对“智慧政务”中的文本挖掘的分析和深度学习

摘要

随着微信、微博、市长信箱等网络问政平台的出现，各类社情民意相关的文本数据量不断攀升。同时，随着大数据、云计算、人工智能等及技术的发展，建立在自然语言处理技术的智慧政务系统将是社会发展的新趋势。

对于问题 1，在数据预处理方面，我们首先通过 python 的 pandas 模块对留言数据去重；然后使用 jieba 分词包对文本进行中文分词；最后采用通用的 stopwords 词典进行过滤。完成数据处理后，我们使用深度学习中的 LSTM 长短期记忆网络来进行文本多分类。其主要过程为：先将处理后的数据向量化，设置停用词；接着进行拆分训练集和测试集，定义 LSTM 模型；定义完成之后，进行训练数据，周期为 100；最后不但计算出 F1-Score 还画出了混淆矩阵对模型进行了评估。

对于问题 2，数据处理完成之后，本文使用 k-means 均值聚类对附件 3 留言详情进行聚类，并采用 LDA 主题模型提取留言主题对其进行改进，接着使用正则表达式提取留言主题中的地点信息，在此基础上使用 Excel 人工筛选，最终完成问题归类。我们给出了适合该问题的留言热度模型，该模型主要考虑了三点因素分别为：初始热度值（分类后的留言条数）、用户交互得分（点赞数和反对数）、随时间衰减的热度分（留言时间）。在综合考量之后，计算出了热度值并最终得到了热度排名前五的留言问题。

对于问题 3，评价模型的建立，为了建立出合适的评价模型并尝试实现，本文使用了层次分析（AHP）模型。选取了四点因素作为方案层分别为：完整性、情感、相关性、可解释性，并对这四点因素进行量化。完整性量化：根据答复意见的语句结构等给出相应的量化指标。情感量化：我们使用 NLP 情感分析模型，在构建情感词典的基础上抽取情感词，得出情感评分。相关性量化：本文使用文本相似度中的余弦相似度进行处理。可解释性量化：我们根据答复意见的合理性等给出量化指标。在建立判断矩阵之后，利用 matlab 建立模型用特征值法求解，得到各影响因素对总目标的影响权重。将量化后的影响因素指标与以上的权重相乘，从而就到达了给答复意见评分的要求，由于部分因素主观评价在大数据量时实现困难，我们用 python 随机抽样了 300 数据进行评价。

关键词：数据预处;LSTM 分类模型;k_means 聚类;LDA 模型;层次分析模型（AHP）

Analysis and deep learning of text mining in "smart government"

Abstract

With the emergence of online questioning platforms such as WeChat, Weibo, and the mayor's mailbox, the amount of text data related to various social conditions and public opinion has been increasing. At the same time, with the development of big data, cloud computing, artificial intelligence, and other technologies, smart government systems based on natural language processing technology will be the new trend of social development.

For question 1, in terms of data preprocessing, we first deduplicate the message data through the python pandas module; then use the jieba word segmentation package to perform Chinese word segmentation; and finally use a general stopwords dictionary to filter. After completing the data processing, we use the LSTM long-short-term memory network in deep learning to perform text multi-classification. The main process is to vectorize the processed data and set stop words. Then split the training set and test set to define the LSTM model. After the definition is completed, the training data is performed with a period of 100. Finally, not only the F1-Score is calculated, but also the confusion matrix is drawn to evaluate the model.

For question 2, after the data processing is completed, we use k-means mean clustering to cluster the attachment 3 message details, and use the LDA theme model for extracting the message subject to improve it. After that we use regular expressions to extract the location information in the message subject, then use Excel to manually filter on this basis, and finally complete the problem classification. We have given a message popularity model suitable for the problem. The model mainly considers three factors: the initial popularity value (the number of messages after classification), the user interaction score (the number of likes and objections), and the decay with time The popularity score (message time). After comprehensive consideration, we calculated the heat value and finally got the top five message questions.

For question 3, the establishment of the evaluation model. In order to establish a suitable evaluation model and try to achieve it, this paper uses an analytic hierarchy process (AHP) model. Four factors were selected as the project layer: integrity, emotion, relevance, and interpretability, and these four factors were quantified. Quantification of completeness: According to the sentence structure of the reply opinion, the corresponding quantification index is given. Sentiment quantification: We use the NLP sentiment analysis model to extract sentiment words on the basis of constructing sentiment dictionaries and obtain sentiment scores. Quantification of relevance: This article uses cosine similarity in text similarity for processing. Interpretability quantification: We give quantitative indicators based on the rationality of the answers. After the judgment matrix is established, the model is established using matlab to solve with the eigenvalue method, and the influence weight of each influencing factor on the overall goal is obtained. Multiply the quantified influencing factor index with the above weights, so as to meet the requirements for scoring the response opinions. Due to the difficulty in achieving a large amount of data for some factors, we randomly sampled 300 data with python for evaluation.

Keywords: data preprocessing LSTM classification model k_means clustering LDA model analytic hierarchy process model (AHP)

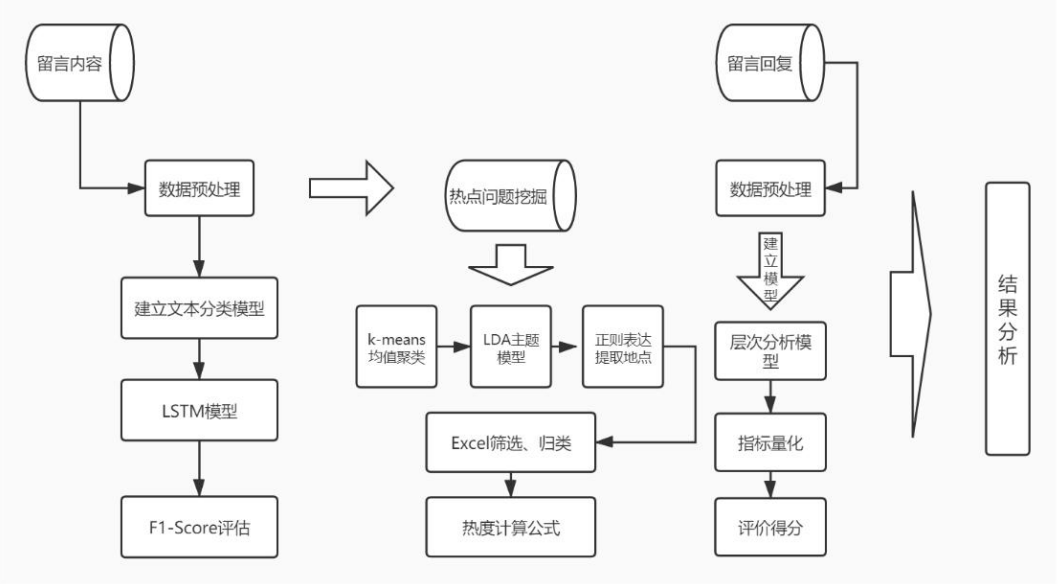
目录

1. 挖掘目标

本次建模目标主要利用互联网公开来源的群众问政留言记录以及相关部门对群众的留言答复，分别对“智慧政务”中文本信息和热点信息进行分类和挖掘，并且对政府给出的回复进行了评价。在问题一中本文首先对所给文本信息进行处理，方便运用数据，然后构建留言内容的一级标签分类模型，最后使用 F1-Socre 对分类方法进行评价。在问题二中先对留言进行归类，随后构建相应的热度模型对热点问题进行挖掘，选取排名前五的热点问题。第三问基于相关部门对留言的答复意见，建立层析分析模型，从答复的语句和与问题的联系两个方面入手，利用相关的分析方法以及自然语言处理给出一个综合评价方案。

2. 分析方法与过程

2.1 总体流程



2.2 问题一的分析方法和过程

2.2.1 对群众留言信息的预处理

在对数据进行提取后,我们发现留言数据中存在一些无关数据和噪声数据，这会影响数据的真实性和可操作性。因此，我们进行了数据清洗，发现并纠正数据文件中可识别的错误，包括检查数据一致性，处理无效值和缺失值等。通过数据清洗，我们会得到更为“干净”的数据，方便进一步处理数据。

① “去重”、“去空”

在导入留言文本之后，本文利用了 python 中的 pandas 模块中的 isnull 函数对空值进行了统计，并舍弃掉为空值的数据；然后对数据中给出的各个一级标签数目进行了统计生成了柱状图；最后进行去重处理，过滤掉重复的留言文本。这里本文仍然采用 pandas 模块，利用 drop-duplicates（）完成剔除操作。

② 删除文本中的无用值

在每一段文本中存在许多无用值，例如：空格、符号、标点等。在这样的处理中我们用 python 中的 re 模块，以及 stripe、sub，complie 等函数对那些无用值进行处理。

③ 中文分词并过滤停用词

在这个问题中本文使用的分词词典是一套通用的中文分词词典，并利用 jieba 分词进行具体操作。

2.2.2 文本分类（分类模型）

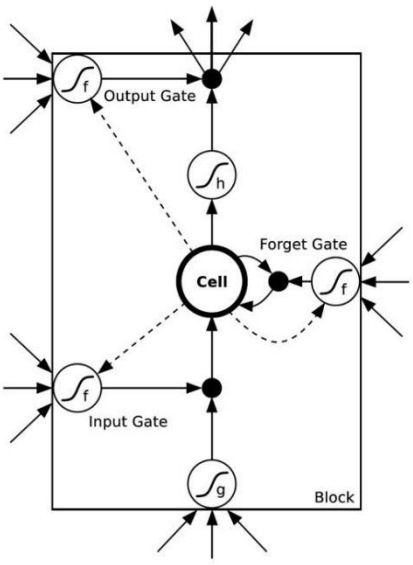
将数据处理完成之后，我们使用 LSTM（Long Short-Term Memory）长短期记忆网络对留言文本进行分类。LSTM 是一种时间循环神经网络，适合用于处理和预测时间序列中间隔和延迟相对较长的重要事件。LSTM 内部主要分为三个阶段：

② 忘记阶段：“忘记不重要的，记住重要的”。

② 选择记忆阶段：“重要的着重记录，不重要的粗略记录”。

③ 输出阶段:“决定哪些将会被当成当前状态输出”。

它的工作原理图如下：



在进行 LSTM 的建模工作中，我们对数据清洗得到的 cut_comment 数据设置保留的最大次数为 60000；然后设置每条 cut_comment 最大的词数为 100 个；在统计总共的不同词语数后，使用 texts_to_sequence 函数对数据进行了向量化处理；最后按 test_size=0.2 的比率拆分训练集和测试集。

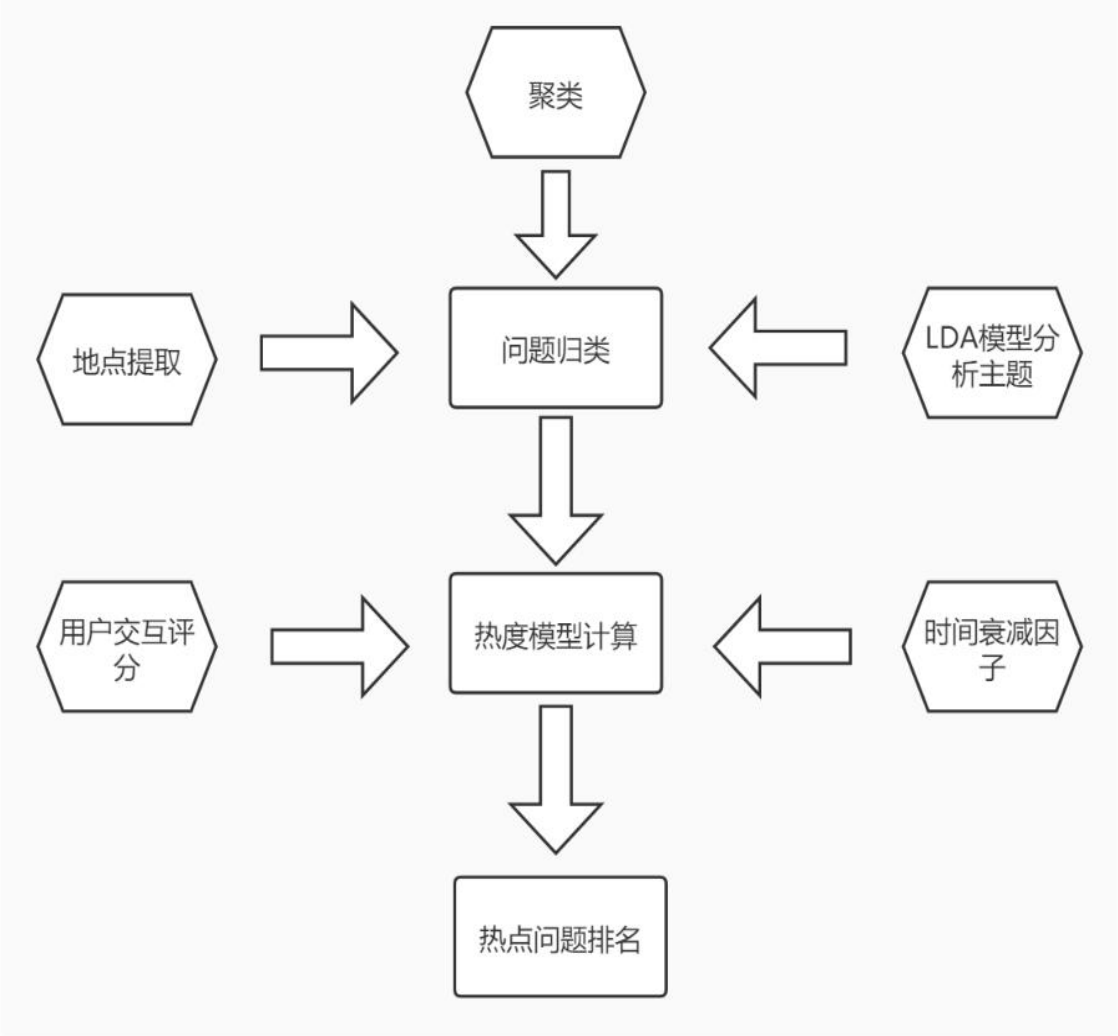
在 LSTM 的序列模型定义中，嵌入层(Embedding)用长度为 100 的向量来表示每一个词。在 SpatialDropout1D 层在训练中每次更新时，随机断开一定比例的输入神经元连接，因此我们设置 dropout 参数为 0.2。为了防止过拟合我们还采

用了早停法（Early stopping)。由于我们最终要将每条留言分类分类到对应的那一个一级标签中去，我们将激活函数设置为“softmax”，与之对应的损失函数为分类交叉熵。

设置完成之后开始进行训练，设置训练周期为 100。在参数选择的过程中，运用了梯度下降法，确定 batch_size=256。最后，通过求出 F-score 分数来评估我们模型的表现。为了更直观地观察到每个类别的准确性，我们还画出了混淆矩阵进行对比。（详细代码见附件 1）

2.3 问题二的分析方法和过程

2.3.1 热点问题总体分析框架



2.3.2 数据预处理

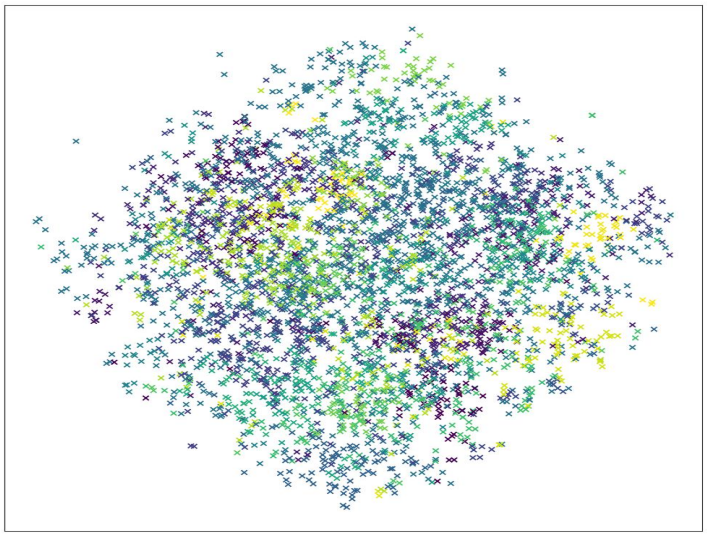
对热点问题挖掘的前提是将文本信息进行相应的预处理。我们需要用到问题一中对文本数据的处理办法对附件 3 中的文本数据进行预处理，处理方法和流程

大致为：

经过处理后得到我们想要的词集。

2.3.3 聚类、主题提取、地点提取

为了能将表示各个问题的所有留言信息归类，我们尝试了 k-means 均值聚类。先把预处理之后的文本数据转化成词向量；再对向量进行聚类；接着使用 k-means 的超参数 n_clusters 对聚类结果进行评估；再使用 T-SNE 算法对向量进行降维处理；对最终运行结果比较之后，将聚类主题数定为 48，生成聚类主题和聚类效果图。其中聚类效果图展示如下（详细代码见附件 2）：



从效果图的呈现和超参数的数值情况来看，聚类算法在该总体中效果一般。我们采用 LDA 主题提取模型进行参考（详细代码见附件 3），并使用正则表达式对留言主题进行遍历提取地点信息（详细代码见附件 4），最后人工利用 excel 进行数据最终筛选与归类总结出了附件 3 中留言表示的几个主要问题。

2.3.4 热度计算

在对附件 3 中的文本数据预处理、聚类、地点提取以及 Excel 筛选之后，需要进一步对热点问题进行量化。我们给出相应的热度模型算法，算出每个留言问题的热度值，利用热度值大小对热点问题进行排名。

① 热度算法基本原理

在观察所给数据后，我们参照新闻的热度分析原理进一步对热点问题进行分析，并给出了相应的算法公式，它的基本原理如下：

留言热度分=初始热度分+用户交互产生的热度分-随时间衰减的热度分

即：

$$Score = S0 + S(Users) - S(Time)$$

1. 初始热度分（S0）：以表述同一留言问题的留言数量(排除同一用户的重复留言数) 作为初始热度分。

2.用户交互产生的热度分(S（Users）)：为了避免用户规模对用户交互产生影响，我们选择固定用户规模。考虑到用户的行为也会对热点问题产生影响，我们将 b 表示的点赞数和反对数总和作为 S（Users）中的重要依据。具体的用户交互产生的热度分公式如下所示：

$$S(Users) = \frac{b}{DAU \times N(\text{固定数})}$$

（DAU 为日活跃用户数量，N 则是为了使三个指标相互独立）

3.随时间衰减的热度分：留言热度是在随时间的衰减而不断下降。根据牛顿冷却定律，时间衰减因子应该是一个类似于指数函数：

$$T(Time) = e^{k(T1-T0)}$$

（T0 是同类留言最后一条留言的时间，T1 是当前时间）

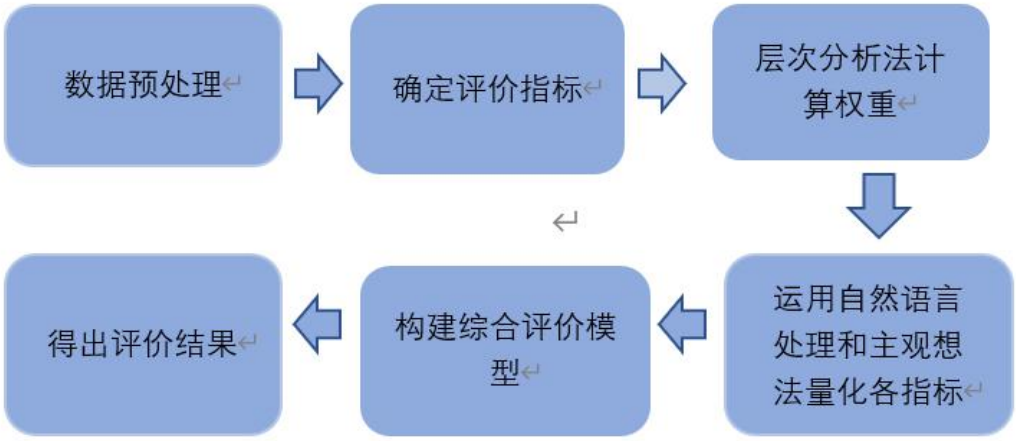
基于该公式，我们运用 MATLAB 软件对留言问题中热度随时间的变化进行绘制（图像如下）。通过图像发现，热度随时间的衰减不是线性的，热度随时间的流逝而衰减，并且衰减趋势越来越快，直至趋近于零热度。

所以本文最终的热度计算公式调整为：

$$Score = \frac{(S0 + S(Users))}{S(Time)}$$

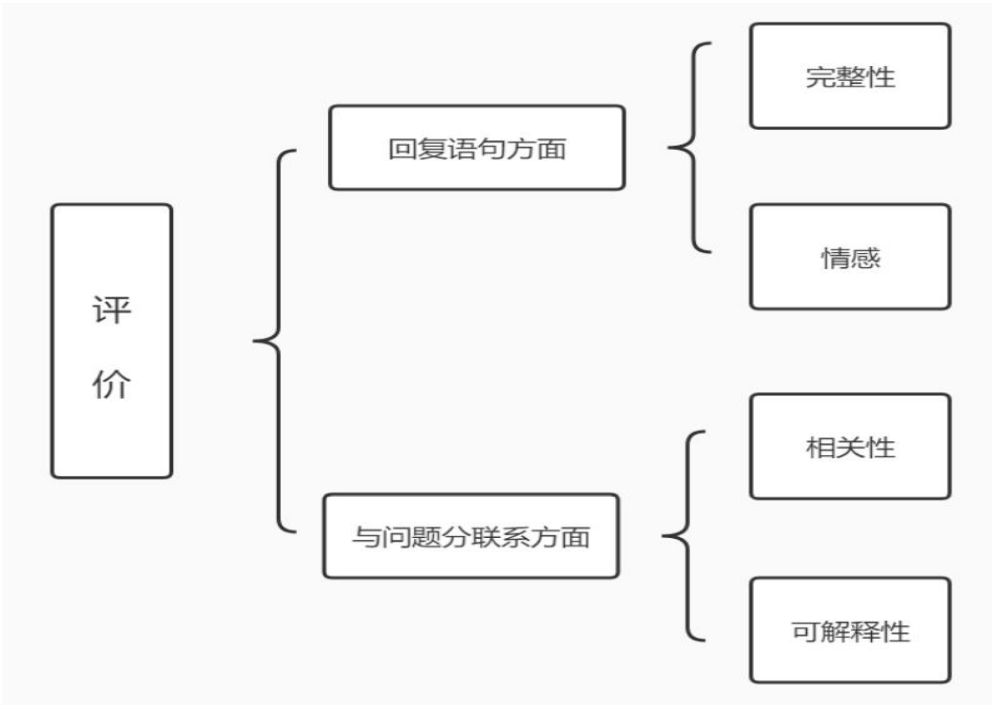
2.4 问题三的分析方法与过程

2.4.1 总流程图



2.4.2 评价指标的确定

评价指标的选取对系统的综合评价起着至关重要的作用。可以说，根据不同的评价指标评价出来的结论之间可能大相径庭。根据对留言答复意见的分析和评价指标选取原则，本文选择的评价指标如下：



2.4.3 层次分析（AHP）模型

由于意见的质量由多个指标确定，则解决此问题的关键就是要确定各个指标在评价体系中所占比重。这里我们使用层次分析法，将与决策有关的因素分解成目标层、准则层、方案层等层次，通过对各因素的计算和比较，得出不同因素的权重，为决策者选择最优方案提供参考依据。鉴于部分因素为主观定义，在大数据量的情况下实现较为困难，在此使用 python 随机抽样三百条数据进行评价(详细代码见附件 5），其余步骤如下：

①建立递阶层次结构模型：

根据对问题的分析，捋清问题所包含的因素，确定出各个因素之间的关联和隶属关系，按这些因素的共同特性，将它们分为目标层、准则层、方案层等层次。本文建立的递阶层次结构模型如下：

意见质量评价体系结构表

意见的质量 A	语句方面 B1	完整性 C1
		情感 C2
	与问题的联系 B2	相关性 C3
		可解释性 C4

②构造出各层次中的所有判断矩阵：

判断矩阵中的 b_{ij} 一般采用九分制标度法（定义详见下表），根据资料数据、专家意见或者系统分析人员的经验，经过反复研究后确定如下：

表 1　标度的含义

标度	含　　义
1	表示两个因素相比，具有相同重要性
3	表示两个因素相比，前者比后者稍重要
5	表示两个因素相比，前者比后者明显重要
7	表示两个因素相比，前者比后者强烈重要
9	表示两个因素相比，前者比后者极端重要
2，4，6，8	表示上述相邻判断的中间值
倒数	若因素 i 与因素 j 的重要性之比为 a_{ij}，那么因素 j 与因素 i 重要性之比为 $a_{ji}=1/a_{ij}$。

本文经过查找资料和分析得出的判断矩阵如下：

判断矩阵 A-B			判断矩阵 B1-C			判断矩阵 B2-C		
A	B1	B2	B1	C1	C2	B2	C3	C4
B1	1	1/4	C1	1	3	C3	1	6
B2	4	1	C2	1/3	1	C4	1/6	1

③层次单排序及一致性检验：

1、计算一致性指标CI：

$$CI=\frac{\lambda_{\max}-n}{n-1}$$

2、查询平均随机一致性指标RI，对应 n=1 到 9，RI 值分别为:

RI 的值									
n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

3、计算一致性比例CR：

$$CR=\frac{CI}{RI}$$

当$CR<0.1$，认为矩阵的一致性是可以接受的。当$n>2$时才应该做一致性检验，当$n\le 2$时，判断矩阵为一致矩阵，因此不需要一致性检验。

④层次总排序及一致性检验：

若需要用到最底层中各因素关于总目标的权重，即和其上一层的对总目标的权重相乘即可，一致性检验和层次单排序的一致性检验相同，这里不再赘述。

⑤用 matlab 建立模型利用特征值法求出对应权重、一致性。（详细代码见附件 6）

2.4.4 评价指标的量化

①完整性的量化(S1):

完整性检测是指对留言回复的语法、语句结构，以及是否对留言信息进行正确解决等方面的量化。我们对答复意见的完整性给出了以下量化指标：

答复意见的完整性等级	A 级	B 级	C 级	D 级	E 级
分值（100 分）	100 分	80 分	60 分	40 分	0 分

为了尽可能出去个人主观因素，采用多人多次对回复意见的完整性进行评分的方式，取平均值作为回复意见完整性的最终得分。

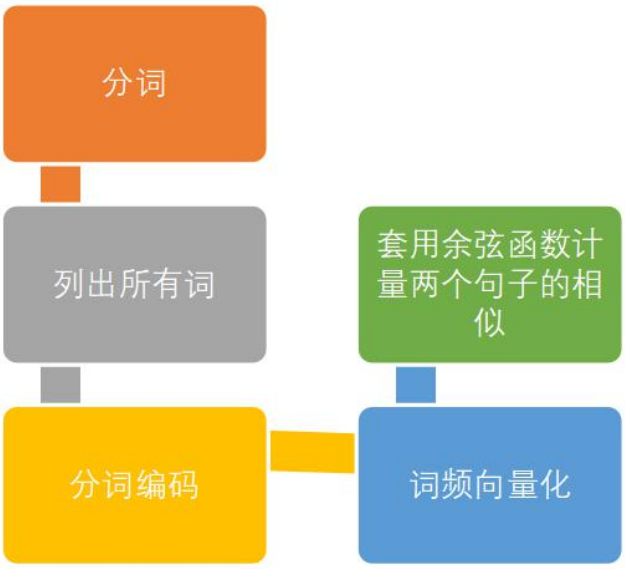
② 情感的量化(S2)：

本文运用 NLP 情感分析模型将回复意见中的中文文本进行情感分析，进而将情感量化。NLP 文本情感分析利用了自然语言处理和文本挖掘技术，对带有情感色彩的主观性文本进行分析、处理和抽取的过程。（详细代码见附件 7）

我们基于构建的情感词典（消极词、积极词、否定词、程度级别词），对回复意见进行文本处理抽取情感词，计算文本的情感倾向，最终将情感分析以情感评分的形式展现出来。

③ 相关性的量化(S3)：

文本相关性一定程度上可用两个文本的相似度替代，本文用留言内容和答复意见的相似度代表两者的相关性。相似度度量指的是计算个体间相似程度，衡量文本相似度最常用的方法是使用余弦相似度（详细代码见附件 8）。流程图如下：



本文通过遍历分别提取两列数据进行分词向量化处理，就变成了计算两个向量之间夹角的余弦值，值越大相似度越高。余弦值计算公式如下：

$$\cos\theta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

④ 可解释性的量化(S4)：

对于可解释性的量化，本文根据答复意见的内容制定了如下标准：

1、 所有答复意见可解释性初始分都为 0 分。
2、 语句进行了具体回复则可得基础分 5 分。
3、 一条答复意见中每出现一次政策依据或具体条例措施即加加 1 分。
4、 通读语句，若答复意见语句合理，无夸张或者假信息可加 2 分。
5、 若回复意见语句中有明显的假信息、夸张不切实际以上情况每出现一次扣 1 分。
6、 若回复意见中出现不合理要求、不符合常人的做法等情况，每出现一次扣 1 分。

7、 根据以上评分细则，多人多次对回复意见的可解释性打分，最终以平均分为最终得分。

通过以上制定的标准，可实现对答复意见的可解释性量化。

3. 结果分析

3.1 问题一的结果分析

3.1.1 群众留言信息的预处理结果

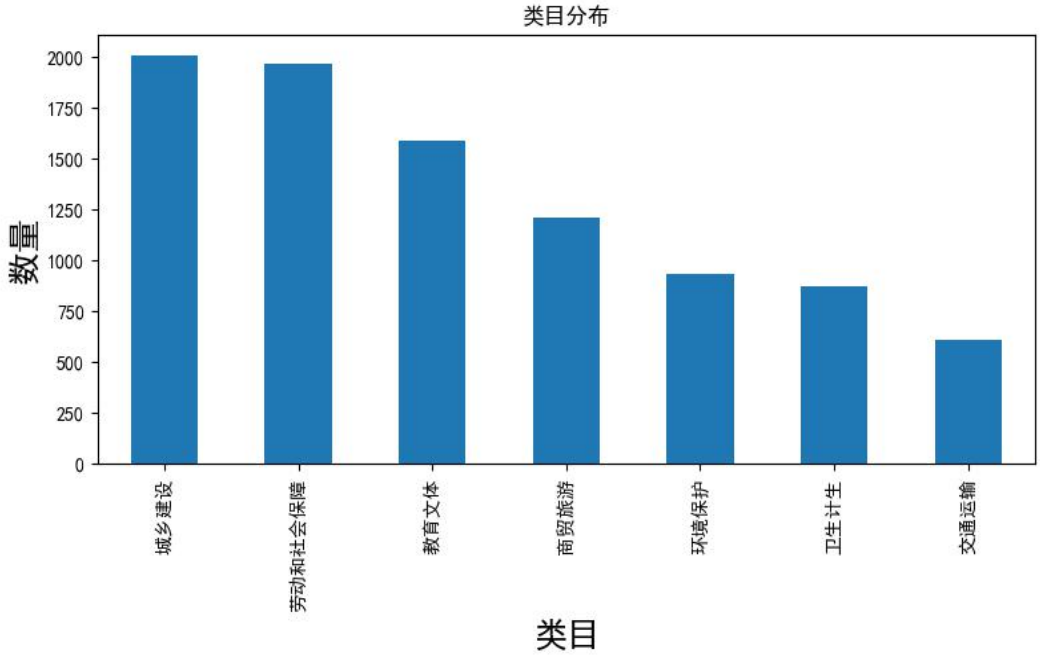
首先我们在去重去空之后，将一级标签中的七个大类进行标记。

	Cat	Cat_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动与社会保障	4
5	商贸旅游	5
6	卫生计生	6

由于留言文本中存在一些无关数据和噪声数据等，会影响数据的可操作性和真实性。我们对数据进行“去重”、“去空”，中文分类、停用词过滤等处理，得到“干净”的数据。结果如下图（部分）：

	comment	...	clean_comment
0	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	...	A3 区 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑 集团...
1	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 烂尾 护栏 围着 占用...
2	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	...	尊敬 A1 区苑 位于 A1 区 火炬 路 物业 A 市程明 物业管理 有限公司 未经 小区...
3	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	...	A1 区 A2 区华庭 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 水...
4	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	...	A1 区 A2 区华庭 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 水...
...	...	...	...
9205	\n \n 我们夫妻都是农村户口，大的是女9岁，小的是儿2岁半，才15斤...	...	夫妻 农村户口 女 9 岁 2 岁 半才 15 斤 治疗 两年 一级 脑瘫 女户 招郎 男方...
9206	\n \n 本人2015年2月16号在B市中心医院做无痛人流手术，手术后...	...	2015 年 2 16 号 B 市中心 医院 做 无痛 人流 手术 手术 怀孕 症状 201...
9207	\n \n 我们是再婚，很想再要一个小孩，不知我省二胎新政策何时出，如果...	...	再婚 想 小孩 不知 我省 二胎 新 政策 先 怀孕 做
9208	\n \n K8县惊现奇葩证明！ 我是西地省K8县人，想生二孩，被告知要...	...	K8 县惊现 奇葩 证明 K8 县人想 生二孩 告知 要开 证明 生二孩 证明 先写 证明 ...
9209	\n \n 领导你好，我们属于未婚生子，但是在2013年已经接受处罚，小...	...	你好 未 婚生子 2013 年 接受 处罚 小孩 上户 小孩 外地 上学 需 办理 积分 入...

在统计各个类别的数据量之后，我们使用图形化的方式对各个类别的分布进行查看，结果如下图：



3.1.2 LSTM 模型流程分析

数据预处理完成之后，进行 LSTM 的建模工作。通过对 cut_review 数据的向量化处理并设置使用词，得到有 82828 个不相同的词语。我们按照 8:2 的比例对训练集和测试集的拆分之后，定义 LSTM 序列模型。定义模型完成之后，设置 100 个训练周期对数据进行训练。接下来通过画混淆阵和求 F1 分数来对模型表现进行评估。其中，混淆矩阵的主对角线表示预测正确的数量。为了更好地反映模型的实际预测精度，我们计算 F1 分数来对模型评估。经过多次运行之后，我们得到该模型的 F-score 大约为 0.83,。

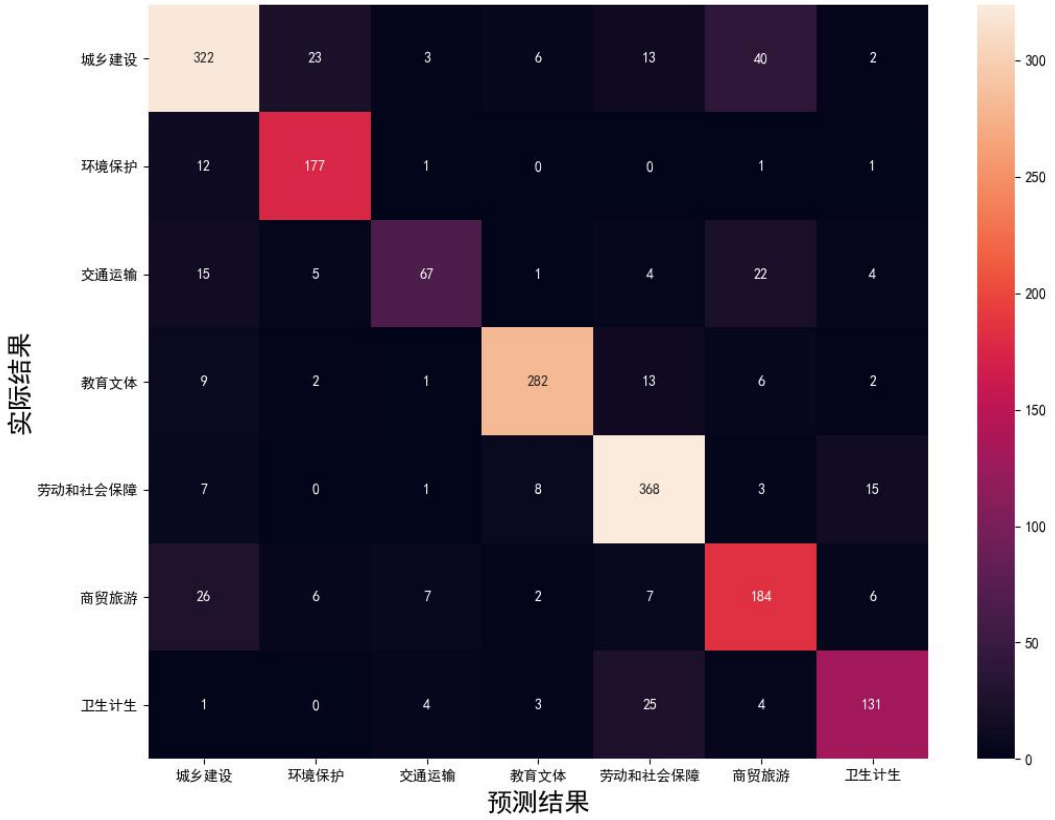
```
accuracy 0.8311617806731814
              precision    recall  f1-score   support

        城乡建设       0.82      0.79      0.80       409
        环境保护       0.83      0.92      0.87       192
        交通运输       0.80      0.57      0.66       118
        教育文体       0.93      0.90      0.91       315
      劳动和社会保障       0.86      0.92      0.88       402
        商贸旅游       0.71      0.77      0.74       238
        卫生计生       0.81      0.78      0.80       168

    accuracy                           0.83      1842
   macro avg       0.82      0.81      0.81      1842
weighted avg       0.83      0.83      0.83      1842
```

通过 LSTM 模型的精准率分析我们得知，样本各分类数量之间的差异会导致

各个分类中准确性有较大的的差异。



3.2 问题二的结果分析

3.2.1 LDA 模型展现主体结果

为了更好的研究热点问题，我们使用 LDA 对数据预处理过后的留言内容和留言主题数据分别进行了运行，均将主题数设为 8。但留言主题的主题抽取结果并不明显，在这里不做展示，留言内容的主题抽取结果如下图：

```
Topic #0:
驾校 成员 医保 集体经济 征地 历史 分配 父亲 山水 尖山 发票 申请人 建房 许可 居委会 蒋户 使用权 摩托车 居住证 纪委
Topic #1:
校区 砂子 精装 月亮 造价 围墙 雅苑 开元 左转 四路 直行 旅游 右转 划分 地铁站 松雅湖 六路 高压线 金星 核算
Topic #2:
公积金 建局 供水 考生 违章建筑 中建 封闭 人防 app 空调 金科 保安 燃气 水费 水表 水管 客服 缴费 菜市场 入户门
Topic #3:
保利 公交线路 窗口 公交站 商户 香樟 评估 木莲 区路 问政 商场 洞井 交汇处 机场 用房 分拣 红星 c5 一段 青竹
Topic #4:
烟道 专用 饭店 烧烤 违停 开设 万科 裂缝 罚款 废气 违章 商住 住宅楼 搭建 分贝 改建 楼板 休闲 空气 达标
Topic #5:
垃圾站 退休 医生 公办 招标 办学 补课 监控 初中 上课 就读 退费 学位 睡眠 数字 离职 中标 患者 岗位 校方
Topic #6:
出借 受害人 投票 办案 股东 兑付 58 业主大会 农民工 信访 鉴定 车贷 公安局 报案 经侦 审计 金融 警官 招聘 签名
Topic #7:
麻将馆 买受人 铁路 大队 绿地 出卖 不动产 市住 起诉 城际 建筑物 判决 买卖合同 广铁集团 搅拌站 民警 房价 认购 认购书 城铁
```

我们利用正则表达式对文本中的地点信息进行提取，在对文本中的地点信息进行观察后，不断完善正则表达式中的位置信息来实现对大部分留言主题的地点提取。

```
<re.Match object; span=(0, 13), match='A3区一米阳光婚纱艺术摄影'>
None
None
<re.Match object; span=(0, 6), match='A2区黄兴路'>
<re.Match object; span=(0, 11), match='A市A3区中海国际社区'>
<re.Match object; span=(0, 18), match='A3区麓泉社区单方面改变麓谷明珠小区'>
<re.Match object; span=(0, 7), match='A2区富绿新村'>
None
<re.Match object; span=(0, 4), match='A市6路'>
<re.Match object; span=(0, 17), match='A3区保利麓谷林语桐梓坡路与麓松路'>
<re.Match object; span=(0, 10), match='A7县特立路与东四路'>
<re.Match object; span=(0, 9), match='A3区青青家园小区'>
```

随着互联网的发展，人们的互动已经不仅仅局限于现实生活中，在网络世界中人们的点赞、转发等行为也是互动的一种方式。人们对于留言回复的互动也会影响着问题的热度，所以根据上面的留言热度分算法和对数据的分析，我们将用户的点赞数和反对数作为 S（Users）中的重要依据。

由于热度随时间的衰减在不断下降，根据热度算法原理中 $T(Time) = e^{k(T1-T0)}$，所以本文将留言时间到现在的时间跨度作为 T，将“月”设为时间跨度的单位，并用 Excel 简单处理。

3.2.2 按照热度模型计算结果

热度模型 $Score = \frac{(S0 + S(Users))}{S(Time)}$，其中根据对数据的分析将初始热度 S0 为某一留言问题的留言数量，将 $T(Time) = e^{k(T1-T0)}$ 的 k 赋值为 1/15、将 $S(Users) = \frac{b}{DAU * N(\text{固定数})}$ 中的 DAU（日活跃用户数）假设为 3000、N（固定数）统一赋值为 1/200。根据算法公式将以上已经计算和赋值出来的各种指标值，利用 Excel 计算并排名，得出热度前五的问题，结果如下表：

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	88.8553894	2019/1/8 至 2019/5/28	A 市 58 车贷	A 市 58 车贷诈骗问题
2	2	85.8458967 9	2019/5/5 到 2019/9/19	A 市 A5 区 K9 县五矿万境	A 市 A5 区五矿万境 K9 县房屋出现质量问题
3	3	52.3987682	2019/4/10 到 2020/1/26	A 市 A2 区丽发新城小区	小区附近搅拌站噪音扰民

4	4	51.77675567	2019/1/16 到 2020/1/6	A 市伊景园·滨河苑小区	房子和车位捆绑违规销售
5	5	35.539953	2019/1/16 到 2019/12/14	A 市 A3 区多处工地	工地施工扰民

3.3 问题三的结果分析

3.3.1 指标所占比重

用层次分析法特征值发求解算得各指标对目标问题的影响权重如下:

判断矩阵 A-B				判断矩阵 B1-C				判断矩阵 B2-C			
A	B1	B2	W1	B1	C1	C2	W2	B2	C3	C4	W3
B1	1	1/4	0.2000	C1	1	3	0.7500	C3	1	6	0.8571
B2	4	1	0.8000	C2	1/3	2	0.2500	C4	1/6	1	0.1429
λ_{max}=7.1623 (为一致性矩阵)				λ_{max}=2 (为一致性矩阵)				λ_{max}=2 (为一致性矩阵)			

由以上结果可以计算出 C1、C2、C3、C4 对总目标的影响的所占权重为:

[C1 C2 C3 C4]=[0.150000 0.05000 0.68568 0.11432]

由此可见，相关性和完整性对于评价结果的影响较大，可解释性和情感对答复意见的影响较小。

3.3.2 答复意见的评价的实现

用[S1 S2 S3 S4]表示完整性、情感、相关性、可解释性量化后的分数，则答复意见评价打分如下：

$$Score = [C1\ C2\ C3\ C4] * [S1\ S2\ S3\ S4]'$$

依层次分析结构模型制定的评价方案对部分抽样得到的答复意见的评价结果如下（完整的答复意见质量评分表见附件 9）：

留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	质量分
UU008190	请求A市允	2014/7/11	各位领	网友：您好	2014/8/13 17:50:52	74.42683
UU008150	反映A3区台	2014/3/27	尊敬的易书	网友：您好	2014/5/9 17:28:09	74.42683
UU008193	反对拆除A	2014/5/9 2	劳动广	网友：您好	2014/6/5 15:41:35	74.42683
UU008287	请求解决A	2014/5/30	易书记	网友：您好	2014/6/16 17:14:57	74.42683
A0003973	请问B9市书	2019/1/13	请问，带小	2019年1月	2019/1/14 16:06:08	63.46663
UU008153	举报C3县书	2019/9/29	本人是长道	你好！201	2019/10/17 12:35:48	52.87701
UU008164	咨询A市残	2018/2/13	请问，你好。如果	2018/4/8 17:06:50	51.57148
UU008145	请求解决L	2018/8/27	L2县竹站村	网友：您反	2018/9/3 10:58:05	49.56194
UU008160	强烈请求L	2018/8/2 1	L10县渠阳	网友：您反	2018/8/9 9:28:25	48.51901
UU008229	咨询乳腺癌	2018/7/23	北京等	“UU00822	2018/7/31 10:06:13	46.62191
UU00863	举报F9市书	2019/3/17	F9市长塘社	您的留言已	2019/4/26 16:37:20	46.59398
UU008164	咨询F9市老	2018/9/19	近期，听F	您的留言已	2018/11/3 14:11:04	45.70435
UU008275	咨询F9市老	2018/9/17	咨询F9市老	您的留言已	2018/9/26 10:00:26	45.53733
UU008199	投诉F3区边	2018/11/3	钱粮湖镇	您的留言已	2018/12/26 9:07:09	45.44574
UU008223	咨询异地公	2019/12/2	我F市户口	您的留言已	2019/12/26 17:36:24	45.44574
UU008678	投诉L7县县	2019/7/7 1	投诉L7县民	网友：你好	2019/8/19 15:44:38	45.44574
UU008560	投诉B市天	2015/12/1	敬启者	网友：	2015/12/17 10:18:09	45.44574
UU008104	咨询K5县老	2019/10/8	你好，我是	网友：你好	2019/10/9 15:09:42	44.79553
UU008217	建议G市地	2019/1/23	由于乡下住	网友：您好	2019/2/22 11:08:10	44.44221
UU008690	反映F市珍	2019/6/20	据了解正在	您的留言已	2019/6/28 12:56:10	43.41785

4. 结论

网络技术水平的发展使政府拓宽了对民意了解的渠道，人们可以相对自由地使用言论来表达自己的意见和想法。但同时，反映社情民意的相关文本数据量也不断攀升，给政府相关部门的工作带来极大的挑战。通过对群众留言文本的分类、热点问题整理以及对政府回复的评价，会给政府相关部门带来极大便利，也为进一步实现“智慧政务”提供方向。

从问题一的研究中我们得出，各个一级标签所对应的留言数量会对该类别准确率产生影响，各分类的准确率也有一定差距，但我们用 LSTM 分类的 F-score 在 0.83 左右。

从问题二的结果分析可以看出，依据评价指标计算出热度指数的高低从而定义热点问题，热度指数排名前五的热点问题为 A 市 58 车贷诈骗问题、A 市 A5 区五矿万境 K9 县房屋出现质量问题、丽发新城小区附近搅拌站噪音扰民、A 市伊景园·滨河苑小区房子和车位捆绑违规销售、A 市 A3 区工地施工扰民为五大热点问题。

从问题三的结果分析可以看出，评价指标完整性、情感、可解释性、相关性对最终的质量得分的影响力各不相同，将四个指标量化后，根据所建立的模型，可以对答复意见的质量做出较为真实合理的评分。

5.参考文献

[1]许学敏. 层次分析法在太阳镜产品质量评价中的应用[D]. 厦门市产品质量监督检验院, 2019(02).

[2]张冲.基于 Attention-Based LSTM 模型的文本分类技术的研究[D].2016.

[3] 公冶小燕,林培光,任威隆.基于改进的 TF-IDF 算法及共现词的主题词抽取算法[J].南京：南京大学报.2017(06).

[4]高星.面向新闻的话题发现和热度评估方法研究[D].2017.

[5]李峰，李明祥，张宇敬.局部迭代的快速 K-means 聚类算法[D].河北市金融学院信息管理与工程系.