基于自然语言处理技术的群众留言分析与挖掘

摘要

近年来，随着大数据的发展，政府管理已经向数字化转型。网络问政是政民互动的一个平台，群众的随时随地留言会产生大量的数据信息。如何快速对留言进行划分和热点问题整理，成为提高政府治理效率的关键！本文基于自然语言处理技术对群众留言分析与挖掘，并解决以下问题。

问题一，目的：按照附件 1 里的标签分类对附件 2 的群众留言进行分类并对分类器进行评估。方法：基于 $pycharm$ 平台，首先对留言详情和留言主题整合为一个文本，然后对文本进行预处理（去特殊字符、去停用词）；第二，基于清洗的文本进行分词并利用 $TF-IDF$ 算法得出文本词的特征；第三，针对特征值对构建朴素贝叶斯、线性分类支持向量机、逻辑回归、随机森林分类模型并对模型训练；第四，模型评估并选择准确率最高的分类模型进行应用。结果：四个分类模型中线性分类支持向量机分类模型准确率最高，$F_1 score$ 值达到 0.94。

问题二，目的：对附件 3 进行分析，得出存在群众中的热点问题。方法：首先基于文本预处理后的数据利用正则表达的规则对群众留言信息的地区/人群提取后，载入 $jieba$ 分词的词典，然后分词、去停用词，利用 $word2vec+idf$ 算法进行特征提取，并根据特征对留言数据进行类的划分。第二，基于划分好类别的留言数据建立热度指数模型，设定相同类别下每增加一条的留言数量加 10 分、留言文本若有点赞/反对数有几票加几分、时间与现在的时间间隔越短时热度值也加一分。结果：已知数据带入热度指数模型后计算出热度指数，并对热度指数进行排序。

问题三，目的：从答复的相关性、完整性、可解释性三个维度建立答复的质量评估模型。方法：首先，答复相关性，设定占比 50%：基于 $TextRank$ 算法对留言和答复关键词提取后并计算关键词的匹配个数，以匹配个数在所有留言关键词数的占比乘以答复相关性的权重；第二，答复完整性，设定占比 30%，以富信息评论标准打分；第三：答复可解释性，设定占比 20%，以答复和留言的时间间隔进行打分；最后，综合这三个维度的分数。结果：见表 3-评价方案评分表。

关键词：分类模型；$TF-IDF$;热点问题；文本聚类；打分模型

Mass Message Analysis and Mining Based on Natural Language Processing Technology

Abstract

In recent years, with the development of big data, government management has transformed into digital. Internet questioning is a platform for the interaction between the government and the people, and the masses' message will generate a lot of data information anytime, anywhere. How to quickly divide the message and sort out the hot issues has become the key to improving the efficiency of government governance! Based on natural language processing technology, this article analyzes and mines the mass message and solves the following problems.

Question 1, Purpose: To classify the messages of the masses in Annex 2 according to the label classification in Annex 1 and evaluate the classifier. Method: Based on the platform, first integrate the message details and message subject into one text, and then preprocess the text (remove special characters, stop words); second, segment the words based on the cleaned text and use the algorithm to derive the text word The characteristics of the third; for the eigenvalue pairs to build naive Bayes, linear classification support vector machine, logistic regression, random forest classification model and model training; fourth, model evaluation and select the classification model with the highest accuracy for application. Results: Among the four classification models, the linear classification support vector machine classification model had the highest accuracy rate, with a value of 0.94.

Question 2: Purpose: To analyze Annex 3 and find out the hot issues existing among the masses. Method: First, based on the text pre-processed data, use regular expression rules to extract the area / crowd of the people's message information, load the word segmentation dictionary, then segment the words and remove the stop words, use the algorithm to perform feature extraction, and according to the feature pairs Message data is divided into categories. Second, based on the classified message data to establish a heat index model, set the number of messages added 10 points for each additional message under the same category, if the message text is a bit like / disagree, there are a few votes plus points, time and current time When the interval is shorter, the heat value is also increased by one point. Results: After the known data was brought into the heat index model, the heat index was calculated and the heat index was sorted.

Question 3: Purpose: To establish a quality assessment model for responses from three dimensions: relevance, completeness, and interpretability. Method: First, the relevance of the reply, set to account for 50%: after extracting the message and reply keywords based on the algorithm and calculating the number of keyword matches, the proportion of the number of matches in the number

of keyword keywords in all messages is multiplied by the reply The weight of relevance; second, the completeness of the reply, set at 30%, and scored with rich information review standards; third: the interpretability of the reply, set at 20%, and scored at the time interval between reply and message ; Finally, integrate the scores of these three dimensions. Result: See Table 3-Evaluation Program Scoring Table.

Keywords: classification model; hot topics; text clustering; scoring model

目录

1 引言

1.1 问题背景

近年来，政民互动的微信、微博、市长信箱、阳关热线等网络问政平台产生大量数据信息，在数据即是“石油”的时代，如何从中挖掘出具有价值的信息，从而提高政府的治理方式和治理的效率是一个值得探索的问题，而基于自然语言处理技术建立智慧政务系统已经是社会治理创新发展的新趋势，对政府的管理水平和管理效率具有极大的推动作用。

1.2 问题重述

我们主要对四个 excel 表数据进行分析和挖掘：群众的留言三级分类标签附件 1、群众留言信息和留言对应的标签表附件 2、群众留言信息和对应留言的点赞数和反对数附件 3、群众留言信息和政府答复信息附件 4。

问题一：按照附件一的三级标签类别对群众留言进行划分，并把群众留言分派至相应的职能部门，即是建立分类模型、模型评估两个过程。

问题二：对附件 3 的留言数据进行处理，从某一时间、特定地点/特定人群中发现群众关心的热点问题，并把前 5 个热点问题按照表 1 给出的格式进行输出展示。

问题三：对附件 4 的数据进行处理，从答复的相关性、完整性、可解释性三个维度建立答复意见质量的评分模型。

1.3 问题分析

问题 1 目的是对附件 2 里的对留言文本数据建立一级标签分类模型，并对分类模型进行评估。文本分类模型的构建需要训练数据，训练的数据需是词向量化后的数据。因此需要对文本进行分词、去停用词等预处理操作后，把分好的词基于向量转化模型进行向量化后，用于模型训练，最后利用划分的测试集对模型进行测试，根据此测试结果可以做混淆矩阵查看预测的误差，并整合查全率和查准率计算 F_1Score 值对模型进行评估。

问题 2 目的是发现热点问题。热点问题应是在某段时间出现的频率高而在之前很少出

现的问题[1];并且针对某一地点、某一人群的主题数量尽可能的多；时间间隔小、地点/人群出现的频率大时，热度指数就越高，针对这个性质，在建立热度指数的模型时，对每一条留言评论热度加分/减分。除此之外，附件 3 里大部分有反对数多的留言评论点赞数也多，不论反对和点赞数，它们都说明了群众有关心此类问题，因此有出现点赞数和反对数的留言都需要对热度指数模型上加分。最后基于热度指数排序，就能得出排名前 5 的热点问题。

问题 3 目的是从答复的相关性、完整性、可解释性三个维度针对每一条答复留言建立评分模型，由模型得出每一条答复的一个综合得分，分值越高、答复质量越好。答复的三个维度中，权重占比最大的是答复相关性，只有相关性存在时，才能定义可解释性和完整性。针对附件 4 表里的属性，答复时间与留言时间的间隔可以作为可解释性的参数，时间间隔越大，不仅影响群众的情绪，并且要解决的问题可能已经变质或变性。答复的完整性应该从群众留言的问题数量与答复数量的对应数量来定义，但是群众留言中一般是反应问题、而不是以问句的形式出现，所以可以由答复文本长度解释完整性，也就是由富信息评论，文本长度越长的答复被认为包含有更多的信息和更细致的描述。

2 模型假设和使用符号

2.1 模型假设

为了简化我们的模型，我们在挖掘群众留言具有价值信息中做了以下假设。

假设 1：所有的群众留言信息之间是独立的。

假设 2: 所有的群众留言、留言点赞数、留言反对数都是他们对现状真实的感受，没有恶意和无聊的留言。

假设 3: 一级标签特征值之间都彼此独立。

2.2 符号使用

表 1 符号意义说明

符号	意义	单位
$TF_{i,j}$	关键词 j 在文档 i 中出现的频率	百分比

$n_{i,j}$	关键词 j 在文档 i 中出现的次数	次
IDF_i	关键词 i 的逆文档频率	百分比
N	文档总数	个
$N(i)$	包含关键词 i 的文档总数	个

2.3 实现平台

基于 *Pycharm* 平台

3 模型建立与求解

3.1 问题一：留言内容的一级标签分类模型

分类模型的数据来源是附件 2 里 9210 条的数据，并且包含留言编号、留言用户、留言主题、留言时间、留言详情、一级标签 6 个属性。每一个类别下的留言数量如下图 1 所示：

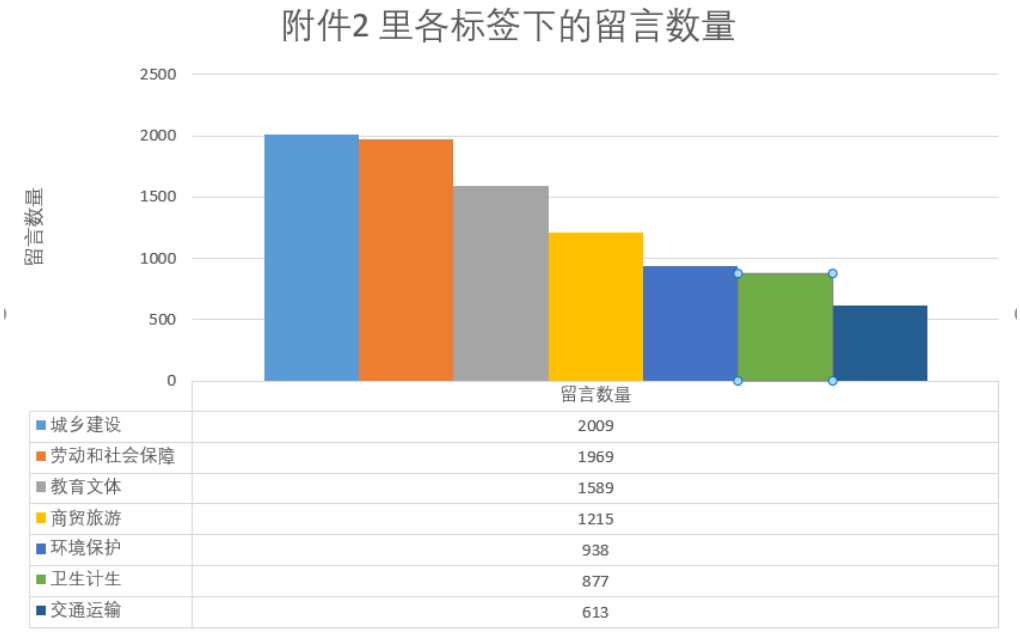


图 1 标签类别留言数目

3.1.1 留言文本的预处理

（1）模型训练数据来源处理

分类模型构建的一个重要的问题是训练数据，也可以说成语料库的选取，并且通常是训练的数据越多越好，可以提高模型的准确率。因此，我们把附件 2 里的留言主题和留言详情内容进行整合，基于整合文本再进行分析，以下文章中所提到的文本均是整合后的文本，流程如下图 2 所示：

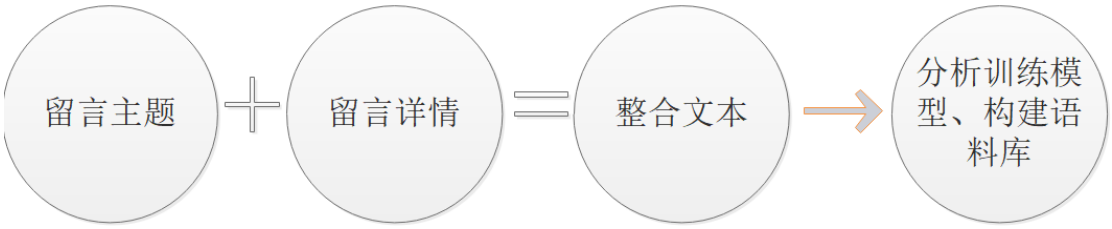


图 2 预处理第一步

（2）去除文本中的特殊字符和停用词

基于整合后的文本，还需去除对留言分类没有贡献的字符，例如：文本排版时的换行符（\n）或者制表符（\t）；还有文本中的各类标点符号，例如：%?!*/#&:，。？！+$()<>\\等。除了各类标点，还需过滤掉存在停用词库中的词语和词长小于 1 的词，以提高分类效率和节省存储空间[2]。

（3）中文分词

词是最小的表达语义的单位，词是文本分析的关键一步，利用 *jieba* 中文分词工具对留言文本进行分词，流程见下图 3，结果见某附件 。

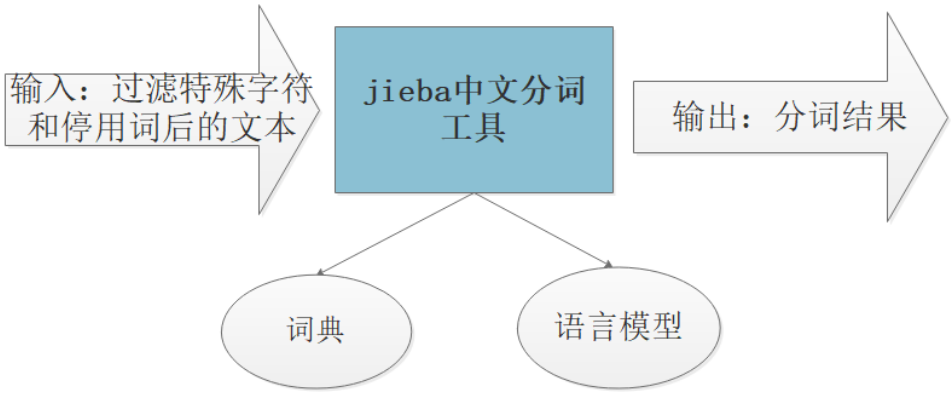


图 3 分词器示意图

分词后，分别提取各类别标签下的词频前 100 的词，利用 *wordcloud* 函数绘制词云如下图 4 所示：

图 4 各类标签的词云图，词云中词的尺寸越大、颜色越深，则词的频率越高，与标签的关联度越大

（4）留言特征提取

群众的留言文本的数据集合是自然语言所构成的，采用 $TF-IDF$ 算法，对分词后的文本进行处理，使字符转化为数字化的结构型数据进行特征挖掘[3]。

$TF-IDF$ 是单词权重的统计方法。TF 表示的是某一个词在文本中出现的次数，即词频。若仅仅用 TF 来评估一个词的权重值大小是没有说服力的，因为中文文本中有大量的“的”、“和”、“它”、“他”、“她”等等词；IDF 是逆文本频率，表示的是某词在文

本中出现的频率取反后的值，也是该词项区别于其他文本的能力。因此，对某词的或者某文本$TF-IDF$向量化后，可以利用此特征进行建立一级标签分类模型。

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{i,k}} \quad \text{公式(一)}$$

$$IDF_i=\log\frac{|N|}{|N(i)|+1} \quad \text{公式（二）}$$

$$TF-IDF=TF*IDF \quad \text{公式（三）}$$

由公式（三）得出的数值才能评估一个词的重要性。

基于 sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取文本的$TF-IDF$特征值，并以卡方检验这一统计学工具，找出每个标签分类中关联度最大的词语和者词语对。这些词语和词语对在构建模型的时候可以增加标签的特征，提高分类器的准确率。下表 2 是每一个标签类关联度最大的词语和词语对：

表 2 标签类别关联词语、词语对

标签类别	关联度最大的词语	关联对最大的词语对
0（城乡建设）	业主；小区	[房产-房产证]；[开发-开发商]
1（环境保护）	环保；污染	[养猪-猪场]；[环保-环保局]
2（交通运输）	租车；出租车	[租车-出租车]；[出租-租车]
3（教育文体）	学校；教育	[幼儿-幼儿园]；[教育-教育局]
4（劳动和社会保障）	劳动；社保	[老金-养老金]；[养老-养老保险]
5（商贸旅游）	传销；电梯	[小区-电梯]；[传销-组织]
6（卫生计生）	医生；医院	[子女-独生子]；[独生-生子]

（5）词向量化

首先，把留言文本转换为词频向量；第二，基于$TF-IDF$算法进行特征处理，把词向量转换为$TF-IDF$向量。（9210 条留言, 170160 词语）转换为$TF-IDF$特征值部分信息如下图 5 所示：

(0, 3459)	0.2802749067661214	(0, 6958)	0.34950247038547866
(0, 6962)	0.2884876633339142	(0, 11681)	0.08885015063386047
(0, 3938)	0.17350946740507942	(0, 3949)	0.11353791383897519
(0, 11721)	0.16270839201132264	(0, 3925)	0.21689689008070798
(0, 3984)	0.17350946740507942	(0, 11623)	0.40198788407389346
(0, 5128)	0.16270839201132264	(0, 7031)	0.21926516705108973
(0, 3934)	0.17350946740507942	(0, 5113)	0.18741393058604577
(0, 11630)	0.21689689008070798	(1, 10602)	0.3818244183480896
(0, 11637)	0.21689689008070798	(1, 7907)	0.3818244183480896
(0, 7081)	0.3254167840226453	(1, 3941)	0.3818244183480896
(0, 5140)	0.15504490706071386	(1, 11116)	0.2468443259739949
(0, 3431)	0.1901467069892193	(1, 10568)	0.22925186186307506
		(1, 7900)	0.2874773825960924

图 5 部分留言 $TF-IDF$ 特征值

3.1.2 一级标签分类模型的构建

基于 $TF-IDF$ 词向量数据，进行模型训练集和测试集的划分：80% 的数据作为训练集、20% 的数据作为测试集。

（1）朴素贝叶斯分类器的构建

留言文本的数据集 $D=\{d_1,d_2,d_3,...,d_n\}$，留言文本数据的特征属性集 X 为 $TF-IDF$ 向量集，标签分类变量为 $Y:\{0,1,2,3,4,5,6\}$，留言文本可以按照标签分类变量进行分类。根据朴素贝叶斯算法得，后验概率可以由先验概率和分类条件概率给出：

$$P_{prior}=p(Y) \qquad \text{公式（四）}$$

$$P_{post}=p(X\mid Y) \qquad \text{公式（五）}$$

有公式（四）和公式（五）得出后验概率：

$$P(Y\mid X)=\frac{P_{prior}(Y)P_{post}(X\mid Y)}{P(X)} \qquad \text{公式（六）}$$

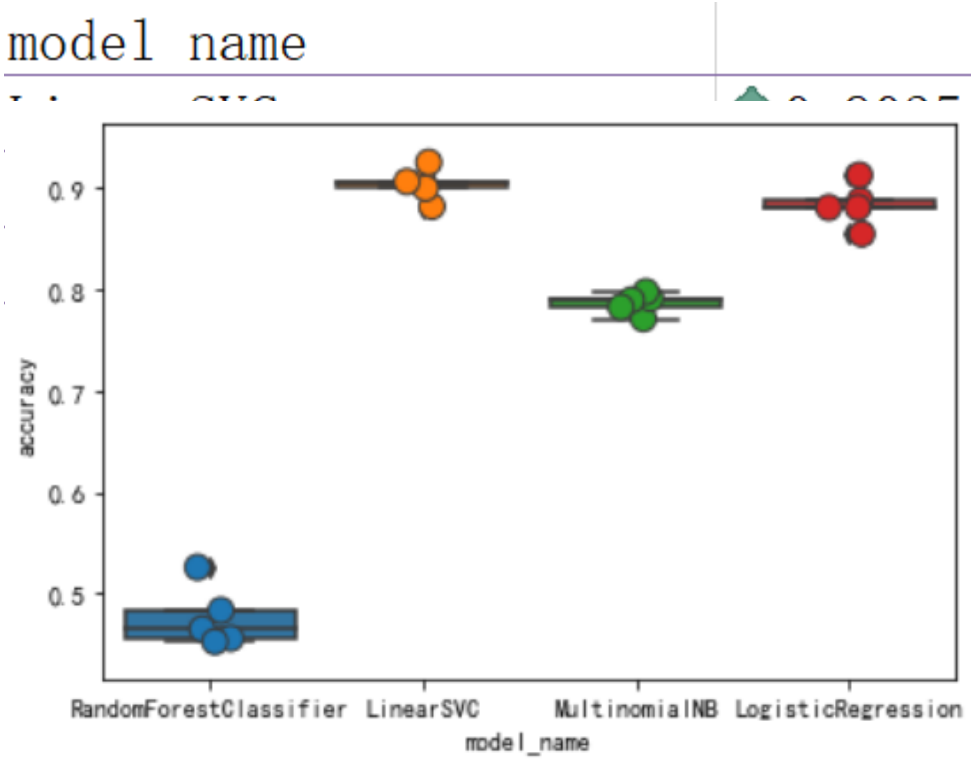
并且朴素贝叶斯各特征间独立，在给定标签类别为 y 的情况下，公式（六）还可以表示成：

$$P(X\mid Y=y)=\prod_{i=1}^{d}P(x_i\mid Y=y) \qquad \text{公式（七）}$$

综上，后验概率为：

$$P(y_i\mid x_1,x_2,...,x_d)=\frac{P(y_i)\prod_{j=1}^{d}P(x_j\mid y_i)}{\prod_{j=1}^{d}P(x_j)} \qquad \text{公式（八）}$$

除此模型外，调用 *sklearn* 库中的线性分类支持向量机（ *LinearSVC* ）函数、逻辑回归（ LogisticRegression ）、随机森林（ RandomForestClassifier ），也进行分类器的构建。对四个分类器的准确率比较，四个分类器中线性分类支持向量机的准确率＞逻辑回归准确率＞朴素贝叶斯分类器准确率＞随机森林分类器的准确率，四个分类器具体的准确率如下图 6



通过对四个分类器的准确率的箱线图分析，得出的结果可以更加直观的看出：四个分类模型中最优的模型，箱线图如下图 7 所示：

线性分类支持向量机是四个分类器中最优的分类模型，因此我们的留言分类基于线性分类支持向量机模型实现，分类器模型评估也是对线性分类支持向量机的评估。

3.1.3 模型的评估

（1） F_1Score 评估

题目基于 F_1Score 对模型进行评估， F_1Score 是查准率和查全率的调和值。 F_1Score 定义如下：

图 7 各分类模型的准确度箱线图

$$F_1 = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{2P_i R_i}{P_i + R_i} \qquad \text{公式（九）}$$

基于 *Python* 软件对查全率（召回率 Re*call*）、查准率（精确率- Pr*ecision*）带入 $F_1 Score$ 模型进行求解，得出各一级分类标签下的 F_1 的分数值及支持度，结果见下：

标签类别	查准率	查全率	F1 分数值	支持度
城乡建设	0.91	0.94	0.93	402
环境保护	0.96	0.96	0.96	187
交通运输	0.94	0.89	0.92	123
劳动社会保障	0.95	0.97	0.96	394
商贸旅游	0.94	0.89	0.91	243
卫生计生	0.95	0.94	0.95	175
平均 $F_1 Score$	**0.94**			

表 3 标签类别分类模型评估

由上表得出，构建的一级标签分类模型 $F_1 Score$ 平均值为 0.94，也就是说我们构建的分类模型在对 100 条留言进行分类时，大约有 94 条可以准确分到正确的标签类别下。

（2）混淆矩阵进行误差分析

由实际结果数和预测结果数利用中 confusion _ matrix 函数得出混淆矩阵，混淆矩阵主对角线代表预测正确的数量，除主对角线以外的格子中的数目就是预测错误的数量。各标签分类的混淆矩阵如下图 8 图 8 所示：

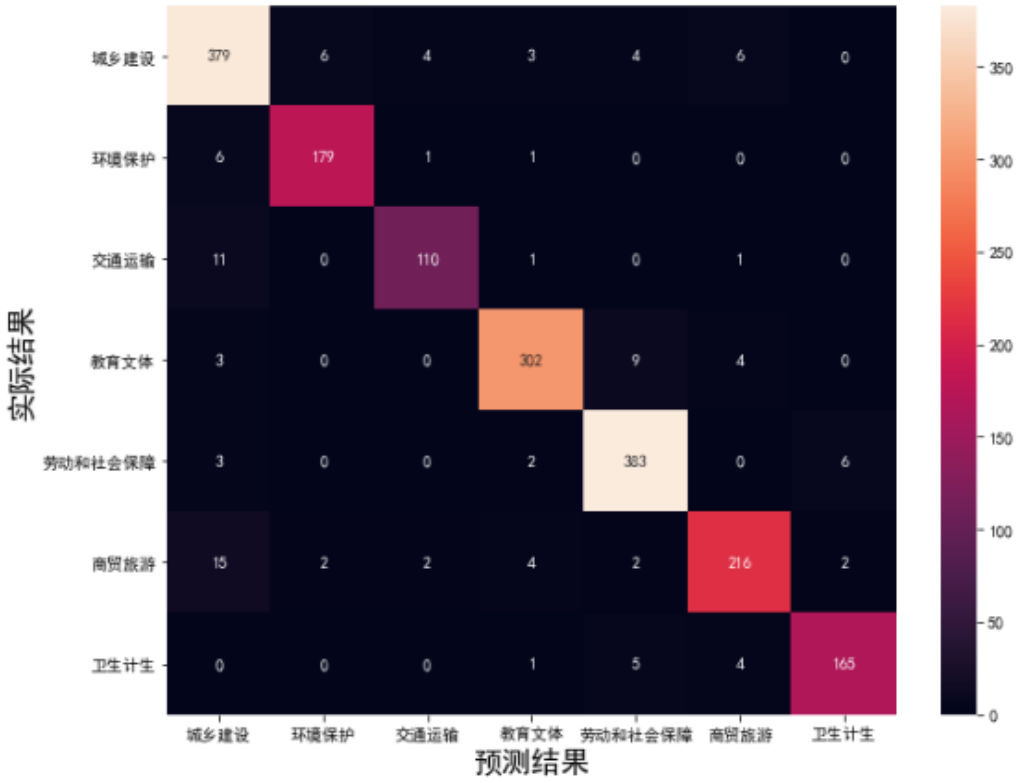


图 8 混淆矩阵：从上面的混淆矩阵可以看出"环境保护"、"交通运输"、"商贸旅游"类预测最准确,无一例预测错误。“城乡建设”和“教育文体”、"劳动和社保问题"有一定量预测的错误。

3.2 问题 2：发现热点问题

3.2.1 模型建立

一个热点问题，应是在某一段时间内某一地区的/某类人群，所关心/所反应的问题尽可能多。针对时间、地点/人群、问题数量三个热度评估的三要素建立一个热度指数评估模型，以热度指数来对热点问题排序，最后得出排名前 5 个热点问题。热点问题发现数据来源是附件 3 里的 4326 条数据。

（1）文本聚类——找出相似的问题

首先，由于问题发现和地点人群的发现只需要对它们进行处理，因此提取出附件 3 表

里的留言主题和留言详情属性的数据。

第二，数据预处理。分词、去特殊字符、去停用词（空格也作为停用词）。

第三，对附件 3 中的留言主题和留言详情进行分析，地点中的市、区、县是以 $A\sim Z$ 和 $0\sim 9$ 的形式出现，例如：{A3区，A7县，C5市}。因此基于正则表达式，对留言主题里的地点进行提取，提取的信息记到一个地点的属性类别下。若留言主题里没有地点的信息，则再去留言详情里以正则表达式提取出地点信息，再把提取的地点信息放到地点这一属性类别下。然后对地点进行去重操作后，把地点这个属性数据添加到 *jieba* 分词工具中。这样，分词后的数据中词向量化时地点也是一个特征项。

第四，词向量化。分别基于 $TF-IDF$ 算法、$word2vec+idf$ 算法对分词后的词进行向量化，得到特征权重值，再把特征值带入 $k-means$ 聚类模型中。

（2）基于聚好类的文本，定义热度指数模型。

①相同类别下，每增加一条留言，热度值在初始值的基础上增加 10，例如：若该类别有三条留言，则热度得分为 30 分。

②针对每一条留言，则统计点赞数和反对数，只要有出现票数，在热度值上加一分。

③时间参数对热度也有影响，当留言实践越接近现在的时间，热度值加一分。

$heat\_value$=某类别下的留言条数*10+点赞数/反对数*1+时间接近度*1

3.2.2 模型求解

基于 *pycharm* 平台实现聚类和热度指数评估后，得到排名前 5 的热点问题如下所示：

① 基于 $TF-IDF+Kmeans$

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	2222	2019/02/18 至 2019/11/29	长房很多	A市五矿万境K9县负一楼面积缩水
2	2	2068	2019/01/11 至 2019/12/12	A市长A4区中学开建	投诉A3区长郡梅溪湖中学附近工地扰民
3	3	1716	2019/01/11 至 2019/07/08	58车贷	西地省A市58车贷恶性退出
4	4	1300	2019/01/04 至 2020/01/07	A市北二环A4区大道三减速带	咨询A市榔梨街道秋江路扩建问题
5	5	830	2019/01/08 至 2019/08/08	58车贷五	58车贷立案五个月过去

图 9 热点问题表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	208069	A00094436	A5区五矿万境K9县的开发商与施工方建房存在质量问题	2019-05-05 13:52:50	本人是A5区	0	2
1	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019-08-19 11:34:04	我是A市A5	0	2097
1	210692	A00073248	A6区新城国际花都楼盘六期存在严重质量问题	2019-06-18 09:56:36	A6区新城国	0	4
1	215507	A000103230	A市五矿万境K9县存在严重的消防安全隐患	2019-09-12 14:48:07	预交房23栋	0	1
1	228216	A00012131	请协调A市375路公交车增停福元路洪山路口站，并增设过街天桥一座	2019-02-18 09:53:36	尊敬的交通	0	2

图 10 热点问题留言明细表

② 基于 $word2vec+TF-IDF+Kmeans$

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	2281	2019/02/03 至 2019/12/14	A5区	再次投诉A市保利中航城业主顶风违建扩违别墅
2	2	1976	2019/01/04 至 2019/12/18	A2区A1区	请解决A2区中建A1区嘉苑的业主小孩入学问题
3	3	1506	2019/01/06 至 2020/01/05	A市连锁店跑路	投诉A市鑫苑木莲世家涉嫌虚假宣传
4	4	1297	2019/01/20 至 2019/12/28	A市有限公司	恳请帮忙解决A市A3区西地省鸟玄科技有限公司拖欠员工薪资问题
5	5	871	2019/01/21 至 2019/10/15	A市中粮四不合理	A市万科金域蓝湾商铺烟道装在住户窗口这种做法合理合法吗

图 11 热点问题表

基于 $word2vec+TF-IDF+Kmeans$ 模型比 $TF-IDF+Kmeans$ 效果更好。

3.3 问题 3 ：答复意见质量评估模型

3.3.1 模型的建立

答复相关性、答复的完整性、答复的可解释性三个维度是答复意见质量的评估模型建立的关键，答复意见质量评估模型的建立，也有待成为监督相关职能部门是否尽职尽责的一个衡量标准。

（1）答复相关性——假设权重占比50%

我们对留言主题和留言详情整合后的文本，基于*TextRank*算法对留言文本和答复文本中的关键词进行抽取，抽取数量为前 20 个关键词，由于答复是针对出现的某一个问题所做出的行动、采取的方法，因此限制抽取的词的词性限制为名词。对抽取后的词进行相互比对，根据群众留言文关键词和答复关键词匹配的个数在对应留言主题中所有关键词总个数的占比，乘以基础分数 50 分，得到每一条答复意见的相关性得分。相关性得分的具体流程如下图 12 所示：

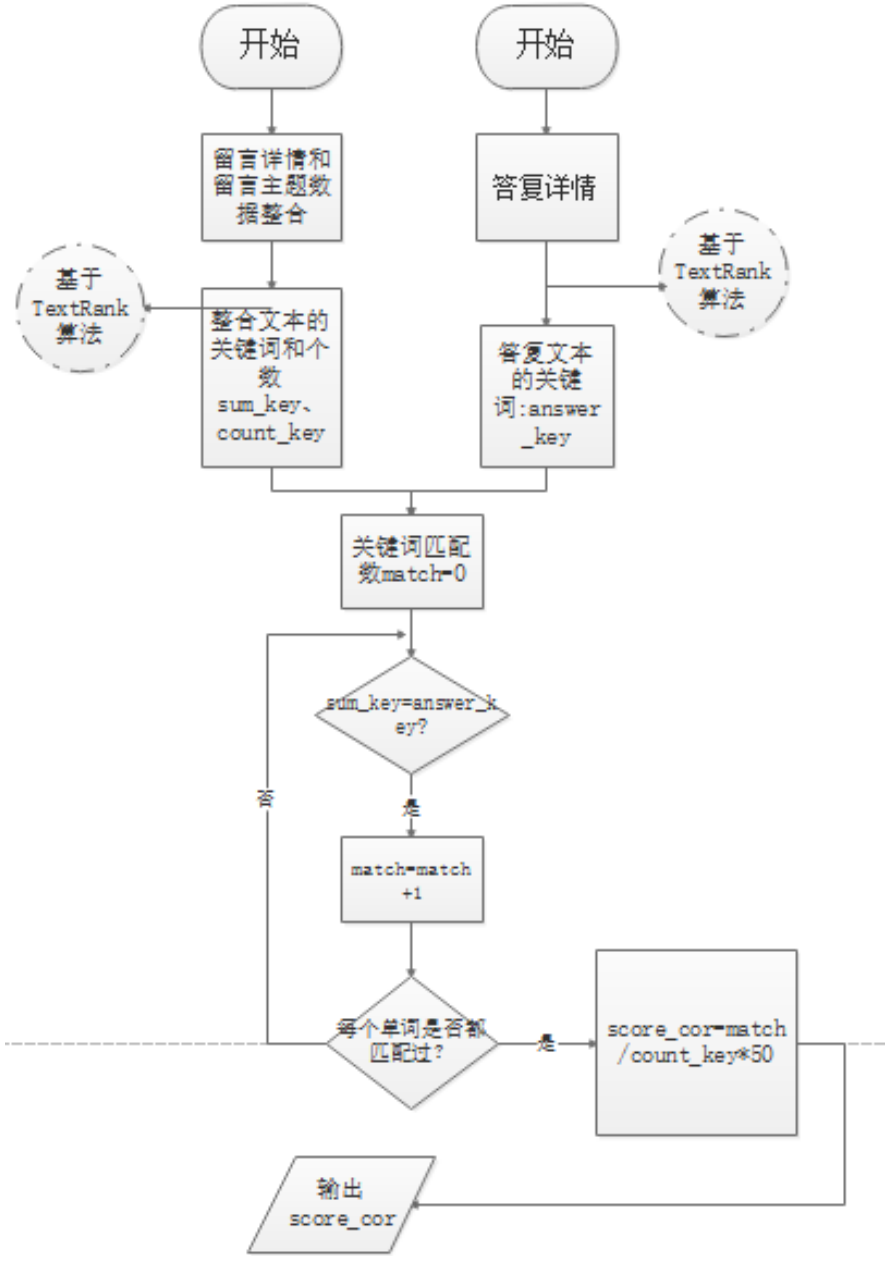


图 12 答复相关性得分计算流程图

$$score\_cor = \frac{match}{count\_key} * 50$$ 公式（十）

（2）答复的可解释性——权重占比 20%

由于留言时间和回复时间间隔会产生一定的信息滞后性。时间间隔越长，答复对解决问题的作用就越小；两者时间间隔越长，群众留言的问题发生变质或者变性的可能性越大。因此，我们设定答复的可解释性这个维度由时间间隔来确定，即建立一个答复时间分数模型。时间间隔越长，扣分越多。

$$score\_explan = (1-(\frac{时间间隔}{答复时间}))*20$$

（3）答复的完整性——权重占比 30%

通常情况下，群众的留言一般是以反应问题的陈述句去留言，不会是以提问的问句去留言，因此在对群众留言和答复信息的问题对应上难以实现，这里我们设定答复的完整性这一维度由富信息评论来确定。富信息评论是指可信度高的评论被定义为文本长度长的评论，因为内容较长的评论被认为包含更多的信息和更细致的描述。先统计评论集合中各评论的字符长度，再计算出单个评论集合的字符长度的平均值，基于此，计算出各评论字符长度的标准差以描述评论的长度特征。

$$score\_complete = \frac{答复的评论长度\text{-}所有答复的平均长度}{总体标准差} \quad 公式（十二）$$

（4）基于以上三个维度，建立答复质量评估模型，模型公式如下：

$$answer\_quality = \lambda_1 score\_cor + \lambda_2 score\_explan + \lambda_3 score\_complete \quad 公式（十三）$$

λ_1、λ_3为正值，λ_2为负值。

3.3.2 模型求解

基于 *pycharm* 平台对已知数据带入到我们建立的模型中进行求解，从而得出各个答复文本的得分，根据得分情况对答复信息进行评估，部分答复信息如下所示。具体求解过程和所有答复信息的得分见附件表 3-评价方案评分表。

4 参考文献

[1] 陈羽中，方明月，郭文忠,等. 基于小波变换与差分自回归移动平均模型的微博话题热度预测

[J]. 模式识别与人工智能, 2015, 28(7):586-594.

[2] 崔彩霞. 停用词的选取对文本分类效果的影响研究[J]. 太原师范学院学报:自然科学版(04):96-98.

[3] Ibrahim Abu El-Khair. TF*IDF[J]. encyclopedia of database systems, 2009, 13(12):3085-3086.

[4] 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, 028(023):5795-5796,5799.

[5] 宫秀军, 刘少辉, 史忠植. 一种增量贝叶斯分类模型[J]. 计算机学报, 2002(6).