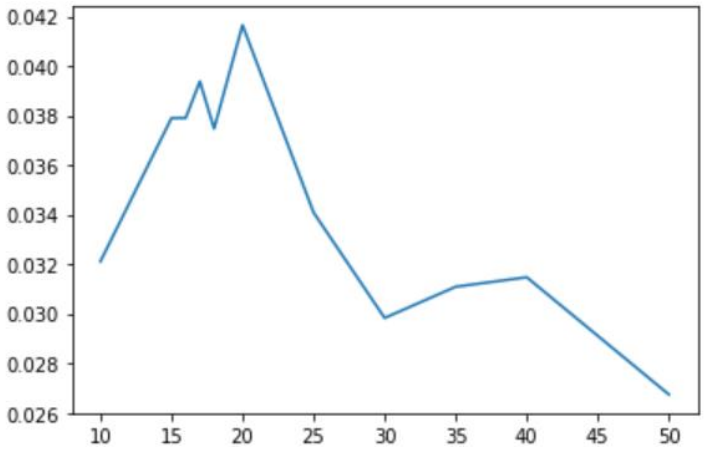


图 23：K-means 算法轮廓系数法结果图

由图可知最高点对应的 K 值为 20，该点对应的平均轮廓系数值为 0.04164990082870862。

综上选择 K=20 进行 K-means 模型训练,并将训练好的模型对测试集进行预测，并得到聚类结果。再把每条数据对应的标签值 cluster.labels_输出到附件 4 中，作为问题 ID 的值。

聚类后 20 个标签的关键词：

```
Cluster 0: 办理 户口 购房 公积金 社保 需要 政策 贷款 请问 工作
Cluster 1: 社区 a3 街道 拆迁 政府 领导 老百姓 违建 部门 餐饮
Cluster 2: 车辆 行人 车道 停车 道路 路段 马路 严重 设置 小区
Cluster 3: 幼儿园 小区 配套 教育局 孩子 招生 业主 收费 家长 a3
Cluster 4: 公交 公交车 出行 线路 方便 地铁 上班 增加 居民 时间
Cluster 5: 居民 生活 垃圾 影响 严重 小区 附近 环境 部门 正常
Cluster 6: 施工 噪音 工地 扰民 凌晨 居民 城管 晚上 部门 严重
Cluster 7: 噪音 严重 晚上 影响 扰民 居民 休息 小区 部门 油烟
Cluster 8: 地铁 规划 公园 建设 号线 城市 请问 建议 周边 出入口
Cluster 9: 小区 业主 居民 物业 严重 生活 社区 新城 影响 环境
Cluster 10: 学生 考试 中学 教师 老师 教育局 教育 学校 学习 领导
Cluster 11: 市委 市政府 国家 中心 希望 建设 尤其 政策 周边 城市
Cluster 12: 领导 现在 请问 西地省 政府 希望 a7 部门 已经 村民
Cluster 13: 公司 工资 员工 有限公司 领导 西地省 2018 自来水 工作 现在
Cluster 14: 开发商 业主 交房 问题 装修 房屋 政府 质量 合同 楼盘
Cluster 15: 路口 大道 车道 通行 建议 道路 行人 公交车 路段 交通
Cluster 16: 学校 学生 孩子 家长 小学 老师 教育 领导 希望 收费
Cluster 17: 医院 小区 a7 请问 居民 人民 西地省 领导 现在 收费
Cluster 18: 车位 职工 销售 购买 购房 12 开发商 资格 商品房 业主
Cluster 19: 业主 物业 小区 电梯 物业公司 业委会 问题 社区 收费 物业费
```

图 24：聚类后 20 个标签的关键词图

获得聚类后含标签的文档部分数据如图所示：

问题ID	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
19	188006	A000102948	阳光婚纱艺术摄影是否合法	2019/2/28 11:25:05	。因为地处居民楼内	0	0
9	188007	A00074795	络命名规划初步成果公示和城	2019/2/14 20:00:00	0年都未曾更换过，	0	1
12	188031	A00040066	华镇金鼎村水泥路、自来水	2019/7/19 18:19:54	，且天还没黑就开	0	1
16	188039	A00081379	络步行街大古道巷住户卫生间	2019/8/19 11:48:23	行清扫。没有解决	0	1
12	188059	A00028571	示社区三期与四期中间空地夜	2019/11/22 16:54:42	投诉业主，态度强	0	0
17	188073	A909164	单方面改变麓谷明珠小区6栋	2019/3/11 11:40:42	何政府调规、改建的	0	0
9	188074	A909092	区富绿新村房产的性质是什么	2019/1/31 20:17:32	让给业主了，然而因	0	0
12	188119	A00035029	对A市地铁违规用工问题的质	2019/5/27 16:04:44	加班还扣钱，扣身份	0	0
18	188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	该司机并未按地面车	0	0
1	188249	A00084085	桐梓坡路与麓松路交汇处地	2019/9/17 4:25:00	边邻居也是苦不堪言	0	0
18	188251	A00013092	东四路口晚高峰太堵，建议调	2019/10/19 11:02:40	下至少两到三个信号	0	0
17	188260	A00053484	小区乐果果零食炒货公共通道	2019/5/31 17:06:13	零食炒货公共通道摆	0	0
10	188396	A00047580	楚在西地省商学院宿舍旁安	2019/4/15 16:23:09	《中小学校校园环境	2	1
4	188399	A00097934	利保壹号公馆项目夜间噪声扰	2019/7/3 6:23:25	2点还在施工中，且	0	0
10	188409	A0003274	线星沙大道站地铁出入口设置	2019/6/19 10:14:39	区、星沙四区、开	0	4
14	188414	A00096844	辰小区非法住改商问题何时能	2019/8/1 7:20:31	处理、不做太多干预	0	0
16	188416	A00029753	3县乡村医生发卫生室执业许	2019/06/20 20:38:47	是证件下来啊。有些	0	0
13	188451	A00013004	春华镇石塘铺村有党员家开麻	2019/4/11 17:54:25	关门了。但是石塘铺	0	2
7	188455	A00035902	咨询异地办理出国签证的问题	2019/5/16 15:20:43	原来的户籍所在地办	0	0
19	188467	A00050188	市温斯顿英语培训学校拖延	2019/3/28 19:57:19	都是推辞的态度！去	0	1
11	188475	A00055810	原国际广场停车场违章乱建现	2019/12/3 15:04:58	车位然后把车位砌围	0	0
6	188535	A00061775	时代星城4幢有非法经营的家庭	2019/6/13 15:28:44	几间旅馆房间，进出	0	0
11	188546	A0006817	佳兆业水新都小区垃圾无人	2019/1/23 13:09:19	有余，周边小区业主	0	0
9	188553	A00092239	平老街上有无证理疗馆骗取老	2019/6/6 21:58:22	力，他们什么松花粉，	0	0

图 25：文档部分数据图

3.2.2 热度评价指标分析

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	10	48.65012871	2019/4/17至2019/10/16	A7县松雅西地省站	3号线松雅西地省站出入口设置不合理
2	16	22.80380602	2019/4/30至2019/10/16	A7县泉星公园	对泉星公园的建议和咨询
3	8	19.73973245	2019/1/17至2019/4/3	A7县东六线榔梨段	建议加大对A7县东六线榔梨段的拆迁力度
4	2	18.185741	2019/1/4至2019/12/30	A市国王陵考古遗址公园	A市国王陵考古遗址公园周边环境恶劣
5	4	14.62820844	2019/3/9至2019/7/23	西地省聚利人普惠投资有限公司	西地省聚利人普惠投资有限公司涉嫌诈骗

图 26：热点问题表

通过层次分析法得到三个热度指标——字数充实数、点赞数、反对数的权重。

几何平均法求权重的结果为：0.1429　0.4286　0.4286

特征值法求权重的结果为：　　0.1429　0.4286　0.4286

一致性指标 CI=-2.2204×10^{-16}

一致性比例 CR=-402701×10^{-16}

因为 CR<0.10，所以该判断矩阵 A 的一致性可以接受!

最终热度指数公式：

热度指数=字数充实度*0.1429+点赞数*0.4286-反对数*0.4286

最终得到热点问题如下：

3 号线松雅西地省站出入口设置不合理
对泉星公园的建议和咨询
建议加大对 A7 县东六线榔梨段的拆迁力度
A 市国王陵考古遗址公园周边环境恶劣
西地省聚利人普惠投资有限公司涉嫌诈骗

表 15：热点问题

3.3 问题三结果分析

3.3.1 各指标量化结果

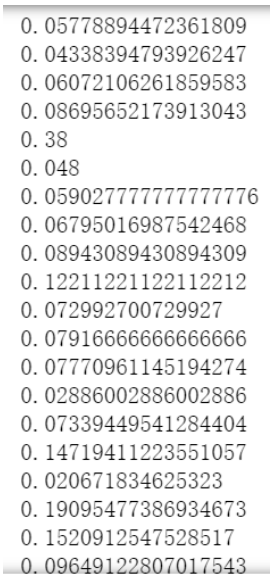
根据附录四所提供的数据得出以下答复详情的量化结果

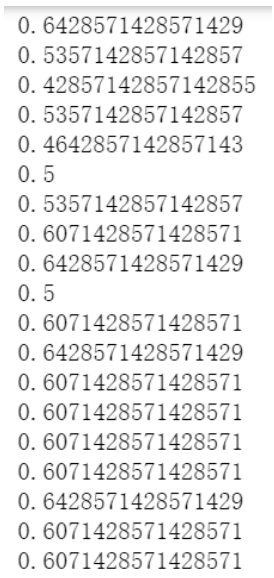
0.2926829268292683
0.3170731707317073
0.2682926829268293
0.34146341463414637
0.1951219512195122
0.2682926829268293
0.3170731707317073
0.36585365853658536
0.21951219512195122
0.12195121951219512
0.2926829268292683
0.36585365853658536
0.0975609756097561
0.12195121951219512
0.12195121951219512
0.3170731707317073
0.24390243902439024
0.12195121951219512
0.1951219512195122

图 27：留言详情与答复详情相似度

0.975
0.7
0.95
0.475
0.2
0.7
1.0
1.0
0.975
1.0
0.975
0.3
0.975
1.0
0.55
1.0
0.975
0.4
0.45
0.6

图 28：留言详关键字匹配度





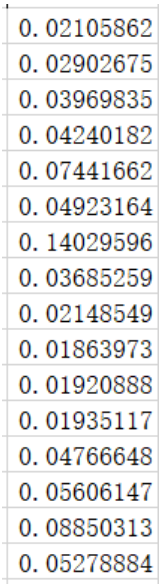


图 29：答复详情规范性　图 30：答复详情可解释性　图 31：字数充实度

经过图 27、28 可以得知每个留言详情和答复详情文本之间的相关性，判断回复的是否是留言中所需要解决的问题。图 29 中各项系数用于判断每个答复文本是否按照规定的答复规范进行回答，并对其规范化程度进行量化评价。图 30 中体现的是意见答复中具有解释性的文本在整个回复中的占比，图 31 字数充实度则为当前文本与充实文本的比例，二者共同诠释答复的可解释性。

3.3.2 层次分析法

将 3.3.1 中的指标进行判断矩阵实例化，将其化为矩阵 A，

A=[1,2,3,3,1;
0.5,1,1,2,0.5;
1/3,1,1,2,1/3;
1/3,1/2,1/2,1,1;
1,2,3,1,1]

算术平均法求权重的结果为：0.3141　0.1564　0.1372　0.1226　0.2697
几何平均法求权重的结果为：0.3281　0.1602　0.1363　0.1120　0.2634
特征值法求权重的结果为：　0.3153　0.1561　0.1377　0.1211　0.2697
一致性指标 CI=0.0770
一致性比例 CR=0.0687
因为 CR<0.10，所以该判断矩阵 A 的一致性可以接受。
最终得出方案：
答复意见的质量评价方法 Y

Y=0.3153* 相似度+0.1561*匹配度+0.1377*回复规范性+0.1211*结论性+0.2697*字数充实度

相似度	匹配度	回复规范性	结论性	字数充实度	评分
0.057789	0.975	0.642857143	0.292683	0.057592287	0.309916325
0.043384	0.7	0.535714286	0.317073	0.038690854	0.2455493
0.060721	0.95	0.428571429	0.268293	0.045287327	0.271158873
0.086957	0.475	0.535714286	0.341463	0.03932513	0.227289956
0.38	0.2	0.464285714	0.195122	0.020423697	0.244103682
0.048	0.7	0.5	0.268293	0.02943042	0.233682028
0.059028	1	0.535714286	0.317073	0.031079538	0.295259028
0.06795	1	0.607142857	0.365854	0.079157681	0.326781965
0.089431	0.975	0.642857143	0.219512	0.064061905	0.312776912
0.122112	1	0.5	0.121951	0.028415578	0.285883954
0.072993	0.975	0.607142857	0.292683	0.062032221	0.310989662
0.079167	0.3	0.642857143	0.365854	0.054167195	0.219226449
0.07771	0.975	0.607142857	0.097561	0.017632881	0.276873134
0.02886	1	0.607142857	0.121951	0.012812381	0.26702693
0.073394	0.55	0.607142857	0.121951	0.026639604	0.21455285
0.147194	1	0.607142857	0.317073	0.082582773	0.34678401
0.020672	0.975	0.642857143	0.243902	0.018394012	0.281734209

图 32：各评价二级指标和总评分的值

4 结论

总结本次比赛，我们基于 TF-IDF 权重法提取特征词，构造词汇-文本矩阵，依据 F-Score 选择基于 SVM（支持向量机）模型。运用 TextRank 提取关键字操作，进一步使用基于向量特征提取算法，基于潜在语义（LSA）分析的奇异值分解算法（SVD）对词汇-文本举证进行空间语义降维，通过 K-means 聚类算法对留言进行聚类得出相同问题进而得出热度评价得出人们最关注的问题留言并排序，定义指标并对其量化使用层次分析法得出合理的占比及结果。分析了留言与答复间的文本相似度，关键字匹配度，引入格式标准，基于层次分析法同样的得出合理的占比与结果，且根据目前社会答复质量整体上的差距，提出参考的答复意见。

但是我们最后得到的聚类，指标得分的准确度不是特别的高，与预期的结果有一定的出入，可能是因为定义规范标准的时候未考虑完全且由于 K 均值算法在计算欧氏距离的时候有一点的误差，这也涉及到了我们当前文档中所提到的挖掘模型的缺点与不足，后期一点会进一部对文本挖掘进行深入探讨。

5 参考文献

[1]李卫东主编 . 应用统计学 . 北京：清华大学出版社，2014
[2]李康康，龙华.基于词的关联特征的中文分词方法[J]. 通信技术，2018，51（10）：2343-2349.
[3]吴帅，潘海珍 . 基于隐马尔可夫模型的中文分词 [J]. 现代计算机（专业版），2018（33）：25-28.
[4]陶伟. 警务应用中基于双向最大匹配法的中文分词算法实现[J]. 电子技术与软件工程, 2016(4).
[5]刘志. 基于用户兴趣的协同过滤算法的广告推荐研究[D]. 昆明理工大学, 2014
[6] 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.
[7]曹洋 . 基于 TextRank 算法的单文档自动文摘研究 [D]. 南京：南京大学，2016.
[8]顾益军，夏天 . 融合 LDA 与 TextRank 的关键词抽取 研究 [J]. 现代图书情报技术，2014（Z1）：41-47.
[9]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.
[10]陈小莉. 基于信息增益的中文特征提取算法研究[D]. 重庆大学, 2008.
[11]徐明, 高翔, 许志刚,等. 基于改进卡方统计的微博特征提取方法[J]. 计算机工程与应用, 2014, 50(19):113-117.
[12]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.
[13]胡学钢, 董学春, 谢飞. 基于词向量空间模型的中文文本分类方法[J]. 合肥工业大学学报:自然科学版, 2007, 30(10):1261-1264.
[14]Dipanjan Das, Andre F.T. Martins,A Survey on Automatic Text Summarization（2007）
[15]郭飞, 张先君, 叶俊. 基于改进互信息的特征提取的文本分类系统[J]. 四川理工学院学报:自然科学版, 2008, 21(3):93-96.
[16]周志华.机器学习[M].北京:清华大学出版社,2016.Zhou Zhihua.Machine Learning[M].Beijing:Tsinghua University Press,2016.
[17]李晓霞,吴薇薇,韩东,石钰婷.基于聚类与贝叶斯网络的航班离港延误预测模型[J].哈尔滨商业大学学报(自然科学版),2020,36(01):110-113+120.