“智慧政务”中的文本挖掘应用

摘要

本文通过对中文留言文本的处理，构建了词向量模型，基于词袋模型，运用朴素贝叶斯分类模型和逻辑斯蒂回归分类模型对留言进行分类预测。再借助词向量的距离来计算文本相似度，依据热度评价指标对问题的热度进行归类排序。最后我们基于 SERVQUAL 理论建立评价指标对答复质量进行评分。

对于留言分类预测，我们对中文文本进行预处理后，利用一级标签作为留言文本标记，构建词向量模型，对分类器进行训练后，利用 F-Score 模型对其预测精度进行评价，最终结果显示精度基本都达到 85%，其中逻辑斯蒂模型效果最佳。

对于留言热度评价指标，我们利用 TF-IDF 技术对文本相似度进行识别，用层次聚类树状图对文本进行归类，最后按照热度评价标准对留言问题热度进行排序，选出前五个热点问题。

对于答复评价，我们依据 SERVQUAL 理论，构建合理的评价指标，然后通过调查问卷收集数据，通过用户打分和综合计算得出答复意见的质量评分。

关键词：词向量 逻辑斯蒂回归分类模型 相似度 TF-IDF doc2bow SERVQUAL

Text Mining Application in "Smart Government Affairs"

Summary

This paper constructs a word vector model by processing the Chinese message text. Based on the word bag model, the naive Bayes classification model and the logistic regression classification model are used to classify the message. Then, the distance between word vectors is used to calculate the text similarity, and the popularity of the question is sorted according to the popularity evaluation index. Finally, we establish evaluation indicators based on SERVQUAL theory to score the quality of responses.

For message classification prediction, we preprocessed Chinese text, used first-level tags as message text tags, built a word vector model, trained the classifier, and evaluated its prediction accuracy using the F-Score model. The final result was displayed The accuracy is basically 85%, and the Logistic model has the best effect.

For the evaluation index of message popularity, we use TF-IDF technology to identify text similarity, use hierarchical clustering tree diagram to classify the text, and finally sort the message issue according to the evaluation criteria

of the heat, and select the top five hot spots problem.

For response evaluation, we construct a reasonable evaluation index based on SERVQUAL theory, and then collect data through a questionnaire, and obtain the quality score of the response opinion through user scoring and comprehensive calculation. .

Keywords: word vector, logistic regession classificantion model,similarty,TF-IDF,doc2bow, SERVQUAL

目录

一、挖掘目标

本次建模的目标是利用自然语言处理和文本挖掘的办法对网络问政平台的管理进行优化。将网络留言文本转化成词向量，利用朴素贝叶斯和逻辑斯蒂等分类模型对留言进行分类并利用一级标签加以区分，以减轻工作人员的依靠人工处理存在的工作量大，效率低等问题。同时借 F-Score 模型对该分类模型的精度进行判断，以选择最合适的分类模型。

针对市民集中反映的热点问题，我们利用 TF-IDF 技术来计算文本内容的相似度，依据相似度对文本内容进行分类，同时按照其热度进行排序，呈现出热点问题明细，便于工作人员答复意见工作的进行，减少重复的无意义工作。

关于答复意见质量我们基于 SERVQUAL 原理建立了合理的评价指标，并对服务质量进行评分，便于工作人员发现问题，解决问题，为更好地服务于大众提供一定的指导作用。

二、问题一的分析方法与过程

2.1 总体流程

本部分利用流程图的表现形式对问题一的建模方法及过程进行描述，并对个部分进行简要说明。

问题一所用步骤主要包括：

步骤一：利用 python 对留言进行去停用词处理，除去无意义的词

步骤二：将留言进行分词，使其组合成新的词序列

步骤三：利用一级标签对留言进行分类标记

步骤四：构建向量空间模型，将中文序列转化成计算机能够理解的语言

步骤五：利用 sk-learn 把数据切分成训练集和测试集

步骤六：利用训练数据构建文本的词袋模型

步骤七：进行模型的训练，测试模型精准度并不断调整优化

步骤八：利用训练后的分类器模型进行分类的精准度预测，选择适合留言系统分类的模型。

具体流程我们利用流程图的形式呈现，流程图如下所示：

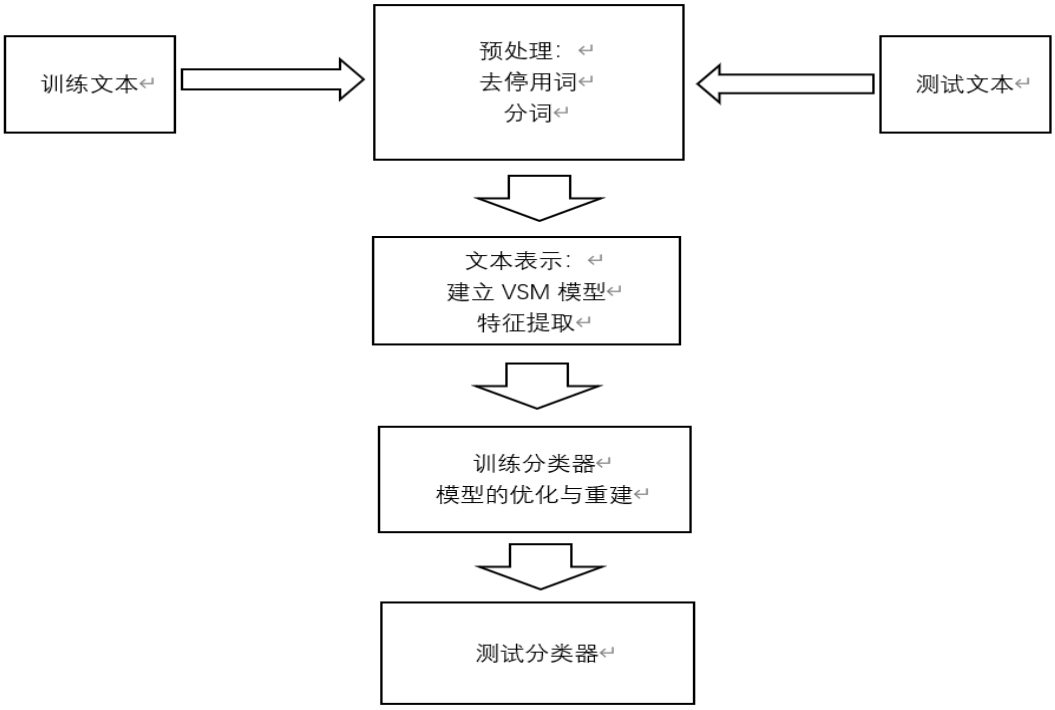

图 1 问题一建模方法的总流程图

2.2 具体步骤

2.2.1 预处理

在进行中文文本的挖掘时，首先需要除去与文本意义无关的内容，也就是对文本进行预处理。在这中间我们使用 jieba 对文本将留言文本的停用词剔除，并且把留言切分，获得一个用空格连接的文本分词，便于进行下一步的向量化。最后，再根据所给数据的一级标签对留言内容进行分类。

去停用词是指将文本中没有意义或者意义很小的词语去除，以保留最能代表该语境的和文本意义的词。Zip 定律表明如果把单词出现的频率按由大到小的顺序排列，则每个单词出现的频率与它的名次的常数次幂存在简单的反比例关系，在英语单词中，只有极少数词会被经常用到，而绝大多数词很少被使用，事实上，包括汉语在内的许多国家的语言都有此特点，这也正是我们去停用词的依据和原理。我们依据网络提供的停用词词库，结合留言系统的实际情况确定了合适的停用词词库，在此我们采用的部分停用词如下表所示

表 1 部分停用词实例

〞	#	$
%	&	(
)	↑	①

一下	一方面	三天两头
万一	不少	严重
已经	尽早	帮助

中文分词分词就是将连续的字序列按照一定的规范切分成单独的词再重新组合成词序列的过程。为了更好地体现留言的内容，我们将数字和字母去除，并且使用 python 对市民留言进行切词，得到独立的词块。

依据规定，我们对每条留言用城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生等一级标签进行处理。

2.2.2 短文本向量化模型

经过上述预处理后，我们接下来就需要将中文文本转化成计算机能够理解的理想模型，也就是文本表示过程。在该过程中我们利用向量空间模型，即 VSM 模型来反映文本的内容与特征，下面我们对该模型进行一些简单介绍。

特征项表示构成向量空间模型的基本单位，而文本就可以用表示为项集，表示如

$$D = (t_1, t_2, t_3, \dots, t_n)$$

其中，t_k为特征项，$1 \le k \le n$

对于有n个项的文本$D = (t_1, t_2, t_3, \dots, t_n)$，项常常被赋予一定的权重来表示它们在文本$D$中的重要程度，及可以表示为

$$D = (t_1, w_1; t_2, w_2;, t_3, w_3; \dots; t_n, w_n)$$

简记为

$$D = (w_1, w_2, w_3, \dots, w_n)$$

其中w_k就是特征项t_k的权。

对于给定的文本$D = (t_1, w_1; t_2, w_2;, t_3, w_3; \dots; t_n, w_n)$，其中既可能出现重复的内容，有存在前后顺序的关系，在分析过程中存在一定的困难，为了简化分析，我们约定各特征项$t_k(1 \le k \le n)$互异，各特征项t_k无先后顺序。此时就可以用 n 维坐标系来表示所有特征项，其中坐标值就是特征项的权重，而该文本就可以被看作一个 n 维向量。

向量之间的相关程度往往与它们的的相似度有关，故将文本转化为向量模型后，我们可以借助向量之间的距离来表示文本之间的相似度。

我们假设

$$D_1 = D_1(w_{11}, w_{12}, w_{13}, \dots, w_{1n})$$
$$D_2 = D_2(w_{21}, w_{22}, w_{23}, \dots, w_{2n})$$

根据 n 维空间原理，我们所求的文本相似度即两向量的距离，一般使用两向量内积来计算，公式表示为

$$\mathrm{Sim}(D_1, D_2) = \sum_{k=1}^{n} w_{1k} \times w_{2k}$$

或者可以用夹角余弦值表示，公式如下

$$\mathrm{Sim}(D_1, D_2) = cos\theta = \frac{\sum_{k=1}^{n} w_{1k} \times w_{2k}}{\sqrt{\sum_{k=1}^{n} {w_{1k}}^2} + \sum_{k=1}^{n} {w_{2k}}^2}$$

其中，相似度原理在我们问题二的处理中有重要应用。

2.2.3 词袋模型

基于上述原理，我们利用词袋模型对留言系统进行分类。词袋模型是一种自然语言处理和信息检索下的简化表达模型，在该留言系统的应用主要有两个环节，一是利用训练集训练分类器，把训练数据转化为词袋模型，二是利用测试集来品评价分类器的分类精度。

针对附件 2 所给的数据集，我们将特征向量的维度设置为 5000，对文本进行处理，抽取词袋模型特征。利用 python 把数据切分，用 sk-learn 把数据切分成训练集和测试集，利用训练数据构建文本的词袋模型。

2.2.4 分类方法

经过上述处理后，我们就可以实现对留言的文本进行分类。分类的基本思路就是利用标记好的训练集对分类器进行训练，生成训练模型，即前面提到的词袋模型。基于此，我们就可以对测试文本进行分类，来检验分类器的预测准确度，而这一准确度就是我们判断该方法是否适合应用的重要标准，也是我们这一问题的最终目标。

目前已经有不少成熟的分类算法，我们在训练中用到的分类器有：朴素贝叶斯(NBM)、逻辑斯蒂(Logistic Regression)、支持向量机(SVM)、随机森林(Random forest)、决策树（Decision Tree)、梯度下降树（GBDT)。下面来具体介绍一下我们所用的分类其中效果较好的几种分类算法：

a.朴素贝叶斯

朴素贝叶斯算法是一个经典的统计学习算法，其主要理论基础就贝叶斯公式，贝叶斯公式表示如下：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

在文本分类的应用中，朴素贝叶斯算法主要用于计算短文本属于某一类别的概率，即我们所求的每条留言属于哪一标签分类。而文本属于哪一类别的概率是要通过文本预处理后的特征项来计算的，这往往是通过我们训练得到的。故我们可以用更加明确的公式来说明文本分类中朴素贝叶斯的含义，留言文本d_i属于一级标签c_j的概率：

$$P(c_j|d_i) = \frac{P(d_i|c_j)P(c_j)}{P(d_i)}$$

在这里我们假设文本之间的特征是相互独立的，不受彼此的影响。但是这在实际语境中并不能完全成立，受语言习惯、表达主题和预警的影响，特征此间可能会存在一定的关联，且该关联程度大小不一。

基于朴素贝叶斯的这一特点，我们尽可能加大了训练的数据集，以保证该算法的准确率。在我们所使用的语料库的训练下，该算法的准确度让人比较满意，在实际应用中具有较高的操作性。

b.逻辑斯蒂回归分类算法

逻辑斯蒂回归是一种二分类模型，用条件概率分布$P(Y|X)$表示，是参数化的逻辑斯蒂分布。训练分类的主要原理就是寻找最合适的拟合参数，优化算法，根据现有数据对分类边界建立回归公式，以此进行分类。

线性分类器是通过假设特征与分类结果存在线性关系的模型，该模型利用累加每个维度的特征和各自权重的乘积来帮助决策，关于特征和权重我们在上述向量空间模型

$$D = (t_1, w_1; t_2, w_2;, t_3, w_3; \dots; t_n, w_n)$$

已经进行处理，为了避免其通过原点的假设，在这里增加一个截距 B。这种线性关系可以表示为

$$f(w, t, b) = wt + b$$

而我们所期望的函数预算结果范围在 0 到 1 之间，于是我们运用逻辑斯蒂函数来构造预测函数：

$$g(z) = \frac{1}{1 + e^{-z}}$$

在这里$g\epsilon\{0,1\}$，函数图像如下图所示

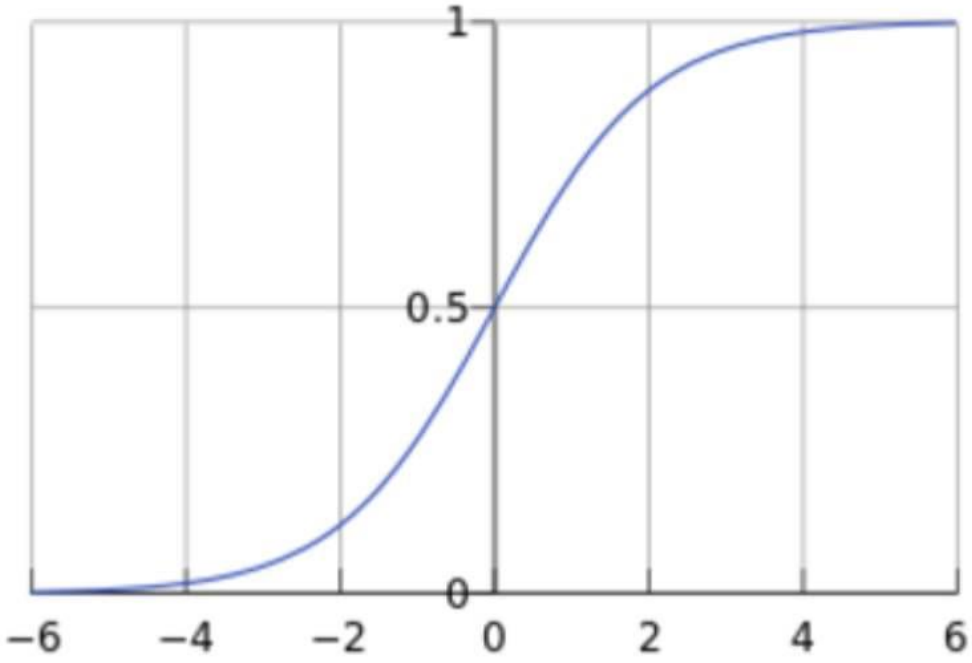

图 2 Logistic 回归方程图像

将以上公式整合后我们就可以得到逻辑斯蒂回归方程

$$g\big(f(w, t, b)\big) = \frac{1}{1 + e^{-f}}$$

据此模型，可以对文本的特征向量进行分类。该模型以样本的最大似然估计为训练目标，避免模型出现过度拟合现象，误判可能小。

同时，实践也表明该模型在众多分类器模型中的准确率较高，在该环境下的文本分类应用中可行性最高。

2.2.5 混淆矩阵分类算法

以词袋模型为基础，用混淆矩阵来验证分类模型的查准率和查全率，由此可以来验证分类器的预测准确度，在这里我们采取的评价模型是题目所示的 F-Score 模型：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

其中P_i表示第i类留言的查准率，R_i表示第i类留言的查全率。

混淆矩阵是模式识别中的一种可视化的分类效果示意图，它可以表示样本数据的真实类别和预测结果之间的关系，故也是我们评价分类器的方法。我们假设一级分类有N类标签，每个训练集文本拥有T_0个样本，每一种一级分类就分别拥有T_i（$i = 1, \ldots, N$）个数据，采用某分类器C，cm_{ij}表示w_i类留言被分类器判断成w_j类标签的留言数据，则可得$N \times N$维混淆矩阵$CM(C, D)$：

$$\boldsymbol{CM}(C,D) = \begin{pmatrix} cm_{11} & cm_{12} & \cdots & cm_{1i} & \cdots & cm_{1N} \\ cm_{21} & cm_{22} & \cdots & cm_{2i} & \cdots & cm_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ cm_{i1} & cm_{i2} & \cdots & cm_{ii} & \cdots & cm_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ cm_{N1} & cm_{N2} & \cdots & cm_{Ni} & \cdots & cm_{NN} \end{pmatrix}$$

我们以逻辑斯蒂分类器训练结果为例

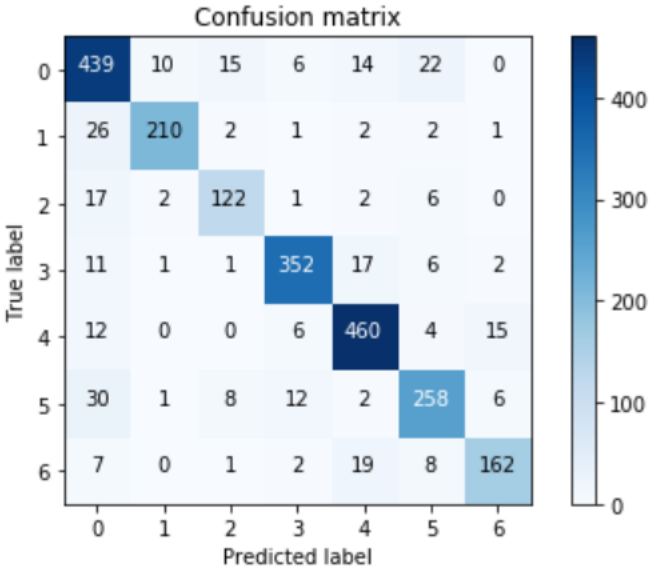

图 3 逻辑斯蒂模型下混淆矩阵训练结果

该混淆矩阵行坐标表示真实标签，列坐标表示预测标签，对角线表示预测标签符合真实标签的数量，我们的精度就可以用对角线数值之和除以矩阵所有元素的和，召回率只需要关注真实值是该标签的那一行，用预测值除以行元素和。

例如对于分类 0，他的召回率为 $\frac{439}{439+10+15+6+14+22+0}$

在此基础上，利用 F-Score 模型，我们可以计算得到分类的预测准确率：

	precision	recall	f1-score	support
0	0.81	0.87	0.84	506
1	0.94	0.86	0.90	244
2	0.82	0.81	0.82	150
3	0.93	0.90	0.91	390
4	0.89	0.93	0.91	497
5	0.84	0.81	0.83	317
6	0.87	0.81	0.84	199
accuracy			0.87	2303
macro avg	0.87	0.86	0.86	2303
weighted avg	0.87	0.87	0.87	2303

图 4 逻辑斯蒂模型下的预测准确率

由此可以看出 F-Score 模型下的分类模型对于各个一级标签的预测准确率都高于 80%，具有较高的参考性。

2.3 结果分析

我们借助短文本向量化模型，利用词袋模型对留言分类进行训练，尝试了多种分类器模型，最后得到的分类器预测准确率如下表所示

表 2 分类器模型预测准确率

模型	预测准确率
朴素贝叶斯	0.8571
支持向量机	0.8385
逻辑斯蒂	0.8697
随机森林	0.8385
决策树	0.7603
梯度下降树	0.8415

从上表可以看出，预测准确率大致稳定在 85%，说明该方法对于留言分类的判断具有较大的参考价值，且处理速度较快，可以满足信息繁杂，样本容量大的留言系统的需求，从而有效减少了工作人员的人工分类的时间，对于针对性回复有较大帮助。

三、问题二的分析与方法与过程

根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按格式给出排名前 5 的热点问题，还有具体留言信息。

3.1 总体流程

任务 1 问题识别：

a.如何从众多留言中识别出相似的留言

b.相似度计算（特征多，计算量大）

任务 2 问题归类:

把特定地点或人群的数据归并，即把相似的留言归为同一问题，结果对应表 2

聚类树状图（gensim corpora Dictionary doc2bow TF-IDF）

任务 3 热度评价:

热度评价指标的定义和计算方法，对指标排名之后得出对应表 1

3.2 具体步骤

(1)利用 TF-IDF 将语料库转换为向量空间（vector space）

(2)计算每个文档间的余弦距离（cosine distance）用以测量相似度

(3)对语料库进行 Ward 聚类算法生成层次聚类（hierarchical clustering）

(4)绘制 Ward 树状图（Ward dendrogram）

(5)利用 jieba+gensim 进行热度评估

3.2.1 转换向量空间

Gensim 用于从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达。它支持包括 TF-IDF，LSA，LDA，和 word2vec 在内的多种主题模型算法，支持流式训练，并提供了诸如相似度计算，信息检索等一些常用任务的 API 接口。

使用 gensim 做自然语言处理的一般思路是：

a.使用处理字典

b.生成处理语料库

c.自然语言处理

在处理文本时，用 TF-IDF 将文字转化为模型可以处理的向量，进行关键字抽取。字词的重要性随着它在某个留言主题出现的次数呈正比地增加，同时也随着它在其他留言主题中出现的频率呈反比地下降。其计算公式：

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

其中，

$tfidf_{i,j}$：指词 i 对文档 j 的重要程度；

$tf_{i,j}$ ：指词 i 在文档 j 中出现的次数占比。

对示例数据进行处理，可得到（30,357）的矩阵：

```
[[0.12428083 0.         0.06677403 ... 0.         0.22741243 0.        ]
 [0.         0.         0.         ... 0.09459323 0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 ...
 [0.03376451 0.         0.         ... 0.         0.         0.        ]
 [0.02416413 0.01590311 0.01298299 ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

图 5 向量矩阵

3.2.2 计算相似度

使用 cosine_similarity，计算余弦相似度，该原理在我们问题一的分析中有提及。

dist 变量被定义为余弦相似度。余弦相似度用以和 tf-idf 相互参照评价。被 1 减去是为了确保我稍后能在欧氏（euclidean）平面（二维平面）中绘制余弦距离。

$$\mathrm{dist}(A,B) = 1 - \cos(A,\ B) = \frac{\|A\|_2\|B\|_2 - A \cdot B}{\|A\|_2\|B\|_2}$$

其中，dist 取值范围为[0,2]。

```
[[ 1.11022302e-16  8.43703826e-01  9.02475136e-01  9.48473135e-01
   9.38944594e-01  8.27231866e-01  8.20206717e-01  6.67675350e-01
   9.04593369e-01  9.50249867e-01  9.39934255e-01  9.30253827e-01
   6.45571743e-01  8.39883552e-01  9.62097262e-01  9.24082514e-01
   9.30423453e-01  8.70113357e-01  8.99641062e-01  9.33260728e-01
   9.37960213e-01  8.24399064e-01  8.16421536e-01  7.80365442e-01
   5.32402915e-01  9.53467045e-01  6.15036087e-01  5.82187057e-01
   7.45982591e-01  8.26081246e-01]
 [ 8.43703826e-01  0.00000000e+00  7.33392799e-01  9.47168733e-01
   8.58358996e-01  8.99578186e-01  8.99820444e-01  9.23770090e-01
   8.84402370e-01  8.49876865e-01  9.24326133e-01  9.20790436e-01
   9.17104763e-01  8.52434474e-01  7.56653342e-01  8.98776645e-01
   8.38051114e-01  8.90867229e-01  7.94370582e-01  8.99229241e-01
   9.27059640e-01  8.60293179e-01  5.72956479e-01  8.20497993e-01
   9.13350873e-01  9.09637876e-01  8.78375268e-01  9.16114633e-01
   5.71532847e-01  9.31108389e-01]
```

图 6 dist 数据

3.2.3 层次聚类

进行层次聚类/凝聚聚类。Ward 聚类属于凝聚聚类算法，在每个处理阶段，聚类间两点距离最小的会被合并成一个聚类。

利用之前计算得到的余弦距离矩阵 dist 来计算 linkage_matrix。

3.2.4 可视化

通过上述原理我们可以将分类结果可视化处理，在这里我们绘制树状图，如下图：

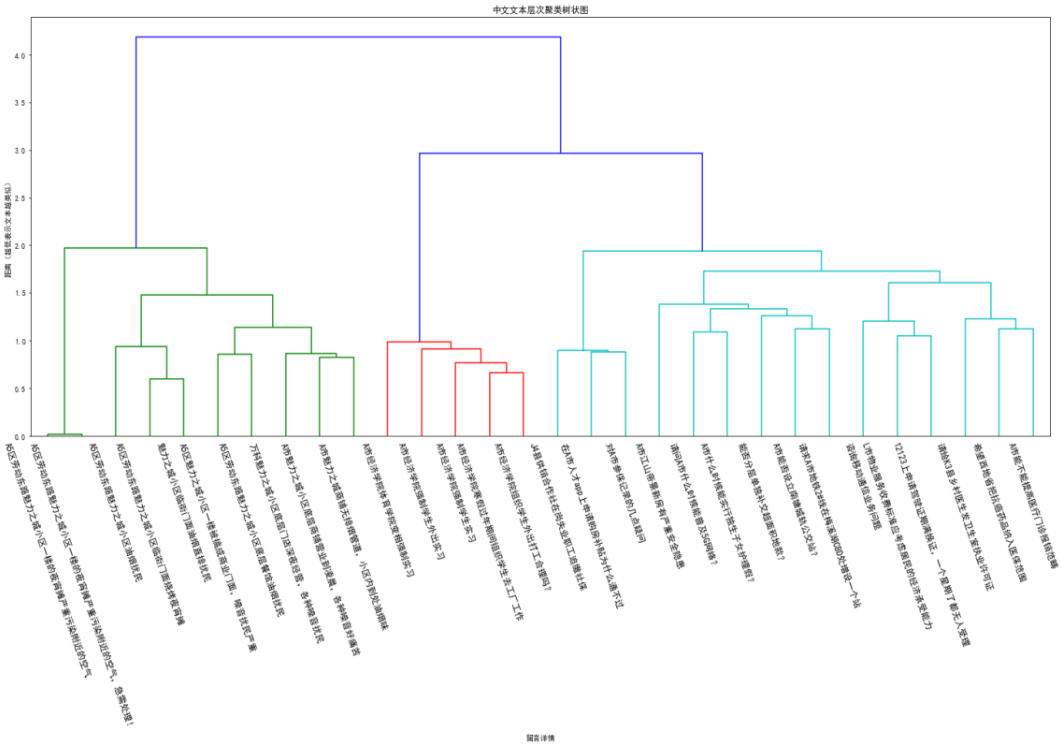

图 7 聚类树状图

3.2.4 热度分析

将距离低的几个问题归为一类问题，再使用 jieba 分词将同一类问题分割成词语，利用 python，使用以下命令

```
Dictionary = corpora.Dictionary(texts)
```

生成词典，并可以使用 save 函数将词典持久化。生成词典以后，将 python 中文档转化为向量形式。

下图是分词处理效果图：

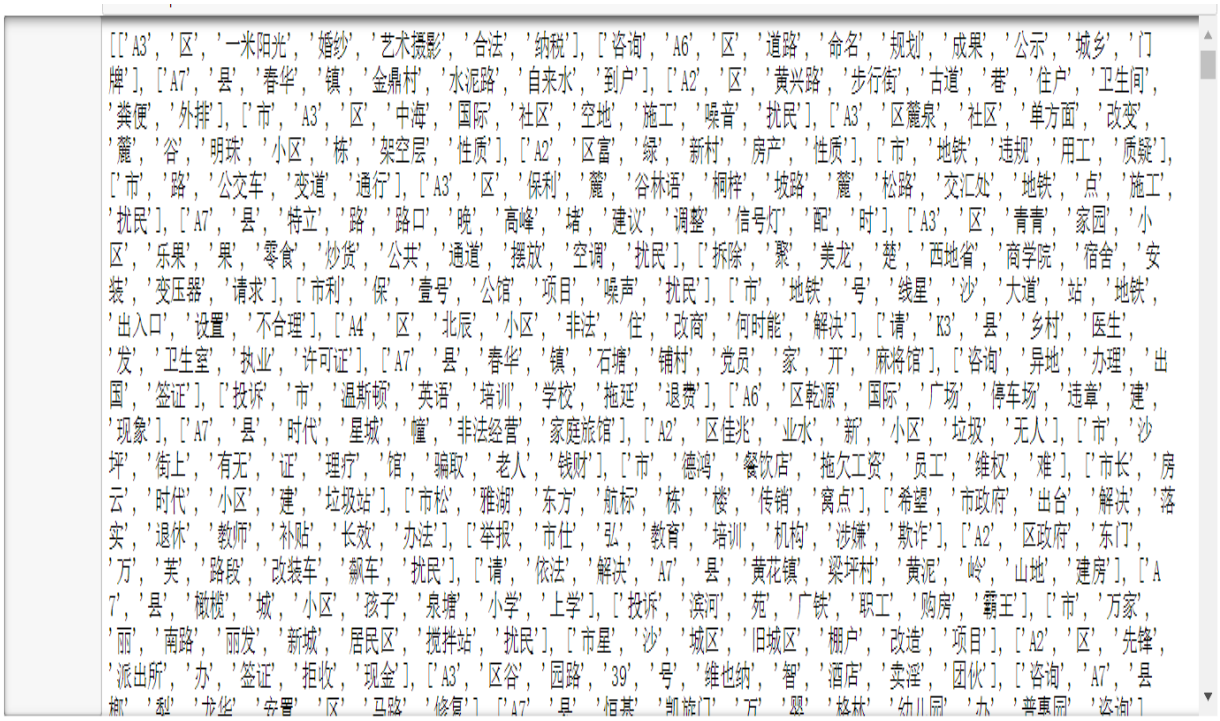

图 8 jieba 分词

在此基础上，再设立一个向量存放词语出现频率，作为热度评价指标。

算法实现介绍:

(1) 输入文件是 Excel,首先通过 pandas 获取 Excel 信息,通过 jieba 分词进行处理,jieba 分词要首先自定义词典以及排除信息,这样效果会差异很大,然后形成一个二维数组

(2) 使用 gensim 中的 corpora 模块，将分词形成后的二维数组生成词典

(3) 将二维数组通过 doc2bow 稀疏向量，形成语料库

(4) 将语料库计算出 Tfidf 值

(5) 获取词典特征数

(6) 计算稀疏矩阵相似度，建立索引

(7) 读取 Excel 行数据，通过 jieba 进行分词处理

(8) 通过 doc2bow 计算测试数据的稀疏向量

(9) 求得测试数据与样本数据的相似度

票牛A市分	51	票牛A市分公司不肯退我草莓音乐节的票钱怎么办_1378529e
寻找A7县	24	寻找A7县退伍人员的下落_68b843d
咨询A市潇	19	咨询A市潇楚卡异地优惠问题_69e9898
是谁将国	19	是谁将国企西地省巴士推向深渊_79aff1f
A市地铁四	15	A市地铁四号线何时会试运营_695ff1f
关于伊景	12	关于伊景园滨河苑捆绑销售车位的维权投诉_8858bc9
请A市坚决	12	请A市坚决取缔校园贷_547ff01
请加快对A	12	请加快对A市不合规网约车的治理_746ff01
建议在"我	11	建议在我的A市app中尽快接入法律服务的意见_99f89c1
老年人到A	11	老年人到A市落户是否可以享受A市城乡基本养老金_1031ff1f
A4区怡然	10	A4区怡然翠园退房事件怎么处理_706ff1f
A2区黄兴	9	A2区黄兴路步行街大古道巷住户卫生间粪便外排_9366392
A市人才购	9	A市人才购房补贴申请是否与单位注册地有关_913ff1f
请A市加快	9	请A市加快轨道交通建设力度_59b5ea6
请问A市各	9	请问A市各大人力资源服务中心的地址都在哪呢_91a5462
西地省科	7	西地省科技职业技术学院女生宿舍条件极差_8715dee
A市社保卡	7	A市社保卡办理能否加快速度_57f5ea6
A7县星沙	7	A7县星沙四区凉塘路的旧城改造要拖到何时_876ff1f
A市大学毕	6	A市大学毕业生买房却迟迟无法入住_6854f4f
请加快A市	6	请加快A市月亮岛片区公共服务力度_68b5ea6
西地省A市	6	西地省A市蜂投网涉嫌诈骗_5289a97
咨询A7县	6	咨询A7县开元互通人行天桥问题_62e9898
咨询婚姻	6	咨询婚姻法中规定的三代以内旁系血亲情况鉴定问题_98e9898
咨询A市转	6	咨询A市转业士官异地安置问题_60e9898
A5区半山	6	A5区半山一号期期延迟没办法管吗_747ff1f
请加快创	6	请加快创建A市食品安全城市建设步伐_7254f10
A7县楚龙	6	A7县楚龙路附近一餐馆的音乐声超级大_7475927

图 9 热度排行

最后根据热度排行最终确定前 5 个热点问题。

四、问题三的分析方法与过程

4.1 总体流程

对于答复评价，我们依据 SERVQUAL 理论，通过调查问卷的形式，让用户对答复的各方面质量进行评分，然后通过调查问卷，用户打分和综合计算得出答复

意见的质量评分。

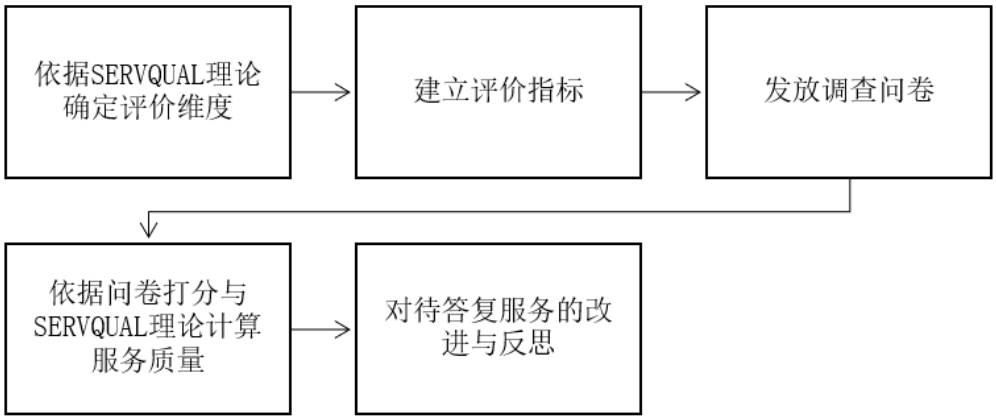

图 10 问题三建模方法的总流程图

4.2 具体步骤

4.2.1SERVQUAL 理论背景

SERVQUAL 是英文“service quality”的缩写，该理论是依据全面质量管理理论在服务业中提出的一种新的服务质量评价体系。

该模型的核心是差距理论，首先确定用户对该服务的期待值，然后调查顾客感知的服务效果，由此计算两者之间的差距，将此差距作为判断服务质量的依据，根据差距大小来判断服务水平的高低。公式表达如下：

$$SQ = P - E$$

式中，SQ=服务质量，P=用户感知的服务质量，E=用户对服务质量的期望值

SERVQUAL 模型衡量服务质量有五个尺度，分别是：有形性、可靠性、响应性、保证性和移情性。

a.有形性

指的是企业的物理设施和设备，以及人员形象等；

b.可靠性

指的是履行服务承诺的能力的可靠程度；

c.响应性

指的是在短时间内能够对用户进行帮助或者提供有效的服务；

d.保证性

指的是工作人员的知识水平、工作能力、服务态度等可以被依赖的能力；

e.移情性

指的是对待用户时，工作人员可以给予真诚的关怀和个性化的关注。

这五个维度下又可以展开多个小问题的指标，对于这些问题一般采用七分制，7 表示完全同意，1 表示完全不同意，借此来将用户对服务的感知值和期待值量化。维度及其相应指标如下图：

维度	项目	序号
有形性	有现代化的服务设施	1
	服务设施具有吸引力	2
	员工有整洁的服装和外表	3
	公司的设施与他们提供的服务相匹配	4
可靠性	公司所承诺的事情都能及时完成	5
	顾客遇到困难时，表现出关心并提供帮助	6
	公司是可靠的	7
	能准时提供所承诺的服务	8
	正确记录相关的服务	9
相应性	不能指望他们告诉顾客提供服务的准确时间	10
	期望他们提供及时的服务是不现实的	11
	员工并不总是愿意帮助顾客	12
	员工因为太忙以至于无法立即提供服务，满足顾客的需求	13
保证性	员工是值得信赖的	14
	在从事交易时顾客会感到放心	15
	员工是有礼貌的	16
	员工可以从公司得到适当的支持以提供更好的服务	17
移情性	公司不会针对不同的顾客提供个别的服务	18
	员工不会给予顾客个别的关怀	19
	不能期望员工会了解顾客的需求	20
	公司没有优先考虑顾客的利益	21
	公司提供的服务时间不能符合所有顾客的需求	22

图 11 SERVQUAL 模型量表

4.2.2 在答复质量评价中的应用

基于以上理论，我们建立了对答复质量的评价模型。在该模型中，我们从答复的有形性、相关性、可靠性、响应性和移情性这五各位对进行评价考量。具体分为 19 小项的评价标准，模型的维度与具体指标如下表所示：

维度	指标
有形性	1.问政平台功能全面
	2.网站设计美观合理
	3.页面可以正常访问，外部链接快捷有效
	4.服务功能清晰，便于查找
相关性	1.工作人员具有良好的相关知识
	2.回复内容与用户意见相匹配
	3.问题反馈及时有效
	4.针对不同意见进行个性化回复
可靠性	1.提供信息及时、完整、全面且充分
	2.问题回复完整合理，接受度高
	3.网站内容实时更新
响应性	1.回复速度快，令人满意
	2.能够提供准确的问题解决时间
	3.网站服务无时间限制
	4．准时兑现回复承诺

移情性	1.工作人员态度良好
	2.隐私保护工作到位
	3.平台提供的服务符合用户需求
	4.具有多种有效的咨询渠道

表 3 模型的维度与指标

围绕这些指标我们面向问政平台用户做一份调查，采用上述七分制评价方法，对回复服务质量的期待值与感知值进行调查，并对各项数据进行分析。具体比较的是各个指标的感知值和期待值以及二者之间的差距，求得各个维度和指标的得分平均值，并对其进行排序。

该模型下数据分析结果有三种：

a.当 P<E 时，感知服务质量小于期待服务质量，说明服务水平交叉，达不到与其标准；

b.当 P>E 时，感知服务质量大于期待服务质量，说明人们对服务质量的满意度较高；

c.当 P=E 时，感知服务质量与期待服务质量相等，说明总体服务令用户满意，是最理想的服务状态。

故我们不难看出，差距值越小，说明政务平台的回复服务质量越差。同时，我们也可以根据这个排序对用户期待高的服务方面提高重视，对服务质量较差的服务加以改进。

在此基础上，我们依据用户期待值，对各个维度在用户心中的重要程度进行评估，进而确定对其权重，在这里我们计划用乘积标度法对各个维度的权重进行计算。

运用上述数据，我们可以根据 SERVQUAL 模型对该问政平台的恢复服务质量进行综合评价，具体公式如下：

$$SQ = \sum_{K=1}^{5} W_k \cdot SQ_k$$

$$SQ_k = \frac{\sum_{i=1}^{m} \overline{P_i} - \overline{E_i}}{m}$$

式中，k表示维度的个数，W_k表示第k个维度的权重，m表示k个维度中指标的数量，$\overline{P_i}, \overline{E_i}$分别表示感知值和期待值的平均数。

根据上述公式，结合本文理论，我们可以计算出该问政平台答复服务质量值，并且可以通过各项指标的分数对问政平台的答复服务进行改进，更好的为公众提供给服务。

五、结论

对于问题一，经过对比，采用逻辑斯蒂分类器模型对留言进行分类，使用 F-Score 对分类方法进行评估，准确率达到了 86.97%，较为准确。解决了人工实现留言分类中工作量大、效率低，且差错率高等问题，效率得到了极大提升。

对于问题二，进行聚类分析，使用 jieba 分词与 gensim 结合的方法，确定热度评价指标与最终热度前 5 的问题，分别是：A 市分公司不肯退音乐节票钱、

寻找 A7 县退伍人员的下落、A 市潇楚卡异地优惠问题、是谁将国企西地省巴士推向深渊、A 市地铁四号线何时试运营，其热度分别为：51/24/19/19/15

热点问题	热度	留言
A 市分公司不肯退音乐节票钱	51	票牛 A 市分公司不肯退我草莓音乐节的票钱怎么办
寻找 A7 县退伍人员的下落	24	寻找 A7 县退伍人员的下落
A 市潇楚卡异地优惠问题	19	咨询 A 市潇楚卡异地优惠问题
是谁将国企西地省巴士推向深渊	19	是谁将国企西地省巴士推向深渊
A 市地铁四号线何时试运营	15	A 市地铁四号线何时会试运营

表 4 前 5 热点问题

对于答复质量的评价，我们在 SERVQUAL 模型基础上进行合理的指标设定，在网络问政平台上发放问卷十分便利，故我们依据这一便利条件可以轻松对获得答复的用户对该服务的评价。再根据数据处理和综合计算，就可以将答复质量用分数的形式量化，达到评价的目的。对于问政服务类平台，及时了解用户满意度，做出合理响应，努力建成真实可靠，令用户信赖的政府形象是十分有必要的。故该答复质量评价指标对问政平台的运作具有一定实用性，为健康的平台发展具有重要意义。

六、参考文献

[1]黄旭 基于机器学习的汉语短文本分类方法研究与实现
[2]彭湃 自然语言处理—中文词和短文本向量化的研究
[3]孔英会 景美丽 基于混淆矩阵和集成学习的分类方法研究
[4]陈春玲 吴凡 余瀚 基于逻辑斯蒂回归的恶意请求分类识别模型
[5]刘芳 政府门户网站的服务质量评价——以忻州市政府门户网站为例
[6]许月美 基于 SERVQUAL 模型的移动政务服务质量评价研究——以南宁市为例