“智慧政务”中的分析与挖掘

摘要：近年来，随着各类社情民意相关的文本数据量不断增加，依靠人工来进行留言划分和热点整理存在工作量大、效率低，且差错率高等问题。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。本文将基于数据挖掘技术对来自互联网公开渠道的群众问政留言记录数据进行内在的信息挖掘，提取我们需要进行分析的部分进行深度挖掘和分析。

对于问题一，在对群众问政留言记录文本数据（附件 2）进行 jieba 进行中文分词、停用词过滤等数据预处理后，基于 TFIDF 权重法提取关键词，用 word2vec 特征提取生成词向量，按照一定的划分体系对留言进行分类，再用 CNN（卷积神经网络）算法进行文本分类，取训练集中的百分之十充当测试集对模型进行检验。最后用 F-Score 对分类方法进行评价。通过评价优劣对分类模型进行调整。

对于问题二，在对群众反映的问题文本数据（附件 3）进行数据预处理后，基于 TFIDF 权重法提取关键词，采用基于 TFIDF 权重法及奇异值分解(SVD)的潜在语义分析的 k-means 聚类算法，对群众集中反映的某些问题进行聚类，根据热度评价指标给出评价结果，并整理出排名前 5 的热点问题以及相应热点问题对应的留言信息。

对于问题三，在相关部门对留言的答复意见（附件 4）进行数据预处理后，利用从答复的相关性、完整性、可解释性三个角度对答复意见的质量给出一套评价方案。

关键词：TFIDF；word2vec；CNN；K-means 文本聚类

Analysis and Digging in "Smart Government Affairs"

Abstract: In recent years, with the increasing amount of text data related to various social conditions and public opinion, relying on manual to divide the message and organize hotspots has problems such as large workload, low efficiency, and high error rate. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of a smart government system based on natural language processing technology has become a new trend of social governance innovation and development. This article will use the data mining technology to carry out internal information mining on the public questionnaire message data from the public channels of the Internet, extract the parts we need to analyze and conduct in-depth mining and analysis.

For question 1, after jieba preprocesses Chinese word segmentation and stop word filtering on text data (Annex 2) of the masses 'message records, extract keywords based on TFIDF weighting method, use word2vec feature extraction to generate word vectors, according to A certain classification system is used to classify the messages, and then the CNN (Convolutional Neural Network) algorithm is used to classify the text. Ten percent of the training set is used as the test set to test the model. Finally, use F-Score to evaluate the classification method. Adjust the classification model by evaluating the advantages and disadvantages.

For question 2, after preprocessing the text data (Annex 3) of the questions reflected by the masses, keywords are extracted based on TFIDF weighting method, and k-means based on latent semantic analysis based on TFIDF weighting method and singular value decomposition (SVD) The clustering algorithm clusters certain problems reflected by the masses, gives the evaluation results according to the heat evaluation index, and sorts out the top 5 hot issues and the corresponding message information of the corresponding hot issues.

For question three, after the data preprocessing of the response comments (Annex 4) of the message by the relevant departments, a set of evaluation schemes are given for the quality of the response comments from the perspectives of relevance, completeness, and interpretability of the responses.

Key words: TFIDF; word2vec; CNN; K-means text clustering

目录

1. 挖掘目标 4
2. 分析方法与过程 4
2.1 问题一分析方法与过程 5
2.1.1 流程图 5
2.1.2 数据分析 5
2.2 问题二分析方法与过程 10
2.2.1 流程图 10
2.2.2 数据分析 10
2.3 问题三分析方法与过程 13
2.3.2 流程图 13
2.3.2 数据分析 13
3. 结果分析 14
3.1 问题一 结果分析 14
3.1.1 结果分析 14
3.1.2 数据验证 15
3.2 问题二 结果分析 16
3.2.1 结果分析 16
3.3 问题三 结果分析 18
3.3.1 结果分析 18
4. 结论 18
5. 参考文献 18

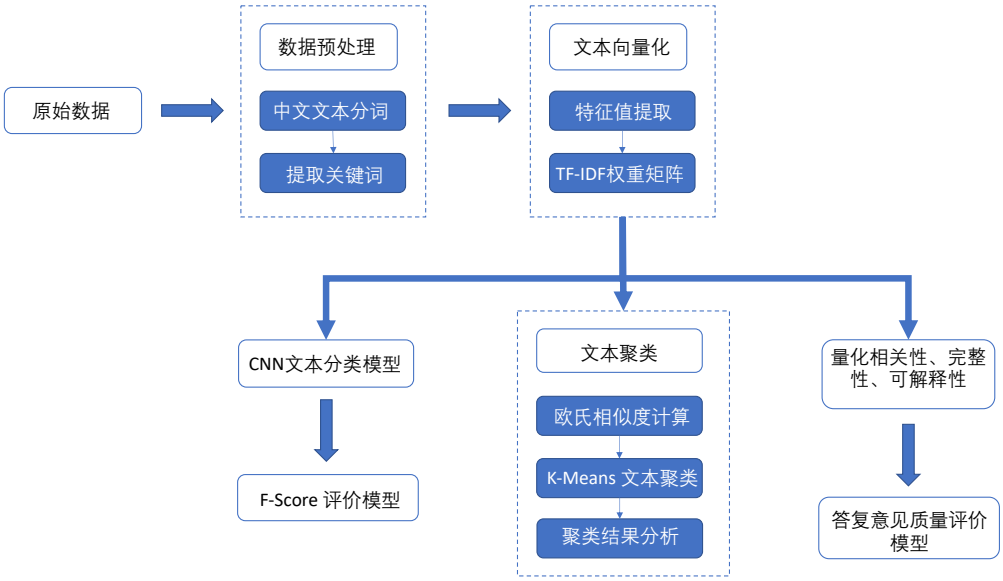
1. 挖掘目标

近年来，随着各类社情民意相关的文本数据量不断增加，依靠人工来进行留言划分和热点整理存在工作量大、效率低，且差错率高等问题。

本次建模目标是利用来自互联网公开渠道的群众问政留言记录数据，在对群众问政留言记录文本数据进行基本的预处理——中文分词、停用词过滤后，第一、按照一定的划分体系对留言进行分类，利用 word2vec 特征提取生成词向量，再用 CNN（卷积神经网络）算法进行文本分类，最后用 F-Score 对分类方法进行评价。第二、采用基于 TFIDF 权重法及奇异值分解(SVD)的潜在语义分析的 k-means 聚类算法，对群众集中反映的某一问题进行聚类，根据热度评价指标给出评价结果，并整理出排名前 5 的热点问题以及相应热点问题对应的留言信息;第三、构造排名算法判断热门行业、职位、地域，并 引入时间因素预测短期人才需求走向;对大数据相关新兴职位，挖掘各个指标内在关联 规则，深入分析其需求增长趋势、行业分布情况、地域分布情况、行业职位特征、行业 薪酬情况以及技能要求;并根据目前高校人才培养方案与实际行业需求的差距，提出参考意见。

2. 分析方法与过程

总体流程图



本用例主要包括如下步骤:

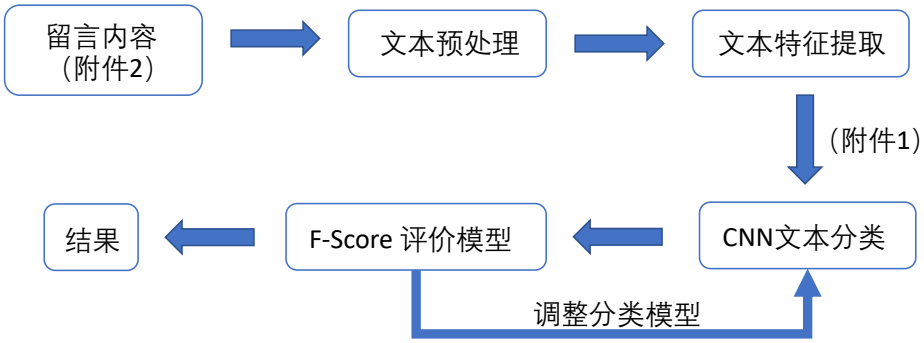
步骤一:数据预处理，对附件 2 非结构化文本去除重复项及空行、中文文本分词、停用词过滤，以便后续分析;

步骤二:数据分析，在对留言内容分词后，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，找出群众留言的关键词，在问题一中，用 word2vec 把群众留言信息转换为权重向量。利用 CNN（卷积神经网络）算法进行文本分类。在问题二中，利用欧氏距离的相似度计算，采用 K-means 算法对群众留言进行分类，根据热度评价指标给出评价结果。在问题三中，用余弦相似度进行相似性的量化。
步骤三：结果分析，在问题一中，取训练集中的百分之十充当测试集对模型进行检验，再用 F-Score 对分类方法进行评价。在问题二中，得到排名前 5 的热点问题以及相应热点问题对应的留言信息。

2.1 问题一分析方法与过程

2.1.1 流程图



2.1.2 数据分析

在题目所给的数据中，数据量较大(共上万多条记录)，且附件 2、3、4 中的字段大多为文本格式，有大量空行以及重复的情况，如果不做处理会对后续分析造成影响，并且留言文本信息存在大量噪声特征，如果把这些数据也引入进行分词、词频统计乃至文本聚类等，则必然会对聚类结果的质量造成很大的影响，于是本文首先要对数据进行预处理。

2.1.2.1 文本预处理
（一）jieba 进行中文分词[1]的原理
1）基于 Trie 树结构[2]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)
2）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
3）对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

（二）TF-IDF 算法[3]
TF（Term Frequency，词频）

一个词在文章中出现的频率，频率越大，说明这个词对这篇文章更重要.。对于某一篇文章这的某一个词来说，其 TF 的表达式如下：

$$TF = \frac{文章中该词出现的次数}{文章中所有词出现的次数之和}$$

但是，当这个词在几乎所有的文章中的频率都很高的时候，那么这个词对这篇文章就不重要了。

IDF（Inverse Document Frequency，逆向文件频率）

在所有文章中，若包含某个词的文章数量越少，则这个词更能有助于在整个语料库中区分出这篇文章，对于某一个词来说，其 IDF 的表达式如下：

$$IDF = \log\left(\frac{语料库的文章总数}{包含该词的文章数 + 1}\right)$$

IDF 越小，则对应词在语料库中的分布越分散，则其对文章归类的作用越小。

TF-IDF

这实际上是结合了词频和逆向文档频率，它的主要思想是，一个词在一片文章中出现的次数越多，而在其它文章中出现的次数越少，则该词具有很好的区分能力。

以下为文本预处理的一个示例：

```
string = data['留言详情'][0]
data = data.astype(str)
```

```
print(string)
```

```
A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市A市，尽快整改这个极不文明的路段。
```

```
# 分词

print(' '.join(jieba.cut(string)))
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\86138\AppData\Local\Temp\jieba.cache
Loading model cost 0.934 seconds.
Prefix dict has been built successfully.
```

```
A3 区 大道 西行 便 道 ， 未管 所 路口 至 加油站 路段 ， 人行道 包括 路灯 杆 ， 被 圈 西湖 建筑 集团 燕子 山 安置 房 项目 施工 围墙 内 。 每天 尤其 上下班 期间 这 条 路上 人流 车流 极 多 ， 安全隐患 非常 大 。 强烈 请求 文明城市 A 市 ， 尽快 整改 这个 极 不 文明 的 路段 。
```

```
# 抽取关键词
print(jieba.analyse.textrank(string, topK=20, withWeight=False,
                             allowPOS=('ns', 'n', 'vn', 'v'), withFlag=False))
```

```
['安置', '路灯', '车流', '请求', '路段', '包括', '人流', '文明城市', '人行道', '集团', '建筑', '项目', '施工', '围墙', '西湖']
```

图 1 文本预处理

2.1.2.2 文本特征提取

（一）用 word2vec[4]特征提取生成词向量

word2vec 工具主要包含两个模型：跳字模型（skip-gram）和连续词袋模型（continuous bag of words，简称 CBOW），以及两种高效训练的方法：负采样（negative sampling）和层序 softmax（hierarchical softmax）。

CBOW 与 Skip-Gram 用于神经网络语言模型

Skip-Gram 的目标函数如下：

$$l(\theta) = \arg\max_{\theta} \prod_{w \in Text} \prod_{C \in Context} P(c|w;\theta) \tag{1}$$

其中$content(w)$为维度((n-1)m,1)，n-1 指的是窗口里面的单词数。w是要预测的下一个词。式（1）取对数之后的形式为

$$L(\theta) = \arg\max_{\theta} \sum_{w \in Text} \sum_{C \in Context} logP(c|w;\theta) \tag{2}$$

(w, c)在上下文出现，说明它们的相似度很高，然后我们就需要$logP(c|w;\theta)$是越大越好的，那么后面就用到的 softmax 的方式来得到一个概率。

其中参数θ是两次词向量的矩阵

$$\theta = [u, v]$$

其中u代表了单词作为上下文的词向量，v代表了单词作为中心词的词向量。这个在 word2vec 的源代码中也是看得到的。

通过 softwmax 处理之后，我们就能得到

$$P(c|w;\theta) = \frac{e^{u_c * v_w}}{\sum_{c' \in corpus} e^{u_{c'} * v_w}} \tag{3}$$

这里的上下文是语料库中的。将公式（3）带入到目标函数式（2）中去，就可以得到

$$L(\theta) = \arg\max_{\theta} \sum_{\omega \in Text} \sum_{C \in Context} [u_c * v_w - log \sum_{c' \in corpus} e^{u_{c'} * v_w}] \tag{4}$$

即得 Skip-Gram 模型的目标函数。为了解决计算的时间复杂度很高的问题，目前使用了两种手段：一个是负采样，另外一个是层次化的 softmax。其中负采样过程中的目标函数为

$$L(\theta) = \arg\max_{\theta} [\sum_{(c,w) \in D} log\sigma(u_c \cdot v_w) + \sum_{c' \in N(w)} log\sigma(-u_{c'} \cdot v_w)] \tag{5}$$

再对各个未知数求一个偏导再采用梯度下降法去更新参数。

2.1.2.3 CNN 文本分类模型[5]

对于文本的分类，我们采用了 CNN（卷积神经网络）的算法。使用了 python 的 keras 和 tensorflow 工具。

下面是对 CNN 的一些简单介绍：

CNN 的结构：（卷积层+非线性激活函数(Relu 或 tanh)+池化层）× n + 几个全连接层。
（一）卷积层输入的是一个表示句子的矩阵，维度为 nd，即每句话共有 n 个词，每个词有一个 d 维的词向量表示。假设 Xi:i+j 表示 Xi 到 Xi+j 个词，使用一个宽度为 d，高度为 h 的卷积核 W 与 Xi:i+h-1(h 个词)进行卷积操作后再使用激活函数激活得到相应的特征 ci，则卷积操作可以表示为：（使用点乘来表示卷积操作）

$$c_i = f(w \cdot x_{i:i+h-1} + b) \tag{6}$$

因此经过卷积操作之后，可以得到一个 n-h+1 维的向量 c 形如$c = [c_1, c_2, \dots, c_{n-h+1}]$。以上是一个卷积核与输入句子的卷积操作，同样的，我们也可以使用更多高度不同的卷积核，且每个高度的卷积核多个，得到更多不同特征。
（二）池化层：对每个不重叠的 n*n 的区域进行降采样，有最大池化、平均池化等。池化层可以提将输出矩阵的尺寸固定不变，减少输出维度，但是保存重要特征。

（三）Channels 通道：通道是输入数据的不同“视图”。在 NLP 中把不同类的词向量表征（例如 word2vec 和 GloVe）看做是独立的通道，或是把不同语言版本的同一句话看作是一个通道。

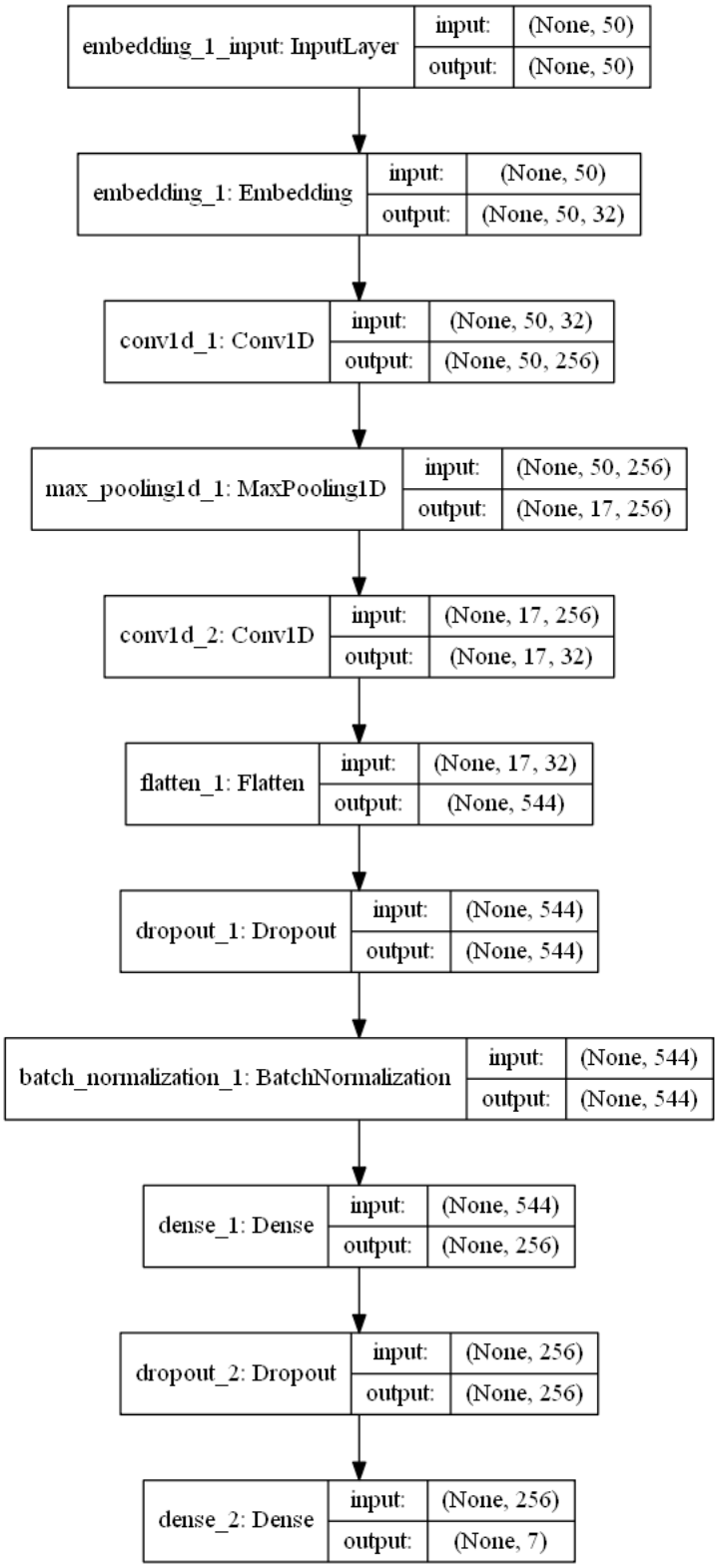


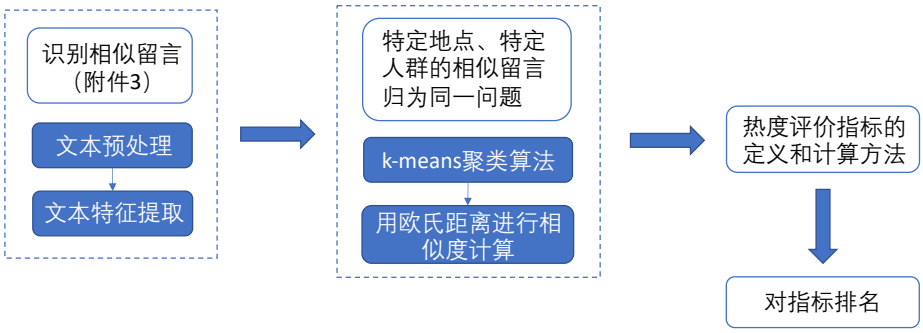
图 2 CNN 网络示意图

上图是我们采用的 CNN 网络示意图。首先进过 Embedding 层降维，之后经过两个卷积层、

一个池化层、一个 Flatten 层。之后的 dropout 层和正则化层都是为了防止缓解过拟合的产生。在附件 2 中，留言一共有 7 个一类标签。因此，我们的 CNN 模型最终输出是取值范围为 0-6 的分类结果。
在损失函数方面，我们采用了稀疏分类交叉熵函数，优化器选择了 adam。

2.2 问题二分析方法与过程

2.2.1 流程图



2.2.2 数据分析

2.2.2.1 文本预处理

附件 3 表格结构如下：

	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
0	188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了？	2019/2/28 11:25:05	\n\t\t\t\t\t\n\t\t\t\t\t座落在A市A3区联丰路米兰春天G2栋320，...	0	0
1	188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	\n\t\t\t\t\t\n\t\t\t\t\tA市A6区道路命名规划已经初步成果公示文件，...	0	1

图 3 附件 3 表格结构

表格的具体信息如下：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4326 entries, 0 to 4325
Data columns (total 7 columns):
留言编号    4326 non-null int64
留言用户    4326 non-null object
留言主题    4326 non-null object
留言时间    4326 non-null object
留言详情    4326 non-null object
反对数     4326 non-null int64
点赞数     4326 non-null int64
dtypes: int64(3), object(4)
memory usage: 236.7+ KB
```

图 4 附件 3 表格详情

一条数据由留言编号、留言用户、留言主题、留言时间、留言详情、赞同数和反对数组成，其中，赞同数、反对数和与之内容相同的留言数目直接反应留言热度，留言时间间接反映留言热度。

	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
430	198766	A000100057	西地省津楚投资有限公司法人失踪半月	2019/11/4 16:23:15	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的A市公安局领导：西地省津楚投资责任有限...	0	0
2114	238649	A000100057	西地省津楚投资责任有限公司资金链断裂，投资者血本无归	2019/10/29 12:45:40	\n\t\t\t\t\t\n\t\t\t\t\t\t西地省津楚投资责任有限公司实际控制人董事长兼...	0	8

图 5 用户就同一点多次留言特例

值得注意的是，表格中存在用户多次为同一件事留言，由于留言是由同一个人产生的，其对留言的热点不具有累加作用。

将留言主题和留言详情的文本结合起来，对其内容进行清洗，消去停用词和格式字符，并进行 jieba 分词，进行词性筛选，保留其中的名词和动词。

2.2.2.2 用 TF-IDF 提取关键词

2.2.2.3 文本特征提取

使用 Countvectorizer（）将文本中的词语转化为词频矩阵，然后运用 TfidfTransformer（）统计每个词的 tf-idf 权值，进而获得句子的特征向量。

2.2.2.4 利用欧氏距离[6]进行文本相似度计算

假设两个文本$X = (x_1, x_2, x_3, \ldots, x_n)$和$Y = (y_1, y_2, y_3, \ldots, y_n)$，其向量表示分别为：$Vec(X) = (v_1, v_2, v_3, \ldots, v_p)$，$Vec(Y) = (l_1, l_2, l_3, \ldots, l_p)$。

欧式距离也称欧几里得距离，是最常见的距离度量，衡量的是多维空间中两个点之间的绝对距离。

计算公式如下：

$$Dist(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{9}$$

2.2.2.5 文本聚类

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特定地点、特定人群，对相似留言进行分类。通过聚类方法，可以快速地将相似问题归并。

2.2.2.6K-means 算法[7]

K-Means 算法是一种无监督分类算法，假设有无标签数据集：

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix}$$

该算法的任务是将数据集聚类成k个簇$C = C_1, C_2, \dots, C_k$，最小化损失函数为：

$$E = \sum_{i=1}^{k}\sum_{x \in C_i} \left||x - \mu_i|\right|^2 \tag{6}$$

其中μ_i为簇C_i的中心点：

$$\mu_i = \frac{1}{|C_i|}\sum_{x \in C_i} x \tag{7}$$

要找到以上问题的最优解需要遍历所有可能的簇划分，K-Mmeans 算法使用贪心策略求得一个近似解，具体步骤如下：

1. 在样本中随机选取k个样本点充当各个簇的中心点$\{\mu_1, \mu_2, \dots, \mu_k\}$
2. 计算所有样本点与各个簇中心之间的距离$dist\left(x^{(i)}, \mu_j\right)$，然后把样本点划入最近的簇中$x^{(i)} \in \mu_{nearest}$，这样就行成了 k 个簇；
3. 根据簇中已有的样本点，重新计算簇中心

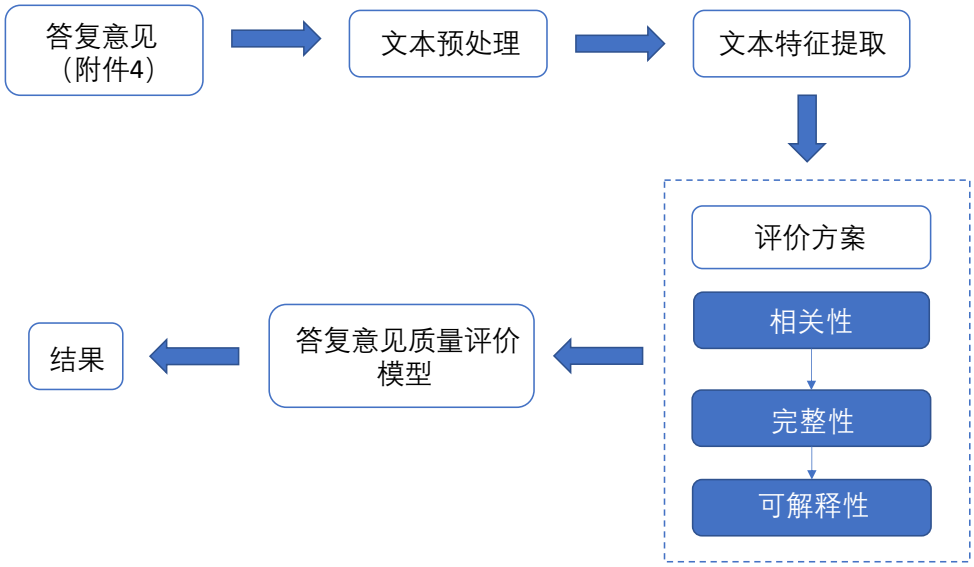
$$\mu_i := \frac{1}{|C_i|}\sum_{x \in C_i} x \tag{8}$$

4. 重复 2、3，直到质心的位置不再发生变化或者达到设定的迭代次数。

2.2.2.7 热度评价指标[8]

2.3 问题三分析方法与过程

2.3.2 流程图



2.3.2 数据分析

2.3.2.1 文本预处理
2.3.2.2 用 TF-IDF 提取关键词
2.3.2.3 量化答复的相关性、完整性、可解释性
相关性：答复意见的内容是否与问题相关
完整性：是否满足某种规范
可解释性：答复意见中内容的相关解释
2.3.2.2 利用余弦距离[9]进行文本相似度计算

假设两个文本$X=(x_1,x_2,x_3,\ldots,x_n)$和$Y=(y_1,y_2,y_3,\ldots,y_n)$，其向量表示分别为：$Vec(X)=(v_1,v_2,v_3,\ldots,v_p)$，$Vec(Y)=(l_1,l_2,l_3,\ldots,l_p)$。

余弦相似度用向量空间中两个向量的夹角的余弦值来衡量两个文本间的相似度，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，一般情况下，用 Embedding 得到两个文本的向量表示之后，可以使用余弦相似度计算两个文本之间的相似度。

计算公式如下：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (9)$$

3. 结果分析

3.1 问题一 结果分析

3.1.1 结果分析

3.1.1.1 F-Score 分类方法评价

F1-measure 的函数为

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

其中 Pi 为第 i 类的查准率， Ri 为第 i 类的查全率。

精确率(precision)的公式是$P = \frac{TP}{TP+FP}$,它计算的是所有被检索到的中,"应该被检索到的 item（TP）" 占的比例。

召回率(recall)的公式是$R = \frac{TP}{TP+FN}$,它计算的是所有检索到的占所有"应该被检索到的 item（TP+FN）"的比例。

将精准率P和召回率R代入 F1-measure 函数得到

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

F1 综合了 P 和 R 的结果，当 F1 较高时则能说明试验方法比较有效。

以下是 CNN 模型训练时 F1-measure 的评价结果：

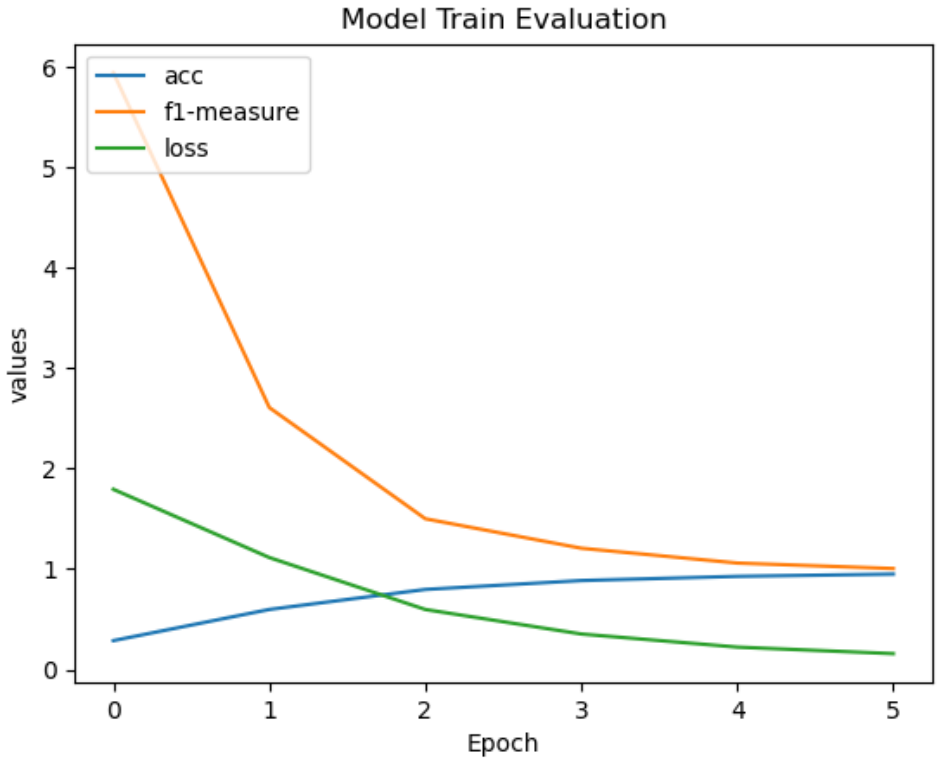


图 6 F1-measure 评价结果

准确率在一次迭代后就已经超过 90%，损失函数的值快速减小，f1 的值快速下降。由此分析，训练次数不宜太高。

3.1.2 数据验证

对于 CNN，我们选取了训练集中的 10%用于测试。对于测试样本的验证结果如下图。可以看到验证的准确率达到接近 80%，损失函数值迅速减小，并在两次迭代后保持平缓。

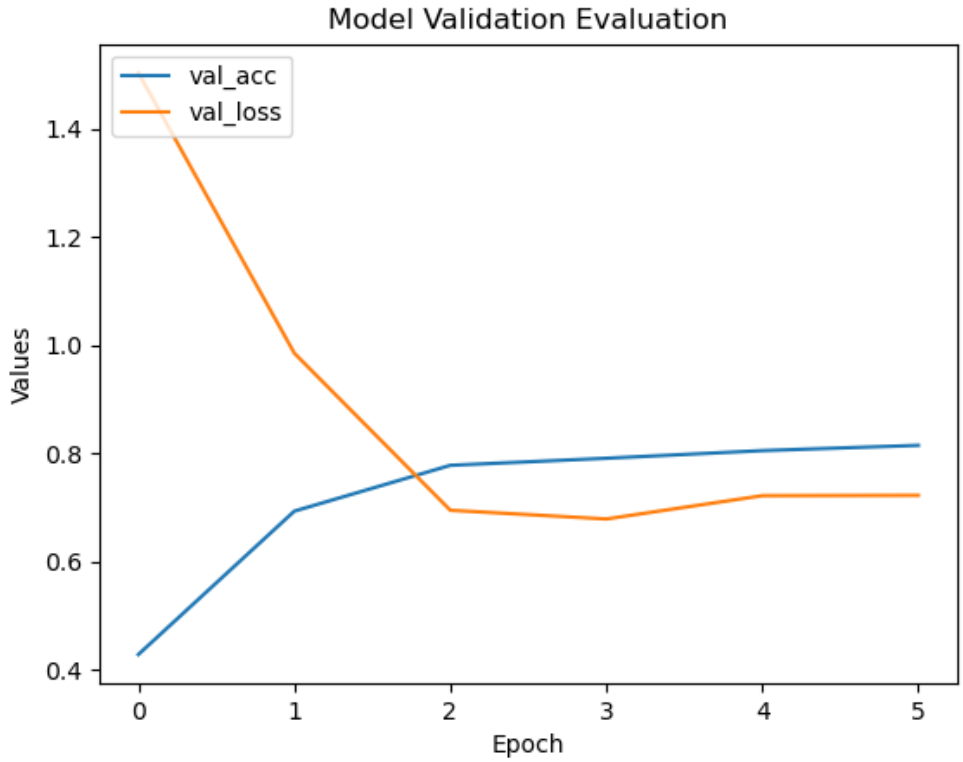


图 7 测试样本的验证结果

3.2 问题二 结果分析

3.2.1 结果分析

	Unnamed: 0	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	主题+详情	jieba	关注数	label
106	106	190812	A00095451	A市江山帝景新房有严重安全隐患	2019/5/30 17:34:02	\n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景a14栋的住户。看到新房，却怎么...	0	0	A市江山帝景新房有严重安全隐患 \n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景a...	市 江山 帝景 新房	0	1
1652	1652	227671	A0008477	A市江山帝景新房脏乱差，有安全隐患	2019/3/20 9:37:45	\n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景哈弗四期A10至A14栋业主，于...	0	0	A市江山帝景新房脏乱差，有安全隐患 \n\t\t\t\t\t\n\t\t\t\t\t我是江山帝...	市 江山 帝景 新房	0	1
2166	2166	239937	A0008477	A市江山帝景新房脏乱差，有安全隐患	2019/3/19 19:37:45	\n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景哈弗四期A10至A14栋业主，于...	0	10	A市江山帝景新房脏乱差，有安全隐患 \n\t\t\t\t\t\n\t\t\t\t\t我是江山帝...	市 江山 帝景 新房	10	1
4301	4301	319659	A023956	A市江山帝景新房有严重安全隐患	2019-05-30 17:34:02	\n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景a14栋的住户。看到新房，却怎么...	0	0	A市江山帝景新房有严重安全隐患 \n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景a...	市 江山 帝景 新房	0	1
4294	4294	289893	A00095451	A市江山帝景新房有严重安全隐患	2019/5/30 17:20:53	\n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景a14栋的住户。看到新房，却怎么...	0	0	A市江山帝景新房有严重安全隐患 \n\t\t\t\t\t\n\t\t\t\t\t我是江山帝景a...	市 江山 帝景 新房	0	1

图 8 留言聚类结果表

表格的左后一列为对应的标签，将同簇留言及其关注度相累加后，可以求出各簇留言对应的热度。

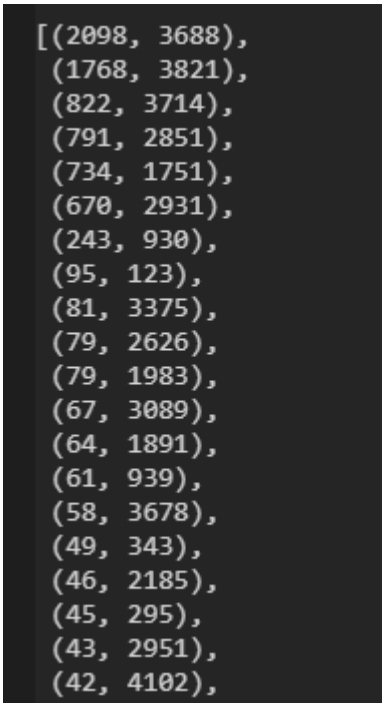


图 9 热点排序图

热度排行列表如上图，元组第一个数字表示热度大小，元组第二个数字表示簇类名。经过整理，可得热点问题明细表格。

表 1 热点问题明细表格

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
1	214238	A00061787	对58车货一	2019/1/20 22:28	标题：恳	1	2
1	220711	A00031682	注A市A4区	2019/2/21 18:45	尊敬的胡	0	821
1	217032	A00056543	贷特大集资	2019/2/25 9:58	胡市长：	0	790
1	194343	A00010616	贷案警官应	2019/3/1 22:12	胡书记：	0	733
2	208636	A00077171	矿万境K9县	2019/8/19 11:34	我是A市	0	2097
3	223297	A00087522	毛湾配套人	2019/4/11 21:02	书记先	5	1762
4	263672	A00041448	赣高铁最近	2019/9/5 13:06	您好，近	0	669
5	193091	A00097965	丽发新城强	2019/6/19 23:28	位于A市	0	242

3.3 问题三 结果分析

3.3.1 结果分析

整体的情况如图所示：

	Unnamed: 0	留言编号	相关性	完整性
count	2816.000000	2816.000000	2816.000000	2816.000000
mean	1407.500000	59776.240767	0.295281	0.045381
std	813.053504	49347.909958	0.218589	0.054562
min	0.000000	2549.000000	0.000000	0.000000
25%	703.750000	17556.000000	0.108640	0.018020
50%	1407.500000	38635.500000	0.268293	0.034645
75%	2111.250000	103807.250000	0.450796	0.055964
max	2815.000000	185986.000000	0.940483	1.000000

图 10 留言回复整体评价

4. 结论

建立基于自然语言处理技术的智慧政务系统，对于处理日益增长的社情民意文本数据量有重大意义，已经成为文本分析的一个课题、一个难题。单纯依靠人工来进行留言划分和热点整理存在工作量大、效率低，且差错率高等问题。本文采用流行的 word2vec、TFIDF 权重法文本预处理方法，采用 CNN（卷积神经网络）建立了分类模型，采用 Kmeans 无监督聚类方法获得留言热点问题。利用此模型能够快速对大量文本进行分类、寻找政府需要集中处理的重点热点问题，从而提供更优质的服务。

由分析结果可以看出，人们更加关注公共设施的完善问题与教育问题，可以看出人们更加追求生活品质。58 车贷一案最受人们关心，人们急切希望得到有关案情的消息，可以看出，人们比较关心对社会影响力大的事件。人们的留言体现了人们日益上升的责任感，这点更督促政府人员更好的为人们服务，通过热点分析，解决人们的燃眉之急，提高城市幸福感。同时对于政府应答的情况来看，政府基本能根据留言的热点进行有针对性、完整的答复

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] 刘志. 基于用户兴趣的协同过滤算法的广告推荐研究[D]. 昆明理工大学, 2014.

[3] 毛郁欣,邱智学.基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J].中国信息技术教育,2020(08):99-101.

[4] 张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[5] 万磊,张立霞,时宏伟.基于 CNN 的多标签文本分类与研究[J].现代计算机,2020(08):56-59+95.

[6] 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.

[7]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.

[8] 梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.

[9] 王志刚,谢恺,朱慧.降成本政策的文本分析——基于文本相似度计算原理[J].地方财政研究,2020(03):90-97.