“智慧政务”中的文本挖掘应用

摘要

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。因此，运用网络文本分析和数据挖掘技术对网络问政平台的应用有着重要的意义。

对于问题 1，本文首先把附件 2 中的数据进行去重去空、中文分词及停用词过滤等数据预处理操作，使用生成的词汇表对原有句子按照单词逐个进行编码，词汇表模型既可以很好的表现文本由哪些单词组成，又能表达出单词之间的前后关系，TensorFlow 默认支持了这种模型，使用 Word2vec 将词转化为稠密向量。再通过 Text-CNN 模型对网络问政平台的群众留言信息进行分类。

对于问题 2，首先把附件 3 中的数据进行去重去空、中文分词及停用词过滤等数据预处理操作，在中文分词操作中利用 jieba 中文分词工具对网络问政平台的群众留言信息进行分词，再利用 TF-IDF 算法得到每个单词的向量表示，然后采用 K-means 进行文本聚类，实现对网络问政平台的留言信息分类，根据制定的热度评价指标，取出排名前五的热点问题。

对于问题 3，针对附件 4 相关部门对留言的答复意见，通过制定一系列的针对答复的相关性、完整性、可解释性等评价方案，对答复意见的质量进行评价并完成实现。

关键词：去重；中文分词；CNN；TF-IDF；词袋模型；K-means 聚类

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on human to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government. Therefore, the application of network text analysis and data mining technology is of great significance to the application of network politics platform.

For question 1, this paper first preprocesses the unstructured data in Annex 2, such as de duplication and de emptiness, Chinese word segmentation and stop word filtering, and uses the generated vocabulary to encode the original sentences one by one according to the words. The vocabulary model can not only show which words the text consists of, but also express the relationship between the words, which is supported by tensorflow by default Models. Then through CNN convolutional neural network to classify the public message information of network politics platform.

For question 2, firstly, we preprocess the unstructured data in Annex 3, such as de duplication and de emptiness, Chinese word segmentation and stop words filtering, and use the Chinese word segmentation tool of Jieba to segment the public message information of the network politics platform, and then use the word bag model to process the text word bag, obtain the corresponding characteristics, and then use TF-IDF The algorithm gets the TF-IDF weight vector of each message classification description. Here, the TF-IDF model and the word bag model are used together to further process the data generated by the word bag model, and then K-means is used The weight vector of TF-IDF is clustered to realize the classification of message

information of network politics platform. According to the established heat evaluation index, the top five hot issues are taken out.

For question 3, in view of the reply opinions of the relevant departments in Annex 4 to the message, the quality of the reply opinions is evaluated and realized through the development of a series of evaluation schemes for the relevance, integrity and interpretability of the reply.

Key words: Keywords de duplication； Chinese word segmentation;；CNN; TF-IDF； word bag model；K-means clustering

目 录

1、挖掘目标

本次建模目标是利用网络问政平台群众留言信息数据，利用 jieba 中文分词工具对留言数据进行分词，采用 CNN 卷积神经网络、K-means 文本聚类算法，达到以下目标：

1）利用词汇表模型、word2vec 和 Text-CNN 的方法对数据进行文本挖掘，将留言信息进行分类，以便后续将群众留言分派至相应的职能部门处理，达到减少工作人员的工作量、降低差错率、提高工作效率等目标。

2）利用文本分词和文本聚类的方法对数据进行文本挖掘，根据聚类结果制定合理的热点评价指标，可以及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

3）针对留言的答复意见，制定系统的评价方案，以达到对相关部门工作人员的答复意见进行质量评估的目标。

2、分析方法与过程

本文的总体架构及思路如下：

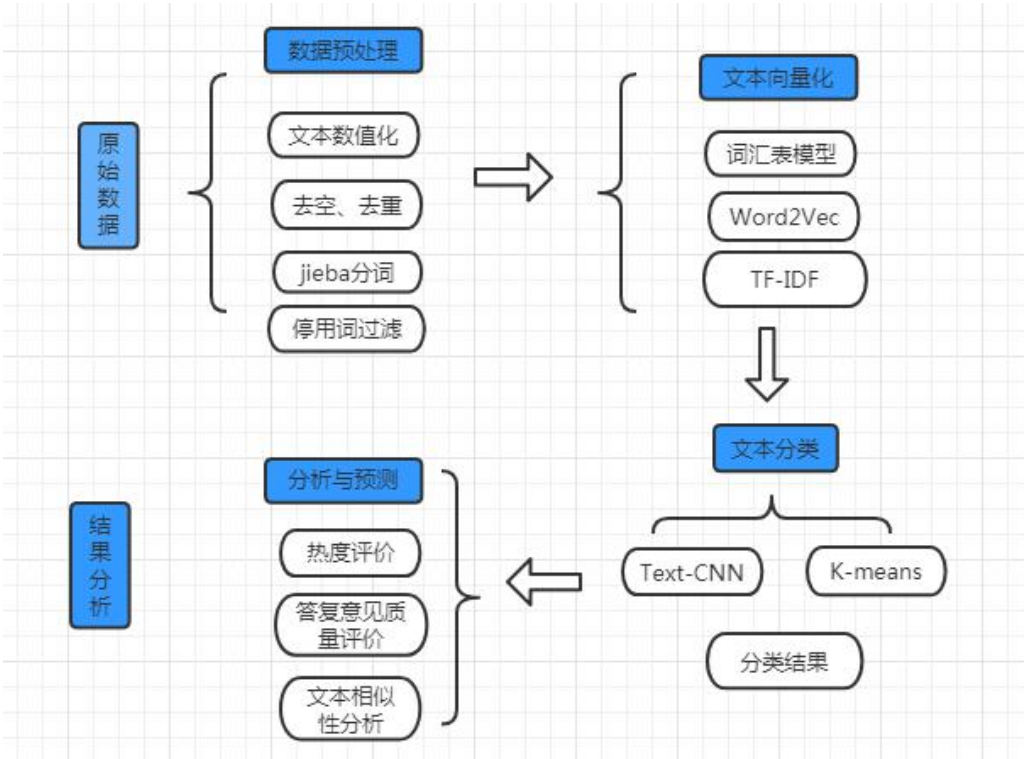

图 1：总流程图

步骤 1：数据预处理，对附件 2、附件 3 及附件 4 中的数据进行预处理，去除重复项以及空数据、中文文本分词、停用词过滤等操作。

步骤 2：文本向量化，基于词汇表模型、Word2Vec、TF-IDF 进行处理。

步骤 3：文本分类，在问题 1 中通过 CNN 卷积神经网络对网络问政平台的群众留言信息进行分类。在问题 2 中采用 K-means 进行文本聚类，实现对网络问政平台的留言信息分类。

步骤 4：分析与预测，根据问题 1 的分类结果，制定热度评价指标，取出排名前五的热点问题。并对答复意见从相关性、完整性、可解释性等角度制定评价方案。

步骤 5：利用以上结果分别的三个问题的结果进行总结。

2.1 问题 1 分析方法与过程

2.3.1 分析过程

1）任务

问题 1 是一个文本多分类的问题，在文本分类中大体上分为基于传统机器学习的文本分类模型和基于深度学习的文本分类模型，需要制定合理的文本分类评估方法。

2）难点

① 文本语义带来的语义交叉，比如：交通局的亲属拖欠我们工资

② 多分类问题带来的难度（转化为多个二分类）

③ 数据不平衡带来的影响（数据增强）

④ 长文本的无意义表达太多（是否转为短文本、关键句）

2.1.2 数据预处理

1）群众留言信息的去重、去空

针对附件 2 中所给数据中的群众留言详细数据，去除数据中的重复数据和空数据，便于后面对数据进行操作。

2）特殊符号过滤

针对题目中所给数据中的群众留言详细数据，去除数据中空和一些符号包括'\t'，'\n'，' '，'\r'，'\xa0'，'\xa9'， '\u3000'

3）对留言信息进行中文分词

在对群众留言信息进行挖掘分析之前，先把文本数据转化为计算机能够识别的结构化信息，为了便于转换，先对留言详细数据进行中文分词操作，这里采用Python 的中文分词包 jieba 进行分词，jieba 采用了基于前缀词典实现的高效词图扫描，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)，同时采用了动态规划 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，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。jieba 分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。

2.1.3 词汇表模型

首先读取文件数据，然后根据训练集构建词汇表并存储，接下来是对文本数据进行编码，把文件里的每一个词对应到词汇表中的每个词，找到词汇表中对应的 id，用这个 id 作为当前词的编码，对于标签采用的是 one-hot 编码，读取分类目录[‘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’]。

2.1.4 word2vec

word2vec 也叫 word embeddings，中文名“词向量”，作用就是将自然语言中的字词转为计算机可以理解的稠密向量（Dense Vector）。Word2Vec 可以将 One-Hot Encoder 转化为低维度的连续值，也就是稠密向量，并且其中意思相近的词将被映射到向量空间中相近的位置。

把文本中的每个词对应到单词表中的每个词，然后取 id，变成数字，完成嵌入，输入到 CNN 中。

word2vec 模型结构如下图所示：

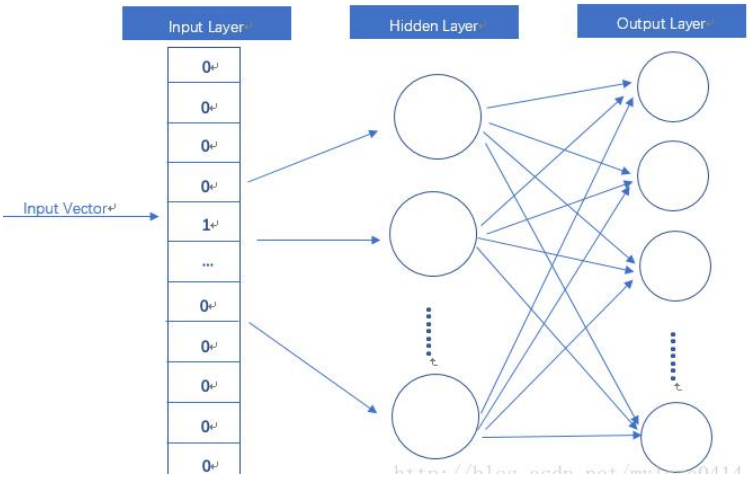

图 2：word2vec 模型架构

输入是 One-Hot Vector，Hidden Layer 没有激活函数，也就是线性的单元。Output Layer 和 Input Layer 具有相同的维度，用的是 Softmax 回归。

2.1.5 Text-CNN 模型

Text-CNN 模型的整体网络架构如图所示：

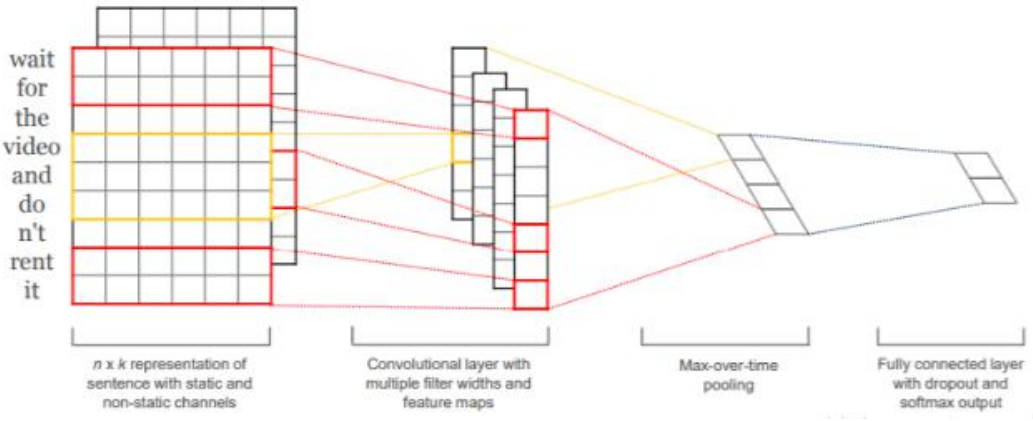

图 3：Text-CNN 模型

整个模型由四部分构成：输入层、卷积层、池化层、全连接层

1.输入层（词嵌入层）：

Text-CNN 模型的输入层需要输入一个定长的文本序列，我们需要通过分析语料集样本的长度指定一个输入序列的长度 L，比 L 短的样本序列需要填充，比 L 长的序列需要截取。最终输入层输入的是文本序列中各个词汇对应的词向量。

2.卷积层：

在 NLP 领域一般卷积核只进行一维的滑动，即卷积核的宽度与词向量的维度等宽，卷积核只进行一维的滑动。在 Text-CNN 模型中一般使用多个不同尺寸的卷积核。卷积核的高度，即窗口值，可以理解为 N-gram 模型中的 N，即利用的局部词序的长度，窗口值也是一个超参数，需要在任务中尝试，一般选取 2-8 之间的值。

3.池化层：

在 Text-CNN 模型的池化层中使用了 Max-pool（最大值池化），即减少了模型 的参数，又保证了在不定长的卷基层的输出上获得一个定长的全连接层的输入。卷积层与池化层在分类模型的核心作用就是特征提取的功能，从输入的定长文本序列中，利用局部词序信息，提取初级的特征，并组合初级的特征为高级特征，通过卷积与池化操作，省去了传统机器学习中的特征工程的步骤。

4.全连接层：

全连接层的作用就是分类器，原始的 Text-CNN 模型使用了只有一层隐藏层的全连接网络，相当于把卷积与池化层提取的特征输入到一个 LR 分类器中进行分类。

模型的效果评估与调优：

模型的效果评估与调优针对分类问题，一般可以使用准确率、召回率、F1 值、混淆矩阵等指标，在文本多标签分类中一般还会考虑标签的位置加权等问题。分类模型中的主要参数：词向量的维度、卷积核的个数、卷积核的窗口值、L2 的参数、DropOut 的参数、学习率等。

2.2 问题 2 分析方法与过程

2.2.1 分析过程

1）任务

子任务 1：问题识别，即如何从众多留言中识别出相似的留言

子任务 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

子任务 3：热度评价，热度评价指标的定义和评价方法，对指标排名之后结果对应表 1

2）难点

问题 2 可能存在的难点有地点、人群的识别，即表达的多样化；相似的计算复杂，即特征多、两两之间计算相似的计算量较大。

2.2.2 数据预处理

1）群众留言信息的去重、去空

针对附件 3 中所给数据中的群众留言详细数据，去除数据中的重复数据和空数据，便于后面对数据进行操作。

2）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）。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。将筛选出的停用词 加入停用词表，再利用停用词表过滤停用词，将分词结果与停用词表中的词语进行匹配，若匹配成功，则进行删除处理。

3）对数据进行中文分词

对附件 3 中的留言详细数据进行中文分词操作，采用 Python 的中文分词包 jieba 进行分词，过程步骤同问题 1。

2.2.3 TF-IDF 算法

在对留言详细分词后，需要把这些词语转换为向量，以供挖掘分析使用。 这里采用 TF-IDF 算法，把留言详细转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数。

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本 中的分布越集中，说明该分词在区分该文本内容属性能力越强。

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某 个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排 序，次数最多的即为要提取的职位描述表中文本的关键词。

生成 TF-IDF 向量：

生成 TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找出每个职位描述的前 5 个关键词；

（2）对每个岗位描述提取的 5 个关键词，合并成一个集合，计算每个岗位描述 对于这个集合中词的词频，如果没有则记为 0；

（3）生成各个岗位描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF）

2.2.4 K-means 聚类

生成留言详细描述的 TF-IDF 权重向量后，根据每个留言详细的 TF-IDF 权重向量，对留言详细进行分类。这里采用 K-means 算法把留言详细分成 7 类。

K-mean 聚类的算法步骤如下：

1、从 X 中随机取 K 个元素，作为 K 个簇的各自的中心。

2、分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度 最低的簇。

3、根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各 自维度的算术平均数。

4、将 X 中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化。

6、将结果输出。

K-mean 聚类的算法流程图如下：

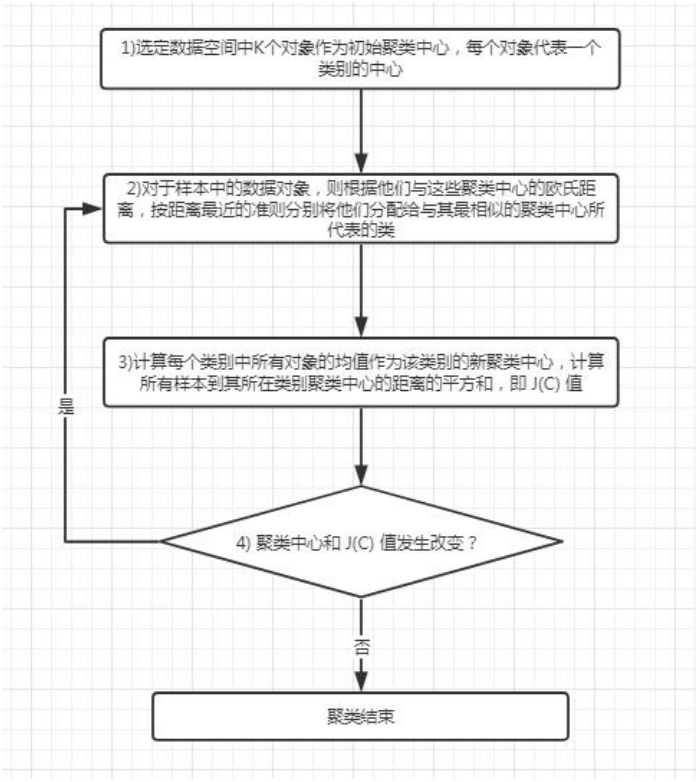

图 4：聚类算法流程图

该算法要求在计算之前给定 k 值。本文通过计算附件 3 给出的所有留言详细的类别数，并以此作 k 的值，这里令 k = 7，即城乡建设、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。

2.2.4 热度评价指标

热度的定义：

“热度”是一个模糊的概念，所以我们首先要定义何为“热度”，本文通过分析，认为热度评价的指标应当是对于同一条留言详细，应具有较高的点赞数和反对数。此外，考虑到所有的样本数据的点赞数和反对数的量，计算单个热点问题的热度在原基础上加 1，每发一条评论说明有一人参加，再进行热度的计算，因此留言中没有被点赞或者反对的评论同样可以计算热度。

2.3 问题 3 分析方法与过程

2.3.1 分析过程

1）相关性：答复意见的内容是否与问题相关

2）完整性：是否满足某种规范

3）可解释性：答复意见中内容的相关解释

难点

1）如何将相关性、完整性、可解释性等描述量化

2）构建何种指标来计算和评价

2.3.2 答复意见评价方案

相似性角度：

采用文本相似度计算，相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体的差异大小，也就是需要对文本信息进行相似度计算，将根据相似特性的信息进行归类。目前相似度计算方法分为距离度量和相似度度量。本文采用的是通过计算两个句子之间的余弦函数来计算相似度得分，其中 similar_score 等于 0.0 表示两个句子完全不相似，二者正交。

在计算文本相似度之前，对附件 4 中的答复意见进行了本文预处理操作，包括数据的去重、去空，停用词过滤，中文分词操作等。

完整性角度：

答复应该具有一定的标准，检查答复意见是否满足标准，比如是否有开头、结尾等等。

可解释性角度：

答复内容里面有没有一些相关的解释或者理论支撑，比如引进据点、一些相关的法律等诸如此类的解释，使群众能够了解此类问题能否被解决的原因。

3、结果分析

3.1 问题 1 结果分析

针对问题 1，对于附件 2 中的留言详细数据，采用 Text-CNN 模型，总迭代轮次为 10，每 100 轮输出一次结果，每 10 轮存入 Tensorboard，模型预测精确率为 90.01%，结果如下图所示：

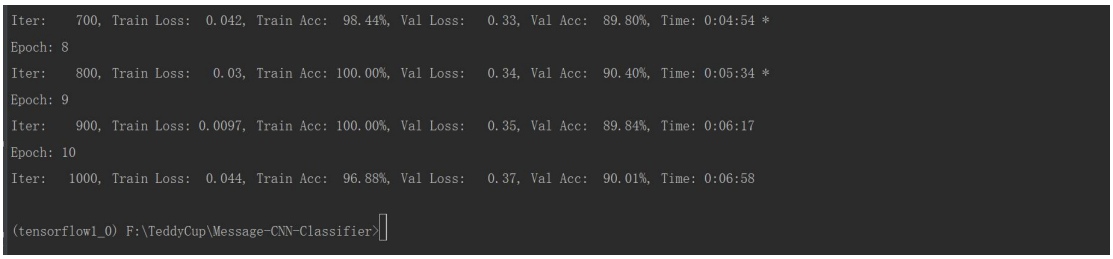

图 5：问题 1 运行结果

3.2 问题 2 结果分析

对于问题 2，针对附件 3 中的留言详细数据以及评论的点赞数和反对数，制定了热度评估指标，对于同一条留言详细，应具有较高的点赞数和反对数。此外，考虑到所有的样本数据的点赞数和反对数的量，计算单个热点问题的热度在原基础上加 1，每发一条评论说明有一人参加，再进行热度的计算，因此留言中没有被点赞或者反对的评论同样可以计算热度。根据题目要求取出排名前五的热点问题，保存在文件“热点问题表.xls”，如下图所示：

	label	hot
0	142	2139
1	325	1778
2	279	1602
3	263	793
4	93	716

图 6：排名前五的热点问题

由于全部数据中时间格式不一致，对时间格式的一致性进行了修改，且 time 一列有日期类型不是字符类型，对每个类别按照时间先后进行排序。

并统计了排名前五的相应热点问题对应的留言信息，保存为“热点问题留言明细表.xls”。

3.3 问题 3 结果分析

通过计算文本相似度，完成答复意见质量的评价，留言详细和答复意见的相似度是通过计算两个句子之间的余弦函数来计算相似度得分，其中 similar_score 等于 0.0 表示两个句子完全不相似，二者正交。

其中计算了相似性答复在全文中的比率，实现过程及结果如下图 7 所示，相似率达到 92.4%，因此，大多数的相关部门工作人员的答复意见的质量还是可以的。

```
zero_count = 0
for score in similar_score_list:
    if score==0.0:
        zero_count += 1

similar_rate = (len(similar_score_list)-zero_count) / len(similar_score_list)
print(similar_rate)
```

```
0.9243607954545454
```

图 7：答复质量的相关性

并取出了前五条留言详细和答复意见不相似的留言数据信息，如图 8 所示：

	id	username	theme	Qtime	content	response	Atime	similar_score
16	3727	UU0081194	投诉A3区桐梓坡路益丰大药房以次充好	2018/12/27 1:55:21	\n\t\t\t\t\n\t\t\t\t12月16日上午，我来到A3区桐梓坡路益丰大...	网友"UU0081194"您好！您的留言已收悉。现将有关情况回复如下：因您未留下联系方式及投...	2019/1/3 14:02:47	0.0
96	4507	UU008310	投诉A市外国语学校国庆补课	2018/10/4 10:53:07	\n\t\t\t\t\n\t\t\t\tA市外国语学校国庆补课，只放四天，剩下三天全...	网友"UU008310"您好！您的留言已收悉。现将有关情况回复如下：经查，该工地是龙湖春江郦...	2018/10/26 16:45:23	0.0
206	6154	UU008707	希望能考虑在A7县泉塘设个地铁	2018/4/29 22:23:27	\n\t\t\t\t\n\t\t\t\t泉塘的小伙伴们恳请各位领导们能考虑泉塘能设个...	网友"UU008707"您好！您的留言已收悉。现将有关情况回复如下：目前，市规划局正在编制《...	2018/5/16 17:22:43	0.0
241	6556	UU0081320	咨询打狂犬疫苗报销比例是多少	2018/3/20 15:19:47	\n\t\t\t\t\n\t\t\t\t 请问领导，农合费用增加了，打狂犬疫苗报销...	已收悉	2018/3/28 16:05:34	0.0
428	9203	UU0081289	希望A市时代阳光大道省质监局段加人行天桥！	2017/3/14 11:15:33	\n\t\t\t\t\n\t\t\t\t 尊敬的领导，我是A5区桃花堰路上的...	网友"UU0081289" 您好！您的留言已收悉。现将有关情况回复如下： 目前，我市正在...	2017/4/12 11:17:58	0.0

图 8：排名前五的不相似留言信息

4、结论

对网络问政平台信息进行分析研究，了解了运用网络文本分析和数据挖掘技术对网络问政平台的应用有着重要的意义，同时也是文本分析的一个课题、一个难题。传统的文本解读已经不能满足数据量庞大的群众留言信息，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。本文采用根据 Text-CNN、K-means 聚类等方法，把群众留言信息进行分类预测，并制定了热度评价方案，对于各职能部分的对群众留言的答复意见制定了质量评价方案，便于对各职能部门的工作人员进行工作质量评估。

由分析结果可以看出，对留言质量通过相似性、完整性、可解释性角度进行评估之后，发现相关部门对群众留言的答复质量还是不错的。

但是，在挖掘分析的过程中也有一些不足之处，我们后期会对文本挖掘任务做进一步的改进，争取把文本挖掘任务做到最好。

5、参考文献

[1]陶伟. 警务应用中基于双向最大匹配法的中文分词算法实现[J]. 电子技术与软件 工程, 2016(4).

[2]郭飞, 张先君, 叶俊. 基于改进互信息的特征提取的文本分类系统[J]. 四川理工学院学报:自然科学版, 2008, 21(3):93-96.

[3]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, 28(23):5795-5796.

[4]周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 中国科学院研究生院(计算技 术研究所), 2005.

[5]胡学钢, 董学春, 谢飞. 基于词向量空间模型的中文文本分类方法[J]. 合肥工业 大学学报:自然科学版, 2007, 30(10):1261-1264.

[6]黄章益, 刘怀亮. 一种基于语义的中文文本特征降维技术研究[J]. 情报杂志, 2011(S2):123-125.

[7]郑翠翠. 面向领域文本的潜在语义分析研究[D]. 南京理工大学, 2010.

[8]邬启为. 基于向量空间的文本聚类方法与实现[D]. 北京交通大学, 2014.

[9]张振亚, 王进, 程红梅,等. 基于余弦相似度的文本空间索引方法研究[J]. 计算 机科学, 2005, 32(9):160-163.

[10]张跃, 李葆青, 胡玲,等. 基于 K-Mean 文本聚类的研究[J]. 中国教育技术装备, 2014(18):50-52.