基于 TF-IDF 算法、LDA、AHP 模型的“智慧政务”文本分析

摘要

目前大数据时代飞速发展，不论是各大企业、中小型企业还是政务部门等等各行各业都应对着相当庞大数据信息的处理。那么如何在新时代背景下，紧跟时代步伐，如何利用现有技术力量，高效有力地处理数以万计的数据……这些问题已然成为各行业、各部门必备的“救生圈”。

本题基于促进政府部门更好地了解民意、服务民生，站在一切为了人民的立场上，将人工智能、大数据挖掘与处理等自然语言处理技术投用于智慧政务系统，进行相关数据的挖掘，以减轻政务相关人员工作压力，并能够有效地处理群众留言信息以及政务部门更好地将群众反映的意见和建议做出合理回应。

针对问题一，在本题所体现的数据挖掘过程中，我们首先结合利用 Python、MATLAB 等计算机工具进行数据预处理、文本中文分词、停用词过滤以及统计文本词频等进一步筛选有效数据，实现对原始数据的进一步优化，并将其进行整合，利用 TF-IDF 算法、LAD 模型对群众的留言内容进行合理分类，进一步结合 F-score 模型对分类方法，计算最终分类模型的精确度，确立精确度 65%以上为合理分类模型。

针对问题二，我们将附件三中给出的数据在利用 Python 强大的数据处理功能对数据进行数据预处理、文本去重、中文分词、脱敏操作、停用词过滤以及词频统计并绘制词云图的操作下，将统计出的词频进行排序，结合题目分析，抽取词频中频数高于 160 的词语进行留言内容的进一步筛选，进而筛选出含有热度高词语的留言内容中留言条目数排名前五的留言作为热度留言，并导出为 Excel 文件。

针对问题三，同理将附件四给出的数据经过数据预处理等操作之后，建立适合本题的 AHP 模型，针对答复意见的相关性、完整性、可解释性以及答复的及时性的四项影响因素，确定各项权重、构造判断矩阵,通过利用 MATLAB 对矩阵的特征值及特征向量计算以及在进行归一化处理后，计算得出 CI，RI 值，进行随机一致性检验，若 RI 值小于 0.1，则通过检验，本题建立的模型具有合理性。将所求得的特征向量作为权向量，分别对应四个因素，最终确立针对答复意见的评价体系权重为相关性占 47.44%，完整性占 20.16%，可解释性占

6.92%，及时性占25.49%。并根据所构建的评价体系，对经过文本处理后的答复意见的数据进行多层次、多方面、全方位的合理分析，深入了解群众关注点，确立新阶段工作目标，更好地达到服务民生、满足群众日益增长的对美好生活的向往的工作宗旨。

关键词：数据预处理 TF-IDF算法 LDA模型 F-score模型 AHP模型
一致性检验 文本分析

Text analysis of "intelligent government affairs" based on tf-idf algorithm, LDA and AHP model

Abstract

At present, the era of big data is developing rapidly, and all walks of life, whether large enterprises, small and medium-sized enterprises or government departments, are dealing with a large amount of data information. So how to keep up with the pace of The Times in the new era, how to make use of the existing technology to efficiently and effectively process tens of thousands of data... These problems have become a necessary "life preserver" for all industries and departments.

Ontology based on promoting a better understanding of public opinion to government departments, services and people's livelihood, standing on the position of everything for the people, such as artificial intelligence, data mining and processing of natural language processing technology for wisdom e-government system, and data mining, to alleviate the government affairs related personnel working pressure, and can effectively deal with the message information and government departments to better reflect the opinions and Suggestions to make a reasonable response to the masses.

For question one, in the process of ontology in data mining, we first combined use of Python and MATLAB computer tools for data preprocessing, text in Chinese word segmentation, stop words filtering and word frequency statistics text further screening data effectively, realize the further optimization of the original data, and integrate it, using TF - IDF algorithm, the LAD model for the message content carries

on the reasonable classification of the masses, further combining with F - score model of classification method, calculate the accuracy of the final classification model, establish the accuracy of more than 65% for reasonable classification model.

In view of the problem, we will be given in annex 3 of the data in the use of Python is a powerful data processing function to data preprocessing, text data to heavy, Chinese word segmentation, desensitization, stop words filtering operation and word frequency statistics and draw the word cloud, under the action of will be ordered by the statistics of word frequency, subject analysis, extracting frequency higher than 160 words in the word frequency further screening of message content, and then select message content in message entries containing high heat words message as heat message number in the top five, and export to Excel file.

For question 3, the same will be presented attached four data after data preprocessing after operation, to establish a suitable for kinds of AHP model, according to the opinions of reply correlation, integrity, interpretability and reply to the timeliness of the four factors, to determine the weight, to construct judgment matrix, by using MATLAB to matrix eigenvalue and characteristic vector calculation and after normalization processing, calculated CI, RI value, random consistency inspection, if the RI value is less than 0.1, by inspection, ontology has established model is reasonable. The obtained feature vectors were taken as weight vectors, corresponding to four factors respectively.

Finally, the weight of the evaluation system for replies was determined to be 47.44% relevance, 20.16% completeness, 6.92% interpretability and 25.49%

timeliness. According to the constructed evaluation system, the author makes a multi-level, multi-dimensional and all-dimensional reasonable analysis of the data of the replies after text processing, so as to deeply understand the concerns of the masses, establish the work objectives in the new stage, and better achieve the work purpose of serving the people's livelihood and satisfying the growing aspirations of the masses for a better life.

Key words: data preprocessing tf-idf algorithm LDA model f-score model AHP modelConsistency checking text analysis

目录

1. 挖掘目标……7

2. 问题一分析流程、具体内容操作……7

2.1. 问题一分析流程图……7

2.2.问题一具体分析过程及操作……8

2.2.1. 附件数据介绍……8

2.2.2. 群众留言数据预处理……8

2.2.3.群众留言文本内容去重……9

2.2.4.群众留言内容去敏操作……9

2.2.5.群众留言内容中文分词……11

2.2.6.群众留言内容停用词过滤……11

2.2.7. 群众留言词频统计……12

2.2.8.TF-IDF 算法…… 13

2.2.9.建立 LDA 模型对群众留言内容分类 ……14

2.2.10. 利用 F-Score 对分类方法进行评价……18

3. 问题二分析方法与过程……21

4. 利用 AHP 模型建立答复意见的评价方案……22

5.结论……29

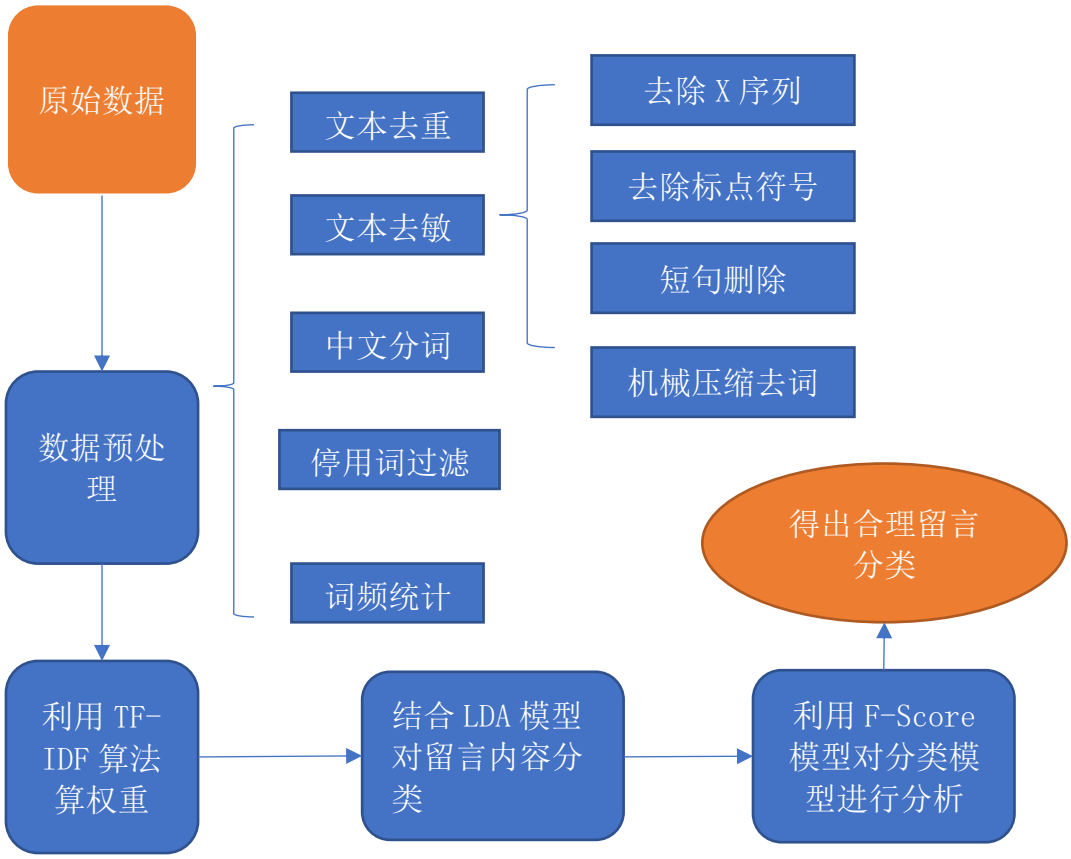
6.参考文献……29

1. 挖掘目标

此次建模针对政府通过微信、微博、市长信箱、阳光热线等平台收集各类社情民意进行文本分析、数据挖掘。通过对数据的预处理、对文本数据中文分词、停用词过滤、文本去重、机械压缩去词、脱敏操作、词频统计等处理后，利用 TF-IDF、LDA 模型、F-Score 评价模型、AHP 评价模型等数据挖掘模型，将附件中所给出的文本数据加以分类并检验分类是否合理，以及利用 Python、MATLAB 等计算机工具对政府的答复意见建立有效的评价方等。通过以上相关对数据及文本分析方法及过程，将大数据处理与挖掘充分利用，在大数据日益发达的当今，更好地提高政府相关部门的工作效率，减轻相关人员的工作压力，以及更好地获取对政府了解民意、汇聚民智、凝聚民气有价值、有意义的群众留言记录，达到政府更好地服务群众的宗旨。

2. 问题一分析流程、具体内容操作

2.1. 问题一分析流程图



本题的分析流程可大致分为以下四步：

第一步：获取附件中全部的分析数据，部分数据需自行爬取。

第二步：将所获得的数据进行基本的预处理。

第三步：针对处理后的数据依据 TF-IDF、LDA 模型将留言内容进行分类。

第四步：利用 F-Score 评价模型对留言分类方法进行评价。

2.2.问题一具体分析过程及操作

2.2.1．附件数据介绍

本题所给数据，来源于微信、微博、市长信箱、阳光热线等众多网络问政平台中全国群众对政府工作的留言，包括留言编号、留言用户、留言时间、留言主题、留言详情、政府答复意见、赞同与反对数等数据，总体来说，数字信息化飞速发展的时代，通过网络各大平台已然成为政府获取非常大量的群众反馈意见及建议，且在各大平台言论用词合理化的管理体制下，高价值、高水准、高质量的留言条目是政府服务民生、了解民意的重要渠道。

2.2.2．群众留言数据预处理

获取全部题目及数据后，首先要将本题附件中给出的全部数据进行预处理，附件二、附件三、附件四中或多或少存在垃圾数据（含金量较低甚至毫无价值的留言条目），若不将数据进行预处理，直接将 Excel 文件导入 Python 中进行分词、统计词频、停用词过滤、短句删除、机械压缩去词、去标点、去敏操作甚至进行后续的情感倾向性分析，得到的结果必然会造成分析结论的不可靠性，甚至是存在相当多的问题。所以，在进行下一步细致性分析之前，必须将数据进行预处理，去除大量对本题终极目标毫无利用价值、参考价值的数据条目。

我们通过 Python 对这些数据文本的预处理主要有这三大块：文本内容去重、文本词频统计及文本分词、根据文本内容各自分析的处理方法，对文本留言内容进行数据的预处理。

2.2.3.群众留言文本内容去重

①文本去重的基本定义

文本去重，顾名思义，去掉文中具有重复的内容，不论是在处理文本内容，还是处理数据条目时，但凡遇到重复的部分，均将其去除。

②文本去重的必要性

从留言者角度分析，为增强其留言内容的可读性，吸引政府部门相关人员的注意力，留言者将尽可能多的重复性的使用相同词语，自故性地增加留言内容字数、篇幅，其目的在于能够引起政务部门注意，来解决其相关问题。

从数据挖掘建模角度分析，若将文本不经过去重处理，必然会因许多重复性词语的出现，造成所构建模型与相关数据内容间的不相容，甚至造成所构建模型根本不具有可行性，其分析结果必然会与期望背道而驰。

从政务部门角度分析，相关人员在回复群众留言内容时，必然在数以万计的留言内容中遇到内容含义大致相同的留言，为方便、节约时间而言，相关人员很可能会将留言内容含义大体相同的回复采取复制粘贴的形式，这样便造成政府部门回复内容的重复性。

经过以上分析，文本去重，将是在分析处理文本、数据时，必然存在且非常重要的一个过程。

③本题针对文本去重所采用的的方法及原因：

本题开始所给出的所有文本数据，在文本去重过程中，很可能会误删掉对政府部门回复时有价值、有意义的数据，那么，我们就需要考虑既简单又不会去除有意义文本的去重方法。很多情况下，由于语义相近的词可能存在大多数为有用的内容，去除这类语义相近的词显然不现实，那么我们只能采取去除语义、词语完全相同的内容，对于去除完全相同的词语，显然方法很简单，利用Python，经过词语间的两两对比，逐一去除便可。

2.2.4.群众留言内容去敏操作

①去敏操作的基本定义

去敏操作，即对相关词、文本数据进行去除操作，本题中，不论在处理留言内容，还是在处理数据挖掘中，均要将其中类型词去除。

②去敏操作的必要性

本题中去敏操作包含去除X序列、去除标点符号、短句删除、机械压缩去词。

针对X序列：在附件给出的文本数据中，对敏感数据进行了脱敏操作，因而生成众多X序列。而这些X序列在我们进行词频统计以及在模型计算过程中

将会造成很大的影响，造成结果的不合理，因此，我们要对附件二中给出的留言内容进行 X 序列的删除操作，来保证下面所进行的数据操作以及根据数据所建模型的合理性。

例如：我是某小区业主，我叫 XXX，在 XXXX 年 XX 月 XX 日，我小区内出现……此句中出现很多 X 序列，在进行词频统计以及模型构建过程中，势必会对结果造成影响，因此我们采取全去除的操作。

针对标点符号：标点符号是我们在书写文本时所必须用到的，但在对数据挖掘、文本处理的过程中，标点符号的多样性将会影响整个分析过程，因此，在大多数情况下，我们将会对标点符号进行去除操作，将原文本中标点符号所在位置，替换为同等字节长度的空格，以保证各词语间的间隔，此操作会极大的便利我们对数据挖掘的整个过程，同时也会使文本数据分析更可靠，更鲜明。

例如：嗨！大家好。我是某小区业主，在这里我感谢我的小区物业给我们的生活环境带来树木、花草、喷泉……在该句中，留言者使用到“！”，“。”“、”“……”些标点符号，在进行数据处理时，我们要将其去除，替换为空格，结果为：嗨 大家好 我是某小区业主 在这里我感谢我的小区物业给我们的生活环境带来树木 花草 喷泉 这样便将一句话分为多个词语，以便词频统计操作。

针对短句删除：从我们日常生活中可以看到，若想明了清晰地表达所要表达的内容，都免不了短句的使用，例如在评价一件事物时，会用到“很不错”“很好”“挺好的”等一系列短句，显然这些短句是几乎没有利用价值的，在文本数据处理过程中，我们要将只是由这一系列短句构成的留言内容进行全部删除。

针对机械压缩去词：譬如一个人要表达某件物品非常好时，会用到“好好好好好好好好好”“非常非常非常好”等一些重复字词，还有“为什么为什么为什么为什么物业费这么贵”等等一系列语句，我们在进行数据分析过程中，只需保留重复字词中的一个即可，多余部分必须将其删除。去掉这些毫无意义的字词，但有些情况下，重复字是不可以删掉的，例如“赵师傅正在滔滔不绝的给我讲述这件物品使用时的注意事项”，在此句中，“滔滔”显然不可以删掉其中任意一个，若删掉，语义将会发生改变，因此我们只对语句开头以及结

尾的连续重复字词进行机械压缩去词处理，使得词频统计数据更有效，模型构建更合理。

2.2.5.群众留言内容中文分词

①分词的基本含义

在我们日常生活中可以发现，在我们所使用的语句中，只有字、句、段落能够通过一些分隔符进行分割，但是“词”和“词组”在形式上，不存在分隔符。分词就是将中文语句切割为单独的词组，英文分词是使用空格来区分词组，而中文我们就用使用 Python 中特定的库来进行词组的切割。

②分词的必要性及操作方法

我们需要做的是将群众留言进行分类，既然要分类，所以一定要将群众留言详情进行分词，我们需要用到具体的一些词来体现整个留言的中心思想与特征词。因为中文的单独的字和词意义不相同，所以我们需要选取 Python 中特定的库来进行中文字词的切割——Jieba 库，我们选取 Jieba 库的原因主要是其具有的易用性、准确率、性能对文本数据处理是很友好的，Jieba 中文分词组件，可用于中文句子、词性分割、词性标注、未登录词识别，支持用户词典等，该组件的分词精度达到了 97%以上，所以我们选用 Jieba 库来进行切割。

2.2.6.群众留言内容停用词过滤

①停用词过滤的定义

显而易见，停用词过滤就是将停用词进行删除，是为了在信息检索过程中提高操作效率在处理文本之前过滤掉某些字和词的操作，来达到提高我们数据处理的效率的目的。

②停用词过滤的必要性及操作方法

由于每个留言详情中的内容繁杂，对最后分类模型没有用的标点字母比比皆是，所以我们可以通过去除操作可以删去没有意义的内容，而不仅仅标点字母没有意义有些停用词比如：“并且”，“不但”，“除此之外”，“等等“，”换句话说“，”即使“，”哼“像上述这些词都是较为常见的无意义的词组，对于后续的操作也无任何意义，某些群众留言内容也只是为了来抒发自己的不满情绪与期待政府的快速解决。为了提高操作效率，我们将停用词进行去除。

我们从网站上查找了停用词列表，通过比较，我们利用一个比较全面的停用词文档——stoplist.txt。但遗憾的是，停用词表并不能完全将我们希望去除的内容清除掉，例如：“A3 区”，“A 市”，“C 市”，等同样也都是不需要的部分，是没有必要存在的字符，但是我们的 stoplist.txt 文档中并不包含某些我们希望同样被清除的词。

所以我们加入自定义的停用词——['会',' ','A','B','C','K',',',',','！',','年','月','日','\n','\t',' ']，通过一系列停用词过滤的操作，我们清除了没有必要的一些字符，这样便达到了对数据过滤的操作，便于我们模型的建立。

2.2.7. 群众留言词频统计

①词频统计的定义

词频统计，即统计整个文本数据文件中所有词语的出现次数，将所有标点符号全部替换为空格,将所有大写转换为小写，生成单词列表，生成词频统计并。排序，排除语法型词汇，代词、冠词、连词，输出词频最大 TOP100。

②词频统计的必要性及操作方法

根据题目所给出的附件中体现的数据，可以看出留言内容条目数的庞大，因此我们基于上文所述的文本去重、文本去敏、文本分词、停用词过滤等前期的数据处理，将处理后的数据进行词频统计，并将其词频数进行降序排序，获取词频排名前 100 的词语，将这些词语所在留言文本内容确定并提取，将文本处理成词的集合，记为 d=（ω_1，ω_2，ω_3，ω_4，……ω_N），其中 N 为文本数据 d 中出现词语的个数，为模型构建建立数据基础。

2.2.8.TF-IDF 算法

在对群众留言内容进行数据预处理之后，进一步将这些词语转化为向量，供我们进行数据挖掘使用，本题首先采用 TF-IDF 算法，将群众留言内容信息转化为权重向量，进行分析。

TF-IDF 算法的基本原理如下：

① 根据上文词频统计的结果，作为 TF 权重。

词频（TF）=该词语在文本内容中出现的次数　　（1）

由于留言内容字数参差不齐，为了便于对群众留言内容的比较，这里将“词频”标准化，标准化后的词频表示为该词在文中出现的次数除以该文本内容的总词数，或者该词在文中出现的次数除以文中出现次数最多的词的出现次数，用式子表示为：

$$\text{词频（TF）}=\frac{\text{该词在文本中出现的次数}}{\text{该文本的总词数}} \quad (2)$$

或

$$\text{词频（TF）}=\frac{\text{该词在文中出现的次数}}{\text{该文本出现次数最多的词的出现次数}} \quad (3)$$

②计算 IDF 权重，即计算逆文档频率，在此过程中，我们需要自行构建一个语料库，其用来模拟留言内容的留言环境。IDF 越大，此特征性词语在文本数据中的分布越集中，说明该词语在区分该文本数据内容的属性能力越强。

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{该语料库的词语总数}}{\text{包含该词的文本内容数}+1}\right) \quad (4)$$

③计算 TF-IDF 值,根据实际的合理分析得出，TF-IDF 值与某个词在群众留言文本数据中出现的次数成正比例关系，某个词在文本中的重要性越高，TF-IDF 值越大。计算群众留言内容文本中每个词的 TF-IDF 值，并按照计算结果值进行排序，次数最多的即为要提取的群众留言内容文本的关键词。TF-IDF 计算公式如下：

$$\text{TF-IDF}=\text{词频（TF）}\times\text{逆文档频率（IDF）} \quad (5)$$

④生成 TF-IDF 向量。我们通过使用 TF-IDF 算法，能够找出每条群众留言内容的前五个关键词；（2）提取在每条群众留言内容中提取的五个关键词，将所有关键词组成一个集合，计算每条群众留言内容对于上述集合中所有词的词频，如果没有，则标记为 0； （3）其次再生成每条群众留言内容文本的 TF-IDF 权重向量，其计算公式如下：

$$\text{TF-IDF}=\text{词频（TF）}\times\text{逆文档频率（IDF）} \quad (6)$$

2.2.9.建立 LDA 模型对群众留言内容分类

在生成群众留言内容的TF-IDF权重向量后，我们可以根据每个留言内容的TF-IDF 权重向量，结合利用LDA模型将群众内容分为七类：城乡建设(0)、环境保护(1)、交通运输(2)、教育文体(3)、劳动和社会保障(4)、商贸旅游(5)、卫生计生(6)。

LDA模型分类的原理如下:LDA模型是通过无督视的学习方法来发掘所研究的文本中深层隐含的相关信息，该模型以无指导学习的途径从所研究的文本中发现隐含语义维度。隐性语义分析的实质为利用文本中字词的共性来发掘文本结构。LDA模型中涉及到贝叶斯模型离不开“先验分布”，“数据（似然）”和"后验分布"三部分。我们知道在LDA模型贝叶斯理论中：

先验分布+数据（似然）=后验分布

在上述该式中，例如当事人对好人和坏人的认知，这里假设先验分布为：100个好人和100个坏人，即好人坏人各占总体的50%，若在某一刻当事人被2个好人帮助了、被1个坏人欺骗了，于是当事人便得到新的后验分布，在此情境下好人数量增加为102个，坏人数量变为101个。此时我们可以看出102＞101，即好人数量多于坏人数量。在新先验分布基础上，当事人在遇到被1个好人帮助了和3个坏人欺骗之后，再次更新先验分布：此时好人数量为103个，相较于前一次先验分布增加一位好人，坏人数量为104个，相较于前一次先验分布增加了3位。如此一直发展下去，我们将得到更多的先验分布。对于我们的数据（似然）即类似于二项分布，其表达式为：

$$Binom(k|n,p)=\binom{n}{k}p^k(1-p)^{n-k}$$

该式中，p我们可以理解为是好人的概率，k为好人的个数，n为好人坏人的总数。对于先验分布，就像上述例子中“102个好人和101个坏人”，它是前一次贝叶斯推荐的后验分布，又是后一次贝叶斯推荐的先验分布，换句话说即先验分布与后验分布形式相同，一般叫共轭分布。在概率分布中，二项分布共轭的分布为Beta分布。Beta分布的表达式为：

$$Beta(p|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

其中是Γ是 Gamma 函数，满足 $\Gamma(x)=(x-1)!$通过仔细观察 Beta 分布和二项分布，可以发现两者的密度函数很相似，区别仅仅在前面的归一化的阶乘项。

后验分布：

$$P(p|n,k,\alpha,\beta)=\binom{n}{k}p^k(1-p)^{n-k}\times\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

将其经过归一化处理后，我们得到的后验分布概率为：

$$P(p|n,k,\alpha,\beta)=\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+k)\Gamma(\beta+n-k)}p^{k+\alpha-1}(1-p)^{n-k+\beta-1}$$

我们通过观察可以看出后验分布为 Beta 分布，且

$$Beta(p|\alpha,\beta)+BinomCount(k,n-k)=Beta(p|\alpha+k,\beta+n-k)$$

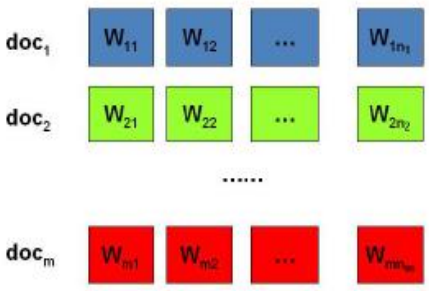
在二维情形下我们使用了 Beta 分布和二项分布来表达这个模型，以此类推，在 N 维情形下，我们可以用 N 维的 Beta 分布来表达先验后验分布，N 项的多项分布来表达数据（似然）。例如在三维情形下，对于三项多项分布的表达，假设数据中的第一类有 m_1 个好人，第二类有 m_2 个坏人，第三类为 $m_3=n-m_1-m_2$ 个不好不坏的人,对应的概率分别为 $p_1, p_2, p_3=1-p_1-p_2$，则对应的三项多项分布式为：

$$multi(m_1,m_2,m_3|n,p_1,p_2,p_3)=\frac{n!}{m1!m2!m3!}p_1^{m1}p_2^{m2}p_3^{m3}$$

维度超过二维的 Beta 分布我们一般称之为 Dirichlet 分布。也可以说 Beta 分布是 Dirichlet 分布在二维时的特殊形式。在三维情形下，分布式如下：

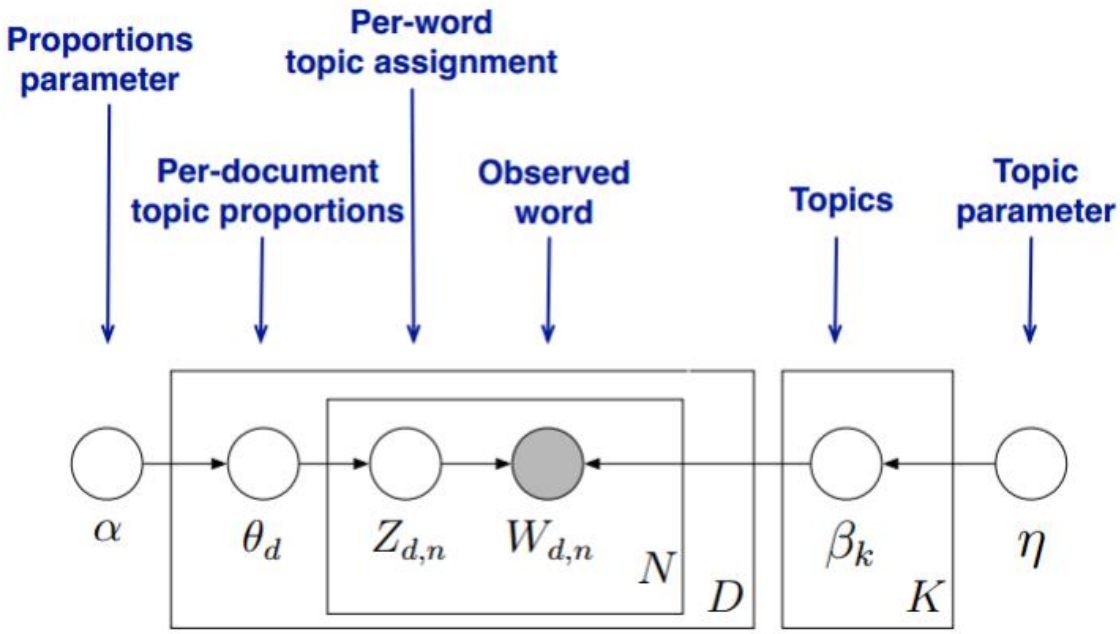
$$Dirichlet(p_1,p_2,p_3|\alpha_1,\alpha_2,\alpha_3)=\frac{\Gamma(\alpha1+\alpha2+\alpha3}{\Gamma(\alpha1)\Gamma(\alpha2)\Gamma(\alpha3)}\ p_1^{\alpha1-1}p_2^{\alpha2-1}p_3^{\alpha3-1}。$$

针对本题中所建立的 LDA 模型，我们有 M 篇文本内容，对应第 d 个文本内容中有 N_d 个词，输入过程矩阵如下图所示：



我们的目标是找到每一篇文档的主题分布和每一个主题中词的分布。在 LDA 模型中，我们需要先假定一个主题数目 K，这样所有的分布就基于这 K 个主题展开。

LDA 模型具体流程如图所示：



LDA 模型假设文档主题的先验分布是 Dirichlet 分布，即对于任一文档 d，其主题分布 θ_d 为：

$$\theta_d = \text{Dirichlet}(\vec{\alpha})$$

其中，α 为分布的超参数，是一个 K 维向量。LDA 模型假设主题中词的先验分布是 Dirichlet 分布，即对于任一主题 k，其词分布 β_k 为：

$$\beta_k = \text{Dirichlet}(\vec{\eta})$$

其中，η 为分布的超参数，是一个 V 维向量。V 代表文本内容中字词的总个数。对于数据中任一一篇文档 d 中的第 n 个词，我们可以从主题分布 θ_d 中得到它的主题编号 z_{dn} 的分布为：

$$z_{dn} = \text{multi}(\theta d)$$

而对于该主题编号，得到我们看到的词 w_{dn} 的概率分布为：

$$w_{dn} = \text{multi}(\beta_{zdn})$$

在这个模型里，我们有 M 个文档主题的 Dirichlet 分布，而对应的数据有 M 个主题编号的多项分布，这样(α → $\theta_d \rightarrow \vec{z}d$)就组成了 Dirichlet-multi 的共轭，可以使用前面提到的贝叶斯推断的方法得到基于 Dirichlet 分布的文档主题后验分布。如果在第个文档中，第 k 个主题的词的个数为：$n_d^{(k)}$,则对应的多项分布的计数可以表示为：

$$\vec{n}_d=(n_{(1)}^d,n_{(2)}^d,...n_{(K)}^d)$$

利用 Dirichlet-multi 共轭，得到 θ_d的后验分布为：

$$Dirichlet(\theta d|\vec{\alpha}+\vec{n}_d)$$

同理，对于主题与词的分布，我们有 K 个主题与词的 Dirichlet 分布，而对应的数据有 K 个主题编号的多项分布，这样(η →$\beta_k \rightarrow \vec{\omega}_k$)就组成了 Dirichlet-multi 共轭，如果在第 k 个主题中，第 v 个词的个数为:$n_k^{(v)}$，则对应的多项分布的计数可以表示为

$$\vec{n}_k=(n_{(k)}^1,n_{(k)}^2,...n_{(K)}^V)$$

利用 Dirichlet-multi 共轭，得到 β_k的后验分布为:

$$Dirichlet(\beta_k|\vec{\eta}+\vec{\eta}_k)$$

由于主题产生词不依赖具体某一个文档，因此文档主题分布和主题词分布是独立的。这里所讲到的 M+K 组 Dirichlet-multi 共轭，也就是我们所讲述的 LDA 的基本原理。

2.2.10. 利用 F-Score 对分类方法进行评价

F-Score 模型介绍

假设我们将数据分为两层，自行规定其中一层数据为正样本，那么自然而然另一层数据为负样本，本赛题中给出的 F-Score 模型计算公式，为取 n 次数据，计算其算数平均数。

我们一般使用四个符号表示预测的所有情况:

- TP(真阳性):正样本被正确预测为正样本
- FP(假阳性):负样本被错误预测为正样本
- TN(真阴性):负样本被正确预测为负样本
- FN(假阴性):正样本被错误预测为负样本

F-Score 模型计算公式如下：

$$F-Score=(1+\beta^2)*\frac{Precision*Recall}{(\beta^2*Precision)+Recall} \quad (1)$$

其中 P（Precision）表示查准率（精确率），其含义预测为在可能含有负样本的数据中,真实正样本所占总样本的比例；计算公式为：

$$Precision=\frac{TP}{TP+FP} \quad (2)$$

其中 R（Recall）表示查全率（召回率），其含义预测为在不包含任何负样本的正样本数据中,正确预测的比例；计算公式为：

$$Recall=\frac{TP}{TP+FN} \quad (3)$$

其中 β 是用来平衡 P（Precision）,R（Recall）在 F-score 计算中的权重,取值情况有以下三种:

- β=1,表示 P(Precision)与 R(Recall)一样重要
- β＜1,表示 P(Precision)比 R(Recall)重要
- β＞1,表示 R(Recall)比 P(Precision)重要

在本题中,令 β=1，此时认为两个指标一样重要，此时的 F-score 计算公式为:

$$\boldsymbol{F-Score=\frac{2*Precision*Recall}{Precision+Recall}} \quad (4)$$

抽取 n 次数据后，计算 F-Score 值取平均，此时计算公式为：

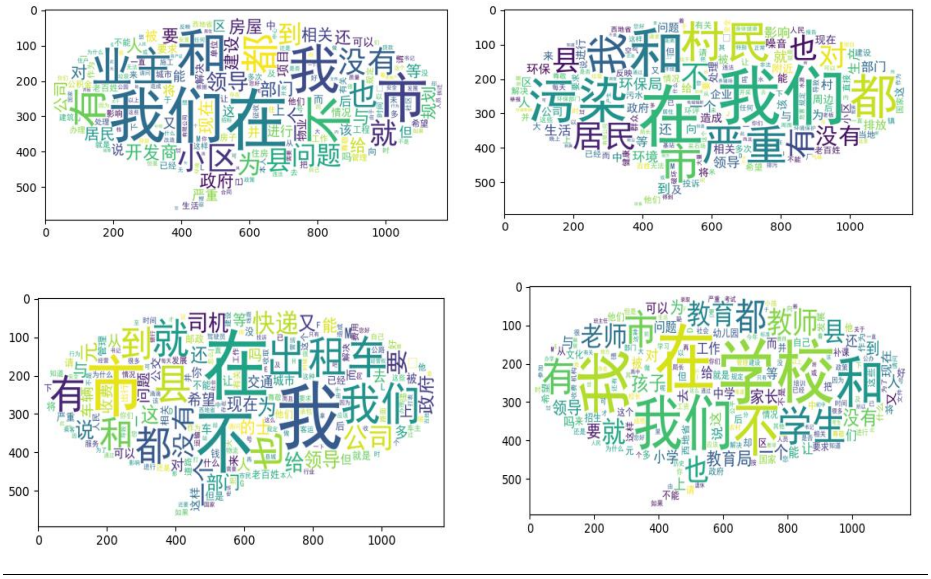
$$\boldsymbol{F1\text{-}Score=\frac{1}{n}\sum_{n=1}^{n}F-Score} \quad (5)$$

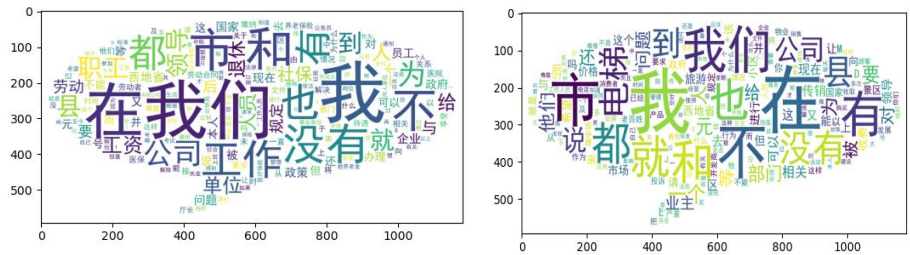
利用 F-Score 模型计算所得结果无非就是合理分类与不合理分类，那么本题中，我们规定将 65%作为两种计算结果的分界线，若计算值大于 65%，那么该分类模型具有合理性，若计算值小于 65%，那么该分类模型不具有合理性。

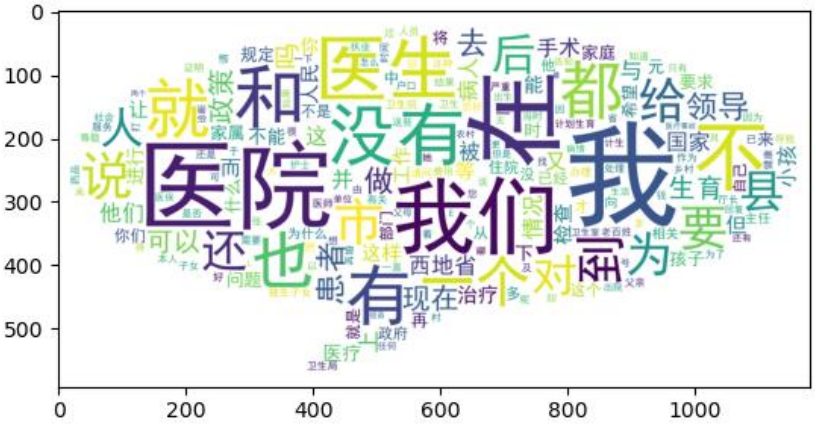
2.3.本题总体结果分析

用 Python 实现 TF-IDF 加权，如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，这样的词条应该给它们赋予较高的权重，并选作为该类文本的特征词以区别与其它类文档。于是我们先建立训练集样本（占总体 20%，随机抽取），获取训练集样本的 TF-IDF 权值向量，建立一个分类标准，最后对测试集样本（剩余 80%部分）获取权值向量从

而达到分类的目的。但在 Python 实现的过程中，训练集样本和测试集的特征维度不同，导致 Python 无法继续运行，这时我们可以让两个 CountVectorizer 共享 vocabulary，使得训练集和测试机的向量长度保持一致。下图为我们在将留言进行分类之后，根据各类词频依次所做出词云图分布。







我们在依照 TF-IDF 算法所计算各类权重后，分别代入 F-Score 模型进行拟合度的计算，下图为各类计算所得 F-Score 值，通过对 F-Score 值进行算数平均值计算所得最终 F-Score 模型计算得数为：0.6968527≈69.69%＞我们所规定的标准 65%，因此该分类体系合理。详细程序代码见附件。

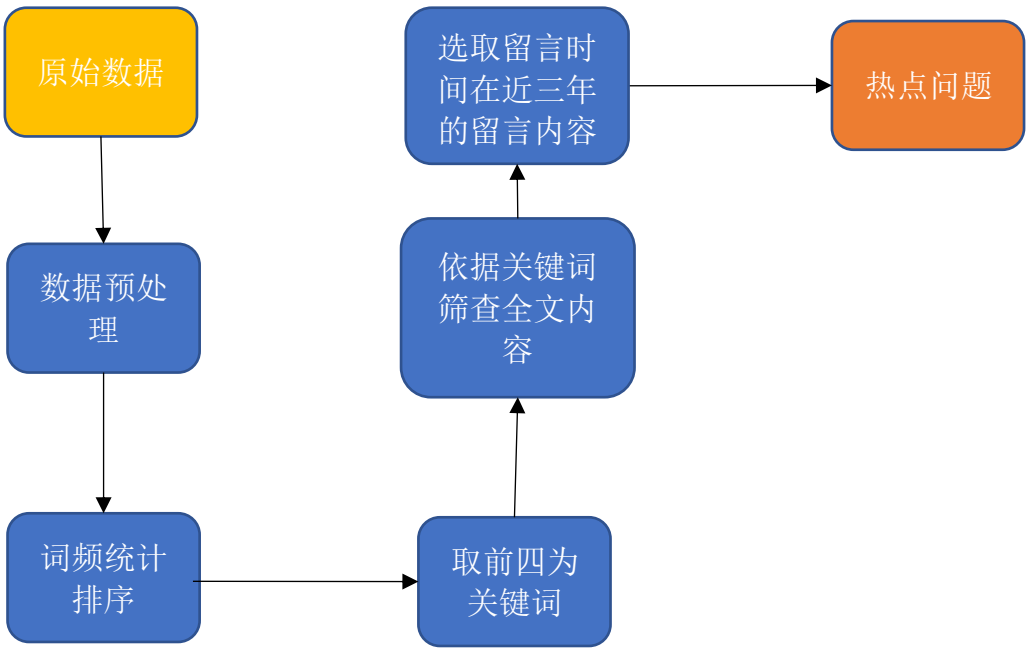
类别	precision	recall	F1-Score 值	support
城乡建设	0.5309	0.682	0.597	1607
环境保护	0.6484	0.3214	0.4298	728
交通运输	0.6923	0.6132	0.6504	457
教育文体	0.5825	0.2463	0.3462	1248
劳动和社会保障	0.4916	0.7103	0.5811	1554
商贸旅游	0.5914	0.4718	0.5249	970
卫生计生	0.6602	0.7488	0.7017	678

3. 问题二分析方法与过程

在我们获取到附件三留言内容时，大致浏览留言内容可以发现，其中存在很多条留言内容地址或相关问题具有极强的相似性甚至完全相同，例如：在同一小区内不同业主针对同一问题的留言等这类留言，在我们对留言内容进行分析时，将他们作为热点内容提取，及时发现热点问题，可以有效地提高政府部门的服务效率，有助于相关部门人员进行针对性的处理，同时，也会给具有相同抱怨的群众获得满足，进而提升政府在群众中的权威力度及口碑。

对于将热点问题提取并分类的问题，我们的评价指标为如下：

首先在进行数据预处理后的留言内容数据中，统计词频并排序，取排名前四的字词语，记为关键词。通过计算机语言实现对处理后的每条留言内容的关键词筛查，若存在多余一个关键词的留言内容，将满足此条件的留言内容作为热度排名为首的信息，其余根据关键词的排序筛选出相应的留言内容，其热度排序与关键词排序一致。分析流程图大致如下：



通过利用 Python，我们对附件 3 经过数据预处理后得到词频在 160 次以上的字词：{'A': 1819, '市': 1283, '的': 1023, '区': 931, 'A7': 662, '，': 633, '小区': 528, '问题': 492, '？': 460, 'A3': 433, '县': 350, '扰民': 265, 'A2': 259, 'A4': 236, '请': 233, '不': 228, '在': 222, '了': 220, '严重': 209, 'A1': 207, '街道': 196, '路': 191, '投诉': 191, '反映': 174, 'A5': 170, '咨询': 168, '建议': 163}，依照词频排序先后，再经过字词的整合，通过 Excel 对工作簿进行关键词筛选，我们得出排名前五的热点问题:（本处只体现关键词，具体见附表）①A 市乱收费现象②A 市众多学校③A 市违法现象④A 市溪湖周边⑤A 市多区麻将馆。针对本题所给出的热点问题见附件“热点问题表.xls”，详情见附件问题二中“热点问题留言明细表.xls”。

4. 利用 AHP 模型建立答复意见的评价方案

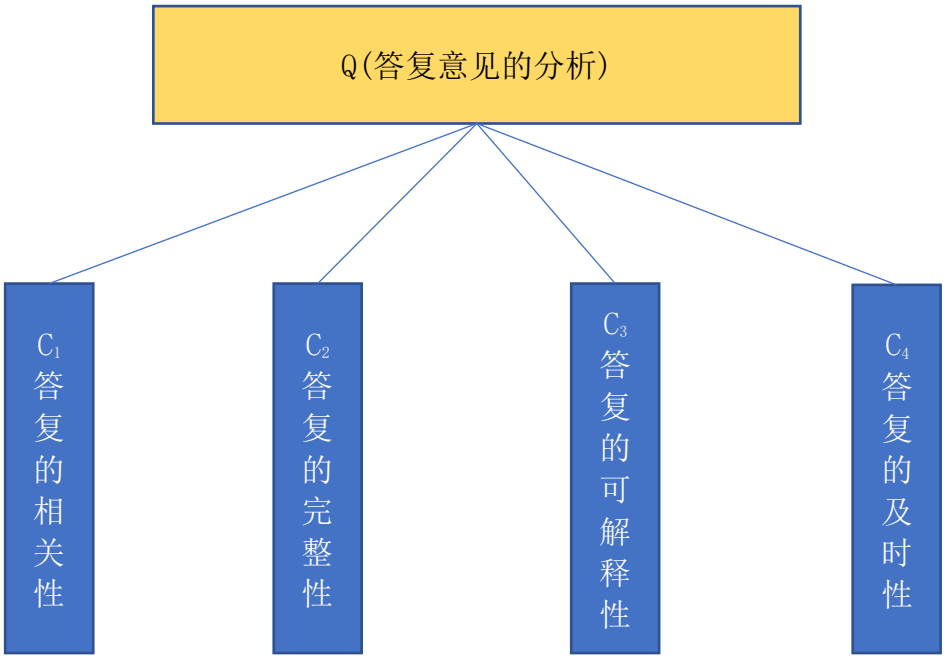
一、针对答复意见的综述

本题附件四中给出了针对群众留言政府相关部门人员给出的回应，我们可以通过统计在经过数据预处理之后的群众留言词频与政府回复词频，两者间的词频、字数对比，进而可以进一步分析出政府对群众呼声的回应度。政府部门

始终秉持着为人民服务的宗旨，通过微信、微博、市长信箱、阳光热线等多种网络平台来获取群众呼声，政府部门根据群众所反映的问题与群众对政府工作所提出的建议，表达出的支持，政府部门分别作出相应的回应，在原有的高度注重民生，服务民生，将工作投身到群众中去，为百姓服务的基础上，更加高效地开展工作，为进一步了解民意、汇聚民智、凝聚民气助力，增强政府在百姓群众间的威信，更好地做到为人民服务。

二．AHP 模型的建立、特征值特征向量的求解及模型的一致性检验

本题利用 AHP 模型对答复意见进行分析，影响答复意见的因素分别是：答复的完整性、相关性、可解释性和及时性。我们所建立的层次结构模型如图所示：



层次分析法所要解决的问题是关于最低层对最高层的相对权重问题，按此相对权重可以对最低层中的各种方案、措施进行排序，从而在不同的方案中做出选择或形成方案的原则。本题中，我们所构建的 AHP 模型的终极目标是对留言的答复意见分析，准则层为四个影响因素，接下来我们将构造判断矩阵，在确定各层次各因素之间的权重时，如果我们只给出定性的结果，则常常不容易被读者接受，因此我们采用一致矩阵法，即：将影响因素两两比较，采用相对尺度，尽可能减少性质不同的诸因素间相互比较的困难，来提高准确度。判断矩阵是表示本层所有因素针对上一层某个相对重要性的比较。

本题中判断矩阵 a_{ij} 采用 Santy 的方法给出：

标度	含义
1	表示两个因素相比，具有同样重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2,4,6,8	上述两相邻判断的中值
倒数	因素 i 与 j 比较的判断 a_{ij}，则因素 j 与 i 比较的判断 $a_{ji}=1/a_{ij}$

设各准则 C_1,C_2,C_3,C_4 对目标 Q 的重要性，C_i：$C_j \rightarrow a_{ij}$，$A=(a_{ij})_{4*4}$，$a_{ij}>0$，$a_{ji}=1/a_{ij}$，本题中，根据自我判断规定 C1：C2=3，C1:C3=5,C1:C4=2，C2:C3=3,C2:C4=1,C3:C4=1/5,写出判断矩阵 A，即

$$A=\begin{pmatrix} 1 & 3 & 5 & 2 \\ \frac{1}{3} & 1 & 3 & 1 \\ \frac{1}{5} & \frac{1}{3} & 1 & \frac{1}{5} \\ \frac{1}{2} & 1 & 5 & 1 \end{pmatrix}$$

在矩阵 A 中，我们利用 MATLAB 求最大特征根 λ 的特征向量，将其作为我们的权向量 ω，即 $A\omega=\lambda\omega$。对应于判断矩阵 A 的最大特征根 λ 的特征向量将其归一化处理后记为 ω，ω 中的元素为同一层因素对于上一层次因素相对重要性的排序权值，这一过程称为层次单排序，我们需要对 ω 进行一致性检验，确定 A 允许不一致的范围。λ 依赖于 a_{ij}，则 λ 比 n（本题中 n=4）越大，不一致性越严重，用最大特征值对应的特征向量作为被比较因素对上层因素影响程度的权向量不一致程度越大，引起的判断误差越大，因而我们利用 λ-n 的数值大小来衡量 A 的不一致程度。我们所使用的一致性指标公式为：

$$CI=\frac{\lambda-n}{n-1}$$

本题中 n=4，若 CI=0，则说明有完全的一致性；若 CI 接近于 0，则说明有一定的一致性，CI 越大，说明不一致性越严重。为衡量 CI 的大小，我们引入随机一致性指标 RI。指标表如下：

随机一致性指标 RI 表

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

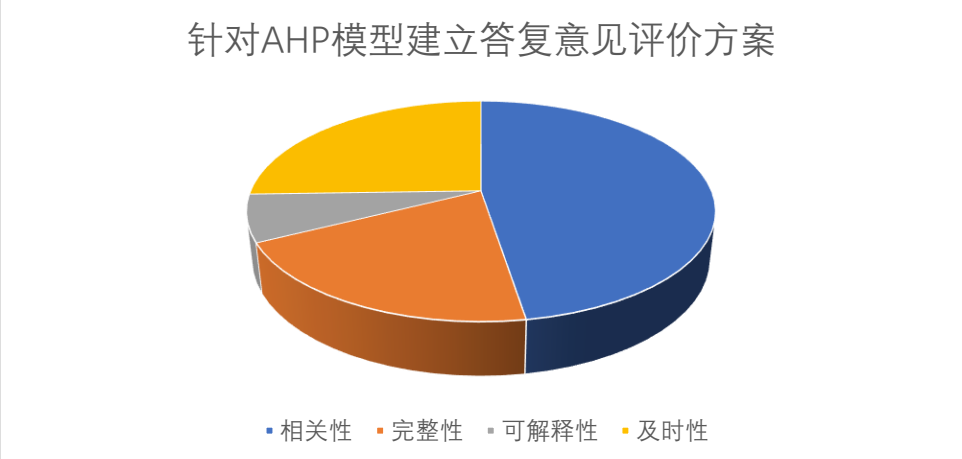
定义一致性比率：

$$CR=\frac{CI}{RI}$$

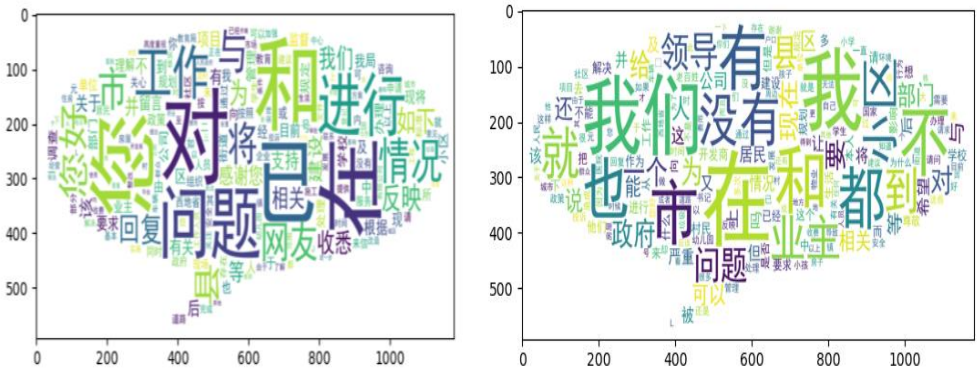
一般情况下，认为 CR＜0.1 时，A 的不一致性程度在允许范围内，通过一致性检验，否则对矩阵 A 进行调整，重新确定权向量。本题 n=4，RI=0.90

本题利用 MATLAB 对矩阵 A 进行最大特征值的计算，求得最大特征根 λ=4.0788.将 λ=4.0788 带入 CI 公式算得 CI=0.02627,将 CI 值带入 CR 得一致性比率 $CR=\frac{CI}{RI}$=0.02627/0.90=0.02919＜0.1 **一致性检验完成。**因此求得的最大特征值对应归一化后的特征向量 ω=(0.4744,0.2016,0.0692,0.2549)，即答复意见的相关性所占评价比重 47.44%，答复意见的完整性所占比重为 20.16%，答复意见的可解释性所占比重为 6.92%，答复意见的及时性所占比重为 25.49%。在我们的评价体系中所占权重最高的因素为答复意见的相关性，其次为答复意见的及时性，其次为答复意见的完整性，最后为答复意见的可解释性。

针对 AHP 模型建立答复意见评价方案	
相关性	47.44%
完整性	20.16%
可解释性	6.92%
及时性	25.49%



通过以上计算我们可以得出一套完整的评价方案体系，我们在针对留言的答复内容进行分析评价时，根据四部分因素所占整个留言的答复意见的评价体系比重，着重关注占比较高的因素，例如答复的相关性以及及时性、完整性。



① 答复意见的相关性：

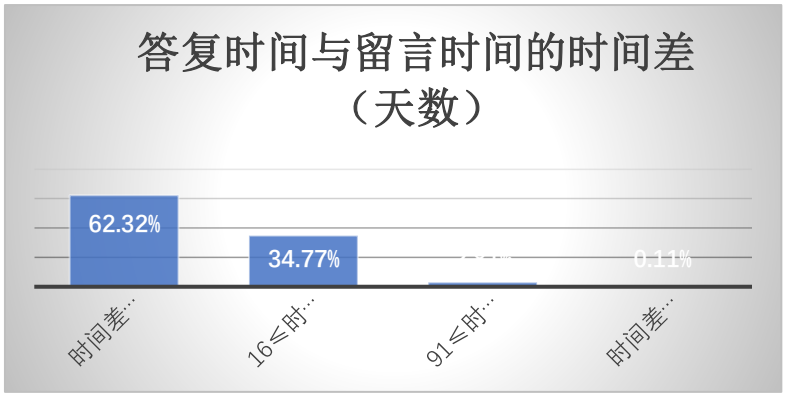
上图分别为群众留言内容的词云图以及政府留言内容的词云图，相关代码见附件。通过观察词云图以及统计的词频可以看到，在群众留言内容与政府相关部门的答复意见词云图以及词频统计表中，有多条重复字词，充分体现了在答复群众留言内容时，政府相关部门针对每条群众留言给予了各自的问题答复，不存在偏离群众留言内容的答复意见，不存在答非所问的情况，从答复意见贴合群众留言内容来看，体现了政府部门的相关工作人员的高度责任心，这也充分体现了政府工作一切为了人民，一切为了群众的宗旨。

② 答复意见的及时性：

在目前大数据时代背景下，信息量与日俱增，政府部门通过各大网络平台、市长热线等渠道收集群众留言，来更好地听取百姓呼声，但在积极广泛的收集留言内容下，出现政府工作人员紧缺的现象，导致一些留言内容未能在信息有效期内得到合理的回复，进而未能解决群众遇到的问题。尤其在信息化高速发达的今天，信息的时效性显得尤为重要，时效性意味着该信息在一定时间内具有效应，超过一定时间范围将失去信息本身的意味。因此，政府部门相关人员对群众留言内容的答复应当在短期合理时间内给出相应的回应，也就是说，在面临人员紧缺、信息剧增、保质保量的给群众解决问题的前提下，政府部门应当合理安排人员分工，保证将群众留言在合理时期内进行答复，本题附件四中给出的数据，附件四共 2816 条留言内容，回复时间差小于等于 15 天的有 1755 条，16—90 天的有 979 条，大于 90 天小于 365 天的 79 条，超过一年

的有 3 条，计算时差公式=DAYS(G:G,D:D)，我们通过 Excel 将政府答复时间与群众留言时间的时间间隔进行统计计算，得出以下统计图：

留言与答复时差（天数）	
时间差≤15	62.32%
16≤时间差≤90	34.77%
91≤时间差≤365	2.81%
时间差≥366	0.11%



通过统计图表我们可以看出，政府在答复群众留言时，有 62.32%的留言在十五天内得到回复，有超过 95%的留言内容都在半年内得到了回复，仅仅有一小部分不超过留言总数 3%的留言超过半年都未得到回复，以上数据可以充分说明，政府在处理这项工作时，积极调动相关人员，在合理时间内，给出高质量高效率的答复工作，这样有利于政府了解民意、汇聚民智、凝聚民气助力，增强政府在百姓群众间的威信。

③ 答复意见的完整性

简单来说，答复意见的完整性意味着，政府相关部门工作人员在回复群众留言时，是否对群众留言所涉及到的问题一一地进行了回复与解决。在针对答复意见完整性的合理分析时，我们将采取以小见大的方式进行分析，通过观察词频统计表，将词频进行降序排序输出后，我们对比留言内容与政府的答复意见词频表中词频在 50 次以下的字词，词频表见附件。通过对比，我们可以很清晰地观察到在群众留言内容中，即便只出现过一次的词语所对应的留言内容，政府部门相关人员都给予合理地回复。另一方面，我们也可以通过统计群众留言字数与政府答复字数对比，在将附件四中的数据经过分词、去重等操作处理之后，我们统计出群众留言字数最终为 84421 个字，政府答复意见为 63833 个

字，通过数据我们可以看出，政府答复重复性较大，因此在去重处理后，字数减少较多，例如附件四中部分答复内容如图所示：

答复意见
您好，你所反映的问题已转交相关单位调查处置。
您好，你所反映的问题已转交相关单位调查处置。

在面对一些具有敏感问题或个人隐私问题的回答时，政府的答复意见具有太强的概括性，且重复条数多，在经过数据处理之后，该内容只出现一次。去除概括性答复意见，政府对群众所提出来的社区等出现的问题都做了详尽的答复。

④ 答复意见的可解释性

可解释性是指在我们需要了解一件事情的时候，我们是否能够得到我们所必需获得的能够让我们理解的信息内容。例如我们在阅读社科性文章、专业学术论文时，遇到的专业性名词，我们需要通过查阅网络，通过网络中很宽泛很通俗的话语来理解专业名词，也比如在某个领域的科学研究中对一个新问题的探究，我们需要通过查阅浏览一些相关资料来了解这个问题的基本定义以及目前的研究现状，来获得我们对与该课题研究方向的正确认识。也就是说，如果在一些生活情境中我们无法得到必需的充足的信息，那么这些事物对于当事人来说就是不可解释的。如果我们对于某一事物了解不够充分，不够透彻，那么我们将会考虑到可解释性的问题。当然，不可解释同样也意味着危险，事实上政府在面对群众留言内容进行的答复意见时的顾虑除了政府部门无法给出足够的信息之外，也有一定的关于安全性的考虑，同时也包含了群众所具有的文化知识水平的限制。因此，政府部门相关人员在进行对群众留言内容的答复时，在避免使用专业性名词以外，更应通过一些通俗易懂却又不失得体的词语对群众留言进行合理地答复，更好地解决群众所提出的问题，这样也有利于政府部门更加高效地开展工作，进一步了解民意、汇聚民智、凝聚民气助力，增强政府在百姓群众间的威信，有利于更好地做到为人民服务。

5. 结论

对于“智慧政务”文本挖掘数据分析，进一步帮助政府了解广大群众的需求及所反映意见、建议，不论是对政府相关部门，还是对广大百姓群众都有着重大意义。目前随着大数据时代的到来，传统的文本解读方式已然不能满足数

据量庞大的网络等各大平台智慧政务的留言信息，本文分别采用了 TF-IDF 算法及 LDA 模型针对群众留言内容进行合理分类，并将各类结果带入 F-Score 模型中检验分类的合理性。并且通过建立 AHP 模型给出关于政府答复意见的评价方案，通过各影响因素的权重比例，深入分析政府答复意见的详细内容及其效用。

通过一系列数据处理后，我们通过词频统计了解到群众在日常生活中，更多的关注点落在各自周边对生活造成一定影响的事物，例如麻将馆非法聚集、xx 楼盘建设严重扰民等等，更多的也会给出政府在一些基础设施上的合理建议。由深入的数据分析可以看出，群众留言内容可以分为城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生七类。我们针对各类留言进行筛选，对词频进行降序排序，进而确定五大热点问题。

我们通过建立的 AHP 模型算出评价的相关性、完整性、可解释性、及时性所占整评价体系权重，得出评价相关性所占权重最高为 47.44%，由此可以看出，在政府部门针对群众留言进行答复时，所答复内容应更多的向群众留言主题靠拢，应严厉杜绝答非所问的现象发生。政府部门答复意见应进一步加强，做到相关性极强，及时性极高，非常完整的详尽的留言答复，来更好地提高政府相关部门的工作效率，有效地获取对政府了解民意、汇聚民智、凝聚民气有价值、有意义的群众留言记录，有助于提高在百姓群众间的威信，达到政府更好地服务群众的宗旨。

6.参考文献

1. 曹卫峰　中文分词关键技术研究.南京理工大学.硕士学位论文.2009
2. 博客网　分类模型的评价指标 F-score
3. 刘建平　文本主题模型之 LDA(一) LDA 基础
4. 博客网　TF-IDF 算法解析与 Python 实现
5. 张拳石　深度神经网络可解释性方法汇总