基于自然语言处理的“智慧政务”中的文本挖掘应用

摘要：

近年来，随着网络问政平台的建设与发展，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。因此，本文通过比较不同的方法，构建基于自然语言处理技术的智慧政务模型，以解决此类问题。

以下是针对三个问题的解题过程：

对于问题 1，通过数据分析与预处理对群众留言记录进行去重，得到不重复的群众留言信息。利用 jieba 中文分词工具对留言描述信息进行分词，成为一个词语集合，这些词语也成为了文本信息表征的基本元素。通过构建停用词表，过滤停用词以降低特征空间的维数。采用 TF-IDF 算法，利用词频统计从众多文本特征中选择出最具有代表性的词汇。通过构建关于留言内容的一级标签分类的朴素贝叶斯模型将群众留言归类为相应的政府职能部门，并对分类方法进行了适当的评价。

对于问题 2，通过统计相关数据，分类筛选汇总，采用 K-means 算法对留言内容进行基于相似度的分类，利用 Knn 算法找出与各中心相似的元素，构建量化热度的评价指标整理出热点问题

对于问题 3，通过挖掘相关部门对留言的答复意见数据，根据答复的及时性、相关性、完整性、可解释性角度对答复意见的质量作出评价，分析答复意见是否及时有效地解决了群众的实际问题。

本文最后作了简要的总结和未来展望，对文章中存在的问题和不足进行了分析，以期在未来不断学习、提升和完善。

关键词：去重 中文分词 TF-IDF 算法 朴素贝叶斯模型 K-means 聚类 KNN 算法

Abstract:

In recent years, with the construction and development of the network political platform, the amount of text data related to various social situations and public opinions has been constantly increasing. It has brought great challenges to the work of relevant departments that used to mainly rely on manual workers to divide messages and sort out hot topics. Therefore, by comparing different methods, this paper constructs an intelligent government model based on natural language processing technology to solve such problems.

The following is the process of solving three problems:

For question 1, the message records of the masses were deduplicated by data analysis and preprocessing to obtain the message information of the masses that was not repeated. Jieba Chinese word segmentation tool is used to segment the message description information into a word set, and these words also become the basic elements of the representation of text information. By constructing a stop word list, the stop words are filtered to reduce the dimension of the feature space. Using tf-idf algorithm, word frequency statistics is used to select the most representative words from many text features. By constructing the naive bayesian model of the first-level label classification of message contents, the message of the masses is classified as the corresponding government functional departments, and the classification method is evaluated appropriately.

For question 2, through statistics of relevant data, classification, screening and summary, k-means algorithm is used to classify message contents based on similarity. Knn algorithm is used to find out elements similar to each center, and quantitative heat evaluation index is constructed to sort out hot issues.

As for question 3, by mining the data of reply opinions from relevant departments, the quality of reply opinions is evaluated according to the timeliness, relevance, integrity and interpretability of the replies, and the actual problems of the masses are solved timely and effectively.

In the end, this paper makes a brief summary and future prospects, and analyzes the problems and shortcomings in the paper, in order to continue to learn, improve and improve in the future.

Keywords: Duplicate removal, Chinese participle, TF-IDF, Naive bayesian model, K-means clustering, KNN algorithm

目录

图目录

一、 引言

近年来，随着互联网数据时代的发展，政务信息逐渐公开透明化，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，如何提升政府的管理水平和施政效率成了一个热门话题。而随着大数据、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。

本文利用文本挖掘技术，以文本分类的方式对群众留言以及政府答复意见数据进行挖掘分析，为政府留言划分和热点问题整理提供有效分析手段，促进政务信息管理朝着智能化方向发展。

二、 分析方法与过程

本用例主要包括如下步骤：

第一步：数据预处理。我们对问题给出的数据集进行统计分析，对重复的群众留言信息进行去重处理，在此基础上进行中文分词。

第二步：数据分析。在对群众留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。采用 TF-IDF 算法，找出每条留言描述的关键词，把留言信息转换为权重向量。建立关于留言内容一级标签分类的高斯朴素贝叶斯模型，对模型进行训练，并评价该分类方法。

第三步：数据整理。统计相关数据，分类筛选汇总，采用 K-means 算法对留言内容进行基于相似度的分类，利用 Knn 算法找出与各中心相似的元素，根据个数多少判定所属类别，并整理出相似问题和热点问题等。

第四步：深入挖掘。通过挖掘相关部门对留言的答复意见数据，根据答复的及时性、相关性、完整性、可解释性角度对答复意见的质量作出评价，分析答复意见是否及时有效地解决了群众的实际问题。

2.1 数据分析与预处理

1、数据的导入

将附件 2 的数据导入到 jupyter notebook 中，命名为 data，截取部分数据，如图 二-1 的形式：

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
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
37	U0008473	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
83	A00063999	投诉A市A1区苑物业违规收停车费	2019/12/30 17:06:14	\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...	城乡建设
303	U0007137	A1区蔡锷南路A2区华庭楼顶水箱长年不洗	2019/12/6 14:40:14	\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...	城乡建设
319	U0007137	A1区A2区华庭自来水好大一股霉味	2019/12/5 11:17:22	\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...	城乡建设
379	A00016773	投诉A市盛世耀凯小区物业无故停水	2019/11/28 9:08:38	\n\t\t\t\t\t\n\t\t\t\t\t我在2015年购买了盛世耀凯小区17栋3楼，...	城乡建设
382	U0005806	咨询A市楼盘集中供暖一事	2019/11/27 17:14:11	\n\t\t\t\t\t\n\t\t\t\t\t由于西地省地区常年阴冷潮湿的气候，加之近年气...	城乡建设
445	A00019209	A3区桐梓坡西路可可小城长期停水得不到解决	2019/11/19 22:39:36	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：您好！家住A市A3区桐梓坡西路...	城乡建设
476	U0003167	反映C4市收取城市垃圾处理费不平等的问题	2019/11/15 11:44:12	\n\t\t\t\t\t\n\t\t\t\t\t我们是梅家田社区辖区内的小区居民，我们每年都...	城乡建设
530	U0008488	A3区魏家坡小区脏乱差	2019/11/10 18:59:24	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的A市政府领导：您们好！我是A市A3区魏...	城乡建设

图 二-1 数据导入图

2、数据的抽样

高质量的数据集是模型匹配和分类的基础，对整个数据集进行分析处理可以促进对数据集的全面认知，从而更好地对数据进行特征工程编码表示，进一步提高数据集的质量。根据分析结果，更容易选择预处理阶段的相关参数，减少重复摸索的概率。根据问题 1 所给的数据集，我们完成了数据集的分析工作，具体结果如下：

对各个留言的数量进行统计，一级标签为城乡建设有 2009 条，一级标签为劳动和社会保障的有 1969 条，一级标签为教育文体的有 1589 条，一级标签为商贸旅游的有 1215 条，一级标签为环境保护的有 938 条，一级标签为卫生计生的有 877 条，一级标签为交通运输的有 613 条。为了得到更多的训练数据，又能使得每个类别的数据数量相同，我们选择每个一级标签中抽取 500 条数据，抽取的城乡建设的数据命名为 sample_1，劳动和社会旅游保障的数据命名为 sample_2，教育文体的数据命名为 sample_3，商贸旅游的数据命名为 sample_4，环境保护的数据命名为 sample_5，卫生计生的数据命名为 sample_6，交通运输的数据命名为 sample_7。将抽取的数据整合，命名为 data_new，截取部分数据，如图 二-2 的形式：

留言编号	liuyanyonghu	liuyanzhuti	liuyanshijian	liuyanxiangqing	label
77862	U0005180	建议F市汽车站建在市区的各个方位	2011/7/27 18:08:56	\n\t\t\t\t\n\t\t\t\t曹书记:　　您好！　　我建言F市应向其他地级...	城乡建设
26389	U0004969	德政园聚心苑小区门口外及院内摊贩得不到彻底根治	2013/11/4 16:27:35	\n\t\t\t\t\n\t\t\t\t尊敬的肖局长:　　您好！有关A1区德政园聚心...	城乡建设
121542	U0007846	通K6县杉木桥乡杉木桥村村民希望能得到政府帮忙	2017/10/26 18:49:01	\n\t\t\t\t\n\t\t\t\t　　尊敬的领导你们好！我是来自杉木桥乡杉木桥...	城乡建设
161173	U0005278	请整顿建筑市场，转变市场风气	2014/3/17 21:39:13	\n\t\t\t\t\n\t\t\t\t厅长:　　请严格抓好施工、监理单位挂靠现象，...	城乡建设
169392	U0006218	Potato）	2019/11/18 10:32:19	\n\t\t\t\t\n\t\t\t\t我是西地省G市G5县锦绣玫瑰园小区的一名业主...	城乡建设
65729	U0006329	E2区体育馆为何如此荒凉?	2013/8/11 10:10:06	\n\t\t\t\t\n\t\t\t\t　　E2区体育馆位于红旗路市中心，虽然早晚有...	城乡建设
32579	U0007802	关于A市住房公积金新政的一点问题	2018/11/1 11:54:34	\n\t\t\t\t\n\t\t\t\t您好，彭菊本人于10月24日，与中介公司签订...	城乡建设
142167	U0004595	M3县应加强荣花乡国家湿地公园厕所配套建设	2018/10/9 8:03:57	\n\t\t\t\t\n\t\t\t\tM3县荣花乡国家湿地公园核心景区荣花村。华天...	城乡建设
168385	U0006366	G1区水榭花城西城管理太差，私家车位经常被占	2017/6/6 8:10:18	\n\t\t\t\t\n\t\t\t\t　　本人是水榭花城西城业主，我们小区管理太差...	城乡建设
171302	U0002257	H4县故事二期3号楼业主乱建铁皮棚严重影响我家住房安全	2019/12/29 10:01:47	\n\t\t\t\t\n\t\t\t\t恳请解决H4县故事二期3号楼业主乱建铁皮棚严...	城乡建设

图 二-2 数据整合图

3、留言的主题和内容合并

留言的主题是对留言内容的概括，留言内容中出现的关键词在留言的主题中出现的可能性会很大，将两块内容合并能够提高分类的准确性。

4、数据的去重

留言数据过多，可能存在重复的留言，需要将这些重复的留言去除，只留下其中的一条，将去重后的数据命名为 data_dup，截取部分数据，如图 二-3 的形式：

```
留言编号
77862     建议F市汽车站建在市区的各个方位\n\t\t\t\t\t\n\t\t\t\t\t曹书记:　　...
26389     德政园聚心苑小区门口外及院内摊贩得不到彻底根治\n\t\t\t\t\t\n\t\t\t\t\...
121542    通K6县杉木桥乡杉木桥村村民希望能得到政府帮忙\n\t\t\t\t\t\n\t\t\t\t\...
161173    请整顿建筑市场，转变市场风气\n\t\t\t\t\t\n\t\t\t\t\t厅长:　　请严格...
169392    Potato） \n\t\t\t\t\t\n\t\t\t\t\t我是西地省G市G5县锦绣玫瑰园...
65729     E2区体育馆为何如此荒凉？\n\t\t\t\t\t\n\t\t\t\t\t　　E2区体育馆位...
32579     关于A市住房公积金新政的一点问题\n\t\t\t\t\t\n\t\t\t\t\t您好，彭菊本...
142167    M3县应加强荣花乡国家湿地公园厕所配套建设\n\t\t\t\t\t\n\t\t\t\t\tM...
168385    G1区水榭花城西城管理太差，私家车位经常被占\n\t\t\t\t\t\n\t\t\t\t\...
171302    H4县故事二期3号楼业主乱建铁皮棚严重影响我家住房安全\n\t\t\t\t\t\n\t\t\...
31056     因为A市政府职能部门的原因，本人失去宝贵的购房摇号资格！\n\t\t\t\t\t\n\t\t...
160929    反映A市楚府名邸一房两本房产证的问题\n\t\t\t\t\t\n\t\t\t\t\t领导：您...
83557     关于G市君香公寓(棚户区改造项目)回迁户的问题\n\t\t\t\t\t\n\t\t\t\t\...
19040     开元路已认定的违章建筑何时解决？\n\t\t\t\t\t\n\t\t\t\t\t尊敬的杨书记...
168194    G1区都市捷座小区现在管理混乱\n\t\t\t\t\t\n\t\t\t\t\t昨天晚上听到小...
168615    G2区公租房使用要求是什么？\n\t\t\t\t\t\n\t\t\t\t\t我想问下相关领导...
68423     E12市龙城居天然气开户费交了一年了，却还没通气\n\t\t\t\t\t\n\t\t\t\t\...
152309    请求领导解决M7市荣庄社区住房困难问题\n\t\t\t\t\t\n\t\t\t\t\t　　我...
84552     G5县城头山镇棚户区改造拖欠工程款\n\t\t\t\t\t\n\t\t\t\t\t请求解决城...
49603     C市到青山桥的县级公路什么时候能维修？\n\t\t\t\t\t\n\t\t\t\t\t尊敬的...
108600    K市龙腾大厦开发商隐瞒房产性质，向购房者强制征收不合理税率\n\t\t\t\t\t\n\t\...
```

图 二-3 数据去重图

5、去除 x 序列

将数据中的 xxx 剔除，如电话号码，日期等。

2.2 jieba 分词操作

在对群众留言信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件 2 中，以中文文本表格的方式给出了留言相关数据。为了便于转换，先要对这些留言信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)，同时采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。

于是通过将数据进行 jieba 分词操作，会将得到以一个个词汇的形式展现出来的新数据集，我们将其命名为 data_cut，截取部分数据，如图 二-4 的形式：

留言编号	
77862	[建议，F，市，汽车站，建，在，市区，的，各个，方位，\n，\t，\...
26389	[德政，园聚，心苑，小区，门口，外及，院内，摊贩，得不到，彻底，根治，...
121542	[通，K6，县，杉木，桥乡，杉木，桥村，村民，希望，能，得到，政府，...
161173	[请，整顿，建筑，市场，，，转变，市场，风气，\n，\t，\t，\t,...
169392	[Potato，），\n，\t，\t，\t，\t，\t，\n，\t，\t...
65729	[E2，区，体育馆，为何，如此，荒凉，？，\n，\t，\t，\t，\t...
32579	[关于，A，市，住房，公积金，新政，的，一点，问题，\n，\t，\t,...
142167	[M3，县应，加强，荣，花乡，国家，湿地，公园，厕所，配套，建设，\n...
168385	[G1，区，水榭，花城，西城，管理，太，差，，，私家车，位，经常，被...
171302	[H4，县，故事，二期，3，号楼，业主，乱建，铁皮，棚，严重，影响，...
31056	[因为，A，市政府，职能部门，的，原因，，，本人，失去，宝贵，的，购房...
160929	[反映，A，市楚府，名邸，一房，两本，房产证，的，问题，\n，\t，\...
83557	[关于，G，市君香，公寓，（，棚户区，改造，项目，），回，迁户，的，...
19040	[开元，路，已，认定，的，违章建筑，何时，解决，？，\n，\t，\t,...
168194	[G1，区，都市，捷座，小区，现在，管理混乱，\n，\t，\t，\t，...
168615	[G2，区公，租房，使用，要求，是，什么，？，\n，\t，\t，\t,...
68423	[E12，市，龙城，居，天然气，开户费，交了，一年，了，，，却，还，...
152309	[请求，领导，解决，M7，市荣庄，社区，住房，困难，问题，\n，\t，...
84552	[G5，县城，头，山镇，棚户区，改造，拖欠，工程款，\n，\t，\t，...
49603	[C，市到，青山，桥，的，县级，公路，什么，时候，能，维修，？，\n...
108600	[K，市龙腾，大厦，开发商，隐瞒，房产，性质，，，向，购房者，强制，征...

图 二-4 中文分词图

2.3 去除停用词

在完成 jieba 分词后，得到的数据都是一个个词汇的形式，同时也存在各种各样的符号，例如\n，\t，？，，（中文的逗号），也存在很多无关的词汇，例如了、的、却，这些词汇和符号不仅不利于分类，还会影响分类结果的准确性，所以需要将这些词汇和符号剔除。需要剔除的词汇和符号构成一个停用词表，数据中停用表中有的便会剔除，得到的数据命名为 data_after_stop，截取部分

数据，如图 二-5 的形式：

留言编号	
77862	[建议，F，汽车站，建，市区，方位，曹，书记，您好，建言，F，市应，...
26389	[德政，园聚，心苑，小区，门口，外及，院内，摊贩，得不到，根治，尊敬，...
121542	[通，K6，杉木，桥乡，杉木，桥村，村民，希望，政府，帮忙，尊敬，领导...
161173	[请，整顿，建筑，市场，市场，风气，厅长，请，抓好，施工，监理，单位,...
169392	[Potato，西地省，G，G5，锦绣，玫瑰园，小区，一名，业主，2010...
65729	[E2，区，体育馆，荒凉，E2，区，体育馆，位于，红旗路，市中心，早晚,...
32579	[住房，公积金，新政，一点，您好，彭菊，10，24，日，中介，公司，签...
142167	[M3，县应，荣，花乡，国家，湿地，公园，厕所，配套，建设，M3，县荣...
168385	[G1，区，水榭，花城，西城，管理，太，差，私家车，位，占，水榭，花...
171302	[H4，故事，二期，号楼，业主，乱建，铁皮，棚，影响，我家，住房，恳请...
31056	[市政府，职能部门，原因，宝贵，购房，摇号，资格，您好，27，日，认筹,...
160929	[市楚府，名邸，一房，两本，房产证，领导，您好，楚府，名邸，同力，家园,...
83557	[G，市君香，公寓，棚户区，改造，项目，回，迁户，领导，您好，\r\n,...
19040	[开元，路，认定，违章建筑，解决，尊敬，杨，书记，您好，开元，路和通，...
168194	[G1，区，都市，捷座，小区，管理混乱，昨天晚上，听到，小区，院子，几十...
168615	[G2，区公，租房，我想，问下，相关，领导，公，租房，用途，可不可以，...
68423	[E12，龙城，居，天然气，开户费，交了，一年，通气，西地省，E12，龙...
152309	[请求，领导，解决，M7，市荣庄，社区，住房，困难，荣庄，社区，一名，...
84552	[G5，县城，头，山镇，棚户区，改造，拖欠，工程款，请求，解决，城头，...
49603	[C，市到，青山，桥，县级，公路，维修，尊敬，领导，您好，C3，花石,...
108600	[K，市龙腾，大厦，开发商，隐瞒，房产，性质，购房者，强制，征收，不合理...

图 二-5 去除停用词图

2.4 数据的优化处理

建立留言所对应的分类标签，命名为 labels，截取部分数据，如图 二-6 的形式：

留言编号	
77862	城乡建设
26389	城乡建设
121542	城乡建设
161173	城乡建设
169392	城乡建设
65729	城乡建设
32579	城乡建设
142167	城乡建设
168385	城乡建设
171302	城乡建设
31056	城乡建设
160929	城乡建设
83557	城乡建设
19040	城乡建设
168194	城乡建设
168615	城乡建设
68423	城乡建设
152309	城乡建设
84552	城乡建设
49603	城乡建设
108600	城乡建设

图 二-6 数据优化处理图

在去除停用词后，得到的数据之间是用逗号分隔的，将数据集中的分词用空格分开，将更加美观且方便后续的数据处理。

2.5 词云图的绘制

对每类标签出现词语进行词频统计，将词频统计结果以词云图的形式展现出来，通过图 二-7 可以看出关键词与我们预计的比较一致。

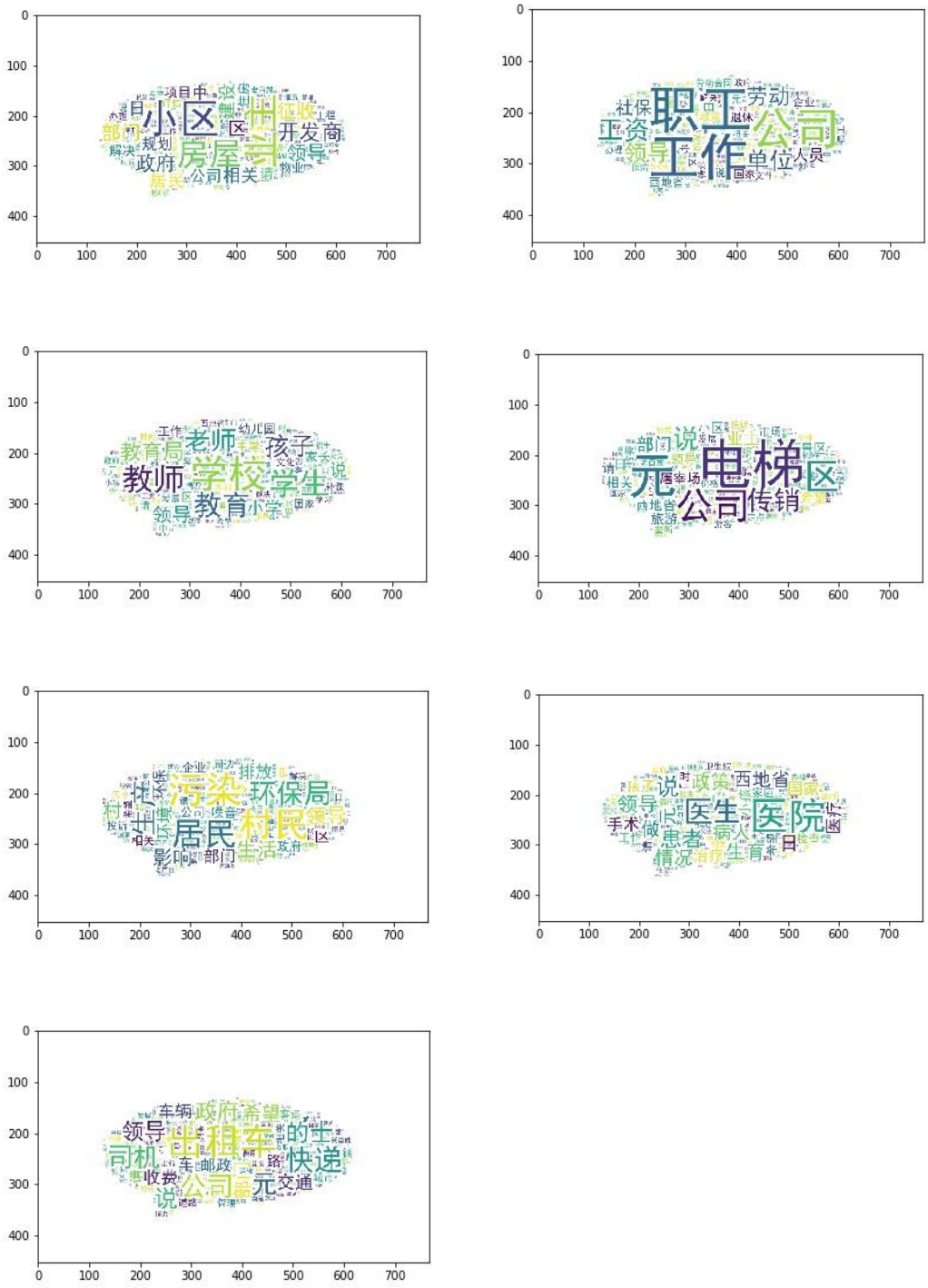

图 二-7 词云图

2.6 K-means 算法分析

K-mean 聚类的原理如下：

假设有一个包含 n 个 d 维数据点的数 $X=\{x_1,x_2,...,x_i,...,x_n\}$，据集其中 $x_i\in R^d$，K-means 聚类将数据集 X 组织为 K 个划分 $C=\{c_k,i=1,2,...,K\}$。每个划分代表一个类 c_k 每个类 c_k 有一个类别中心 μ_i。选取欧式距离作为相似性和距离判断准则，计算该类内个点到聚类中心 μ_i 的距离平方和：

$$J(c_k)=\sum_{x_i\in c_k}\left\|x_i-\mu_k\right\|^2$$

聚类目标是使各类总的距离平方和最小，

$$J(C)=\sum_{k=1}^{K}J(c_k)=\sum_{k=1}^{K}\sum_{x_i\in c_k}\left\|x_i-\mu_k\right\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\left\|x_i-\mu_i\right\|^2$$

其中，$d_{ki}=\begin{cases}1,若x_i\in c_i\\0,若x_i\notin c_i\end{cases}$，所以根据最小二乘法和拉格朗日原理，聚类中心 u_k 应该取为类别 c_k 类各数据点的平均值。

K-mean 聚类的算法步骤如下：

1、从 X 中随机取 K 个元素，作为 K 个簇的各自的中心。

2、分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

4、将 X 中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化。

6、将结果输出。

2.7 数据整理

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，采用 K-

means 算法对留言内容进行基于相似度的归类，利用 Knn 算法找出与各中心相似的元素，根据个数多少判定所属类别，根据相似留言的记录数量和点赞量按照一定权重定义合理的可量化的热度评价指标，进而整理出相似问题和热点问题。

2.8 深入挖掘

通过挖掘相关部门对留言的答复意见数据，根据答复的及时性、相关性、完整性、可解释性角度对答复意见的质量作出评价，分析答复意见是否及时有效地解决了群众的实际问题。

三、 结果分析

3.1 训练集与测试集的分割

将训练集与测试集采用 2,8 分割，其中的 20%作为训练集，80%作为测试集，训练集数据 data_tr，测试集数据 data_te，训练集标签 labels_tr，测试集标签 labels_te。

3.2 获取训练样本的 TF-IDF 权值

在对群众留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把留言信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（TermFrequency）。

$$词频（TF）=某个词在文本中出现的次数$$

考虑到文本有长短之分，为了便于不同文本的比较，进行“词频”标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即：

$$词频（TF）=\frac{某个词在文本中的出现次数}{文本的总词数}$$

或：

$$词频（TF）=\frac{某个词在文本中的出现次数}{该文本出现次数最多的词的出现次数}$$

第二步，计算 IDF 权重，即逆文档频率（InverseDocument Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$逆文档频率(IDF)=\log(\frac{语料库的文本总数}{包含该词的文本数+1})$$

第三步，计算 TF-IDF 值（TermFrequencyDocumentFrequency）。

$$TF-IDF=词频(TF)\times 逆文档频率(IDF)$$

实际分析得出 TF-IDF 值与一个词在留言中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的留言的关键词。

生成 TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找出每条留言的前 5 个关键词；

（2）对每条留言中提取的 5 个关键词，合并成一个集合，计算每条留言对于这个集合中词的词频，如果没有则记为 0；

（3）生成每条留言的 TF-IDF 权重向量，计算公式如下：

$$TF-IDF=词频(TF)\times 逆文档频率(IDF)$$

这样便得到一个 TF-IDF 的权值矩阵，是一个很稀疏的矩阵，将其转换为 array 形式，命名为 X_tr，如图 三-1 的形式：

```
array([[0.        , 0.        , 0.02033797, ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

图 三-1 训练样本 TF-IDF 权值矩阵图

3.3 获取测试样本的 TF-IDF 权值

测试样本与训练样本的维度不同，需要进行维度共享，经过操作得到与训练样本维度形同的 TF-IDF 权值矩阵，同样是一个稀疏的矩阵，将其转换为 array 形式，命名为 X_te，如图 三-2 形式：

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

图 三-2 测试样本 TF-IDF 权值矩阵图

3.4 训练模型

高斯朴素贝叶斯分类算法是以贝叶斯理论为基础并假设特征词在类别确定的条件下都是相互独立的，其基本思想是利用特征项和类别的联合概率来估计给定文档的类别概率。朴素贝叶斯的条件独立性假设虽然忽略了文本中词语之间的相关性，但是其分类效果十分出色，故这里的分类模型我们采用高斯朴素贝叶斯模型，对模型进行训练。

查看准确率，发现准确率有 74.1%，经过多次测试，准确率稳定在 72%。

四、 总结与展望

4.1 总结

为了便利政府相关部门在群众问政留言划分和热点问题整理的工作，促进政务信息管理朝着智能化方向发展。本文基于自然语言处理的相关理论和实验，构建了一套划分留言归属和热点问题挖掘的智慧政务系统模型。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，基本实现本赛题设立的目标。

4.2 展望

本文仍然有很多需要改进的不足之处，日后的研究工作可以从以下几个方面做深入和改进：

（1） 本文从数据预处理到分类模型构建实现的是一个半自动化的文本分类模型。从应用角度来说，可以建立一套更加完备的自动化处理流程完成文本数据的自动管理更新功能，为政务智能化提供更实用的服务。

（2） 本文所构建的模型仍需要通过一定的主观分析得到分类结果，实用性不强。可以将模型分类后的数据实现可视化，使热点问题更加直观地呈现出来，更加便于政府工作者在第一时间直观了解社情民意。

（3） 本文仅从答复意见的及时性和可读性对答复意见的质量进行了简单评价，可以更深入地分析答复意见与留言的相关性，以及基于文本情感分析的可解释性，建立一套答复意见可量化的评价模型。

五、 参考文献

[1]孙海锋 网络招聘信息的数据挖掘与综合分析 北京林业大学

[2]http://www.52nlp.cn/ 中英文维基百科语料上的 Word2Vec 实验

[3]M.Abadi,P.Barham,J.Chen,Z.Chen,A.Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In OSDI, 2016. 3

[4]张彬城 一种基于潜在语义索引和卷积神经网络的智能阅读模型 暨南大学

[5]朱正 政府通告文本分类系统的设计与实现 东南大学

[6]翟东海，鱼江，高飞，于磊等。最大距离法选取初始簇中心的 K_means 文本 聚类算法的研究.西南交通大学.2014

[7]L. Sha, B. Chang, Z. Sui, and S. Li. Reading and thinking: Re-read lstm unit for textual entailment recognition. In COLING, 2016. 3