基于“智慧政务”的分析与挖掘

摘 要

网络问政平台凭借运行成本低，方便快捷等优势，已成为政府了解民意，居民反馈问题的重要渠道。因此，快速解决对大量留言数据归类和热点问题进行整理，提高相关部门及时处理群众反映的问题的效率等问题就显得极为迫切。本文将基于自然语言处理技术，对智慧政务系统平台收集的群众留言记录信息进行智能的归类处理。主要内容如下：

针对问题一，本文首先将附件 2 中的留言进行文本预处理，预处理步骤包括：中文分词，停用词过滤及绘制词云图等，然后利用 TFIDF 权重法对文本数据进行文本特征提取，同时将文本数字化，构造词汇—文本矩阵，最后，将其划分为 80%的训练集和 20%的测试集用贝叶斯算法进行模型的训练测试，并采用 F-Score 系数进行模型的评价。

针对问题二，本文首先将附件 3 中的数据进行预处理，用 TFIDF 进行文本特征提取 TOP 关键词并转换成矩阵。利用余弦性相似度对附件 3 进行文本相似度计算，并利用自定义的热度指数公式得出相应热点问题的热度指数，得出词频矩阵，最终导出“热点问题表”（排名前五的热点问题）和“热点问题留言明细表”。

针对问题三，本文将附件 4 中的答复意见进行数据预处理，通过导入情感评价表，对答复意见”进行评价。相关性：利用 BM25 算法，计算答复关键词与群众留言关键词之间的相关性分数，从而给出答复的评价一；完整性：利用自定义的分词库和基于空间向量的 KNN 算法和余弦算法，判断答复中关键词与规范分词库中关键词的相似度，余弦计算值越接近 1 则评价二结果越好 ；可解释性：构建 LDA 模型，提取关键字，建立字典和语料库，将答复意见的关键词和答复中引用的相关法律法规的关键词进行比对，从而给出答复的评价三。自定义评价指标表进行打分，最终得到加权总分，给出结果评价。

最后我们基于对“智慧政务”的留言处理效率分析得出结论，并提出针对性建议。

关键词：TFIDF、贝叶斯算法、F-Score、BM25 算法、KNN 算法、LDA 模型

Analysis and mining based on "smart government affairs"

Abstract

With the advantages of low operating cost, convenience and quickness, the network administration platform has become an important channel for the government to understand public opinion and residents' feedback. Therefore, it is extremely urgent to quickly sort out a large number of message data classifications and hot issues, and improve the efficiency of relevant departments in handling the problems reflected by the masses in a timely manner. This article will use the natural language processing technology to intelligently classify the mass message record information collected by the smart government system platform.

Aiming at the problem one: This article firstly deduplicates and blanks the message text in Attachment 2, pre-processes data such as Chinese word segmentation and stopword filtering, then extracts features based on TFIDF, digitizes the text, and constructs a vocabulary-text matrix Then, the dimensionality reduction is performed according to the latent semantic analysis LSA decomposition algorithm SVD, and finally the text vector is classified according to the K-means clustering algorithm. Finally, based on the data in Annex II, 70% were randomly selected for model training, 30% were used for model testing, and F-Score coefficients were used for model evaluation.

For problem two: This article first preprocesses the data in Annex 3, uses TFIDF for text feature extraction, and converts TOP keywords into matrices. Use cosine similarity to calculate the text similarity of Annex 3, and use the custom heat index formula to get the heat index of the corresponding hotspot problem, get the word frequency matrix, and finally derive the "hotspot problem table" (the top five hotspot problems) And "List of Hot Messages".

Aiming at question three: This article pre-processes the responses to the comments in Annex 4 and evaluates the responses by importing an emotional evaluation form. Relevance: Use the BM25 algorithm to calculate the correlation score between the keywords of the response and the keywords of the mass message, so as to give an evaluation of the relevance of the response; completeness: use a custom word segmentation library an,the KNN algorithm and the cosine algorithm based on space vectors , To judge the similarity between the keywords in the response and the keywords in the canonical lexicon, the closer the cosine calculation value is to 1, the

better the completeness; interpretability: construct the LDA model, extract the keywords, build a dictionary and corpus, and respond to the comments The keywords are compared with those of relevant laws and regulations cited in the reply, so as to give an evaluation of the interpretability of the reply. Customize the evaluation index table to score , and finally get the weighted total score , and give the result evaluation .

Finally, we draw conclusions based on the analysis of the message processing efficiency of "smart government affairs" and make targeted suggestions.

Keywords: TFIDF , F-Score, SVD algorithm, BM25 algorithm, KNN algorithm, LDA model

目　录

1 挖掘目标

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，过去依靠人工来进行留言划分和热点整理的相关部门的工作急需技术提高效率。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。

本文正是基于这一背景，同时利用主办方提供的附件信息中收集的来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数据，再利用自然语言处理和文本挖掘的方法解决相关问题。前期我们仅抽取部分样本进行建模，基本模型建成后，将对全部数据进行测试，再对其进行优化，在模型基本成型后我们将导入全部数据分析。

本次文本挖掘我们首先将对留言问题进行一级标签划分并进行分类结果评价。再者我们将对热点问题进行排序以便政务部门在今后的公务处理中更加高效地解决群众所反映的问题并在过去群众反映较多的问题领域出台新的改进措施，提前预防或者减少相关问题的再发生。接着，我们将对有关部门给予的答复进行分析，了解答复意见的质量并进行评价。最后，我们将根据上述分析结果对“智慧政务”项目开展结果的结论，并给出参考建议。

2 总体流程与步骤

2.1 总体流程

本文的总体框架及思路如下：

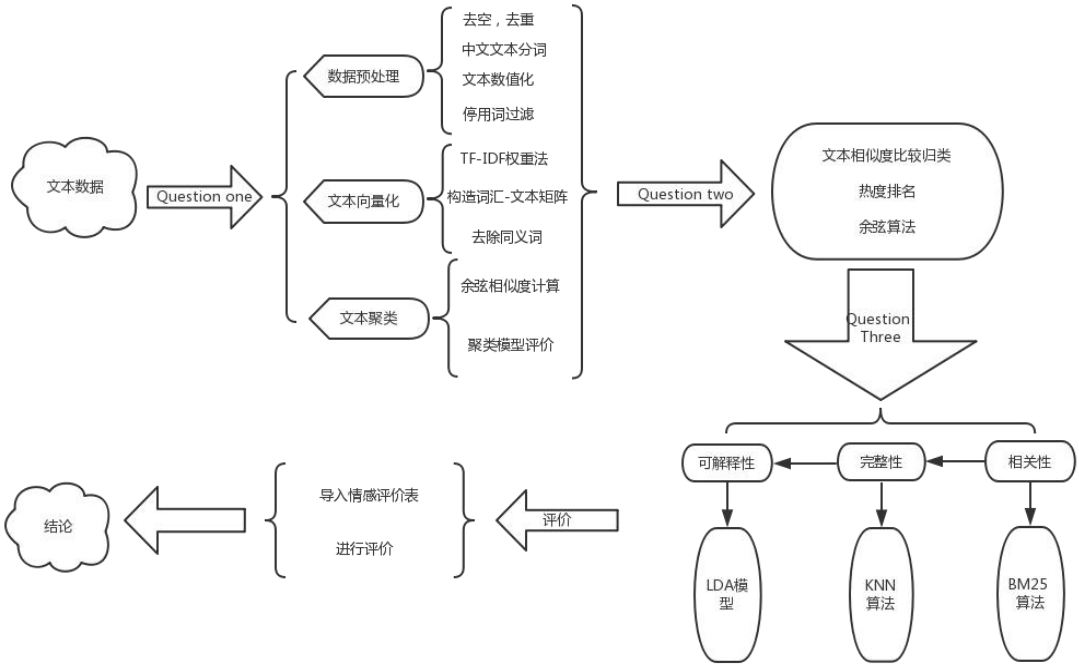


图 2-1 总体框架图

对应的步骤如下：

（1） 数据预处理，对附件 2 非结构化文本进行去掉重复项和空行处理，再对中文文本进行分词、停用词过滤操作，以便后续分析;

（2） 文本向量化，利用 TFIDF 权重法对关键词进行提取，得到词汇-文本矩阵，并对同义词造成的影响进行去除，以简化计算。

（3） 制定热度评价指标（自定义热度指数计算公式），利用 TFIDF 权重法提取 TOP 关键词，利用余弦相似度对附件三进行文本相似度计算。去重分类构建词频矩阵。取前五个热点问题导出“热点问题表.xlsx”。

（4） 进行热度排名，导出“热点问题留言明细表.xlsx”。

（5） 利用 BM25 算法计算相关性；通过 KNN 计算，判断完整性；构建 LDA 模型判断可解释性。

（6） 导入情感评价表，对答复意见进行评价。

3 文本聚类

3.1 数据预处理

3.1.1 数据描述

通过附件中已获得的相关数据，已知数据整体量较大，前期我们仅抽取部分样本进行建模，基本模型建成后，再对其进行优化。由于数据中存在大量口语化，情绪化，过多描述个人经历的无意义数据，因此我们将利用去重去空，中文分词及停用词过滤等数据预处理手段初步处理数据，再利用 TF-IDF 算法将文本数据化。附件 2：留言主题和留言详情，附件 3：数据热度，附件 4：答复意见，都存在大量无意义重复形式表达（噪声特征），如果不进行处理将对后续分析造成一定影响，同时影响模型计算效率及质量，因此本文先进行相关数据预处理。

3.1.2 文本预处理

我们把这些文本数据的预处理分为三个部分：

（一）去空，去重

对于附件 2 中的群众留言进行筛选，对于留言相同的文本或者留言详情为空的文本进行去除处理。

（二）中文分词

在中文文本中，我们可以明显看出词与词之间的界限并不明显，我们在进行文本的挖掘时需要从文本中提取关键词，因此在预处理时我们需要对文本进行分词，本文直接采用由 Python 开发的便捷中文分词库——jieba 分词库，分别对附件 3 中的留言主题和留言详情进行中文分词，以下是 jieba 分词库的基本计算方法：

（1） 利用动态规划以查找最大概率路径和基于词频的最大切分组合。

（2） 采用基于汉字成词能力的 HMM 模型，并通过 Viterbi 算法进行处理（主要针对未登陆词）。

Jieba 库主要提供分词、进行词性标注和未登录词的识别等功能，并且支持用户自定义词典，进行关键词提取。

部分分词结果显示如图 3-1：

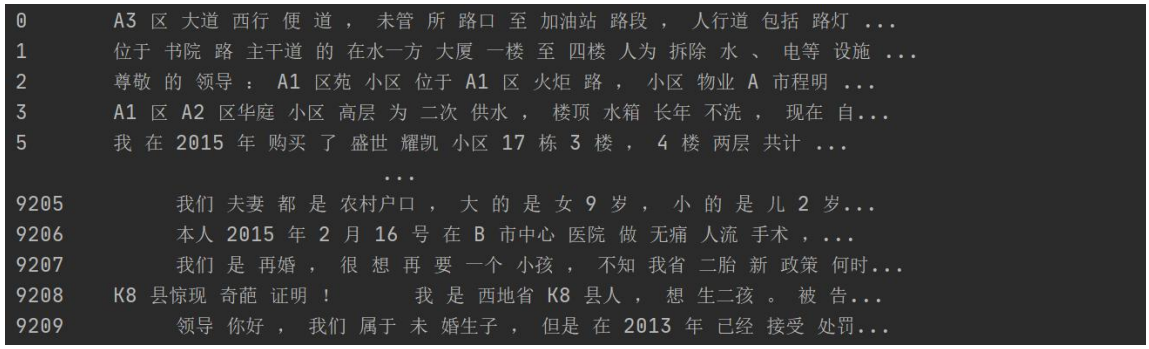


图 3-1 部分分词结果

图 3-1 为部分分词结果，此时还未进行停用词过滤，因此，我们可以发现其中存在大量标点和无意义的带有明显个人情感色彩的词语，此类带有明显个人情感的单词会对后续文本挖掘分析造成不良影响。因此，接下来我们将对这类无意义表达单词进行停用词过滤。

（三）停用词过滤

在实际文本挖掘项目中存储空间和搜索效率十分重要，在分析挖掘文本数据之前，我们需要自动过滤掉某些表达无意义的字或者词，即停用词。停用词具有以下两个特征显著特征：具有普遍性或出现频率较高和信息量低，带有个人情感色彩或对关键词主题无意义。

为了提高在后续分析中提取到的关键词的有效性，我们利用单词在各文档中的出现频率作为噪声词的判定标准。经分析发现，高频词通常具有高噪声值，但实际上在少数文本中出现的部分高频词并不应该被划分为噪声词。因此我们利用以下指标判断词语是否为噪声词：

（1） TF

TF 即简单评估函数，取值为训练集合中某单词的词频数。该评估函数的判断意义为当某个词大量出现时，一般我们判定该词为噪声词。

（2） DF

DF 是不同于 TF 的另一种简单评估函数，取值为训练集合中包含此单词的文本数。该评估函数的判断意义是当某个词在大量文档中出现时，一般我们判定该词为噪声词。

将筛选处理过的词加入自定义的停用词表，再利用自定义的停用词表对附件 2 和附件 3 中的文本数据进行停用词过滤处理，将已得到的分词结果与自定义停用词表中的单词进行比较，若匹配相符，则采用删除处理。

去除停用词后的分布结果如图 3-2：

```
0       A3 区 大道 西行 便 道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建...
1       位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水 电等 设施 烂尾 多年 ...
2       尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 A 市程明 物业管理 有...
3       A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 ...
5       2015 年 购买 盛世 耀凯 小区 17 栋 3 楼 4 楼 两层 共计 2 千 平方 一...
                                ...
9205    夫妻 农村户口 女 9 岁 2 岁 半 15 斤 治疗 两年 一级 脑瘫 纯 女户 招郎 男...
9206    2015 年 2 月 16 号 B 市中心 医院 做 无痛 人流 手术 手术 怀孕 症状 2...
9207                再婚 想 一个 小孩 不知 我省 二胎 新 政策 出 先 怀孕 会 做 处理
9208    K8 县惊现 奇葩 证明 西地省 K8 县人 想 生二孩 告知 要开 证明 没生 二孩 证明...
9209    领导 你好 属于 未 婚生子 2013 年 已经 接受 处罚 小孩 上户 小孩 外地 上学 ...
```

图 3-2 停用词过滤后分词结果

3.2 文本特征抽取

在完成上述文本预处理后，虽然部分停用词已经去掉，但还有部分词语对文本向量化处理产生了不良影响，因此我们将进行特征抽取。该处理的主要目的在于不改变文本原有核心信息的情况下，尽量减少词语的数量，尽可能地降低向量空间的维数，达到简化计算的目的，提高文本挖掘处理效率。本文将使用词频 TF-IDF 对文本进行降维处理。

TF-IDF 中文名称维词频-逆向文档频率，词频（TF）即词语在文本中出现的频率（次数），一个词语的词频越高其权重就越高，TF-IDF 基本公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3-1}$$

以上的式子中 $n_{i,j}$ ， j 是该词在文件中的出现次数，分母为文件中所有字词的词频之和。

逆向文档频率（IDF）即在少数文本中出现某一个词的频率比在多数文本中出现该词的频率高，原因为在聚类中这些词更具有区分能力。

以下为基本公式：

$$idf_i = \log \frac{N}{\left|\{j : t_i \in d_j\}\right|} \tag{3-2}$$

其中， N 语料库中的文件总数， $\left|\{j : t_i \in d_j\}\right|$：包含词语 t_i 的文件数目，如果该词语不在语料库中被除数记为 0，我们在一般情况下将使用 $1+\left|\{j : t_i \in d_j\}\right|$。

最后，经过总结我们可以得出：

$$w_{ij} = tf_{ij} \times idf_i \tag{3-3}$$

以上公式（3-3）即为词的权重。通过利用 TF-IDF 权重法抽取特征词条，我

们可以依据从大到小的顺序队词的权重进行排序，最终将权重较大的特征词抽取出来。作为本文的候选特征词。

以下为抽取特征词条后得到的词云图：

图 3-3 词云图一

图 3-4 词云图二

3.3 文本的空间向量模型

众所周知，计算机对于文本信息无法进行直接处理，因此在进行文本分析的过程中我们首先将对文本进行数字化处理。在文本特征表达中，要求文档的内容能够被准确地反映出来，且具有对不同文档的区分能力。本文将采用向量空间模型来进行文本表示。

向量空间模型：将不同的每个文本表示为向量空间中的一个向量，以每一个不同的特征项作为向量空间的维度，每一个维度的值就是对应的特征项在文本中的权重，此处的权重主要由 TF-IDF 权重法计算处理。其基本形式如下：

$$V(d)=(t_1,w_1(d),t_2,w_2(d),\ldots,t_n,w_n(d)) \qquad (3-4)$$

其中，$t_i(i=1,2,...,n)$为文档中的特征项，$w_i(i=1,2,...,n)$为特征项的权值，二者皆由 TF-IDF 权重法计算得出。

3.3.1 文本的向量化表示

本文在文本特征抽取过程中将对全部特征项进行筛选，并构建出一个词袋，规定文本的特征性对应词袋中的位置，组成统一维度的向量：

$$C=(t_1,t_2,...,t_n) \tag{3-5}$$

其中 C 为词袋集合，t_n是每个词在向量中的位置。

通过上述处理，留言文本信息根据词袋得出了同一维数的词向量，我们将再次运用 TF-IDF 权重法将文本数据向量化得到类似于以下的词汇-文本矩阵：

$$\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array}\begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{pmatrix} \end{array} \tag{3-6}$$

上述过程通过代码实现的详细步骤见附件，图 3-5 为部分结果截图：

	0	1	2	3	4	5	6	7	8	9	10
0	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.1017682900	0.0000000000	0.0000000000
1	0.2704207898	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0842273419	0.0000000000	0.0000000000
2	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
3	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
4	0.1239377183	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0772051926	0.0000000000	0.0000000000
5	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
6	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
7	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
8	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
9	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
10	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
11	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0338201664	0.0000000000
12	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
13	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
14	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0738561353	0.0000000000
15	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

图 3-5 词汇-文本 TFIDF 权重矩阵图

图 3-6 为代码实现导出的词频矩阵部分结果截图：

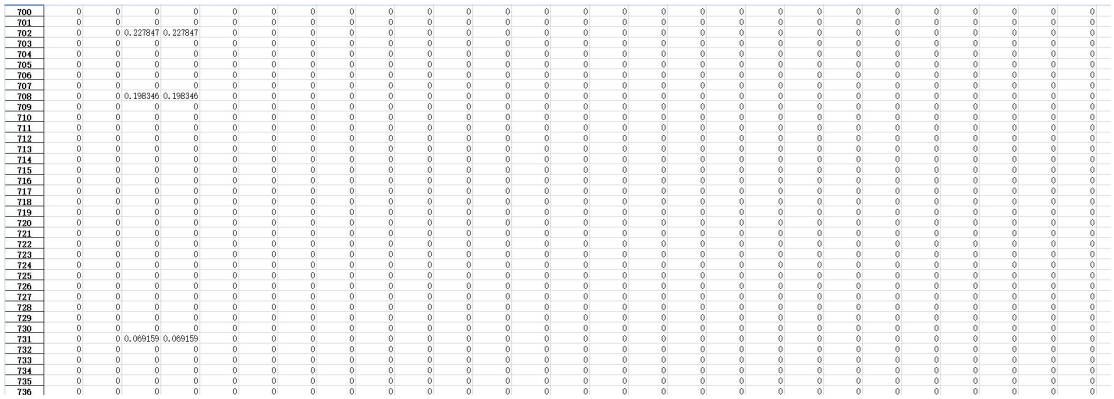


图 3-6 导出的部分词频矩阵图

3.4 贝叶斯分类算法

贝叶斯分类是以贝叶斯定理为基础的同一类分类算法的总称，在该类分类算法中朴素贝叶斯分类是最简单常见的一种分类方法。本文将采用贝叶斯分类算法解决问题一（对群众留言的分类）。

3.4.1 贝叶斯算法基本原理

分类问题在数学领域有如下定义：已知集合 $C=y_1,y_2,...,y_n$ 和 $I=x_1,x_2,...,x_n$，确定映射规则 $y=f(x)$，使得任意 $x_i \in I$ 有且仅有一个 $y_i \in C$，使得 $y_i \in f(x_i)$ 其中 C 叫做类别集合，其中每一个元素是一个类别，而 I 叫做项集合（特征集合），其中每一个元素是一个待分类项，f 叫做分类器。分类算法的任务就是构造分类器 f 成立。[1]

贝叶斯分类算法是基于贝叶斯定理衍生出的一种文本分类算法，其基本公式如下：

$$P(B \mid A)=\frac{P(A \mid B)P(B)}{P(A)} \qquad (3-7)$$

3.4.2 朴素贝叶斯分类原理

朴素贝叶斯分类的基本原理：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

以下为正式定义：

（1）设 $x=\{a_1,a_2,..,a_m\}$ 为一个待分类项，而每一个 a 为 x 的一个特殊属性。

---

[1] 详解朴素贝叶斯分类算法 - 这里记录着我一点一滴的进步 - CSDN 博客 《互联网资源库 (https://blog.csdn.net/cc)》

（2）有类别集合$C=\{y_1,y_2,...,y_n\}$。

（3）计算$P(y_1|x),P(y_2|x),...,P（y_n|x)$。

（4）如果 $P(y_k|x)=\max\{P(y_1|x),P(y_2|x),...,P(y_n|x)\}$，则$x\in y_k$。

以下为第（3）步中条件概率的计算方式：

① 找到一个已知分类的待分类项集合，设为训练样本集。

② 统计得到在各类别下各个特征属性的条件概率估计，基本公式如下：

$$\begin{aligned}&P(a_1|y_1),P(a_2|y_1),...,P(a_m|y_1);\\&P(a_1|y_2),P(a_2|y_2),...,P(a_m|y_2);\\&P(a_1|y_n),P(a_2|y_n),...,P(a_m|y_n)\end{aligned}\qquad(3-8)$$

③ 若各个特征属性是独立的，则根据贝叶斯定理有如下推导：

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)}\qquad(3-9)$$

其中分母对于所有类别的情况都为常数，故我们只需要对分子进行最大化处理。同时我们知道各特征属性是条件独立的，所以有：

$$P(x|y_i)P(y_i)=P(a_1|y_i)P(a_2|y_i)....P(a_m|y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j|y_i)\qquad(3-10)$$

3.4.3 估计类别下特征属性划分的条件概率及 Laplace 校准

朴素贝叶斯分类的关键性步骤即计算各个划分的条件概率 P(a|y)，当特征属性为离散值时，统计训练样本中各个划分在每个类别中出现的频率即可用来估计 P(a|y)，下面将主要讨论特征属性是连续值的情况。

（一）特征属性为连续值时

本文假定其值服从高斯分布（也称正态分布）。即：

$$g(x,\eta,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\eta)^2}{2\sigma^2}}\qquad(3-11)$$

而$P(a_k|y_i)=g(a_k,\eta_{y_i},\sigma_{y_i})$ (3-12)

经过上述计算，我们还需求出训练样本中各个类别中此特征项划分的各均值和标准差，再代入上述公式即可得到需要的估计值。

（二）当 P(a|y)=0

该情况主要出现于当某个类别下某个特征项划分没有出现时，此类情况下分类

器质量降低无法满足需求。因此我们利用 Laplace 校准（即对没类别下所有划分的计数加 1）进行改进，当训练样本集数量足够大时，不会对结果产生不良影响影响，并且解决了当频率为 0 时可能引发错误的情况。

3.5 一级分类标签

通过上述数据处理和建立模型分类，我们根据附件一提供的内容分类三级标签体系对群众在“智慧政务”网络平台的留言进行分类，在实际分类过程中，我们该内容分类三级标签体系作出了部分修改以提高体系标签的明确性。

表 3-1 一级分类标签表

一级分类标签		
城乡建设	党务政务	国土资源
纪检监察	交通运输	经济管理
民政	农村农业	商贸旅游
政法	教育文体	劳动和社会保障
环境保护	科技与信息产业	卫生计生

3.6 F-Score 评估方法

3.6.1 引入背景

（一） 查准率（精度）是判断某一检索系统的信号噪声比的一种指标，即检出的相关文献量与检出的文献总量的百分比。公式表示为：

查准率=（检索出的相关信息量/检索出的信息总量）x100%

使用专指性较强的检索语言(如上位类、上位主题词)能提高查准率，但查全率下降。

（二） 查全率（召回率）是判断某一检索系统从文献集合中检出相关文献成功度的一项指标，即检出的相关文献量与检索系统中相关文献总量的百分比。公式表示为：

查全率=（检索出的相关信息量/系统中的相关信息总量）x100%

使用泛指性较强的检索语言(如上位类、上位主题词)能提高查全率，但查准率下降。

以上提到的查准率和查全率评估指标从公式上来看没有相关关系，但在实际处理大规模数据集合中，这两个指标相互制约。一般情况下，查准率提高，查全率就降低；查全率提高，查准率就降低。所以在实际代码运用中，我们需要综合权衡查准率和查全率指标，因此我们引入一个新的指标 F-score。本文将利用 F-Score 评估方法综合考虑查全率和查准率的调和值。

3.6.2. 基本原理

（一）TP、TN、FP 与 FN

①行表示预测的 label 值，列表示真实 label 值

②TP（True Positive）：被判定为正样本，事实上也是正样本。

③FP（False Positive）：被判定为正样本，但事实上是负样本。

④TN（True Negative）：被判定为负样本，事实上也是负样本。

⑤FN（False Negative）：被判定为负样本，但事实上是正样本。

表 3-2

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

（三） 查准率和查全率的计算公式

$$\text{精确度} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3-13}$$

$$\text{查准率} = \frac{TP}{TP + FP} \tag{3-14}$$

$$\text{查全率} = \frac{TP}{TP + FN} \tag{3-15}$$

（四） P-R 曲线和平均精度

P-R 曲线：选取不同阈值时对应的查全率与查准率，在坐标系中描点并用平滑曲线连接得到 P-R 曲线。（如下图 3-7 P-R 曲线）

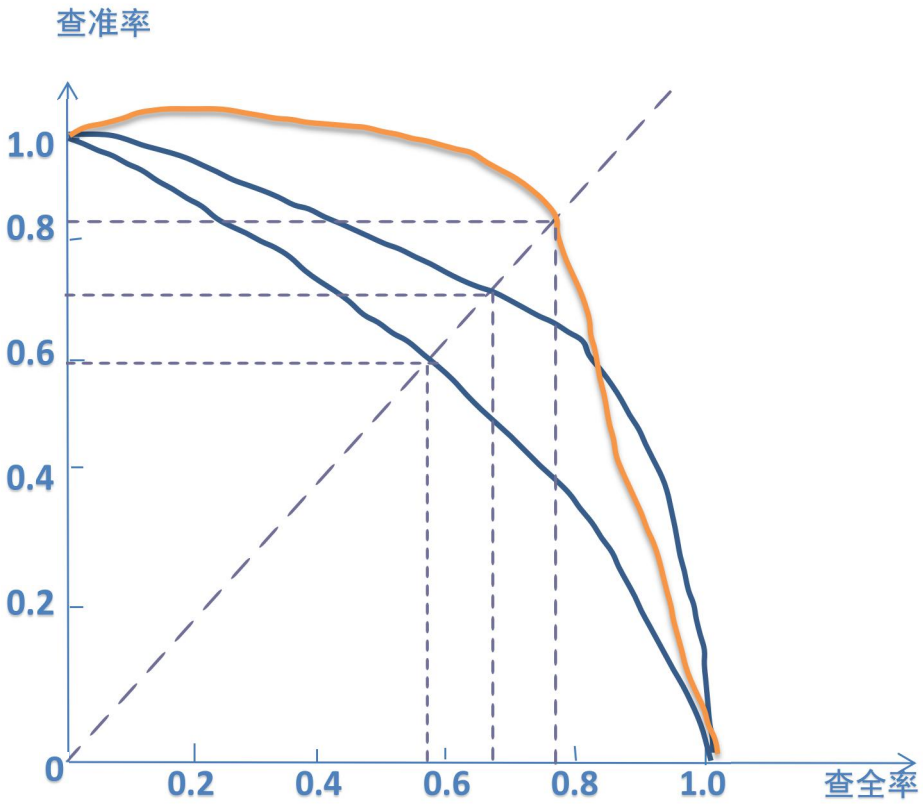


图 3-7 P-R 曲线图

图中与斜线的三个交点即为“平衡点”（BEP）。

（五） F-Score

F-Score（非模型评价打分）是一种评价特征分辨能力的方法，此方法主要用于实现特征选择，较同类方法，F-Score 可以完成最有效的特征选择。F-Score 基本公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \qquad (3-16)$$

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

3.6.3 评估结果

我们利用 F-Score 评估方法对本节一级标签分类体系模型进行准确度评估。经多次测试（测试次数大于 50 次），模型分类准确度在 80.4%至 83.5%间浮动，主要聚集在 82%-83%之间，具体准确度分布见下图 3-8：

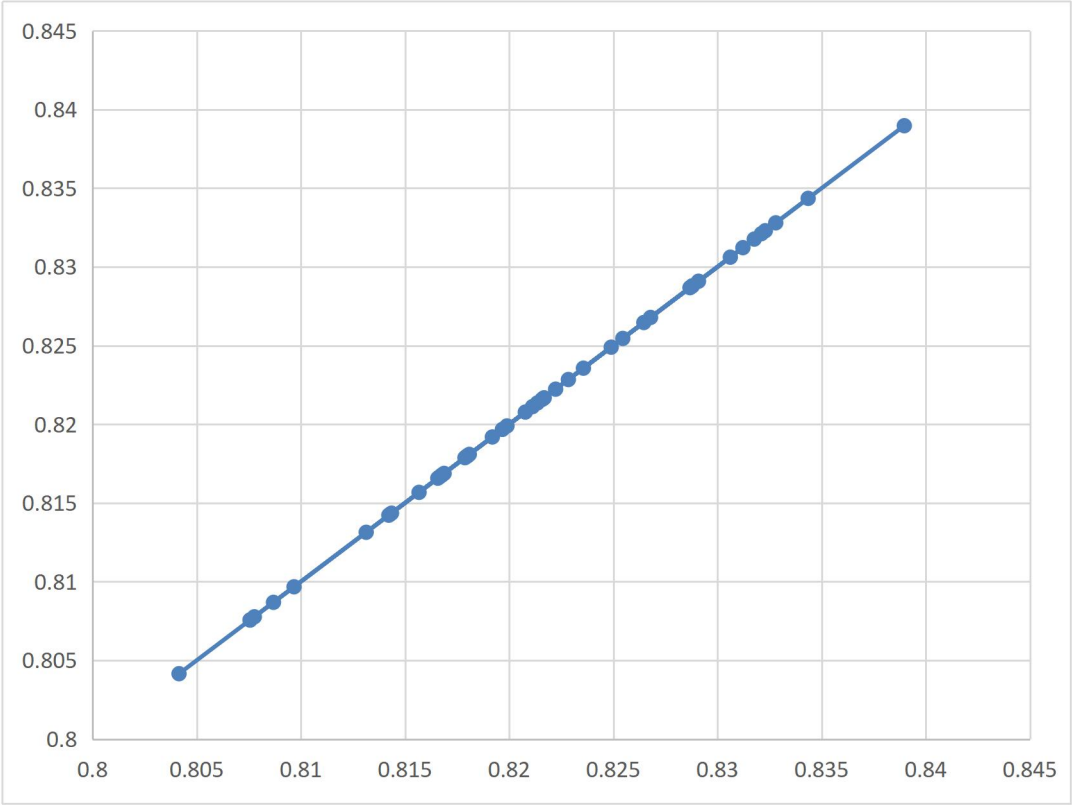


图 3-8 测试准确度分布散点图

下图 3-9 为代码实现 F-Score 评估结果截图：

```
adata, data ,labels=DataProgress(file='D:/contest/total/附件2.xlsx')
data_tr, data_te, labels_tr, labels_te = train_test_split(adata, labels, test_size=0.2)
overlap = CountVectorizer()
data_tr = overlap.fit_transform(data_tr)
X_tr = TfidfTransformer().fit_transform(data_tr.toarray()).toarray()
data_te = CountVectorizer(vocabulary=overlap.vocabulary_).fit_transform(data_te)
X_te = TfidfTransformer().fit_transform(data_te.toarray()).toarray()
model = GaussianNB()
model.fit(X_tr, labels_tr)
model.score(X_te, labels_te)
```

```
0.8142857142857143
```

图 3-9 分类方法准确率截图

4 热点问题分析

4.1 分析热点问题

4.1.1 热点问题的定义

“热点问题”，即某一时段内群众集中反映的某一问题。但此概念较为模糊，并不能直接用于模型分析，因此，我们需要重新分析所谓“热点问题”的基本特征。本文通过主观分析，认为热点问题需具备以下几个特征：

1） 某一时间段，同一地点出现的同一种类问题;

2） 多次被群众反映;

3） 影响群众日常生活、出行或工作。

针对上述几大特征，我们设立关键词并结合所给数据（“附件 3.xlsx”），本文用时间范围、地点/人群、问题描述作为关键，利用自定义的热度指标公式计算所有热点问题的热度指数，最终得出“热点问题留言明细表”。

4.1.2 数据预处理

（一）“智慧政务”平台留言详情（附件 3）中，给出了留言主题、留言时间及留言详情，并在表格末端给出点赞数和反对数。

（二）对于留言时间、点赞/反对数本文已转化为数值型数据，而对于留言详情，本文先进行去空去重、停用词过滤处理（图 4-1），再利用 TFIDF 权重法提取 TOP 关键词，得到词频矩阵（图 4-2）。

0	A3 区 大道 西行 便 道 未管 路口 加油站 路段 人行道 包括 路灯 杆 …
1	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水 电等 设…
2	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 A 市程明 物…
3	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来…
5	2015 年 购买 盛世 耀凯 小区 17 栋 3 楼 4 楼 两层 共计 2 千 平方 …
6	西地省 地区 常年 阴冷 潮湿 气候 近年 气候 逐渐 更加 恶劣 地处 月…
7	尊敬 胡书记 您好 家住 A 市 A3 区 桐梓 坡 西路 可可 小城 居民 长期…
8	梅家田 社区 辖区 小区 居民 每年 依法 依规 小区 物业公司 交纳 城市…
9	尊敬 A 市政府 领导 您们好 A 市 A3 区 魏家坡巷 业主 多年 小区 脏 …
11	请求 依法 监督 泰华 一村 小区 第四届 非法 业主 委员会 涉嫌 侵占 …
12	住 梅 溪湖 壹号 御湾 4 楼 2019 年 8 月份 住 进来 每天晚上 会 停…
13	尊敬 领导 你们好 A 市 A4 区 捞刀河 镇 彭家巷 社区 鸿涛 翡翠 湾 一…
14	地铁 5 号线 施工 导致 万家 丽路 锦楚 国际 星城 小区 三期 一个月 …
15	尊敬 领导 你好 A6 区润 紫 郡 业主 今年年初 小区 周边 竖起 一道道 …

图 4-1 Spyder 处理后的部分结果截图

	0	1	2	3	4	5	6	7	8	9	10
0	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.1017682900	0.0000000000	0.0000000000
1	0.2704207898	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0842273419	0.0000000000	0.0000000000
2	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
3	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
4	0.1239377183	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0772051926	0.0000000000	0.0000000000
5	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
6	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
7	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
8	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
9	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
10	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
11	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0338201664	0.0000000000
12	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
13	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
14	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0738561353	0.0000000000
15	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

图 4-2 TFIDF 词频矩阵图

以下图 4-3 为关键词提取结果：

	A	B	C	D
1		id	title	key
2	0	188006	A3区一米	一米阳光
3	1	188007	咨询A6区	成果
4	2	188031	反映A7县	到户
5	3	188039	A2区黄兴	外排
6	4	188059	A市A3区中	空地
7	5	188073	A3区麓泉	架空层
8	6	188074	A2区富绿	开关厂
9	7	188119	对A市地铁	上班
10	8	188170	A市6路公	指示
11	9	188249	A3区保利	松路
12	10	188251	A7县特立	信号灯
13	11	188260	A3区青青	冰柜
14	12	188396	关于拆除	变压器
15	13	188399	A市利保壹	市利
16	14	188409	A市地铁3	星沙
17	15	188414	A4区北辰	改商
18	16	188416	请给K3县	证件
19	17	188451	A7县春华	党员
20	18	188455	咨询异地	出国
21	19	188467	投诉A市漏	退费
22	20	188475	A6区乾源	违章
23	21	188535	A7县时代	旅馆
24	22	188546	A2区佳兆	业水
25	23	188553	A市沙坪老	理疗
26	24	188560	A市德鸿馨	和解
27	25	188592	A市长房云	垃圾站

图 4-3 关键词提取截图

4.2 热度评价指标

及时发现热点问题对提高政务部门开展线上政务答复处理的效率有重大意义。本文已根据“附件三”将反映某一时间段内特定地点或特定人群问题的留言进行分类，并通过定义评价热度指标公式对热点问题进行热度排序。

定义计算热度指数公式如下：

$$热度指数 = 5 \times 问题条数 + 总点赞数 - 总反对数 \qquad (4-1)$$

本文得出的“热点问题留言明细表”即基于上述公式进行热度排序。

4.3 余弦相似度计算

相似度是用来衡量文本间相似程度的一个标准。在文本分类中，需要研究文本个体将的差异大小，也就是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。相似度计算有距离度量和相似度度量两种方法。本文采用的是余弦相似度算法。

4.3.1 最长公共序列

余弦相似度 (Cosine Similarity)：

通过计算两个向量的夹角余弦值来评估他们的相似度。余弦相似度将向量根据坐标值，绘制到向量空间中。

最长公共子序列（基于权值空间、词条空间）

（1）将两个字符串分别设做行和列，组成矩阵。

（2）计算每个节点行列字符是否相同（相同则为 1）。

（3）通过找出值为 1 的最长对角线即可得到最长公共子串。

（4）为进一步提升该算法，我们可以将字符相同节点的值加上左上角（d[i-1，j-1]）的值，这样即可获得最大公共子串的长度。如此一来只需以行号和最大值为条件即可截取最大子串。

4.3.2 最小编辑距离

最小编辑距离算法（基于词条空间）

（1）狭义编辑距离

设 A、B 为两个字符串，狭义的编辑距离定义为把 A 转换成 B 需要的最少删除（删除 A 中一个字符）、插入（在 A 中插入一个字符）和替换（把 A 中的某个字符替换成另一个字符）的次数，用 ED（A，B）来表示。

（2）步骤

a）对两部分文本进行处理，将所有的非文本字符替换为分段标记“#”

b）较长文本作为基准文本，遍历分段之后的短文本，发现长文本包含短文

本子句后在长本文中移除，未发现匹配的字句累加长度。

c）比较剩余文本长度与两段文本长度和，其比值为不匹配比率。

4.3.3 余弦相似值

根据坐标值，将向量绘制到向量空间中。如最常见的二维空间。求出夹角对应的余弦值。

以下图（图 4-4）为例，我们通过向量的方法求夹角θ的值，通过基本的数学知识我们知道可通过余弦定理来解决：

余弦定理公式：

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}} \qquad (4-2)$$

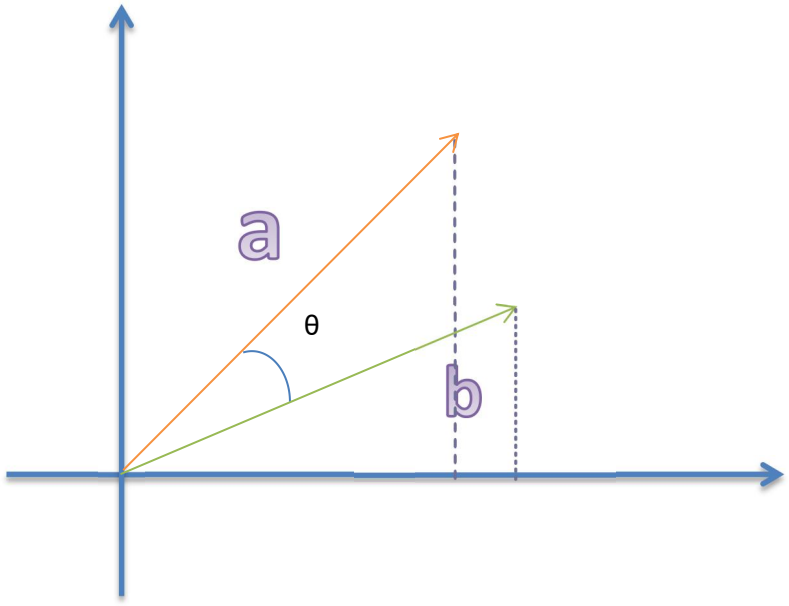


图 4-4 余弦示例

同理，将余弦定理引申至 n 维向量，我们可以得出下列结论：

假定有同一坐标系中的两个 n 维向量 A 和 B，其中 A 向量的坐标为 $(A_1, A_2, ..., A_n)$，B 向量的坐标为 $(B_1, B_2, ..., B_n)$，则 n 维向量的夹角θ的余弦可表示为：

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} + \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{\vec{A} \bullet \vec{B}}{|\vec{A}| \times |\vec{B}|} \qquad (4-3)$$

将余弦定理思想运用至文本相似度计算，即余弦相似度计算算法。在余弦算法中，此余弦值就可以用来表示这两个向量的相似性。夹角越小，余弦值越接近于 1，则越相似。因此，我们将利用夹角的大小，来判断相似度。夹角越小，就代表越相似。

4.4 热点问题排名列举

通过代码分析，我们发现不同阈值下，热点问题排名有差异，故我们对不同阈值进行测试，以下将列出不同阈值下的代码运行结果并对其作出比较，最终将根据结果，选择最理想的阈值。经统计，部分热点问题排名（热度排名前五的热点问题，即“热点问题表.xlsx”）列举如下：

（一）阈值为 0.5 时

下表 4-2 给出在阈值为 0.5 的情况下热度排名前五的热点问题：

表 4-2 热点问题表（阈值为 0.5）

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	2102	2019/8/1 2019/8/1	A 市 小区	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	2	1767	2019/3/1 2019/4/1	A 市 金毛	反映 A 市金毛湾配套入学的问题
3	3	831	2019/2/2 2019/8/2	A 市 受害人	请书记关注 A 市 A4 区 58 车贷案
4	4	795	2019/2/2 2019/2/2	A 市 涉案人员	严惩 A 市 58 车贷特大集资诈骗案保护伞
5	5	738	2019/3/1 2019/3/1	A 市 留言	承办 A 市 58 车贷案警官应跟进关注留言

以下图 4-5 给出在阈值为 0.5 时生成的“热点问题留言明细表.xlxs”的部分结果截图：

	A	B	C	D	E	F	G	H
1	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
2	1	208636	A00077171	A市A5区汇金	2019/8/19 11:34:04	我是A市A5区	2097	0
3	2	223297	A00087522	反映A市金毛	2019/4/11 21:02:44	书记先生：您	1762	5
4	2	255733	A00061749	A市旭辉御府	2019/3/12 3:18:35	胡书记您好：	0	0
5	3	220711	A00031682	请书记关注A	2019/2/21 18:45:14	尊敬的胡书记	821	0
6	3	229783	A00063792	请关注A4区	2019/8/2 19:10:27	A4区幸福桥社	0	0
7	4	217032	A00056543	严惩A市58车	2019/2/25 9:58:37	胡市长：您好	790	0
8	5	194343	A000106161	承办A市58车	2019/3/1 22:12:30	胡书记：您好	733	0

图 4-5 热点问题留言明细表（阈值为 0.5）

（二）阈值为 0.4 时

下表 4-3 给出在阈值为 0.4 的情况下热度排名前五的热点问题：

表 4-3 热点问题表（阈值为 0.4）

热度排名	问题 ID	热度指数	时间范围（始-终）	地点 人群	问题描述
1	1	2134	2019/6/2 2019/9/1	A 市 装修	A 市五矿万境 K9 县交房后仍存在诸多问题
2	2	2119	2019/5/5 2019/9/1	A 市 小区	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
3	3	1767	2019/3/1 2019/4/1	A 市 小区	反映 A 市金毛湾配套入学的问题
4	4	836	2019/2/2 2019/8/2	A 市 受害人	请书记关注 A 市 A4 区 58 车贷案
5	5	818	2019/2/2 2019/8/2	A 市 涉案人员	严惩 A 市 58 车贷特大集资诈骗案保护伞

以下图 4-6 给出在阈值为 0.4 时生成的“热点问题留言明细表.xlxs”的部分结果截图，完整版热点问题留言明细表见附件“热点问题留言明细表”。

	A	B	C	D	E	F	G	H
1	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
2	1	208636	A00077171	A市A5区汇	2019/8/19	我是A市A5	2097	0
3	1	215507	A00010323	A市五矿万	2019/9/12	预交房23栋	1	0
4	1	234086	A00099869	A市五矿万	2019/6/20	五矿万境K	6	0
5	1	252650	A00010531	A市五矿万	2019/9/11	尊敬的相关	0	0
6	1	262599	A00010042	A市五矿万	2019/9/19	我是西地省	0	0
7	1	275491	A00061339	A市五矿万	2019/9/10	关于五矿万	0	0
8	2	208069	A00094436	A5区五矿万	2019/5/5	本人是A5区	2	0
9	2	208636	A00077171	A市A5区汇	2019/8/19	我是A市A5	2097	0
10	2	252650	A00010531	A市五矿万	2019/9/11	尊敬的相关	0	0
11	2	262599	A00010042	A市五矿万	2019/9/19	我是西地省	0	0
12	3	223297	A00087522	反映A市金	2019/4/11	书记先生：	1762	5
13	3	255733	A00061749	A市旭辉御	2019/3/12	胡书记您好	0	0
14	4	220711	A00031682	请书记关注	2019/2/21	尊敬的胡书	821	0
15	4	220961	A00015167	A4区的植被	2019/6/4	请问A4区栋	0	0
16	4	229783	A00063792	请关注A4区	2019/8/2	A4区幸福栋	0	0
17	5	217032	A00056543	严惩A市58	2019/2/25	胡市长：您	790	0
18	5	232063	A00083732	请A市依法	2019/8/21	尊敬的市委	20	2

图 4-6 热点问题留言明细表（阈值为 0.4）

（三）阈值为 0.3 时

下图 4-7 给出在阈值为 0.3 的情况下热度排名前五的热点问题：

	A	B	C	D	E	F	G	H	I	J	K
1	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述					
2	1	1	2158	2019//1/1	A5区重点	A5区五矿万境K9县的开发商与施工方建房存在质量问题					
3	2	2	2146	2019//1/1	A市装修	A市五矿万境K9县交房后仍存在诸多问题					
4	3	3	2134	2019//6/	A市墙面	A市五矿万境K9县房子的墙壁又开裂了					
5	4	4	1767	2019//3/	A市金毛	反映A市金毛湾配套入学的问题					
6	5	5	892	2019//1/	A4区改商	A4区北辰小区非法住改商问题何时能解决？					

图 4-7 热点问题表（阈值为 0.3）

以下图 4-8 给出在阈值为 0.3 时生成的“热点问题留言明细表.xlxs”的部分结果截图，完整版热点问题留言明细表见附件“热点问题留言明细表”。

	A	B	C	D	E	F	G	H
1	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
2	1	199836	A0006793	A3区江山	2019/11/1	您好！本	3	0
3	1	201691	A00044429	A3区欧城	2019/12/3	尊敬的相	0	0
4	1	207882	A00041577	A市明昇壹	2019/5/23	周边市民	0	0
5	1	208069	A00094436	A5区五矿	2019/5/5	本人是A5	2	0
6	1	208636	A00077171	A市A5区江	2019/8/19	我是A市A5	2097	0
7	1	232594	A909154	A5区创元	2019/5/13	尊敬的胡	0	0
8	1	251828	A00030641	A市朝阳时	2019/5/24	朝阳时代	0	0
9	1	251901	A00021433	A市南山十	2019/3/28	领导好，	1	0
10	1	252650	A00010531	A市五矿万	2019/9/11	尊敬的相	0	0
11	1	262599	A00010042	A市五矿万	2019/9/19	我是西地	0	0
12	1	285756	A00058732	A市橘郡开	2019/8/17	2014年买	0	0
13	2	208069	A00094436	A5区五矿	2019/5/5	本人是A5	2	0
14	2	208636	A00077171	A市A5区江	2019/8/19	我是A市A5	2097	0
15	2	215507	A00010323	A市五矿万	2019/9/12	预交房23	1	0
16	2	228825	A909142	A市禹泰云	2019/11/5	禹泰云开	0	0
17	2	234086	A00099869	A市五矿万	2019/6/20	五矿万境	6	0
18	2	252650	A00010531	A市五矿万	2019/9/11	尊敬的相	0	0
19	2	262599	A00010042	A市五矿万	2019/9/19	我是西地	0	0
20	2	275491	A00061339	A市五矿万	2019/9/10	关于五矿	0	0

图 4-8 热点问题留言明细表（阈值为 0.3）

（四）阈值为 0.2 时

下图 4-9 给出在阈值为 0.2 的情况下热度排名前五的热点问题：

	A	B	C	D	E	F	G	H	I	J	K
1	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述					
2	1	1	6867	2019//0/7	A4区改商	A4区北辰小区非法住改商问题何时能解决？					
3	2	2	4680	2019//1/	A7县到户	反映A7县春华镇金鼎村水泥路、自来水到户的问题					
4	3	3	4191	2019//0/7	A市物业	A市A5区桃花苑二期房屋存在质量安全问题！					
5	4	4	3499	2019//0/7	A5区违建	A5区佳兆业水岸新别墅一期违建成风					
6	5	5	3432	2019//1/	A7县围档	A7县黄兴镇主干道黄江大道路边围档严重影响居民出行					

图 4-9 热点问题表（阈值为 0.2）

以下图 4-10 给出在阈值为 0.2 时生成的“热点问题留言明细表.xlxs”的部分结果截图，完整版热点问题留言明细表见附件“热点问题留言明细表”。

	A	B	C	D	E	F	G	H
1	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
2	1	188073	A909164	A3区麓泉	2019/3/11	作为麓泉	0	0
3	1	188260	A00053484	A3区青青	2019/5/31	还我宁静	0	0
4	1	188414	A00096844	A4区北辰	2019/8/1	您好！我	0	0
5	1	188535	A00061775	A7县时代	2019/6/13	尊敬的各	0	0
6	1	188546	A0006817	A2区佳兆	2019/1/23	敬爱的领	0	0
7	1	188592	A00039456	A市长房云	2019/6/18	长房云时	0	0
8	1	188679	A00010384	希望A市政	2019/12/1	2019年12	5	0
9	1	188780	A00094754	请依法解	2019/10/2	尊敬的县	0	0
10	1	188799	A00010734	A7县橄榄	2019/5/22	我是橄榄	1	0
11	1	189029	A00080952	A4区楚江	2019/4/3	从2018年	0	0
12	1	189113	A00011205	A市融圣国	2019/7/1	尊敬的领	1	0
13	1	189245	A00042948	A市北辰三	2019/3/19	A市北辰三	0	0
14	1	189345	A00077163	A4区东风	2019/7/31	我是1996	0	0
15	1	189381	A00010981	A市万科魁	2019/12/4	A市万科魁	0	0
16	1	189456	A00028347	A3区云栖	2019/11/1	A3区云栖	0	0
17	1	189635	A00029819	A1区桐阴	2019/7/15	桐阴里夜	0	0
18	1	190019	A00010428	A7县星沙	2019/6/4	领导好，	0	0
19	1	190077	A00011291	A3区欣胜	2019/3/14	尊敬的胡	0	0
20	1	190108	A909240	丽发新城	########	丽发新城	1	0

图 4-10 热点问题留言明细表（阈值为 0.2）

经对上述不同阈值下的运行结果进行比较，我们最终得出当阈值为 0.4 时，导出结果最为准确，故本文将选取 0.4 作为最理想阈值，并将阈值为 0.4 时得到的“热点问题表”和“热点问题留言明细表”导出作为本节解题的最终结果。

5 留言回复的评价

5.1 分析答复意见

5.1.1 答复意见分析的意义

“智慧政务”平台存在的意义不仅是为普通群众对政府及事业单位提出意见提供了方便，同时对普通群众意见的答复情况也应当有所重视。因此，对答复意见进行文本分析评价也是本文进行文本挖掘的一大重点，本节将从答复意见的相关性、完整性、可解释性这三方面制定评价方案，并对“附件 4.xlsx”中答复意见质量进行评价。

5.1.2 数据预处理

本节主要对问题三进行分析解答，首先我们对“附件 4.xlsx”的留言和答复意见进行数据预处理，预处理基本流程与问题一和问题二基本一致。以下为本节问题三的数据预处理流程框架图：

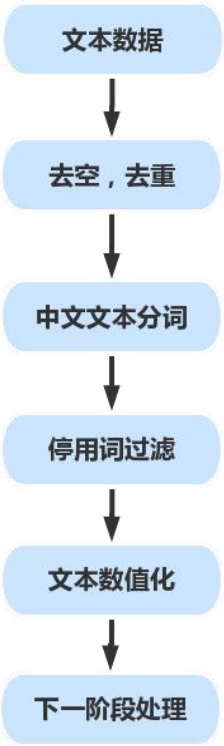


图 5-1 数据预处理流程图

5.2 相关性

5.2.1 相关性定义

答复意见的相关性，即答复意见与对应群众留言问题是否相关。通过对答复意见相关性进行分析，我们可以判断出政务部门是否认真查阅群众意见并给出有针对性的答复。本文对答复意见的相关性定义如下：

（1）当出现多个（个数 $n \geq 5$ ）相同关键词时，我们可初步判定答复意见与群众留言相关。

（2）对同义词进行判断，若同义词（尤其是时间地点人群这三类关键词）基本相符可判定答复意见与群众留言相关。

本文将采用 BM25 算法（非语义匹配算法）对答复意见的相关性进行评价。

5.2.2 BM25 算法

BM25 算法是一种用来评价搜索词和文档之间相关性的算法，它是一种基于概率检索模型提出的算法。

BM25 算法主要思想为：首先对 Query 进行语素解析，生成语素 q_i 词，；再计算每个搜索结果 d 中每个语素 q_i 与 d 的相关性得分；最后将“一个 Query 各个 q_i 相对于 d 的相关性得分”加权求和，得到“Query 与 d 的相关性得分”。

一条语素与任意文档之间的相关性分数公式如下：

$$Score（Q，d）=\sum_{i}^{n} W_i * R(q_i，d) \qquad (5-1)$$

其中 Q 表示语素，$R（q_i,d)$ 是查询语句中每个词 q_i 和文档 d 的相关度值，W_i 是该词的权重，q_i 表示根据语素解析获得的语素，d 表示搜索结果的一条文档。

（1）W_i 的定义

计算一个语素与一个文档的相关性权重，比较常用的公式是：

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \qquad (5-2)$$

上式，N 是搜索结果的全部文档数目，$n(q_i)$ 是包含每个词 q_i 的文档数目。可以看出，$n(q_i)$ 与 $IDF(q_i)$ 是反相关关系，就是说当给定的文档集里，有很多文档都包含 q_i 时，q_i 的区分度就不高，那么使用这个语素 q_i 来判断相关性的重要性就比较低。

（2）$R（q_i,d)$ 的定义

计算一个语素与一个文档的相关性分数，比较常用的公式是：

$$R(q_i,d)=\frac{f_i(k_1+1)}{f_i+K}\frac{qf_i(k_2+1)}{qf_i+k_2} \qquad (5-3)$$

$$K=k_1\left(1-b+b\frac{dl}{avgdl}\right) \qquad (5-4)$$

上式中，k_1，k_2，b 是可根据经验设置的调节因子，一般 $k_1 \in [1.2,2]$，b=0.75；f_i 是 q_i 在文档中出现的频率，qf_i 为 q_i 在语素中的出现频率。dl 为文档的长度，$avgdl$ 为文档集中所有文档的平均长度。在通常情况下，q_i 在语素中只会出现一次，即 qf_i=1，那么公式就可以简化为：

$$R(q_i,d)=\frac{f_i(k_1+1)}{f_i+K} \qquad (5-5)$$

通过K的表达式知，b的作用是调整文档长度 dl 对“相关性的影响”的大小，也就是说 b 越大，文档长度 dl 对“相关性分数的影响”也就越大；而文档长度 dl 越长，计算出来的 K 值也就越大，相关性的分数就越小。

通俗的来讲，当文档较长时，那么包含 q_i 的机会也就越大，在同等出现频率 f_i 的情况下，“长文档与 q_i 的相关性”应该会比“短文档与 q_i 的相关性”弱一些。

5.2.3 相关性评价关键词分值表

基于 5.2.1.相关性的定义，我们对回复与群众留言之间的关键词进行了相似度分析，并制定了评价回复相关性的四个评价关键词：好，较好，一般，差。

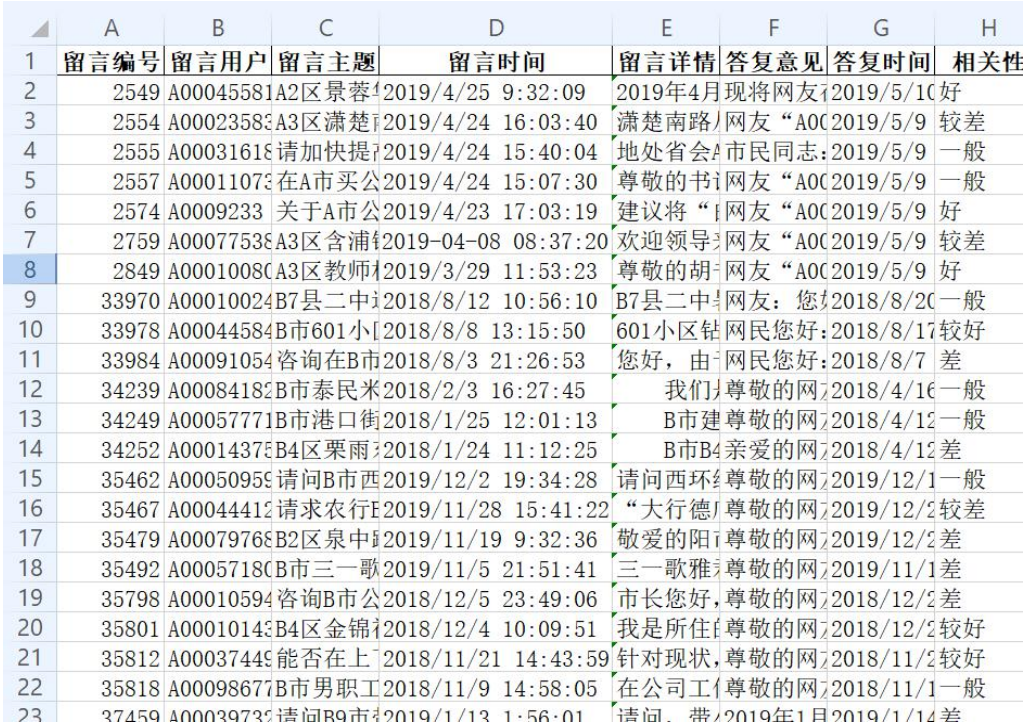
具体的评价标准如下表 5-1：

表 5-1 相关性评价关键词分值表

关键词	好	较好	一般	差
分值（分）	10	8	5	3

5.2.4 相关性代码实现结果

根据 5.2.1 相关性定义得到以下结果：



	A	B	C	D	E	F	G	H
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	相关性
2	2549	A00045581	A2区景蓉	2019/4/25 9:32:09	2019年4月	现将网友	2019/5/10	好
3	2554	A00023583	A3区潇楚	2019/4/24 16:03:40	潇楚南路	网友“A0	2019/5/9	较差
4	2555	A00031618	请加快提	2019/4/24 15:40:04	地处省会	市民同志：	2019/5/9	一般
5	2557	A00011073	在A市买公	2019/4/24 15:07:30	尊敬的书	网友“A0	2019/5/9	一般
6	2574	A0009233	关于A市公	2019/4/23 17:03:19	建议将“	网友“A0	2019/5/9	好
7	2759	A00077538	A3区含浦	2019-04-08 08:37:20	欢迎领导	网友“A0	2019/5/9	较差
8	2849	A00010080	A3区教师	2019/3/29 11:53:23	尊敬的胡	网友“A0	2019/5/9	好
9	33970	A00010024	B7县二中	2018/8/12 10:56:10	B7县二中	网友：您	2018/8/20	一般
10	33978	A00044584	B市601小	2018/8/8 13:15:50	601小区钻	网民您好：	2018/8/17	较好
11	33984	A00091054	咨询在B市	2018/8/3 21:26:53	您好，由	网民您好：	2018/8/7	差
12	34239	A00084182	B市泰民米	2018/2/3 16:27:45	我们	尊敬的网	2018/4/16	一般
13	34249	A00057771	B市港口街	2018/1/25 12:01:13	B市建	尊敬的网	2018/4/12	一般
14	34252	A00014375	B4区栗雨	2018/1/24 11:12:25	B市B4	亲爱的网	2018/4/12	差
15	35462	A00050959	请问B市西	2019/12/2 19:34:28	请问西环	尊敬的网	2019/12/1	一般
16	35467	A00044412	请求农行B	2019/11/28 15:41:22	“大行德	尊敬的网	2019/12/2	较差
17	35479	A00079768	B2区泉中	2019/11/19 9:32:36	敬爱的阳	尊敬的网	2019/12/2	差
18	35492	A00057180	B市三一歌	2019/11/5 21:51:41	三一歌雅	尊敬的网	2019/11/1	差
19	35798	A00010594	咨询B市公	2018/12/5 23:49:06	市长您好，	尊敬的网	2018/12/2	差
20	35801	A00010143	B4区金锦	2018/12/4 10:09:51	我是所住	尊敬的网	2018/12/2	较好
21	35812	A00037449	能否在上	2018/11/21 14:43:59	针对现状，	尊敬的网	2018/11/2	较好
22	35818	A00098677	B市男职工	2018/11/9 14:58:05	在公司工	尊敬的网	2018/11/1	一般
23	37459	A00039732	请问B9市	2019/1/13 1:56:01	请问，带	2019年1月	2019/1/14	差

图 5-2 相关性评价结果（部分）

5.3 完整性

5.3.1 完整性定义

答复意见的完整性，即答复意见与是否具有完整的回复格式（特定的开头与结尾）及是否对居民留言的所有问题作出回复。通过对答复意见完整性分析，我们可以判断出政务部门是否对群众留言中涉及的问题进行全面的答复。本文对答复意见的完整性有如下定义：

（1）具有标准的开头与结尾（规定的公式化的起始语句和结束语句）；

（2）具有特定的被留言栏目、部门或有明确被留言人的姓名及官职；

（3）具有对留言问题进行调查并给出调查结果；

（4）回复意见中引用与留言问题相关法律规章或出台的文件条例（书面语）；

（5）回复意见中具有对条例的白话解释（口语化）；

（6）文末附答复时间或相关部门咨询电话。

本文将采用 KNN 算法对答复意见的完整性进行评价。

5.3.2 KNN 算法

5.3.2.1 KNN 算法描述

K 最近邻（K-Nearest Neighbor,KNN）算法，即 KNN 算法。

KNN 算法通过测量不同特征值之间的距离，根据距离之间的差异来进行分

类。在输入没有标签的数据后，将这个数据的每个特征与样本集中的数据对应的特征进行相互比较，再提取出样本中特征最相近的数据的分类标签。

KNN 算法的中心思想是如果一个样本在特征空间中的 k 个最临近的样本中的大多数属于某一个类别，那么这个样本也属于这个类别。K 通常取不大于 20 的整数。

算法描述如下：

（1）输入训练集的数据和标签，输入测试的数据；

（2）计算测试数据与训练集数据之间的距离；

（3）按照这些距离的递增关系进行大小排序，选取距离最小的 K 个点；

（4）确定前 K 个点所归属类别的出现频率，返回前 K 个点中出现频率最高的类别，以此作为测试数据的预测分类。

KNN 算法的三个要素有：K 值的选取，度量距离的方式和分类决策的规则。

5.3.2.2 K 值的选取

k 值的选取，通常根据样本的特征分布，选取一个比较小的值，也可以通过交叉验证选取一个合适的 k 值，K 通常取不大于 20 的整数。

K 值的选取会对 KNN 算法的结果产生影响：

（1）当选取比较小的 k 值时，也就是说使用较小的邻域中的训练实例进行预测，“学习”的近似误差会相应地减小，但“学习”的估计误差会增大。预测的结果会对近邻的实例点非常敏感，只有与输入实例较近的训练实例才会对预测结果起作用，那么 k 值的减小就意味着整体模型变得非常复杂，也会更加容易产生过拟合的情况。

（2）当选取比较大的 k 值时，也就是说用较大的邻域中的训练实例进行预测，虽然减少了估计误差，但是却增大了近似误差。这时与输入的实例较远的训练实例也会对预测起作用，就会使得预测发生错误。k 值的增大虽然意味着整体的模型变得更加简单，但是 K 值也应该控制在 20 以内的整数。

（3）当 k 值选取样本数时，无论输入的实例是什么，都将简单地预测它属于在训练实例中的最多的类。这时，模型变得过于简单，完全忽略训练实例中的大量有用的信息，是非常不可取的。

因此 K 值的选取对于 KNN 算法来说，变得非常关键，经过重复分析和不断调试，本文的 K 值最终选取为 15。

5.3.2.3 欧式距离

KNN 算法中关于测量距离的方法通常来说，有欧式距离、曼哈顿距离、闵可夫斯基距离这三种。

（一）欧式距离（Eucliden Distance）：

令 $i=(x_1,x_2,...,x_p)$ 和 $j=(y_1,y_2,...,y_p)$ 是两个被 p 个数值属性标记的对象，则对象 i 和 j 之间的欧式距离定义为：

$$dis(i,j)=\sqrt{\sum_{k=1}^{p}(x_k-y_k)^2} \tag{5-6}$$

（二）曼哈顿距离：

$$d(x,y)=\sum_{k=1}^{n}|x_k-y_k| \tag{5-7}$$

（三）闵可夫斯基距离：

$$D(x,y)=\sqrt[p]{(|x_1-y_1|)^p+(|x_2-y_2|)^p+\cdots+(|x_n-y_n|)^p}=\sqrt[p]{\sum_{i=1}^{n}(|x_i-y_i|)^p} \tag{5-8}$$

通过上述三个公式的对比分析，我恩们可以发现闵可夫斯基距离即欧氏距离在 p=2 时的特例，当 p=1，欧式距离的特例即曼哈顿距离。

本文采用最简便常见的欧式距离来计算文档中的距离度量。

5.3.2.4 分类决策规则

本文的分类决策规则选用的是多数表决法。也就是说，如果一个样本在特征空间中的 k 个特征空间中最邻近的样本的大多数属于某一个类别，那么这个样本也属于这个类别。

多数表决规则的表述：如果分类的损失函数为 0-1 损失函数，分类函数为：$f:R^n\rightarrow\{c_1,c_2,\cdots,c_k\}$

那么误分类的概率是：

$$P(Y1=f(X))=1-(Y=f(X)) \tag{5-9}$$

对给定的实例x，其最邻近的k个训练实例点构成集合 $N_k(x)$。如果涵盖 $N_k(x)$ 的区域的类别是 c_j，那么误分类率是：

$$\frac{1}{k}\sum_{x_i\in N_k(x)}I(y_i1=c_j)=1-\frac{1}{k}\sum_{x_i\in N_k(x)}I(y_i=c_j) \tag{5-10}$$

要使得误分类率最小，也就是说经验风险最小，那么就要使 $\frac{1}{k}\sum_{x_i\in N_k(x)}I(y_i=c_j)$ 最大，

所以多数表决规则也可以等价理解为经验风险最小化。

5.3.3 完整性评价关键词分值表

基于 5.3.1.完整性定义的六个标准，我们对留言答复意见进行了分析，得出每条留言中符合标准的项数，并根据符合标准的项数制定了评价回复完整性的四个评价关键词：好，较好，一般，差。

具体的评价标准如下表 5-2：

表 5-2 完整性评价关键词分值表

符合条数	6 条	4-5 条	2-3 条	0-1 条
关键词	好	较好	一般	差
分值（分）	10	8	5	3

5.3.4 完整性代码实现结果

根据 5.3.1 完整性定义得到以下结果：

	A	B	C	D	E	F	G	H
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	完整性
2	2549	A00045581	A2区景蓉	2019/4/25 9:32:09	2019年4月	现将网友	2019/5/10	好
3	2554	A00023583	A3区潇楚	2019/4/24 16:03:40	潇楚南路	网友“A0	2019/5/9	好
4	2555	A00031618	请加快提	2019/4/24 15:40:04	地处省会	市民同志：	2019/5/9	好
5	2557	A00011073	在A市买公	2019/4/24 15:07:30	尊敬的书	网友“A0	2019/5/9	好
6	2574	A0009233	关于A市公	2019/4/23 17:03:19	建议将“	网友“A0	2019/5/9	好
7	2759	A00077538	A3区含浦	2019-04-08 08:37:20	欢迎领导	网友“A0	2019/5/9	好
8	2849	A00010080	A3区教师	2019/3/29 11:53:23	尊敬的胡	网友“A0	2019/5/9	好
9	33970	A00010024	B7县二中	2018/8/12 10:56:10	B7县二中	网友：您	2018/8/20	好
10	33978	A00044584	B市601小	2018/8/8 13:15:50	601小区钻	网民您好：	2018/8/17	好
11	33984	A00091054	咨询在B市	2018/8/3 21:26:53	您好，由	网民您好：	2018/8/7	差
12	34239	A00084182	B市泰民米	2018/2/3 16:27:45	我们	尊敬的网	2018/4/16	一般
13	34249	A00057771	B市港口街	2018/1/25 12:01:13	B市建	尊敬的网	2018/4/12	一般
14	34252	A00014375	B4区栗雨	2018/1/24 11:12:25	B市B4	亲爱的网	2018/4/12	一般
15	35462	A00050959	请问B市西	2019/12/2 19:34:28	请问西环	尊敬的网	2019/12/1	一般
16	35467	A00044412	请求农行E	2019/11/28 15:41:22	“大行德	尊敬的网	2019/12/2	好
17	35479	A00079768	B2区泉中	2019/11/19 9:32:36	敬爱的阳	尊敬的网	2019/12/2	一般
18	35492	A00057180	B市三一歌	2019/11/5 21:51:41	三一歌雅	尊敬的网	2019/11/1	一般
19	35798	A00010594	咨询B市公	2018/12/5 23:49:06	市长您好，	尊敬的网	2018/12/2	好
20	35801	A00010143	B4区金锦	2018/12/4 10:09:51	我是所住	尊敬的网	2018/12/2	一般
21	35812	A00037449	能否在上	2018/11/21 14:43:59	针对现状，	尊敬的网	2018/11/2	一般
22	35818	A00098677	B市男职工	2018/11/9 14:58:05	在公司工	尊敬的网	2018/11/1	一般

图 5-3 完整性评价结果截图（部分）

5.4 可解释性

5.4.1 可解释性定义

答复意见的可解释性，即答复意见中是否对引用的法律条文进行解释，即白话化。通过对答复意见可解释性的分析，我们可以判断普通群众是否能够得到通俗易懂的答复意见。本文将采用 LDA 模型对答复意见的可解释性进行评价。

5.4.2 LDA 模型

5.4.2.1 LDA 概念

潜在狄利克雷分布 LDA 模型(Latent Dirichlet Allocation)

LDA 模型是一种文档生成的模型，也是一种非监督机器学习模型。它认为一篇文档有多个主题，而其中每个主题又对应着不同的词，构造一篇文档的过程，以一定的概率选择某个主题，再在这个主题下以一定的概率选取某一个词，这样就生成了这篇文档的第一个词。通过不断重复地生成，就构造出了整篇文档。

LDA 模型运用的过程就是上述文档生成的逆过程，就是说根据一篇得到的文档，去寻找这篇文档的主题，以及这个主题所对应的词。LDA 模型主要用于文本的分类，因此本文将采用 LDA 模型对答复意见的可解释性进行评价。

5.4.2.2 LDA 生成流程

对于语料库中的每篇文档，LDA 模型定义了如下生成过程（generativeprocess）：

（1） 对每一篇文档，从主题分布中抽取一个主题；

（2） 从被抽到的主题所对应的词的分布中抽取一个词；

（3） 重复上述过程直至遍历文档中的每一个词。

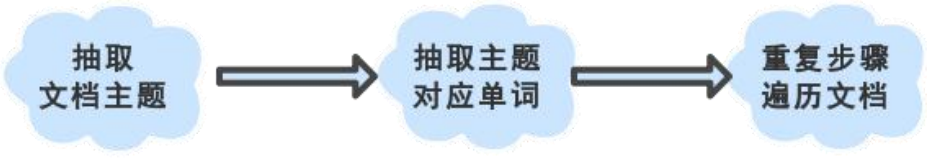


图 5-4 LDA 模型生成流程图

语料库中的每一篇文档与 T 个主题的一个多项分布相对应，将这个多项分布记为θ。每个主题又与词汇表中的 V 个单词的一个多项分布相对应，将这个多项分布记为 ϕ 。

5.4.2.3 LDA 整体流程

定义文档集合 D，主题集合 T。D 中每个文档 d 看做一个单词序列 $(w_1, w_2, .., w_n), w_i$ 表示第 i 个单词, d 有 n 个单词（词袋）。

D 中涉及的所有不同单词构成一个集合词汇，LDA 以文档集合 D 作为输入，训练出的两个结果向量（设聚成 k 个主题，词汇中共包含 m 个词）。

对每个 D 中的文档 d，对应到不同主题的概率 $\theta d(p_{t_1}, p_{t_2}, ..., p_{t_k})$，其中，$p_{t_i}$ 表示 d 对应 T 中第 i 个主题的概率。计算方法为：

$$p_{t_i} = \frac{n_{t_i}}{n} \tag{5-11}$$

其中 n_{t_i} 表示 d 中对应第 i 个主题的词的数目，n 是 d 中所有词的总数。

对每个 T 中的主题，生成不同单词的概率 $\phi t(p_{w_1}, p_{w_2}, .., p_{w_m})$，其中，$p_{w_i}$ 表示 t 生成词汇中第 i 个单词的概率。计算方法为：

$$p_{w_i} = \frac{N_{w_i}}{N} \tag{5-12}$$

其中 N_{w_i} 表示对应到主题的词汇中第 i 个单词的数目，N 表示所有对应到主题的单词总数。

由此我们得出 LDA 模型的计算公式：

$$p(w|d) = p(w|t) \times p(t|d) \tag{5-13}$$

5.4.3 可解释性评价关键词分值表

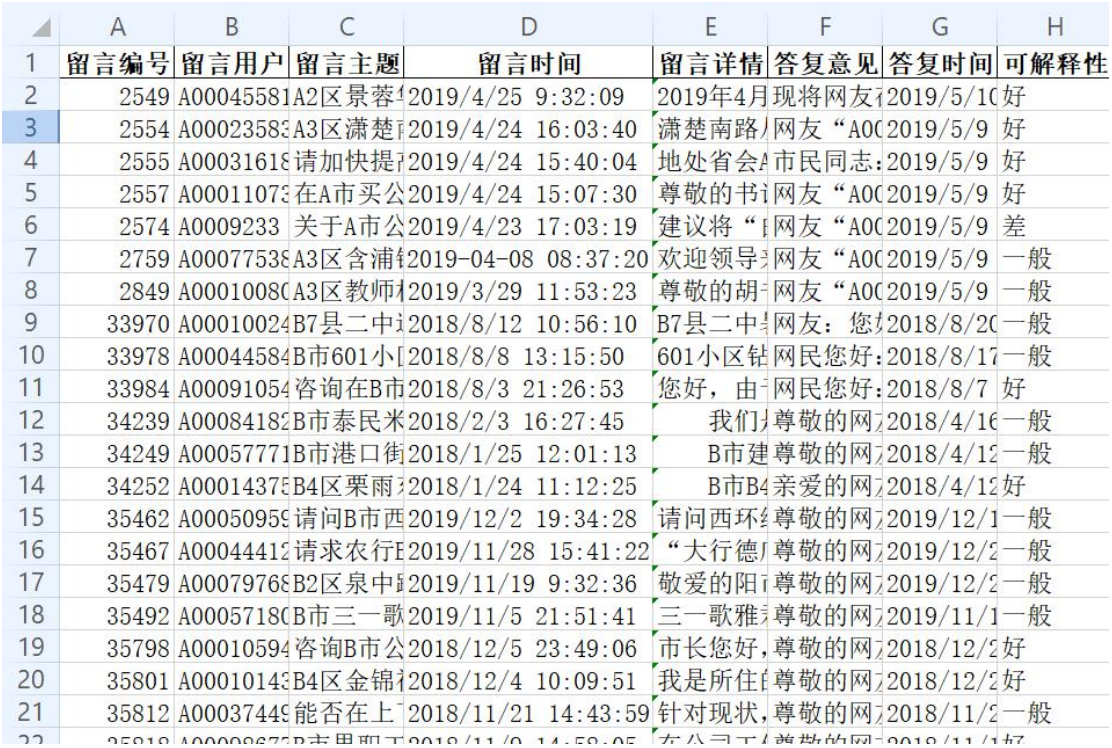
基于 5.4.1.可解释性的定义，我们对回复与群众留言之间的间隔时间进行了分析，并制定了评价回复效率（快慢）的四个评价关键词：好，较好，一般，差。具体的评价标准如下表 5-3：

表 5-3 可解释性评价关键词分值表

关键词	好	较好	一般	差
分值（分）	10	8	5	3

5.4.4 可解释性代码实现结果

根据 5.4.1 可解释性定义得到以下结果：



	A	B	C	D	E	F	G	H
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	可解释性
2	2549	A00045581	A2区景蓉	2019/4/25 9:32:09	2019年4月	现将网友	2019/5/1(	好
3	2554	A00023583	A3区潇楚	2019/4/24 16:03:40	潇楚南路	网友“A0(	2019/5/9	好
4	2555	A00031618	请加快提	2019/4/24 15:40:04	地处省会	市民同志：	2019/5/9	好
5	2557	A00011073	在A市买公	2019/4/24 15:07:30	尊敬的书	网友“A0(	2019/5/9	好
6	2574	A0009233	关于A市公	2019/4/23 17:03:19	建议将“	网友“A0(	2019/5/9	差
7	2759	A00077538	A3区含浦	2019-04-08 08:37:20	欢迎领导	网友“A0(	2019/5/9	一般
8	2849	A00010080	A3区教师	2019/3/29 11:53:23	尊敬的胡	网友“A0(	2019/5/9	一般
9	33970	A00010024	B7县二中	2018/8/12 10:56:10	B7县二中	网友：您	2018/8/20	一般
10	33978	A00044584	B市601小	2018/8/8 13:15:50	601小区钻	网民您好：	2018/8/17	一般
11	33984	A00091054	咨询在B市	2018/8/3 21:26:53	您好，由	网民您好：	2018/8/7	好
12	34239	A00084182	B市泰民米	2018/2/3 16:27:45	我们	尊敬的网	2018/4/16	一般
13	34249	A00057771	B市港口街	2018/1/25 12:01:13	B市建	尊敬的网	2018/4/12	一般
14	34252	A00014375	B4区栗雨	2018/1/24 11:12:25	B市B4	亲爱的网	2018/4/12	好
15	35462	A00050959	请问B市西	2019/12/2 19:34:28	请问西环	尊敬的网	2019/12/1	一般
16	35467	A00044412	请求农行E	2019/11/28 15:41:22	“大行德	尊敬的网	2019/12/2	一般
17	35479	A00079768	B2区泉中	2019/11/19 9:32:36	敬爱的阳	尊敬的网	2019/12/2	一般
18	35492	A00057180	B市三一歌	2019/11/5 21:51:41	三一歌雅	尊敬的网	2019/11/1	一般
19	35798	A00010594	咨询B市公	2018/12/5 23:49:06	市长您好，	尊敬的网	2018/12/2	好
20	35801	A00010143	B4区金锦	2018/12/4 10:09:51	我是所住	尊敬的网	2018/12/2	好
21	35812	A00037449	能否在上	2018/11/21 14:43:59	针对现状，	尊敬的网	2018/11/2	一般

图 5-5 可解释性评价结果截图（部分）

5.5 及时性

5.5.1 及时性定义

为更加精准地对政务部门工作人员给出的留言回复质量进行评价，我们引入了第四个自定义的评价关键词——及时性。现将及时性定义如下：

（1）留言回复时间与群众留言时间的时间间隔在 10 天以内的，我们称之为答复群众及时。

（2）留言回复时间与群众留言时间的时间间隔在 20 天以上的，我们称之为答复群众不及时。

（3）同类问题反馈我们依据群众留言时间的先后，对比回复时间的先后，

筛选出部分因某种原因（人工判断误差、程序出错等）导致的回复延后，这样的延后我们定义在不及时的范围内。

根据上述定义我们继续进行对答复意见的情感分析。

5.5.2 及时性评价关键词分值表

基于 5.5.1.及时性的定义，我们对回复与群众留言之间的间隔时间（天数）进行了分析，并制定了评价回复效率（快慢）的四个评价关键词：好，较好，一般，差。

具体的评价标准如下表 5-4：

表 5-4 及时性评价关键词分值表

时间间隔	5 天内	5-10 天	11-20 天	20 天以上
关键词	好	较好	一般	差
分值（分）	10	8	5	3

5.5.3 及时性代码实现结果

根据 5.5.1 及时性定义得到以下结果：

	A	B	C	D	E	F	G	H
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	及时性
2	2549	A00045581	A2区景蓉	2019/4/25 9:32:09	2019年4月	现将网友	2019/5/10	一般
3	2554	A00023583	A3区潇楚	2019/4/24 16:03:40	潇楚南路	网友“A0	2019/5/9	一般
4	2555	A00031618	请加快提	2019/4/24 15:40:04	地处省会	市民同志：	2019/5/9	一般
5	2557	A00011073	在A市买公	2019/4/24 15:07:30	尊敬的书	网友“A0	2019/5/9	一般
6	2574	A0009233	关于A市公	2019/4/23 17:03:19	建议将“	网友“A0	2019/5/9	一般
7	2759	A00077538	A3区含浦	2019-04-08 08:37:20	欢迎领导	网友“A0	2019/5/9	差
8	2849	A00010080	A3区教师	2019/3/29 11:53:23	尊敬的胡	网友“A0	2019/5/9	差
9	33970	A00010024	B7县二中	2018/8/12 10:56:10	B7县二中	网友：您	2018/8/20	较好
10	33978	A00044584	B市601小	2018/8/8 13:15:50	601小区钻	网民您好：	2018/8/17	较好
11	33984	A00091054	咨询在B市	2018/8/3 21:26:53	您好，由	网民您好：	2018/8/7	好
12	34239	A00084182	B市泰民米	2018/2/3 16:27:45	我们	尊敬的网	2018/4/16	差
13	34249	A00057771	B市港口街	2018/1/25 12:01:13	B市建	尊敬的网	2018/4/12	差
14	34252	A00014375	B4区栗雨	2018/1/24 11:12:25	B市B4	亲爱的网	2018/4/12	差
15	35462	A00050959	请问B市西	2019/12/2 19:34:28	请问西环	尊敬的网	2019/12/1	较好
16	35467	A00044412	请求农行E	2019/11/28 15:41:22	“大行德	尊敬的网	2019/12/2	好
17	35479	A00079768	B2区泉中	2019/11/19 9:32:36	敬爱的阳	尊敬的网	2019/12/2	一般
18	35492	A00057180	B市三一歌	2019/11/5 21:51:41	三一歌雅	尊敬的网	2019/11/1	较好
19	35798	A00010594	咨询B市公	2018/12/5 23:49:06	市长您好，	尊敬的网	2018/12/2	一般
20	35801	A00010143	B4区金锦	2018/12/4 10:09:51	我是所住	尊敬的网	2018/12/2	一般
21	35812	A00037449	能否在上	2018/11/21 14:43:59	针对现状，	尊敬的网	2018/11/2	好
22	35818	A00098677	B市男职工	2018/11/9 14:58:05	在公司工	尊敬的网	2018/11/1	好

图 5-6 及时性评价结果截图（部分）

5.6 留言回复评价分值表

综合以上四个评价关键词（相关性、完整性、可解释性和及时性），我们给出了对留言答复意见质量的完整评分表，如下表 5-5：

表 5-5 答复意见质量评分表

关键词（分）	好（10）	较好（8）	一般（5）	差（3）
相关性				
完整性				
可解释性				
及时性				
附：打分表相应项勾选出来即可				

最终得分的计算公式本文自定义为：

$$总分 = 相关性 \times 0.3 + 完整性 \times 0.3 + 可解释性 \times 0.3 + 及时性 \times 0.1 \tag{5-14}$$

6 结论

本节将对本次比赛做一个小总结。我们选择的题目即对“智慧政务”留言平台的文本数据进行文本挖掘分析。整个解题过程中，我们主要是运用 TFIDF 权重法提取关键词，构建词频矩阵，同时运用了大量的文本相似度计算算法对留言进行分类。本次文本挖掘项目中，我们自定义了部分概念性质，如相关性、完整性、可解释性和及时性等。通过自定义部分概念，我们能够更好的运用代码实现对文本信息的分析与挖掘，同时也能使读者更好的了解相关概念并获得一定的启发。

本文主要分析内容：

问题一： 一级分类标签，对分类方法进行评价；

问题二： 热点问题（热度指数）排序，热点问题留言明细（主要成果为附

件中的表格"热点问题表.xlxs"、"热点问题留言明细表.xlxs")

问题三: 从四个评价关键词（相关性、完整性、可解释性和及时性）分别对留言答复意见质量进行分析评价，最终根据"表 5-5 留言答复意见质量评分表"和 5.6 留言回复评价分值表中自定义的答复意见质量评分公式（5-14）得出最终质量评分，并给出评价和参考性建议。

通过以上分析，我们对“智慧政务”平台有着以下几点建议：

（一）聘请相关技术人才对“智慧政务”平台作出改进，如：增加自动分类机制以减轻政务部门工作人员的劳动量或改进留言入口，实现分部门留言（可行性较前者较低）。

（二）集中定期对给出群众留言答复的岗位人员进行学习培训，提高其工作素养。

（三）加大对普通民众的普法宣传。

（四）相关部门（如城管部门、监察部门等）应加强对城市民生的巡查，多下基层了解普通民众的日常生活，做到防患于未然而非“亡羊补牢”。

（五）改进留言处理平台，在工作人员处理界面优先显示未答复的留言时间较早的群众留言。

在本次解题过程中，我们经历了在离比赛结束还剩不到两周的中途更换了编程队员，重新开始代码的实现。由于部分未预料因素，我们的问题一分类模型准确率并没有取得预期的效果（预期为准确率 85%，实际只有 82%左右），经过反思我们认为在算法的选择和部分阈值的选取上有待改进。这也让我们深深地体会到了在文本挖掘知识的储备方面还太过浅显，因此，在后续的学习生活中，我们将会继续深入了解文本分析挖掘的有关知识，对文本挖掘问题有更深刻的思考。最后在此感谢在竞赛期间给予我们帮助的老师和主办方，感谢你们无私的帮助和提供的免费学习资源！鸣谢！

7 参考文献

[1] 吴兴蛟 .K-Means 聚类算法研究与改进[D]:[硕士学位论文].昆明：昆明理工大学,2003.

[2] 刘峰,蔡志杰,乐斌.基于市场资金流向分析的商品期货量化交易策略[J].数学建模及其应用,2017,5(1):3-12.

[3] 帕提.胡赛因.哈萨克文信息检索停用词表的统计方法[J].电脑知识与技术,2013,8(1):6-8.

[4] 刘惠,赵海清.以数据挖掘为导向的应用型统计人才培养的思考[J].科技视界,2019,2(1):4-8

[5] 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.

[6] 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.

[7] 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.
[8] 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.
[9] 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).
[10] 许洁.基于大间隔最近邻的度量学习算法研究[D]:[硕士学位论文].西安：西安电子科技大学,2019.
[11] 邬启为。基于向量空间的文本聚类方法与实现[D].北京：北京交通大学，2014.
[12] 张振亚,王进，程红梅,等.基于余弦相似度的文本空间索引方法研究[J].计算机科学，2005,32(9):160-163.
[13] 蔡珺哲.基于文本挖掘的中国社交平台用户情感分析[D].上海师范大学，2014.
[14] 约尔尼萨·吾不力卡司木，玉素甫·艾白都拉.基于最小编辑距离和词汇库的维吾尔语文本校对系统的设计与算法实现[J].信息与电脑(理论版),2013,8(6):78-82.
[15] 战扬,金英,杨丰.基于监督的距离度量学习方法研究[J].信息技术,2011,5(1），34-41.