第八届“泰迪杯”
全国数据挖掘挑战赛

作品名称：基于智慧政务中群众留言及反馈的文本挖掘

基于智慧政务中群众留言及反馈的文本挖掘

摘要

近几年，网络问政平台逐步成为政府了解民意汇、汇集民智、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。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。本文将基于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，利用自然语言处理和文本挖掘的方法进行深度挖掘和分析。

针对问题一，关于群众留言分类，在处理网络问政品台的群众留言时，首先按照一定的划分体系对留言进行分类，划分体系需要参考附件一提供的内容分类三级标签体系，以便后续将群众留言分派至相应的职能部门处理。由于目前大部分电子政务系统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。根据附件二给出的数据，建立关于留言内容的以及标签分类模型。问题一需要注意的问题有文本语义的词语交叉，将多分类问题转化为二分类，数据增强以及将长文本转为短文本和关键句。

针对问题二，将民众留言进行归类，定义热度评价指标，发现热点问题。首先将附件 3 中非结构的化数据进行去重去空、中文分词及停用词过滤等数据预处理，然后基于 TF－IDF 权重策略形成词频统计，筛选高频率的词语--通过 k-means 聚类算法对人群和地点进行划分。根据热度算法定义热度评价指标，得出热点问题前五名。

针对问题三：对于群众留言针对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。相关性即答复意见的内容是否与问题相关。完整性即是否满足某种规范。可解释性即答复意见中内容的相关解释。本问题虽然属于开放性问题，但重点是对文本进行相关性、完整性、可解释性等描述量化，以及构建指标来计算和评价，对此也要先进行文本挖掘。

关键词：多分类模型 F-Score TF－IDF k-means 聚类算法 热度算法 向量量化 构建指标

Text mining based on the comments and feedback of the masses in the intelligent government

Abstract

In recent years, the online political platform has gradually become an important channel for the government to understand public opinions, gather the wisdom of the people, and gather the people's spirit. With the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government. In this paper, based on the records of people's political comments from open sources on the Internet and the replies from relevant departments to some people's comments, the author makes in-depth exploration and analysis by using the methods of natural language processing and text mining.

In view of question 1, regarding the classification of comments from the masses, when dealing with comments from the masses on the network information platform, the comments are first classified according to a certain classification system. The classification system needs to refer to the three-level label system of content classification provided in appendix 1, so as to assign comments from the masses to the corresponding functional departments for subsequent processing. At present, most e-government systems rely on manual handling according to experience, there are problems such as large workload, low efficiency and high error rate. According to the data given in annex ii, establish the model of message content and label classification. One of the issues to note is the semantic crossing of text words, the conversion of multiple classification problems into dichotomies, data enhancement, and the conversion of long text into short texts and key sentences.

For question two, the public comments were classified, the heat evaluation index was defined, and hot issues were found. Firstly, the non-structured data in annex 3 were preprocessed with data such as de-duplication and de-nullification, Chinese word segmentation and stop word filtering, and then word frequency statistics were formed based on the weight strategy of tf-idf to screen high-frequency words -- population and location were divided by k-means clustering algorithm. Heat evaluation index is defined according to heat algorithm, and the top five hot issues are obtained.

In view of question three: for the comments of the masses in response to the comments of the relevant departments, from the relevance, integrity, interpretability and other aspects of the reply to the quality of the comments to give a set of evaluation program, and try to achieve. Relevance is whether the content of the response is relevant to the question. Completeness is whether a specification is met. Interpretability is the relative interpretation of the content of the replies. Although this question is an open one, it focuses on the description and quantification of the relevance, integrity, interpretability and other aspects of the text, as well as the construction of indicators for calculation and evaluation.

Key words: Multi-classification model f-score tf-idf K-means clustering algorithm heat algorithm vector quantization construction index

目录

1 挖掘目标

本次建模的目标是利用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，结合分词工具对留言和答复进行分词，使用 F-Score 对分类方法进行评价，K-means 聚类的方法，热度计算，构建指标进行评价并且对答复进行量化达到以下三个目标：

（1）利用数据预处理和文本分词的方法对非结构化的数据进行文本挖掘，并且按照一定的划分体系对留言进行分类，建立分类模型分析智慧政务中群众留言的问题分类，通过对留言的划分分析文本挖掘提升政府管理水平。

（2）根据四千多条群众留言的数据，将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，分析数据来源城市的热点问题。

（3）针对附件四相关部门答复意见，对答复意见给出一套评价方案，并且对答复进行量化。基于词向量的方法计算短文本相似度，用短文本主题建模解释完整性，尝试实现短文本聚类，从而使智慧政务存在的问题得到有效改善，互动更透明，决策更精准。

2 总体流程与步骤

整体流程如下：

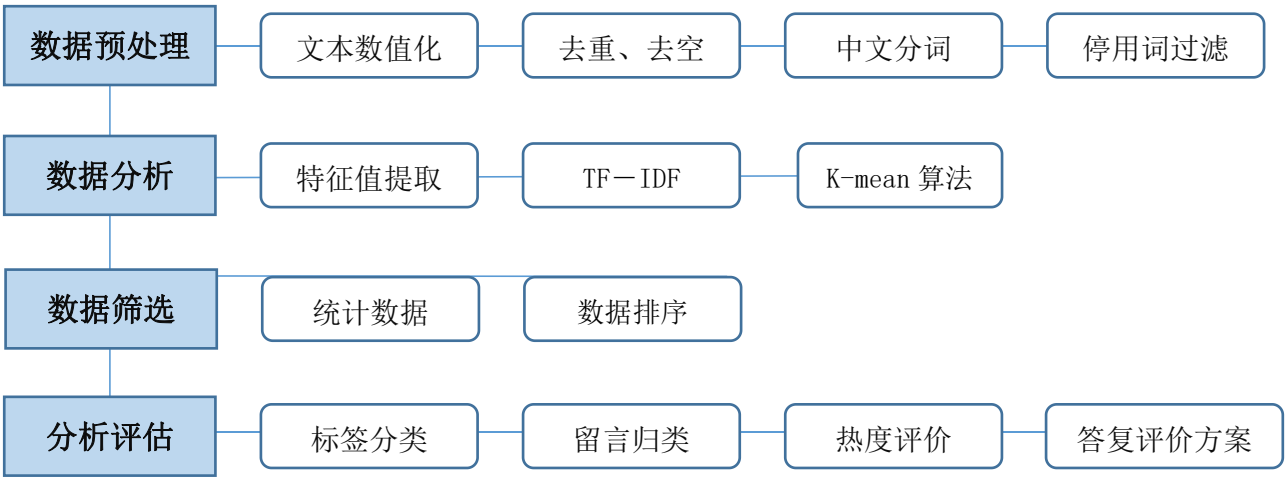

图 2.1 流程图

本例主要包括的步骤：

步骤一：数据预处理，在题目给出的数据中将重复的留言数据进行去重，而后在此基础上进行中文分词。

步骤二：数据分词，对留言进行分词后，讲这些词语转换为向量，供之后使用。采用 TF－IDF 算法，找出留言词频率高的词汇，把留言转换为权重向量。采用 K-mean 算法对留言进行分类，找出各条留言的相似元素，根据数量多的判别其属性。

步骤三：数据筛选，统计相关数据，分类汇总。

步骤四：分析评估，基于步骤三，建立关于留言内容的一级标签分类模型。将留言按地区、人群，问题而归类，定义热度评价指标，给出评价结果。针对群众留言的政府反馈留言制定评价方案。

3 群众留言分类

3.1 文本数据预处理

3.1.1 正则表达式

正则表达式是一种可以用于模式匹配和替换的工具，可以让用户通过使用一系列的特殊字符构建匹配模式，然后把匹配模式与待比较文本进行比较，根据比较对象中是否包含匹配模式，执行相应的程序。对于去除数字、连字符、标点符号、特殊字符，所有大写字母转换成小写字母，实现方法是通过正则表达式：

```
String res[]=line.split( " [^a-zA-Z] " )
```

3.1.2 jieba 中文分词

对群众留言进行文本挖掘之前，威力便于转换，要先对留言中的文本数据进行分词。以词作为基本单元，使用 Python 自动对中文文本进行词语的切分，即使词之间有空格，这样方便计算机识别出各语句的重点内容。

(一)算法：

(1)基于前缀词典实现高效的词图扫描，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)；

(2)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；

(3)对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用 Viterbi 算法。

(二)Jieba 库：

(1)jieba.cut 方法接受三个输入参数：需要分词的字符串；cut_all 参数用来控制是否采用全模式；HMM 参数用来控制是否使用 HMM 模型；

(2)jieba.cut_for_search 方法接受两个参数：需要分词的字符串；是否使用 HMM 模型，该方法适合用于搜索引擎构建倒排索引的分词，粒度比较细；

(3)待分词的字符串可以是 unicode 或 UTF-8 字符串、GBK 字符串；

(4)jieba. Tokenizer (dictionary=DEFAULT_DICT)新建自定义分词器，可用于同时使用不同词典，jieba.dt 为默认分词器，所有全局分词相关函数都是该分词器的映射；

(5)jieba.load userdict (path): 添加用户字典。

得到的分词结果如图 3.1 所示：

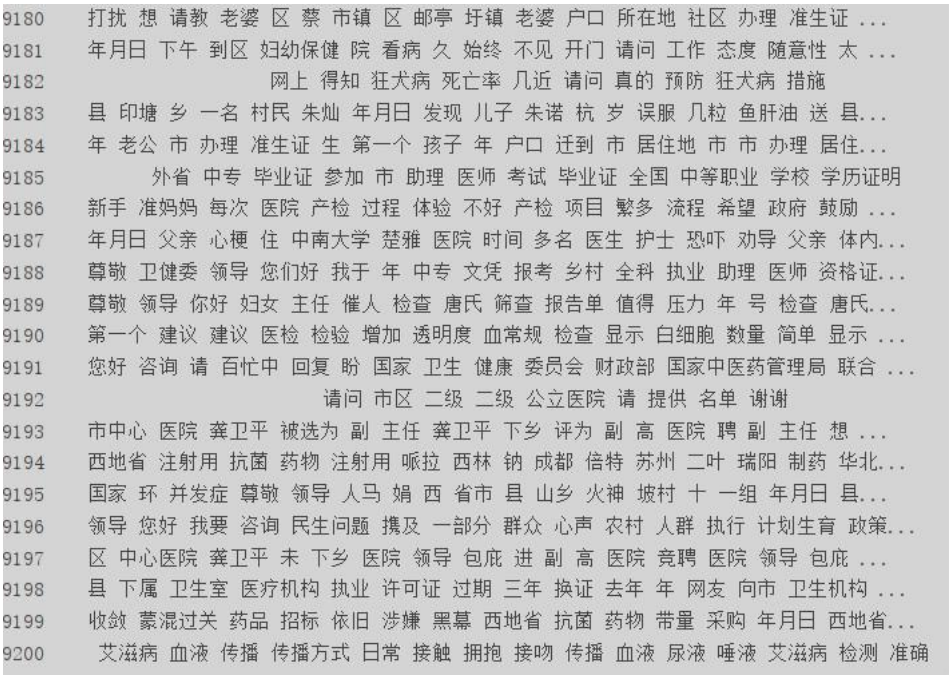

图 3.1 分词结果

3.1.3 群众留言的词频统计

针对群众留言的词频统计，在中文文本分词后使用字典表达词频。统计留言中各种词出现的频率，实现按词表对词语进行分类，并按字典序给出的词表中各词语的出现次数通过运行结果展现出来。这里需要用到的算法是 TF-IDF 算法，首先对词频统计规律进行研究，推导同频词数 In 计算公式、探究各频次词语所占比重，进而将词频统计规律应用于文本关键词提取，提出基于词频统计的 TF-IDF 算法。某个词对留言的重要性越高，它的 TF-IDF 值就越大。

步骤包括：第一步，计算词频：词频(TF)=某个词在留言中出现的次数。第二步，计算逆文档频率：逆文档频率(IDF)=log(留言数据中的文本总数/包含该词的留言数+1)；TF-IDF=词频(TF)*逆文档频率(IDF)，得到的结果如图 3.2 所示：

图 3.2 词频统计

3.1.4 绘制词云图

词云图是文本结果展示的有利工具，通过词云图的展示可以对群众留言数据分词后的高频词予以视觉上的强调突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，其中频率越高的字体越大。首先需要安装 WordCloud，在使用 WordCloud 词云之前，需要使用 pip 安装相应的包。得到的结果如图 3.3 所示：

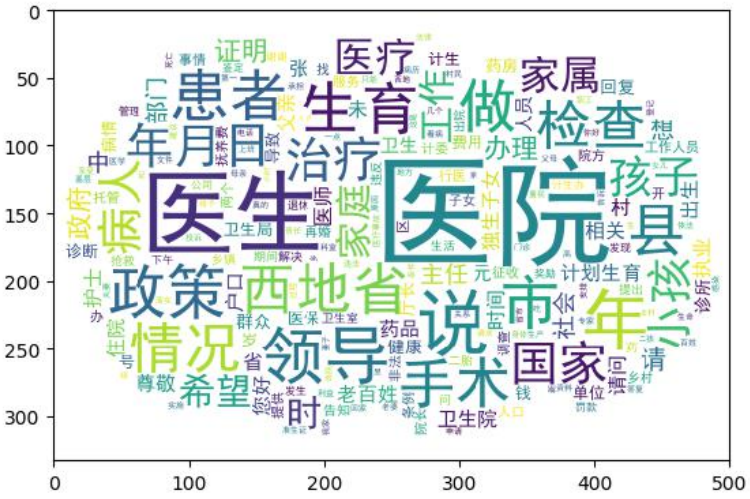

图 3.3 词云图

通过不同的标签，我们可以绘制出不同的词云图，并且可以通过词云图直观的看出不同标签中留言的特征，由此可以帮助我们进行有效的分类。

由上图对标签“卫生计生”进行词频统计和词云的绘制可以看出，在该标签中“医生”、“医院”、“手术”、“生育”等词语出现的频率较高，所以通过建立模型，我们就可以将那些含有该类词语的未分类留言进行高效、准确的分类。

3.2 构建模型

根据附件二给出的数据，建立关于留言内容的一级标签分类模型。对于文本分类而言，可构建的模型有很多，这里我们用到 LDA 模型，也被称为三层贝叶斯概率模型，包含文档(d)、主题(z)、词(w)能够有效对文本建模，和传统的空间向量模型(SVM)相比，增加了概率的信息，通过 LDA 主题模型，能够挖掘数据集中潜在地主题，进而分析数据集的集中观点及其相关特征词。LDA 模型采用词袋模型(Bag Of Words ,BOW)将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。首先将 sklearn.naive_bayes 导入，构建模型。

可以用生成模型来看文档和主题这两件事，所谓生成模型，就是说一篇文档的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到的，要生成一篇文档，它里面的每个词语出现的概率为：

P(词语|文档)=Σ(主题)P(词语|主题)*P(主题|文档)

3.3 分类模型评估方法

本文使用 F-Score 对分类方法进行评价：

基于 F-Score 的特征评价准则，F-Score 是度量特征在不同类别间的区分度的一种指标，F-Score 值越大，代表该特征在不同类别之间的区分度越强。假设 x 代表数据集中的样本(k=1, 2, ・・・, N) .n+为正类样本的数量，n-为负类样本的数量，则数据集中第 i 个特征的 F-Score 可由计算得到。

$$F_i = \frac{(\overline{x_i}^{(+)} - \overline{x_i})^2 + (\overline{x_i}^{(-)} - \overline{x_i})^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} + \overline{x_i}^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} + \overline{x_i}^{(-)})^2} \tag{3-1}$$

4 热点问题挖掘

4.1 数据预处理

4.1.1 数据描述

数据包含 4326 个留言信息，每条信息包含留言编号、留言用户、留言主题、留言时间、留言详情、反对数和点赞数共计 7 个标签。

4.1.2 文本预处理

(一) 文本数值化

将“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”编码转换为名“massageid、userid、topic、time、details、agree、oppose”。

(二) 去重、去空

单独取出“topic”列，进行去除空格、去除 x 序列、文本去重后，数据剩余 4326 项。得到的结果如图 4.1 所示：

0	A区一米阳光婚纱艺术摄影是否合法纳税了？
1	咨询A区道路命名规划初步成果公示和城乡门牌问题
2	反映A县春华镇金鼎村水泥路、自来水到户的问题
3	A区黄兴路步行街大古道巷住户卫生间粪便外排
4	A市A区中海国际社区三期与四期中间空地夜间施工噪音扰民
	...
4321	A市经济学院寒假过年期间组织学生去工厂工作
4322	A市经济学院组织学生外出打工合理吗？
4323	A市经济学院强制学生实习
4324	A市经济学院强制学生外出实习
4325	A市经济学院体育学院变相强制实习

图 4.1

(三) 中文文本分词

在中文里，“词”和“词组”边界模糊。现代汉语的基本表达单元虽然为“词”，且以双字或者多字词居多，但由于人们认识水平的不同，对词和短语的边界很难去区分。

本文采用 Python 开发的中文分词模块---jieba 分词，对附件 3 中“topic”列进行中文分词，jieba 分词用到算法：

基于 Trie 树结构实现高效的词图扫描，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)；

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，使用了 Viterbi 算法；

得到的结果如图 4.2 所示：

```
0                [A, 区, 一米阳光, 婚纱, 艺术摄影, 是否, 合法, 纳税, 了, ？]
1       [咨询, A, 区, 道路, 命名, 规划, 初步, 成果, 公示, 和, 城乡, 门牌, 问题]
2            [反映, A, 县, 春华, 镇金鼎村, 水泥路, 、, 自来水, 到户, 的, 问题]
3             [A, 区, 黄兴路, 步行街, 大, 古道, 巷, 住户, 卫生间, 粪便, 外排]
4       [A, 市, A, 区, 中海, 国际, 社区, 三期, 与, 四期, 中间, 空地, 夜间...
                                    ...
4321        [A, 市, 经济, 学院, 寒假, 过年, 期间, 组织, 学生, 去, 工厂, 工作]
4322             [A, 市, 经济, 学院, 组织, 学生, 外出, 打工, 合理, 吗, ？]
4323                       [A, 市, 经济, 学院, 强制, 学生, 实习]
4324                    [A, 市, 经济, 学院, 强制, 学生, 外出, 实习]
4325                 [A, 市, 经济, 学院, 体育, 学院, 变相, 强制, 实习]
```

图 4.2 分词结果

(四)停用词过滤

中文表达中最常用的功能性词语是限定词，如“的”、“一个”、“这”、“那”等。这些词语的使用较大的作用仅仅是协助一些文本的名词描述和概念表达，并没有太多的实际含义。而大多数时候停用词都是非自动生产、人工筛选录入的，因为需要根据不同的研究主题人为地判断和选择合适的停用词语。

词频（TF），某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否。）

文档频率（DF），最为简单的一种特征选择算法，它指的是在整个数据集中有多少个文本包含这个单词。在训练文本集中对每个特征计一算它的文档频次，并且根据预先设定的阈值去除那些文档频次特别低和特别高的特征。

本文利用停用词表过滤停用词，将分词结果词语和停用词中词语进行匹配，匹配则删除该词语。得到的结果如图 4.3：

```
0                [A, 区, 一米阳光, 婚纱, 艺术摄影, 合法, 纳税]
1       [咨询, A, 区, 道路, 命名, 规划, 初步, 成果, 公示, 城乡, 门牌]
2                [A, 县, 春华, 镇金鼎村, 水泥路, 自来水, 到户]
3             [A, 区, 黄兴路, 步行街, 古道, 巷, 住户, 卫生间, 粪便, 外排]
4       [A, 市, A, 区, 中海, 国际, 社区, 三期, 四期, 空地, 夜间, 施工, 噪...
                                ...
4321        [A, 市, 经济, 学院, 寒假, 过年, 期间, 组织, 学生, 工厂, 工作]
4322             [A, 市, 经济, 学院, 组织, 学生, 外出, 打工]
4323                   [A, 市, 经济, 学院, 强制, 学生, 实习]
4324                [A, 市, 经济, 学院, 强制, 学生, 外出, 实习]
4325             [A, 市, 经济, 学院, 体育, 学院, 变相, 强制, 实习]
```

图 4.3 停用词过滤

4.2 特征值抽取

特征选择方法的目标是找到原始数据集的一个合适的子集，使该子集用在算法模型上的效果不差于原始数据集。和特征选择方法不同，特征提取，又称为特征学习(Feature Learning)，通过一定的规则对原始数据中的特征维度进行变换组合以及抽象，生成新的特征。常用方法如下：

TF-IDF(term frequency - inverse document frequency)词频-逆文本频率。TF 词频：容易理解，频率高能够在一定程度上反应该词的重要性。IDF 逆文本频率：若某一个词在所有文本

中都出现，或出现的频率过高，则也有可能是虚词这种重要性不高却频率很高的词，此时单纯依靠词频来判断词的重要性就不可靠了．所以引入了 IDF 逆文本频率这一个量与 TF 共同判断词的重要性。

互信息是信息论中对变量之间相互依赖性的量度。互信息可以简单地理解为一个随机变量中包含另一个随机变量的的信息量。

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{4-1}$$

标准化互信息，要对熵的整体大小标准化可以消除熵大小的影响.需要针对 H(A) 和 H(B) 的整体大小标准化，以消除 H(A)和 H(B) 大小的影响。

本文采用 TF-IDF 算法抽取特征词条，将权重按照有大到小的顺序排列。

4.3 文本向量空间模型

向量空间模型(VSM)给定一个文档，$D=D(t_1,w_1;t_2,w_2;...;t_n,w_n)$，D 符合以下两条约定：各个特征项 tk(1 小于等于 k 小于等于 n)互异；各个特征项 tk 无先后顺序关系(即不考虑文档的内部结构)在以上两个约定下，可以把特征项$t_1,t_2,\ldots,t_n$看成一个 n 维坐标系，而权重$w_1,w_2\ldots,w_n$为相应的坐标值，因此一个　文本就表示为维空间中的一个向量，我们称$D=D(t_1,w_1;t_2,w_2;...;t_n,w_n)$为文本 D 的向量或向量空间模型。

4.3.1 文本的向量化表示

上述文本将抽取全部文本为候选特征项，需要构造一个词袋，根据群众留言的特征项对应词袋中的位置，组成统一维数的向量。群众留言即可根据词袋组成同一个维数的词向量，通过 TF-IDF 将其向量化得到词汇文本矩阵。

词频矩阵得到的结果如图 4.4 所示：

```
[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 4.4

关键词及权重得到的结果如图 4.5 所示：

```
学院 0.48800995214024995
实习 0.3302597794188751
 0.2988691875725
学生 0.288164091096
强制 0.283511938059375
经济 0.2233862382575
外出 0.20022583600375002
黄兴路 0.17375847065
一米阳光 0.1650941308925
水泥路 0.15642979113625
艺术摄影 0.14943459378625
镇金鼎村 0.14943459378625
```

图 4.5 关键词及权重计算结果

统计词频得到的结果如图 4.6 所示：

A	区	一米阳光	婚纱	艺术摄影	合法	纳税	咨询	...	磁悬浮	贷为	死灰复燃	不查	结案	时刻表	风气	三月
4171	926	1	1	1	24	2	171	...	1	1	1	1	1	1	1	1

图 4.6 词频统计结果

4.3.2 语义空间降维

因为数据量大，构造词汇-文本矩阵过于庞大，计算很困难。并且留言信息中含有同义词的词语，通过特征抽取转化文本向量无法达到自然语言的要求。所以接下来需要潜在语义分析(LSA)理论讲留言信息中文本向量空间中非完全正交的多维特征投影到维数较少的潜在语义空间上，因此用到奇异值分解(SVD)。

4.3.2.1 词汇-文本矩阵的奇异值分解

正交矩阵是在欧几里得空间里的叫法，在酉空间里叫酉矩阵，一个正交矩阵对应的变换叫正交变换，这个变换的特点是不改变向量的尺寸和向量间的夹角。假设二维空间中的一个向量 OA，它在标准坐标系也即 e1、e2 表示的坐标是中表示为(a,b)'（用'表示转置），现在把它用另一组坐标 e1'、e2'表示为(a',b')'，存在矩阵 U 使得(a',b')'=U(a,b)'，则 U 即为正交矩阵。正交变换只是将变换向量用另一组正交基表示，在这个过程中并没有对向量做拉伸，也不改变向量的空间位置，加入对两个向量同时做正交变换，那么变换前后这两个向量的夹角显然不会改变。正交矩阵的行（列）向量都是两两正交的单位向量，正交矩阵对应的变换为正交变换，它有两种表现：旋转和反射。正交矩阵将标准正交基映射为标准正交基。

特征值分解（EVD），选择一种特殊的矩阵——对称阵。对称阵有一个性质：它总能相似对角化，对称阵不同特征值对应的特征向量两两正交。一个矩阵能相似对角化即说明其特征子空间即为其列空间，若不能对角化则其特征子空间为列空间的子空间。从对称阵的分解对应的映射分解来分析一个矩阵的变换特点是非常直观的。假设对称阵特征值全为 1 那么显然它就是单位阵，如果对称阵的特征值有个别是 0 其他全是 1，那么它就是一个正交投影矩阵，它将 m 维向量投影到它的列空间中。

奇异值分解(SVD)，作为一个基本的算法，在很多机器学习算法中都有存在，特别是在现在的大数据时代，由于 SVD 可以实现并行化，因此更是大展身手。我们称利用 SVD 的方法为潜在语义索引(Latent Semantic Indexing ，LSI)或潜在语义分析(Latent Semantic Analysis ，LSA)。

SVD 的基本公式: $A=U\sum V^T$ 。其中，$A\in R^{m*n}$ ，$U\in R^{m*n}$，$\sum \in R^{m*n}$ 且除了主对角线上的元素以外全为 0，主对角线上的每个元素都称为奇异值，且已按大小拍好序，$V\in R^{m*n}$ 。其中，U 的列向量即是 AA^T 的特征向量，一般我们将 U 中的每个特征向量叫做 A 的左奇异向量；V 的列向量即是 AA^T 的特征向量，一般我们将 V 中的每个特征向量叫做 AA^T 的右奇异向量。

对于奇异值，它跟我们特征分解中的特征值类似，在奇异值矩阵中也是按照从大到小排列，而且奇异值的减少特别的快，在很多情况下，前 10%甚至 1%的奇异值的和就占了全部的奇异值之和的 99%以上的比例。也就是说，我们也可以用最大的 K 个的奇异值和对应的左右奇异向量来近似描述矩阵。也就是说，可以对 $U\sum V^T$ 三个矩阵进行裁剪，比如将特征 n

维降至 k 维，那么$U \in R^{m*k}$，$\sum \in R^{k*k}$，$V \in R^{n*k}$即可。奇异值分解图如图 4.7 所示：

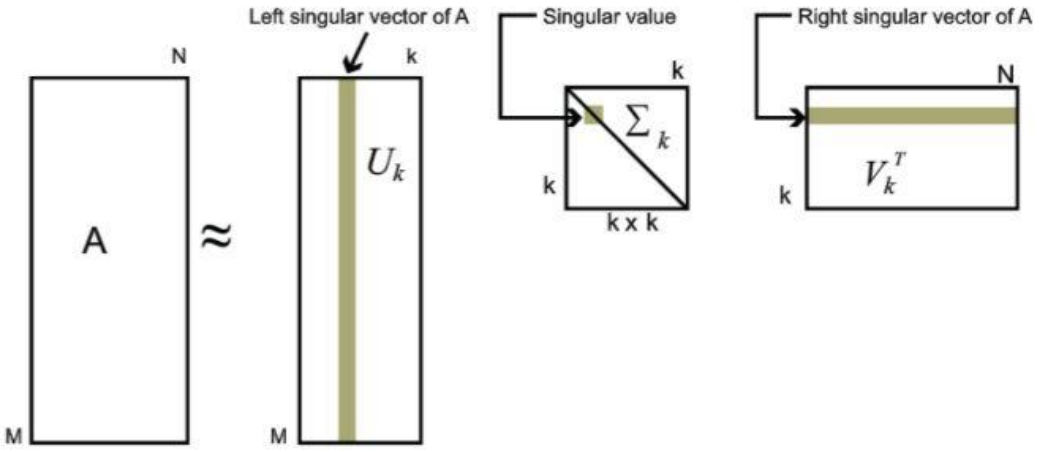

图 4.7 奇异值分解图

4.3.2.1 向量语义化

对特征项为 n 的文本向量 w 进行奇异值分解得到$\mathrm{w} = \vec{\mathrm{w}}\sum \mathrm{U}$，t 在 k 维映射后的向量

$\vec{\mathrm{w}} = \mathrm{w}\mathrm{U}_{\mathrm{k}}^{T}\sum_{\mathrm{k}}^{-1}$，然后进行文本聚类。

4.4 文本聚类

4.4.1 文本相似度计算

文本相似度问题包含：词与词、句与句、段落与段落、篇章与篇章之间的相似度问题；以及词与句、句与段落、段落与篇章等之类的相似度问题，这里的相似指的是语义的相似。这些问题的难度递增。文本相似度计算方法可分为四大类：基于字符串的方法（String-Based）、基于语料库的方法（Corpus-Based）、基于世界知识的方法（Knowledge-Based）和其他方法。

基于字符串的方法，该方法从字符串匹配度出发，以字符串共现和重复程度为相似度的衡量标准。根据计算粒度不同，可以将该方法分为基于字符的方法和基于词语的方法。

基于语料库的方法，该方法利用语料库中获取的信息计算文本相似度。

基于世界知识的方法，该方法是利用具有规范组织体系的知识库计算文本相似度，一般分为两种：基于本体知识和基于网络知识。

4.4.2 文本聚类文本

文本聚类是将一个个文档由原有的自然语言文字信息转化成数学信息，以高维空间点的形式展现出来，通过计算那些点距离比较近来将那些点聚成一个簇，簇的中心叫做簇心。一个好的聚类要保证簇内点的距离尽量的近，但簇与簇之间的点要尽量的远。文本聚类过程如图 4.8 所示：

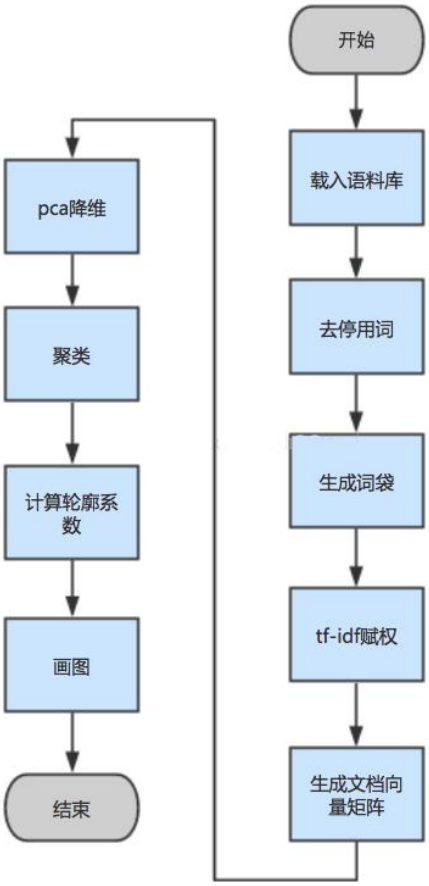

图 4.8 文本聚类流程

4.4.3 K-mean 聚类算法

K-means 算法是非监督学习（unsupervised learning）中最简单也是最常用的一种聚类算法，具有的特点是:对初始化敏感。初始点选择的不同，可能会产生不同的聚类结果；最终会收敛，不管初始点如何选择，最终都会收敛。针对每个点，计算这个点距离所有中心点最近的那个中心点，然后将这个点归为这个中心点代表的簇。一次迭代结束之后，针对每个簇类，重新计算中心点，然后针对每个点，重新寻找距离自己最近的中心点。如此循环，直到前后两次迭代的簇类没有变化。

K-means 算法是将样本聚类成 k 个簇（cluster），具体算法描述如下：

1、 随机选取 k 个聚类质心点（cluster centroids）为$\mu_1,\mu_2,\ldots,\mu_k \in \mathbb{R}^n$

2、 重复下面过程直到收敛 {

对于每一个样例 i，计算其应该属于的类 $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

对于每一个类 j，重新计算该类的质心 $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

}

流程图如图 4.9 所示：

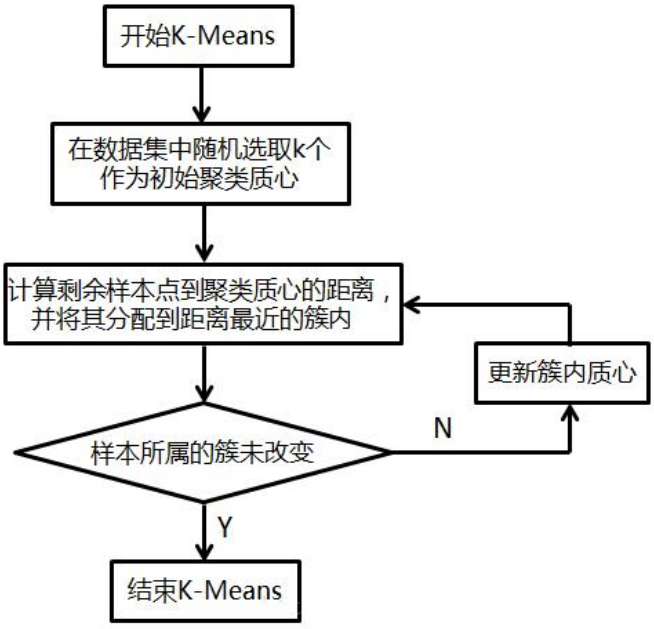

图 4.9 K-mean 聚类算法流程图

4.5 热度评价指标

个性化推荐是基于一定的 数据之上使用的，无论是基于用户行为的个性化，还是基于内容相似度的个性化，都建立在大量的用户数和内容的基础上。产品发布之初，一般两边的数据都有残缺，因此个性化推荐也无法开展。所以在产品发展的初期，推荐内容一般采用更加聚合的“热度算法”，顾名思义就是把热点的内容优先推荐给用户。虽然无法做到基于兴趣和习惯为每一个用户做到精准化的推荐，但能覆盖到大部分的内容需求，而且启动成本比个性化推荐算法低很多。因此内容型产品，在发布初期用热度算法实现冷启动，积累了一定数量级以后，才逐渐使用个性化推荐算法。

热度算法也是需要不断优化去完善的，其基本原理为：新闻热度分 = 初始热度分 + 用户交互产生的热度分 - 随时间衰减的热度分 $S\text{core} = S_0 + S_{User} - S_{Time}$。

留言词条入库后，为之赋予一个初始热度值，该留言词条就进入了热度列表进行排序；随着留言词条不断被用户点击点赞，反对等，这些用户行为被视作帮助留言词条提升热度，系统需要为每一种新闻赋予热度值。按照类别给予留言不同的初始热度，让用户关注度高的类别获得更高的初始热度分，从而获得更多的曝光，参考新闻热度初始值的赋分标 S(Type)。参考类别赋分方式如 4-2 所示：

(4-2)

解决留言词条入库的初始分之后，接下来是热度分的变化。先要明确用户的的哪些行为会提高新闻的热度值，然后对这些行为赋予一定的得分规则。例如对于单条词条，用户可以点击阅读(click)，收藏(favor)，分享(share)，评论(comment)这四种行为，我们为不同的行为

赋予分数，就能得到新闻的实时用户行为分。我们根据“附件 3”给出的值设定适于本题的用户行为分：S(Users)(用户互动分)=20*分享数(发布次数)+10*回答数(点赞量)-10*反对量。

由于新闻的强时效性，已经发布的新闻的热度值必须随着时间流逝而衰减，并且趋势应该是衰减越来越快，直至趋近于零热度。参考牛顿冷却定律，时间衰减因子应该是一个类似于指数函数：$T_{Time}=e^{k(T_1-T_0)}$。根据本题，将时间对词条热度的影响定义为：

$$T_{Time}=\frac{T_t(\text{问题持续时长})}{T(\text{所有数据总时长})}$$

综上设定的热度评价指标为：$Score=(S_{Type}+S_{Users})*T_{Time}$

5 答复意见的评价

5.1 数据预处理

5.1.1 jieba 中文分词

首先进行数据读取与数据清洗，使用正则表达式去重，去除特殊字符。利用一个中文词库，确定中文字符之间的关联概率。中文字符间概率大的组成词组，形成分词结果。进入代码后新建一个脚本文件有关于分词的操作，然后导入 jieba 库。将政府答复意见进行分词，方便后续操作。分词的过程中我们会发现实际上有些词实际上意义不大，比如标点符号以及语气词，这个时候我们需要将停用词去除掉，自定义下载 StopwordCN.txt 停用词组。得到的结果如图 5.1 所示：

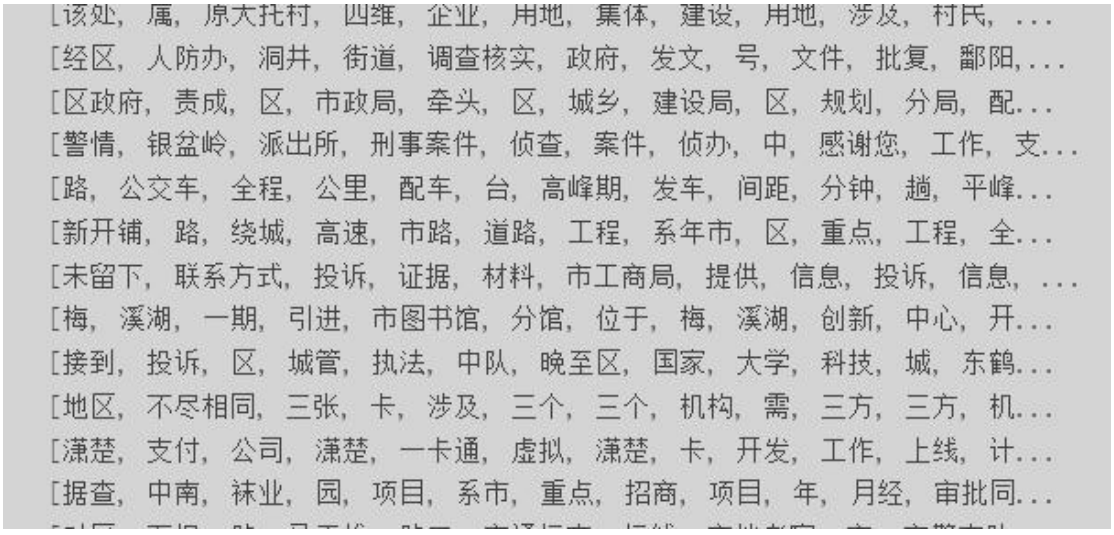
[该处，属，原大托村，四维，企业，用地，集体，建设，用地，涉及，村民，...
[经区，人防办，洞井，街道，调查核实，政府，发文，号，文件，批复，鄱阳，...
[区政府，责成，区，市政局，牵头，区，城乡，建设局，区，规划，分局，配...
[警情，银盆岭，派出所，刑事案件，侦查，案件，侦办，中，感谢您，工作，支...
[路，公交车，全程，公里，配车，台，高峰期，发车，间距，分钟，趟，平峰...
[新开铺，路，绕城，高速，市路，道路，工程，系年市，区，重点，工程，全...
[未留下，联系方式，投诉，证据，材料，市工商局，提供，信息，投诉，信息，...
[梅，溪湖，一期，引进，市图书馆，分馆，位于，梅，溪湖，创新，中心，开...
[接到，投诉，区，城管，执法，中队，晚至区，国家，大学，科技，城，东鹤...
[地区，不尽相同，三张，卡，涉及，三个，三个，机构，需，三方，三方，机...
[潇楚，支付，公司，潇楚，一卡通，虚拟，潇楚，卡，开发，工作，上线，计...
[据查，中南，袜业，园，项目，系市，重点，招商，项目，年，月经，审批同...

图 5.1 群众答复意见 jieba 分词

5.1.2 群众答复意见词频统计

利用 jieba 分词，排除停用词 stopword 之后，对回复留言进行词频统计，利用 jieba 分词，排除停用词 stopword 之后，对回复留言进行词频统计。关于评价方案的制定，我们需要一个重要性调整系数，衡量一个词是不是常见词。如果某个词比较少见，但是他在文本中多次出现，那就是我们要找的关键词。用统计学表达就是在词频的基础上，对每一个词分配一个重要性权重，这里需要用到的算法是 TF-IDF 算法。

得到的结果如图 5.2 所示：

图 5.2　群众答复意见词频统计

5.2 相关性

文本相似度算法很多，我们使用了余弦距离方法，余弦可以通过向量公式计算。两个向量夹角越小，无限接近于 0 则 cos 近似于 1，则极相似；而两个向量互相垂直，cos=0，则互不相关。首先使用大规模语料库通过 word2vec 训练出词向量，然后将短文本进行分词操作，并找出每个词对应的词向量，最后对短文本的所有词的词向量进行求和(也可以根据词性或规则进行加权求和)操作，获得该短文本的句子向量。对两个短文本句子向量进行距离度量，最终获得其相似度值。

余弦距离，即使用两个向量的夹角余弦来衡量两个向量方向的差异。规定夹角余弦取值范围为[0,1]。余弦值越大表示两个向量的夹角越小，即两个向量越相似。

二维空间中向量 $a(x_{11},x_{12})$ 与 $b(x_{21},x_{22})$ 的夹角余弦公式：

$$cos\theta = \frac{x_{11}x_{21} + x_{12}x_{22}}{\sqrt{x_{11}^2 + x_{12}^2}\sqrt{x_{21}^2 + x_{22}^2}} \tag{5-1}$$

N维向量空间中向量 $a(x_{11},x_{12},\cdots,x_{1N})$ 与 $b(x_{21},x_{22},\cdots,x_{2N})$ 间的夹角余弦公式为：

$$cos\theta = \frac{\sum_{k=1}^{N} x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{N} x_{1k}^2}\sqrt{\sum_{k=1}^{N} x_{2k}^2}} \tag{5-2}$$

5.3 完整性

主题提取模型通常包含多个流程，比如文本预处理、文本向量化、主题挖掘和主题表示过程。每个流程中都有多种处理方法，不同的组合方法将会产生不同的建模结果。首先，我们需要考虑下如何评估一个主题模型建模效果的好坏程度。多数情况下，每个主题中的关键词有以下两个特征： 关键词出现的频率得足够大、足以区分不同的主题。

主题模型将选择主题词语分布中频率最高的词语作为该主题的关键词，但是对于 SVM 和 K-mean 算法来说，模型得到的主题词语矩阵中既包含正向值也包含负向值，我们很难直接从中准确地提取出主题关键词。为了解决这个问题，我选择从中挑出绝对数值最大的几个词语作为关键词，并且根据正负值的情况加上相应的标签，即对负向词语加上 "^" 的前缀，比如"^bergers"。

5.4 可解释性

广义上的可解释性指在我们需要了解或解决一件事情的时候，我们可以获得我们所需要的足够的可以理解的信息。反过来理解，如果在一些情境中我们无法得到相应的足够的信息，那么这些事情对我们来说都是不可解释的。而具体到机器学习领域来说，以最用户友好的决策树模型为例，模型每作出一个决策都会通过一个决策序列来向我们展示模型的决策依据，而且决策树模型自带的基于信息理论的筛选变量标准也有助于帮助我们理解在模型决策产生的过程中哪些变量起到了显著的作用。本文中用到的是基于实例的方法，主要是通过一些代表性的样本来解释聚类、分类结果的方法。当然基于实例的方法的一些局限在于可能挑出来的样本不具有代表性或者人们可能会有过度泛化的倾向。

一些基于实例的表达法能够更进一步对实例进行显式的泛化。典型的方法是通过建立矩形区域来包围属于同一类的实例。如果一个未知类的实例落入某一矩形区域内，它将被赋予相应的类，而落在所有矩形区域外的样本将服从最近邻规则。当然这将产生与直接的最近邻规则不同的决策边界。落入矩形的多边形部分将被砍掉，而由矩形边界取代。在实例空间的矩形泛化就像是包含特殊条件形式的规则，它对一个数值变量进行上、下边界的测试，并选择位于其间的区域。不同尺寸的矩形对应于由逻辑与组合在一起的在不同属性上的测试。

选择一个最适合的矩形作为测试边界所产生的规则，将比由基于规则的机器学习方案产生的规则更为保守，因为对于区域的每一个边界，都有一个真正的实例落在边界上(或边界内)。而像 x<a(是一个属性值，a 是一个常量)的测试将包围一半的空间，不管 x 有多小只要它小于 a。当在实例空间中运用矩形泛化时，能够做到保守，因为如果一个新的样本落在所有区域以外，还可以求助于最近邻的度量方法。而采用基于规则的方法时，如果没有规则适用于这个样本，那么它将不能被分类，或者仅得到一个缺省的分类。更加保守的规则的优点是尽管保守的规则并不完整，但它也许比一个覆盖所有事件的规则集的表达更为清楚。最后，要保证区域之间不重叠，也就是保证最多只能有一个规则适合应用于一个样本，这样就避免了在其他基于规则学习系统中，多个规则适用于一个样本的难题。

5.5 评价方案

智慧政务的本质是指以政务服务平台为基础，以公共服务普惠化为主要内容，以实现智慧政府为目标。我们这里用到 python 实现向量量化(LVQ) 算法对群众答复意见的评价进行量化。

向量量化算法和 K 均值算法类似，是找到一组原型向量来聚类，每一个原型向量代表一个簇，将空间划分为若干个簇，从而对于任意的样本，可以将它划入到与它距离最近的簇中。特别的是 LVQ 假设数据样本带有类别标记，可以用这些类别标记来辅助聚类。

大致思想如下：

1、 统计样本的类别，假设一共有 q 类，初始化为原型向量的标记为{t1, 12, ... ,tq}. 从样本中随机选取 q 个样本点位原型向量{p1,p2, ... ,pq}.初始化一个学习率 a,a 取值范围(0, 1) .

2、从样本集中随机选取一个样本(x,y),计算该样本与 q 个原型向量的距离(欧几里得距离)， 找到最小的那个原型向量 P, 判断样本的标记 y 与原型向量的标记 t 是不是一致。若一致则更新为 p'=p+a* (x-p) , 否则更新为 p'=p-a* (x-p) .

3、重复第 2 步直到满足停止条件。(如达到最大迭代次数)

4、返回 q 个原型向量。

定义一个答复量化类，三个属性，分别为相关性、完整性、可解释性，评价原则：

1、相关性即答复意见的内容是否与问题相关。

2、完整性即是否满足某种规范。

3、可解释性即答复意见中内容的相关解释。

评价质量：

根据对附件四经过一系列文本挖掘发现相关部门对与群众生产、生活、工作密切相关的热点、难点问题回复的非常详细。对于政府及其工作部门工作作风、质量、效率的投诉、批评、意见以及建议都虚心接受，且对此做出解释，并且反思出现此类问题的原因，以及会积极出台改善政策。对于群众对政府部门工作业务表达的支持与认可，相关部门表示对群众配合工作的感谢以及表达为人名服务的宗旨。因此，根据向量的量化和指标的构建对评价质量整体呈满意状态。回复及时、答复质疑，且能切实解决问题。

希望智慧政务通过统一规范服务标准、优化服务流程和网上办理、开展网上咨询，构建起一套公开透明、高效便捷的政府服务体系，让群众办事更方便，解决问题更顺畅。

6 群众留言分类

在留言分类中朴素贝叶斯是比较常用的算法，朴素的含义是假设特征之间相互独立。假设词汇表中有 1000 个单词，要得到好的概率分布，就需要足够的数据样本，假设每个特征需要 N 个样本，那么对于 1000 特征的词汇，则需要 N 的 1000 次幂。随着特征数目增大样本数会迅速增长。但如果假设特征之间相互独立，则样本数可以由 N 的 1000 次幂减少到 1000*N 个。所谓的独立指的是统计意义上的独立，即一个特征或者单次出现的可能性与它和其他单词相邻没有关系。比如，单词 bacon 出现在 unhealthy 和出现在 delicious 后面的概率相等，当然这个假设并不正确，两者肯定不相等，这个假设正是朴素贝叶斯中的朴素的概念。朴素贝叶斯的另外一个假设是每个特征同等重要。其实这个假设也有问题，如果要判断留言板的留言是否恰当，可能并不需要看完所有的 1000 个单词，而只需要看 10-20 个单词(特征)即可做出判断。

本题就是运用附件 1 中的标签对群众留言进行分类，通过词频统计和词语图的绘制可以直观的看出不同标签下留言的特征。得到的结果如图 6.1、图 6.2 所示：

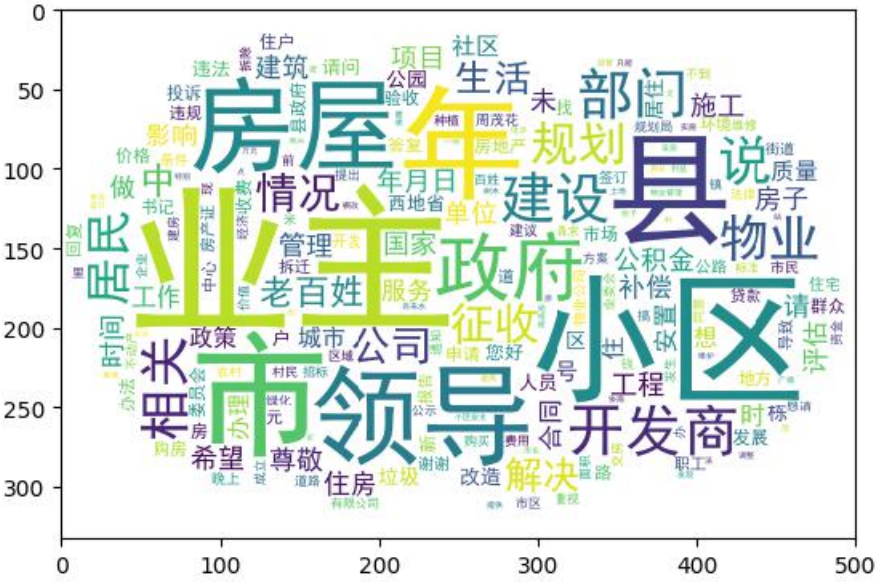

图 6.1 城乡建设词云图

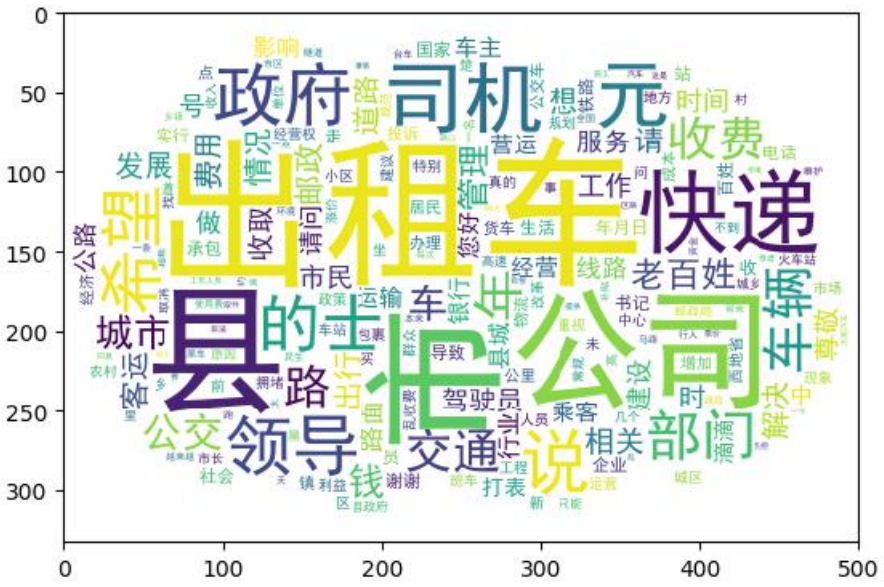

图 6.2 交通运输词云图

在城乡建设标签的留言中，“业主”、“房屋”、“小区”等词汇出现的较为频繁，而在交通运输留言中，“出租车”、“快递”、“司机”等词汇出现的较为频繁，将这种方法运用到不同的标签中，可以由此建立模型，将出现这些词汇的相关留言赋予对应的标签。这样我们就可以极大提高留言分类的效率以及准确率。

7 热点留言问题挖掘

随着 Internet 的发展和普及，尤其是 web2.0 时代之后，普通网络用户已经从信息接收者转变为信息的生产者，网络中充斥着各种数据。其中有很多具有评论性和主观倾向性的文本，这些网络数据能够反映出发表者对于评论对象的观点态度。例如，本次题目收集的 A 市群众意见留言中含有用户发表的大量评论，包含着用户对各事件、人物的观点态度。然而，绝这些网上数据大多是计算机无法直接处理的非结构化文本数据。在这种情形之下，如何通过相关技术分析文本中表达的观点与情感极性是很必要的。众多问题意见中被一次又一次提及，被其他用户所点赞、评论的即成为热点问题。热点问题需要相关部门重点关注、最先解决。

本题中所有意见文本的词语出现的频率可以重点关注，即热点话题的雏形。以下仅展示部分有效词频率高的词语，如图 7.1 所示：

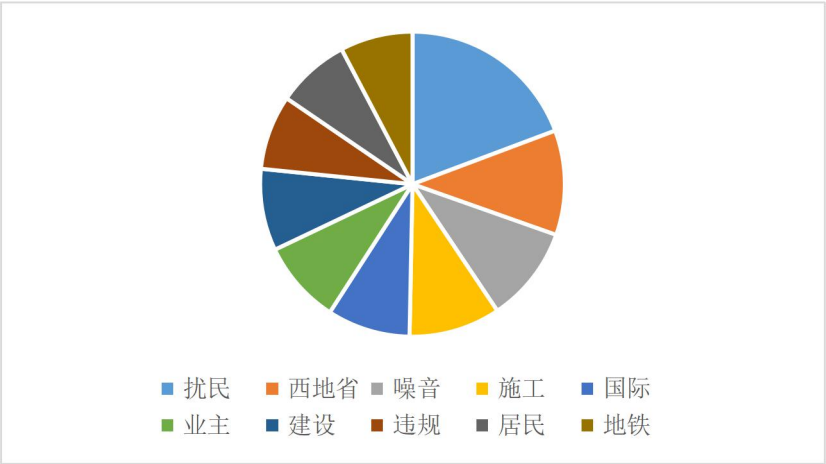

图 7.1 有效词语词频前十名

根据次频率高的“地点”、“人群”词语可筛选出热门事件，通过合理的热度计算得出 A 市需要密切关注的排名在前五的热点问题。

其一，“58 车贷案”，在五个月内被群众西地省展星投资有限公司涉嫌诈骗 58 车贷案件案发后并未得到及时却透明化的解决，受案的群众众多，涉案资金庞大，涉案人员并未抓获，调查案件报告也未公开，需要相关部门特别关注，尽快解决，给予受害人员公正解决。

其二，A5 区汇金路五矿万镜 K9 县的楼房原本定位高端别墅小区，实行人车分流管理，标榜五星级服务，但是屡次出现群租房泛滥成灾，毛坯房擅自改格子间，整个楼道烟味呛鼻的，这些群租房的存在给合法的住户安全带来极大的隐患以及不确定性。相关部门需合理整顿此片区域的楼房，物业管理提高。

其三，经济学院的学生在 2017 年至 2019 年多次持续提出学校强制去鼎点公司实习的问题，此问题持续时间之长，每年都有学生去非本专业的单位实习，强制要求实习时长，浪费学生的时间。需要教育部门重视，尽快解决。

其四，A2 区丽发新城小区旁边建立一个非法搅拌站，污染周边的卫生环境，扰乱居民正常作息。需要有关部门尽快取缔此非法搅拌站。

其五，渝长厦高铁最新的红线征地范围以及走向经过，紧挨着绿地海外滩小区二期，最近的位置只有 30 米不到，因按技术条件此小区和高铁太近，已经不适用于居住，无法消除噪声问题。希望相关部门尽快解决小区居民生活影响的问题。

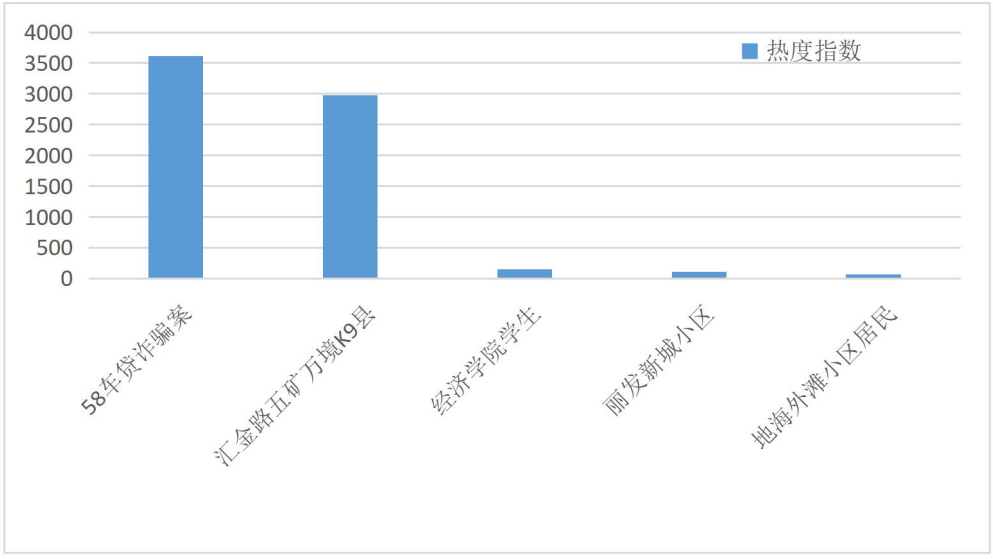

图 7.2 热点问题前五名热度指数

8 答复群众留言的评价

网民留言，是人民群众表达诉求的新途径；相应地，回复网民留言，成为党委和政府密切联系群众、回应百姓诉求的新渠道。因此加大统筹协调力度，及时化解矛盾和纠纷，努力做到网民留言回复率和群众满意度“百分百”。留言内容中涉及的每一类问题在处理过程中都形成了较为科学的解决方案，大大提高了留言回复效率。根据词频统计绘制的词云图可以明显看出针对群众留言回复的热点，得到的结果如下图：

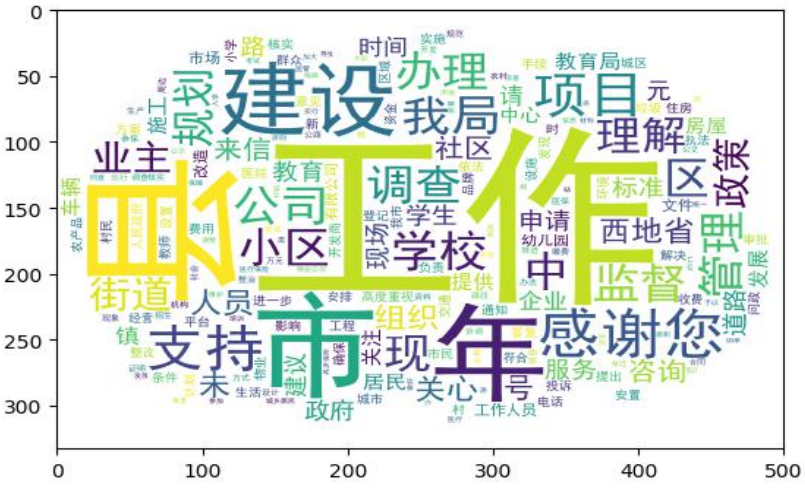

图 8.1 群众留言回复词云图

民政业务现状呈现出多元化、多层次和社会性。社会福利以及社会救助问题成为热点留言问题，信息资源共享弱，公共便民服务缺乏，跨部门资源互通和业务协同机制不完善。对于留言总结做出需求分析为以下要点：

1、统一的公共服务门户功能

提供以公开透明为目的的政府综合信息服务功能，为社会公众获取民政事业信息和各类事项办理提供优质、方便、快捷的服务，实现一站式的信息服务体验，增强透明度，提高民生服务满意度。

2、以规范高效为目的的业务网上办理功能

通过对流程进行梳理使之程序化、规范化、无纸化，实现民政业务的在线工作，提高行为的透明度和效益。通过建立电子化信息的管理方式、以及电子印章和电子签名等电子凭证审核控制机制，进一步提高业务经办效率和管理水平。

3、以高效协作为目的的安全共享与协同功能

通过建立统一的协同办公系统，实现无纸化网上办公、公文流转等功能。实现内部各部门间有效衔接，部门间全流程贯通互动；外部与相关部门系统衔接，推动民政核心业务高效开展。

4、以有效监管为目的的政府管理监督功能

通过各项业务的全面网上电子化办理，实现规范业务、实时监控，在保障安全的基础上，实现自动化管控，强化政策的落实和执行，强化业务跟踪与监管。

5、以科学决策为目的的政府决策支持功能

在强大的技术平台和决策支持系统的支撑下，基于数据挖掘技术，以数据分析为基础，科学、全面地掌握发展趋势，实现辅助决策进一步准确、智能、高效。

6、延伸和拓展功能

随着民政事业的不断发展和创新，更好的适应未来业务的变动和各类信息化的挑战，系统要具有延伸、拓展、不断优化服务的功能。

7、无缝对接功能

制定与国家民政部、省民政厅相关的信息化系统进行系统对接的可行方案，同时与社保、公安、卫生、房管、车管、社区等部门信息化系统数据接口的可行方案，做到“数据一次采集、一源多用”， 避免同一事务，多套系统重复管理，实现相关系统间的数据自动同步，提高数据的对内、对外交互共享能力。

9 结论

机器学习的一个重要应用就是文档的自动分类，比如一封电子邮件、新闻报道、用户留言、政府公文等。在文档分类中，比如一封电子邮件就是一个实例，而电子邮件中的某些元素(词语)则构成特征。我们可以观测文档中出现的词，并把每个词的出现或者不出现作为一个特征，这样得到的特征数目就会跟词汇表中的词目一样多。

伴随着大数据、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。本文采用 TF-IDF 算法、k-mean 聚类实现热点问题的筛选。由分析结果可见居民小区日常生活遇到的问题集中包括环境、噪音、房屋的问题。定义热度指标能更快筛选出最紧迫的问题。

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，以及通过文本挖掘构建指标，通过对评价的量化来衡量智慧政务中群众所反映的问题能否得到回复以及解决方案。对于群众意见的答复评价能够有效地提升政府智慧服务水平，有效预测民众所需要的服务，而且有针对性的解决民众的实际问题，引导群众办理有关事项。

10 参考文献

[1] TAHANIH KELLER JM. Infomation fusion in computer vision u-sing the fuzzy integral [J] IEEE Transactions on Systems, Man and Cybemetics, 1990, 20 (3) : 733-741.

[2] 黄萱菁，吴立德 独立于语种的文本分类方法 [期刊论文] －中文信息学报 2000 (06)

[3]徐文海,温有奎.一种基于 TFIDF 方法的中文关键词抽取算法[J],情报理论与实践,2008, 31 (2) :298-302.

[4] Kim B, Koyejo O, Khanna R, et al. Examples are not enough, learn to criticize! Criticism for Interpretability[C]. neural information processing systems, 2016: 2280-2288

[5] 杨超.贝叶斯统计在文本挖掘的若干研究.中国博士学位论文全文数据库.2019

[6] 会议论文.基于词频统计分析国内外文本挖掘的研究热点.第十二届（2017）中国管理学年会论文集.2017