“智慧政务”中的文本挖掘应用

摘要

近年来，建立基于自然语言处理技术的智慧政务系统对提升政府的管理水平和施政效率具有极大的推动作用。本文运用了 LSTM 建模和 LDA 主题模型[1]的方法，对“智慧政务”的热点问题进行了挖掘和处理。

针对问题一，我们先对数据进行预处理：去特殊符号、分词、去过滤词以及将“一级标签”转化为相对的 id 数字；然后，我们进行 LSTM 建模，将“留言详情”进行向量化处理，得到最频繁使用的 50000 个词；而后，按照 9:1 划分训练集和测试集，定义一个 LSTM 的序列模型，求得 Loss 与模型准确率。最后，我们通过画混淆矩阵和求 F1 分数来评估我们模型，得出模型的准确性较高。

针对问题二，在第一小题，我们运用了 LDA 主题模型进行文本主题关键字的提取，通过数据挖掘和文本处理，将文本内容（即留言主题）向量化，利用 gensim 库进行了文本向量的计算，文档 TF-IDF 值的计算以及 LDA 模型拟合推断。我们得到的训练模型选取了 30 个主题，我们对 30 个主题的强度进行模型计算，通过 LDA 主题模型的可视化分析，我们得到：LDA 模型训练提取关键字是符合要求的，主题关键字出现的概率呈现一个稳定上升的趋势。因此，我们找出了热度值前 5 的在某一时段特定地点的留言热点问题，分别是（1）A4,A7 区地铁交通问题（2）A 市住宅区业务收房问题（3）A3,A7 小区新城大道扰民问题（4）A 市 A7 区社区建设发展问题（5）A7 地铁涉嫌诈骗和物业房产拖欠。在第二小题，我们结合第一小题 LDA 主题模型提取出的关键字，以及热度值前五的留言热点问题，利用 Python 的第三方库 pandas 进行数据挖掘，查找“留言主题”，“留言详情”中涉及的关键词，进行数据提取。根据五个热点问题分布的不同时间段，进行多次数据分析和提取，将提取到的数据进行整合，数据提合结果见“热点问题留言明细表.xls”。

针对问题三，我们分析了附件 4 的数据，对答复时间和留言时间的差距进行数据的可视化分析。通过答复时间和留言时间的差距情况图，我们得到的结论是：各地政府对不同时间段的留言信息和留言详情都能做到一个快速的处理，能够在相应的时间段给出问题的答复意见，对热点问题的处理也比较好，使得民众的问题能够得到对应政策的解决。

在改进措施和政策的同时，利用好互联网渠道，将对提升政府的管理水平和施政效率具有极大的推动作用。

关键词：LSTM 建模　LDA 主题模型　数据挖掘和分析

Text mining application in "smart government"

abstract

In recent years, the establishment of intelligent government system based on natural language processing technology has played a great role in promoting the management level and efficiency of the government. This paper uses the methods of LSTM and LDA to mine and deal with the hot issues of "smart government".

To solve the first problem, we first preprocess the data: to remove special symbols, segmentation, filter words, and transform the "first level label" into relative ID numbers.Then, we model LSTM and vectorize "message details" to get the most frequently used 50000 words.Then, the training set and test set are divided according to 9:1, and a series model of LSTM is defined to obtain the loss and model accuracy. Finally, we draw confusion matrix and calculate F1 score to evaluate our model, and get the accuracy of the model.

Aiming at the second problem, in the first topic, we use LDA topic model to extract the text topic keywords, through data mining and text processing, we vectorize the text content (i.e. message topic), use gensim library to calculate the text vector, the TF-IDF value of the document, and LDA model fitting and inference.The training model we get selects 30 topics, and we calculate the intensity of 30 topics. Through the visual analysis of LDA topic model, we get that: the key words extracted by LDA model training meet the requirements, and the probability of subject key words appears a steady rising trend.Therefore, we found out the top 5 hot topics of message in a specific place in a certain period of time, which are (1) A4, metro traffic problems in A7 district (2) business room collection problems in residential area of city a (3) A3, new town avenue disturbance problems in A7 district (4) community construction and development problems in A7 District of city a (5) A7 Metro suspected of fraud and property default. In the second topic, we combine the keywords extracted from LDA topic model of the first topic and the top five hot topics of message, use the third-party library pandas of Python for data mining, find the key words involved in "message topic" and "message details", and extract the data.According to the different time periods of five hot issues, data analysis and extraction are carried out for many times, and the extracted data are integrated. See "hot issues message list. XLS" for the data extraction results.

In response to question 3, we analyzed the data in Annex 4, and made a visual analysis of the gap between the response time and the message time. Through the

picture of the gap between the response time and the message time, we can get the conclusion that the local governments can do a fast processing for the message information and message details in different time periods, can give the reply opinions in corresponding time periods, and can deal with the hot issues better, so that the problems of the people can be solved by corresponding policies.

At the same time of improving measures and policies, making good use of Internet channels will greatly promote the management level and governance efficiency of the government.

key word:LSTM modeling LDA theme model Data mining and analysis

目录

1、挖掘目标

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，本次建模目标是建立基于自然语言处理技术的智慧政务系统，提升政府的管理水平和施政效率。本题提供的数据，其中包含结构化和非结构化文本数据，在对数据进行基本的预处理、中文分词、停用词过滤后，一方面根据进行，采用算法，构造类型模型；另一方面构造排名算法判断热门留言，并通过排名情况了解民众的实际需要。

2、分析方法与过程

2.1 问题 1 的分析方法与过程

2.1.1 数据的分析与处理

先进行数据预处理，读取附件 2 中的数据去除重复项及空行，并取“一级分类”及“留言详情”这两列，数据如下：

0	城乡建设	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...
1	城乡建设	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水、电等设施后，烂尾多年，用护栏...
2	城乡建设	尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物业管理有限公司，未经...
3	城乡建设	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味...
4	城乡建设	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味...
5	城乡建设	我在2015年购买了盛世耀凯小区17栋3楼，4楼两层共计2千平方，一直以来我们按...
6	城乡建设	由于西地省地区常年阴冷潮湿的气候，加之近年气候逐渐更加恶劣，地处月亮岛片区近年规...
7	城乡建设	尊敬的胡书记：您好！家住A市A3区桐梓坡西路可可小城的居民长期以来经常停水。小区...
8	城乡建设	我们是梅家田社区辖区内的小区居民，我们每年都依法依规向小区物业公司交纳了城市垃圾...
9	城乡建设	尊敬的A市政府领导：您们好！我是A市A3区魏家坡巷的业主，多年以来，我们小区的脏...
10	城乡建设	尊敬的A市政府领导：您们好！我是A市A3区魏家坡巷的业主，多年以来，我们小区的脏...

图 1：读取原数据

接着，对一级标签的类别进行统计，可以一共有六个类别：

	count	一级标签
0	2009	城乡建设
1	1969	劳动和社会保障
2	1589	教育文体
3	1215	商贸旅游
4	938	环境保护
5	877	卫生计生
6	613	交通运输

图 2

接下来我们要将一级标签类转换成 id，这样便于以后的分类模型的训练。我们将‘一级标签’转换成了 id(0 到 6)

	一级标签			留言详情	一级标签_id	\
4721	教育文体	\n	\n	蒋厅长:您好!我有几个问题想请教一下蒋厅长。1.西地省E12市小...	3	
7812	商贸旅游	\n\t\t\t\t\t\n\t\t\t\t\t		出租车，车用天然气二个月内共提价3次，共涨价...	5	
1738	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t		蒋厅长： 您好，我要举报西地省楚诚嘉园二期...	0	
2192	环境保护	\n\t\t\t\t\t\n\t\t\t\t\t		我家楼上建了个移动的基站，，有辐射吧，，...	1	
1387	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t		官庄镇人民政府:\r\n 介亭驿村民委员会...	0	
9077	卫生计生	\n	\n	别嘌醇片在2013年12月A市药店价为6元/瓶，近两年涨到30元...	6	
6996	劳动和社会保障	\n	\n	本人因在A市工作期间，不慎受伤被路人送往A市区三甲医院西地省旺旺...	4	
4865	教育文体	\n	\n	以公补私行为好不好？不好！道理人人都懂，但这样的现象还是在我们身...	3	
3541	交通运输	\n\t\t\t\t\t\n\t\t\t\t\t		1.隧道技术已成熟。建成合武高速那样连续...	2	
5562	劳动和社会保障	\n	\n	我们是B7县乡镇水管站正式在编在岗职工，2008年B市实行农...	4	

图 3：一级标签转化结果图

由于我们的‘留言详情’都是中文,所以要对中文进行一些预处理工作,这包括删除文本中的标点符号,特殊符号,还要删除一些无意义的常用词。我们过滤掉了‘留言详情’中的标点符号和一些特殊符号，并生成了一个新的字段 clean_留言详情。接下来我们要在 clean_留言详情的基础上进行分词,把每个评论内容分成由空格隔开的一个一个单独的词语，得到以下内容：

	clean_留言详情 \
0	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...
1	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...
2	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...
3	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...
4	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...
5	我在2015年购买了盛世耀凯小区17栋3楼4楼两层共计2千平方一直以来我们按时足额缴纳物业费...
6	由于西地省地区常年阴冷潮湿的气候加之近年气候逐渐更加恶劣地处月亮岛片区近年规划有楚江集中供暖...
7	尊敬的胡书记您好家住A市A3区桐梓坡西路可可小城的居民长期以来经常停水小区业主业委会多次找物...
8	我们是梅家田社区辖区内的小区居民我们每年都依法依规向小区物业公司交纳了城市垃圾处理费我也认为...
9	尊敬的A市政府领导您们好我是A市A3区魏家坡巷的业主多年以来我们小区的脏乱差多次向社区反映都...

图 4：去特殊符号图

	cut_留言详情
0	大道 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 燕...
1	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
2	尊敬 领导 区苑 小区 位于 火炬 路 小区 物业 A 市程明 物业管理 有限公司 未经 小...
3	区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 水是 日常生...
4	区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 水是 日常生...
5	2015 购买 盛世 耀凯 小区 17 栋 楼 楼 两层 共计 千 平方 按时 足额 缴纳 ...
6	西地省 地区 常年 阴冷 潮湿 气候 近年 气候 恶劣 地处 月亮 岛 片区 近年 规划 楚...
7	尊敬 胡书记 您好 家住 A 桐梓 坡 西路 可可 小城 居民 长期以来 停水 小区业主 业...
8	梅家田 社区 辖区 小区 居民 依法 依规 小区 物业公司 交纳 城市 垃圾处理 费 环卫局...
9	尊敬 A 市政府 领导 您们好 A 魏家坡巷 业主 多年 小区 脏乱差 社区 得不到 解决 ...

图 5：过滤停用词，分词结果图

2.1.2 模型的分析与建立

LSTM 建模

A.数据预处理完成以后，接下来我们要开始进行 LSTM 的建模工作：将‘cut_留言详情’，数据进行向量化处理,我们要将每条‘cut_留言详情’转换

成一个整数序列的向量设置最频繁使用的 50000 个词设置每条 cut_留言详情最大的词语数为 250 个。经过代码运行，我们可得到共有 83356 个不相同的词语.

B.训练和测试的数据集都准备好以后,接下来我们要定义一个 LSTM 的序列模型:

模型的第一次是嵌入层，它使用长度为 100 的向量来表示每一个词语；SpatialDropout1D 层在训练中每次更新时， 将输入单元的按比率随机设置为 0，这有助于防止过拟合 LSTM 层包含 100 个记忆单元；输出层为包含 10 个分类的全连接层；

由于是多分类，所以激活函数设置为'softmax'由于是多分类, 所以损失函数为分类交叉熵 categorical_crossentropy；

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_4 (Embedding)      (None, 250, 100)          5000000
_________________________________________________________________
spatial_dropout1d_3 (Spatial (None, 250, 100)          0
_________________________________________________________________
lstm_3 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_3 (Dense)              (None, 10)                1010
=================================================================
Total params: 5,081,410
Trainable params: 5,081,410
Non-trainable params: 0
_________________________________________________________________
None
```

C. 定义好 LSTM 模型以后，我们要开始训练数据:设置 5 个训练周期，batch_size 为 64

```
Train on 7460 samples, validate on 829 samples
Epoch 1/5
7460/7460 [==============================] - 74s 10ms/step - loss: 1.6572 - acc: 0.3572 - val_loss:
1.1929 - val_acc: 0.5694
Epoch 2/5
7460/7460 [==============================] - 93s 12ms/step - loss: 0.9247 - acc: 0.6799 - val_loss:
0.7717 - val_acc: 0.7612
Epoch 3/5
7460/7460 [==============================] - 80s 11ms/step - loss: 0.4913 - acc: 0.8558 - val_loss:
0.6345 - val_acc: 0.8154
Epoch 4/5
7460/7460 [==============================] - 76s 10ms/step - loss: 0.2365 - acc: 0.9395 - val_loss:
0.5783 - val_acc: 0.8384
Epoch 5/5
7460/7460 [==============================] - 77s 10ms/step - loss: 0.1527 - acc: 0.9629 - val_loss:
0.5881 - val_acc: 0.8323
921/921 [==============================] - 2s 2ms/step
Test set
  Loss: 0.641
  Accuracy: 0.807
```

下面我们画损失函数趋势图和准确率趋势图:

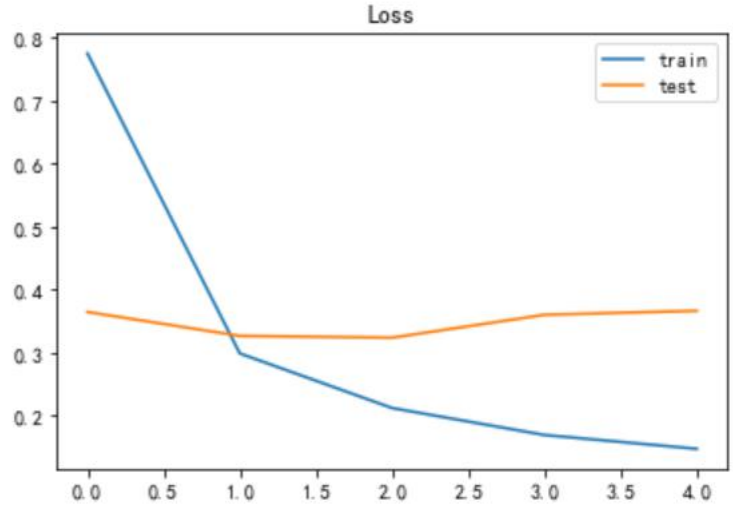

图 6：损失函数趋势

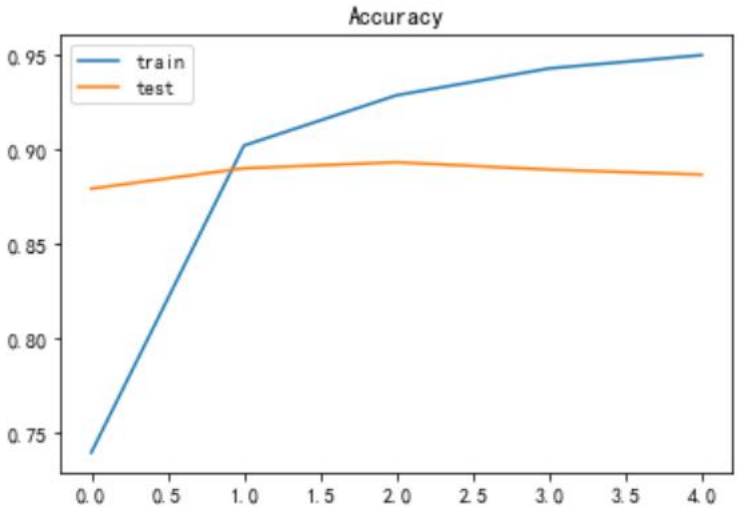

图 7：准确率趋势

从上图中我们可以得出以下两点：

a.随着训练周期的增加,模型在训练集中损失越来越小,这是典型的过拟合现象,而在测试集中, 损失随着训练周期的增加由一开始的从大逐步变小,再逐步变大;

b.随着训练周期的增加,模型在训练集中准确率越来越高,这是典型的过拟合现象,而在测试集中, 准确率随着训练周期的增加由一开始的从小逐步变大，再逐步变小。

2.1.3 模型的评估

LSTM 模型的评估 ： 接下来我们通过画混淆矩阵和求 F1 分数来评估我们模

型的表现，得到图 8：

	precision	recall	f1-score	support
城市建设	0.91	0.96	0.95	366
劳动和社会保障	0.78	0.82	0.80	1000
教育文本	0.87	0.81	0.84	270
商贸旅游	0.93	0.90	0.92	965
环境保护	0.82	0.83	0.83	1006
卫生卫计	0.88	0.54	0.67	54
交通运输	1.00	1.00	1.00	207
avg / total	0.89	0.89	0.89	6278

图 8：各类标签的各项指标

使用 F-Score 对分类方法进行评价：

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

由图 8 中的数据，我们可以计算得到 $F_1 = 0.87$

2.2 问题 2 的分析方法与过程

2.2.1 数据的分析与处理

首先，根据附件 3 的数据，我们对数据内容进行总结和整理。对每一个月出现的群众热点问题进行分析，删除数值为空的数据和重复的数据行，对留言主题内容进行文本预处理。

接着，利用 pandas 第三方库进行数据分析，把热点问题的留言主题分为对 12 个月份，进行文本分词，去除停用词，删除无关词汇等一些对提取留言文本有影响等操作。接着，统计词频，对分词的结果进行词频统计，根据词频在前 300 的词汇做出词云图，如附件（1-12），对 12 个词云图展现出来的群众集中热点问题进行分类和归纳处理。

根据数据分析的结果，我们用 pandas 数据筛选的方式，数据处理选出“附件 3”中一月份的留言内容数据，挑选点赞数排名前 20 的用户留言情况，进行分析，观察“反对数”和“点赞数”两列数值。如图 9:

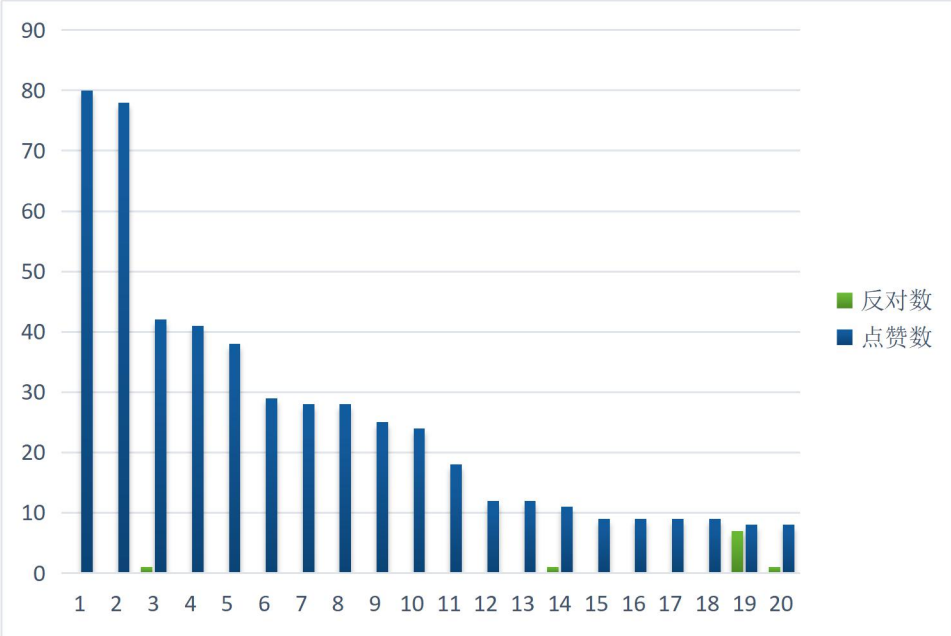

图 9:1 月份点赞数前 20 的留言点赞数、反对数的情况图

最后，我们发现各个时间段对于某个热点的留言主题，反对数都是小于 10 的，由于热点问题应该反映大众的集中问题，方便政府机构和相对应的部门进行有处理和管理，所以我们选取的热点问题不考虑反对数带来的影响。

2.2.2 模型的分析与建立

根据数据的分词结果，我们采用 LDA 主题模型：

文本分类任务中，LDA 主题模型可以用来选择特征，因为训练数据中含有类别信息，可以在不同类别的结果中，删除相同的、比较常见的主题词，选取关键词，为主题类别提供更好的特征。

我们对 LDA 主题模型进行简单分析。由于提供的“留言主题”和“留言详情”这两列数据的文本长又多，如果只进行简单的分词，再进行词频统计，会出现一些不必要的词汇，无法提取重要信息。因此，我们采用 LDA 主题模型进行建模处理，通过 python 运算找出具有重要意义的关键词，对后续操作带来便利，易于理解。

在此我们利用 python 第三方开源库 Gensim,用于从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达，直接调用进行数据处理和模型训练。

由于附件 3 的数据是一个表格的形式，其中的“留言主题”和“留言详情”有很多行，我们把这些数据行当成文档来处理，读入数据集；用 jieba 库进行文本分词，去除停用词和无关干扰词，完成文本内容的预处理。

训练算法。我们通过比较分析 TF-IDF 算法，LSI 主题算法和 LDA 主题算法的差异，如表 1：

算法	训练过程
TF-IDF 算法	根据数据集生成对应的 IDF 值字典，计算 TF-IDF 值时，直接从字典中读取
LSI 主题算法	根据现有的数据集，生成文档-主题分布矩阵和主题-词分布矩阵
LDA 主题算法	

因此，我们决定选用 LDA 主题算法，处理得到的数据集较直观，公式的符号说明如下表 2：

符号	符号说明
w_i	第i个单词
d_i	第i个文档
t_i	第i个主题
a	每个热度问题发生在同一地点（人群）的概率
b	在同一个地区内出现相同群众问题的概率
p	一个热点问题出现的热度值
p_{ti}	表示d_i对应 k 个主题中第 i 个主题的概率
p_{wm}	表示主题t_i生成 V 中第 i 个单词的概率
θ_k	不同主题的概率
θ_t	不同单词的概率向量

LDA 链式关系：

$$P(w_i \mid d_i) = P(w_i \mid t_i)P(t_i \mid d_i) \quad (2)$$

即

将每篇文档d_i看作个单词序列：

$$W = (w_1, w_2, ..., w_n) \quad (3)$$

所有文档涉及的所有不同单词组成一个词汇表大集合V，LDA 以文档集合D作为输入，训练出的两个结果向量 (假设形成k个主题，V中共有m个词)。

（1）对每个文档d_i，对应到不同主题的概率$\theta_k = (p_{t1}, p_{t2}, ..., p_{tk})$

（2）对每个 T 中的主题中的主题t_i，生成不同单词的概率向量$\theta_t = (p_{w1}, p_{w2,}, ..., p_{wm})$

LDA 算法开始时，先随机地给θ_k, θ_t赋值

对文档集 D 中的所有文档d_i中的所有w_i进行一次公式（2）计算，并重新选择主题这是一次迭代过程，多次迭代得到模型拟合。

运用 LDA 主题模型，将文本内容（留言主题）向量化，利用 gensim 库进行了文本向量的计算，文档 TF-IDF 值的计算，文本相似度[2]的计算 LDA 模型拟合推断。

训练模型选取了 30 个主题，我们对 30 个主题的强度进行模型计算，得到各个主题的强度如下：

[0.0118327 0.01491317 0.0156075 0.01784284 0.01822634 0.01856311 0.01906548 0.01984119 0.02072324 0.02090891 0.02142221 0.02412721 0.02481293 0.02825058 0.02935638 0.03015174 0.03071561 0.03123711 0.03306191 0.03521906 0.03906774 0.04094503 0.04297512 0.04335678 0.04656486 0.0552396 0.05706833 0.05909378 0.06058098 0.08922856]

对应的主题出现频率如图 9：

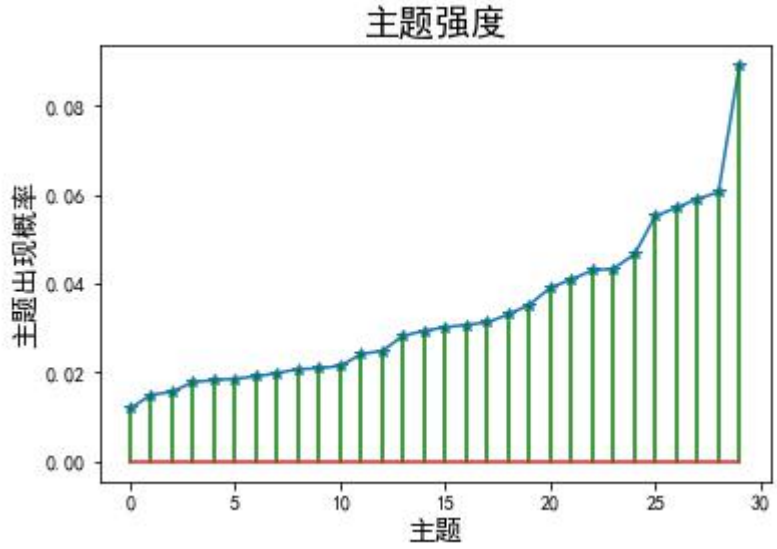

图 10：30 个主题文本的 LDA 主题模型训练

由图 10 我们可以得到，LDA 主题模型训练提取关键字是符合要求的，主题出现的概率呈现一个稳定上升趋势。

2.2.3 定义热度评价

首先，结合 LDA 主题模型，提取文本（即留言问题）的关键字，再根据文本分词，对文本进行预处理之后，得到了 12 个词云图，选择 2 月份和 5 月份的词云图，如下图：

图 11：2 月份词云图

图 12：5 月份词云图

我们总结了 2019 年 12 个月份中的群众反映问题，并整理热度比较大的群众问题。对出现的问题进行量化，将同一个热点问题出现的总频率作为热度评价指标，假设每个热度问题发生在同一地点（人群）的概率为 $a=\frac{i}{N},(i=0,1,2,...,N)$，在同一个地区内出现相同的群众问题的概率为

$\mathrm{b}=\frac{j}{M},(j=0,1,...,M)$，因此，一个热点问题出现的热度值为$P=ab,(0\le P\le 1)$。

综上，我们计算出对应问题的热度评价表，如表 3：

时间范围	地点/人群	问题描述	热度评价值
2019.01.01-2019.01.31	A 市地铁	A 市地铁运输对群众生活的影响	0.00143472
2019.01.01-2019.01.31	A 市 A7 区	A 市 A7 区社区建设发展问题	0.032281205
2019.01.01-2019.01.31	A 市小区	A 市业主与拖欠服务	0.011477762
2019.01.01-2019.01.31	A 市西北省	西北省加大城区建设问题	0.008608321
2019.02.01-2019.02.28	A 市 A7 区	A7 小区扰民问题	0.06456241
2019.02.01-2019.02.28	A 市 A7 区幼儿园	A7 区幼儿园学校周边建设问题	0.021520803
2019.02.01-2019.02.28	A 市车站	A 市公交车违规相关政策和力度	0.00071736
2019.02.01-2019.02.28	A 市 A7 区住宅区	A 市 A7 区住宅区反映业主收房问题	0.043041607
2019.02.01-2019.02.28	A 市 A5，A2 区学校周边	A5，A2 区学校周边建设问题	0.005738881
2019.03.01-2019.03.31	A3,A7 街道小区	A3,A7 街道投诉问题	0.006097561
2019.03.01-2019.03.31	A 市幼儿园	A 市幼儿园业主和物业问题	0.00286944
2019.03.01-2019.03.31	A3,A7,A4 区	A3,A7,A4 区咨询问题	0.015064562
2019.03.01-2019.	A4 区教师处	A4 区教师处社区购房咨询	0.008608321

03.31			
2019.03.01-2019.03.31	A 市 A7 区	A 市 A7 区加快城区建设问题	0.032281205
2019.04.01-2019.04-30	A1,A3 区地铁	A1,A3 区地铁扰民投诉问题	0.00286944
2019.04.01-2019.04-30	A3,A7 区街道	A3,A7 区街道违规投诉	0.005738881
2019.04.01-2019.04-30	A 市西北省	A 市西北省社区扰民严重问题	0.017216643
2019.05.01-2019.05-31	A 市	A 市区违规，安全隐患严重	0.005738881
2019.05.01-2019.05-31	A 市小区	A 市小区街道扰民问题严重	0.017216643
2019.06-01-2019.06.30	A3,A7 小区	A3,A7 小区噪音扰民问题	0.053802009
2019.06-01-2019.06.30	A4,A7 区地铁	A4,A7 区地铁和物业医疗问题	0.020444763
2019.06-01-2019.06.30	A 市 A1,A7 区购房区	A 市 A1,A7 区购房周边，违规违规	0.019368723
2019.07.01-2019.07.31	A 市	A 市扰民噪音严重问题	0.014347202
2019.07.01-2019.07.31	A1 区街道	A1 区街道垃圾投诉	0.00215208
2019.08.01-2019.08.31	A7,A4 小区	A7,A4 小区车位违规投诉	0.006814921
2019.08.01-2019.08.31	A 市 A7 区	A 市 A7 区销售捆绑问题	0.010760402
2019.08.01-2019.08.31	A 市滨河	A 市滨河地铁扰民问题	0.00286944

2019.09.01-2019.09.30	A市西北省中路公园	西北省中路公园施工存在安全隐患	0.005738881
2019.09.01-2019.09.30	A市小区街道	西北省扰民投诉问题	0.017216643
2019.10.01-2019.10.31	A市小区街道	A市小区街道施工路段扰民问题	0.014347202
2019.10.01-2019.10.31	A市A7地铁	A7地铁涉嫌诈骗和物业房产拖欠	0.043041607
2019.10.01-2019.10.31	A市A7区	A市A7区存在安全隐患扰民问题	0.021520803
2019.11.01-2019.11.30	A2区附近	A2区附近修建厂噪音问题	0.014347202
2019.11.01-2019.11.30	A3新城大道	A3新城大道涉嫌扰民噪音问题	0.007173601
2019.11.01-2019.11.30	A2小区附近	A2小区附近社区街道扰民	0.017216643
2019.12.01-2019.12.31	A市西北省小学	A市西北省小学业主投诉墙面	0.005738881
2019.12.01-2019.12.31	A3,A7新城大道	A3,A7新城大道涉嫌扰民	0.036585366
2019.12.01-2019.12.31	A市地铁	A市地铁国际政策建设	0.00143472
2019.12.01-2019.12.31	A市A2区	A市A2区停车场附近噪音扰民	0.014347202

接着，对表3的热点问题的热度评价值进行排序，我们选取了排名前五的热点问题，得到的评价结果如图13：

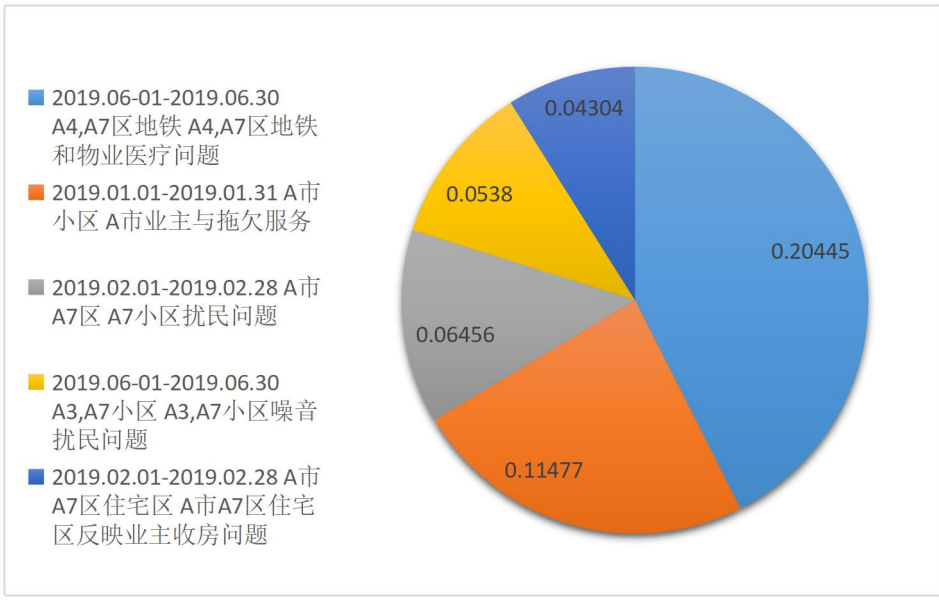

图 13:热度评价值排名前五的热点问题分布图

将热点问题表保存为“热点问题表.xlsx”

另外，在第一小题的条件下，我们经过数据挖掘和文本处理，找出了五个在同一时段特定地点问题的留言热点问题。在某一阶段，若可以及时发现相关热点问题，有助于相关部门进行针对性处理，提升服务效率。

对于问题 2 的第二个小问题，我们主要依据 LDA 主题模型提取出的关键字，以及前五大留言热点问题，利用 Python 的第三方库 pandas 进行数据挖掘，搜索“留言主题”，“留言详情”中涉及的关键词，进行数据提取。根据五个热点问题分布的不同时间段，进行多次数据分析和提取，将提取到的数据进行整合，得到相应热点问题对应的留言信息，数据提取结果见“热点问题留言明细表.xls”.

2.3 问题 3 的评价方案

结合附件 4 的数据，我们作出了答复时间和留言时间差距图，如图 13：

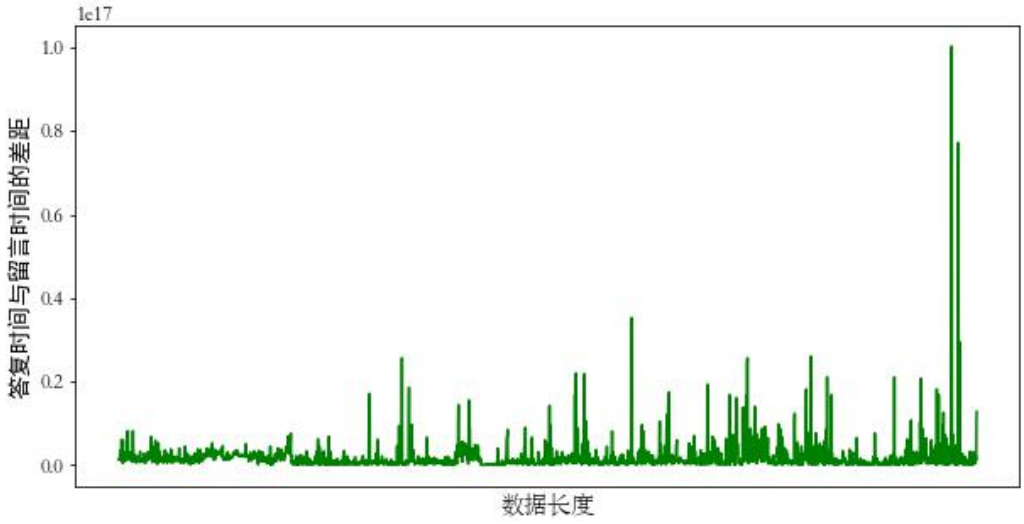

图 14：答复时间和留言时间的差距情况图

（1）从答复时间上看，我们对附件 4 的数据进行可视化处理，计算了留言时间和答复时间之间的差值，从图 13 来看，我们可以看出，各地政府对不同时间段的留言信息和留言详情都能做到一个快速的处理，能够在相应的时间段给出问题的答复意见，对热点问题的处理也比较好，使得民众的问题能够得到对应政策的解决。从图上看，有少部分时间还是处于一个缓慢回复群众问题的状态，如果非热点问题，对社情民意问题的解决可能影响不大；如果恰好这部分时间是高频率发生的热点问题，对社会民众一些生活物质经济方面可能会带来很大的影响。因此，各个地方政府以及相关部门的工作可以合理分配，不断提高政府的管理水平和施政效率。

（2）答复问题的相关性和完整性。关于社会状态下的民生问题，一直以来都围绕着住房，就业，医疗，教育，环境这几个方面。对于附件 4 给的答复意见，政府和各个部门的联手合作，改善民生问题，缓解广大群众的精神压力等都做的很好，回复的内容与热点相关，具有一定的完整性。热点问题下政府给予的答复建议，从可行措施和不可行措施出发，对可以及时处理的问题，采取对应措施，在一定时间内处理民众问题；对于不可行措施，政府及相关部门可以告知民众：该政策不可实施的原因以及后续采取的改进方案，给予民众精神上的慰藉。

（3）改进方案：

政府应坚持从腐败现象发生的经济、政治、文化、社会等原因入手，从宏观上进行总体设计；坚持统筹兼顾的原则，使各方面制度紧密配合、综合发挥作用。具体而言，应突出抓好以下几方面的制度建设：

第一，进一步加强反腐倡廉教育制度建设。增强对应社区公民和政府从事的工作人员的法律观念和制度意识，是加强反腐倡廉制度建设的根本。应通过

制度建设，加强和改进反腐倡廉宣传教育工作，提高国家工作人员廉洁从政的素质和公民依法监督的自觉性，使预防腐败工作更有成效，也可以减少民众对腐败现象不满带来的社会热点的弊端。

第二，进一步加强监督制度建设，形成健全完善的监督制度。完善的监督制度，可以保证权力的授予和运行受到必要的监督和制约。政府应通过进一步细化和完善监督制度的有关规定，改革和完善监督体制，拓宽监督渠道，号召民众共同监督，也便于及时处理问题，逐步减少热点问题的产生。

第三，教育，住房，医疗等方面的措施逐步完善。

A、政府和相关部门要努力改善现有的教育不公平现状，让教育资源更加合理配置，减少关于教育不平衡和不合理配置的热点问题。促进教育公平，提倡教育家办学，打破潜规则、斩断利益链，去除学校行政级别以及如何应对和改善困扰多年的教育发展不均衡的问题。

B、对待医疗方面的问题，打破“看病难、看病贵”这一不合理说法，使民众在医疗方面的困难可以解决。

第四，生活环境的改善。政府及相关部门对一些不合理，导致民众生活环境受到了影响和干扰的行为，要采取一定措施。从地方民警处落实，面对深夜扰民和不合理的业务拖欠，应制定相关惩罚措施。

解决民生问题，减少热点问题的产生。关于互联网逐渐作为政府了解民意、汇聚民智、凝聚民气的重要渠道，在改进措施和政策的同时，利用好互联网渠道，将对提升政府的管理水平和施政效率具有极大的推动作用。

3. 结果分析

3.1 问题 1 的结果分析

问题 1 采用 LSTM 模型进行建模，设置五个训练周期，每一次的训练结果，损失率 loss 下降，准确率上升；模型的评估 F_1 达到 87%，说明分类结果较好。如果想取得更加准确的分类效果，我们可以多设置几个训练周期。

3.2 问题 2 的结果分析

（1）第一小题，我们运用了 LDA 主题模型进行文本主题关键字的提取，将文本内容（即留言主题）向量化，利用 gensim 库进行了文本向量的计算，文档 TF-IDF 值，文本相似度的计算，以及 LDA 模型拟合推断。通过 30 个主题文本的 LDA 主题模型的可视化分析，得到结论：LDA 模型训练提取关键字是符合要求的，

主题关键字出现的概率呈现一个稳定上升的趋势。

其次，进行热度值指标量化，找到了热度值前 5 的在某一时段特定地点的留言热点问题，如下表 4：

时间范围	地点/人群	问题描述
2019.06-01-2019.06.30	A4,A7 区地铁	A4,A7 区地铁交通问题
2019.01.01-2019.01.31，2019.02.01-2019.02.28	A 市小区	A 市住宅区业务收房问题
2019.02.01-2019.02.28，2019.12.01-2019.12.31，2019.06-01-2019.06.30	A 市 A3,A7 区	A3,A7 小区新城大道扰民问题
2019.01.01-2019.01.31，2019.03.01-2019.03.31	A 市 A7 区	A 市 A7 区社区建设发展问题
2019.10.01-2019.10.31	A 市 A7 地铁	A7 地铁涉嫌诈骗和物业房产拖欠

（2）第二小题，利用 Python 的第三方库 pandas 进行数据挖掘，查找“留言主题”，“留言详情”中涉及的关键词，进行数据提取。根据五个热点问题分布的不同时间段，进行多次数据分析和提取，将提取到的数据进行整合，数据整合结果见“热点问题留言明细表.xls”。

缺点：数据提取和整合过程存在一定的误差，但对大文本主题关键字的读取影响不大。

4.参考文献

[1]晓辉,孙静.LDA 主题模型[J].智能计算机与应用,2014(5):105-106.

[2]王振振,何明,杜永萍.基于 LDA 主题模型的文本相似度计算[J].计算机科学,2013(12):229-232.