

“泰迪杯” 全国数据挖掘挑战赛

二〇二〇 年 五 月 七 日

网络智慧政务的分析与挖掘

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。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。本文将基于信息挖掘技术对智慧政务的相关信息和数据进行挖掘，提取我们需要进行分析的部分进行深度挖掘和分析。

针对问题一，本文首先将附件 2 中的非结构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 TFIDF 权重法提取出了候选特征词，形成词袋，构造词汇-文本矩阵，再通过训练朴素贝叶斯模型来进行分类算法对群众留言问题进行划分。

针对问题二，我们也是先对数据进行去重去空、中文分词及停用词过滤等数据预处理，然后使用 K-means 聚类方法将所有的文本数据划分为 30 类，之后定义热门问题，采用主成分分析法对划分的 30 类文本数据进行定义分析，分别统计 30 类问题的，点赞数量，时间跨度，反应人数，并将其做表，导入 SPSS 中，进行主成分分析计算得出热度分数，根据分数排序选出热度前五的问题结果。

针对问题三，我们从附件 4 中提取出留言详情和答复意见两列，同样，我们首先对文本数据经过数据清洗，中文分词等操作，然后对居民留言以及政府回复进行数据清洗，再对每一个居民的留言以及政府的回复一一对应用 word2vec 将文本向量化进行对比相似度的计算，相似度越高说明回复的相关性以及可解释性越高，用 word2vec 得出的结果评判文本完整性进行模板建立。再用 LDA 主题特征提取留言详情与答复意见的主题词，相比较 2 个主题词的联系程度定义可解释性。

最后结合我们对各项数据的挖掘分析，解决群众留言分类问题，做出答复意见的评价，最后根据居民留言和政府回复与模板之间的差异，提出针对性建议。

关键词：去重 中文分词 K-means 聚类 LDA 主题特征提取 TF-IDF

目录

摘要.....	2
一、项目背景.....	4
1.1 挖掘目标.....	4
二、总体流程.....	4
2.1 总体流程.....	4
2.2 (流程图)	4
三、文本聚类.....	5
3.1 群众留言分类.....	5
3.1.1 数据预处理:	5
3.1.2 数据清洗:	6
3.1.3 中文分词:	6
3.1.4 特征提取与文本向量化:	7
3.1.5 模型建立:	8
3.1.6 评价模型:	10
四、分析热门需求.....	10
4.1 热点问题挖掘.....	10
4.1.1 数据预处理.....	10
4.1.2 中文分词.....	10
4.1.3 提取特征.....	10
4.1.4 文本聚类.....	11
4.2 热点问题分析.....	12
五、结果处理与分析.....	18
5.1 答复意见的评价.....	18
5.2 数据预处理.....	18
5.3 word2vec 算法.....	18
5.4 文本相关性.....	19
5.5 文本完整性.....	19
5.6 文本可解释性.....	19
参考文献.....	20

一、项目背景

1.1 挖掘目标

近年来,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。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。

本次建模目标是利用智慧政务网络数据,经过数据清洗后,利用 jieba 中文分词工具对智慧政务描述进行分词、文本向量化、K-means 聚类、PCA 主成分分析法, LDA 主题特征提取 等方法,达到以下三个目标:

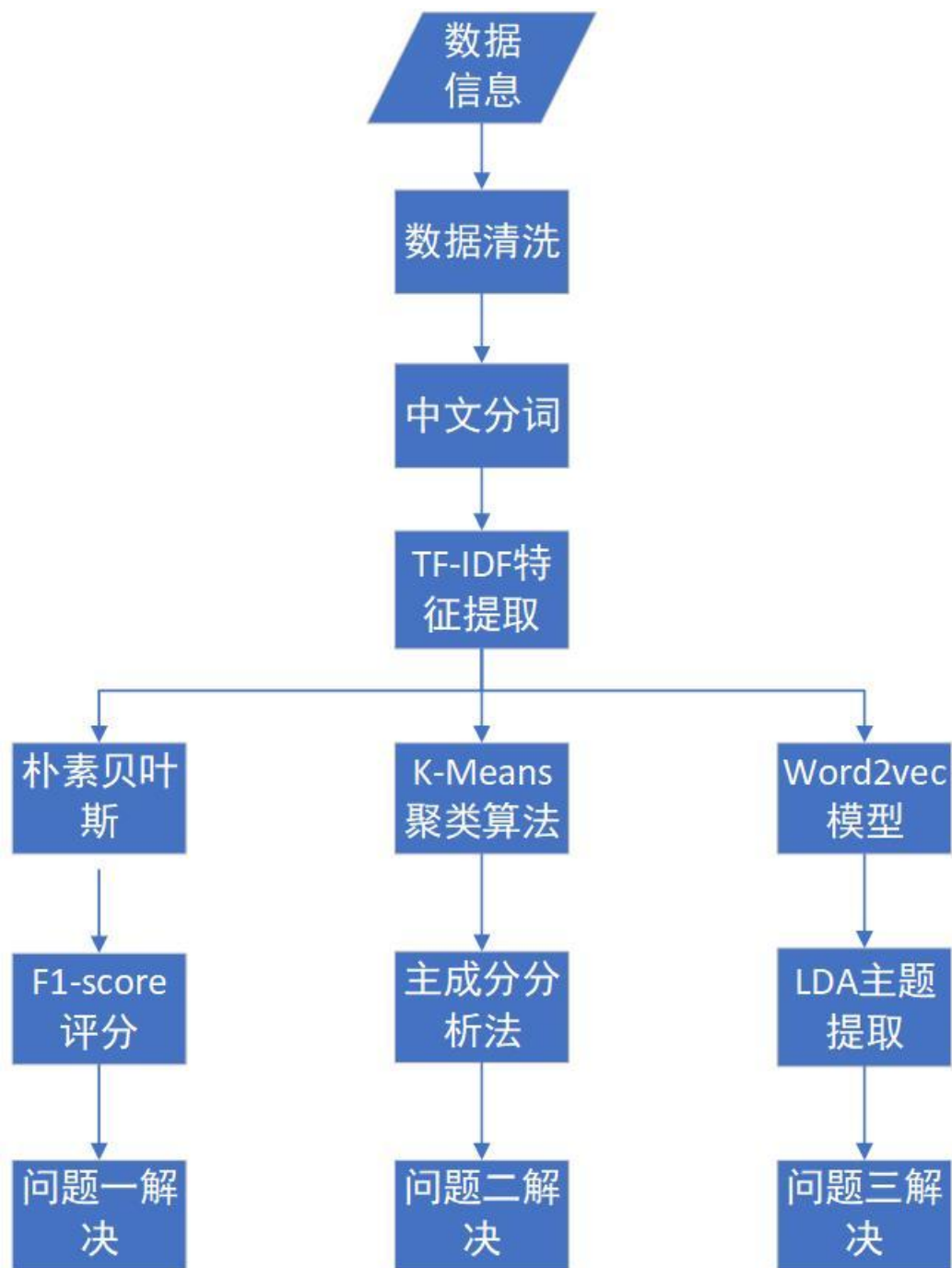
- (1) 解决群众留言分类问题
- (2) 解决热点问题的挖掘
- (3) 做出答复意见的评价

二、总体流程

2.1 总体流程

将所有数据进行清洗,中文分词,然后进行 TF-IDF 特征提取。对于解决问题一:训练朴素贝叶斯模型来进行分类,然后使用 F1-score 进行评价。对于问题二:先用 K-Means 聚类方法将数据进行聚类,然后使用主成分分析法选出热度前五的问题。对于问题三:使用 Word2vec 模型评价相关性与完整性,使用 LDA 主题特征提取评价可解释性

2.2 (流程图)



三、文本聚类

3.1 群众留言分类

3.1.1 数据预处理：

留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
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	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目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市A市，尽快整改这个极不文明的路段。	城乡建设

附件 2 中给出了群众留言与其所属的类别。但是点开一个仔细观察，不难发现其中都是长句，会出现许多重复词，无意词，错误词。如图 1，其中许多需要删除，清洗的文字。图 1

所以首先需要将其中的文本清洗、去重、去空，然后进行分词，方便下一步的进行，减小、误差，防止对分类结果造成较大的误差。而且数据中皆为文本，我们要将其进行计算，就需要将其转换为数值。

3.1.2 数据清洗：

编写代码，消除数据中所有的空格，换行符，数字，以及大小写字母。如图 1 可见数据中有地点等出现了数字与大小写字母，这类信息对我们分类是没有帮助的，属于无意义的词语，所以在数据清洗环节就要将其去除，这样可以减小无关因素对结果产生的误差。

3.1.3 中文分词：

我们知道结巴分词(Jieba)是一个强大的分词库,它的开发者经过了大量的训练之后,向里面录入了有两万多条词语组成了基本的库,并且 jieba 的实现原理也比较完善,设计的算法有基于前缀词典的有向无环图、动态规划、HMM 模型等。Jieba 分词支持了三种分词的模式：

①精确模式。这种模式试图以最高的精度来对句子进行划分，比较适用于文本的分析任务。

②全模式。这种模式是可以扫描出句子中可以划分成词的词语，它的速度非常快，但是不可以用来解决歧义问题。

③搜索引擎模式。这种模式是基于精确模式对长词在进行切分，可以将这种模式用于搜索引擎分词。

我们使用 jieba 分词的第一种模式也即精准模式，分词以后可以便于以后的计算。部分数据分词结果如图 2

```
0      [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      [市政府, \, , 市, 交警支队, \, , 市, 安监局, \, , 市, 环保局, \, , 区政府...
3      [胡书记, , , 您好, , , 感谢您, 百忙之中, 查看, 这份, 留言, 。, 我, 的...
4      [ , , 县, 丁字街, 的, 商户, 乱, 摆摊, , , 前段时间, 丁字街, 的, ...

...

490    [邓, 局长, :, , , , , , , , , 市, 卫生局, 卫生, 监督, 所...
491    [ , , 山门, 镇, 人民, 医院, 挂号费, 这几年来, 一直, 都, 是, 高额, ...
492    [我, 是, 县, 人民, 医院, 的, 一个, 普通, 医务人员, , , 我, 想, 请问...
493    [张, 厅长, :, , , , 您好, !, 我, 是, 一名, 医务, 工作者, 。, ...
494    [书记, :, , , , 你好, , , 我们, 是, 县乡镇, 卫生院, 的, 一名, 普...
Name: 留言详情, Length: 495, dtype: object
```

图 2

在分词以后我们发现图 2 中还有许多的空格和符号还有一些无意义的词语如我、是，所以我们将此类无意义的停用词编入 `stopword.txt` 中。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。为了找出这些停用词，需要一些标准估计词的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看作是噪声词。编入代码去除掉 `stopword` 里面包括的停用词，能提高精准度。

3.1.4 特征提取与文本向量化：

经过数据清洗后，我们发现虽然去掉了一些停用词，但是还有许多的词语，其中的一部分才是对我们来说有价值的词语来降低向量空间维数，简化我们的计算，提高计算的速度与效率，所以我们使用特征提取，在不改变文本有价值词语的情况下尽量减少范围，以此增强结果的可信度。

特征提取是在特征独立的假设基础上,通过构造评估函数,对特征集合中的每个特征进行独立的评估,并对每一个特征进行打分，然后再将所有特征按分值的大小排序,再提取预定数目的最优特征作为提取结果的特征子集。我们常见的评估函数有以下几种：

①文档频数(Document Frequency)。

$$\text{DocFreq}(F) = P(W|C_i) = \frac{DF}{|C_i|}$$

②信息增益(Information Gain)。

$$\begin{aligned} \text{InfGain}(F) = & P(W) \sum_i P(C_i|W) \log \frac{P(C_i|W)}{P(C_i)} + \\ & P(\bar{W}) \sum_i P(C_i|\bar{W}) \log \frac{P(C_i|\bar{W})}{P(C_i)} \end{aligned}$$

③期望交叉熵(Expected Cross Entropy)。

$$\text{CrossEntryTxt}(F) = P(W) \sum_i P(C_i|W) \log \frac{P(C_i|W)}{P(C_i)}$$

④互信息(Mutual Information)。

$$\text{MutualInforTxt}(F) = \sum_i P(C_i) \log \frac{P(W|C_i)}{P(W)}$$

⑤文本证据权(The Weight of Evidence for Text)。

$$\text{WeightofEvidTxt}(F) = P(W) \sum_i P(C_i) \left| \log \frac{P(C_i|W)(1-P(C_i))}{P(C_i)(1-P(C_i|W))} \right|$$

⑥优势率(Odds Ratio)。

$$\text{OddsRatio}(F) = \log \frac{P(W|pos)(1-P(W|neg))}{P(W|neg)(1-P(W|pos))}$$

⑦词频(Word Frequency)。

$$\text{Freq (F) =TF (W)}$$

以上几种方法都是比较常见的特征提取方法,由于我们需要处理的数据类型都是文本,而且考虑到每一个方法的运行速率,以及运行效率,我们选择使用了词频特征提取法来解决问题。

运用 TF-IDF 来提取文档中的关键词。Tf 也就是词频,也就是词语在文章中出现的频率,在使用 tf-idf 算法之前要先把一些常用的词语给洗了如‘的’,‘我’,‘你’之类的词语,因为这一类词虽然出现的频率很高但是很明显不是我们需要的高频词,在统计出文章的词频后,可能会有好几个高频率的词频,但是其中可能只有一个是我们所需要的此时我们就需要引用 idf 也就是逆文档频率,意思是说某些词在平常出现的比较少但是在改文档中出现的频率又比较高,那么该词就是我们该文档的关键词。

TF-IDF 具体的计算过程为:

(1) 计算词频

词频 (tf) = 某个词语在文章中出现的次数

还可以对词频进行标准化处理

词频 (tf) = 某个词语在文章中出现的次数 / 文章的总词数

(2) 计算逆文档频率

需要使用一个语料库来模拟周围的语言环境

逆文档频率 (idf) = $\log (\text{语料库的文档总数} / (\text{包含该词的文档数} + 1))$

(3) 计算 tf-idf

Tf-idf = tf * idf 也就是用词频乘以逆文档频率,可以很清楚的明白,tf-idf 算法就是与一个词语在文档中出现的次数成正比,与该词语在我们建立的语料库中出现的次数成反比,计算出每个词的 tf-idf 后,按照降序排列,在前面的几个词语就是我们所需要提取的文档的关键词。

3.1.5 模型建立:

根据问题我们需要对数据进行分类,分类的标签题目已经给好,我们需要根据现有的数据建立分类模型。常见的分类模型有:朴素贝叶斯算法,支持向量机,Rocchio 算法等等,综合利弊后,我们选用了朴素贝叶斯算法。

朴素贝叶斯算法用到了统计学中的全概公式和贝叶斯公式。我们在概率论课上学习过,其中全概公式的描述是:设随机实验 E 的样本空间为 U, A 为样本空间中的一个事件, B₁, B₂, ..., B_n 为样本空间 U 的一个划分即 $P(B_1) + P(B_2) + \dots + P(B_n) = 1$, 那么有 $P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)$ 。贝叶斯公式的描述如下:

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^n P(A|B_i)}$$

所以可以知道朴素贝叶斯算法在文本分类中的应用是这样的:假定把文本分类语料库中的所有文本预先分为 m 个类别,分别记为 C_1 、 C_2 、...判定类别的文档记作 D , 如果用特征词来表示文档 D , 那么 $D=\{X, X_1, \dots\}$, 那么判定文档 D 是否属于类别 C_i 的概率可以表示为:

$$P(C_i|D) = \frac{P(D|C_i)P(C_i)}{P(D)},$$

首先其中 $P(D)$ 对于任何类别来说都是一个常数在判定文档 D 属于何种类别时可以暂时忽略。 $P(C)$ 即为类别 C_i 在语料库中出现的概率, 可以表示为 $P(C_i) = N_i/N$, 其中 N 表示文本语料库中的文档总数, N_i 表示类别 C_i 中的文档数。当假定文档 D 中的各个特征词汇相互独立时 $P(D|C_i) = \prod_{j=1}^n P(X_j|C_i)$, 其中 $P(X_j|C_i)$ 表示类别 C_i 中包含特征词 X_j 的概率, 可以用 A_{ij}/N_i 表示, 这里 A_{ij} 表示类别 C_i 中包含特征词 X_j 的文档数。那么判定文档 D 属于类别 C_i 的概率可以推导为:

$$P(C_i|D) = P(C_i) \prod_{j=1}^n P(X_j|C_i) = \frac{N_i}{N} \times \prod_{j=1}^n \frac{A_{ij}}{N_i}$$

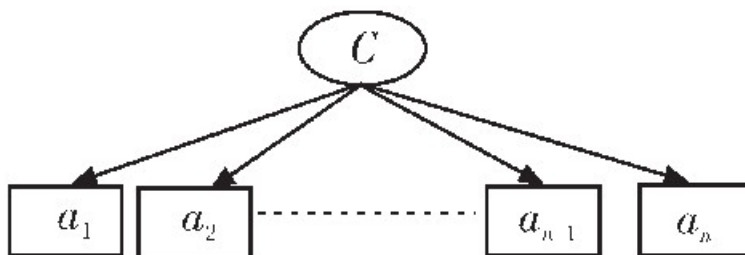
我们发现在朴素贝叶斯算法中, 如图 3 所示, 需要将训练实例表示成属性(特征)向量 A 和决策类别变量 C 。同时, 在该算法中, 假定每个特征之间相互独立对于决策变量它们具有独立作用。

$$P(A|C_i) = \prod_{k=1}^n P(a_k|C_i) \quad (4)$$

则对于后验概率 $P(C_i|A)$, 表示特征 A 属于类别 C_i 的概率, 根据贝叶斯准则, 可以由下式计算:

$$P(C_i|A) = \frac{P(C_i)}{P(A)} \prod_{j=1}^m P(a_j|C_i) \quad (5)$$

其中, C 的先验概率 $P(C_i)$ 很容易求得: 训练集中属于 C 的数量/训练集的总量。



图三

在知道了朴素贝叶斯算法的原理后我们随机将数据按百分之二十比例分为测试集，百分之八十比例分为训练集，进行了朴素贝叶斯模型的建立。

3.1.6 评价模型：

模型建立好后，模型的效果怎么样，需要确立一个标准来评价他。F-score

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i},$$

评价是比较普遍的一种评价方法，

(其中

P 为第 i 类的查准率，R_i 为第 i 类的查全率。)

查准率= (预测为真且正确预测的样本数) / (所有预测为真的样本数)

查全率= (预测为真且正确预测的样本数) / (实际情况中为真的样本数)

编入代码后 F1 评分：

0.6828730069367097

可见我们所创立模型的 F-score 评分比较平庸，此模型的精准度比较一般，可能是因为数据处理方面有待提高。日后完善应该增加数据加强方面的过程。

四、分析热门需求

4.1 热点问题挖掘

4.1.1 数据预处理

附件 3 中也是有许多的重复词，无意词，错误词。

A	B	C	D	E	F	G
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
188006	A000102948	A3区一米阳光婚纱摄影是否合法纳税了？	2019-02-28 11:25:05	座落在A市A3区联丰路米兰春天G2栋320，一家名叫一米阳光婚纱摄影的影楼，据说年里这一个工作室营业额就上百万，因为地处居民楼内部，而且有蛮长的时间了，请税务局和工商局查一下，看看这个一米阳光有没有正常纳税！如果没有，应该怎么操作！	0	0

其中许多需要删除并清洗。所以我们同样使用代码清除数据中所有的空格，换行符，和其他符号。此次没有删除数字与大小写字母，因为需要回答后面地点人群的问题，所以地点等信息对我们来说是有用的，不能删除。

4.1.2 中文分词

使用 jieba 中文分词将其分词，便于以后的计算(同样使用停用词过滤)。

4.1.3 提取特征

我们再次选择 TF-IDF 算法进行特征提取，然后进行文本向量化。方便我们对文本的计算，提高效率，减小误差。

4.1.4 文本聚类

K-means 的基本思想是，已知一个包含 N 个样本数据的数据集，以及给定聚类数目 K, 首先随机选取 K 个样本分别作为初始划分的簇类中心，然后根据相似性度量函数采用迭代的方法，计算未划分的样本数据到每个聚类中心点的距离，并将该样本数据划分到与之最近的那个聚类中心所在的簇类中，对分配完的每一个簇类，通过计算该簇类内所有数据平均值不断移动聚类中心，重新划分聚类，直到类内误差平方和最小且没有变化时为止。该算法有一个特点，就是每一次迭代过程中都要判断每个样本数据是否正确划分到簇类中，若不正确，重新调整。当全部数据调整完后，再修改簇类中心，进行下一次迭代计算。如果某一次迭代过程中每个数据样本都分配到正确的簇类中，则不再调整聚类中心。聚类中心稳定不再变化，标志目标函数收敛，算法结束，最后评价聚类结果。为了方便我们理解聚类后的结果我们要明白 k-means 算法中的定义，为了方便算法描述，定义已知一个含有 n 个数据的样本集合为 Q, 即:

$$\Omega = \{x_i | x_i = (x_{i1}, x_{i2}, \dots, x_{id}), i = 1, 2, \dots, n\}$$

其中， $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ 是一个 d 维向量，表示第 i 的

$$dis(x_i, c_j) = \sqrt{\sum_{l=1}^d (x_{il} - c_{jl})^2}, i = 1, 2, \dots, n; j = 1, 2, \dots, K$$

数据的 d

个不同属性，n 是样本容量。

聚类的中心为:

$$C = \{c_j | c_j = (c_{j1}, c_{j2}, \dots, c_{jd}), j = 1, 2, \dots, K\}$$

其中， $c_j = (c_{j1}, c_{j2}, \dots, c_{jd})$ 是第 j 个簇类的中心点，每个中心点 C, 都含有 d 个不同属性，K 是簇类个数。

定义 1 两个数据 x 和 C, 之间的欧几里得距离为 $dis(x, c)$, 其表示如下:

其中， $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$, $c_j = (c_{j1}, c_{j2}, \dots, c_{jd})$, K 是簇类个数。

定义 2 同一簇类的中心点为 c, 表示如下:

$$c_{jl} = \frac{1}{N(\phi_j)} \sum_{x_i \in \phi_j} x_{il}, l = 1, 2, \dots, d; j = 1, 2, \dots, K$$

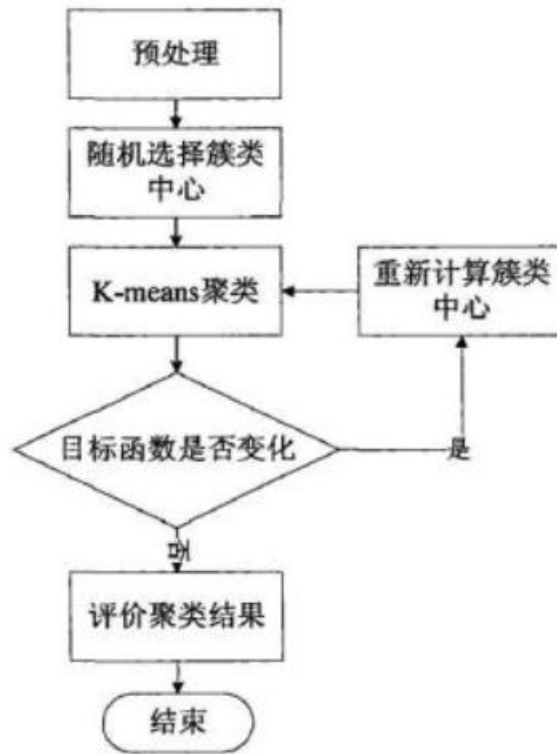
其中， $N(\phi_j)$ 是同一簇类 ϕ 中数据的个数， $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ 。

定义 3 准则函数为类内误差平方和 SSE 表示如下:

$$SSE = \sum_{j=1}^K \sum_{x_i \in \phi_j} dis(x_i, c_j)$$

其中， $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ ， $c_j = (c_{j1}, c_{j2}, \dots, c_{jd})$ ，K 是簇类个数。

k-means 聚类方法的大致流程如下图



我们使用 K-means 聚类方法将所有的文本数据划分为 30 类。

4.2 热点问题分析

首先我们定义什么是热点问题：

- I、点赞数量多的问题
- II、时间集中的问题
- III、反应人数多的问题

在这个三个因素上我们主要采用了主成分分析法





因为主成分分析法旨在利用降维的思想，将分散在一组变量上的信息转换到某几个综合变量(线性组合)上，并且得到的综合变量它们之间没有关联。从几何的观点来看，主成分分析是对原坐标轴作了一个坐标旋转，然后获得相互正交的坐标轴，使该坐标轴的方向为所有数据点分散最开的方向，依照得到的特征值的大小排列这些新的坐标轴。

主成分分析法的操作流程图如下：



根据热点问题的定义，我们分别统计 30 类问题的，点赞数量，时间跨度，反应人数，并将其做表，导入 SPSS 中，进行主成分分析计算得出结果。

导入到 SPSS 中进行降维因子分析

	 V1	 点赞数	 时间跨度天	 总数
1	第1组	136	367	133
2	第2组	2101	370	170
3	第3组	65	363	93
4	第4组	2188	352	38
5	第5组	1041	366	177
6	第6组	1030	386	174
7	第7组	68	339	43
8	第8组	34	382	25
9	第9组	68	345	55
10	第10组	150	363	117
11	第11组	47	349	49
12	第12组	158	360	99
13	第13组	85	365	115
14	第14组	133	362	134
15	第15组	318	367	53
16	第16组	3029	379	1385
17	第17组	218	367	152
18	第18组	50	432	52
19	第19组	237	359	137
20	第20组	623	371	459
21	第21组	116	371	76
22	第22组	76	359	37
23	第23组	66	356	52
24	第24组	120	360	68
25	第25组	60	361	58
26	第26组	3	260	17
27	第27组	74	349	74
28	第28组	15	368	51
29	第29组	44	358	70
30	第30组	29	359	46

具体结果如下

KMO 和巴特利特检验

KMO 取样适切性量数。		.531
巴特利特球形度检验	近似卡方	19.281
	自由度	3
	显著性	.000

公因子方差

	初始	提取
点赞数	1.000	.796
时间跨度/天	1.000	.174
总数	1.000	.813

提取方法：主成分分析法。

总方差解释

成分	总计	初始特征值		提取载荷平方和		
		方差百分比	累积 %	总计	方差百分比	累积 %
1	1.783	59.420	59.420	1.783	59.420	59.420
2	.916	30.545	89.965			
3	.301	10.035	100.000			

提取方法：主成分分析法。

成分矩阵^a

	成分 1
点赞数	.892
时间跨度/天	.417
总数	.902

提取方法：主成分分析法。

a. 提取了 1 个成分。

旋转后的成分矩阵^a

a. 仅提取了一个成分。无法旋转此解。

成分得分系数矩阵

	成分 1
点赞数	.500
时间跨度/天	.234
总数	.506

提取方法：主成分分析法。

旋转方法：凯撒正态化最大方差法。

成分得分协方差矩阵

成分	1
1	1.000

提取方法：主成分分析法。

旋转方法：凯撒正态化最大方差法。

确定权重

用主成分分析确定权重有:指标权重等于以主成分的方差贡献率为权重,对该指标在各主成分线性组合中的系数的加权平均的归一化。

(1) 计算出各个指标在不同主成分线性组合中的系数(用载荷数除以对应特征根的开方)

(2) 用初始特征值的方差表示各个主成分方差贡献率,方差贡献率越大则该主成分的重要性越强。

由此计算出综合得分模型中的系数

	A	B	C	D	E
1		线性组合中的系数	方差贡献率	综合得分模型中的系数	指标权重
2	点赞数	0.668		0.668	0.403382
3	时间跨度	0.312		0.312	0.188406
4	总数	0.676	0.594	0.676	0.408213

将其归一化后得到各个因素的权重

	A	B	C	D	E
1		点赞数	时间跨度/总数	得分	
2	第1组	136	367	133	178.068
3	第2组	2101	370	170	985.623
4	第3组	65	363	93	132.383
5	第4组	2188	352	38	963.444
6	第5组	1041	366	177	560.547
7	第6组	1030	386	174	558.65
8	第7组	68	339	43	108.68
9	第8组	34	382	25	95.718
10	第9组	68	345	55	114.704
11	第10组	150	363	117	176.43
12	第11组	47	349	49	104.545
13	第12组	158	360	99	171.746
14	第13组	85	365	115	149.795
15	第14组	133	362	134	176.327
16	第15组	318	367	53	218.774
17	第16组	3029	379	1385	1857.019
18	第17组	218	367	152	218.866
19	第18组	50	432	52	122.582
20	第19组	237	359	137	218.899
21	第20组	623	371	459	508.089
22	第21组	116	371	76	147.504
23	第22组	76	359	37	113.216
24	第23组	66	356	52	114.742
25	第24组	120	360	68	143.784
26	第25组	60	361	58	115.712
27	第26组	3	260	17	57.025
28	第27组	74	349	74	125.626
29	第28组	15	368	51	96.037
30	第29组	44	358	70	113.596
31	第30组	29	359	46	97.947
32	权重	0.403	0.188	0.408	

所以我们获得评价的模型 $F=0.403 \cdot X_1 + 0.188 \cdot X_2 + 0.408 \cdot X_3$

用此方法分别计算 30 类问题的得分,按得分大小排序,获得热度前五的问题

	A	B	C	D	E
1		点赞数	时间跨度/	总数	得分
2	第16组	3029	379	1385	1857.019
3	第2组	2101	370	170	985.623
4	第4组	2188	352	38	963.444
5	第5组	1041	366	177	560.547
6	第6组	1030	386	174	558.65
7	第20组	623	371	459	508.089
8	第19组	237	359	137	218.899
9	第17组	218	367	152	218.866
10	第15组	318	367	53	218.774
11	第1组	136	367	133	178.068
12	第10组	150	363	117	176.43
13	第14组	133	362	134	176.327
14	第12组	158	360	99	171.746
15	第13组	85	365	115	149.795
16	第21组	116	371	76	147.504
17	第24组	120	360	68	143.784
18	第3组	65	363	93	132.383
19	第27组	74	349	74	125.626
20	第18组	50	432	52	122.582
21	第25组	60	361	58	115.712
22	第23组	66	356	52	114.742
23	第9组	68	345	55	114.704
24	第29组	44	358	70	113.596
25	第22组	76	359	37	113.216
26	第7组	68	339	43	108.68
27	第11组	47	349	49	104.545
28	第30组	29	359	46	97.947
29	第28组	15	368	51	96.037
30	第8组	34	382	25	95.718
31	第26组	3	260	17	57.025
32	权重	0.403	0.188	0.408	

根据图前五的热度问题分别为，第 16 组，第 2 组，第 4 组，第 5 组，第 6 组做成热点分析表

	A	B	C	D	E	F
1	热点问题表					
2	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
3	1	16	1857.019	2019-01-02至2020-01-25	A市社区	生活不便问题
4	2	2	985.263	2019-01-02至2020-01-08	A2区居民	存在危险隐患扰民问题
5	3	4	963.444	2019-01-16至2020-01-04	A7县小区居民	垃圾清理，扰民，违法停车问题
6	4	5	560.547	2019-01-06至2020-01-07	A4区居民	黑帮势力威胁，扰民问题
7	5	6	558.65	2019-01-05至2020-01-26	A市社区	扰民问题

然后代码导出相应问题的详细信息为热点问题留言明细表

	A	B	C	D	E	F	G	H	I
1		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
2	0	1	188006	A00010294	A3区一米路	2019-02-28 11:25:05	座落在A市	0	0
3	1	1	188170	A88011323	A市6路公	2019-12-23 08:50:24	12月21日	0	0
4	2	1	188399	A00097934	A市利保宣	2019-07-03 06:23:25	您好，我想	0	0
5	3	1	188416	A00029755	请给K3县乡	2019-06-20 20:38:47	K3县的乡村	0	0
6	4	1	188467	A00050188	投诉A市温	2019-03-28 19:57:19	退费之日起	1	0
7	5	1	188553	A00092235	A市沙坪老	2019-06-06 21:58:22	在沙坪老街	0	0
8	6	1	188560	A00075321	A市德鸿餐	2019-10-06 15:01:36	a我姐杜四	0	0
9	7	1	188665	A00010623	A市松雅湖	2019-03-26 00:18:26	您好 谢谢	1	0
10	8	1	188679	A00010384	希望A市政	2019-12-19 12:48:05	2019年12月	5	0
11	9	1	188801	A909180	投诉滨河苑	2019-08-01 00:00:00	尊敬的张市	0	0
12	10	1	188820	A00028135	A市星沙城	2019-02-21 11:38:34	领导好：巨	0	0
13	11	1	188856	A00010423	A3区谷园路	2019-08-14 12:12:40	高新区谷园	0	0
14	12	1	188895	A00063285	票牛A市分	2019-06-28 09:31:25	22日A市草	0	0
15	13	1	188930	A00035285	A市C5市中	2019-04-09 19:14:41	投诉C5市中	0	0
16	14	1	188941	A00018142	中建嘉和城	2019-01-07 10:16:45	近日A市普	1	0
17	15	1	188972	A00041922	A市内道路	2020-01-06 10:58:57	您好！第一	0	0
18	16	1	189176	A00093880	A3区威嘉湖	2019-07-24 11:04:12	A3区威嘉湖	0	0
19	17	1	189180	A00010651	A市人才购	2019-06-18 09:51:36	胡书记，您	3	0
20	18	1	189245	A00042948	A市北辰三	2019-03-19 11:22:30	A市北辰三	0	0
21	19	1	189247	A00010786	建议在“手	2019-12-21 09:45:46	互联网时代	0	0
22	20	1	189294	A00083527	A市南站里	2020-01-01 18:00:16	1、进候车	0	0
23	21	1	189313	A00010375	A市磁悬浮	2019-05-21 18:51:29	A市南站到	2	0

五、结果处理与分析

5.1 答复意见的评价

5.2 数据预处理

首先我们从附件 4 中提取出留言详情和答复意见两列，同样其中不乏许多重复词，无义词，以及错误词。经过数据清洗，中文分词后，留作后用。

5.3 word2vec 算法

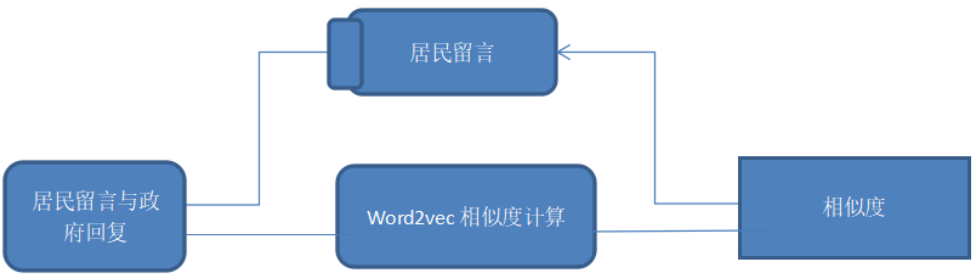
我们首先要用到模块 Gensim。它是一款开源的第三方 Python 工具包，用来从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达。它可以支持包括 TF-IDF，LSA，LDA 等算法，并且提供了诸如相似度计算，信息检索等一些常用任务的 API 接口。

Word2Vec 是 Google 公司于 2013 年发布的一个开源词向量工具包。Word2vec 可以根据给定的语料库，通过优化后的训练模型快速有效地将一个词语表达成向量形式，为自然语言处理领域的应用研究提供了新的工具。

Word2vec 算法的步骤：

- (1) 先对居民留言以及政府回复进行数据清洗
- (2) 对每一个居民的留言以及政府的回复一一对应
- (3) 用 word2vec 将文本向量化然后进行对比相似度的计算
- (4) 相似度越高说明回复的相关性以及可解释性越高

具体流程图如下



5.4 文本相关性

将处理好的数据用于训练 gensim 中的 similarities 模型，得出的结果即可为相关性的评判标准。

5.5 文本完整性

文本的完整性只需要将居民留言换成我们想要回复的模板如（网民，你好，对于你提出的什么问题，我们现在给出如下答复，谢谢您的提议）然后与政府回复进行对比即可得出结果的相似度越高那么回复的完整性也就越高。

5.6 文本可解释性

我们使用了 LDA 主题提取。
LDA 模型的主要思想是：首先将每一篇文章看作所有主题的一个混合概率分布，而其中的每一个主题都要视为在单词上的一个概率分布。因此当有 D 篇文档、T 个主题和 W 个单词时,在一篇文档中的第 i 个单词的概率可以表示为：
$$P(w_i) = \sum_j P(z_i=j)P(w_i|z_i=j)$$

将处理好的留言详情与答复意见分别导入形成语料库，然后训练 LDA 模型，分别提取留言详情与答复意见的主题特征词，形成 5 个主题，部分结果如图

```
0.021*年" + 0.010*您好" + 0.009*网友" + 0.009*工作" + 0.008*情况" + 0.008*县" + 0.007*回复" + 0.006*收悉" + 0.006*建设" + 0.006*2019"
0.011*业主" + 0.008*年" + 0.008*小区" + 0.006*开发商" + 0.005*领导" + 0.005*部门" + 0.004*县" + 0.004*政府" + 0.004*市" + 0.004*相关"
```

然后对比留言详情与答复意见主题特征词的相同个数，得到解释评分，然后分别取得 5 个特征的评分，算取平均值得到最终的可解释性评分。

	A	B	C	D
1		留言详情与答复意见关联程度		评价
2	主题1		0.2	0.44
3	主题2		0.4	
4	主题3		0.4	
5	主题4		0.6	
6	主题5		0.6	

参考文献

- [1] 祝永志,荆静.基于 Python 语言的中文分词技术的研究[J].通信技术,2019,52(07):1612-1619.
- [2] 庞景安.Web 文本特征提取方法的研究与发展[J].情报理论与实践,2006(03):338-340+367.
- [3] 屈探春.基于空间向量模型的先秦文献相似性研究[J].文教资料,2014(30):160-163.
- [4] 唐勇.基于朴素贝叶斯算法对论坛文本分类的技术实现 [J]. 电脑知识与技术,2014,10(32):7612-7615.
- [5] 王斌.基于朴素贝叶斯算法的垃圾邮件过滤系统的研究与实现 [J]. 电子设计工程,2018,26(17):171-174.
- [6] 李荟娆. K-means 聚类方法的改进及其应用[D].东北农业大学,2014.
- [7] 陈佩. 主成分分析法研究及其在特征提取中的应用[D].陕西师范大学,2014.
- [8] 周练.Word2vec 的工作原理及应用探究[J].科技情报开发与经济,2015,25(02):145-148.
- [9] 唐晓波.基于 LDA 模型和微博热度的热点挖掘.武汉大学信息系统研究中心 2014