“智慧政务”中的文本挖掘应用

摘要

在我国，随着各大运营平台及时事新闻传播渠道的出现，各类社情民意相关的文本数据量不断攀升，给进行留言划分和热点整理的相关部门的工作带来了极大挑战。相比传统的人工整理文本数据，机器学习算法能从大量的文本数据中快速、精准地抽取有价值的信息和知识，对提升政府的管理水平和施政效率具有极大的推动作用。因此，找到一种令人信服的算法和模型，能较好的普适的解决文本数据相关问题具有重大的意义。

针对问题一，首先将附件2的数据(列'留言主题'和列'留言详情')进行中文分词及停用词过滤等数据预处理，然后基于TFIDF算法提取特征词，形成词袋，构成词向量矩阵。接下来使用KNN、决策树、贝叶斯分类器对列'留言主题'生成的词向量矩阵进行训练，再使用逻辑回归算法对列'留言详情'生成的词向量矩阵(此处的词向量矩阵由word2vec算法生成)进行训练，对比四类模型的分类效果，使用F1值进行评价，最后选用分类效果更佳的逻辑回归算法。

针对问题二,首先将附件3的数据(列'留言主题')进行中文分词及停用词过滤等数据预处理，然后基于TFIDF算法提取特征词。然后使用余弦距离计算词向量两两之间的相似度，设定阈值，当相似度＞阈值时，将两文本归为同一类。最后利用每个类别中的文本数作为热度指标，输出排名前5的热点问题。

针对问题三，本文对答复评价模型进行了分析，从相关性，完整性，可解释性和信息量来构造这个模型。首先运用向量化算法doc2vec计算附件4中的数据(列'留言详情'和列'答复意见')的相似度作为答复的相关性的评价指标；而信息量通过文本字数来定义，可解释性我们定义为可读性，并借鉴了英文的可读性定义来确定这个值。完整性则通过答复中是否有数值数据来确定。得到这四个值确定一个回归模型。接着我们通过专家打分法确定初始权重系数，再通过灰色关联分析引入关联度，再对初始权重系数进行更改得到最终权重系数。

关键词

KNN算法　Logistic回归　朴素贝叶斯算法　doc2vec　word2vec　TF-IDF算法

目录

1.挖掘目标

本次的建模目标针对来自收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，在对文本进行基本的机器预处理、中文分词和停用词过滤后，通过 KNN 算法、TF-IDF 算法及 doc2vec 等多种数据挖掘模型，实现对文本数据的倾向性判断及所隐藏的信息的挖掘并分析，以期望得到有价值的内在内容。

2.分析步骤

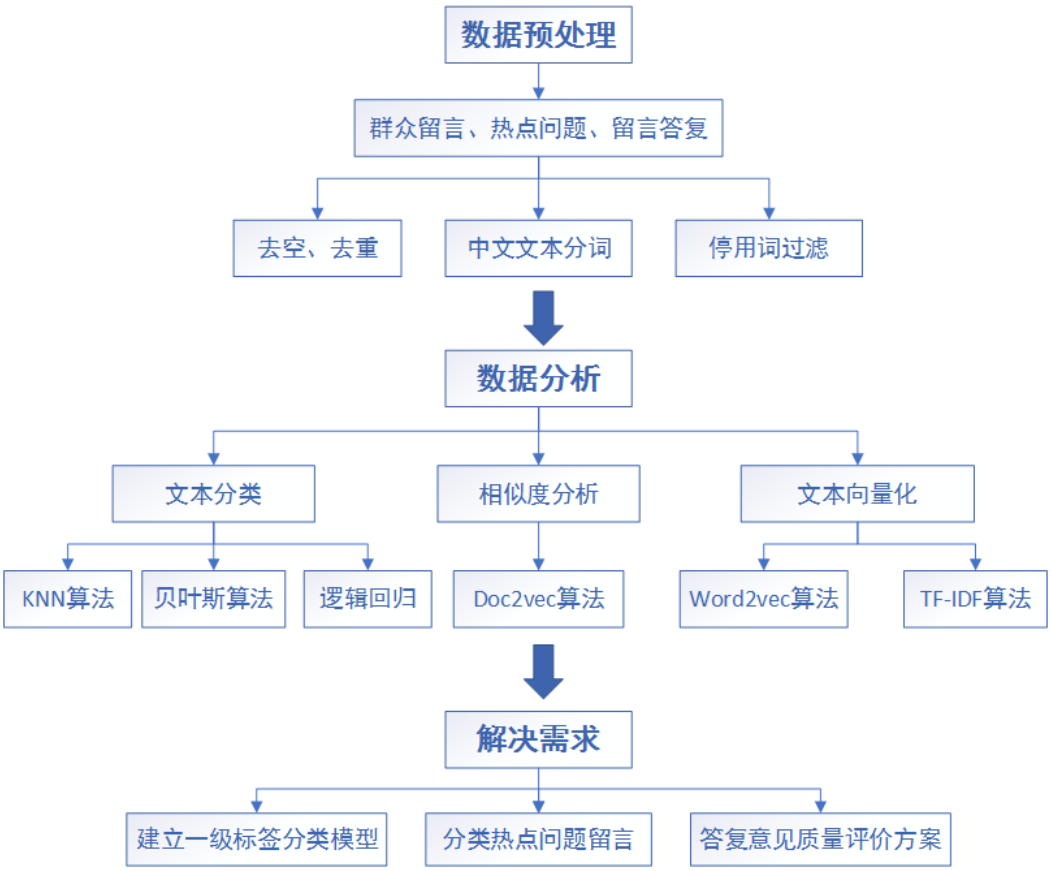


图 1 本文分析步骤

步骤 1：数据预处理。对附件 2、附件 3 和附件 4 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进行数值化处理，对其进行去除重复项及空行、中文文本分词、停用词；过滤，以便后续分析；

步骤 2：数据分析。通过 word2vec 和 TFIDF 将文本向量化，得到向量矩阵，去除同义词影响，简化计算；通过 KNN 算法、贝叶斯算法及逻辑回归对数据进

行分析分类，得到准确率及回召率，用加权平均法获得 F 值，比较 F 值，综合得出最适合的分类模型；最后，通过 doc2vec 算法获取文本数据之间的相关度。

步骤 3：解决需求。根据以上算法得出的数据结果 ，建立一级标签分类模型、挖掘热点问题并提出一套合适的评价方案，解决政府所需，改善管理水平，提升施政效率。

3.分析方法与过程

3.1 问题一“群众留言分类”分析方法与过程

3.1.1 数据预处理

针对题目中所给的原始数据而言，原始数据中存在着大量的冗余信息、不一致信息、 非结构化数据和有异常的数据，这将会严重的影响数据挖掘和建模的执行效率，甚至可 能会导致数据挖掘的结果出现偏差，因此，必须先对给定的原始数据进行预处理操作， 提高数据集的质量，并使得数据能够更好地适应我们的挖掘算法和数学模型

1) “去重”、“去空”

对于存储了全部网络商业评论的 txt 文件，每行代表了一个评论文本但是难免会出现两个完全一样的文本和一些空行。所以本文首先进行了“去重”、“去空”的预处理工作。在导入评论文本时，同时进行了是否为空的判断，只导入不为空的文本，从而过滤掉了空白文本。

2)中文分词

分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂得多、困难得多。主要分词结果如下所示：

>>>text = 'A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这

条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。'

>>>jieba.lcut(text)['A3', '区', '大道', '西行', '便', '道', '，', '未管', '所', '路口', '至', '加油站', '路段', '，', '人行道', '包括', '路灯', '杆', '，', '被', '圈', '西湖', '建筑', '集团', '燕子', '山', '安置', '房', '项目', '施工', '围墙', '内', '。', '每天', '尤其', '上下班', '期间', '这', '条', '路上', '人流', '车流', '极', '多', '，', '安全隐患', '非常', '大', '。', '强烈', '请求', '文明城市', 'A', '市', '，', '尽快', '整改', '这个', '极', '不', '文明', '的', '路段', '。']

3)停用词过滤

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词），停用词具备两个特征，一是极其普遍，出现频率高；而是包含信息量低，对文本标识无意义。为了找出这些停用词，需要一些标准估计此的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看做是噪声词，可以通过词频（TF）和文档频数（DF）指标去衡量词语的有效性。

3.1.2 文本向量化

3.1.2.1word2vec 算法

word2vec 是 Google 在 2003 年开源的一款将词表征为实数值向量的高效算法，采用的模型有 CBOW【Continuous Bag-Of-Words 连续的词袋模型】和 Skip-Gram 两种。

word2vec 通过训练，可以把文本内容的处理简化为 k 维向量空间中的向量运算，二向量空间上的相似度可以用来表示文本语义上的相似度。

3.1.2.2TF-IDF 算法

TF-IDF（term frequency–inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用

加权技术。

TF-IDF 是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词

（1）TF 是词频(Term Frequency)

词频表示词条（关键字）在文本中出现的频率。这个数字通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。公式：

，即：

其中是该词在文件中出现的次数，则是文件中所有词汇出现的次数总和。

（2）IDF 是逆向文件频率(Inverse Document Frequency)

某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。公式：

其中是语料库中的文件总数。表示包含词语的文件数目。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用 1+。

。

（3）TF-IDF 实际上是：TF * IDF

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。公式：

3.1.3 分类模型研究

3.1.3.1 朴素贝叶斯法（Naive Bayes）

是基于贝叶斯定理与特征条件独立假设的分类方法。贝叶斯算法起源于古典

数学理论，是一种分类算法的总称。它以贝叶斯定理为基础，假设某待分类的样本满足某种概率分布，并且可以根据已观察到的样本数据对该样本进行概率计算，以得出最优的分类决策。通过计算已观察到的样本数据估计某待分类样本的先验概率，利用贝叶斯公式计算出其后验概率，即该样本属于某一类的概率，选择具有最大后验概率的类作为该样本所属的类。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入，利用贝叶斯定理求出后验概率最大的输出。

表 1 朴素贝叶斯法（Naive Bayes）优缺点

优点	缺点
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，尤其是数据量超出内存时，我们可以一批批的去增量训练。	理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型给定输出类别的情况下,假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。
朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。	需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。
对缺失数据不太敏感，算法也比较简单，常用于文本分类。	对输入数据的表达形式很敏感。
	由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

3.1.3.2 逻辑回归

逻辑回归的假设函数为：

其中为样本输入，为模型输出，为要求解的模型参数。设 0.5 为临界值，当时，即>0 时，为 1；时，即<0 时，为 0；模型输出值在区间取值，因此可以从概

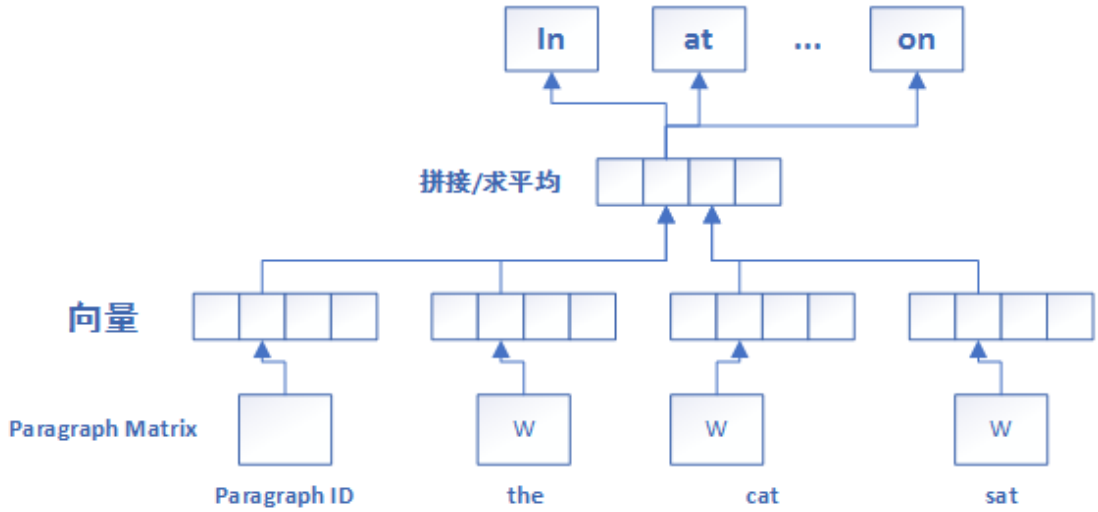
率角度进行解释：越接近 0，则分类为 0 的概率越高；越接近 1，则分类为 1 的概率越高；越接近临界值 0.5，则无法判断，分类准确率会下降。

3.1.3.2 doc2vec

Word2vec 技术也用于计算句子或者其他长文本间的相似度，其一般做法是对文本分词后，提取关键词，用词向量表示这些关键词，接着对关键词向量求平均或者将其拼接，最后利用词向量计算文本间的相似度。这种方法丢失了文本中的语序信息，而文本的语序包含重要信息。为了充分利用文本语序信息，有研究者在 Word2vec 的基础上提出了文本向量化 doc2vec，doc2vec 技术存在两种模型——Distributed Memory（DM）和 Distributed Bag Of Words（DBOW），分别对应 word2vec 技术里的 CBOW 和 Skip-gram 模型。

DM 模型增加了一个与词向量长度相等的段向量，也就是和 DM 模型结合词向量和段向量预测目标词的概率分布，如图 所示。在训练的过程中，DM 模型增加了一个 paragraph ID，和普通的 word2vec 一样，paragraph ID 也是先映射成一个向量，即 paragraph vector。Paragraph vector 与 word vector 的维数虽然一样，但是代表两个不同的向量空间。在之后的计算里，paragraph vector 与 word vector 累加或者连接起来，将其输入 softmax 层。在一个句子或者文档的训练过程中，paragraph ID 保持不变，共享着同一个 paragraph vector，相当于每次在预测单词的概率时，都利用了整个句子的语义。在预测阶段，给待遇测的句子新分配一个 paragraph ID，词向量和输出层 softmax 的参数保持训练阶段得到的参数不变，重新利用随机梯度算法训练待遇测的句子。待误差收敛后，即得到待遇测句子的 paragraph vector。

图 2 DM-模型示意图



DM 模型通过段落向量和词向量相结合的方式预测目标词的概率分布，即 DBOW 模型的输入只有段落向量，具体如图 3 所示。DBOW 模型通过一个段落向量预测段落中某个随机词的概率分布。doc2vec 不仅提取了文本的语义信息，而且提取了文本的语序信息，在一般文本处理任务中，会将词向量和段向量相结合使用以期获得更好的效果。

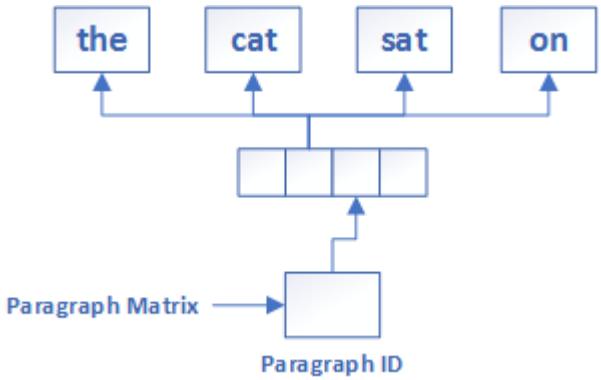


图 3 Paragraph ID 模型示意图

3.1.3.3KNN 算法

KNN 的全称是 K Nearest Neighbors，它的原理是当预测一个新的值 x 的时候，根据它距离最近的 K 个点是什么类别来判断 x 属于哪个类别。主要由两部分组成： K 值的选取和点距离的计算。KNN 是一种非参的，惰性的算法模型。非参的意思意味着这个模型不会对数据做出任何的假设，与之相对的是线性回归（我们总会假设线性回归是一条直线）。也就是说 KNN 建立的模型结构是根据数据来决定的，这也比较符合现实的情况，毕竟在现实中的情况往往与理论上的

假设是不相符的。而惰性主要体现在逻辑回归需要先对数据进行大量训练，最后才会得到一个算法模型。而 KNN 算法却不需要，它没有明确的训练数据的过程，或者说这个过程很快。主要体现在两个方面：距离计算及 k 值选取。

1.距离计算

要度量空间中点距离的话，有好几种度量方式，比如常见的曼哈顿距离计算，欧式距离计算等等。不过通常 KNN 算法中使用的是欧式距离，这里只是简单说一下，拿二维平面为例，二维空间两个点的欧式距离计算公式如下：

拓展到多维空间，则公式是：

这样我们就明白了如何计算距离，KNN 算法就是将预测点与所有点距离进行计算，然后保存并排序，选出前面 K 个值看看哪些类别比较多。但其实也可以通过一些数据结构来辅助，比如最大堆等。如图 4 所示：

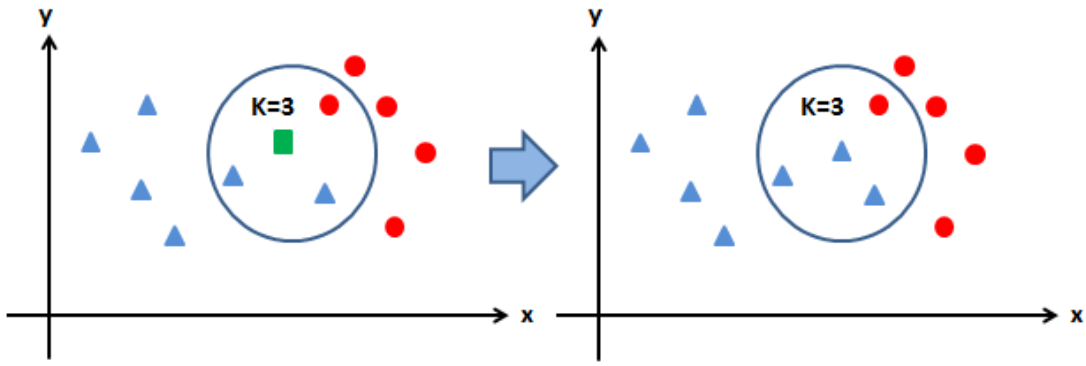


图 4 K 值选取示意图

2. k 值选取

通过上面那张图我们知道 K 的取值比较重要，那么该如何确定 K 取多少值好呢？答案是通过交叉验证（将样本数据按照一定比例，拆分出训练用的数据和验证用的数据，比如 6：4 拆分出部分训练数据和验证数据），从选取一个较小的 K 值开始，不断增加 K 的值，然后计算验证集合的方差，最终找到一个比较合适的 K 值。

表 2 KNN 算法优缺点

优点	缺点
简单易用，简洁明了	对内存要求较高,因为该算法存储了所有
模型训练时间快	训练数据
模型训练时间快	预测阶段可能很慢
对异常值不敏感	对不相关的功能和数据规模敏感

3.1.4 选取最优模型运算

3.1.4.1 不同分类器的分类效果

通过运算得到四种分类器的分类时间以及对应的 F1 值，由此我们可以看出，通过逻辑回归进行分类所得到的 F1 值更高，因此我们采用逻辑回归法对文本数据进行分类。

如下表 3 所示：

表 3 不同分类器分类时间以及对应的 F1 值

分类器	分类时间	F1 值
贝叶斯	11.8399923620223999s	0.636722979
KNN	1320.1659944057465s	0.701689534
决策树	58.92043137550354s	0.636722979
逻辑回归	15.006829753979423s	0.791022641

3.1.4.2F-Score 分类模型的评估方法

精确率（Presion）和召回率（Recall）评估指标，理想情况下做到两个指标都高当然最好，一般情况下精确率高，召回率就低，召回率高，精确率就低。

所以在实际中常常需要根据具体情况作出取舍，例如在保证召回率情况下，尽量提升精准率。而像癌症检测、地震监测、金融欺诈，则在保证精确率的条件下，尽量提升召回率，因此引出一个新的指标 F-Score，综合考虑精确率和召回

率的调和值：

题目中所求的是 F1，此时，对应的公式为

3.2 问题二“热点问题挖掘”分析方法与过程

3.2.1 分析过程

首先将附件 3 中“留言主题”列导入 Python 中，通过运用数据预处理中的“中文分词”以及“去停用词”，将该列数据向量化，生成 TF-IDF 向量化矩阵。通过利用每行留言主题的词向量去计算每个词向量之间的相似度，如果相似度>阈值，则将他们归入同一类。以该相似度作为热度评价的标准

3.2.2 模型建立与求解

3.2.2.1 TF-IDF 词向量矩阵生成

将“留言主题”列生成词向量矩阵，结果如下所示：

array([[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.],
...,
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.]])

3.2.2.2 余弦相似度

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90°时，余弦相似度的

值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是与向量的长度无关的，仅仅与向量的指向方向相关。余弦相似度通常用于正空间，因此给出的值为-1 到 1 之间。2

两个向量间的余弦值可以通过使用欧几里得点积公式求出：

给定两个属性向量，A 和 B，其余弦相似性θ由点积和向量长度给出，如下所示：

这里的和分别代表向量 A 和 B 的各分量。

给出的相似性范围从-1 到 1：-1 意味着两个向量指向的方向正好截然相反，1 表示它们的指向是完全相同的，0 通常表示它们之间是独立的，而在这之间的值则表示中间的相似性或相异性。

对于文本匹配，属性向量 A 和 B 通常是文档中的词频向量。余弦相似性，可以被看作是在比较过程中把文件长度正规化的方法。

在信息检索的情况下，由于一个词的频率（TF-IDF 权）不能为负数，所以这两个文档的余弦相似性范围从 0 到 1。并且，两个词的频率向量之间的角度不能大于 90°。

以下列两个“留言主题”为例：

句子 A：A 市万家丽南路丽发新城居民区附近搅拌站扰民。

句子 B：投诉 A2 区丽发新城附近建搅拌站噪音扰民。

步骤一：分词、去停用词：

句子 A：万家 南路 丽发 新城 居民区 附近 搅拌站 扰民

句子 B：投诉 A2 区丽发 新城 附近 搅拌站 噪音 扰民

步骤二：使用 TF-IDF 得到词向量：

句子 A：

0.0	40123733358447	40740123733358447
0.0	40123733358447	0.0

0.40740123733358447	0.28986933576883284	0.0
0.28986933576883284	0.28986933576883284	0.28986933576883284

句子 B：

0.40740123733358447	0.0	0.0
0.40740123733358447	0.0	0.40740123733358447
0.0	0.28986933576883284	0.40740123733358447
0.28986933576883284	0.28986933576883284	0.28986933576883284

步骤三：计算两个向量的余弦值来确定他们的相似度：

```
>>> print(cos_sim(weight[0],weight[1]))
0.6680484636381288
```

即可得到这两个句子的相似度约为 0.668

3.2.2.3 阈值的取值

对于阈值的取值，选择 Top K (K 由人为给出)个相似度最高（余弦相似度大于阈值）的留言主题。在本文中，我们通过多次试验，最终将 K 值取为 0.65。

由图 阈值测试表可知，当阈值取得 0.65 时，根据所分类的具体留言主题可知分类结果明显优于取值为 0.6 时。

将每类中含有的“留言主题”的数目进行排序，去除重复的类别，然后取前五个热点问题类别，即可得到热点问题表，附件 1；将每个类别中的“留言主题”依次输出，即可得到热点问题留言明细表，附件 2。

万科魅力之城小区底层门店深夜经营，各种噪音扰民		A市经济学院强制学生实习	
sim>0.65	sim>0.6	sim>0.65	sim>0.6
A5区劳动东路魅力之城小区油烟扰民	市丽发小区建搅拌站，噪音污	西地省科技职业学院未经沟通就	西地省科技职业学院未经沟通就强制
A5区魅力之城小区一楼被搞成商业门面	A市明昇壹城开发商施工占用学	西地省财政经济学院食堂只开放	A6区二中强制学生暑假补课且收费
A5区劳动东路魅力之城小区临街门面烧	A9市釜家桥小区楼下的夜宵店	A市涉外经济学院强制学生实习	A市涉外经济学院组织学生外出打工
A5区劳动东路魅力之城小区底层餐馆油	A5区劳动东路魅力之城小区油	西地省财政经济学院涉嫌宽带垄	中南林科大强制拆除学生空调回购价
A5区劳动东路魅力之城小区一楼的夜宵	A市A2区丽发新城小区遭搅拌站	西地省涉外经济学院变相强制当	投诉A市电子科技职业学院要求学生
A5区劳动东路魅力之城小区一楼的夜宵	A5区魅力之城小区一楼被搞成	A市商贸旅游职业技术学院强制	西地省财政经济学院校园宽带被垄断
A5区劳动东路魅力之城小区油烟扰民	A1区解放东路49号一楼住改商	A市经济学院寒假过年期间组织	西地省财政经济学院食堂只开放十余
A5区劳动东路魅力之城小区底层餐馆油	A5区劳动东路魅力之城小区临	A市经济学院组织学生外出打工	A市涉外经济学院强制学生实习
A5区劳动东路魅力之城小区临街门面烧	A市丽发小区建搅拌站，噪音污	A市经济学院强制学生实习	西地省财政经济学院涉嫌宽带垄断
A5区魅力之城小区一楼被搞成商业门面	A5区劳动东路魅力之城小区底	A市经济学院强制学生外出实习	西地省涉外经济学院变相强制学生“
A5区劳动东路魅力之城小区一楼的夜宵	A5区劳动东路魅力之城小区一	A市经济学院体育学院变相强制	A市商贸旅游职业技术学院强制学生
A5区劳动东路魅力之城小区一楼的夜宵	A市幕云街道丽发新城社区附近		A市长郡楚府中学强制学生周六补课
	A5区劳动东路新城新世界开发		西地省财政经济学院以报名名额已满
	反映A市A5区劳动东路昇明壹城		A市涉外经济学院寒假过年期间组织
	希望在A5区万科魅力城和阳光		A市雅礼中学强制高二学生周六补课
	A5区劳动东路魅力之城小区一		A市经济学院寒假过年期间组织学生
	A5区劳动东路魅力之城小区油		A市经济学院组织学生外出打工合理
	A5区劳动东路魅力之城小区底		A市经济学院强制学生实习
	A5区劳动东路魅力之城小区临		A市经济学院强制学生外出实习

图 5 阈值测试表

3.3 问题三“答复意见的评价”的分析方法与过程

3.3.1 分析过程

我们综合考虑了答复的相关性，完整性，可解释性和信息量大小来对答复意见给出一套评价方案。

首先我们对文本的完整性定义为：答复是否包含数值数据和文本数据，若包含两类数据则为 1，否则标为 0；

其次，对信息量的定义为：答复的文本长度，以答复的字数为计；

接着，由于中文文本中可解释性没有统一的定论和成型的算法，所以我们采用可读性指标来替代可解释性，本文借鉴了英文可读性指标的公式，即：

其中，一个中文汉字可按照两个字符计算。

最后，由相关性指标可定义为：两个文本向量的余弦相似度。于是我们得到四个指标，如表 4 所示：

表 4 指标表

指标	权重
相关性(Relevancy)	
完整性(Complete)	
可读性(Readability)	
信息量(Words)	

3.3.2 初步评价模型建立

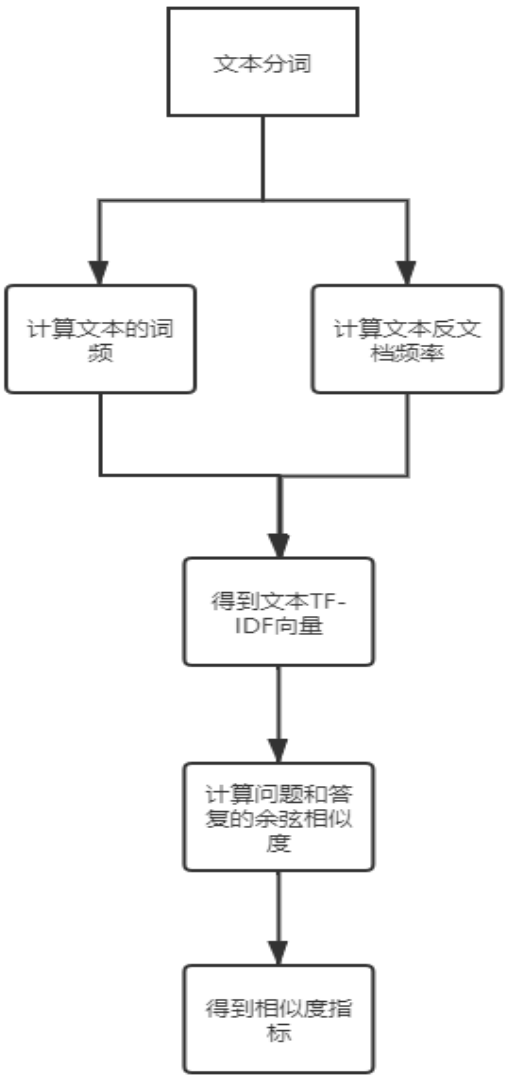


图 6 初步评价模型

通过以上推导，我们得到初步的一个评价模型：

通过以上模型我们可以对答复的效果进行初步的评价，但是我们的权重系数还没有很好的确定下来。我们通过以下几种方法来确定权重。

1. 专家打分法

专家打分法是指通过匿名方式征询有关专家的意见，对专家意见进行统计、处理、分析和归纳，客观地综合多数专家经验与主观判断，对大量难以采用技术

方法进行定量分析的因素做出合理估算，经过多轮意见征询、反馈和调整后，对债权价值和价值可实现程度进行分析的方法。常常应用于招标过程的评标阶段。

2. 灰度关联分析

对于两个系统之间的因素，其随时间或不同对象而变化的关联性大小的量度，称为关联度。在系统发展过程中，若两个因素变化的趋势具有一致性，即同步变化程度较高，即可谓二者关联程度较高；反之，则较低。因此，灰色关联分析方法，是根据因素之间发展趋势的相似或相异程度，亦即“灰色关联度”，作为衡量因素间关联程度的一种方法。

3. 通过灰度关联分析修正的专家打分法

通过专家打分法得到一个权重系数 W1，然后引入关联度。

得到最终的权重系数 W2。

参考文献

[1]梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.

[2]郭卫丽. 文本评论数据质量分析方法研究[D].重庆大学,2016.

[3]钟将,张淑芳,郭卫丽,李雪.主题特征格分析:一种用户生成文本质量评估方法[J].电子学报,2018,46(09):2201-2206.

[4 阮光册,谢凡,涂世文.基于 Word2vec 的图书馆推荐系统多样性问题应用研究[J].图书馆杂志,2020,39(03):124-132.

[5]徐蕾,张科伟.基于文本挖掘的京东商品评论分析[J].内蒙古科技与经济,2020(03):41+43.

[6]高峥,徐震.基于多元回归 KNN 的油田缺失数据填充方法[J].信息技术,2020,44(04):79-83.

[7]刘福刚.一种适用于中文博客自动分类的贝叶斯算法[J].长春师范大学学报,2019,38(12):36-43.

[8]李自昂. 基于逻辑回归的微博流行程度预测[D].河南大学,2017.

[9] 涂铭，刘祥，刘树春.Python 自然语言处理实战：核心技术与算法.机械工业出版社，2018.4

[10]郭银灵. 基于文本分析的在线评论质量评价模型研究[D]. 2017.

[11]刘思峰，杨英杰，吴利丰等．灰色系统理论及其应用［M］．北京：科学出版社．2014.

[12]朱磊. 基于 word2vec 词向量的文本分类研究[D].

[13]李荣陆. 文本分类及其相关技术研究[D]. 复旦大学.