“智慧政务”中的文本挖掘应用

摘要

“智慧服务”其实就是“互联网+政务服务”，政府利用互联网思维、技术和资源实现融合创新的过程,除了通过“连接”提升运作效率、服务能力，更重要的是通过“化学反应”和"基因再造”，重构流程，重塑公共及政务服务，实现政府服务体系的“升级和重塑”。在实现政府服务体系的升级和重塑过程中，群众的留言以及相关部门的回复是最直接有效的依据。所以本文主要通过对这些留言以及回复利用自然语言处理以及文本挖掘，来构造更好的服务平台。

首先对群众留言进行处理，选择逻辑回归模型对留言进行训练，将训练好的模型在利用 F-Score 对模型进行评价，之后利用 word2vec 和 Tfidf 特征值提取的方法实行聚类，找出群众留言中的热点问题，再通过相关部门的回复信息，以“相关性”、“完整性”、“可解释性”制订一套评估方案。

关键词：群众留言；文本分类；F-Score；模型评价；聚类

Abstract

"Intelligent Service" is actually "internet + Government Service" . The government uses Internet thinking, technology and resources to realize the process of integration and innovation, more importantly, through "chemical reaction" and "gene reengineering" , we can reconstruct the process, reshape the public and government service, and realize the "upgrade and remodeling" of the government service system. In the process of upgrading and reshaping the government service system, the message of the masses and the reply of

the relevant departments are the most direct and effective basis. So this paper mainly through the use of natural language processing and text mining, to build a better service platform.

Firstly, the mass message is processed, and the logistic regression model is selected to train the message, then F-Score is used to evaluate the model, and then the method of extracting the characteristic values of word2vec and TFIDF is used to cluster the model to find the hot issues in the mass message, and then a set of evaluation plan is made with "relevance" , "integrity" and "interpretability" through the response information of relevant departments.

Keywords: comments; text categorization; F-Score; model evaluation; clustering

目 录

一、问题重述

1、问题背景

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，给以往主要依靠人工来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。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。

2、解决问题

（1）对网络问政平台的群众留言利用一级标签建立分类模型，并对该模型利用F-Score 对模型进行评价。

（2）通过群众反映的热点问题，对这些热点问题进行分析，根据所定义的热度评价指标，筛选出排名前五的热点问题。

（3）对相关部门对留言的相关答复意见进行分析，并根据答复的三大特性对答复意见的质量给出一套评价方案。

二、问题一的分析与求解

问题一主要是通过群众留言建立分类模型，再对分类模型进行评价，建立分类模型之前需要对群众留言进行相关处理，对群众留言进行去重、分词、去停用词等的操作提取群众留言中的关键词，对处理完后的留言进行测试集、训练集的划分，挑选合适模型进行训练，训练完后对模型进行评价。

1、数据处理

对于长文本，在操作之前需要对数据进行预处理，查看群众留言可以看出有部分数据是重复的，则需要对这些数据进行去重处理，处理完后对文本进行分词、去停用词等常规操作。由于长文本的内容冗长且好多无用的词，所以需要对处理后的长文本进行关键词提取，及计算文本 TF-idf 值，的对处理完的留言按照 2:8 的比例进行训练集和测试集的划分，接着就可以进行模型的建立。

2、模型的构建

（1）逻辑回归模型

逻辑回归又称 logistic 回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。

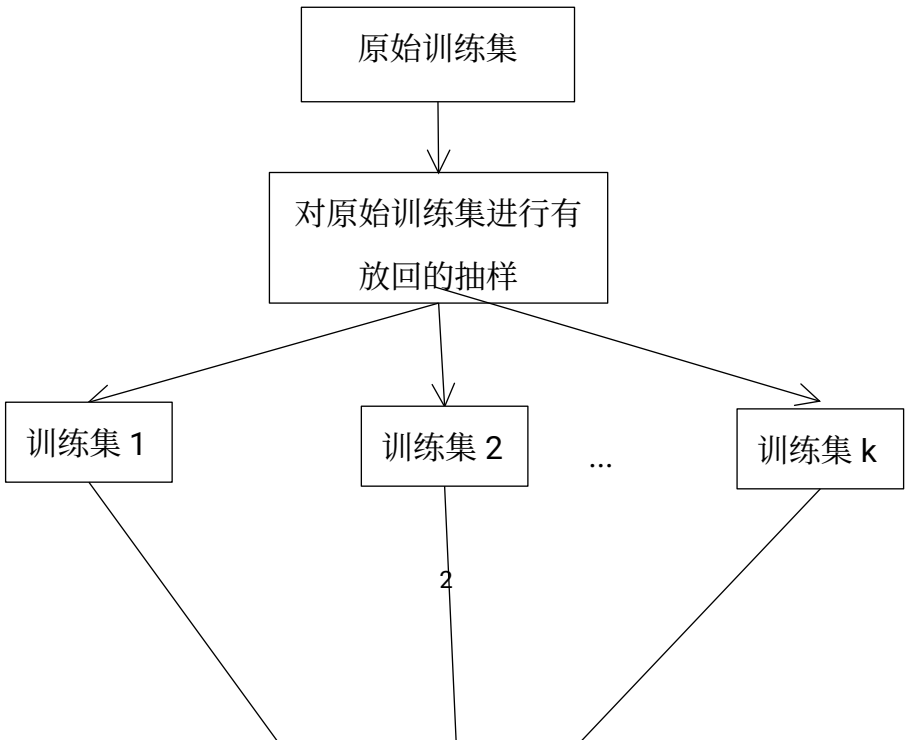
回归函数：

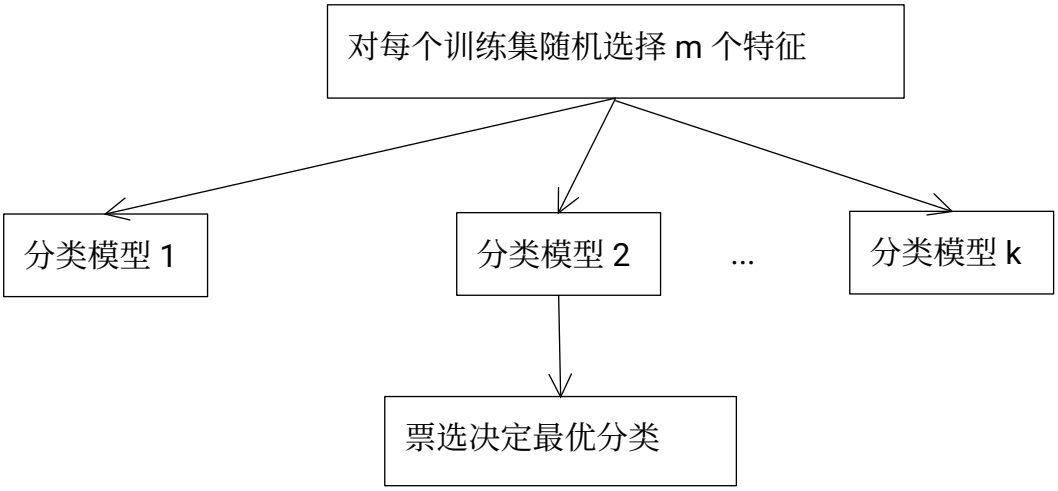
$$g(z)=\frac{1}{1+e^{-z}}$$

（2）随机森林模型

在机器学习中，随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。

随机森林的计算步骤：





通过使用逻辑回归模型以及随机森林模型对数据集进行训练, 逻辑回归模型训练完后的结果为 0.9102144674663759，随机森林模型的结果为 0.8720465285350781，可以逻辑回归模型的训练效果好于随机森林模型的效果，所以选择选择逻辑回归模型为最终模型。

3、模型评价

（1）计算混淆矩阵

在进行模型评价之前, 我们先对数据进行混淆矩阵的计算并绘制热力图如下图所示绘制混淆

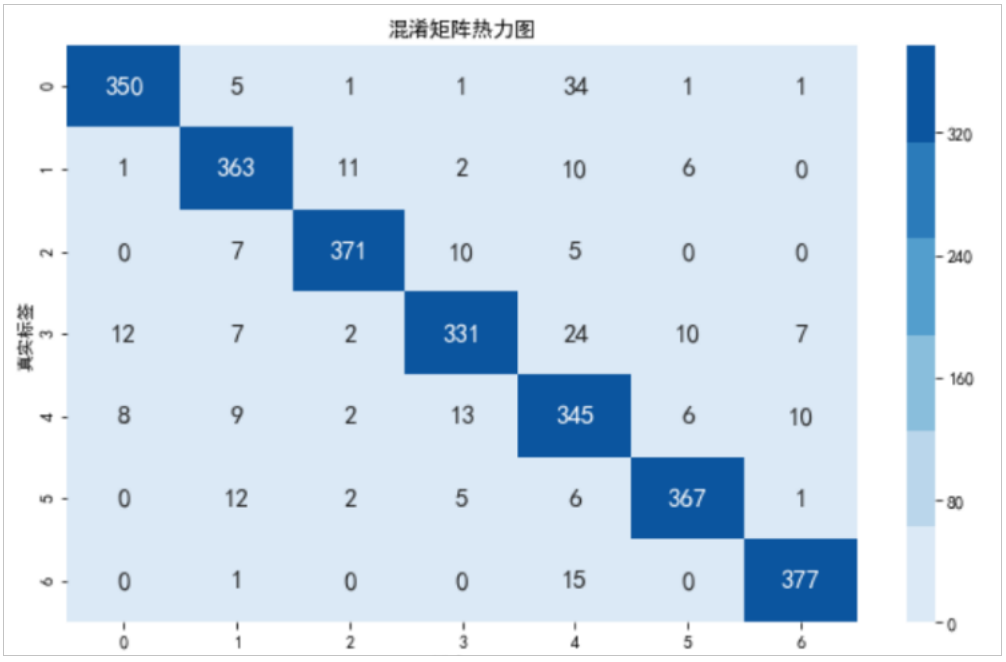


图 1 混淆矩阵热力图

（2）利用 F-Score 对模型评价

表 1 模型评价计算结果图

	precision	recall	f1-score	support
交通运输	0.94	0.89	0.92	393
劳动和社会保障	0.90	0.92	0.91	393
卫生计生	0.95	0.94	0.95	393
商贸旅游	0.91	0.84	0.88	393
城乡建设	0.79	0.88	0.83	393
教育文体	0.94	0.93	0.94	393
环境保护	0.95	0.96	0.96	393
avg	0.91	0.91	0.91	2751

从上表可以看出以及标签的平均精确度、召回率以及 F1 分数都为 0.91。

4、难点

（1）在进行训练集和测试集的划分时，训练集和测试集的列数不同，在计算 tf-idf 值之前应将训练集和测试集的列数统一。

（2）在进行模型训练的时候，由于数据间存在不平衡问题，对后期模型的训练和评价都有很大影响，所以在进行模型训练之前必须先对数据类别进行平衡处理来增加模型的准确率。

三、问题二的分析和求解

第二题对热点问题的挖掘，为了便于操作，我们对这 4326 条数据的“留言主题”内容进行了提取，生成包含 4326 条信息的列表。然后对每一条信息进行分词、去停用词处理等操作，将分好的词集中起来，进行 word2vec 词向量处理，训练词袋模型，运用词相似度计算得出与所需要的特点词相似的词语以确定大致的热

点问题的方向，然后利用 sklearn 中的 Tf-idf 对各条信息进行特征值提取，得到每一条信息的特征值后运用 sklearn.metrics.pairwise 中的 cosine_similarity 来计算各条信息之间的相似性，给出 dist = 1 - cosine_similarity(tf-idf_matrix)这样的公式计算信息之间的“距离”。然后进行聚类分类的操作，将文本相似度极高聚为一类，这样就做好了热点问题的提取，得到了一个距离的矩阵。随后运用可视化操作进行信息聚类的显示。最后按照题目的要求，整理好数目最多的前五名生成“热点问题表.xls”，以及“热点问题留言明细表.xls”。

1、文本提取

为了提取数据每一行内容大致反映了什么地点出了什么事件，“留言详情”内容太多，还是“留言主题”以简短的话语概括了所需要的内容，所以选择“留言主题”进行数据提取。运用 python 语句对各条“留言主题”进行提取，保存为 files。

2、分词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，本文采用 Python 开发的一个中文分词模块——jieba 分词，对之前所提取的“留言主题”进行中文分词。

jieba 分词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所构成的有向无环图 (DAG)，它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（即没有被收录在分词词表中但必须切分出来的词），采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。对提取的各条“留言主题”进行 jieba 分词后，保存到一个列表中。

3、去停用词

由于文本分词后会有很多功能及其普通、没有什么实际意义的词语或是标点符号，比如中文中的“我、你、的、他、她、是、了”等，英文中的“is、the、this、those、an、a、of” 等。对于停用词一般在预处理阶段就将其删除，避免对文本，特别是短文本，造成负面影响。本文所使用的停用词表“stoplist.txt”取自四川大

学机器智能实验室停用词表。

4、Word2Vec

Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

5、Tf-idf 特征值提取

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 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。除了 TF-IDF 以外，因特网上的搜索引擎还会使用基于链接分析的评级方法，以确定文件在搜寻结果中出现的顺序。

TF-IDF 的主要思想是:如果某个词或短语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。TFIDF 实际上是:TF * IDF，TF 词频(Term Frequency)，IDF 逆向文件频率(Inverse Document Frequency)。TF 表示词条在文档 d 中出现的频率。IDF 的主要思想是:如果包含词条 t 的文档越少，也就是 n 越小，IDF 越大，则说明词条 t 具有很好的类别区分能力。

词频(term frequency, TF)指的是某一个给定的词语在该文件中出现的频率，逆向文件频率(inverse document frequency，IDF)是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取以 10 为底的对数得到。

词频公式：$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$

逆文档频率公式：$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$，其中|D|：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$1 + |\{j : t_i \in d_j\}|$，然后$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$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。可以用 Tf-Idf 进行特征值的提取。

在 python 中运用 sklearn.feature_extraction.text 模块中的 TfidfVectorizer 最终提取出 4326 条留言主题中的 8145 个特征值。

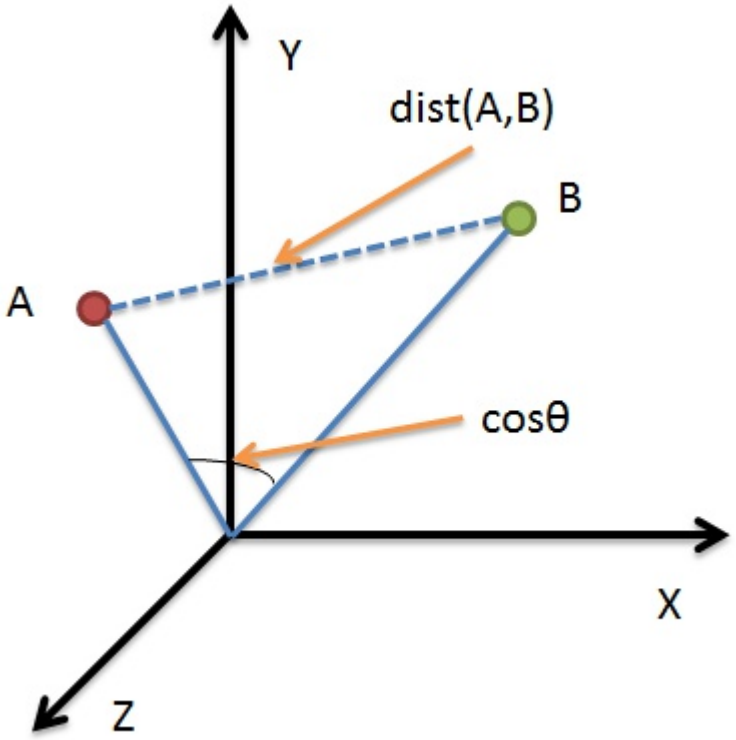
6、文档相似性计算

进行文本相似度计算的算法有很多, 这里我们选取了最常规的余弦相似度算法。

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。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，而非距离或长度上。公式如下所示：

$$sim(X,Y) = cos\theta = \frac{\vec{x} \cdot \vec{y}}{||x|| \cdot ||y||}$$

与欧几里德距离类似，基于余弦相似度的计算方法也是把用户的喜好作为 n-维坐标系中的一个点，通过连接这个点与坐标系的原点构成一条直线（向量），两个用户之间的相似度值就是两条直线（向量）间夹角的余弦值。因为连接代表用户评分的点与原点的直线都会相交于原点，夹角越小代表两个用户越相似，夹角越大代表两个用户的相似度越小。同时在三角系数中，角的余弦值是在[-1, 1]之间的，0 度角的余弦值是 1，180 角的余弦值是-1。



余弦相似度更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分用户兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦相似度对绝对数值不敏感）。

这里我们直接运用 python 中 sklearn.metrics.pairwise 模块中的 cosine_similarity 包进行余弦相似度的计算。

7、聚类

计算好文本之间的相似性后，需要对相似度高的文本进行聚类操作。

这里选取的是 Hierarchical Clustering 层次聚类法。

Hierarchical Clustering(层次聚类)：就是按照某种方法进行层次分类，直到满足某种条件为止。

主要分成两类：

a）凝聚：从下到上。首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有的对象都在一个簇中，或者某个终结条件被满足。

b）分裂：从上到下。首先将所有对象置于同一个簇中，然后逐渐细分为

越来越小的簇，直到每个对象自成一簇，或者达到了某个终止条件。（较少用）

算法步骤：

a）将每个对象归为一类，共得到 N 类，每类仅包含一个对象．类与类之间的距离就是它们所包含的对象之间的距离．

b）找到最接近的两个类并合并成一类，于是总的类数少了一个．

c）重新计算新的类与所有旧类之间的距离．

d）重复第 2 步和第 3 步，直到最后合并成一个类为止(此类包含了 N 个对象)．

在 python 中调用 scipy.cluster.hierarchy 模块中的 ward, dendrogram, linkage 包，最后得到一个联系矩阵，部分如下图所示：

```
[[4.34000000e+02 3.61800000e+03 0.00000000e+00 2.00000000e+00]
 [7.13000000e+02 7.69000000e+02 0.00000000e+00 2.00000000e+00]
 [3.53100000e+03 4.12000000e+03 0.00000000e+00 2.00000000e+00]
 ...
 [8.64300000e+03 8.64700000e+03 5.78968835e+00 4.07700000e+03]
 [8.61100000e+03 8.64600000e+03 6.33335956e+00 2.49000000e+02]
 [8.64800000e+03 8.64900000e+03 6.44835011e+00 4.32600000e+03]]
```

通过可视化操作输出一个联系绘图：

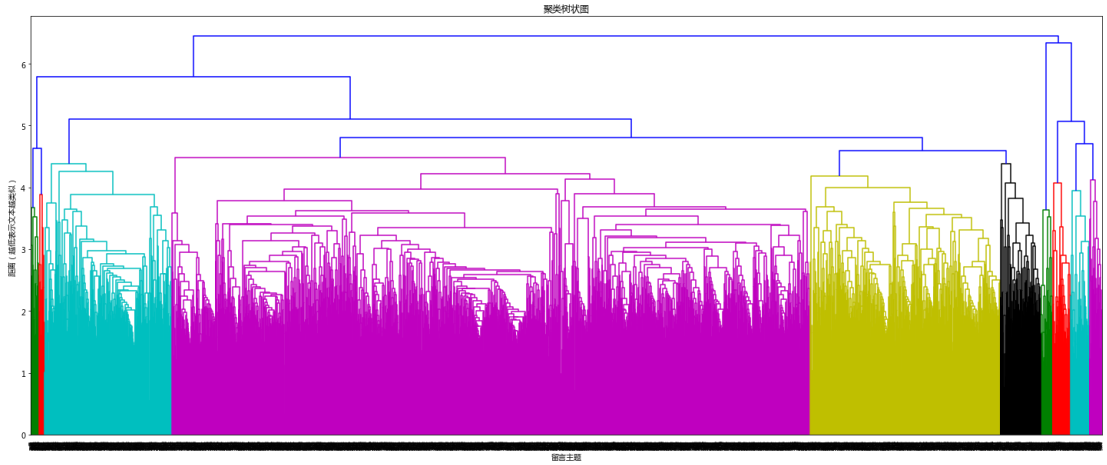


图 2 聚类网状图

8.难点

这道题的难点在于原始数据有 4000 多条，其中反映的具体问题形形色色，一个地点甚至可能对应很多个不同的问题，这就对聚类和分类造成了很大的困难。所以，我们需要在分词时对这些特定的地点词，比如：丽发新区、魅力之城、滨河苑等进行词典编辑，能够把这些地点词汇分好，然而，这类的词语是在是太多了，有着较大的困难。

四、问题三的分析和求解

1. 评价方法

问题三主要是需要建立一套完整合理的答复意见评价方法。我们需要对问题答复的多个角度分别对数据进行分析，选取多个不同的指标进行相关的计算，最后对各角度得到的结果进行加权计算得出最终的结论。

2. 主要参数及计算方法

（1）相关性：相关性主要考察答复的内容是否与群众提出的问题相关。计算两个文本的相关性，我们可以对留言主题和答复意见分别进行分词和去停用词，得到两个文本的关键词。然后我们可以计算留言主题中词语在答复意见去停用词后得到的词语中出现的个数，并最终用公共词语的个数除以留言主题分词后得到的个数得出相关性。

（2）完整性：完整性主要考察答复意见是否很礼貌恰当的回答了群众提出的问题，一般问题回复都满足某种语言规范，答复意见首先应该明确问题，然后提出解决方法，最后展示解决效果。除此之外开头应该是问题答复群体，最后应该是祝福语，所以我们可以将答复意见分成这五个方面，每个部分都有代表性的词语，比如说“您好”，“解决方法”，“祝您”等等，我们根据这些部分出现的数量计算完整性。

（3）可解释性：可解释性主要考察答复意见中内容的相关解释，考察答复意见是否合理。可解释性评价了答复意见的实用程度，对评价答复意见的质量评价很重要。

计算方法：对前面的三个指标得出的数据按照 3：3：4 的权重进行加权，得出最终的结果。

3. 难点

（1）计算答复意见的完整性时很难对答复意见的文本进行合理的分割。我们是根据一些特殊的词语进行分割，但是有的词语会重复出现，而有的不会出现。

（2）对可解释性的计算缺乏指标以及参数。可解释性考验问题回复的精准实用性，通过深度学习很难做到。

五、总结

本文从三个模块解决所需解决的三个问题。

第一是留言的分类问题。先是对原始数据进行分词、去停用词处理，运用 Tfidf 方法划分训练集测试集并构建了权值矩阵，随后就是模型的构建了，在随机森林模型、贝叶斯模型和逻辑回归模型的训练中最终选取了效果较好的逻辑回归模型。接着进行模型评估，得到所需的 F-Score。

第二是热点问题的挖掘。运用到了 Word2vec 词向量化，使用 Tfidf 进行特征值的提取，随后用余弦相似度算法对文本间相似度进行计算，最后用到 Hierarchical Clustering 聚类法，提取得到排名前五的热点问题。

第三是答复意见的评价。通过相关性、完整性、可解释性三个角度，给出适当的评价方案。

参考文献：

[1]https://blog.csdn.net/ustbbsy/article/details/80960652

[2]https://blog.csdn.net/rachel715/article/details/51700931

[3]https://www.jianshu.com/p/705b3776a808

[4]https://baike.so.com/doc/433640-459181.html

[5]https://www.sohu.com/a/279136744_163476

[6]秦彩,杰管强.一种基于 F-Score 的特征选择方法[J].宜宾学院学报,2080(6):4-6