基于自然语言处理技术的智慧政务系统研究报告

Research report on intelligent government system based on natural language processing technology

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也有了发展基础，目前来看，智能政务系统正处在方兴未艾的发展中。

对于本次赛题针对智能政务系统对群众留言分类、热点问题挖掘以及答复意见评价三方面问题，我们分别通过文本特征词分析、文本相似度分析、时间排序、热度计算及排序、文本主题模型、文本结构分析，对应问题要求达成目标。

对于赛题复杂的要求，其中通过使用数据框工具 Pandas、jieba 分词包、tfidfvectorizer 方法、abelencoder 方法、逻辑回归模型方法、TF-IDF 模型、apply 函数、LDA 方法、Excel 软件的 GetMatchingDegree 函数、doc2bow、TFIDF 等达成以上各类功能。最后通过使用 F-Score 进行了对分类模型的评价，通关输出结果对热点问题挖掘模型与答复意见评价模型进行了检测，并最终得到赛题要求效果，有利证明了三个模型的可行性、实用性。

关键词：文本特征词分析；文本相似度分析；文本结构分析；TF-IDF ；Doc2Bow.

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 intelligent government system based on natural language processing technology also has a development basis. At present, intelligent government system is in the ascendant.

For the three aspects of intelligent government system, i.e. message classification, hot issue mining and response evaluation, we achieved the goal through text feature word analysis, text similarity analysis, time sorting, heat calculation and sorting, text subject model and text structure analysis.

For the complex requirements of the competition, the above functions are achieved by using the data frame tools pandas, Jieba word segmentation package, tfidfvector method, abelencoder method, logical regression model method, TF-IDF model, apply function, LDA method, getmatchingdegree function, doc2bow, TFIDF of Excel software. Finally, through the use of F-score to evaluate the classification model, the output results of customs clearance test the hot issues mining model and the reply evaluation model, and finally get the effect of the requirements of the game, which is beneficial to prove the feasibility and practicability of the three models.

Key words: Text feature word analysis; text similarity analysis; text structure analysis; TF-IDF; doc2bow.

目 录

一、引言

1.1 问题重述

1.1.1 群众留言分类

首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。 通常使用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n_1}\sum_{i=1}^{n}\frac{2P_iR_i}{P_iR_i}$$

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

1.1.2 热点问题挖掘

请根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

1.1.3 答复意见的评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

1.2 研究背景及意义

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。

推进“互联网+政务服务”，促进部门间信息共享，是深化简政放权、放管结合、优化服务改革的重要内容。为进一步推动部门间政务服务相互衔接，协同联动，打破信息孤岛，变“群众跑腿”为“信息跑路”，变“群众来回跑”为“部门协同办”，变被动服务为主动服务，建设政务服务平台，具有以下意义：

1、有助于拓展服务渠道，体现服务便民

2、有助于抢占舆论阵地，扩大政府影响

3、有助于加强舆情控制，树立政府威信

4、有助于强化政民互动，拉近政民距离

即为，建立完善的政务服务平台，有助于政府各部门更高效、便民、创新、透明、规范、廉洁的处理信息，群众也更容易感受到政府为人民服务的决心。

（1）杜绝重复建设，节约财政支出。智慧政务系统可以充分利用现有的基础资源，有效促进各种资源整合，为政府部门提供资源、安全、运维和管理服务，能够提升基础设施利用率，减少运维人员和运维费用。

（2）促进信息共享，实现业务协同。通过智慧政务系统，在政府部门之间、政府部门与社会服务部门之间建立“信息桥梁”，通过系统内部信息驱动引擎，实现不同应用系统间的信息整合、交换、共享和政务工作协同，将大大地提高各级政府机关的整体工作效率。

（3）优化资源配置，提升服务能力。通过智慧政务系统，政府的程序和办事流程更加简明、畅通，使人力和信息资源得到最充分的利用和配置。同时，采用智慧政务系统集约化模式建设电子政务项目，可以使政府部门从传统的硬件采购、系统集成、运行维护等工作中解脱出来，转而将更多的精力放到业务的梳理和为民服务上来，极大提升为民服务的能力和水平。

1.3 研究现状及发展态势

智慧政府建设需要综合运用云计算、物联网、移动互联网、大数据、人工智能、语义网络、实境网络、Web3.0 等技术工具。这方面代表性的研究成果有：以智慧政务云平台规划设计为背景，提出智慧政务 Web 服务的语义模型，研究了分阶段多策略的服务发现方法，分析了智慧政务中系统集成和业务协同面临的关键问题，并针对 SOA 和 Web 服务技术提出了有效的解决方法；研究了云计算、大数据、物联网和移动互联网技术对智慧政府建设带来的便利和推动作用。

智慧政务系统应具有移动性、无缝性、实时性、集成性、泛在性、可视化、透彻感知、需求预测、快速反应、个性化订制、主动服务、场景导航、无障碍服务、基于位置的服务等特征。

在新兴信息技术的驱动下，伴随着激增的网络用户群体、泛在的网络接入服务、丰富的信息提供形式、多样的在线公共服务以及活跃的政民网络互动，智慧政府在整个政府信息化过程中将扮演更加积极的角色，并显现出强大的可持续发展能力。这样一个复杂的系统必然和信息生态环境有物质、能量和信息的交换，所以智慧政务系统未来可以定义为开放的复杂系统。按照系统论的思想，智慧政府协同发展是一个有机动态系统，要研究智慧政府的未来发展问题，需要构建科学、严谨并具有普世意义的理论模型。当然，新兴信息技术的应用在进一步凸显智慧政务系统特征的同时，也面临着日益严峻的挑战。

综上所述，智慧政务系统是一个比较新兴的研究领域，发现和诠释智慧政府的运行规律，解决智慧政府建设中基础性、开拓性的科学问题，构建智慧政府研究的理论方法体系，对于未来相关研究领域的进步，以及推进中国政府信息化建设工作都具有十分重要的理论与实践意义[1]。

二、问题的解答与模型建立

2.1 群众留言分类

对于问题一在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且 差错率高等问题。由此我们可以借助强大的 NLP[2]（自然语言处理）来解决这一问题。

2.1.1 结构图

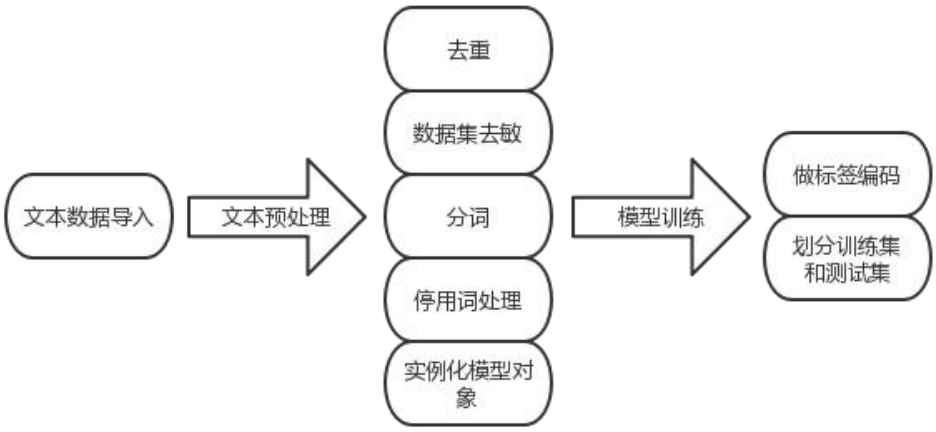


图1 问题一的研究分析路径图

Fig.1 Research and analysis path of question one

2.1.2 实现方法

第一步：文本预处理

1、 首先我们将数据集以留言详情进行初步的去重操作。

2、 数据集去 min 操作，对过滤后的数据会出现特殊符号来代替一些涉及个人信息的文本内容。

3、 对数据做分词处理，由于中文文本中词与词之间没有明显的分隔。要从众多的语句中提出关键词，就可以用上 Python 开发的中文分词模型 jieba 分词，对其进行留言主题分词。

4、 停用词处理，将分词完的结果用中文停用词表将分词后没有实际意义的词语去掉。此处停用词表取自文献 3[3]。

5、 调用 sklearn feature extration text[4]库中的 tfidfvectorizer 方法实例化模型对象。

第二步：模型训练

1、 调用 preprocessing 库中的 abelencoder 方法对文章分类做标签编码。以便于对后续的分类进行划分。

2、 调用逻辑回归模型方法划分训练集和测试集。本次将按照 2:8 的比例将数据划分为测试集和训练集。

2.1.3 结果说明

1、经过交叉检验可以得到 3 次检测值，公式为：

$$F_1 = \frac{1}{n_1}\sum_{i=1}^{n}\frac{2P_iR_i}{P_iR_i}$$

的返回值为 0.44，其中利用绘制混淆矩阵可以对模型进行可视化评估，具体参见表一：

表1 模型可视化评估

Tab.1 Model visual evaluation

	交通运输	劳动和社会保障	卫生计生	商贸旅游	城乡建设	教育文体	环境保护
交通运输	24	11	0	6	73	9	0
劳动和社会保障	1	170	14	9	168	23	2
卫生计生	4	22	43	9	89	9	4
商贸旅游	3	20	6	74	123	11	1
城乡建设	3	42	5	12	299	17	13
教育文体	1	46	3	11	157	110	5
环境保护	0	15	5	5	93	7	65

2.2 热点问题挖掘

对于热点问题的讨论，现代社会很多方面都涉及关于热点问题的排序，其代表的是问题本身带来的影响力和渗入时间和数量对热点形成一个定义。如本题中某一时段内群众集中反映某一问题，可称为热点问题，及时发现热点问题将有助于相关部门进行有针对性的处理。

2.2.1 结构图

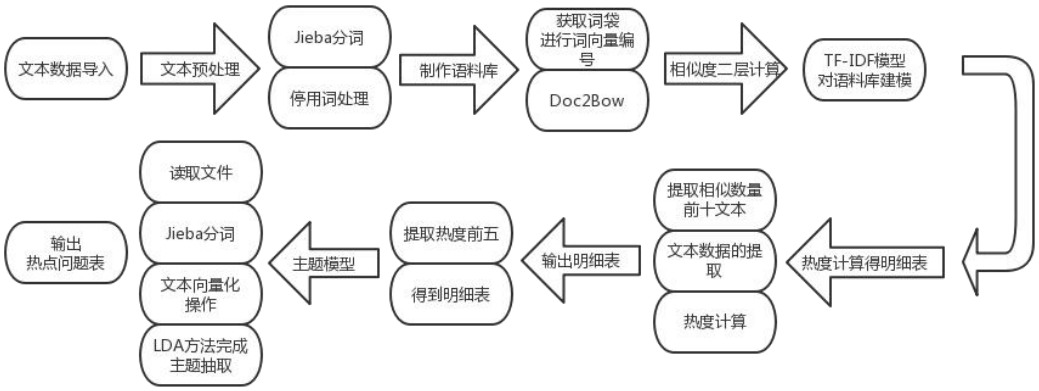


图2 问题二的研究结构图

Fig.2 Question 2 Research Structure

2.2.2 实现方法

第一步：文本预处理

1、 Jieba 分词，对于题二中的文本预处理，在 jieba 分词表的基础上添加了文本，赛题对城市地区字母化后产生的地点名，在分词时无法合并，由此在 jieba 上添加 A1 市—A9 市等数据。

2、 停用词处理。

第二步：制作语料库

1、 首先用 dictionary 方法获取词袋（bag-of-words）—在信息检索中，假定对于一个文本忽略其词序、语法及句子结构。将其看做成一个词的集合，文本中每个词不依附于其他词，其出现都是独立的。在做成磁带之后，用数字对所有词进行编号。通过 token2id 得到特征数，一个词仅对应一个编号。

2、 调用 Gensim 中封装的一个方法 Doc2Bow。Bow 使用一组无序的单词(Words）来表达一段文字和一个文档。以同样的方法把测试文档也转为二元组的向量。

第三步：相似度二层计算

1、使用 TF-IDF 模型对语料库建模，从中获取测试文档中每个词的 TF-IDF 值，即可求出测试文本（text1）和文本集的相似度[5]。通过计算文本之间的相似度，其中设置相似度超过 0.5 归成一类。根据词袋进行构建语料库，判断出文本之

间的相似。通过余弦定理进行计算，发现很多同一类的问题会因同义词的出现，而出现断层，因此可用循环函数对问题进行二层相似，思路构建如下：

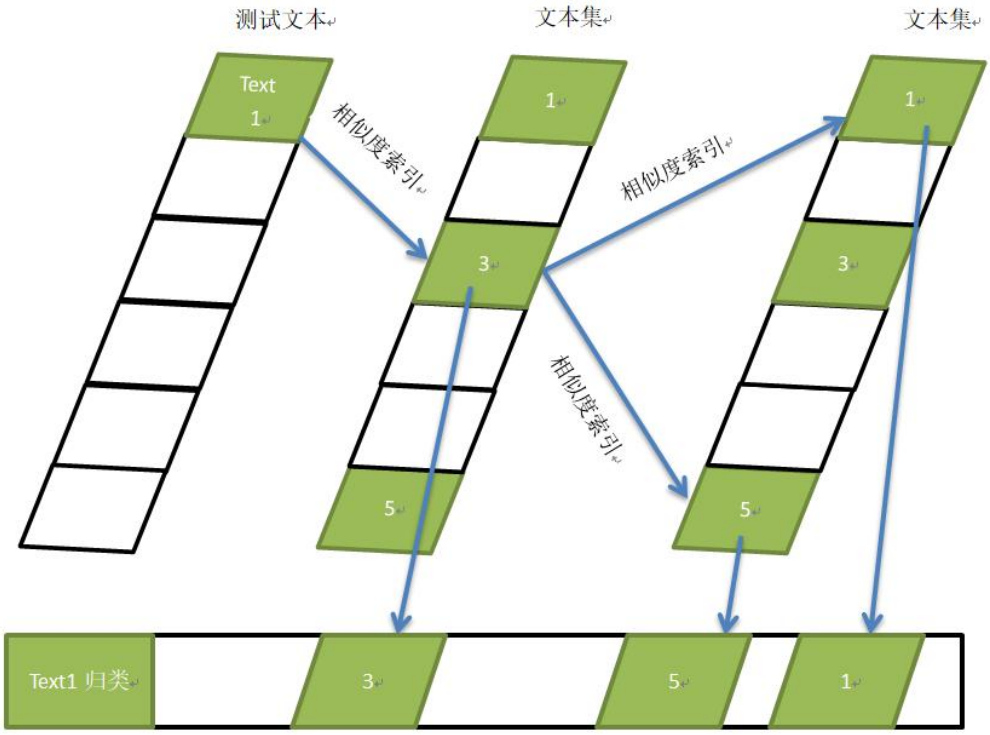


图3 相似度的思路构建

Fig.3 Construction of similarity

第四步：对相似文本数量前十的组别进行热度指数计算，得出“热点问题明细表”

1、提取相似文本数量前十的所有编号。用常用的排序方法 sorted 对所有组别每组相似数量进行降序。

2、对文本数据进行提取。在这我们做了一个 for 循环，将提取出来的数据用主键合并的函数 concat，对其参数 join 选择 inner 进行上下的、保留原序号的形式进行合并，放到新的 DataFrame 中，这可以更直观的看到所匹配出来的数据是否符合我们的预期。同时将各组的点赞数、反对数、留言时间用 append 函数整理出来，用于计算热度需要的各组点赞总数、反对总数、时间范围。再修改新表中的时间格式为 object，方便下面对时间进行排序的操作。

3、计算热度。将时间转成 datetime 的格式后，计算每组时间跨度，用一个嵌套的 for 循环求和，得点赞总数和反对总数，设计热度公式为：

$$热度=\frac{(相似文本篇数*0.7+(总反对数+总赞数)*0.3)*1000}{(时间跨度+2)^{1.2}}$$

注：热度公式设计理由：根据（文献[6]）对应实际变换得出，相似文本篇数占比应比总反对数与总赞数大，相似文本篇数可看作在人群中影响程度，总反对数与总赞数之和可看做该问题在人群中的关注度。

第五步：造热点问题留言明细表和热点问题表

1、按热度提取前五。将‘热度指数’、‘问题 ID’的空值列加到“留言相似排名前十数据表”，再对热度进行排序。再次利用与上面文本数据提取中相似的 for 循环，将提取出来的前五的数据用主键合并放到新的 DataFrame 中，并同时将每组文本内的多条文本按时间升序进行排序、用一个列表推导式添上每行‘问题 ID’的值，得到包含热度的“热点问题留言明细表”。

2、得到“热点问题留言明细表”。将“热点问题留言明细表（含热度）”中的热度用 del 去掉，得到最终的“热点问题留言明细表”。

第六步、主题模型的主题数确定和可视化：

1、通过使用数据框工具 Pandas 读入已经完成热度计算及排序的“热点问题留言明细表”（热点问题留言明细表.xlsx）编辑命令检测数据读取的正确性与完整性；

2、调用 jieba 分词包，把文本数据并行化处理并利用 apply 函数对文本进行分词，编辑命令检测其正确性；

3、进行文本向量化操作，从中提取 1010 个最重要的特征关键词进行向量转换；

4、应用 LDA[7]方法完成主题抽取，指定 50 个主题进行刻画，编辑程序定义函数，定每个主题输出前 10 个关键词；编辑可视化模型，输出交互动态图，在对应集合圆上悬停手动提取高频词语组合成“热点问题表”中的问题描述。

第七步、输出“热点问题表”

1、得到热点问题表。将“热点问题留言明细表”（热点问题留言明细表.xlsx）中的每组相似文本的时间范围、地点 / 人群、问题描述用代码结合人工的方式提取出来。

2.2.3 结果说明

相似度二层计算部分结果：

表 2 相似前十部分编号

Tab.2 Similar to the first ten part numbers

```
[3636, 9, '相似文本编号：', [47, 64, 386, 1100, 1110, 1684, 2193, 2757, 3308, 3636]],
[4280, 8, '相似文本编号：', [826, 2739, 3741, 4280, 1905, 1992, 2386, 2448, 3082]],
[4055, 8, '相似文本编号：', [1965, 3411, 3638, 4055, 827, 924, 3008, 3045, 2607]],
```

表 3 留言相似排名前十数据表

Tab.3 Top 10 data table for similar messages

	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
88	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019-08-23 12:22:00	投诉伊景园.滨河苑开发商捆绑销售车位！A市武广新城片区下的伊景园.滨河苑是广铁集团铁路职工的...	0	0
318	196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019-08-07 19:52:14	\n\t\t\t\t\t\n\t\t\t\t\tA市伊景园.滨河苑现强制要求购房者捆绑购买1...	0	0
702	205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2019-08-14 09:28:31	广铁集团强制要求职工购买伊景园滨河苑楼盘时捆绑购买12万一个的车位，不买车位，不能购买房子！...	0	1
			坚决反对伊景园滨河苑强制捆绑销售	2019-08-03	我坚决反对伊景园滨河苑捆绑销售车位！原本广铁集团与市政府和		

提取文本用于计算热度的数据：

表 4 赞成数

Tab.4 Number of favor

```
: shufan
: [1, 3, 5, 19, 20, 20, 20, 21, 21, 21]
```

表 5 反对数

Tab.5 Number of objections

```
: shufan
: [1, 3, 5, 19, 20, 20, 20, 21, 21, 21]
```

表 6 排名前十的时间详表

Tab.6 Time table of the top ten

```
: jian
: [DatetimeIndex(['2019-08-23 12:22:00', '2019-08-07 19:52:14',
                  '2019-08-14 09:28:31', '2019-08-03 10:03:10',
                  '2019-08-23 12:16:03', '2019-08-01 22:42:21',
```

表 7 热度计算

Tab.7 Heat calculation

```
hot
[277.8388884092327, 217.37964361649662, 26.699353695028467, 271.6581180439992, 65.9625890488162, 62.842
9265122874497, 48.63088181024607, 263.96099582003154]
```

按热度排序提取前五，输出“热点问题明细表”部分结果：

表 8 热点问题明细表

Tab.8 Hot issues list

	问题ID	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	热度指数
4297	1	289950	A00044759	投诉A市伊景园滨河院捆绑销售车位	2019-07-07 07:28:06	\n\t\t\t\t\t\n\t\t\t\t\t\t提问A市政府就广州局集团公司与A市政府及A市...	0	0	277.838888
4038	1	283879	A00044759	A市伊景园滨河苑项目捆绑销售车位	2019-07-18 20:27:40	\n\t\t\t\t\t\n\t\t\t\t\t\t关于铁广集团铁路职工定向商品房伊景园滨河苑项...	0	0	277.838888
1482	1	223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019-07-23 17:06:03	\n\t\t\t\t\t\n\t\t\t\t\t\t关于铁广集团铁路职工定向商品房伊景园滨河苑项...	0	0	277.838888

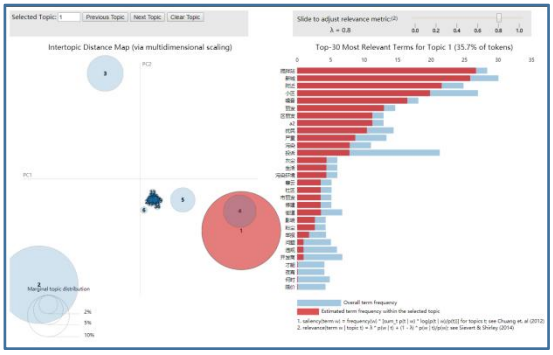
输出热点问题表部分结果：

表 9 热点问题表

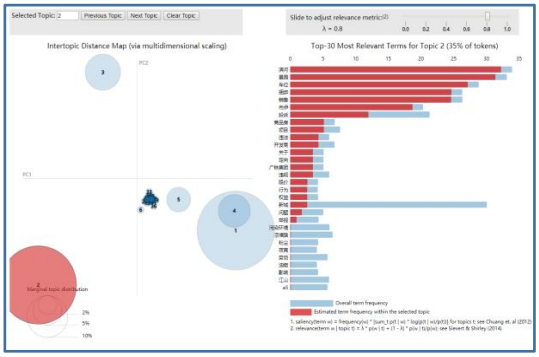
Tab.9 Hot issues table

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	1	277.838888	2019-07-07至2019-09-01	A市伊景园滨河苑	商品房销售项目捆绑销售车位
1	2	2	256.852107	2019-07-21至2019-09-25	A5区劳动东路魅力之城小区	小区夜宵摊油烟对附近造成严重污染
2	3	3	255.962178	2019-03-19至2019-06-13	A市江山帝景	新房有严重安全隐患
3	4	4	200.911489	2019-11-02至2020-01-26	A市A2区丽发新城小区	附近建搅拌站噪音扰民，灰尘严重污染生活环境
4	5	5	86.984540	2019-02-14至2019-09-02	A7县星沙四区凉塘路的旧城	旧城改造何时进行

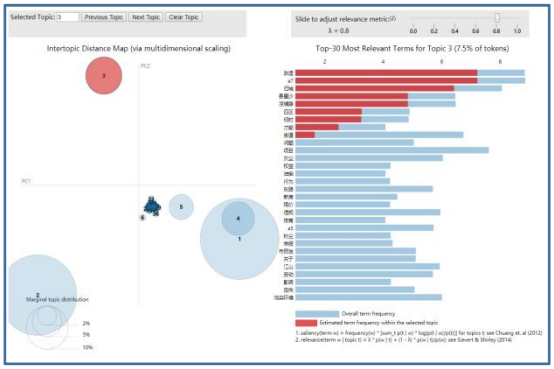
主题模型的主题数确定和可视化：



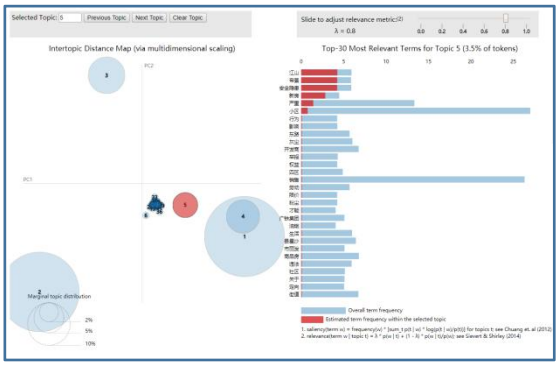
（1）



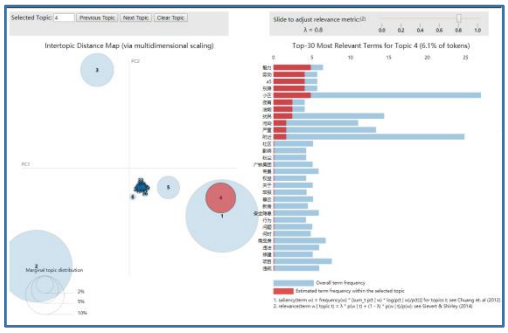
（2）



（3）



（4）



（5）

图 4 主题数的确定和可视化

Fig.4 Determination and visualization of the number of topics

把鼠标依次悬停在 1、2、3、4、5 号上可知，当前位置主题依次为：

附近建搅拌站噪音扰民；商品房销售项目捆绑销售车位；旧城改造何时进行；小区夜宵摊油烟对附近造成严重污染；新房有严重安全隐患。

2.3 答复意见评价

网络信息层出不穷，对于复杂繁琐的信息各式各样，要想从中通过人工进行信息的评价，将会耗费大量的人工资源而拥有极低的效率。根据此题针对答复意见，从答复的相关性、完整性、可解释性等角度定对答复意见的质量给出一套评价方案。

2.3.1 结构图

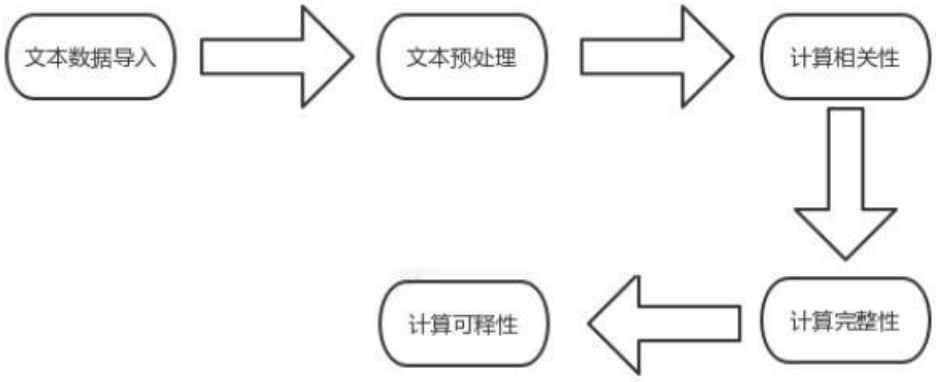


图5 问题三的研究结构图

Fig.5 Question 3 research structure

2.3.2 实现方法

第一步：文本预处理

1、 jieba 分词处理。

2、 停用词处理。

第二步：相关性计算

1、安装 Excel 网络函数库。

2、直接使用 Excel 内置函数公式=GetMatchingDegree（‘留言详情列（E）’，‘答复意见（F）’）直接得出相似度数据，取数据大于 0.5 为答复意见与留言详情相关。

第三步：完整性计算

1、 利用正则表达式剔除日期，防止分句时的干扰。

2、 对导入的文本进行分句处理，按照断句的标点符号对句子进行切分。

3、 人为筛选出句频较多的四种类型句子（问候语、答复语、收悉语、感谢语）。

4、 把分句后的每一个小句独立形成个体句子，并把全部聚集一个列表，并利用 jieba 分词和停用词库对句子进行处理，去除词频数小于 1 的词。

5、 调用 TF-IDF 模型对分句语料库建模，从中获取测试文本中每个词 TF-IDF 值。既可用四个文字样本结构测试集分别设置对应的阈值来求出相似的句子，形成模板集。

6、 调用 dictionary 方法获取词袋(bag-of-words)，通过 token2id 得到特征数，调用 Gensim 中封装的 Doc2Bow 方法，用来把测试文档和原料文档转化为二元向量。

7、 调用 TF-IDF 模型对分句语料库建模，从中获取测试文档中每个词 TF-IDF 值。既可用四个文字样本结构测试集分别设置对应的阈值来求出相似的句子，形成模板集。

8、 设置五个类型各占 2 分，满分 10 分用来判断句子的完整性，利用 for 循环来判断每一个内容分句结果是否存在与模板集，进行相应的加分，利用正则表达式去检测日期是否存在。

第四步：可解释性计算

1、利用文献和法律的特殊格式进行正则表达式[8]的匹配，如匹配到则返回可解释，否则返回不可解释。

2.3.3 结果说明

筛选高频句部分结果如下图所示：

```
pd.Series(_flatten(fenju_list)).value_counts()[:100]
```

```
现将有关情况回复如下：          645
现回复如下：                    555
您的留言已收悉。                554
网友：                          520
您好！                          452
目前，                          262
来信收悉。                      240
感谢您对我们工作的支持、理解与监督！    221
同时，                          192
    感谢您对我们工作的支持、理解与监督！    145
```

图6 筛选高频句结果

Fig.6 Filter high-frequency sentence results

相关性计算部分结果如下表所示：

表 10 相关性计算结果

Tab.10 Correlation calculation results

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	E与F相似度计算
2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	业公司却以交20万保证金，不	管理费，在业主大会结束	2019/5/10 14:56:53	2.352422907
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好？	2019/4/24 16:03:40	面的生意带来很大影响，里	换填，且换填后还有三趟	2019/5/9 9:49:10	1.009836066
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	同时更是加大了教师的工作压	聘任教职工要依法签订劳动	2019/5/9 9:49:14	1.282913165
2557	A000110735	A市买公寓能享受人才新政购房补贴吗	2019/4/24 15:07:30	落户A市，想买套公寓，请问	岁以下（含），首次购房	2019/5/9 9:49:42	0.396774194
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	“马坡岭小学”，原“马坡岭	坡岭”的问题。公交站点的	2019/5/9 9:51:30	0.726708075
2759	A00077538	A3区含浦镇马路卫生很差	2019/4/8 8:37	每把泥巴冲到石边，越是上下	中没有说明卫生较差的具	2019/5/9 10:02:08	0.556034483
2849	A000100804	A3区教师村小区盼望早日安装电梯	2019/3/29 11:53:23	为老社区惠民装电梯的规范	人民政府办公室下发了《关	2019/5/9 10:18:58	1.325
3681	UU00812	映A5区东澜湾社区居民的集体民生诉	2018/12/31 22:21:59	好远，天寒地冻的跑好远，非	备及设施设备采购等工作	2019/1/29 10:53:00	1.431089744
3683	UU008792	A市美麓阳光住宅楼无故停工以及质量	2018/12/31 9:55:00	也没得到相关准确开工信息。	分户检查后，西地省楚江嘉	2019/1/16 15:29:43	1.255445545
3684	UU008687	新城和顺路洋湖壹号小区路段公共绿	2018/12/31 9:45:59	立交桥等地方做立体绿化，取	规划要求完成了建设，其	2019/1/16 15:31:05	1.518248175
3685	UU0082204	反映A2区大托街道大托新村违建问题	2018/12/30 22:30:30	规划局审批通过《温室养殖	笔耕地征收补偿款给原大托	2019/3/11 16:06:33	0.971370143
3692	UU008829	A5区鄱阳村D区安置房人防工程的咨询	2018/12/29 23:27:51	区安置房地下室近两万平方米	人防发[2014]7号文件要求	2019/1/29 10:52:01	0.468384075
3700	UU00877	城小区段请求修建一座人行天桥或者	2018/12/29 11:55:34	锋，大量从小区开车出去的业	进行具体选址，招标（邀	2019/1/14 14:34:58	0.667546174
3704	UU0081480	举报A市芒果金融平台涉嫌诈骗	2018/12/28 17:18:45	省相关政府部门的大力支持	警情，已由银盆岭派出所	2019/1/3 14:03:07	0.479871176
3713	UU0081227	建议增开A市261路公交车	2018/12/28 7:53:25	小时以上！天寒地冻，其他公	驾驶员工作时间长，劳动	2019/1/14 14:33:17	0.871428571
3720	UU008444	开铺路与披塘路交叉路口通行安全问	2018/12/27 15:18:07	址：https://baidu.com/。	塘路路口两端各拆除20米中	2019/3/6 10:26:14	2.99078341
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	便以各种理由拒绝退货，并	提供的信息进行投诉信息的	2019/1/3 14:02:47	0.52398524
3733	UU008706	建议在A市梅溪湖开办一个图书馆	2018/12/26 16:51:40	称。建议在艺术中心先期借	梅溪湖二期金菊路与雪松路	2019/1/14 14:32:40	0.588652482
3747	UU008201	治理A3区中海国际社区一期旁边工地	2018/12/25 19:35:12	很早就施工，严重影响居民	单位由于需要夜间连续作	2019/1/8 16:19:16	0.416666667
3755	UU0081681	市社保卡、医保卡、居民健康卡尽快	2018/12/25 16:23:27	希望可以尽快合一，让社保	机构，需三方或三方以上	2019/1/4 15:48:23	0.443609023
3756	UU0081681	市满楚一卡通尽快支持手机nfc虚拟公	2018/12/25 16:19:49	华为、苹果等手机都无法开通	时间请关注潇洒支付公司	2019/1/4 15:49:46	0.62755102
3760	UU0081500	盛镇对泉水村塘下组土地征收存在的	2018/12/25 14:40:13	本农田。根据《土地管理法》	订了土地补偿协议，并按	2019/1/8 16:18:00	2.39546131
3762	UU0081057	5区交警大队纠正电子交通警察的错误	2018/12/25 13:56:31	直行车辆和行人通行，此路口	》第三十八条第一款第三	2019/1/16 15:22:16	1.00887574
3777	UU008162	轨道8号线北段在楚江北路上设江湾路	2018/12/23 21:47:34	事故频发。如果8路线设立站	您好，非常感谢您对于	2019/1/29 10:50:31	1.255131965
3788	UU0081604	询A市商业住房贷款转公积金贷款问题	2018/12/21 11:01:00	金，是否能在A市办理商业住	不支持非本中心的缴存人以	2019/1/3 14:00:47	0.28440367
3791	UU008694	东六线（劳动东路-机场高架）段目前	2018/12/20 17:28:09	在到A市国际会展中心非常不	已完成约800米路基，其余	2019/1/4 15:47:36	0.606694561
3797	UU008765	咨询A3区西湖街道茶场村公路规划问题	2018/12/20 11:16:07	政府修A3区山景区西大门，拆	投资计划调整，该项目已	2019/1/3 13:59:33	1.086466165
3838	UU0082119	反映A3区新江洋湖集体资产的有关问题	2018/12/15 15:17:53	是一个多亿好远，这笔大资金	举办的西地省洋兴置业公	2019/1/4 15:44:31	1.265873016
3848	UU008233	质疑A市佳兆业云项小区筹建的干洗店	2018/12/14 14:29:25	店就是这样操作的。梅溪湖的	名称为A市A3区那么好干洗	2018/12/29 15:05:11	0.778195489

可释性部分结果如下表所示：

表 11 可释性结果

Tab.11 Releasable result

	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	完整性分数	是否可解释
0	2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	10	可解释
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	6	不可解释
2	2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	6	不可解释
3	2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	10	可解释
4	2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30	6	不可解释
5	2759	A00077538	A3区含浦镇马路卫生很差	2019-04-08 08:37:20	\n\t\t\t\t\t\n\t\t\t\t\t欢迎领导来A市泥泞不堪的小含浦镇滚泥巴，这个...	网友“A00077538”：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、...	2019/5/9 10:02:08	6	不可解释

三、全文总结与展望

3.1 全文总结

为了提升政府的管理水平和施政效率，本文基于自然语言处理技术的相关理论和实验，分别解决了关于群众留言分类、热点问题挖掘以及答复意见评价三个问题，让政府的管理与政务的处理“更智能”、“更高效率”、“更低成本”、“更直观”地实现。

针对问题一，我们成功撰写了对于群众留言内容进行一级标签分类的模型，最终以“清晰简明”、“可视化”的图表进行了有效输出，并通过了 F-Score 对本次分类模型的检验。通过最终结果，可直观了解到各类问题涉及数量及各问题归属类别，可使得问题更快速地交至政府对应部门进行处理，提高政务处理效率。

针对问题二，我们使用逆向解题思路，首先实现了对“热点问题留言明细表.xls”的精准输出，再由该表成功进行了对“热点问题表.xls”的输出，在此过程中我们自主编撰了相似度二层计算模型，成功对多文本进行相似分类精确处理。通过最终结果，可智能明确各时间段最热点问题的精确总结，使得政府部门

更高效率明确问题轻重缓急，安排问题解决先后顺序，有针对性地处理问题，及时高效地对市民所提出的问题进行准确有效的施政，提升服务效率。

针对问题三，结合问题一与问题二，我们快速确定了相关性、完整性及可释性的判断标准，最终成功对‘答复意见’这三个性质进行了“智能”、“合理”、“准确”的判断。通过最终结果，通过数字化更直观地感受到政府部门答复问题的相关性、完整性及可释性，有助于提升政府的管理水平，促进政府有关部门自我完善，对政府有关部门处理政务能力进行“更智能”、“更高效”、“更直观”的监督。

综上三个问题模型都分别以数据预处理、关键词匹配以及精准匹配作为主要的三个阶段，在对赛题研究的基础上，我们根据研究思路撰写本论文，通过实验验证了本模型的可行性，成功实现本赛题设立的目标。

3.2 后续展望

在接下来的研究中，我们将主要攻克在实验中遇到优化的难题。其中利用逻辑回归模型方法对留言进行分类，从总体上缺乏稳定性，在多分类的问题上可以尝试结合深度学习浅层神经网络对多种模型的规划调优。提高题二中相似度二层计算的效率，加入浅层语义的挖掘使出现的交集情况减少，使问题的初步归类效果更加准确。在题三中我们将加强对三种性能关系网，使结果输出更直观清晰，期待日后能把模型变成一种常态化的工具，可以运用到生活的各个方面。

参考文献

[1] 张建光.朱建明.尚进 国内外智慧政府研究现状与发展趋势综述[D].电子政务,2015 年第 8 期.
[2]《Python 自然语言处理》,Steven Bird,人民邮电出版社
[3]《2750 个通用停用词表整理，免费下载》https://blog.csdn.net/weixin_35757704/article/details/91948847
[4] http://sklearn.apachecn.org/cn/0.19.0/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[5] 艾楚涵.姜迪.五建德 基于主题模型和文本相似度计算的专利推荐研究[D].昆明理工大学知识产权发展研究院.2020.
[6] 《排行榜热度公式计算》https://blog.csdn.net/liushuwei0224/article/details/39895969
[7] 牛力强 基于神经网络的文本向量表示与建模研究[D].南京大学.2020.
[8] Purva Goyal, Dr. Akash Saxena. Design and Implementation of Network Data Reptile. 2014, 1(4)
[9] 王英豪 网络评论的情感分析方法研究[D].重庆邮电大学.2020.
[10]张辛 基于 TFIDF 算法的全面从严治党重要论述关键词共现分析[D].江苏建筑职业技术学院纪委办公室.2019.
[11]《基于 python 的智能文本分析》,Benjamin Bengfort,Rebecca Bilbro,Tony Ojeda,中国电力出版社
[12]Humberto Pérez-Espinosa,Juan Martínez-Miranda,Ismael Espinosa-Curiel,Josefina Rodríguez-Jacobo,Luis Villaseñor-Pineda,Himer Avila-George. IESC-Child: An Interactive Emotional Children’s Speech Corpus[J]. Computer Speech & Language,2020,59.
[13]Shahbaz Hassan Wasti,Muhammad Jawad Hussain,Guangjian Huang,Aftab Akram,Yuncheng Jiang,Yong Tang. Assessing semantic similarity between concepts: A weighted‐feature‐based approach[J]. Concurrency and Computation: Practice and Experience,2020,32(7).