留言信息的文本挖掘与综合分析

摘要：近年来，随着各大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，因此，运用网络文本分析和数据挖掘技术对市民的留言进行分类、汇总，对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题一：本文首先通过 python 对附件 2 中的留言详情去重去空，并利用 jieba 中文分词工具对留言详情分词及停用词过滤等数据预处理，然后使用 TextRank 提取每一个留言详情的前 12 个关键词，接着使用 token 字典将文字转化为数字列表，再用 Embedding 层将数字列表转化为向量列表，最后将向量列表送入 CNN 卷积神经网络进行训练，在 epoch=1 时训练准确值和损失值趋于平稳，建立留言内容的一级标签分类模型比较优。

针对问题二：通过利用 pyltp 对分词以及停用词过滤后的附件 3 进行命名实体识别，识别出地点、机构名和人名，对这三个类别进行数据数值化处理后，利用余弦相识度计算文本相识度，再通过 k-means 聚类算法对热点问题进行划分，根据热度评价指标，筛选出“开发商捆绑车位销售”、“小区附近违建搅拌站”，“污染环境和噪音扰民”、“58 车贷案件近半年毫无进展”、“人才购房补贴申请失败”、“小区临街餐饮店油烟噪音扰民”这五个热点问题，并且统计出热点指数统计表。

针对问题三，利用 Python 对附件 4 进行去重去特殊字符后，对附件 4 中的留言详情及答复意见进行 jieba 分词，再用停用词表对分词后的信息进行过滤，最后提取关键词，对留言详情中的关键词和答复意见中的关键词进行余弦相似度处理，最后得出其相关性指标为 0.824，完整性指标为 0.867，可解释性指标为 0.815，根据得出的三个指标，可见答复意见的质量之高。

最后，根据研究分析，建立了一级标签分类模型，找出了排名前五的热点问题以及制定了对答复意见质量的评价方案。

关键词：中文分词　TextRank　K-means 文本聚类 CNN 卷积神经网络

Abstract: Ask ZhengPing Taiwan in recent years, with the network gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge, therefore, using the network text analysis and data mining technology to the citizens of message classification and summary, to enhance the management level of government and governance efficiency has a great role in promoting.

According to the problems: firstly,This paper through the python in annex 2 a message details to heavy to empty, and use the jieba Chinese word segmentation tools for message details points and stop words filtering data preprocessing, and then use the TextRank extract every message details before 12 key words, then use the Token dictionary to translate words into digital list, with a list of Numbers can be converted to vector Embedding layer list, the final list into the vector convolution neural network (CNN) training, training in epoch = 1 accurate value and loss leveled off, It is better to establish a level 1 label classification model for message content.

To solve the second problem: By using pyltp to word segmentation and the annex 3 after the stop words filtering for named entity recognition, to identify the location, institutions and names, to quantize the three categories of data processing, using cosine met computed text met degrees, again through the k - means clustering algorithm on hot issues, according to the evaluation index, heat out "developers bundle car sales", "village near the neo-treasure hill station", "Pollution of the environment and noise nuisance", "58 car loan case nearly half a year no progress", "talent purchase subsidy application failure", "community on the street restaurant lampblack noise nuisance" these five hot issues, and the statistics out of the hot spot index statistical table.

For three problems: annex 4 to make use of Python to heavy to special characters, to leave a message in annex 4 details and reply to jieba participles, reoccupy after the stop list of word segmentation information filtering, and finally to extract the keywords, to the key words in the message for details and reply to key words in the opinion of cosine similarity processing, finally it is concluded that the correlation index of 0.824, integrity index is 0.867, interpretability index is 0.815, according to conclude three indicators, the high quality of the opinions of visible reply.

Finally, according to the research analysis, the paper establishes the model of first-level label classification, finds out the top five hot issues and develops the evaluation scheme for the quality of the replies.

Key words: Chinese Word Segmentation K-means text clustering TextRank convolution neural network(CNN)

目录

1. 挖掘目标.....4
2. 总体流程与步骤.....4
3. 问题 1 的分析方法与过程.....4
3.1 流程图.....5
3.2 数据预处理.....5
3.2.1 留言详情去重去空.....5
3.2.2 中文文本分词.....5
3.2.3 停用词过滤.....5
3.3 文本向量化.....6
3.3.1 TextRank 提取关键词.....6
3.3.2 建立 Token 字典并将文字转换成数字列表.....7
3.3.3 用 Embedding 把数字列表转换成向量.....7
3.3.4 将向量送入 CNN 卷积神经网络进行训练.....8
3.3.5 Tensorflow.....9
3.4 建立模型.....9
3.4.1 保存模型与模型可视化.....9
3.4.2 训练过程可视化.....9
4. 问题 2 的分析方法与过程.....10
4.1 流程图.....10
4.2 数据预处理.....11
4.3 pyltp 的命名实体识别.....11
4.4 TF-IDF 权值向量.....12
4.4.1 传统的 TF-IDF：.....12
4.4.2 TF-IDF 权值向量:.....13
4.4.3 文本向量化表示:.....13
4.5 余弦相识度.....14
4.6 Kmeans 聚类.....16
4.7 结果分析.....16
5. 问题 3 过程分析与流程图.....17
5.1 流程图.....17
5.2 数据预处理.....18
5.2.1 留言详情、答复意见的去重去特殊字符.....18
5.2.2 停用词过滤.....18
5.2.3 关键词提取.....19
5.2.4 分析关键词得出各项指标.....19

1. 挖掘目标

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。因此，人们把目光转向效率高的人工智能技术方面，建立基于自然语言处理技术的智慧政务系统也越来越迫切。

本次建模目标是利用附件 1、附件 2 以及附件 3 的数据，利用 jieba 中文分词对留言详情、留言主题和意见答复这些长句子进行分词、TextRank 提取关键词以及 K-means 聚类的方法，达到以下三个目标：

(1) 对附件数据预处理后，把文本数据转换成向量列表进行 CNN 卷积神经网络进行训练、分类，并建立留言内容的一级标签分类模型。

(2) 对已经处理好的附件 3，利用余弦相识度计算文本相识度，再通过 k-means 聚类算法对热点问题进行划分。

(3) 对留言详情中的关键词和答复意见中的关键词进行余弦相似度处理，根据相关性、完整性、可解释性三个指标，对答复意见的质量进行合理性评价。

2. 总体流程与步骤

2.1 总体流程

本文的总体架构及思路如下：

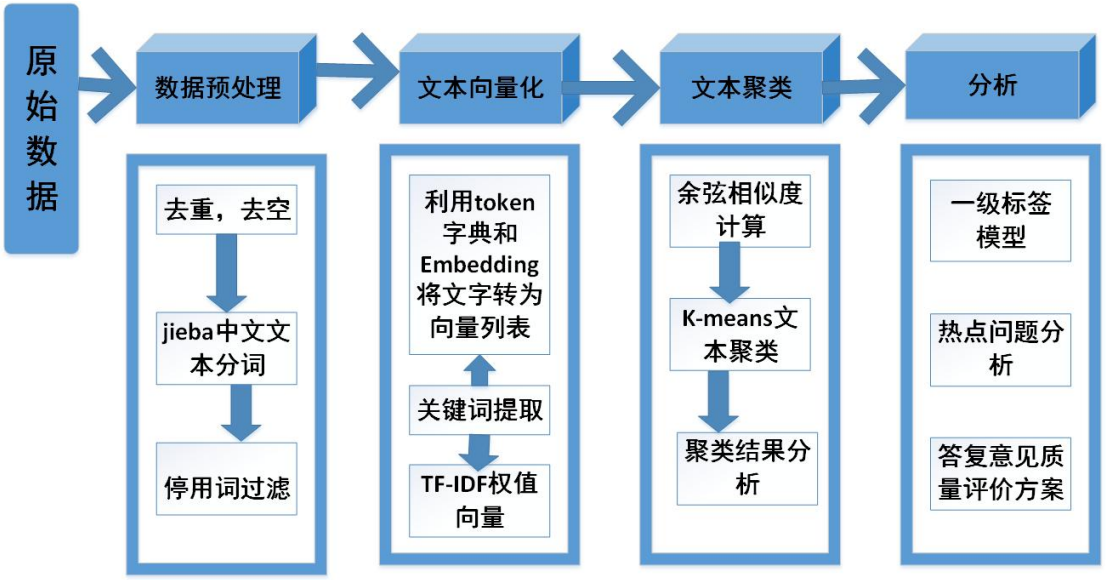

图 1 总流程图

步骤一：数据预处理，题目所给的文本数据较长，所以需要对附件 1 结构化文本数据值化处理，对附件 2、附件 3、附件 4 非结构文本去除重复项及空行、中文文本分词、停用词过滤，以便后续分析。

步骤二：文本向量化，基于 TextRank 提取关键词，进而利于用 token 字典和 Embedding 把文字转换成向量列表，以便后续建立模型。

步骤三:对附件 2 建立一级标签分类模型，把向量送进 CNN 卷积神经网络进行训练，利于后续的多分类，最后保存模型。

步骤四：

3. 问题 1 的分析方法与过程

3.1 流程图

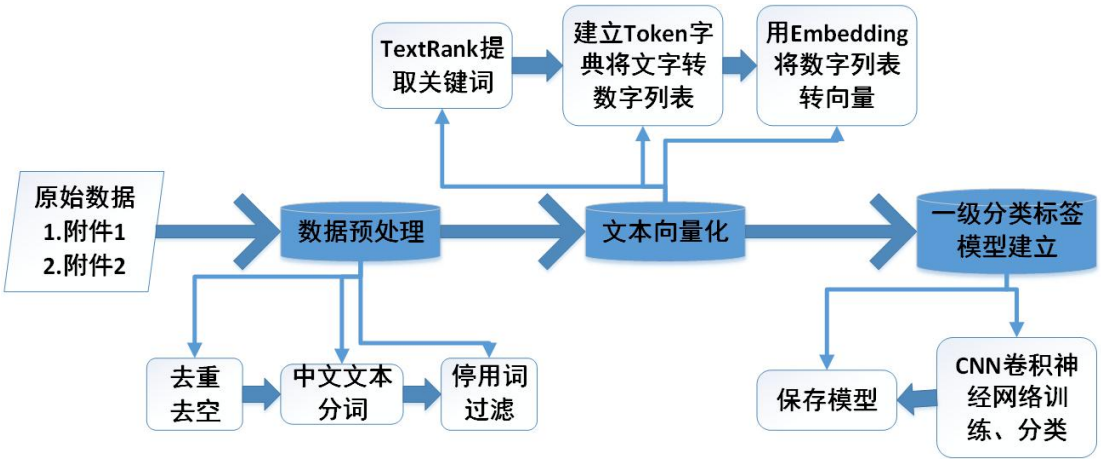

图 2 问题 1 流程图

3.2 数据预处理

3.2.1 留言详情去重去空

附件 2 留言详情、附件 3 留言主题、附件 4 留言详情和答复意见存在大量空行和描述文本完全一致的样本，例如同一个市民多次进行内容重复留言，会给后续分类造成更大的误差，因此，在去重的时候保留一条留言即可。

3.2.2 中文文本分词

在对智慧家政进行文本挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件 2 留言详情、附件 3 留言主题、附件 4 留言详情和答复意见中，以中文文本的方式给出了数据。为了便于转换，先要对这些长句子的进行中文分词[1]。本例采用了 Python 中的中文分词库 jieba 进行分词。jieba 库，是一个强大的分词库，它的开发者通过大量的训练后，向其录入了两万多条词语进行基本的库。因此，jieba 的实现原理也比较完善，设计的算法有基于前缀字典的有向无环图、动态规划、HMM 模型等。

部分分词结果示例如图：

```
0       [A3, 区, 大道, 西行, 便, 道, , , 未管, 所, 路口, 至, 加油站, 路段...
1       [位于, 书院, 路, 主干道, 的, 在水一方, 大厦, 一楼, 至, 四楼, 人为, 拆...
2       [尊敬, 的, 领导, :, , A1, 区苑, 小区, 位于, A1, 区, 火炬, 路, , ...
3       [A1, 区, A2, 区华庭, 小区, 高层, 为, 二次, 供水, , , 楼顶, 水箱,...
5       [我, 在, 2015, 年, 购买, 了, 盛世, 耀凯, 小区, 17, 栋, 3, 楼...
                                    ...
9205    [我们, 夫妻, 都, 是, 农村户口, , , 大, 的, 是, 女, 9, 岁, , , 小...
9206    [本人, 2015, 年, 2, 月, 16, 号, 在, B, 市中心, 医院, 做, 无...
9207    [我们, 是, 再婚, , , 很, 想, 再, 要, 一个, 小孩, , , 不知, 我省, ...
9208    [K8, 县惊现, 奇葩, 证明, !, , , , 我, 是, 西地省, K8, 县...
9209    [领导, 你好, , , 我们, 属于, 未, 婚生子, , , 但是, 在, 2013, 年,...
Name: 留言详情, Length: 9052, dtype: object
```

图 3 部分分词结果

3.2.3 停用词过滤

为了节省储存空间，提高搜索效率，在处理文本之前会自动过滤某些无表达意义的字或词，这些字或词被称为停用词[3]。停用词有两个特征，第一极其普遍出现频率高；第二是包含信息量低，对文本标识无意义。本文所用停用词表示引用哈工大停用词表，在此基础上结合留言文本信息进行增加或删减，促使后续提取关键词更加有利。

部分停用词结果示例如图：

```
0       [A3，区，大道，西行，道，未管，路口，加油站，路段，人行道，包括，路...
1       [位于，书院，路，主干道，在水一方，大厦，一楼，四楼，人为，拆除，水，...
2       [尊敬，领导，A1，区苑，小区，位于，A1，区，火炬，路，小区，物业,...
3       [A1，区，A2，区华庭，小区，高层，二次，供水，楼顶，水箱，长年，不...
5       [2015，年，购买，盛世，耀凯，小区，17，栋，3，楼，4，楼，两...
                                ...
9205    [夫妻，农村户口，女，9，岁，2，岁，15，斤，治疗，两年，一级，脑...
9206    [2015，年，2，16，号，B，市中心，医院，做，无痛，人流，手术,...
9207              [再婚，想，小孩，不知，我省，二胎，新，政策，先，怀孕，做]
9208    [K8，县惊现，奇葩，证明，  ，  ，西地省，K8，县人，想，生二孩，告...
9209    [领导，未，婚生子，2013，年，接受，处罚，小孩，上户，小孩，外地，...
Name: 留言详情，Length: 9052，dtype: object
```

图 4 部分停用词结果

3.3 文本向量化

3.3.1 TextRank 提取关键词

经过上述文本预处理后，虽然已去掉了部分停用词，但还是包括大量词语给，文本向量化过程带来困难，所以提取关键词，主要目的是不改变原文核心信息的情况下，尽量减少出力的词次，以此来降低向量空间维数，从而简化计算，提高文本处理速度和效率。

TextRank 算法[2]是将一篇文档转换成一张有向带权的词图模型，是将文本进行分割，分割成基本单元，即词语，每个基本单元看作是一个节点，每个节点之间的边由词节点之间的共现关系决定，而节点的重要性又由相邻节点指向数量决定。

TextRank 算法计算公式流程图：

$$WS(V_i)=(1-d)+d\bullet\sum_{V_j\in In(V_i)}\frac{W_{ji}}{\sum_{V_k\in Out(V_j)}W_{ji}}WS(V_j)$$

TextRank 算法提取关键词流程图：

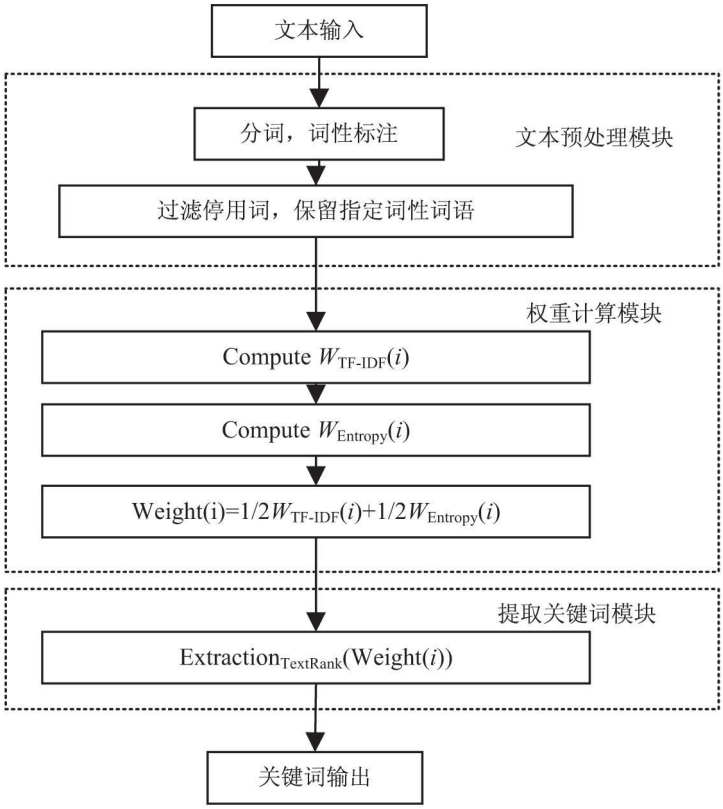

图 5TextRank 算法提取关键词流程图

部分提取关键词结果示例如图：

```
0          路段 A3 未管 文明城市 车流 安全隐患 西行 加油站 上下班 人行道 路灯 整改
1        护栏 在水一方 电等 过往行人 锈迹斑斑 烂尾 四楼 主干道 一楼 人行道 牵头 倒塌
2            业主 物业 车位 收费 小区 停车 停车费 A1 公摊 程明 西地省 小区业主
3            不洗 A1 A2 区华庭 水是 健康 霉味 致癌物 水箱 环保部门 楼顶 长年
5       小区 业主 物业公司 2015 耀凯 17 水电费 业委会 物业费 为所欲为 问一问 足额
                                ...
9205           招郎 15 女户 50 70 农村户口 脑瘫 生子 男方 多岁 再生 夫妻
9206         手术 医生 2015 看病 病人 宫腔积 不用 16 25 药流 宫腔镜 经验
9207                          二胎 再婚 怀孕 小孩 我省 不知 政策
9208          证明 二孩 奇葩 生育 K8 审批 盖章 西地省 计生委 群众 计生办 签字
9209           抚养费 再交 小孩 交清 老百姓 农村 户口 外地 缴纳 告知 社会 办理
Name: data_qudou, Length: 9052, dtype: object
```

图 6 部分提取关键词结果

3.3.2 建立 Token 字典并将文字转换成数字列表

深度学习模型中，需要将文字转化为数字列表，这时候就可以借助 Token 字典。Token 字典。Token[4]的鉴权机制类似于 http 协议也是无状态的，它不需要在服务端去保留用户的认证信息或者会话信息。这就意味着基于 Token 认证机制的应用不需要去考虑用户在哪一台服务器登录了，因此，为此应用的拓展提供了便利。

对附件 2 留言详情和附件 3 留言主题分别建立一个 2000 字的 Token 字典，并将文字转换成数字列表，方便后续转换成向量。

3.3.3 用 Embedding 把数字列表转换成向量

我们定义的首层是 embedding layer，它将词汇表的词索引映射到低维向量表示。它基本上是我们从数据中学习到的 lookup table。embedding 是一种矩阵，其中每列是与词汇表中的 item 对应的向量，要获得表示多个词汇 item（例如句

子或段落中的所有单词）的稀疏向量的稠密向量，可以检索每个单独 item 的 embedding，然后将它们相加。最后查找乘法和加法过程等同于矩阵乘法。 给定 1×N 稀疏表示 S 和 N×M 的 embedding 表 E，矩阵乘法 S X E 给出 1×M 稠密向量。

3.3.4　将向量送入 CNN 卷积神经网络进行训练

卷积神经网络[6][5]是由卷积层，池化层和全连接层组成。卷积层通过卷积计算来提取数据的特征。池化层则从卷积层提供的特征中选取最优特征，之后输出给全连接层进行处理。图 7 是本研究所采用的卷积神经网络的结构

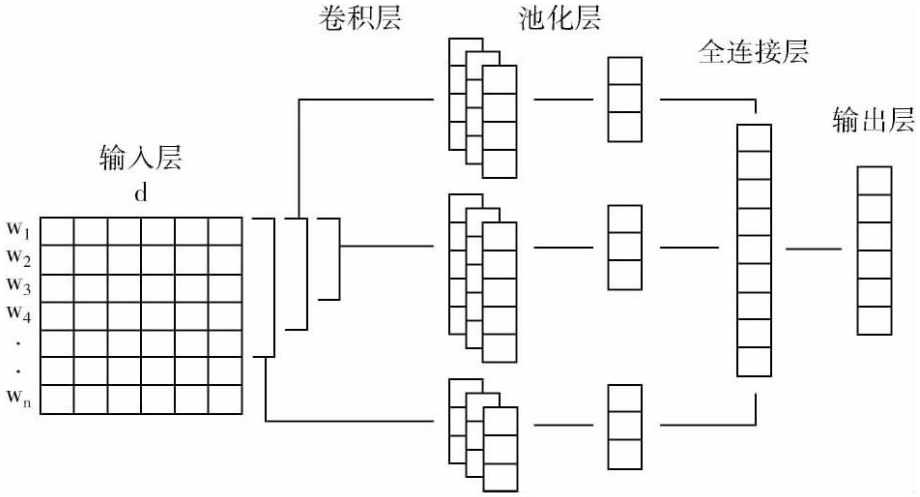

图 7 卷积神经网络的结构

输入层：输入层的输入是一个代表文本的矩阵，d 代表词向量的维度，n 代表每个数据所包含的词向量的数量。

卷积层：卷积层涉及卷积核 $W \in R^{hk}$,h 表示卷积窗口大小，k 为卷积维度，等于词向量的维度。一般来说，$W_{i:i+h}$ 表示单词 W_i，W_{i+1}，W_{i+h}。所以生成一个文本特征的表达式为 $c = f(W \bullet W_{i:i+h})$，其中 b 为偏置，f 为非线性函数。将此卷积核应用于 $(W_{1:h}, W_{2:h+1} \cdots W_{N-h+1:N})$ 生成一个特征映射 $c = (c_1, c_2 \cdots c_{N-h+1})$。

池化层：池化层采用最大池化的方法对特征映射进行特征采样，仅保留每个特征组最重要的特征：$C_{\max} = \max_{c_i \in c} c_i$

全连接层：将池化层输出的多个特征向量进行拼接并输入到全连接层的输入。

输出层：使用 softmax 层进行输出，输出结果是所有类别的概率分布。系统采用自定义的数据集作为网络的输入层，对其输出进行 fine-tuning，并重新定义全连接层。通过全连接层变换公式，选用 tf.nn.softmax 函数作为输出的分类器，

$$\mathrm{softmax}(y_i) = y_i^{'} = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_i}}$$

中i即某一物品类别，从而输出概率分布，实现多类。

3.3.5 Tensorflow

在迁移训练完成之后,在 TensorBoard(Tensorflow[7]的训练可视化工具)当中可视化训练过程。之前在 Summary 数据集中保存的节而 TensorBoard 将会解析这个数据包,并绘制模型的训练图表。TensorBoard 的输出表格数据较为繁复,将其简化后重新绘制为图 8、9。

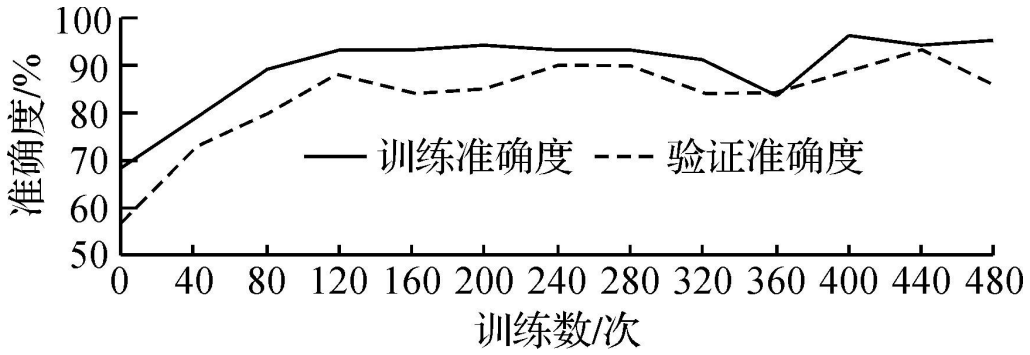

图 8 准确度

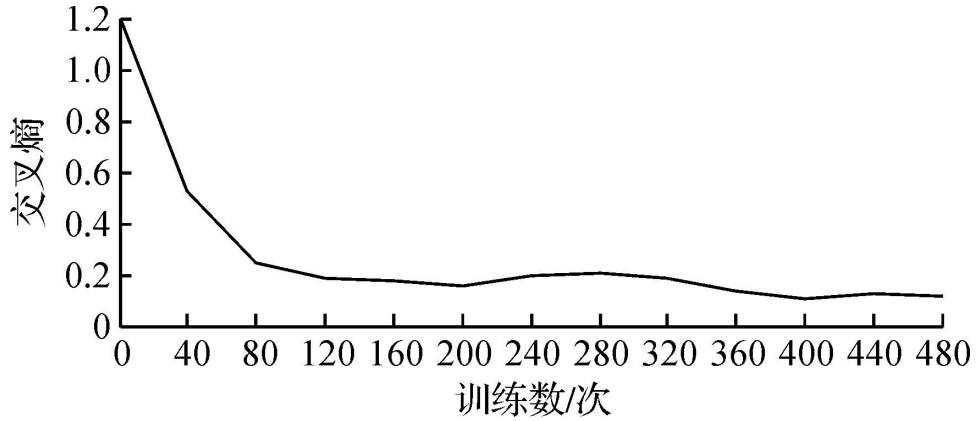

图 9 交叉熵

3.4 建立模型

3.4.1 保存模型与模型可视化

附件 2 经过的数据预清理、文本向量化以及 CNN 卷积神经网络训练、分类已经得到了一级标签分类模型,利用 from keras.utils import plot_model 导出模型。

3.4.2 训练过程可视化

对 CNN 卷积神经网络训练过程进行检测评价，如图 10、11 是训练过程的损失值图和训练过程的准确值图，在 epoch=1 时训练准确值和损失值趋于平稳，建立留言内容的一级标签分类模型相对较优。

图 10 训练准确值

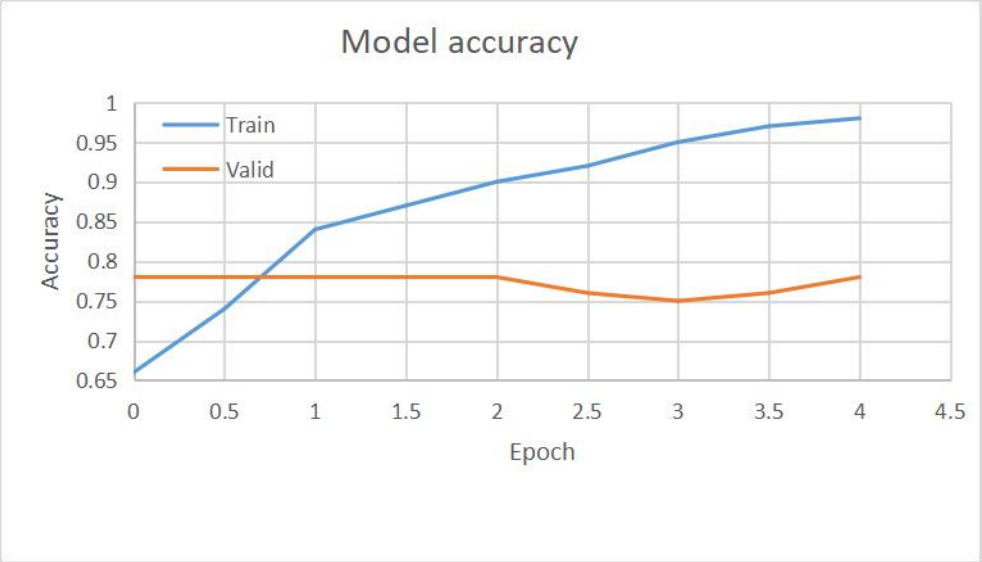

图 11 训练损失值

4. 问题 2 的分析方法与过程

4.1 流程图

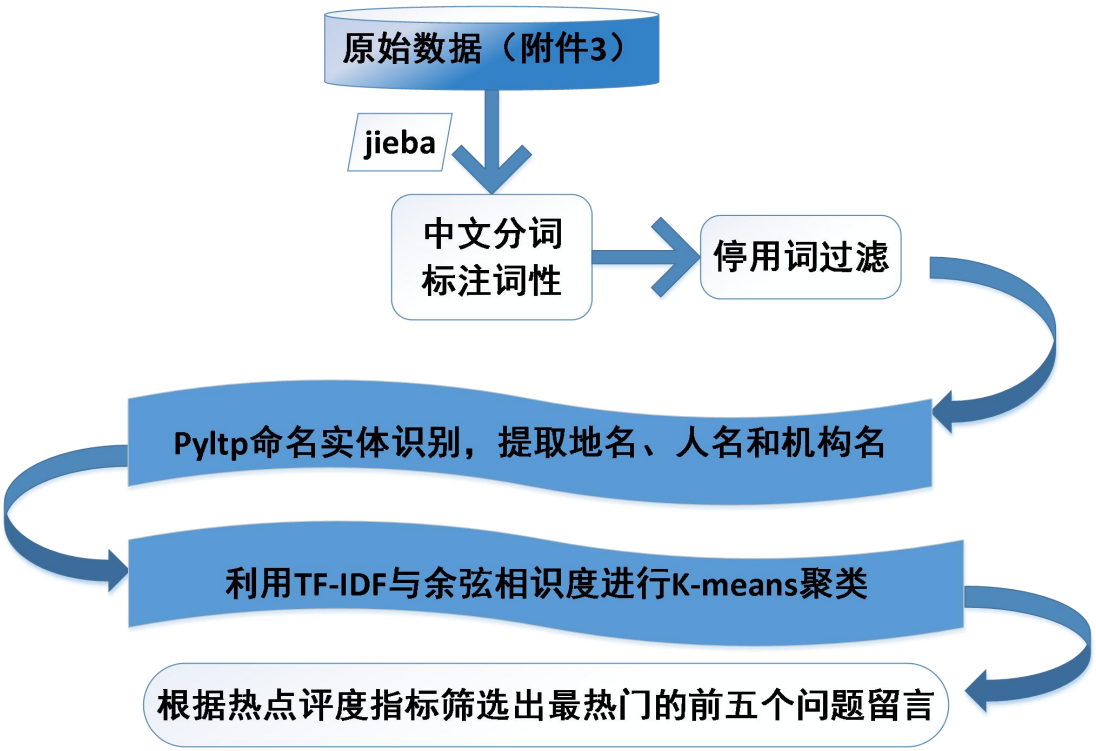

图 12 问题 2 流程图

4.2 数据预处理

在对热点问题进行挖掘分析之前，先采用 python 的中文分词包 jieba 对留言主题进行分词和词性标注，为后面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做准备，然后进行停用词过滤，，过滤掉无意义的词，提高后面搜索的效率。

4.3 pyltp 的命名实体识别

命名实体识别（Named Entity Recognition，简称 NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。简单的讲，就是识别自然文本中的实体指称的边界和类别。

Pyltp 简介[8]：

pyltp 是 LTP 的 python 封装，提供了分词，词性标注，命名实体识别，依存句法分析，语义角色标注的功能，LTP 采用 BIESO 标注体系。pyltp 在命名实体识别中，B 表示实体开始词，I 表示实体中间词，E 表示实体结束词，S 表示单独成实体，O 表示不构成命名实体，它提供的命名实体类型为:人名（Nh）、地名（Ns）、机构名（Ni）。

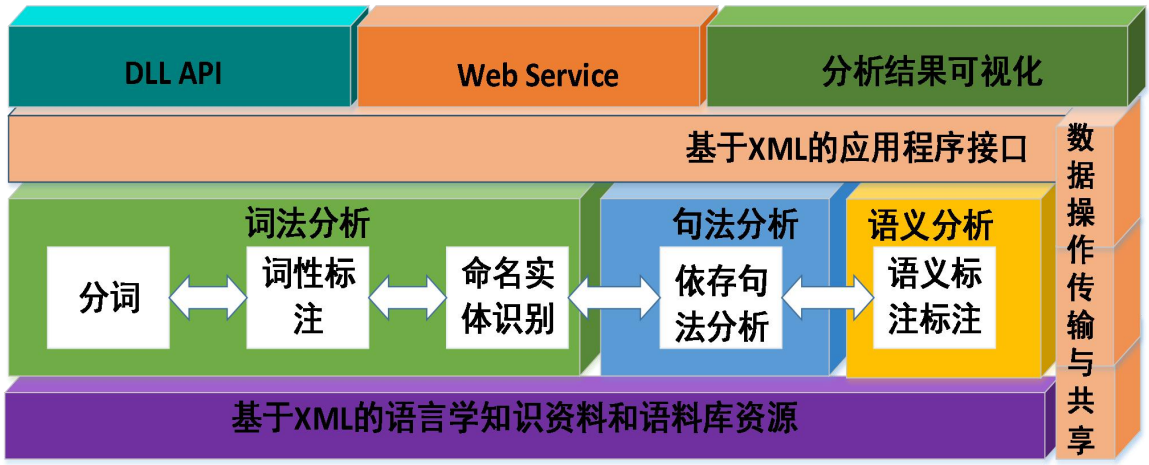

图 13pyltp 简介图

地名识别结果：

留言主题	地名
A3区一米阳光婚纱艺术摄影是否合法纳税了？	A3区
咨询A6区道路命名规划初步成果公示和城乡门牌问题	A6区
反映A7县春华镇金鼎村水泥路、自来水到户的问题	A7县春华镇金鼎村
A2区黄兴路步行街大古道巷住户卫生间粪便外排	A2区黄兴路步行街大古道
A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	A市A3区中海国际社区
A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质	A3区麓泉社区
A2区富绿新村房产的性质是什么？	A2区富绿新村
对A市地铁违规用工问题的质疑	A市
A市6路公交车随意变道通行	A市

图 14 地名识别部分结果显示

4.4 TF-IDF 权值向量

词频-逆向文档频率[9] (Term Frequency-Inverse DocumentationFrequency,TF-IDF)

4.4.1 传统的 TF-IDF：

词频(Term Frequency，TF)是词语在文本中出现的频率，如果某一个词在一个文本中出现的越多，它的权重就越高，基本公式：

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

以上式子$n_{i,j}$中 是该词在文件d_j中的出现次数，而分母则是在文件d_j中所有字词的出现次数之和。

逆向文档频率(Inverse Documentation Frequency，IDF)是指在少数文本中出现的词的权重比在多数文本中出现的词的权重高，因为在聚类中这些词更具有区分能力。它的基本公式如下：

$$\mathrm{id}f_i = log\frac{N}{|\{j:t_i \in d_j\}|}$$

其中， N ：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 $n_{i,j} = 0$ 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$1+|\{j:t_i \in d_j\}|$。

在 Shannon 的信息论的解释中:如果特征项在所有文本中出现的频率越高，它所包含的信息熵越小；如果特征项集中在少数文本中，即在少数文本中出现频率较高，则它所具有的信息熵也较高。

最后可以得出：

$$w_{ij} = tf_{ij} \times idf_i$$

就是词的权重。上述方法各有利弊，IG 计算量相对其它几种方法较大；对于 MI 方法，在相同的条件概率下，稀有名词会比一般词获得更高的得分；x^2方法基于x^2分布，如果这种分布被打破，则对低频词不可靠。因此本文采用 TF-IDF 算法抽取特征词条，将权重按照从大到小的顺序排列，抽取权重最大的前 5000 个特征词作为候选特征词。

4.4.2 TF-IDF 权值向量:

由于计算机不能够直接处理文本信息，我们需要对文本进行处理，将文本表示成为计算机能够直接处理的形式，即文本数字化。文本表示[10] (Test Expression)也称为文本特征表达，它不仅要求能够真实准确的反映文档的内容，而且要对不同的文档具有区分能力。目前常用的文本表示模型有布尔模型、向量空间模型和概率模型等。

而向量空间模型[11] (Vector Space Model, VSM)最早是由 Salton 和 McGill 于 20 世纪 60 年代末提出的，是目前在文本挖掘技术中最常用的表示模型。

其主要思想：将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项(词条)对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重，这里的权重可以由 TF-IDF 等算法得到。向量空间模型就是将文本表示成为一个特征向量：

$$V(d) = (t_1, w_1(d), w_2(d), \ldots, t_n, w_n(d))$$

其中$t_i(i = 1,2,\ldots,n)$为文档 d 中的特征项，$w_i(i = 1,2,\ldots,n)$为特征项的权值，可由 TF-IDF 算法得出。

4.4.3 文本向量化表示:

上述文本特征抽取，这时需要构建一个词袋，根据留言主题的特征项对应词袋中的位置，组成统一维数的向量：

$$B = (t_1, t_2, \ldots, t_n)$$

其中 B 为词袋集合，t_n是每个词在向量中对应的位置。

这样网络招聘文本信息根据词袋组成了同一维数的词向量，再通过 TF-IDF 将它们

向量化得到一个词汇-文本矩阵：

$$\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix}$$

部分结果显示：

	A	B	C	D	E	F	G	H	I	J
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	1
3	0	0	0	0	0	3	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	1	0
6	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	4	0
9	0	0	0	2	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	1
12	1	1	0	0	0	1	0	0	0	0
13	0	0	0	0	1	1	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0
15	1	0	0	1	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	0	0	1	0
19	0	0	0	0	0	1	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0

表 1 词频矩阵

	A	B	C	D	E
1	0.0000000000	0.0000000000	0.0000000000	3.2354761985	0.0000000000
2	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
3	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
4	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
5	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
6	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
7	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
8	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
9	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
10	0.5890324110	0.0000000000	0.0000000000	0.0000000000	0.0000000000
11	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
12	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
13	0.0000000000	0.0000000000	0.0000000000	0.0854714869	0.0000000000
14	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
15	0.0000000000	0.0000000000	0.1257894110	0.0000000000	0.0000000000
16	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
17	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
18	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
19	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
20	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

表 2 TF-IDF 权重矩阵

4.5 余弦相识度

余弦相似度[12]通常用在文档相似度判断上，是利用两个向量夹角的余弦值来衡量两个向量的差异的大小，余弦值越靠近 1，就表明夹角越接近于0°，也就是两个向量越相似，余弦相似度不考虑两个数据对象的量值。

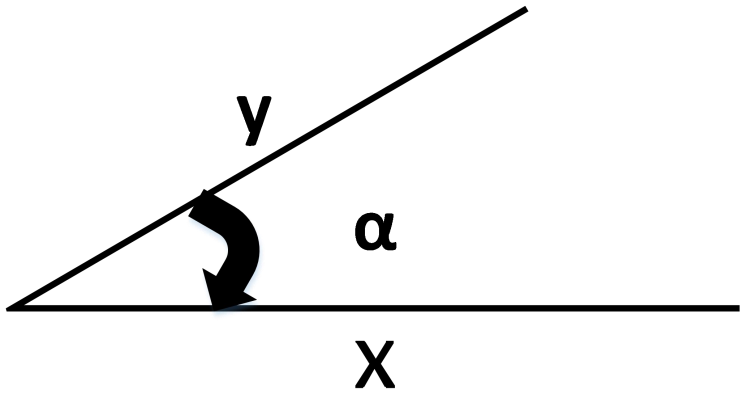

图 15 余弦相似度的几何解释

如图所示，边x与y的余弦相似度是边x与y之间的夹角α的余弦值。因此，如果余弦相似度为 1，则x与y之间的夹角为0°，此时除了长度外x和y是相同的，如果余弦相似度为 0，则x与y的夹角为90°，x与y完全不相似。

在二维空间，根据向量点积公式，显然：

$$\cos\alpha = \frac{x \cdot y}{\|x\|\|y\|}$$

其中:$\|x\|$是向量x的长度，$\|x\|=\sqrt{x^2}$，$\|y\|$是向量y的长度，$\|y\|=\sqrt{y^2}$。

假设向量x、y的坐标分别为(x_1,y_1)、(x_2,y_2)。则：

$$\cos\alpha = \frac{x_1x_2 + y_1y_2}{\sqrt{{x_1}^2 + {y_1}^2} \times \sqrt{{x_2}^2 + {y_2}^2}}$$

推广到多维：$x = (x_1,x_2,\dots,x_n)$,$y = (y_1,y_2,\dots,y_n)$，则：

$$\cos\frac{\sum_1^n(x_iy_i)}{\sqrt{\sum_1^n {x_i}^2} \times \sqrt{\sum_1^n {y_i}^2}}$$

注：

算法：对任意两行数据（m 行、n 行）执行下面程序：

$\mathrm{CompCos}\alpha(m,n,\cos\alpha)$

1. 当 i 小于指标数时继续执行，否则跳到第 5 步（i从 0 开始记录已计算指标数）
2. 计算 m 行与 n 行对应 i 指标数值的乘积$(x_{mi} \times y_{ni})$，并累加。
3. 计算 m 行对应 i 指标的平方${x_i}^2$，并累加。
4. 计算 n 行的对应 i 指标的平方${y_i}^2$，并累加。
5. 对 2、3、4 步加结果计算$\cos\alpha$。
6. 返回$\cos\alpha$。

4.6 Kmeans 聚类

所谓文本聚类就是将无类别标记的文本信息根据不同的特征，将有着各自特征的文本进行分类，使用相似度计算将具有相同属性或者相似属性的文本聚类在一起。这样就可以通过文本聚类的方法根据岗位职责与任职要求，对不同职位进行分类。通过聚类方法，求职者可以结合自身状况更加快捷地获取相关信息资源。

K-means 聚类原理：

K-means 算法[13]是很典型的基于划分的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似性就越大。

K-means 算法的基本思想是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

假设要把样本集分为 k 个类别，算法描述如下：

（1）适当选择 k 个类的初始中心；

（2）在第 k 次迭代中，对任意一个样本，求其到 k 个中心的距离，将该样本归到距离最短的中心所在的类；

（3）利用均值等方法更新该类的中心值；

（4）对于所有的 k 个聚类中心，如果利用(2)(3)的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

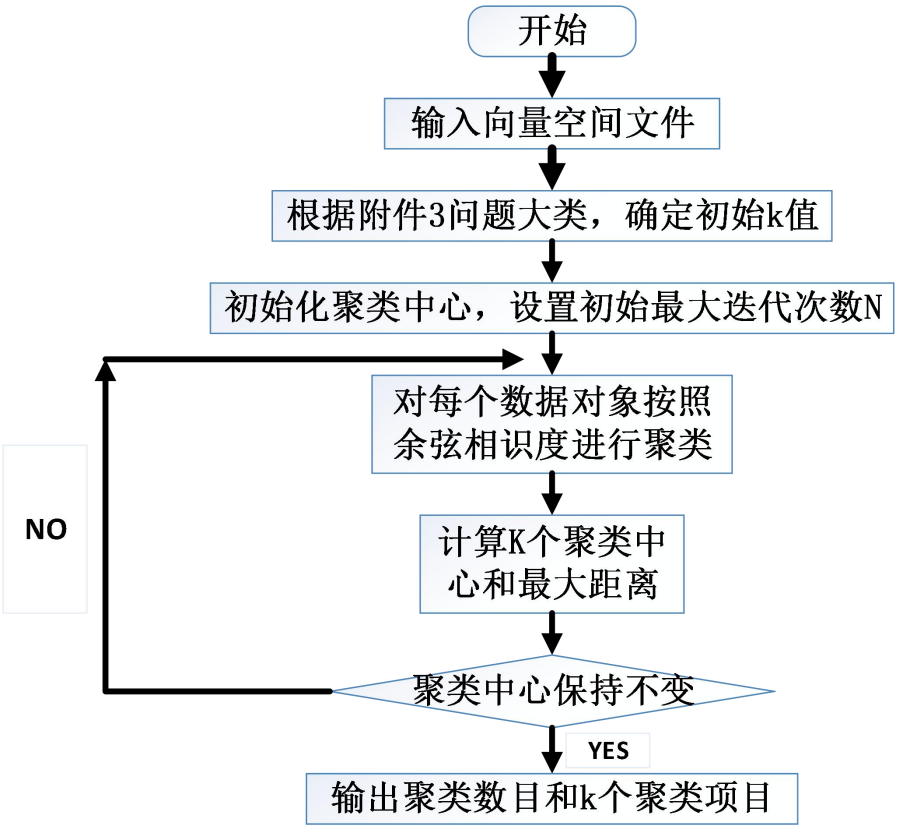

图 16K-means 聚类流程

4.7 结果分析

通过上述方法，将留言问题进行分类，统计出每个问题的留言数目，但每个问题的点赞数和反对数在一定程度上对热度有影响，因此需要制定热度指标评价方案。

热度指标评价方案制定根据了每个问题的留言数、反对数和点赞数，规定留

言数是评价热度的重点指标，点赞数与反对数作为参考指标。

因此，热度评价指标为：

$$S = n + \frac{i - j}{100}$$

其中，S 为热度指数，i 为点赞数，j 为反对数。

前五热度问题结果显示：

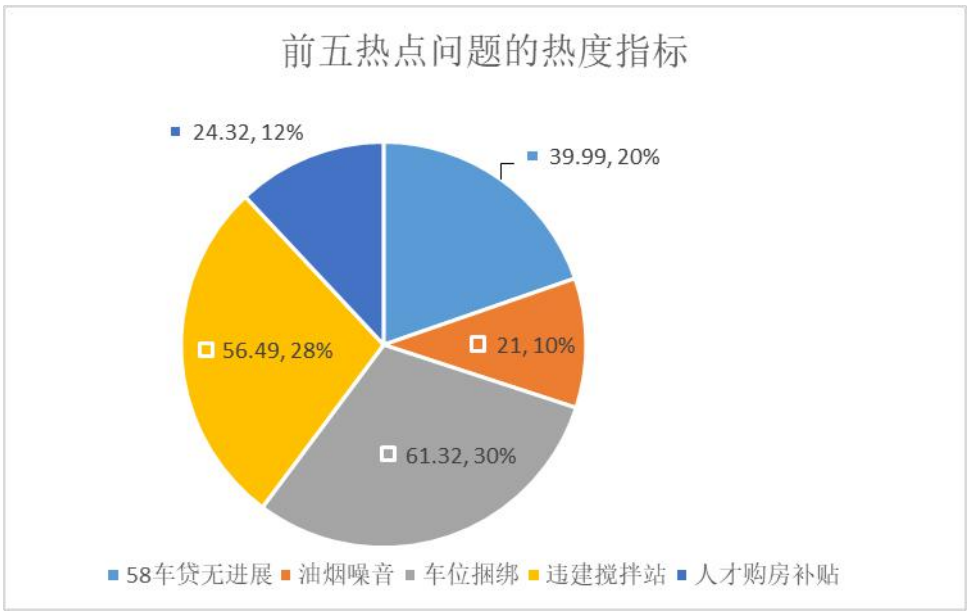

图 17 前五热度指标结果显示

	A	B	C	D	E	F
1	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
2	1	1	61.32	2019/07/01至2019/09/01	伊景园滨河苑广铁集团铁路职工	开发商捆绑车位销售
3	2	2	56.49	2019/04/20至2020/01/26	A市A2区丽发新城小区	小区附近违建搅拌站，污染环境和噪音扰民
4	3	3	39.99	2019/01/08至2019/07/8	A市A4区西地省	58车贷案件近半年毫无进展
5	4	4	24.32	2018/11/15至2019/12/2	A市人才	人才购房补贴申请失败
6	5	5	21	2019/04/20至2020/01/26	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民

表 3 前五热点问题结果显示

	A	B	C	D	E	F	G	H
1	问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
2	1	190337	A00090519	园滨河苑捆绑销售车位的	2019-08-23 12:22:00	投诉伊景园.滨河苑开发商捆绑	0	0
3	1	191001	A909171	滨河苑协商要求购房同时必	2019-08-16 09:21:33	商品房伊景园滨河苑项目是由A	12	1
4	1	195511	A909237	车位捆绑违规销售	2019-08-16 14:20:26	对于伊景园滨河苑商品房，A市	0	0
5	1	196264	A00095080	A市伊景园滨河苑捆绑车位	2019/8/7 19:52:14	A市伊景园·滨河苑现强制要求	0	0
6	1	199190	A00095080	武广新城违法捆绑销售车	2019/8/1 22:32:26	武广新城为铁广集团的定向商品	0	0
7	1	204960	A909192	本来就困难,还要捆绑买卖	2019-08-21 18:12:20	我是广铁集团铁路职工,因家人	0	0
8	1	205277	A909234	滨河苑捆绑车位销售合法	2019-08-14 09:28:31	广铁集团强制要求职工购买伊景	1	0
9	1	205982	A909168	对伊景园滨河苑强制捆绑销	2019-08-03 10:03:10	我坚决反对伊景园滨河苑捆绑销	2	0
10	1	206355	A909189	售卖内部福利房强行捆绑	2019-08-29 18:26:03	一、做为A市伊景园滨河苑项目	0	0
11	1	207243	A909175	滨河苑强行捆绑车位销售	2019-08-23 12:16:03	您好！A市武广新城片区的伊景	0	0
12	1	209506	A909179	城坑客户购房金额并且捆	2019-08-02 16:36:23	您好！由A市广铁集团发起的定	0	0
13	1	209571	A909200	对苑项目绑定车位出售是不	2019-08-28 19:32:11	广铁集团铁路职工定向商品房供	0	0
14	1	212323	A00020702	要求员工购房时必须同时	2019-07-11 00:00:00	尊敬的领导：您好！我是一名新	0	0
15	1	218709	A000106692	伊景园滨河苑捆绑销售车	2019/8/1 22:42:21	伊景园滨河苑作为广铁集团定向	1	0
16	1	222209	A00017171	苑定向限价商品房项目违	2019-08-28 10:06:03	广铁集团与伊景园滨河苑开发商	0	0
17	1	223247	A00044759	A市伊景园滨河苑捆绑销售	2019/7/23 17:06:03	关于铁广集团铁路职工定向商品	0	0
18	1	224767	A909176	车位捆绑销售！广铁集团	2019-07-30 14:20:08	伊景园滨河苑车位捆绑销售！广	0	0
19	1	225479	A00043800	发有限公司违规操作铁广	2019/7/5 1:55:26	A市市政建设开发有限公司违规	0	0
20	1	229731	A00043800	得购房资格的职工购房为	2019/7/2 11:38:30	为什么现在开发商要绑定销售车	0	0

表 4 热点问题明细部分结果显示

5. 问题 3 过程分析与流程图

5.1 流程图

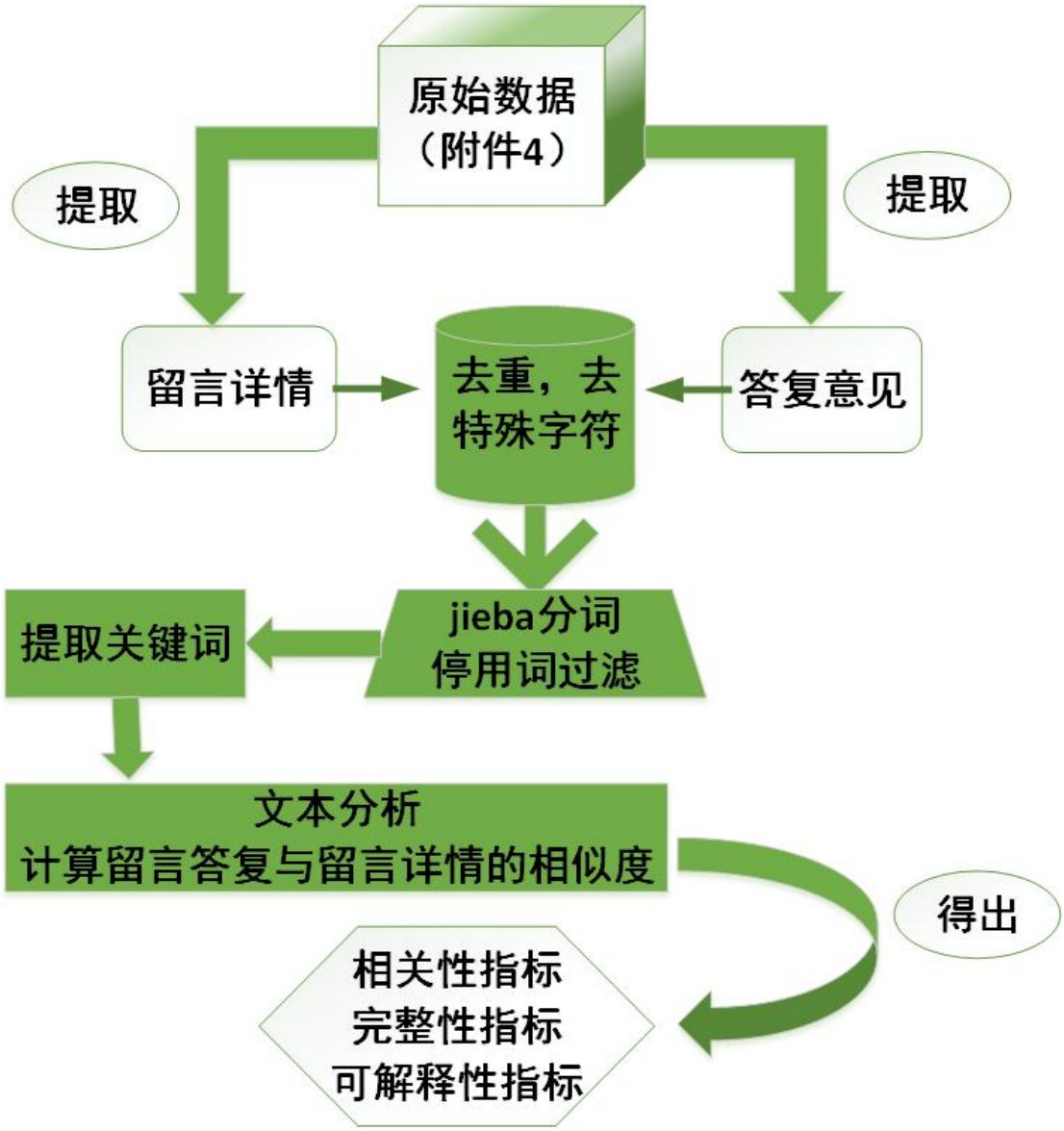

图 18 问题 3 流程图

5.2 数据预处理

5.2.1 留言详情、答复意见的去重去特殊字符

在题目给出的信息中，出现了很多重复的留言及答复信息，考虑到对信息处理的方便性及准确度，便使用 Python 进行了去重。信息读取后出现了很多特殊字符，为了后续处理的方便性，也使用 Python 将其进行去掉。去重后的信息及去特殊字符后的留言详情及答复意见分别保存在附件 4 去重.csv，附件 4 去特殊字符 1.csv，附件 4 去特殊字符 2 中。

5.2.2 停用词过滤

去重去特殊字符后的信息仍然存在大量的无意义词语，这对我们后续提取关键词造成了很大麻烦，因此，需要对其进行停用词过滤处理，去除“的”“了”等字词，要想去除这类词，需要先分词，使用 Python 进行 jieba 分词后，利用去停用词表进行去停用词，使信息篇幅简短了很多的同时，对我们后续对于关键词的提取也方便了很多。停用词过滤后的留言详情和答复意见分别保存在附件 4 去停用词 1.csv，附件 4 去停用词 2.csv。

5.2.3 关键词提取

去重去特殊字符及停用词过滤后的文本信息看起来简洁了不少，但是要判断留言详情和答复意见的相似度，这些文本信息还是太多了，所以要对这些文本信息进行关键词提取，使用 Python 对每一条留言详情和答复意见分别提取 10 个关键词，以便后续进行相关性处理。提取的留言详情和答复意见的关键词保存在附件 4 关键词.csv

5.2.4 分析关键词得出各项指标

比较留言详情和答复意见的关键词，可得出留言详情和答复意见的相关性指标、答复意见的完整性和可解释性指标如下

相关性指标	**0.824**
完整性指标	0.867
可解释性指标	0.815

表 5 相关性、完整性和可解释性指标表

柱状图如下:

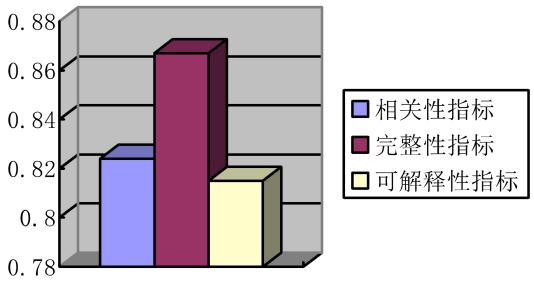

图 18 相关性、完整性和可解释性指标柱状图

由以上留言详情和答复意见的相关性指标、完整性指标、可解释性指标可以看出，对于相对应留言的答复意见的质量还是很高的，高质量的答复意见，高质量的办事态度，高质量的服务水平，给政府点赞！

总结本次比赛，我们用 jieba 中文文本分词、去停用词等数据预处理，基于 TextRank 提取关键词，用 Token 字典和 Embedding 将中文文本转化成数字列表，通过 CNN 卷积神经网络进行分类和建立一级标签分类模型；并根据对附件 3 进行命名实体识别，识别出地点、机构名和人名，对这三个类别进行数据数值化处理后，利用余弦相识度计算文本相识度，再通过 k-means 聚类算法对热点问题进行划分；同样利用余弦相似度对留言详情和答复意见中的关键词处理，并根据相关性、完整性、可解释性三个指标，对答复意见的质量进行合理性评价。

但是我们最后的用 k-means 聚类时以后，在涉及热点评价排名是技术不够，只能利用 Excel 进行排名，对在定义意见答复的相关性、准确性、可解释性指标时不够合理。我们后期也会进一步第文本挖掘进行深入讨论。

[1]祝永志,荆静.基于 Python 语言的中文分词技术的研究[J].通信技术,2019,52(07):1612-1619.
[2] 李志强,潘苏含,戴娟,胡佳佳.一种改进的 TextRank 关键词提取算法[J].计算机技术与发展,2020,30(03):77-81.
[3] 哈工大停用词表.https://wenku.baidu.com/view/b3275a66f5335a8102d2200d.html
[4] Token 的使用的整体流程.https://blog.csdn.net/weixin_30628077/article/details/101859541
[5] [经典论文]CNN 文本分类.https://mp.weixin.qq.com/s/6_mvPGq-b1p7Ci_FoqAZ3g
[6] 吴碧程,邓祥恩,张子憧,唐小煜.基于卷积神经网络的智能垃圾分类系统[J].物理实验,2019,39(11):44-49
[7] 彭宏,庄宁.基于 Tensorflow 的 Android 端相册分类 App 设计与实现[J].浙江工业大学学报,2020,48(02):165-172.
[8] Pyltp 简介：原文链接：https://segmentfault.com/a/1190000018081013
[9] 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.
[10] 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.
[11] 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.
[12]基于余弦相识度的聚类算法在统计调查对象分类中的应用研究_王习涛
[13] 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.