

“智慧政务”中的文本挖掘应用

摘要

本文将基于数据挖掘技术对附件中的文本数据进行内在信息的挖掘与分析。在本次数据挖掘过程中，我们首先对获取到的文本数据利用 R 以及 Excel 工具进行数据预处理、分词以及停用词过滤操作，实现了对数据的优化并提升了其可建模度。

针对问题一，我们首先对附件 2 中给出的数据进行了预处理，分词以及去停用词，根据得到的数据绘制了词云图，对数据特点有一个基本的认识。接着，我们使用数据挖掘程序，利用 LDA 主题模型的思想，结合统计学的角度实现文本主题模型的构建，并通过 F-Score 对模型进行了评价。

针对问题二，首先，我们使用 R 软件对附件 3 的留言主题进行分词处理和词频汇总，然后根据词频结果，绘制了词频图，利用词频图中的一些高频词汇并结合相关的软件进行查找，最终绘制了热点问题表以及热点问题留言明细表。

针对问题三，分别对附件 4 中的留言主题和答复详情进行分词处理，使用余弦相似度对答复的相关性，完整性，可解释性进行评价，其基本过程为：用 jiebaR 分词，再向量化（构建 TermDocumentMatrix 矩阵），再用 TF-IDF 赋权，最后计算余弦相似度。

关键词： LDA 主题模型 jiebaR 分词 F-Score 余弦相似度

Abstract

This article will mine and analyze the intrinsic information of the text data in the attachment based on data mining technology. In this data mining process, we first used R and Excel tools for data preprocessing, word segmentation and stopword filtering operations on the obtained comment data, which optimized the comment data and improved its modelability.

Regarding question one, we first preprocessed the data given in Annex 2: word segmentation, stop words, and a word cloud map based on the obtained data, with a basic understanding of the characteristics of the data. Next, we used a data mining program, using the idea of the LDA topic model, combined with a statistical perspective to achieve the construction of the review topic model, and evaluated the model through F-Score.

Regarding question two, first, we use R software to perform word segmentation processing and word frequency summary on the message topic of Annex 3, and then draw a word frequency graph based on the word frequency result, use some high-frequency words in the word frequency graph and combine with related software to find A list of hot issues and a list of hot issues can be obtained.

Regarding question three, the word subject and the details of the reply in Annex 4 are separately segmented. The cosine similarity is used to evaluate the relevance, completeness and interpretability of the reply. The basic process is: use jiebaR to segment the word and then vectorize (Build TermDocumentMatrix matrix), and then use TF-IDF weighting, and finally calculate the cosine similarity.

Key words: LDA topic model jiebaR segmentation F-Score cosine similarity

目录

1. 研究背景	4
2. 挖掘目标	4
3. 分析过程与方法	5
3.1 总体分析思路	5
3.2 具体步骤	6
3.2.1 文本预处理	6
3.2.2 基于 LDA 模型的主题分析	8
3.2.3 热点问题挖掘	12
3.2.4. 答复意见评价	13
4. 总结与展望	17
参考文献	18

1. 研究背景

自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融合了语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系，但又有重要的区别。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分。

自然语言处理产生发展至今可以大致上划分为两个阶段：第一个阶段是上世纪 50 年代到 70 年代基于规则的处理方式；第二个阶段是 70 年代至今基于数学模型和统计的处理方法。在第一个阶段的二十多年里，自然语言处理领域并未取得实质性的进展，而在第二个阶段的三十多年来，该领域逐步取得了突破性的进展，并在很多产品中得到了广泛的应用[1]。

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。

2. 挖掘目标

本篇论文针对附件给出的来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见数据，对文本进行基本的预处理、分词以及停用词过滤后，首先建立 LDA 主题模型，然后对附件 2 给出的数据，建立关于留言内容的一级标签，并使用 F-Score 对分类方法进行评估；对附件 3 的留言主题进行分词处理和词频汇总，然后根据词频结果，绘制了词频图，利用词频图中的一些高频词汇并结合相关的软件进行查找，可以得到热点问题表以及热点问题明

细表；对附件 4 中的留言主题和答复详情进行分词处理，使用余弦相似度对答复的相关性，完整性，可解释性进行评价。

3. 分析过程与方法

3.1 总体分析思路

首先用一个流程图对本文的分析过程做一个说明：

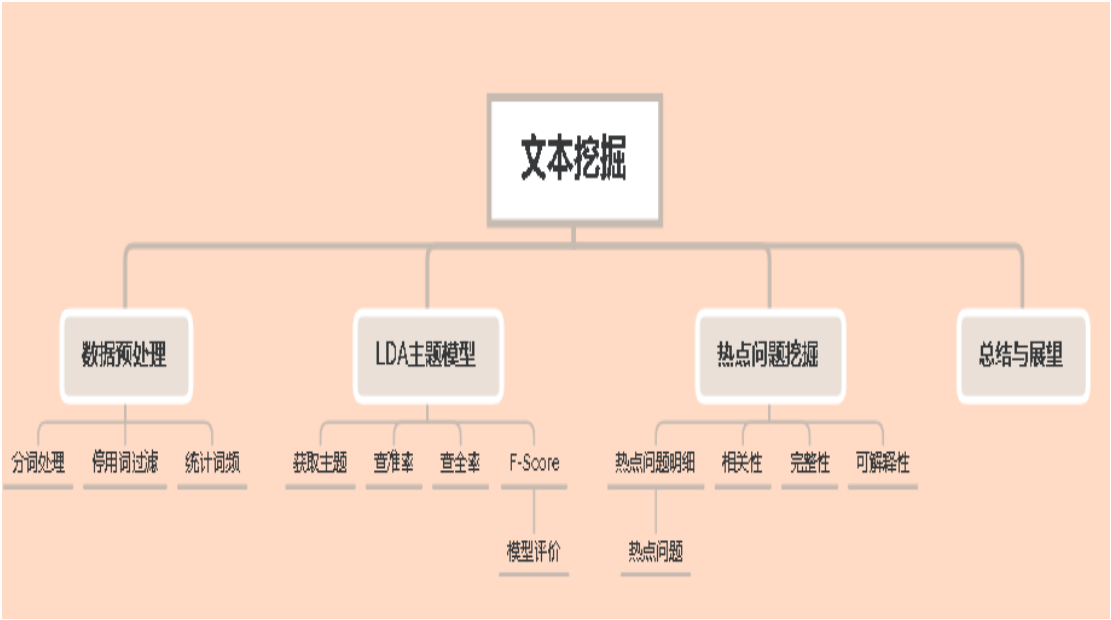


图 1 本文的分析过程流程图

本文的分析大致可以分为以下几步：

1. 首先利用 jiebaR 分词对附件 2 的文本进行分词处理，然后进行去停词。
2. 基于预处理后的文本数据建立 LDA 主题模型，采用 Gibbs 算法得到七大主题，计算 F-得分来对模型进行评价。
3. 针对附件 3 我们使用 R 软件对留言主题进行分词处理和词频汇总，然后根据词频结果，绘制了词频图，利用词频图中的一些高频词汇并结合相关的软件进行查找，最终绘制了热点问题表以及热点问题明细表。
4. 对附件 4 中的留言主题和答复详情进行分词处理，使用余弦相似度对答复的相关性，完整性，可解释性进行评价，其基本过程为：用 jiebaR 分词，再向量化，然后再用 TF-IDF 赋权，最后计算余弦相似度。

3.2 具体步骤

3.2.1 文本预处理

因为对于文本类型的数据来说，如果不首先对数据进行预处理，会导致后面的模型建立与求解变得困难，进而导致模型评价的得分偏低。我们运用 R 对文本进行预处理的步骤包括以下几个方面：分词，去停用词，统计词频。

在中文中，只有字、句和段落能够通过明显的分界符进行简单的划界，而对于“词”和“词组”来说，它们的边界模糊，没有一个形式上的分界符。因此，进行中文文本挖掘时，首先应对文本分词，即将连续的字序列按照一定的规范重新组合成词序列的过程。分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择[2]。

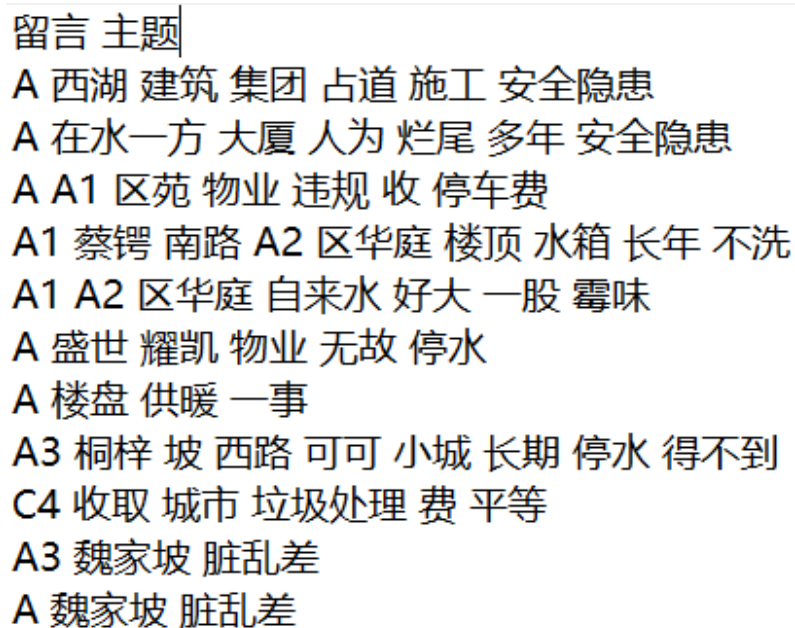
本文利用 R 里面的中文分词包 **jiebaR** 对附件 2 中的留言主题进行粗略的分词处理，然后对词频为前 100 的词汇做词云图如下图 2 所示：



图 2 词云图

为了建立主题模型，我们需要对文本进行更加精确的分词处理，分词结果如

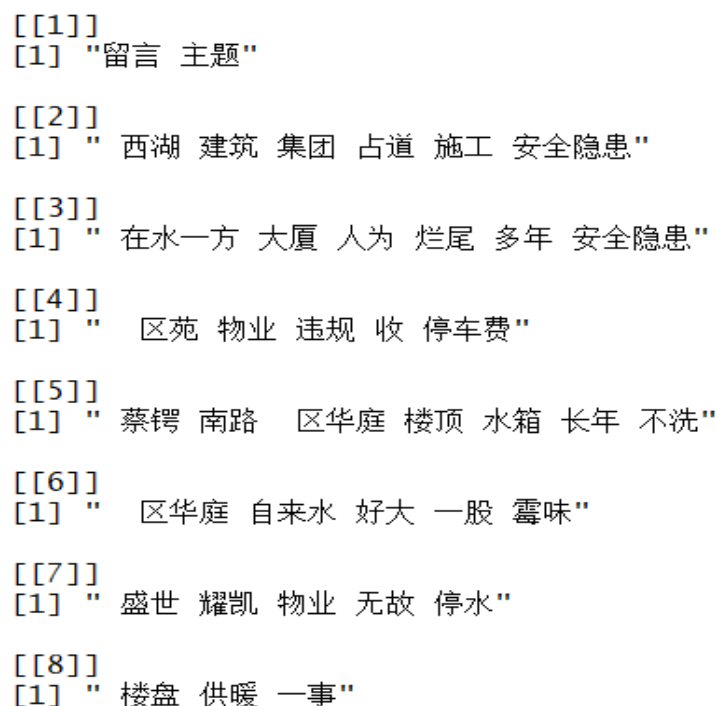
下图 3 所示：



```
留言 主题
A 西湖 建筑 集团 占道 施工 安全隐患
A 在水一方 大厦 人为 烂尾 多年 安全隐患
A A1 区苑 物业 违规 收 停车费
A1 蔡锷 南路 A2 区华庭 楼顶 水箱 长年 不洗
A1 A2 区华庭 自来水 好大 一股 霉味
A 盛世 耀凯 物业 无故 停水
A 楼盘 供暖 一事
A3 桐梓 坡 西路 可可 小城 长期 停水 得不到
C4 收取 城市 垃圾处理 费 平等
A3 魏家坡 脏乱差
A 魏家坡 脏乱差
```

图 3 分词结果图

可以看出 jiebaR 分词较为完整，但是一些无关紧要的词对我们后面的建立模型有影响，所以我们需对分词后的文本进行去停词处理效果如下图 4 所示：



```
[[1]]
[1] "留言 主题"

[[2]]
[1] " 西湖 建筑 集团 占道 施工 安全隐患"

[[3]]
[1] " 在水一方 大厦 人为 烂尾 多年 安全隐患"

[[4]]
[1] " 区苑 物业 违规 收 停车费"

[[5]]
[1] " 蔡锷 南路 区华庭 楼顶 水箱 长年 不洗"

[[6]]
[1] " 区华庭 自来水 好大 一股 霉味"

[[7]]
[1] " 盛世 耀凯 物业 无故 停水"

[[8]]
[1] " 楼盘 供暖 一事"
```

图 4 去停词图

在完成去停词这一步骤之后，进行词频统计，得到的排名前 19 的词频表如

下表 1 所示：

表 1 排名前 19 的词

Var	镇	公 司	教 师	医 院	社 保	违 规	污 染	职 工	村	电 梯	学 校	收 费	退 休	人 员	居 民	有 限 公 司	工 资	中 学	办 理
Freq	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	6	3	2	0	9	9	9	8	7	7	7	6	6	6	6	6	5	5
	4	2	8	9	5	7	5	2	1	7	4	1	9	8	8	6	6	7	7

下面我们将通过分词和去停词处理后的文本进行主题建模，并对文本分类结果进行评价。

3.2.2 基于 LDA 模型的主题分析

主题模型是文本挖掘的重要工具，近年来在工业界和学术界都获得了非常多的关注。在文本挖掘领域，大量的数据都是非结构化的，很难从信息中直接获取相关和期望的信息，一种文本挖掘的方法：主题模型能够识别在文档里的主题，并且挖掘语料里隐藏信息，并且在主题聚合、从非结构化文本中提取信息、特征选择等场景有广泛的用途。

LDA(Latent Dirichlet Allocation) 是其中最具代表性的模型，它是一种无监督的生成式主题概率模型。

3.2.2.1 LDA 主题模型介绍

LDA 主题模型是一个生成概率模型，能够对语料库进行建模，达到对文档降维的效果。它的基本思想是文档可以用潜在主题的随机混合表示，而每个主题是在词语上的概率分布。LDA 的文档生成过程主要有下面三个步骤[3]：

- 1. 生成 $N \sim Poission(\xi)$;
- 2. 生成 $\theta \sim Dir(\alpha)$;
- 3. 对于 N 词中的每个词 w :
 - (a) 生成一个主题 $z \sim Multinomial(\theta)$,
 - (b) 以概率 $P(w|z, \beta)$ 生成词 w ，这个概率是主题 z 的条件概率。

LDA 主题模型做了一些基本假设：第一，Dirichlet 分布的维度 K (主题 z 的个数) 是已知且固定的；第二，词语的概率通过 $K \times V$ 的矩阵 β 参数化， $\beta_{ij} = p(w_j = 1|z_i = 1)$ ；第三，Poisson 分布没有严格要求需要按照现实语料的长度。

生成步骤中的第二步， K 维的 Dirichlet 随机变量 θ 拥有下面的概率强度函数：

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$

α 是 K 维向量，等式右边分子是 Gamma 函数。给定参数 α 和 β ，主题混合 θ ， N 个主题 z 的集合， N 个词的集合的联合概率分布公式：

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^N p(z_n|\theta) p(w_n|z_n, \beta)$$

集成 θ 和主题 z ，可以得到一篇文档的原始分布：

$$p(d|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^N \sum_{z_n} \prod_{n=1}^N p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta$$

对所有单篇文档分布求积，就能够得到整个语料库的概率：

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta|\alpha) \left(\prod_{n=1}^N \sum_{z_n} \prod_{n=1}^N p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta_d$$

LDA 主题模型如下图 5 所示：

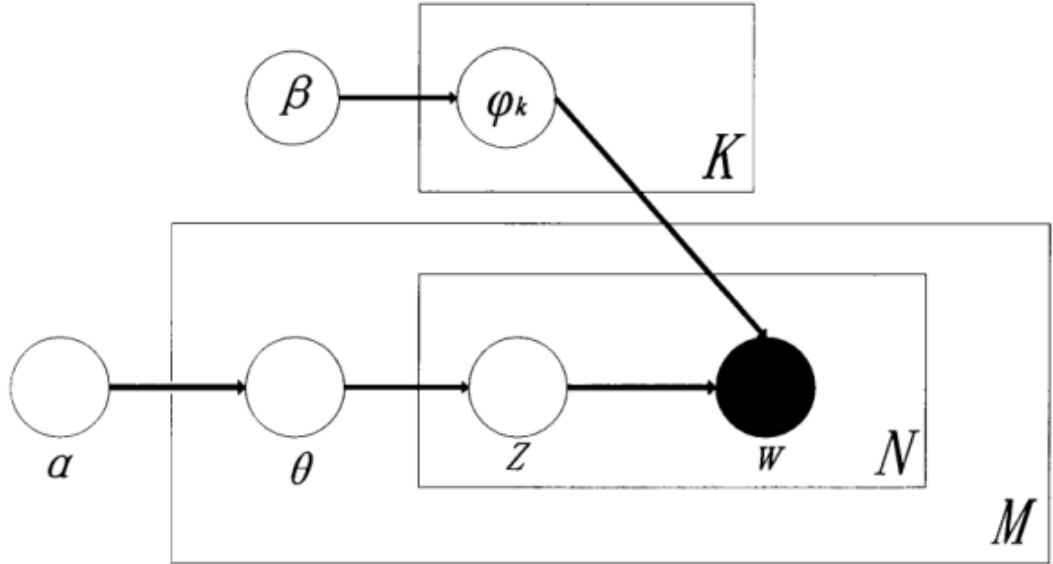


图 5 LDA 主题模型图

3.2.2.2 LDA 模型参数估计

我们可以采用 Gibbs 抽样公式来训练语料得到 LDA 模型，LDA 中 Gibbs 抽样公式如下：

p(z_i = k | z_{-i}, d) \propto \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^K (n_{m,-i}^{(k)} + \alpha_k)} \cdot \frac{n_{k,-i}^{(i)} + \beta_i}{\sum_{i=1}^V (n_{k,-i}^{(i)} + \beta_i)}

公式的右边可以看成是 p(topic|word) × p(word|topic)，表达的含义就是选择 doc → topic → word 的路径，因为 topic 一共有 K 个，所以 Gibbs 抽样就是从这 K 个路径中进行抽样。LDA 主题模型的学习过程，就是通过 Gibbs 抽样获取语料库中的 (z,w) 二元组的样本，根据采样得到的样本来估计模型中需要的参数[3]。

3.2.2.3 LDA 模型的实现

本文利用 R 建立 LDA 模型，根据附件 2 我们依次根据排名前 10 的高频词得出了 7 个主题（Topic），结果如下图 6 所示：

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
[1,]	"消费者"	"屠宰场"	"廉租房"	"公租房"	"公务员"	"幼儿园"	"公务员"
[2,]	"劳动者"	"公租房"	"农民工"	"工作者"	"加班费"	"采石场"	"管理员"
[3,]	"管理区"	"乡镇卫"	"回龙镇"	"湖壹号"	"湖小区"	"残疾人"	"采石场"
[4,]	"采石场"	"驾驶员"	"加班费"	"消费者"	"体育馆"	"居民区"	"第小区"
[5,]	"公共场"	"旅游区"	"保障性"	"职教师"	"幼儿园"	"运动员"	"公共卫"
[6,]	"玫瑰园"	"资格证"	"残疾人"	"服装费"	"防疫员"	"冲十字"	"共华镇"
[7,]	"农产品"	"外务工"	"工党员"	"劳动者"	"桐林坳"	"家坊镇"	"劳动合"
[8,]	"泰阳商"	"翡翠湾"	"公共场"	"若罔闻"	"违规施"	"考试院"	"没拿到"
[9,]	"文昌阁"	"农民工"	"检察院"	"市华尔兹"	"委治安"	"内槟榔企业"	"违规养"
[10,]	"保护性"	"土家织"	"欺骗性"	"卫生室"	"污染源"	"市镇弹"	"怎么样"

图 6 模型结果图

通过查找资料，我们将得到的 7 个主题进行分析，可以将 Topic1, Topic2, Topic3, Topic4, Topic5, Topic6, Topic7 分别对应于附件 2 中的一级标签：环境保护，商贸旅游，劳动和社会保障，卫生计生，教育文体，城乡建设，交通运输。为了计算 F1 得分，首先我们给出查准率和查全率的定义如下：

查准率：又称精度是衡量某一检索系统的信号噪声比的一种指标，即检出的相关文献量与检出的文献总量的百分比。普遍表示为：

$$\text{查准率} = (\text{检索出的相关信息量} / \text{检索出的信息总量}) \times 100\%$$

查全率：又称召回率，是衡量某一检索系统从文献集合中检出相关文献成功度的一项指标，即检出的相关文献量与检索系统中相关文献总量的百分比。普遍表示为：

$$\text{查全率} = (\text{检索出的相关信息量} / \text{系统中的相关信息总量}) \times 100\%$$

可表述为如下的矩阵：

表 2 矩阵表

模型预测 真实情况	True	False
True	TP	FN
False	FP	TN

则有查准率和查全率的计算式如下：

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$

再通过对得到的数据进一步的处理，处理的结果见附件，我们可以得到每一类的查准率和查全率如下表 3 所示：

表 3 计算结果表

	1	2	3	4	5	6	7
P	0.411	0.412	0.363	0.788	0.595	0.588	0.592
R	0.584	0.409	0.772	0.502	0.538	0.328	0.734

利用题目中给出的 $F - Score$ 得分的计算公式：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。将上述表格数据代入可得 F_1 约为 0.52。

3.2.3 热点问题挖掘

首先，我们使用 R 软件对附件 3 的留言主题进行分词处理和词频汇总，得到结果如下表 4 所示：

表 4 留言主题词频表

扰民	业主	噪音	施工	社区	建设
278	152	147	143	135	134
物业	违规	居民	安全	规划	幼儿园
120	119	107	96	92	86

我们根据词频结果，绘制了图 7

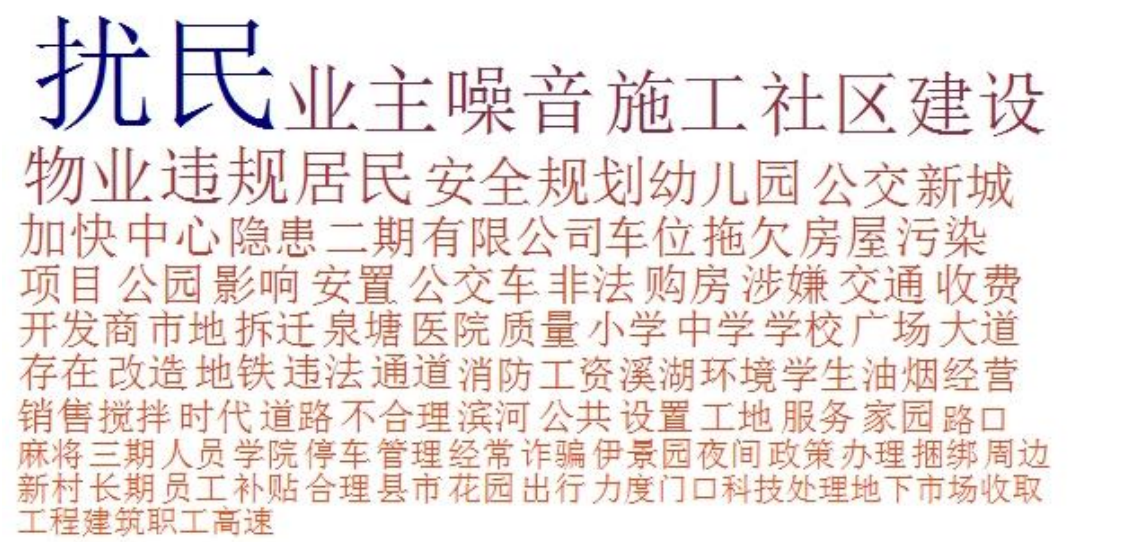


图 7 词频图

根据上图，我们可以清晰的看到，扰民、噪音、施工、物业以及违规这几个词的出现频率最高，所以我们根据这几个高频词汇查找了附件三所对应的留言主题，发现它们所对应的留言条数较多，且内容相似度较高。根据进一步的分析，我们确定了最热点的五个话题，以下是部分结果：

表 5 热点问题表

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	45	2019/11/19至2020/1/26	A市丽发新城附近搅拌站	搅拌站噪音、灰尘污染影响居民生活
2	2	35	2019/5/27至2019/9/1	A市伊景园滨河苑	捆绑销售车位
3	3	21	2019/8/18至2019/9/4	A市魅力之城小区	小区临街餐饮店油烟噪音扰民
4	4	10	2019/2/14至2019/12/28	A7县星沙旧城	旧城改造工期拖延质量不佳
5	5	9	2017/6/8至2019/11/22	A市经济学院学生	学校强制学生去定点企业实习

表 6 热点问题留言明细表

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
1	243692	A909201	城小区附近的搅拌站噪音严重	2019-11-15 11:23:21	的灰尘极大，都飘到	0	2
1	244335	A909135	丽发新城社区搅拌站灰尘，吵	2019/12/2 12:11:23	型搅拌站，水泥厂从	0	0
1	244512	A00094706	丽发新城小区粉尘大的孩子生	2019-12-05 20:57:50	，大人都无法正常呼	0	1
1	253040	A909202	区丽发新城附近建搅拌站噪	2019-12-04 12:10:21	里根本无法正常休息	0	0
1	255276	A909219	望领导“拯救”丽发新城小	2019-12-11 00:00:00	皮肤受尘土影响皮肤	0	0
1	258242	A909220	丽发新城社区搅拌站灰尘，吵	2019-12-02 12:23:11	严重影响附近居民	0	0
1	258378	A00084226	社区附近搅拌站修建严重影响	2019-11-23 00:00:00	境，还严重影响了手	0	0
1	259788	A909221	街道丽发新城社区搅拌厂危害	2019-12-07 00:00:00	染严重，危害附近小	0	0
1	260979	A909229	暮云街道丽发新城小区环境污	2019-12-04 14:21:32	还得了疾病住院，	0	0
1	264944	A0004260	新城附近修建搅拌厂噪音、	2019-11-02 14:23:11	修建搅拌厂，请问环	0	0
1	267050	A909227	污染的A2区丽发新城附近已扒	2019-11-02 10:18:00	居民不能正常休息，	0	0
1	268109	A909230	A2区丽发新城小区开发商违	2019-12-05 18:21:32	的噪音污染小区居民	0	0
1	268300	A909225	附近修建搅拌厂噪音污染导	2019-11-25 10:17:58	拌厂，请问是谁审批	0	0
1	272224	A909224	城小区噪音大粉尘大，求撤走	2020-01-09 19:46:10	粉尘太大无法呼吸，吵	0	1
1	272361	A909242	发新城小区旁建搅拌厂严重扰	2019-12-04 08:46:20	平常上班不在家我	0	1
1	273282	A909226	城附近修建搅拌厂烟尘滚滚	2019-12-25 10:17:59	政府，我们该怎么生	0	0
1	274004	A00026895	道丽发新城社区附近搅拌站噪	2019-12-21 10:11:09	居民区附近建搅拌	0	0
1	281348	A909219	望领导“拯救”丽发新城小区	2019-11-24 00:00:00	影响，更有多名业主	0	0

3.2.4. 答复意见评价

经过查询相关资料，我们获取了计算文本相似度的一些常用方法，如：余弦相似性、简单共有词、编辑距离、Jaccard 相似性系数等。

1. 相似度计算方法

（1）余弦相似度：余弦（余弦函数），三角函数的一种。在 Rt△ABC（直角三角形）中，∠C = 90°，角 A 的余弦是它的邻边比三角形的斜边，即 $\cos A = b/c$ ，也可写为 $\cos A = AC/AB$ 。余弦函数： $f(x) = \cos(x)(x \in R)$

$$\cos\theta = \frac{a^2 + b^2 - c^2}{2ab}$$

假设向量a、b的坐标分别为 (x_1, y_1) 、 (x_2, y_2) 。则：

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

(2) 简单共有词：通过计算两篇文档共有的词的总字符数除以最长文档字符数来评估他们的相似度。假设有 A、B 两句话，先取出这两句话的共同都有的词的字数然后看哪句话更长就除以哪句话的字数。

(3) 编辑距离：编辑距离 (Edit Distance)，又称 Levenshtein 距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。

(4) Jaccard 系数，又叫 Jaccard 相似性系数，用来比较样本集中的相似性和分散性的一个概率。Jaccard 系数等于样本集交集与样本集合集的比值，即 $J = |A \cap B| \div |A \cup B|$ 。

问题三要求，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，首先我们对这三个词进行解释：

相关性：答复意见的内容是否与问题相关，这里我们采用余弦相似度作为衡量答复相关性的指标。

完整性：答复内容是否包含对于所有留言问题的回答。

可解释性：答复是否满足某种规范，这里我们采用答复是否依据法律法规来进行衡量。

根据对附件 4 的留言主题和答复详情进行筛选，我们得到分别取出前 3 条绘制了表 7 和表 8：

根据对相关性、完整性以及可解释性的理解，我们对附件 4 的用户留言以及答复意见的分析，我们认为如果答复意见能够紧扣留言主题，且按照相关的法律法规来进行回答，那么是比较有解释性和可信度的。所以，根据这些标准，我们对附件 4 的数据进行了分析处理以及提取，获得了一些较好的例子，这些留言与答复之间，无论从相关性还是完整性以及可解释角度来说都是较好的示例。

表 7 示例表

留言 编号	留言主题	答复意见
2557	在 A 市买公寓能享受人才新政购房补贴吗？	现将相关情况回复如下：按照《A 市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在 A 市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄 35 周岁以下（含），首次购房后，可分别申请 6 万元、3 万元的购房补贴。“首次购房”是指在 A 市限购区域内首次购买商品住房（含住宅类公寓）。
3762	呼吁 A5 区交警大队纠正电子交通警察的错误执法	根据《中华人民共和国道路交通安全法实施条例》第三十八条第一款第三项“红灯亮时，禁止车辆通行”之规定，左转弯信号灯为红灯时，禁止车辆越过停止线行驶。
4173	投诉 A3 区 A 市艺苑培训学校乱收费	根据楚发改价费〔2017〕771 号《关于进一步规范我省中小学（幼儿园）教育收费管理有关事项的通知》文件第四条规定：“进一步推进民办学校收费管理市场化改革，必须坚持教育公益导向，有效引导民办学校科学合理制定收费标准，坚持推进市场形成机制，健全完善主要由市场决定民办学校收费标准的机制”、“民办学校制定的收费标准，需在招生简章和宣传广告明示，实际收费标准应与招生简章及宣传广告收费标准一致”，民办学校的收费标准是由学校根据市场需求和办学成本自主制定。

表 8 示例表

留言 编号	留言主题	答复意见
11924	呼吁 A 市将山 水洲城独特的 文化和资源打 造成精品	网友：您好！留言已收悉
11927	A 市轴承厂工 矿棚户区改造 项目问题的举 报	网友：您好！留言已收悉
11955	关于尽快开通 12 路公交线 路的情况反映	网友：您好！留言已收悉

根据表 8 我们可以看出，对于编号为 11924，11927,11955 的留言，有关部门对用户留言的回复的相关性，完整性和可解释性均没有满足，所以对于这一类答复，可以视为无效回复。

根据有关部门对用户留言的回复的相关性，完整性以及可解释性，我们提出了以下评价指标：

我们首先对相关性，完整性，可解释性这三个指标分别进行评分，每个指标的最高得分为 10 分，最低得分为 0 分。

考虑到要综合相关性，完整性，可解释性这三个角度对答复意见质量进行评价，我们定义了 T-得分，其定义如下：

$$T = \frac{1}{3} \times \text{相关性得分} + \frac{1}{3} \times \text{完整性得分} + \frac{1}{3} \times \text{可解释性得分}$$

最终根据 T-得分，我们对答复意见的质量做出评价。

表 9 答复意见质量评价等级表

T-得分	0-3	3-6	6-10
答复质量	差	中等	好

4. 总结与展望

本篇论文针对附件给出的来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见数据，对文本进行基本的预处理、分词以及停用词过滤后，得到 LDA 主题模型，根据指标对模型进行评价；绘制了热点问题明细表和热点问题留言明细表；给出了一套评价方案并对部门回复意见进行了评价。

但是从我们的分析结果当中也可以看出总体来讲效果还不是特别的好，比如我们基于 LDA 主题模型计算出来的 F-Score 仅为 0.52，可以看出得分不高，可能是我们的主题分类没有定好，所以我们在后期应该改善和优化我们的程序。对于问题三，我们也只是大致的提供了一些方案，经过尝试，并没有实现，所以在后期进一步的对中文文本数据的研究过程中可以继续深入探讨。

参考文献

- [1] 吴军. 数学之美[M]. 人民邮电出版社, 2012. 29-32.
- [2] 邓尧. 基于 R 语言的网络文本分析[D]. 华中科技大学, 2017.
- [3] 王军. 热门微博话题事件主题聚类分析[D]. 安徽大学, 2016.
- [4] 王春柳, 杨永辉, 邓霏, 赖辉源. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37(03):158-168.
- [5] 张金鹏. 基于语义的文本相似度算法研究及应用[D]. 重庆理工大学, 2014.
- [6] 康东. 中文文本挖掘基本理论与应用[D]. 苏州大学, 2014.
- [7] 刘三女牙, 彭晔, 刘智, 孙建文, 刘林, 郑年亨. 基于文本挖掘的学习分析应用研究[J]. 电化教育研究, 2016, 37(02):23-30.
- [8] 陈大力, 沈岩涛, 谢槟竹, 马颖异. 基于余弦相似度模型的最佳教练遴选算法[J]. 东北大学学报(自然科学版), 2014, 35(12):1697-1700.