第八届“泰迪杯”

全国数据挖掘挑战赛

作品名称：“智慧政务”中的文本挖掘应用

“智慧政务”中的文本挖掘应用

摘要

近年来，政府部门为了深入了解民情民意，结合当前科学技术，微信、微博、市长信箱、阳光热线等网络问政平台已经逐步成为了政府了解民意的重要渠道，各类社情民意相关的文本数据量在不断攀升，给以往主要依靠人工进行留言划分和热点整理的部门的工作带来了极大挑战，建立基于自然语言处理技术的智慧政务系统具有重大意义。

对于问题 1，利用 drop_duplicates 对附件 2 中的留言信息进行去重，得到不重复的留言信息，然后利用 jieba 库对留言信息进行分词，加入停用词表进行停用词的去除，以一级标签分类通过 TF-IDF 算法提取各类关键词并转化为权重向量，再利用朴素贝叶斯构建模型，最后用 F-Score 方法对构建的模型进行检验。

对于问题 2，首先对数据进行一系列的预处理操作，把分词后的文本和留言时间、留言详情保存到数据框中以便后面操作运用，将分词后的留言主题通过 TF-IDF 算法转化为权重向量，得到稀疏向量，对其进行 K-means 聚类，观察聚类结果，在问题集中的聚类结果总结聚类结果的问题描述，后利用 gensim 包计算向量间的相似性，进行两两比较，得到相似度，并考虑留言时间段，留言人数等因素，找出热点问题，结合热度评价指标，得出排名前 5 的热点问题，得到文件“热度问题表”。根据热度问题表的结果，按照热度指数从高到低将每条留言详情保存到另一个表格中，得到“热点问题留言明细表”。

对于问题 3，我们需要对留言的回复进行综合评价，然后通过评价结果给出建议，在这里我们决定对留言回复的相关性、时效性进行评价，并对回复进行情感分析。相关性主要是看留言回复与留言详情之间的主题是否相吻合，这里通过留言回复和留言详情的关键词进行比较得到；时效性主要由回复留言时间与发出留言时间的天数长短决定，时间越长，分数越低；情感评价分析，利用 BosonNLP_sentiment_sorce 情感评价表对答复意见进行情感分析，得出情感评分。

关键词：去重 jieba 分词 TF-IDF 算法 朴素贝叶斯 K-means 聚类 Gensim 热点问题 Difflib BosonNLP_sentiment_sorce 答复建议

Application of text mining in "intelligent government affairs"

Abstract

In recent years, the government departments in order to further understand the mood of public opinion, combined with the current science and technology, WeChat,such as weibo, mayor mailbox, sun hotline network asked ZhengPing Taiwan has gradually become the important channel for the government to understand public opinion, all kinds of public opinion related to the amount of text data in rising, leave a message to past mainly rely on artificial division and hot finishing department's work has brought great challenges, build wisdom e-government system based on natural language processing technology is of great significance.

For question 1, attachment 2 in the message information using drop_duplicates to heavy, get not to repeat the message information, then use jieba library to participle message information, join the stop list of stop words, by level 1 label through the TF - IDF algorithm to extract keywords and translated into the weight vector, using simple bayesian model build, in the end, the F - test Score method for the construction of the model.

For question 2, the first of a series of pretreatment, the text after the word segmentation and message time, message behind the details data saved in the box so that operation is applied, the message theme through the TF - after word IDF algorithm into the weight vector, and get the sparse vector, K means clustering, to observe the clustering results, the problem of clustering results summary clustering problem description, after using gensim package to calculate the similarity between the vector, the pairwise comparison, similarity, and consider the message time, factors such as the number of messages, find out the hot issues, evaluation index, combined with the heat Get the top 5 hot issues, get the file "hot issues list". According to the results of the hot topic list, save the details of each message in another table from high to low according to the heat index, and get the "hot topic list"

For question 3, we need to make a comprehensive evaluation of the replies to the messages, and then give Suggestions based on the evaluation results. Here we decide to evaluate the relevance and timeliness of the replies to the messages, and conduct emotional analysis on the replies. Relevance is mainly to see whether the topic between message reply and message details is consistent, here through message reply and message details of the keyword comparison; The timeliness is mainly determined by the time for replying to messages and the number of days for sending messages. The longer the time, the lower the score. Emotional evaluation analysis, using BosonNLP_Sentiment_sorce emotional evaluation form emotional analysis of the

response to the opinion to obtain an emotional score.

Key words: deduplication jieba segmentation tf-idf algorithm naive bayes Gensim k-means clustering GensimHot issues Difflib BosonNLP_sentiment_sorce responded to Suggesti

目录

一、挖掘目标

本次建模目标是对给予的留言信息，对其进行去重、jieba 中文分词工具对留言信息进行分词、去除停用词、TF-IDF 算法提取关键词、朴素贝叶斯模型构建、转化为特征向量等方法，达到以下目标：

1、得到对留言信息进行一级标签归类的依据，建立关于留言内容的一级标签分类模型，并对所建立的模型进行检验。

2、得到某一时间段内群众集中反映的某一问题，即热点问题，根据热度指数高低对热点问题进行排名，得到热点问题表。再将具体的留言信息提取出来得到热点问题留言明细表。

3、从答复的相关性、完整性、可理解性、时效性等多角度对答复的意见进行评分，对评分的结果做出相应的结论，提出真实可靠的建议。

二、分析方法与过程

总体流程图：

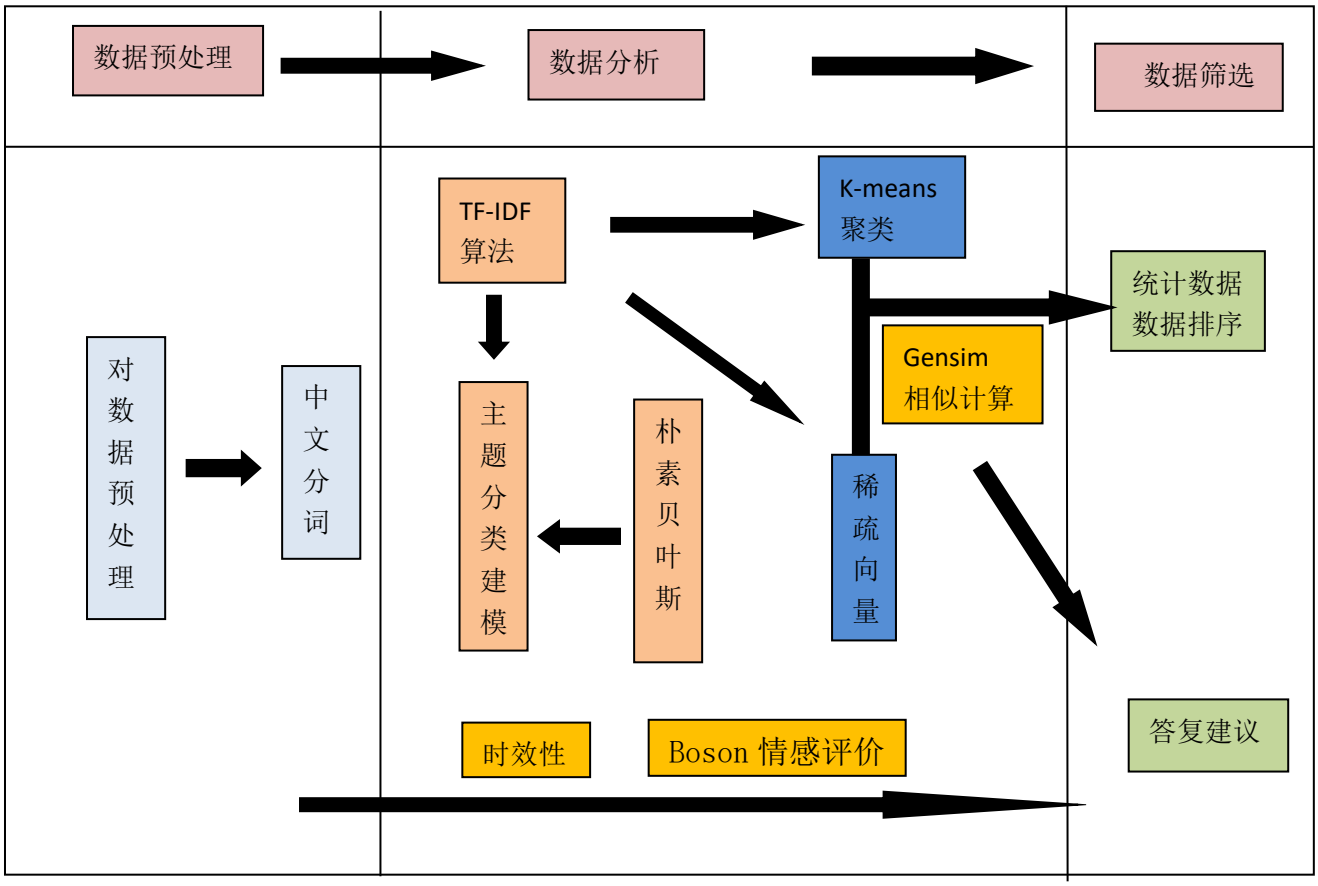

图 1 总体流程图

主要包括如下步骤：

步骤一：数据预处理，在题目给出的数据中，出现了部分重复的留言数据，在原始数据上进行去重处理，再进行中文分词、去除停用词操作。

步骤二：数据分析，在对留言信息分词后，需要将这些词语转化为向量，以供挖掘分析使用。在这里采用 TF-IDF 算法，找出不同标签下留言信息的关键词，把提取的各类标签下的关键词转化为权重向量。利用朴素贝叶斯进行模型构建。

步骤三：对留言信息权重处理后，转为稀疏矩阵，结合 K-means 聚类分类结果，进行 Gensim 相似计算得出热点问题，根据热度评价指标，得出排名前五的热点问题。

步骤四：对留言信息与答复意见利用 Difflib 相似行计算，留言时间与答复时间间的时间间隔作为评价时效性的依据，Boson 情感评价表对留言答复给出情感评价，结合三个指标给出答复建议。

1、问题 1 分析方法与过程

1.1 流程图

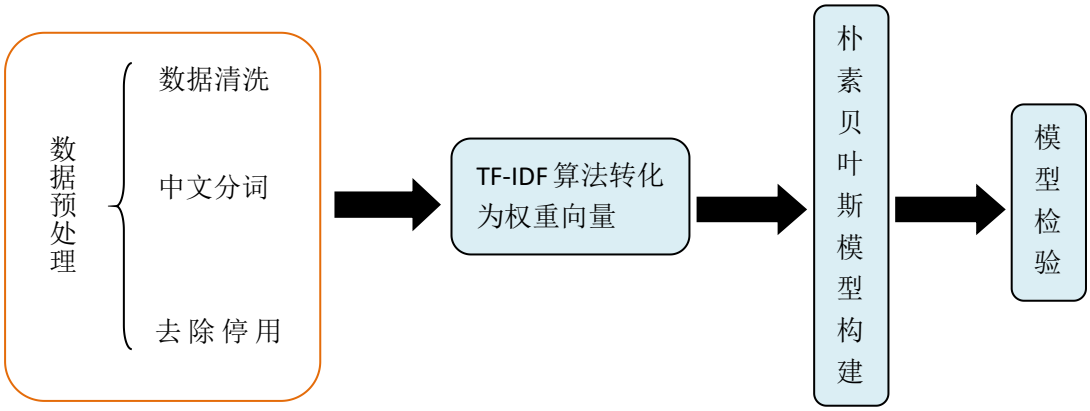

图 2 问题 1 流程图

1.2 数据预处理

1.2.1 留言信息的去重

在题目给出的数据中，出现了部分留言信息重复的情况。后面我们需要对留言信息进行词频统计来区分各类一级标签，而重复的留言信息对留言信息进行一级标签归类无任何意义，所以我们要去除重复信息，在这里我们采用较为简单的 drop_duplicates 来进行操作。另外，我们初步观察数据并未发现空缺的留言信息，

因此不需要对留言信息进行去空操作。

1.2.2 对留言信息进行中文分词

在对留言信息进行挖掘分析之前，我们先要把非结构化的文本信息转化为计算机能够识别的结构化信息。在附件 2 中以中文文本方式给出了留言信息数据，为了便于转换，我们首先要对留言信息进行中文分词。在这里我们采用中文分词包 jieba 进行分词，jieba 分词的原理是基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况构成的有向无环图，然后采用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，而对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。但是 jieba 分词也存在着一些缺陷，即它是固定的一些分词模式，无法跟随世界潮流，因此一些地名、专有名词可能会被分开，因此在这里我们根据留言信息内容中的专有名词进行自定义词典，然后将自定义的词典加入到 jiaba 库中，以保证分词有更高的正确率，对后续的操作有着重要影响。

1.2.3 TF-IDF 算法

在对留言信息分词后，需要将这些词语转化为向量，以供挖掘分析使用。在这里我们采用 TF-IDF 算法，把留言信息转化为权重向量。TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术，主要思想是如果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，则认为此词或者短语具有很好的类别区分能力，适合用来分类。基本原理如下：

（1）TF 是词频，表示某一个给定的词语在该文本中出现的次数。

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。

公式为：$$TF_{\omega}=\frac{\text{在某一类中词条}\omega\text{出现的次数}}{\text{该类中所有词条数目}}$$

（2）IDF 是逆向文件频率，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。

公式为：$$IDF=\log\left(\frac{\text{语料库的文档总数}}{\text{包含词条}\omega\text{的文档数}+1}\right)$$

注：这里分母加 1，避免分母为 0

（3）TF-IDF 是词频-逆文件频率，根据某一特定文件内的高词语频率，以及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，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

公式为：$TF-IDF=TF*IDF$

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。

1.2.4 生成 TF-IDF 权重向量

生成 TF-IDF 向量的具体步骤如下：

（1）根据一级标签进行分类，运用 TF-IDF 算法提取各类一级标签下留言信息的关键词；

（2）对每个一级标签对应的留言信息提取的关键词，合并成一个集合，分别计算每个一级标签对应的留言信息对于这个集合中的词频；

（3）生成每个一级标签下留言信息的 TF-IDF 权重向量，计算公式如下：

$$TF-IDF = TF * IDF$$

1.3 构建模型

1.3.1 朴素贝叶斯算法

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。原理如下：

（1）贝叶斯定理：是关于随机事件 A 和 B 的条件概率（或边缘概率）的一则定理。

贝叶斯公式：$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

当 A 与 B 相互独立时，$P(AB) = P(A)P(B)$

（2）条件独立性

两个随机变量 X 和 Y，若对于所有 x,y 有 $P(X = x, Y = y) = P(X = x)P(Y = y)$，则称随机变量是相互独立的，记作 $X \perp Y$，如果关于 X 和 Y 的条件概率对于 Z 的每一个值有 $P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$，则称随机变量 X 和 Y 在给定随机变量 Z 时是条件独立的，记作 $X \perp Y \mid Z$。

（3）朴素贝叶斯算法

由于上面两个特征，根据全概率公式展开，得到朴素贝叶斯法分类的基本公式为：$$P(C_k \mid X) = \frac{P(X \mid C_k)P(C_k)}{P(X)} = \frac{\coprod_{k=1}^{M} P(X_i \mid C_k)P(C_k)}{\sum_k P(C_k)\prod_{k=1}^{M} P(X_i \mid C_k)}$$

朴素贝叶斯构建模型步骤如下：

（1）准备工作阶段。

根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工

对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出的是特征属性和训练样本。即根据已经对留言信息进行了一级标签分类后的数据提取各类样本的样本特征属性。

（2）分类器训练阶段。

计算每个类别在训练样本中出现的频率以及每个特征属性划分对每个类别的条件概率估计，并记录结果。这一阶段输入的是特征属性和训练样本，输出的是分类器。

（3）应用阶段。

使用分类器对待分类项目进行分类。这一阶段输入的是分类器和待分类项，输出的是待分类项与类别的映射关系。即做到给定一个留言信息输入，然后经过分类器处理，最终可以得到该条留言信息所属的一级标签。

朴素贝叶斯算法流程图如下：

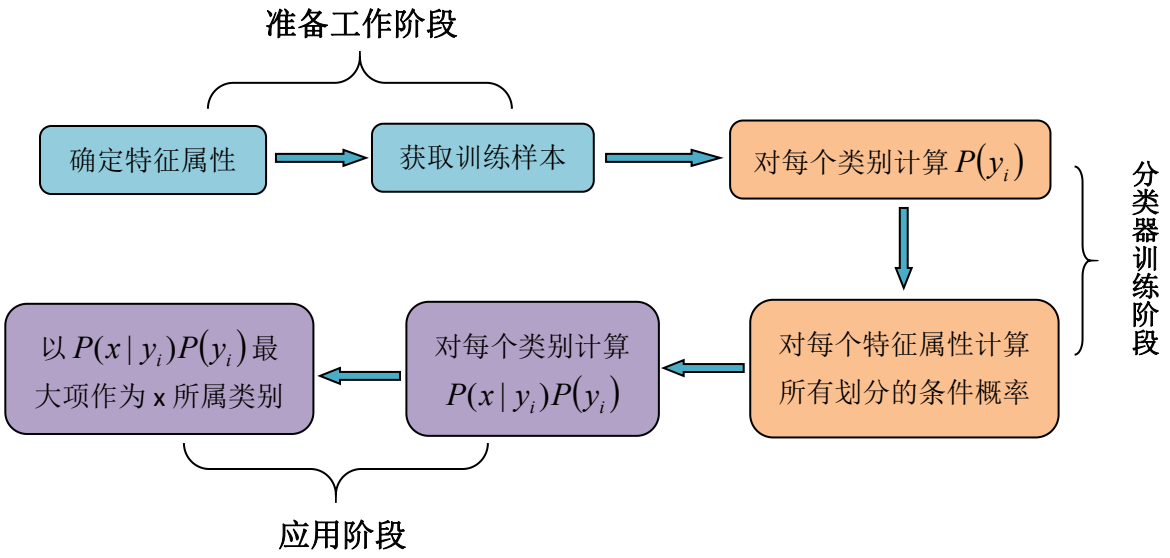

图 3 朴素贝叶斯算法流程图

1.4 模型检验

利用朴素贝叶斯建立分类模型，得出模型的预测结果，使用 F-Score 对分类方法进行评价：（数据见分类结果.csv）

$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$，其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率

计算结果为 70.08%。

1.5 模型改进

朴素贝叶斯分类原理是一种统计学方法。贝叶斯定理是分类器建模的理论基础，它利用条件概率原理，利用先验信息和样本数据信息确定时间的发生概率。

运用朴素贝叶斯分类模型进行分类时，首先需要利用数据集中的训练集部分构造分类模型，然后利用该模型对待分类的样本进行分类。因此，训练集的优劣

对分类的性能有很大影响。训练集在形成过程中可能会因为输入错误，测量时设备故障等原因产生噪声实例。改进方法：

(1)使用集成技术的分类器

分类器集成学习，也就是把不同的分类器进行组合，形成的最终分类器。Adaboost 算法的核心内容：第一步，给出一个弱学习算法和训练集合 $<(x_1,y_1),(x_2,y_2),.....,(x_N,y_N)>$,其中 $x_t \in X$,X 表示某个实例空间，含义是：在贝叶斯分类器中一个带有类别标志的几何，$y_t \in Y=\{+1,-1\}$；第二步初始化，让每一个训练例的权重都是相同的；第三步，将弱学习算法迭代 T 次，每次迭代后，按照训练更新训练集权重，训练失败的训练例更新为较大的权重，这些训练例在下一次迭代中会更加关注，从而得到一个权重的预测函数序列 $h_1,h_2,.....h_t$，每个预测函数 h_t 也赋予一个权重，h_t 的值有预测效果好坏决定。T 次迭代之后，采用带权重的投票法产生最终的结果，也就是预测函数 H。

(2)基于属性间关系的改进算法

朴素贝叶斯网络要求的条件独立的假设在现实世界很难满足，在工程上，为了提高贝叶斯分类器的性能，降低条件之间的联系。国内外机器学习的工作者先后提出了不同的方法，其中数扩展的朴素贝叶斯（TAN）是一类高效的、主流的方法。基本思想是：以图论为基础，把不包括类的属性组织为树，每个节点不但拥有类属性作为父节点，而且最多拥有一个其他属性作为父节点，而类作为根节点，没有父节点。目前实现 TAN 的常用方法有：由 Ftiedman 和 Goldszmidt 提出的 TANF 算法；由 Keogh 和 Pazzani 提出的 HCS 及 SP 算法。[1]

2、问题 2 分析方法与过程

2.1 流程图

[1] 高晓利 王维 赵火军 几种改进的朴素贝叶斯分类模型

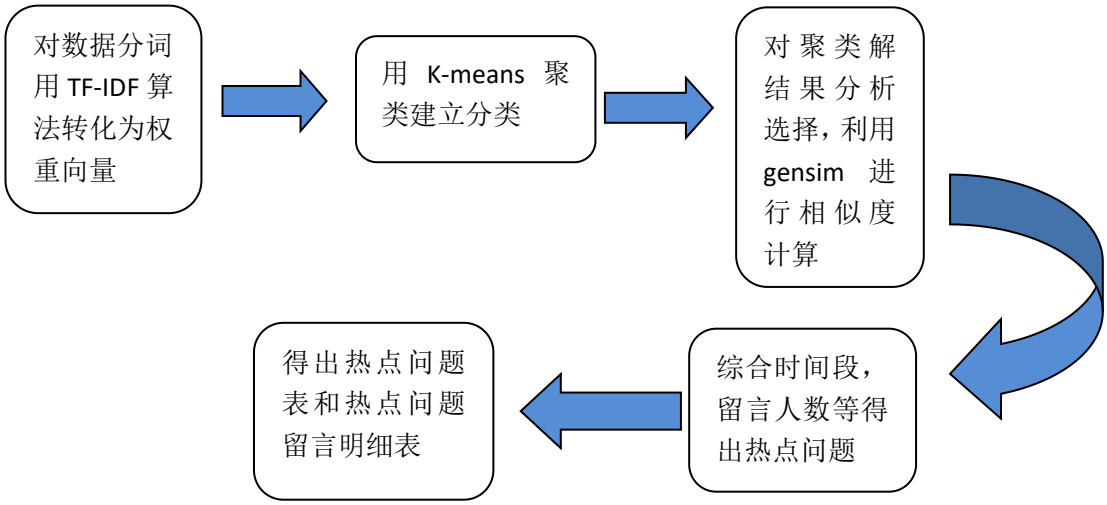

图 4 问题 2 流程图

2.2 数据预处理

首先需要对附件 3 中的留言内容进行预处理，预处理过程与上题预处理过程相似，另外我们需要将留言时间、留言详情保存在数据框中，方便后续制作表格时进行调用。

2.3 计算相似度

2.3.1 gensim 包

Gensim 是一个 python 的自然语言处理库，所用到的算法，如 TF-IDF，隐含狄利克雷分配，潜在语义分析或随机预测等，是通过检查单词在训练语料库的同意文档中的统计贡献模式来发现文档的语义结构，最后转化为向量模式，以便进行进一步的处理。此外，Gensim 还实现了 Word2vec 功能，能够将单词转化为词向量。

语料是一组原始文本的集合，用于无监督地训练文本主题的隐层结构。语料中不需要人工标注的附加信息。在 Gensim 中，corpus 通常是一个可迭代的对象。每一次迭代返回一个可用于表达文本对象的稀疏向量。

向量是由一组文本特征构成的列表。是一段文本在 Gensim 中的内部表达。

词典是所有文档中所有单词的集合，并且记录了各词的出现次数等信息。

模型是一个抽象的术语，定义了两个向量空间的变换，即从文本的一种向量表达变换为另一种向量表达。

2.3.2 使用流程

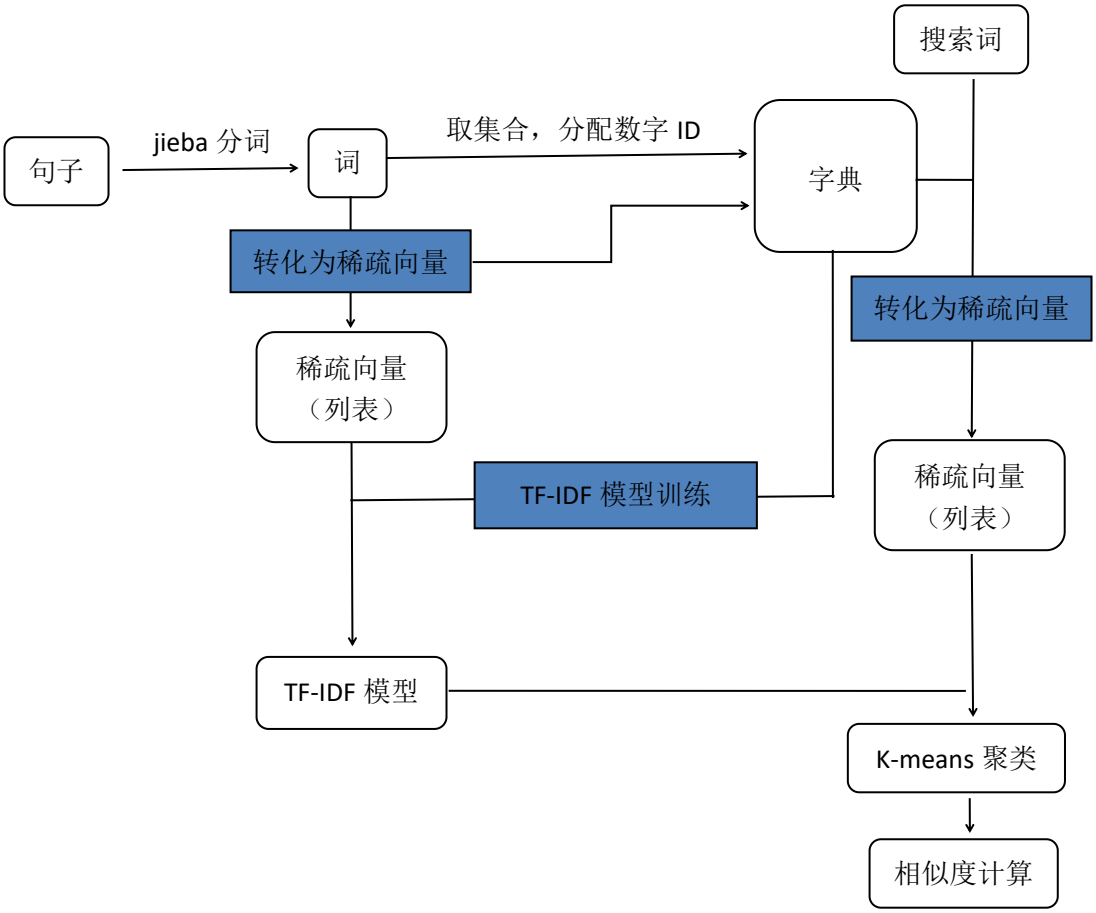

图 5 gensim 使用流程

2.4 K-means 聚类

聚类是一个将数据集中在某些方面相似的数据成员进行分类组织的过程，聚类就是一种发现这种内在结构的技术，聚类技术经常被称为无监督学习。

k 均值聚类是最著名的划分聚类算法，由于简洁和效率使得他成为所有聚类算法中最广泛使用的。给定一个数据点集合和需要的聚类数目 k，k 由用户指定，k 均值算法根据某个距离函数反复把数据分入 k 个聚类中。

2.4.1 算法流程

对于 K-Means 算法，首先要注意的是 k 值的选择，一般来说，我们会根据对数据的先验经验选择一个合适的 k 值，如果没有什么先验知识，则可以通过交叉验证选择一个合适的 k 值。

在确定了 k 的个数后，我们需要选择 k 个初始化的质心，就像上图 b 中的随机质心。由于我们是启发式方法，k 个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的 k 个质心，最好这些质心不能太近。

传统的 K-Means 算法流程：输入是样本集 D={x1,x2,...xm},聚类的簇树 k,最大迭代次数 N，输出是簇划分 C={C1,C2,...Ck}

1) 从数据集 D 中随机选择 k 个样本作为初始的 k 个质心向量：{μ1,μ2,...,μk}

2）对于 n=1,2,...,N，将簇划分 C 初始化为 Ct=∅t=1,2...k；对于 i=1,2...m,计算样本 xi 和各个质心向量 μj(j=1,2,...k)的距离：dij=||xi−μj||22，将 xi 标记最小的为 dij 所对应的类别 λi。此时更新 Cλi=Cλi∪{xi}c)；对于 j=1,2,...,k,对 Cj 中所有的样本点重新计算新的质心 μj=1|Cj|∑x∈Cj；如果所有的 k 个质心向量都没有发生变化，则转到步骤 3）

3）输出簇划分 C={C1,C2,...Ck}[2]

2.5 热点问题判定

利用 K-means 聚类，对筛选得到的结果(数据见聚类结果.excel)，用 gensim 相似度计算本题中相似度大于等于 0.3 则认为是对同一个问题的反映（数据见 sim2.csv）。进一步我们考虑时间跨度问题，经过热度指标计算，得到五类问题的热度高低，进而得出热点问题表和热点问题留言明细表。(表格见热点问题表.excl 热点问题留言明细表.excel)

在热度值的评价中，我们可以考究的因素有留言条数，留言时间跨度和点赞与反对数三个，我们以这三个为出发点对热度评价进行定义：

(1)计算以 10%为门槛的各类留言密集程度

一般随着时间跨度的拉长，话题的热度也会降低，因此我们尝试把留言数量作为纵坐标，月份作为横坐标制作条形图，通过观察留言密集程度来判断各类别的热度，以下我们以其中一类留言为例子作出以下条形图。

[2]刘健平 K-means 聚类算法原理

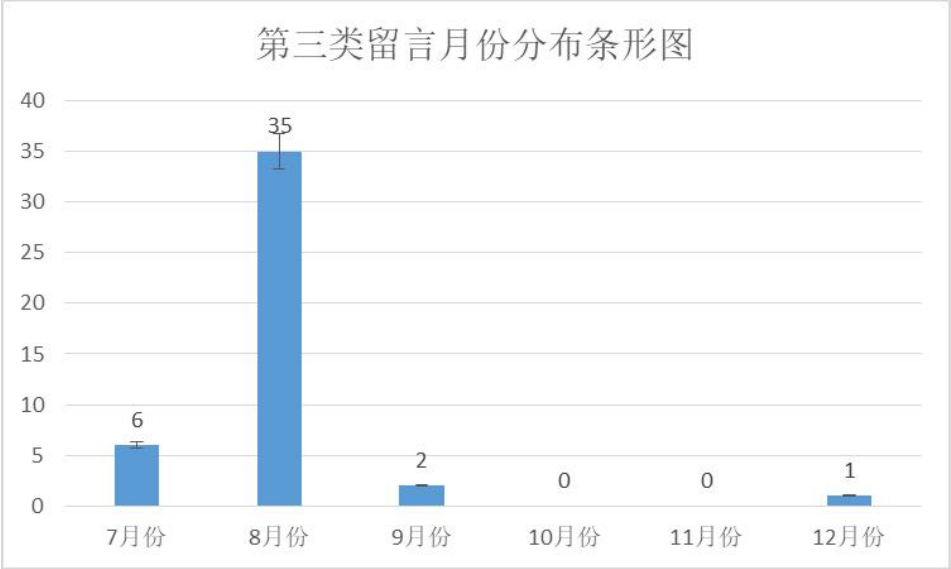

图 6 第三类留言时间分布

由以上条形图我们可以看出留言集中在 8 月份，然而在 12 月份依旧有少量留言，如果我们单纯以均值来代表留言密集程度的话，数值等于 44/4=11，会因为个别留言而被拉小，因此我们以类别留言数量的 10%为门槛，对留言数量超过类别留言数量 10%的月份才进行密集程度计算，此类别 10%为 4.4，因此只有 7 月份和 8 月份符合条件，这样以来，密集程度等于 41/2=21.5，增大了很多。

表 1 各类问题热度密集程度

类别	0	1	2	3	4
条/月	1.25	5.34	9.34	21	2.75

公式：留言密度 c 公式：$c = \frac{a}{b}$

留言数量超过本累留言总数 10%的月份数量 a
留言数量超过本累留言总数 10%的月份的留言总数 b

（2）计算各类留言条均点赞反对数

在热度值的考究上，点赞数和反对数也是很大的一个因素，而点赞与反对在情感上不存在多大的情感比重，因此可以将两者数量相加起来共同比较。为了便于与其他类别相比较，我们将其除以留言条数，以均值进行比较。

表 2 各类问题评论指标

类别	0	1	2	3	4
个/条	2.6	1.64	0.82	0.58	2

公式：条均点赞反对数$\mathrm{f}=\frac{\mathrm{d}}{\mathrm{e}}$

类别点赞加反对数总和 d

留言条数 e

（3）计算密集程度和条均点赞反对数占总体比例

有了密集程度和条均点赞反对数，由于两个数据的量级不同，我们需要对两者进行处理，才能使两者有可加性。

表 3 密度密集程度与评论数综合表

类别	0	1	2	3	4
密集程度	0.03	0.13	0.24	0.53	0.07
条均点赞反对数	0.34	0.21	0.11	0.08	0.26

公式：密集程度所占比例$\mathrm{g_i}=\frac{c_i}{\Sigma c_i}$

条均点赞反对数所占比例$h_i=\frac{f_i}{\Sigma f_i}$

（4）加权后得出结果

对处理完的密集程度和条均点赞反对数加权，以便于最终算出热度值，因为密集程度相对条均点赞反对数来说更重要，因此我们觉得给予密集程度 70%的权重，给予条均点赞反对数 30%的权重。最终得出热度值评价结果。

表 4 热度值评价

结果 第几类	0	1	2	3	4
加权后分数	0.123	0.154	0.201	0.395	0.127
排名	5	3	2	1	4

公式：$\mathrm{S}=\mathrm{g_i}*0.7+\mathrm{h_i}*0.3$

3、问题 3 分析方法与过程

3.1 流程图

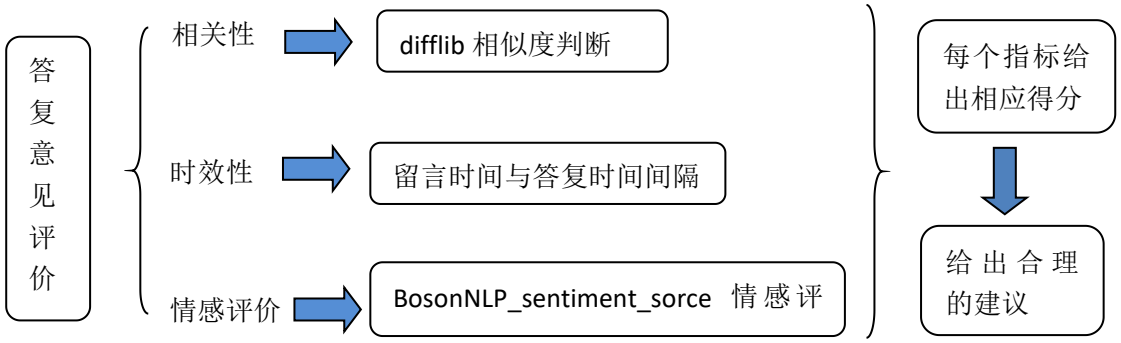

图 7 问题 3 流程图

3.2.相关性

在题目给出的数据中，包含留言详情与答复意见，分析答复意见是否与留言详情中反应的问题时是否一致，对其进行相似度比较。将留言详情和答复意见放到两个列表中，然后用 Difflib 库进行相似度比较。

3.2.1 Difflib

Difflib 这个模块提供的类和方法用来进行差异比较，它能够比对文件并生成差异结果文本或者 html 格式的差异化比较页面，如果需要比较目录的不同，可以使用 filecmp 模块。

Class difflib.SequenceMatcher 此类提供了比较任意可哈希类型序列对方法。此方法将寻找没有包含‘垃圾’元素的最大连续匹配序列。

通过对算法的复杂度比较，它由于原始的完形匹配算法，在最坏情况下有 n 的平方次运算，在最好情况下，具有线性的效率。它具有自动垃圾启发式，可以将重复超过片段 1%或者重复 200 次的字符作为垃圾来处理。可以通过将 autojunk 设置为 false 关闭该功能。

3.2.2 分析流程

图 8 相关性流程图

3.2.3 相似度判定

对得到留言详情和答复意见的相似度结果进行排序，最大的相似度是 0.708，对此制定一个评分标准，对其进行相似度的评分。（数据见 sim3.csv）

表 5 相似度评分标准

相似度	评分
>0.5	10
0.3~0.5	7
0.1~0.3	4
0.1 以下	1

3.2.4 相似度分析

对得出留言详情和答复意见相似度评分后，绘制扇形图，分析答复意见相似度情况。

表 6 相似度得分情况

得分	答复数
10	34
7	914
4	1578
1	290

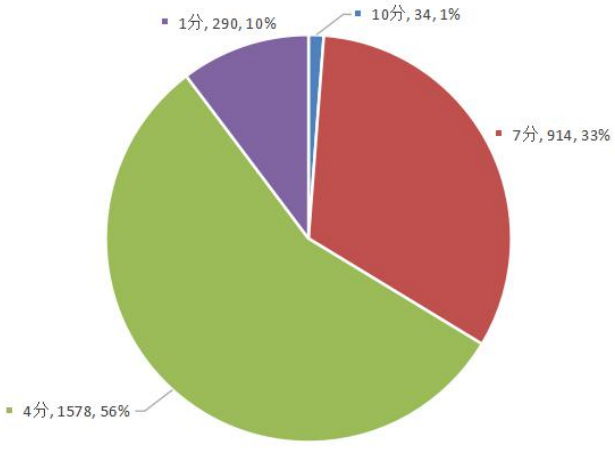

图 9 相似度得分扇形图

从相似度评分结果来看，相似度并不是很理想，相似度在 0.5 以上的只有 34 条答复意见，大部分答复意见在集中在 7 分和 4 分，少部分答复意见只有得到 1 分。

3.3 时效性

答复是否及时也是答复意见质量一个重要指标，分析留言时间与答复时间的时间跨度，来分析答复是否及时有效。提取留言时间和答复时间，转变时间类型为 datetime 作出时间差。

3.3.1 Datetime

Datetime 是一个关于时间的库，主要包含的类有：

date	日期对象，常用的属性有 year,month,day
time	时间对象，hour,minute,second,毫秒
datetime	日期时间对象，hour，minute，second，microsecond
timedelta	时间间隔，即两个时间点之间的长度

3.3.2 分析流程

图 10 时效性流程图

3.3.3 时效性判定

对得到留言时间和答复时间的时间差结果进行排序，最快的回复时间在留言当天就进行了回复，回复效率很高，但也存在答复时间过慢的，答复时间超过一年甚至更久，没有及时答复，群众等待时间过长，造成问题没有及时解决的情况。对此制定一个评分标准，对其进行相似度的评分。(数据见 time.csv)

表 7 相似度评分标准

时间差（天）	评分
<30	10
31~60	8
61~90	6
91~120	4
>120	2

3.3.4 时效性分析

对得出留言时间和答复时间差评分后，绘制扇形图，分析答复时间时效性情况。

表 8　时效性得分情况

得分	时间间隔数
10	2361
8	299
6	74
4	34
2	48

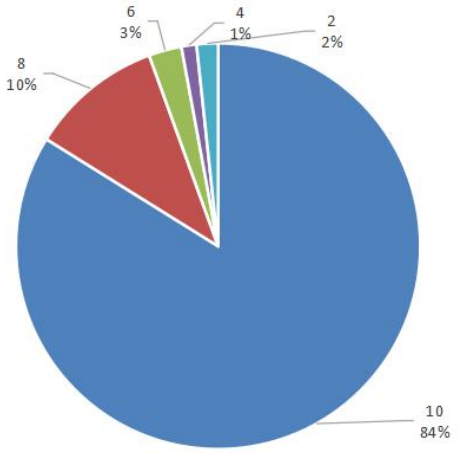

图 11　时效度得分扇形图

从时效性评分结果来看，在一个月内回复留言占 84%，回复留言的效率都较高，对群众的留言回复及时，存在少数答复时间过长的情况。

3.4 情感评价

答复时情感是否积极，也是提升群众满意度的重要指标。通过对答复意见情感分析，分析答复意见的情感态度，进一步给出建议。对答复意见进行分词、去停用词操作，然后导入 BosonNLP 表，对每条答复意见匹配计算，得出每条答复意见的情感评分。

3.4.1 BosonNLP_sentiment_sorce

BosonNLP 表包含不同情感词的评分，根据这个表对答复意见进行评分。

3.4.2 分析流程

图 12 情感评分流程图

3.4.3 情感分析

对得出答复意见情感得分后，分析情感得分情况，对得分进行划分，统计答复意见在各区间的个数，分析分布情况，得出总体答复意见的情感倾向。(数据见 mood.csv)

表 9 答复意见情感得分情况

分数	数量
<0	107
[0,50)	2132
[51,100)	419
[101,150)	87
[151,200)	22
[201,250)	13
>250	15

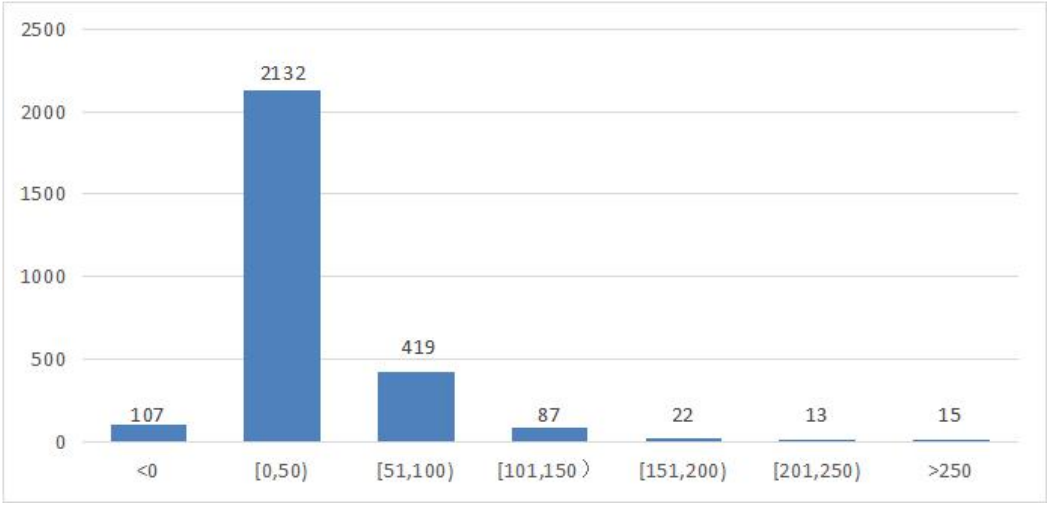

图 13 答复意见情感得分条形图

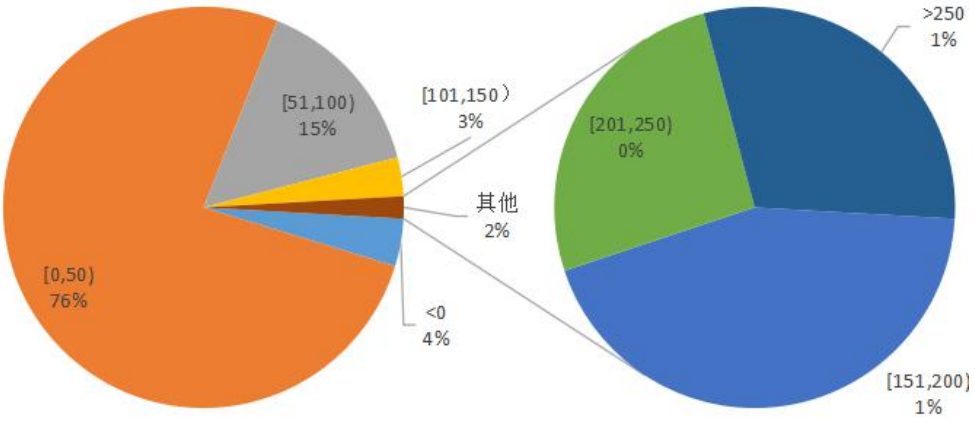

图 14 答复意见情感得分扇形图

从条形图可看出，答复意见的情感得分集中分布在[0,50)这个情感区间，有少部分的情感得分小于 0，情感态度较差；少数答复留言的情感得分较高，超过 150 分。结合以上分析，答复意见的情感得分总体偏低。

3.5 结合研究结果，给政府部门答复意见提建议

随着网络时代的到来，网络问政平台逐渐成为政府了解民意、汇聚民智、凝聚民气的重要渠道。民众也通过问政平台，拉近与政府间的距离，更加方便、快捷的与政府部门沟通，向政府部门反映社会现象与生活中遇到问题。政府部门对民众的答复是否有效、及时，侧面反映了政府部门的工作效率与服务质量，是否及时的帮助民众解决问题，规范社会秩序，提升城市形象，增加民众的幸福感。

综合上面的分析，给政府部门的建议：

(1)精简答复，揭示问题本质

面对成百上千的留言信息，政府部门答复意见简短有效，既可以提高工作效率，也给民众一个确切答复，了解问题的进展程度与解决方案，提升政府形象。

(2)及时答复，了解民意与民情

对民众的留言，要及时处理，及时处理可以了解最近社会存在的问题现象与民众存在的一些困惑，有利于对下一阶段政府部门城市规划提供方向。及时回复也侧面反映了政府部门的工作效率，提高民众对政府部门的满意度。

(3)提升情感，共筑和谐社会

政府部门答复时，针对民众提出的合理问题，积极礼貌的给予回应，拉近与民众间的距离。政府部门文明礼貌，也为人民群众树立一个好榜样，每个人在在周围讲礼貌的环境下成长，我们的社会也会变得更加和谐美丽。

向政府部门工作反映问题的同时，人民群众也应该自己自觉履行好自己的责任与义务，留言时应礼貌，提供的问题与信息应准确真实，不能隐瞒欺骗，直击问题本质。

4、 结论

对网络留言进行分析与研究，了解社会主要存在的问题，民众的生活质量，对政府部门制定城市发展，提升管理水平，提高施政效率有重要意义，同时也是文本分析的一个课题、一个难题。传统的人工解读分类已经不能满足当前社会发展的需要。本文采用朴素贝叶斯构建模型，对目前收集到的留言中，确定特征属性，获取训练样本得到每个类别标签划分的条件概率，从而对留言信息实现分类，减轻工作负担提高效率。

通过分析，得出最近时间民众较多反映的问题，可以看出在哪些地区，民众反映的问题较多，民众大多反映什么问题，归纳出热点问题表。了解民意，为城市下一步发展提供方向。

政府部门对民众的答复进行分析，发现政府部门工作中可能存在的一些问题，需要不断完善，更好的为人民服务，提升服务品质，树立一个好榜样，共同构建和谐社会。

5、参考文献

[1] 高晓利 王维 赵火军 几种改进的朴素贝叶斯分类模型
[2] 刘健平 K-means 聚类算法原理