第八届“泰迪杯”

数据挖掘挑战赛

C

题

基于“智慧政务”中的文本挖掘应用

摘要

本文根据附件给出的收集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。进行相关的群众留言分类，热点问题挖掘以及对答复意见的评价的文本挖掘。

针对群众留言分类，按照题目给出的分类体系，先基于 TD-IDF 权重法提取特征，再基于 k-means 方法进行聚类，再进行进一步分类。

针对热点问题挖掘，先构建 DBSCAN 模型聚类，根据聚类结果创建“热点问题留言明细表”。再基于 LDA 进行留言相关主题的提取，创建“热点问题表”。

针对答复意见的评论，进行词频统计以及文本相似度分析，进行相关的答复评价。

关键词：

TF-IDF 权重　　k-means 方法　　DBSCAN 聚类　　LDA 模型　　文本相似度分析

Text mining application based on "smart government"

Abstract:

In this paper, according to the attachment, we collect the records of the public political message from the public sources of the Internet, and the response opinions of the relevant departments to some of the public message. Carry on the relevant mass message classification, hot issues mining and the text mining of the evaluation of the reply.

According to the classification system given by the topic, we first extract features based on td-idf weight method, then cluster them based on K-means method, and then further classify them.

In order to mine hot issues, we first construct DBSCAN model clustering and create "hot issues message list" according to the clustering results. Then based on LDA, we extract the related topics of the message and create the "hot issues table".

According to the comments of the replies, we make word frequency statistics and text similarity analysis to evaluate the replies.

Key word:

TF-IDF weight k-means method DBSCAN clustering LDA model text similarity analysis

目录

摘要.....1
1.群众留言分类.....4
1.1 挖掘目标.....4
1.2 分析方法与过程.....4
1.3 分析.....5
1.4 结论.....5
2.热点问题挖掘.....5
2.1 挖掘目标.....5
2.2 分析流程及步骤.....6
2.3 结论.....7
3.对答复意见评价的文本挖掘.....8
3.1 进行数据读取及预处理.....8
3.2 进行词频的统计.....9
3.3 进行词云的绘制.....9
3.4 对“留言详情”和“答复意见”进行相似度分析.....10
3.5 小结.....11
参考文献.....11

1.群众留言分类

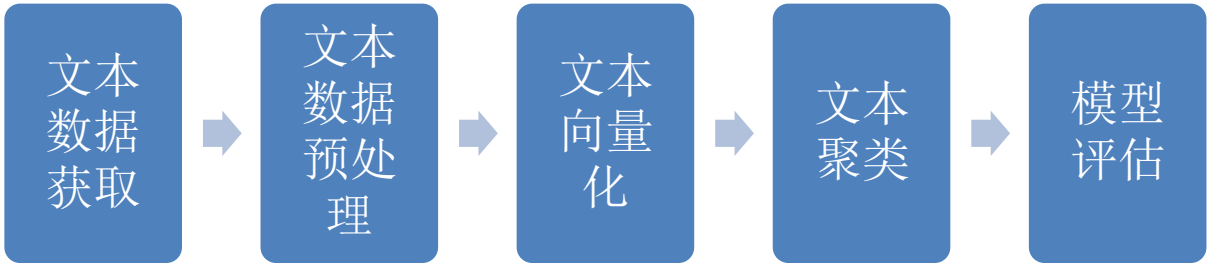
1.1 挖掘目标

首相将附件 3 中的文本数据进行去重去空、文本分词以及去除停用词等数据预处理，然后用 TFIDF 权值矩阵法提取 4000 个特征词，构成文本权值矩阵，由于 TFIDF 也有一定的缺点，比如高维度等等，因此，进一步用 PCA 降维方法进行空间词汇文本降维，再通过 k-means 聚类算法对分类标签划分。

1.2 分析方法与过程

1.2.1 总体流程

总体框架及思路



（1）文本数据获取：下载附件，读取。

（2）文本数据预处理：对附件 2 的留言详情内容进行基本的数据处理

（3）文本向量化：基于 TFIDF 权值矩阵法提取关键词，进一步用 PCA 降维方法进行数据降维。

（4）文本聚类：根据文本向量，把留言详情的分类问题用 k-means 的算法来解决

（5）模型评估：用$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_I}$对分类进行评价。

1.2.2 具体步骤

（1）文本数据获取，首先读取要处理的文本如表 1。

表 1 读取文本表

	留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
0	24	A00074011	A市西湖建筑集团占道施工有安全隐患	2020/1/6 12:09:38	\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，...	城乡建设
1	37	A000107866	A市在水一方大厦人为烂尾多年，安全隐患严重	2020/1/4 11:17:46	\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...	城乡建设
2	3742	A00013884	A3区杜鹃文苑小区外的非法汽车检测站要开业了！	2018/12/26 10:13:37	\n\t\t\t\t\t\n\t\t\t\t\tA市政府、市交警支队、市安监局、市环保局、A...	城乡建设
3	5315	A0009647	民工在A6区明发国际工地受伤，工地方拒绝支付医疗费	2018/7/17 7:13:58	\n\t\t\t\t\t\n\t\t\t\t\t胡书记，您好，感谢您百忙之中查看这份留言。我...	城乡建设
4	100026	A00047862	K8县丁字街的商户乱摆摊	2017/10/14 14:25:14	\n\t\t\t\t\tK8县丁字街的商户乱摆摊，前段时间丁字街的交通好了几天，最近那些在丁...	城乡建设

（2）文本数据预处理：

1. 对文本的留言详情内容进行去除空白留言以及重复的留言

2. 中文分词结果如表 2：

表 2 分词结果表

```
0    [A3, 区, 大道, 西行, 便, 道, ，, ，, 未管, 所, 路口, 至, 加油站, 路段...
1    [位于, 书院, 路, 主干道, 的, 在水一方, 大厦, 一楼, 至, 四楼, 人为, 拆...
2    [尊敬, 的, 领导, ：, ，, A1, 区苑, 小区, 位于, A1, 区, 火炬, 路, ，, ...
3    [A1, 区, A2, 区华庭, 小区, 高层, 为, 二次, 供水, ，, ，, 楼顶, 水箱,...
5    [我, 在, 2015, 年, 购买, 了, 盛世, 耀凯, 小区, 17, 栋, 3, 楼...
Name: 留言详情, dtype: object
```

3.停用词去除：

a.去除较为大众化的词，对文本影响不大的词。

b.文本中出现频率较高，但是对文本无影响的词。

（3）文本向量化，基于 TF-IDF 提取关键字

TF-IDF 是一种使人更容易接受的一种权重法，理解起来更容易。

第一步：计算词频

词频（TF）=某个词在文章中的出现次数

不同的文章总的字数不一样，但为了方便计算，一般情况下都有一个标准。

$$词频（TF）=\frac{某个词在文章中的出现次数}{文章的总次数}$$

第二步：计算逆文档频率

$$逆文档频率（IDF）=\log（\frac{语料库的文档总数}{包含该次的文档数+1}）$$

第三步：计算 TF-IDF

TF-IDF=词频（TF）×逆文档频率（IDF）

PCA 分析：

PCA 分析：将空间词汇文本权值矩阵降维。

文本聚类：k-means 聚类

算法如下：

1.假设有 k 个不同类，因此有 k 个初始中心；

2.在第 k 次迭代中，对任意一个样本，比较其到 k 个中心的距离，该样本归为离中心最近的一类中

3.再均值、方差等方法更换新的类的中心值；

4.经过多次迭代之后，k 个中心值不再变则结束，否则继续。

（4）模型评估:：用$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_I}$对分类进行评价。

P：查准率　R：查准率　P 与 R 相互制约

P-R 曲线：

①以二分类为例，按照模型输出大小预测，计算出查全率与查准率，最后连线。

②P-R 图一般为非光滑非单调曲线

F1 系数：

综合查准率与查全率：

$$F_1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{样例总数+TP-TN}$$

更一般的形式：

$$F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}$$（其中β为正数）

$\beta=1$：标准的 F1 系数

$\beta>1$：查全率有更大影响

$\beta<1$：查准率有更大影响

1.3 分析

针对留言详情，我们基于 TD-IDF 权重法提取特征，再基于 k-means 方法进行聚类，再往下进行了进一步分类。

1.4 结论

文本中，通过对留言分析数据挖掘，得知在处理网络问政平台的群众留言的分类类型，通过大数据，更加直接、清晰的看到群众的留言聚集问题，政府部门对群众留言问题更有针对性，办事效率也就自然而然提升起来。但是，由于技术问题，群众的留言不能准确的分类到相应的部门，中间存在误差。所以，要想更好的为群众服务，我们的技术提高。

2.热点问题挖掘

2.1 挖掘目标

对题目所给的留言信息将某一时间段内反映特定地点或特定人群问题的留言进行归类，根据合理的热度评价指标给出排名前五的热点问题和具体的留言信息。

2.2 分析流程及步骤

2.2.1 进行数据读取及预处理

（1）在 Python 语言中，使用 pandas 库将题目给出的 excel 文件进行导入并读取对应的留言详情列。

（2）利用 Python 语言中 jieba 库的精确模式对留言详情内容进行分词，去除“了”、“的”、“在”等停用词和相关标点符号，得到一个元素用空格链接的列表。如：

```
0    座落在 市 A3 区联 丰路 米兰 春天 G2 栋 320 一家 名叫 一米阳光 婚纱 艺术...
1    市 A6 区 道路 命名 规划 初步 成果 公示 文件 转化 正式 成果 希望 加快 路名 ...
2    系 春华 镇金鼎村 七里 组 村民 不知 相关 水泥路 到户 政策 自来水 到户 政策 政府...
3    靠近 黄兴路 步行街 城 南路 街道 古道 巷 一步 两 搭桥 小区 停车场 东面 围墙 外...
4    市 A3 区 中海 国际 社区 三期 四期 蓝天 璞 洲 幼儿园 旁边 块 空地 处于 三不...
```

（3）使用 sckikit-learn tfidf 计算词语权重[1]将文本进行量化，得到 tf-idf 的权值

2.2.2DBSCAN 文本聚类

基于密度的 DBSCAN 聚类算法[1]是一种无监督方法，能够有效地解决 k-means 等分区算法需要人为事先确定聚类个数，无法适用于具有任意形状的簇和内存占用资源较大以及层次划分需要确定停止分裂的条件的问题。

在参数 Eps 的确定上，作 sorted k-dist 图[2]，即对于留言中的每个留言点，分别计算他们与第 k 个最近邻的距离，然后将这些距离进行逆排序，并绘制他们的距离分布曲线。根据留言的文本数据得到如下图 1。

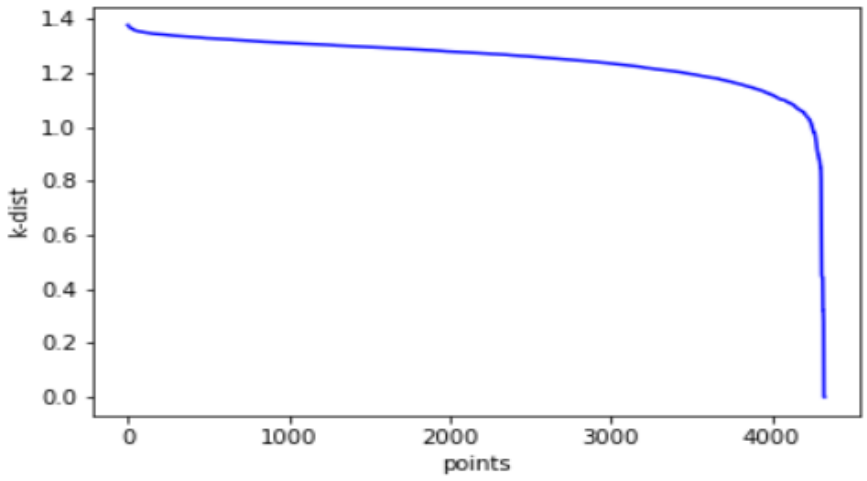


图 1 k-dist 图

根据做出的图可观察到曲线的分布状况，以此确定 Eps 参数为 1.1，MinPts 参数为 3。

（3）考虑题中所给的留言数据是否需要进行 pca 降维。

（4）将量化后的留言词语进行 DBSCAN 聚类。得到聚类后的留言情况，其中某类如图 2：

```
labels_original[0]
```

```
['桐梓 坡 589 号 白鹤 咀 停车场 由聚 美龙楚 新能源 公司 建 商学院 新能源 汽车 充电站 项目 告知 情况 商学院 宿舍 栋 外墙 安装 1000 千伏安 变压器 距离 商学院 宿舍 栋 不到 米 距离 A3 区 一小 不到 10 米 违反 住宅 建筑 电气设计 规范 4.2 变压器 外侧 住宅 外墙 间距 小于 20 米 中小学校 校园环境 管理 暂行规定 16 条 单位 校门口 200 米 设置 停车场 违背 备案 项目 承诺书 不建 居住 文化教育 功能 区域 充电站 违法 违规 项目 欺骗 政府 违背 承诺书 周边 4000 多名 小学生 大学生 居民 带来 电 辐射 污染 噪音 污染 消防 隐患 请求 拆除',
 '桐梓 坡 589 号 白鹤 咀 停车场 由聚 美龙楚 新能源 公司 建 商学院 新能源 汽车 充电站 项目 告知 情况 商学院 宿舍 栋 外墙 安装 1000 千伏安 变压器 距离 商学院 宿舍 栋 不到 米 距离 A3 区 一小 不到 10 米 违反 住宅 建筑 电气设计 规范 4.2 变压器 外侧 住宅 外墙 间距 小于 20 米 中小学校 校园环境 管理 暂行规定 16 条 单位 校门口 200 米 设置 停车场 违背 备案 项目 承诺书 不建 居住 文化教育 功能 区域 充电站 违法 违规 项目 欺骗 政府 违背 承诺书 周边 4000 多名 小学生 大学生 居民 带来 电 辐射 污染 噪音 污染 消防 隐患 请求 拆除',
 'A3 区 一小 紧临 地铁 号线 项目 施工 剩望 岳路 一条 断头 路 接送 学生 望岳 路上 接送 拥挤不堪 车流 人流 空空荡荡 1500 平方 收费 咀 鹤 停车场 停车场 一部分 租给 美 聚龙 楚 新能源 公司 建 商学院 新能源 汽车 充电站 项目 紧临 商学院 宿舍 距离 A3 区 一小 不到 10 米 营利性 停车场 交通 火灾 触电 噪音 隐患 白咀 鹤 停车场 市 城建 投 项目 白咀 鹤 停车场 商学院 新能源 汽车 充电站 项目 未 验收 运营 拨打 12345 市 城建 投 答复 自行 找美 聚龙 楚 公司 解决 市发 改局 答复 项目 验收 2020 年 下半年 平台 问政 西地省 24 A3 区委 办 答复 时过 半年 网上 答复 相关 单位 现场 查看 落实工作',
 '桐梓 坡 589 号 白鹤 咀 停车场 由聚 美龙楚 新能源 公司 建 商学院 新能源 汽车 充电站 项目 告知 情况 商学院 宿舍 栋 外墙 安装 1000 千伏安 变压器 距离 商学院 宿舍 栋 不到 米 距离 A3 区 一小 不到 10 米 违反 住宅 建筑 电气设计 规范 4.2 变压器 外侧 住宅 外墙 间距 小于 20 米 中小学校 校园环境 管理 暂行规定 16 条 单位 校门口 200 米 设置 停车场 违背 备案 项目 承诺书 不建 居住 文化教育 功能 区域 充电站 违法 违规 项目 欺骗 政府 违背 承诺书 周边 4000 多名 小学生 大学生 居民 带来 电 辐射 污染 噪音 污染 消防 隐患 请求 拆除',
 '领导 报告 A3 区 第一 小学 门口 违规 建设 咀 鹤 停车场 安全隐患 A3 区 一小 紧临 地铁 号线 项目 施工 剩望 岳路 一条 断头 路 接送 学生 望岳 路上 接送 拥挤不堪 车流 人流 空空荡荡 1500 平方 收费 咀 鹤 停车场 停车场 一部分 租给 美 聚龙 楚 新能源 公司 建 商学院 新能源 汽车 充电站 项目 紧临 商学院 宿舍 距离 A3 区 一小 不到 10 米 营利性 停车场 交通 火灾 触电 噪音 隐患 白咀 鹤 停车场 市 城建 投 项目 白咀 鹤 停车场 商学院 新能源 汽车 充电站 项目 办理 施工 许可 环评 商学院 新能源 汽车 充电站 项目 未 验收 运营 法律依据 住宅 建筑 电气设计 规范 4.2 变压器 外侧 住宅 外墙 间距 小于 20 米 中小学校 校园环境 管理 暂行规定 16 条 单位 校门口 200 米 设置 停车场 商学院 新能源 汽车 充电站 项目 违背 备案 项目 承诺书 不建 居住 文化教育 功能 区域 市政府 投资 建设 项目管理 办法 通知 第六条 基本建设 程序 办理 施工 许可 第十七条 含 充电 桩 项目 西地省 电动汽车 充电 基础设施 项目 验收 办法 第一条 验收 合格 报告 方可 投入使用 拨打 12345 市 城建 投 答复 自行 找美 聚龙 楚 公司 解决 市发 改局 答复 项目 验收 2020 年 下半年 请求 胡书记 解决 A3 区 第一 小学 门口 违规 建设 咀 鹤 停车场 安全隐患']
```

图 2 聚类留言情况图

2.2.3 热度指数计算

由于已将留言详情聚类，则无需再将每条留言之间进行相似度计算。则按照每类元素的多少进行热度指数排行。除去所得值为-1 的噪点类，取留言热度指数的前五类。

2.2.4 构建热点问题留言明细表

根据热度指数排名前五的相关信息构建 DateFrame 并将其写入 excel 保存。创建“热点问题留言明细表” ，部分数据表 3：

表 3 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	点赞数	反对数
0	1	188801	A909180	投诉滨河苑针对广铁职工购房的霸王规定	2019-08-01 00:00:00	尊敬的张市长，您好！我叫李建义，来自湖北仙桃，虽然已经在广铁集团A市公司工作了十几年，但是依...	0	0
1	1	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
2	1	195511	A909237	车位捆绑违规销售	2019-08-16 14:20:26	对于伊景园滨河苑商品房，A市广铁集团违规捆绑车位销售至今，买房必须买车位我们反映多次一直没有...	0	0
3	1	195528	A0006953	A3区车塘河路公园尚小区物业强制买车位	2019-04-17 10:47:26	\n\t\t\t\t\t\n\t\t\t\t\tA市A3区车塘河路公园尚小区物业强制买车位，...	0	0
4	1	195995	A909199	关于广铁集团铁路职工定向商品房伊景园滨河苑项目的问题	2019-08-10 18:15:16	尊敬的市政府领导，您好！我是广铁集团基层职工，我要反应的问题是关于广铁集团铁路职工定向商品房...	0	0

2.2.5 热点问题主题提取

（1）根据 pandas 库读入“热点问题留言明细表.xls”

（2）将留言详情和留言主题进行数据处理，将文本进行量化。

（3）构建 LDA 模型

① LDA 生成过程

首先，从每条留言详情的主题分布中抽取一个留言主题。根据被抽到的主题所对应的语句中抽取一个词语，重复上述的过程直至遍历每条留言中的每个词语[3]。

将留言详情中的每一条留言看做一个词语序列，每个词语出现的位置对于 LDA 算法是没有影响的。每一条留言语句对应不同的主题概率，其计算方法则是第某条留言的某个词语的概率=对应某个主题词的数目/该条留言中的所有词的总数。

②将所得到的留言词语向量进行 LDA 主题提取，得到相应的热点问题主题和留言问题详情。

2.2.6 时间范围

根据 dt.data 对留言时间列进行日期和具体时间时间的拆分，得到留言的日期后，找出日期的最大最小值得到热点问题的时间范围。

2.2.7 构建热点问题表

时间范围，热点问题的主题及问题详情等相关信息构建 DateFrame 并将其写入 excel 保存，得到“热点问题表”。

2.3 结论

对留言进行归类时，不需要让各条留言之间进行相似度的计算，采用 DBSCAN 聚类，减少了计算所需的内存以及时间。并根据 k-dist 图对模型进行了参数调整，使得聚类结果较为理想。对于特定地点或人群的确定，在聚类的基础上对每一类留言其进行主题提取，使所得到的结果更加精确。所得部分结果如表 4：

表 4 热点问题表

热度排名	问题ID	时间范围	地点/人群	问题描述
1	1	2019-01-08至2020-01-06	A市A3区伊景园滨河苑	开发商广铁集团捆绑销售车位
2	2	2019-01-04至2019-12-30	A市国王陵国家考古公园	加快周边建设力度
3	3	2019-11-13至2020-01-09	A市A2区暮云街道丽发新城小区	小区附近建搅拌站影响居民生活
4	4	2019-01-07至2019-07-31	A市小区配套幼儿园	幼儿园性质何时改为惠普或公办
5	5	2019-01-04至2019-09-09	A市国家中心城市	加快国家中心城市建设刻不容缓

3.对答复意见评价的文本挖掘

针对附件 4 相关部门对留言的答复意见，从答复的相关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。

3.1 进行数据读取及预处理

（1）在 Python 语言中，使用 pandas 库将题目给出的 excel 文件进行导入并读取对应的留言详情列。

表 5 留言答复详情表

```
df=pd.read_excel(".\C题全部数据\附件4.xlsx",index_col=0)
df.head()
```

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
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
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\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
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\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
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\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
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\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

（2）将文本进行中文分词

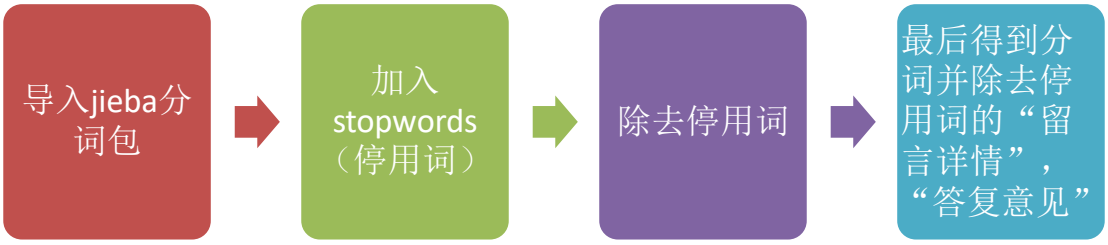


表 6 留言详情分表 1

```
留言编号
2549    [位于, A2, 区, 桂花, 坪, 街道, A2, 区, 公安分局, 宿舍区, 景蓉华苑,...
2554    [潇楚, 南路, 修, 一年, 路, 挖, 稀烂, 围栏, 围, 动工, 有时候, 来台, ...
2555    [地处, 省会, 民营, 幼儿园, 众多, 小孩, 祖国, 未来, 民营, 幼儿园, 教师,...
2557    [您好, 研究生, 毕业, 人才, 新政, 落户, 想, 买, 套, 公寓, 请问, 购买,...
2574    [白竹坡, 路口, 更名, 马坡岭, 小学, 原, 马坡岭, 小学, 取消, 保留, 马坡岭]
Name: 留言详情, dtype: object
```

表 7 留言详情分表 2

```
留言编号
2549    [现将，平台，问政，西地省，栏目，胡华衡，书记，留言，A2，区景蓉，花苑...
2554    [A00023583，A3，区潇楚，南路，洋湖，段，修好，A3，区洋湖，街...
2555    [市民，同志，请，加快，提高，民营，幼儿园，教师，待遇，来信，收悉，现...
2557    [A000110735，平台，问政，西地省，留言，收悉，市住，建局，交由，...
2574    [A0009233，留言，收悉，现将，具体内容，答复，来信，建议，白竹坡，...
Name: 答复意见，dtype: object
```

3.2 进行词频的统计

导入 NumPy，tkinter 分别将“留言详情”，“答复意见”进行统计得到最高的词语是“元”，“收悉”；

表 8 答复意见分词表 1

```
        202
        105
        44
元       35
部门      34
dtype: int64
```

表 9 答复意见分词表 2

```
       275
        56
收悉      46
支持      36
单位      34
dtype: int64
```

3.3 进行词云的绘制

导入 matplotlib.pyplot，wordcloud 和一张名为 dashu 的背景图，分别对 “留言详情” 和 “答复意见” 进行词云的绘制。

图 3 留言详情词频统计图

图 4 答复意见词频统计图

3.4 对“留言详情”和“答复意见”进行相似度分析

首先从文章中提出这两列，然后导入“difflib”包计算，把计算后的数命名为“相似度”，再创建一个新的 Excel 文件并命名为“outdata”

表 10 文本相似度表

	留言详情	答复意见
0	[位于, A2, 区, 桂花, 坪, 街道, A2, 区, 公安分局, 宿舍区, 景蓉华苑,...	[现将, 平台, 问政, 西地省, 栏目, 胡华衡, 书记, 留言, A2, 区景蓉, 花苑...
1	[潇楚, 南路, 修, 一年, 路, 挖, 稀烂, 围栏, 围, 动工, 有时候, 来台, ...	[A00023583, A3, 区潇楚, 南路, 洋湖, 段, 修好, A3, 区洋湖, 街...
2	[地处, 省会, 民营, 幼儿园, 众多, 小孩, 祖国, 未来, 民营, 幼儿园, 教师,...	[市民, 同志, 请, 加快, 提高, 民营, 幼儿园, 教师, 待遇, 来信, 收悉, 现...
3	[您好, 研究生, 毕业, 人才, 新政, 落户, 想, 买, 套, 公寓, 请问, 购买,...	[A000110735, 平台, 问政, 西地省, 留言, 收悉, 市住, 建局, 交由, ...
4	[白竹坡, 路口, 更名, 马坡岭, 小学, 原, 马坡岭, 小学, 取消, 保留, 马坡岭]	[A0009233, 留言, 收悉, 现将, 具体内容, 答复, 来信, 建议, 白竹坡, ...
...	...	...
115	[朱, 局长, 您好, 派出所, 办理, 新, 身份证, 工作人员, 告诉, 快递, 身份证...	[A00085038, 问政, 西地省, 感谢您, 理解, 支持]
116	[， , 您好, 西地省, 某县, 一名, 基层干部, 岁, 多年, 纪检, 政法, 经...	[基层干部, ， , 感谢, 森林, 公安, 事业, 热爱, ， , 国家, 录警,...
117	[一名, 一级, 肢体, 残疾, 低保, 度日, 残联, 买, 台, 电动, 轮椅, 代步,...	[第一, 你试, 修, 厂家, 保修期, 厂家, 购买, 残联, 补贴]
118	[国家, 残, 联网, 发出通知, 组织, 各项, 技能, 比赛, 包含, 盲人, 保健, ...	[建议, 建议, 转给, 部门, 研究]
119	[G, 残疾人, 创业, 协会, 创业, 残疾人, 城乡, 青年, 摆摊, 创业, 市场经济...	[澄清, 十九, 全国政协, 会议, 所说, 安排, 固定, 摊点, 精神, 双联, 双百,...

表 11outdata 部分数据表

df

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	相似度
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	0.038136
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	0.037383
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	0.026810
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	0.030675
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	0.067797
...	...	...	...	...	...	...	...
181267	UU008766	汽车北站进站口附近居民强烈反对建设I市平康肾病医院!	2018/12/12 15:20:46	\n\t\t\t\t\t\n\t\t\t\t\t我们是I市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转I1区委、区人民政府调查处理。	2019/1/8 16:54:53	0.039216
181603	UU008194	强烈反对I市9路公交车改线路	2018/6/12 8:51:03	\n\t\t\t\t\t\n\t\t\t\t\t强烈反对I市9路公交车改线路获悉从7月1日起...	“UU008194”您的留言已收悉。关于您反映的问题，已转市交通运输局调查处理。	2018/7/4 16:55:53	0.105263
184423	UU0082115	对G7县文盛小学特色班的一点质疑	2018/10/11 20:02:52	\n\t\t\t\t\t\n\t\t\t\t\tG7县文盛小学引入特色班，每个家长必须参加...	“UU0082115”您好！获悉关于“对G7县文盛小学特色班的质疑”的网帖后，我局领导高度重...	2018/10/24 9:22:07	0.018321

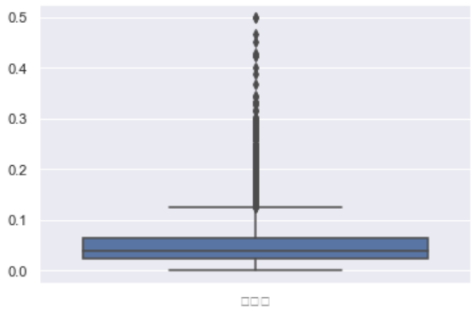


图 5 相似度—箱式图

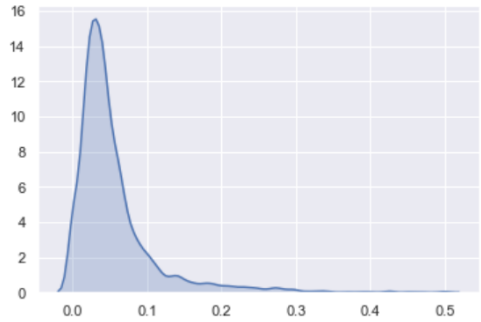


图 6 相似度—kde 图

3.5 小结

问题三是一个开放性的问题，是对“答复意见”和“留言详情”的相似度的程度进行的计算，我认为这种文本相似度的计算可以运用到生活的方方面面，如网购中商品的商家描述程度和顾客的评价之间进行相似度计算，可以判断该商品是否真的符合商家的描述；对文档相似度计算的研究可以应用到很多自然语言处理任务中，例如信息检索，机器翻译，自动问答，复述问题，对话系统等等[4]。

参考文献

[1]荣秋生，颜君彪，郭国强. 基于 DBSCAN 聚类算法的研究与实现[J]. 计算机应用，2004(04):47-48+63.
[2] 虞倩倩. 基于数据划分的 DBSCAN 算法研究[D]. 江南大学.
[3] 唐晓波，向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作，2014(05):60-65.
[4]李兰君. 《面向法律案例检索的文档相似度计算研究》[D].南京师范大学.2018