基于自然语言处理技术的智慧政务系统

摘要

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。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，对提升政府的管理水平和施政效率具有极大的推动作用。因此，对各类社情民意相关的留言和热点进行分析研究有着非常重要的意义。

针对问题一：首先基于附件 2 中的内容建立识别模型，准确地识别出内容对应的分类标签，再将非结构化数据进行去重去空、结巴分词 jieba（采用前缀词典实现高效的词图扫描、动态规划查找最大概率路径以及使用了 Viterbi 算法)及停用词过滤等数据预处理，然后基于 TFIDF 权重法提取候选特征词，形成词袋，建立高斯朴素贝叶斯模型，再利用处理后的训练集通过训练朴素贝叶斯模型，并由测试集进行分类得到模型分类结果。使用 F-Score 对这个一级标签分类模型进行评价，得到评分。

针对问题二：对预处理后的文本内容转换成文本-词条矩阵，运用 K-Means 聚类模型得出文本聚类的结果，选择其中排名前五的类别构成热点问题，筛选得出热度问题留言表，并根据在某段时间内某地方热点问题出现的频率，定义热度评价指标。

针对问题三：计算留言与相关部门对留言的答复意见的相似程度即相关性，将附件文本中词语，映射到向量空间，形成文本中文字和向量数据的映射关系，通过计算几个或者多个不同的向量的差异的大小，来计算文本的相似度。同时要知道其覆盖率（答复意见中对留言内容的覆盖情况），即答复的完整性。再构建 LDA 主题模型，将文档集中每篇文档的主题按照概率分布的形式给出。结合这三个角度对答复意见的质量制定一套评价方案。

关键词:TFIDF 权重法，高斯朴素贝叶斯模型，K-Means 聚类模型，LDA 主题模型，

目录

1. 介绍

1.1 背景

近年来，随着互联网技术的高速发展，数字化、信息化的时代早已降临。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。

1.2 我们的任务

Task 1　群众留言分类

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型（参考附件 1 提供的内容分类三级标签体系）。再使用 F-Score 对分类方法进行评价。

Task 2　热点问题挖掘

根据附件 3 将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

Task 3　答复意见的评价

根据附件 4 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。

2. 术语

2.1 术语

- P:为第 i 类的查准率
- R:为第 i 类的查全率
- IDF：Inverse document frequency 指逆向文本频率
- TF：表示词条在文档中出现的频率
-

3. 符号

3.1 符号

表 1

符号	*定义*
N	单词在某文档中的频次
M	该文档的单词数
D	总文档数
D_w	出现了该单词的文档数

4. 模型

4.1 问题 1：群众留言分类

基于群众留言详情内容，建立识别模型，准确地识别出内容对应的分类标签，以便后续部门解决群众留言中的问题

4.1.1 模型准备

1. 数据展示

In [8]:
```
1 data
```

Out[8]:

	number	message	label
0	留言编号	留言详情	一级标签
1	24	A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项...	1
2	37	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水、电等设施后，烂尾多年，用护栏围着，不但占...	1
3	83	尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物业管理有限公司，未经小区业主同意...	1
4	303	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道...	1
...	...	...	...
9063	11763690	我们夫妻都是农村户口，大的是女9岁，小的是儿2岁半，才15斤，治疗两...	7
9064	11977590	本人2015年2月16号在B市中心医院做无痛人流手术，手术后还有怀孕...	7
9065	14697490	我们是再婚，很想再要一个小孩，不知我省二胎新政策何时出，如果先怀孕了...	7
9066	15397690	K8县惊现奇葩证明！ 我是西地省K8县人，想生二孩。被告知要开证明，...	7
9067	16135190	领导你好，我们属于未婚生子，但是在2013年已经接受处罚，小孩已上户...	7

9068 rows × 3 columns

图 1

2. 数据分布

对原始 9067 条数据进行数据探索，发现数据中并无存在空值，进一步查看七个分类标签留言的分布情况。

3. 数据抽取

随机抽取 700 条文本处理后的数据的 80%作为训练样本，其余作为测试集样本。

欠抽样：通过减少多数类样本提高少数类的分类性能

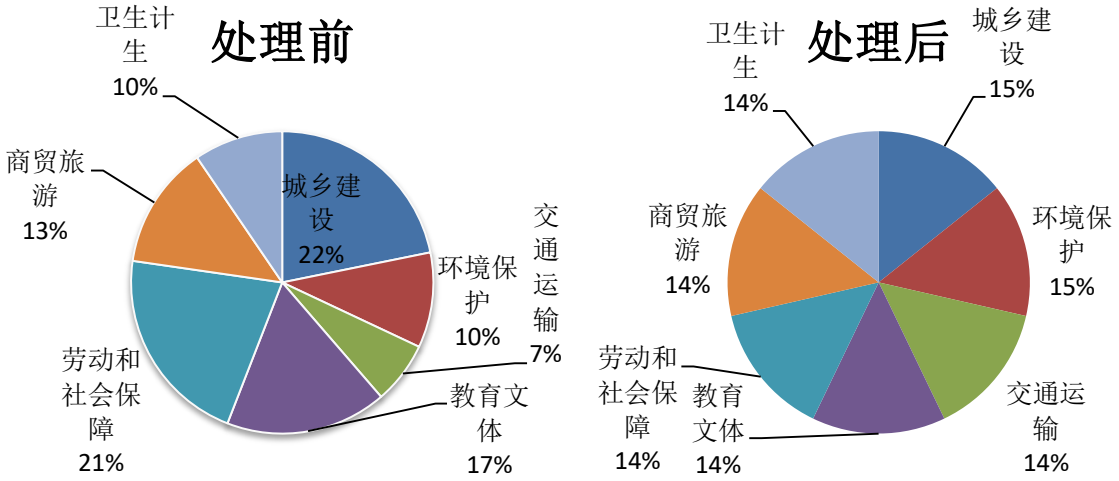

图 2

4. 数据预处理

数据清洗：去除空格、换行符

文本去重：在数据的储存和提取过程中，由于技术和某些客观的原因，造成了相同短信文本内容缺失等情况，因此需要对文本数据进行去重，去重即仅保留重复文本中的一条记录。

分词：结巴分词 jieba——“结巴”中文分词算法

- 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)
- 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

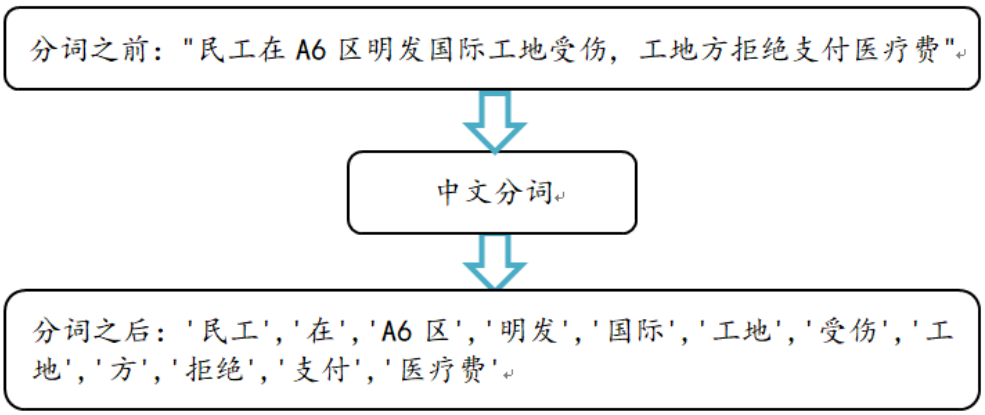

图 3

添加词典去停用词：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。

以上步骤处理部分展示：

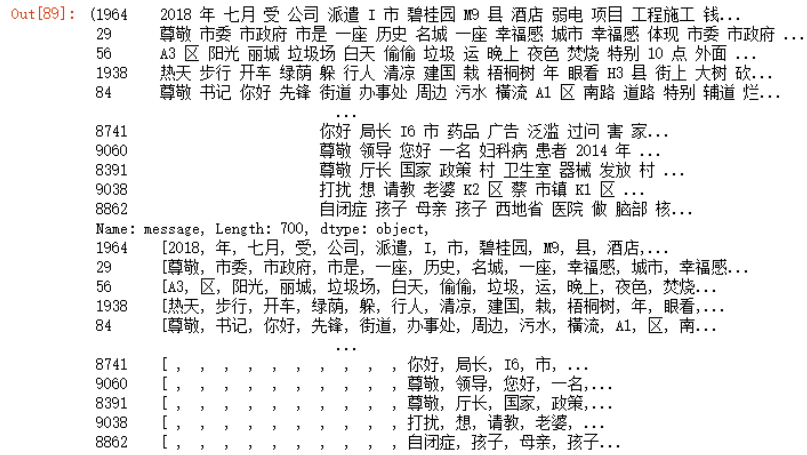

图 4

5. *文本的向量表示*

TF－IDF 权重策略

- 权重策略文档中的高频词应具有表征此文档较高的权重，除非该词也是高文档频率词
- TF：Term frequency 即关键词词频，是指一篇文档中关键词出现的频率

$$TF = \frac{N}{M} \tag{1}$$

IDF：Inverse document frequency 指逆向文本频率，是用于衡量关键词权重的指数，由公式

$$IDF = \log\left(\frac{D}{D_W}\right) \tag{2}$$

运算得到关键词出现的频率（部分展示）

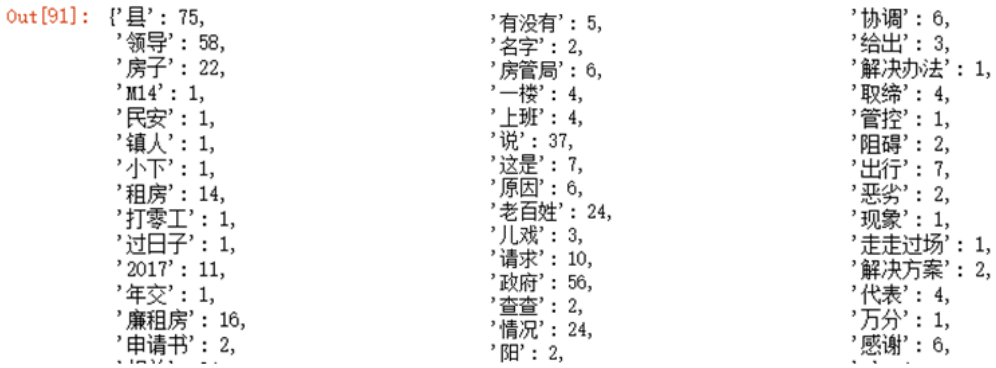

图 5

4.1.2模型建立

高斯朴素贝叶斯模型

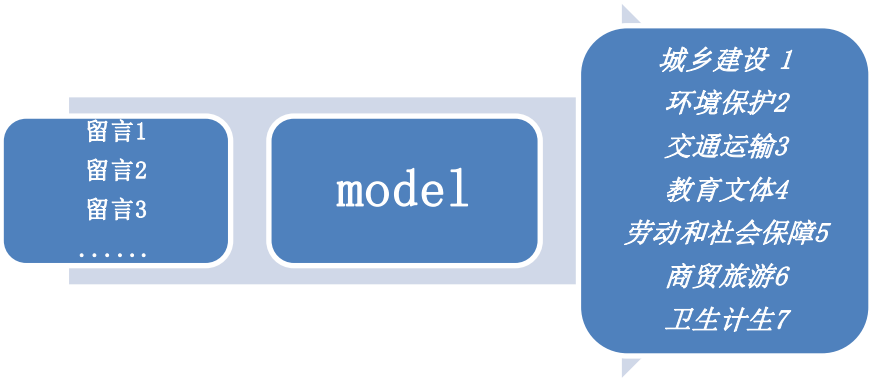

图 6

贝叶斯公式：

$$P(AB) = P(A)P(B \mid A) = P(B)P(A \mid B) --> P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad (3)$$

当 A 与 B 相互独立时：$P(AB) = P(A)P(B)$

朴素贝叶斯：

假设样本属性相互独立，$P(x \mid y) = P(x_1 \mid y)P(x_2 \mid y)P(x_3 \mid y)P(... \mid y)$（4）

朴素贝叶斯的表达式：

$$P(y \mid x) = \frac{P(x)P(x \mid y)}{P(x)} = \frac{p(y)}{p(x)} \prod_{i=1}^{d} P(x_i \mid y) \quad (5)$$

高斯朴素贝叶斯：

原始的朴素贝叶斯只能处理离散数据，当 $x_1, \ldots, x_n$ 是连续变量时，我们可以使用高斯朴素贝叶斯（Gaussian Naïve Bayes）完成分类任务。

当处理连续数据时，一种经典的假设是：与每个类相关的连续变量的分布是基于高斯分布的，故高斯贝叶斯的公式如下：

$$P(x_i = v \mid y_k) = \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} \exp\left(-\frac{(v - \mu_{y_k})}{2\sigma_{y_k}^2}\right) \quad (6)$$

其中 $\mu_y, \sigma_{y_k}^2$ 表示全部属于类 y_k 的样本中变量 x_i 的均值和方差。

4.1.3结果

混淆矩阵：利用处理后的训练集通过训练朴素贝叶斯模型，并由测试集进行分类得到模

型分类结果，整理汇总如下混淆矩阵：

表 2

		分类结果						
		类 1	类 2	类 3	类 4	类 5	类 6	类 7
真实类别	类 1	12	0	0	1	1	2	1
	类 2	7	21	17	0	0	3	0
	类 3	2	2	17	0	0	3	0
	类 4	1	2	0	10	1	1	1
	类 5	2	1	0	0	15	0	0
	类 6	2	0	0	1	0	14	3
	类 7	0	0	1	0	2	0	12

4.1.4 结果分析

精确度：表示的是分类为负类的样本中实际为负类的样本所占的比例，精确度越高，模型某类的分类效果越好。

$$精确度(\mathrm{Pr}\,ecision) = P = \frac{TN}{TN + FN} \tag{7}$$

召回率：

表示被正确分类的负类的比例，召回率越高，表示模型将负类勿分为正类的模型概率越低，模型效果越好。

$$召回率(recall) = R = \frac{TN}{TN + FP} \tag{8}$$

F1 值：

F-Measure（又称为 F-Score）综合考虑精确度与召回率，其中 P 指精确率即查准率，R 指召回率即查全率。F-Measure 是精确度和召回率的加权调和平均：

$$F = \frac{(\alpha^2 + 1) * P * R}{\alpha^2 * (P + R)} \tag{9}$$

当参数$\alpha = 1$时，就是最常见的 F1 值，即：

$$F1 = \frac{2 * P * R}{P + R} \tag{10}$$

表 3

	precision（精确度）	recall（召回率）	f1-score（F1）
1	0.4615	0.7059	0.5581
2	0.8077	0.7000	0.7500
3	0.8947	0.7083	0.7907

4	0.7692	0.6250	0.6897
5	0.7895	0.8333	0.8108
6	0.7000	0.7000	0.7000
7	0.7059	0.8000	0.7500
macro avg	0.7327	0.7246	0.7213

P=0.7327

R=0.7246

F1=0.7213

4.2 问题 2：热点问题挖掘

对某一特定时段内反映特定地点或特定人群问题的留言归类，定义热度评价指标，给出评价结果和对应留言信息“热度问题表”。

4.2.1 模型准备

对文本内容进行预处理后，转换为文本-词条矩阵。

4.2.2模型建立

建立基于 K－Means 的文本聚类模型：

聚类分析指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程。它是一种重要的人类行为。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。

简单理解，如果一个数据集合包含 N 个实例，根据某种准则可以将这 N 个实例划分为 m 个类别，每个类别中的实例都是相关的，而不同类别之间是区别的也就是不相关的，这个过程就叫聚类了。

聚类过程：

1) 特征选择(feature selection)：就像其他分类任务一样，特征往往是一切活动的基础，如何选取特征来尽可能的表达需要分类的信息是一个重要问题。表达性强的特征将很影响聚类效果。

2) 近邻测度(proximity measure)：当选定了实例向量的特征表达后，如何判断两个实例向量相似呢？这个问题是非常关键的一个问题，在聚类过程中也有着决定性的意义，因为聚类本质在区分相似与不相似，而近邻测度就是对这种相似性的一种定义。

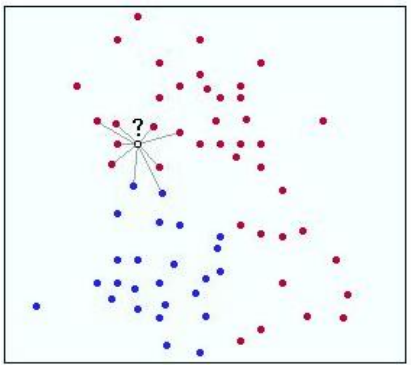

图 7

3) 聚类准则(clustering criterion)：定义了相似性还不够，结合近邻测度，如何判断相似才是关键。直观理解聚类准则这个概念就是何时聚类，何时不聚类的聚类条件。当我们使用聚类算法进行计算时，如何聚类是算法关心的，而聚与否需要一个标准，聚类准则就是这个标准。

4）聚类算法(clustering algorithm)：利用近邻测度和聚类准则开始聚类的过程。

4. 2. 3结果

定义热度评价指标

$$热度=点赞率*时间长*词频*留言占比数 \tag{11}$$

可得热点问题表以及热点问题留言明细表：

热点问题表如图：

问题ID	热度指数	时间范围	地点/人群	问题描述
1		2019/7/21至2019/9/25	A市A5区魅力之城小区	小区临街餐饮店油烟噪声扰民
2		2017/6/18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
3		2019/6/12至2020/1/06	J4县A市和西地省	社保和医保存在诸多问题
4		2018/10/27至2019/10/31	A市	交通不便利
5		2019/5/14至2019/10/18	A市A2区	通讯业务存在骚扰或技术不足

图 8

热点问题留言明细表如图：

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
1	189247	A000107866	pp中尽快接	2019/12/21 9:45:46	的平台，并且保证有三个律师解答（全国其它平台一个律师解答都不保证），便	0	0
1	190538	A00051614	旺龙路增加	2019/1/30 11:30:34	，老百姓上班、吃饭原本就在对面，却因为增加护栏而绕很大一个圈，建议中间	0	0
1	190669	A00084854	园小区相关	2019/1/8 17:37:48	，进行开发成商业，因为此区域缺少大的商业中心，缺少基础设施，或者建成学	0	6
1	191111	A00072923	市公交线路的	2019/2/10 18:02:44	—洞井路）、曙光南路（南二环—香樟路），洞井路（劳动路—香樟路段），需	0	0
1	191249	A0009233	交站点名称	2019/4/23 17:03:19	建议将“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”	0	0
1	191381	A0008159	西片区城铁	2019/7/4 18:52:39	院等带动发展，有金阳大道直达A市，并且规划有长浏城际铁路，直达金阳新城	0	0
1	194260	A00077323	场站一二号	2019/1/18 10:25:27	然后再从地下一层下到地下三层来乘坐二号线。原本地下二层和地下三层有扶梯	0	2
1	195198	A00026503	期A市地铁的	2019/3/22 18:03:51	处）-黎家冲（湖高路和欣盛路交汇处）-曾家冲（麓延路和金南路交汇处）-大	2	0
1	195915	A00018309	地铁7号至	2019/2/26 8:22:42	商业次中心。泉塘街道有众多楼盘，企业，十几万人口，出行不便，公交拥挤缓	1	21
1	197527	A000106250	展与交通安全	2019/3/29 11:13:55	）2.最近几年如一日，园内建设停滞不前，基本没有开发其它地块建设，尤其是	2	9
1	200206	A0003730	阳高路以南	2019/7/3 14:51:45	雨天还不需要很大的雨，就可以看到大量的黄土和污垢随雨水从空地内流出，排	0	3
1	201854	A00042351	建高速公路	2019/1/6 14:14:21	印寺等大汤山4A级景区，是一个成熟的旅游景点，随着A市，A8县工业的迅速发	0	0
1	202054	A0006300	星沙大道道路	2019/1/8 14:20:57	不及提前变道，直到看到了离红绿灯较近的地面标识才匆忙强行变道，严重影响	0	0
1	203070	A00017571	地铁站公交	2019/8/19 8:21:27	仅有的225路要30分钟才一趟，使得从A6区坐地铁出行几不方便，希望政府多开	0	2
1	203208	A00043904	华都学校西	2019/2/25 23:00:53	转车道上的直行车辆挡着不走，直行灯变绿时，左转车道上的直行车辆又插队到	0	0

图 9

该热度评价指标方便我们及时发现热点问题，也有助于相关部门进行有针对性地处理，提升服务效率。

4. 3 问题3：答复意见的评价

4.3.1 模型准备

- 答复的相关性：可理解为计算留言与相关部门对留言的答复意见的相似程度，相似度度量的值越小，说明个体间相似度越小，相似度的值越大说明个体差异越大。对多个不同的文本或者短文本对话消息计算他们之间的相似度，将这些文本中词语，映射到向量空间，形成文本中文字和向量数据的映射关系，通过计算几个或者多个不同的向量的差异的大小，来计算文本的相似度。
- 答复的完整性：覆盖率（答复意见中对留言内容的覆盖情况）

- 答复的可解释性：构建 LDA 主题模型(Latent Dirichlet Allocation 主题模型)，它可以将文档集中每篇文档的主题按照概率分布的形式给出。同时它是一种无监督学习算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集以及指定主题的数量k 即可。主题模型是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

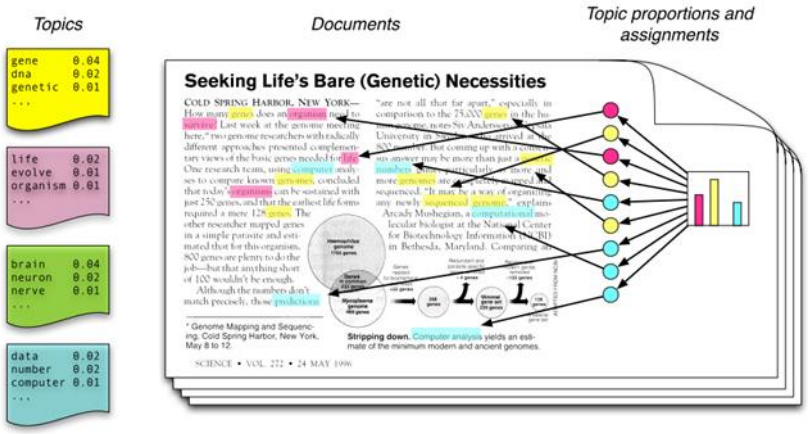

图 10

4.3.2 模型建立

1.特征抽取（TF-IDF）

在一份给定的文件里，词频 (term frequency, TF) 指的是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（分子一般小于分母 区别于 IDF），以防止它偏向长的文件。

逆向文件频率 (inverse document frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

TFIDF 的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TFIDF 实际上是：TF * IDF，TF 词频(Term Frequency)，IDF 反文档频率(Inverse Document Frequency)。

2.相似度计算（余弦相似度）

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。

3.完整性（覆盖率）

$$答复意见覆盖率 = \frac{答复意见涉及高频词频}{留言详情词频} \times 100\% \tag{12}$$

5. LDA 主题模型

在主题模型中，主题表示一个概念、一个方面，表现为一系列相关的单词，是这些单词的条件概率。形象来说，主题就是一个桶，里面装了出现概率较高的单词，这些单词与这个主题有很强的相关性。词语出现的概率公式为：

$$p(词语|文档)=\sum_{主题} p(词语|主题)\times p(主题|文档) \tag{13}$$

这个概率公式可以用矩阵表示：

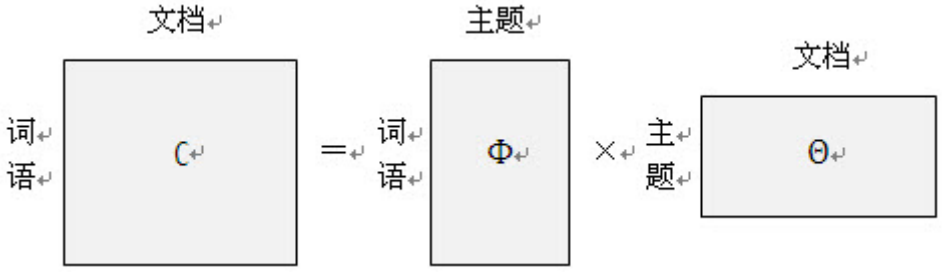

图 11

其中”文档-词语”矩阵表示每个文档中每个单词的词频，即出现的概率；”主题-词语”矩阵表示每个主题中每个单词的出现概率；”文档-主题”矩阵表示每个文档中每个主题出现的概率。

6. 模型的评估和推广

6.1 优缺点

6.1.1 优点

- 高斯朴素贝叶斯模型优点：有稳定的分类效率。对小规模的数据表现很好，能个处理多分类任务，适合增量式训练。对缺失数据不太敏感，算法比较简单。
- 主题模型算法是文本处理与数据挖掘中一个非常重要的方法，它可以有效地从文本语义中提取主题信息，也是对文字隐含主题进行建模的方法。它克服了传统信息检索中文档相似度计算方法的缺点。

6.1.2 缺点

- 高斯朴素贝叶斯模型缺点：在属性个数比较多或者属性之间相关性较大时，分类效果不好。分类决策存在一定的错误率。对输入数据的表达形式很敏感。
- 主题模型缺点：词袋方法没有考虑词与词之间的顺序，简化了问题的复杂性。

7. 结论

7.1 问题的结论

- 问题一：对附件 2 中的内容建立识别模型，准确地识别出内容对应的分类标签，再将非结构化数据进行去重去空、结巴分词 jieba（采用前缀词典实现高效的词图扫描、动态规划查找最大概率路径以及使用了 Viterbi 算法)及停用词过滤等数据预处理，然后基于 TFIDF 权重法提取候选特征词，形成词袋，建立高斯朴素贝叶斯模型，再利用处理后的训练集通过训练朴素贝叶斯模型，并由测试集进行分类得到模型分类结果。使用 F-Score 对这个一级标签分类模型进行评价，得到评分 0.7213。用这个模型来进行留言划分和热点整理的工作希望能对提升政府的管理水平和施政效率。
- 问题二：用 K-Means 聚类后得出的热点问题会有误差，一些关联性不大的问题也会分在相同的类别，为了解决这个问题，将聚类分成 1000 类再从其中选出热度最高的五类，进行热度评价，精度也有了较大的提升。
- 问题三：计算留言与相关部门对留言的答复意见的相似程度即相关性，将附件文本中词语，映射到向量空间，形成文本中文字和向量数据的映射关系，通过计算几个或者多个不同的向量的差异的大小，来计算文本的相似度。同时要知道其覆盖率（答复意见中对留言内容的覆盖情况），即答复的完整性。再构建 LDA 主题模型，将文档集中每篇文档的主题按照概率分布的形式给出。结合这三个角度对答复意见的质量制定一套评价方案。

7.2 我们的模型中使用的方法

- 结巴分词 jieba
- Viterbi 算法
- TFIDF 权重法
- 高斯朴素贝叶斯模型
- K-Means 聚类模型
- LDA 主题模型

8. 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] 杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007
[3] 王美方, 刘培玉.朱振方. 基于 TFIDF 的特征选择方法[J]. 计算机工程与设计, 2007, 28(23):5795-5796
[4] 陈春燕.计算机信息完整性度量与保护方法研究. 北京信息职业技术学院.2017
[5] 全文君.数据挖掘过程中的可解释性问题研究. 重庆大学博士学位论文.2018
[6] 张敏. 基于文本内容的垃圾短信识别
[7] David M. Blei, AndrewY. Ng, Michael I. Jordan, LatentDirichlet Allocation, Journal of Machine Learning Research 3, p993-1022,2003
[8] Jiawei Han, Micheline Kamber, Jian Pei 著 范明 孟小峰译 数据挖掘—概念与技术 北京：机械 工业出版社 2012 年

9. 附录

9.1 附录 1

```
## 数据预处理
data=pd.read_excel('附件 2(1).xlsx',header=None)
data.columns=['number','message','label']
data['message']=data['message'].dropna().apply(lambda x:re.sub('\t\t\t\t\t\t\t\t\t\t',",x))
n=100
a=data[data['label']==1].sample(n)
b=data[data['label']==2].sample(n)
c=data[data['label']==3].sample(n)
d=data[data['label']==4].sample(n)
e=data[data['label']==5].sample(n)
f=data[data['label']==6].sample(n)
g=data[data['label']==7].sample(n)
data_new=pd.concat([a,b,c,d,e,f,g],axis=0)
data_dup=data_new['message'].drop_duplicates()
jieba.load_userdict('newdic1.txt')
data_cut=data_dup.apply(lambda x:jieba.lcut(x))
stopwords=pd.read_csv('stopword.txt',encoding='GB18030',sep='hahaha',header=None)
stopwords=['      ','^[0-9a-zA-Z]+$','[0-9]+[a-zA-Z]+[0-9a-zA-Z]*|[a-zA-Z]+[0-9]+[0-9a-zA-Z]*','\u3000','会','
月','\t','\r']+list(stopwords.iloc[:,0])
data_after_stop=data_cut.apply(lambda x:[i for i in x if i not in stopwords])
adata=data_after_stop.apply(lambda x:' '.join(x))
labels=data_new.loc[data_after_stop.index,'label']

## 计算词频（label=1）
word_fre={}
adata
data_after_stop
labels=data_process()
for i in data_after_stop[data['label']==1]:
      for j in i:
            if j not in word_fre.keys():
                  word_fre[j]=1
            else:
                  word_fre[j]+=1
word_fre

## 建立训练集和测试集
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
data_tr, data_te, labels_tr, labels_te = train_test_split(adata, data_new['label'], test_size=0.2)
## 让训练集和测试集向量长度一致
```

```
countVectorizer = CountVectorizer()
data_tr = countVectorizer.fit_transform(data_tr)
X_tr = TfidfTransformer().fit_transform(data_tr.toarray()).toarray()
data_te = CountVectorizer(vocabulary=countVectorizer.vocabulary_).fit_transform(data_te)
X_te = TfidfTransformer().fit_transform(data_te.toarray()).toarray()
```

9.2 附录 2

```
import re
import jieba
import pandas as pd
data=pd.read_excel('附件 3(3).xlsx',header=None)
data.columns=['num','mes']
content = data['mes']
content.head()
# 分词
content = content.apply(jieba.lcut)
content.head()
# 去除停用词
with open('stopword.txt', 'r', encoding='GB18030') as f:
    stop = f.read()
stop = stop.split()
content = content.apply(lambda x: [i for i in x if i not in stop])
content.head()
# 转换为文本-词条矩阵
x = content.apply(lambda x: ' '.join(x))
y = data['num']
print(x.head(), y.head())
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
x_data = cv.fit_transform(x)
cv.vocabulary_  # 字典
cv.get_feature_names()  # 词条
x_data.toarray()
# 使用 KMeans 聚类

from sklearn.cluster import KMeans
model = KMeans(n_clusters=40).fit(x_data)
y_pre=model.labels_
# 评价聚类效果
from sklearn.metrics import silhouette_score
# help(silhouette_score)
silhouette_score(x_data, y_pre)
```