摘要

本文针对“智慧政务”的文本挖掘问题，建立一级标签分类模型、基于熵值法的综合评价模型，并采用潜在语义分析、聚类分析、主成分分析，实现了留言的一级标签分类、热点问题的挖掘和排名、留言答复质量的评价。

针对问题一：为实现留言的一级标签分类并作评价，建立**一级标签分类模型**。构建并划分数据集，剔除文本标记和特殊字符，使用 *jieba* 分词进行中文分词并去停用词。基于 $TF-IDF$ 对文本特征提取，以**向量空间模型**表示留言文本，用 CHI 选择特征。为进一步对词向量降维，再结合基于文本余弦相似度的**一对一** SVM **多分类**方法，建立留言的一级标签分类模型，实现留言的一级标签分类。引入 $F-Score$ 评价分类模型，求得 $\boldsymbol{F_1}$**=0.8683**，表明分类效果较好。

针对问题二：针对热点问题的挖掘和排名，采用**聚类分析**、**主成分分析**。对留言向量化后，由词汇-文本矩阵的奇异值分解对向量语义化，实现**潜在语义分析**对文本向量进行语义空间降维，再计算文本的**余弦相似度**，结合 $K-means$ **聚类算法**建立文本聚类模型，实现热点问题的挖掘，并按热点汇编。对已挖掘的热点进行主成分分析，以热点的留言数、留言时间密集度、点赞数、反对数为评价指标。以指标的信息贡献率为权重计算主成分综合得分，以此为热度指数，对热点问题进行排名，截取排名前5的热点并制作热点问题表、热点留言明细表（详见附件），热点问题表概要如下：

热度排名	1	2	3	4	5
热度指数	0.89	0.62	0.53	0.48	0.46
时间范围	2019.1.6至2020.1.6	2019.11.13至2020.1.25	2019.7.21至09.25	2019.2.14至12.3	2019.1.6至9.12
地点/人群	A1-A7此7区	A2区丽发新城小区	A5魅力之城小区	A7县星沙镇四区	A3区西湖街道
问题描述	停车位贵、难求	搅拌站噪音扰民	临街烧烤夜宵摊油烟污染、噪音扰民	凉塘路建设、改造不合理	茶场村拆迁未果

针对问题三：为对留言的答复意见给出评价方案，建立**综合评价模型**。对留言和答复预处理后，选取每条答复包含的问候、对点回答、感谢与致歉，答复紧凑度作为评价指标。采用极差化法对总体的指标标准化，再用**熵值法**计算指标的权重系数。构建一个偏离理想范围的加权距离，运用**改进后的** $TOPSIS$ **法**，定义并计算**质量劣值，以其正负对答复意见质量的好坏进行区分，负值的大小可清晰区别质量较差的答复意见质量不好的程度**。

关键词：文本多分类、 SVM **、余弦相似度、** $K-means$ **聚类、综合评价**

目录

一、问题重述

1.1挖掘意义

在大数据和信息化的时代特征下，网络平台无疑为收集海量的文本数据提供了便捷，如何快速有效精确地筛选出主要信息以及对其分类、答复，不得不说这对人类来说一直是一个需要不断精化、持续进步的难题。为了构建智能文本挖掘模型，学界对于机器读取的研究从未止步。

机器读取技术的发展对信息检索、答复系统、机器翻译等自然语言处理研究任务有积极作用，同时也能够直接改善搜索引擎、智能助手等产品的用户体验。因此，以读取筛选、文本挖掘为契机研究机器自然语言的技术，在有限的信息范围内要做到准确全面的处理，具有重要的研究与应用价值。

网络问政平台作为一种新兴模式，以其快捷、不受时空限制等优点而受到政府机构的青睐。借助网政平台收集群众反馈的海量信息数据，实是了解民意、汇聚民智、凝聚民气的重要渠道。如果能从群众留下的信息，敏锐地捕捉信号，不仅能够提升政府的管理水平及，同时也能更好地为群众百姓提供服务，进行互赢模式间的双向信息传递。

1.2题设数据

已知该数据来源于互联网公开渠道，关于留言者信息及留言内容的4个附件及题目所给信息。具体如下：

一级标签分类模型：F-Score（留言内容） 。

附件一：体系为内容分类三级标签的样例；

附件二、三、四：热点问题留言的分类、反馈、答复时间的具体信息；

1.3需解决的问题

针对本题相关留言数据信息，基于利用自然语言处理和文本挖掘的方法现提出以下几个需要解决的问题：

问题一：群众留言分类。针对处理网政平台的群众留言时，按照题目所给划分体系（附件一）对留言进行分类并分派至相应的职能部门处理，据附件二建立一级标签分类模型；

问题二：热点问题挖掘。针对附件三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”；

问题三：答复意见的评价。针对附件 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。

二、模型假设

假设一：所有数据是真实可靠的；
假设二：数据信息中不带个人偏向的主观意愿；
假设三：所有的文本都是有意义的，没有无意义的乱码或者骚扰留言。

三、符号说明

符号	名称
n	留言容量
d_i	第i条留言
m	类别标签容量
D	训练集中的文本总数
t_k	为d中的特征项
w_{td}	为t_k对应的权重
df_t	训练集中包含特征项t的文档数
l	特征空间的大小
c_j	类别
A	表示类别c_j中出现特征项t_k的文本数
B	表示除了类别c_j的其他类别中出现特征项t_k的文本数
C	表示类别c_j中没有出现特征项t_k的文本数
D	表示类别c_j之外的其他类别中没有出现特征项t_k的文本数
N	训练集的文本数
$P(c_j)$	概率
N_j	类别c_j中的文本数
（H_1、H_2）	几何间隔
r_{ij}	表示分类器区分类别i和j时偏向类别i的概率
P_j	模型分类的准确度
$n_q(c_j)$	本属于c_j类的全部留言
$U_{m\times m}$	词汇矩阵中的每一行主要承载着文本关键词信息
$V^T_{n\times n}$	文档矩阵
r_{ij}	第i个评价对象的第j项指标的相关系数
F_j	第j主成分
(l_j,u_j)	评价指标设定的一个理想范围

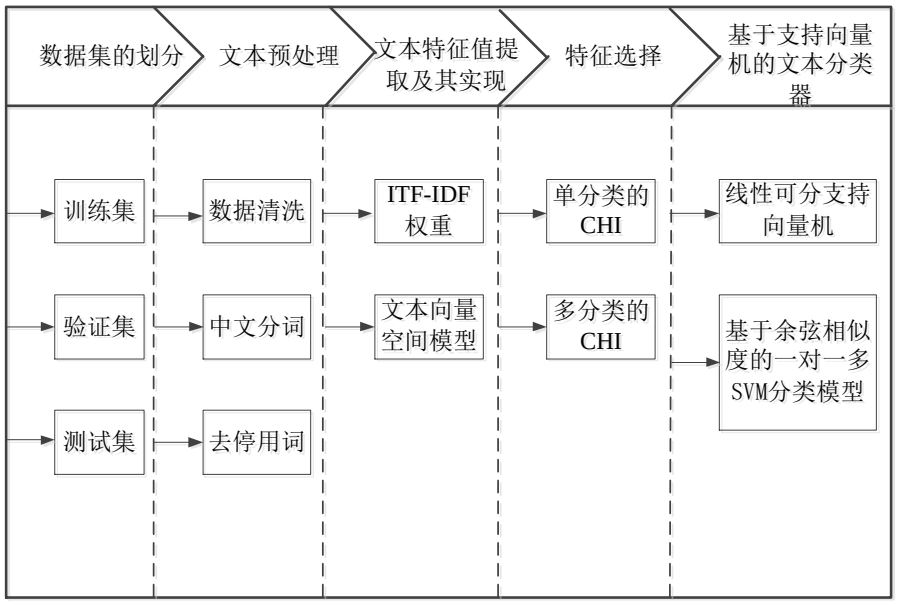
四、问题分析

题目给出该网政平台上的群众留言以及相关部门对部分群留言的答复意见的相关信息数据，要求利用以上条件建立模型，通过分析和相关计算对留言进行分类、归类及评价。

即解决政府在网政平台上从接收到群众留言，到发现集中的热点问题，再分类反馈给各相关部门，并给出统一式高质量的答复的这一系列为方便行政效率的自动化操作的实际问题。具体问题的分析如下所示。

4.1对问题一的分析

对于问题一，根据附件1的分类体系标准对附件2的留言进行一级标签分类，并评价分类方法。据题目已知信息可构建数据集并划分，进而预处理文本。基于 $TF-IDF$ 对文本特征提取，构建空间模型，用 χ^2 统计选择特征。为进一步对词向量的降维，进而结合一对一 SVM 多分类模型，从而建立留言的一级标签分类模型，实现一级标签分类。具体流程步骤如图所示：



图：问题一思路流程分析图

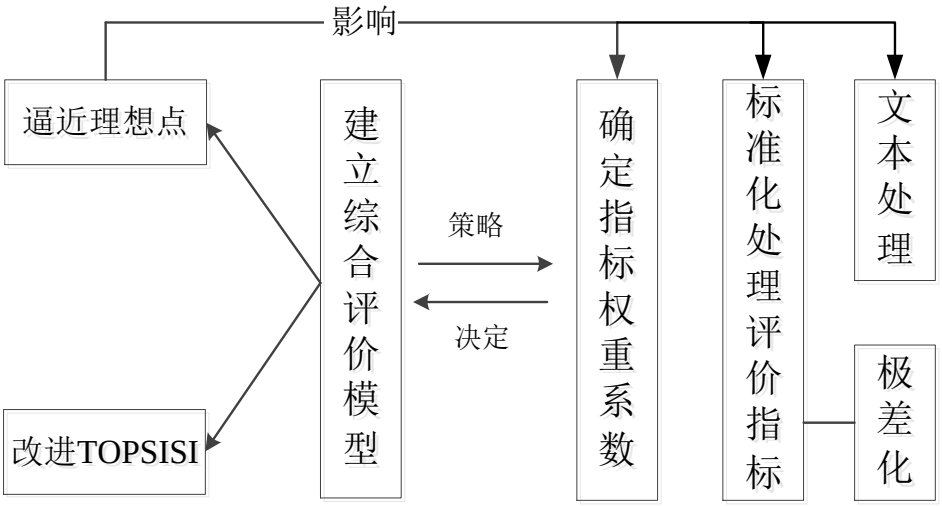
4.2对问题二的分析

对于问题二，基于留言信息挖掘热点，给出排名前 5 的热点问题。在此对留言预处理后，采用潜在语义分析 LSA 对文本向量进行语义空间降维，计算文本余弦相似度，建立基于 $K\text{-}means$ 聚类的文本聚类模型，实现热点问题挖掘，并按热点汇编。

采用主成分分析法，将每个热点中的留言数、留言时间密集度、点赞数、反对数作为评价指标，分析已挖掘的热点样本。基于每个热点权重计算后的主成分综合得分，作为热度指数进行前5的排名。

4.3对问题三的分析

对于问题三，评价留言答复意见的质量。经预处理留言与答复文本后，选取每条答复中问候、对点回答、感谢与致歉，答复紧凑度作为评价指标。采用极差化法归一指标，再用熵值法确定评价指标的权重系数。基于改进逼近理想点法，构建一个偏离理想范围的加权距离，定义并计算质量劣值，其正负值对应答复意见好坏。



图：问题三思路流程分析图

五、数据分析

5.1数据的预处理

5.1.1数据集的划分

先将附件2中由所有“留言详情”栏的文本构成的数据集，容量为n，每条留言表示为$d_i(i=1,2,\cdots,n)$。一级标签共m类，类别集合为$C=\{c_1,c_2,\cdots,c_m\}$。

为训练后续模型、分类留言，将数据集划分为3部分：

（1）60%为训练集，用于训练不同的分类模型。

（2）20%为验证集，用于调整模型。使用交叉验证来挑选最优的模型，通过不断的迭代来改善模型在验证集上的性能。

（3）20%为测试集，用于评估模型的性能。

在划分中，采用*sklearn*提供的函数，实现分层抽样，以保证3部分数据的一级标签分布均匀性。

5.1.2文本预处理

（1）数据清洗

①文本标记处理

经浏览发现附件2中留言的“留言详情”栏附有*HTML*标签、*URL*地址等文本标记，没有有效的分类信息，所以需要清除这部分多余的内容，去除噪声，为

后续分类奠定基础。

②特殊字符处理

为得到纯文本，考虑将“留言详情”栏中的标点符号去除，留言便只剩下文字部分。

（2）中文分词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，本文采用 *Python* 开发的一个中文分词模块—— *jieba* 分词，对附件2中每一条精处理过的留言进行中文分词， *jieba* 分词用到的算法详见模型的求解。

（3）去停用词

关于“是”、“的”等停用词，主要包括一些副词、形容词及其一些连接词，这些停用词对表达文本语义没有贡献，而且出现的频率较高，增加了分类器的消耗。因此，将这些词清理掉，提高文本分类效率，降低处理复杂度。在此基础上，能够使模型更好地去拟合实际的语义特征，从而增加模型的泛化能力。

本文采用先导入停用词库形成列表，接下来对一个单独句子处理，先通过 *re.findall* 提取出句子中的每一个单独汉字，再用 *join* 函数把汉字连接成没有分隔符的句子，再用 *jieba.lcut* 将句子分词形成列表，这里使用的是精准切割，最后通过 *for* 循环，倒序检查列表的每一个元素，如果这个元素出现在停用词列表中，者将其删除，最后返回这个列表。

以附件2的部分留言为样本，去除停用词后的部分结果示例如下图所示：

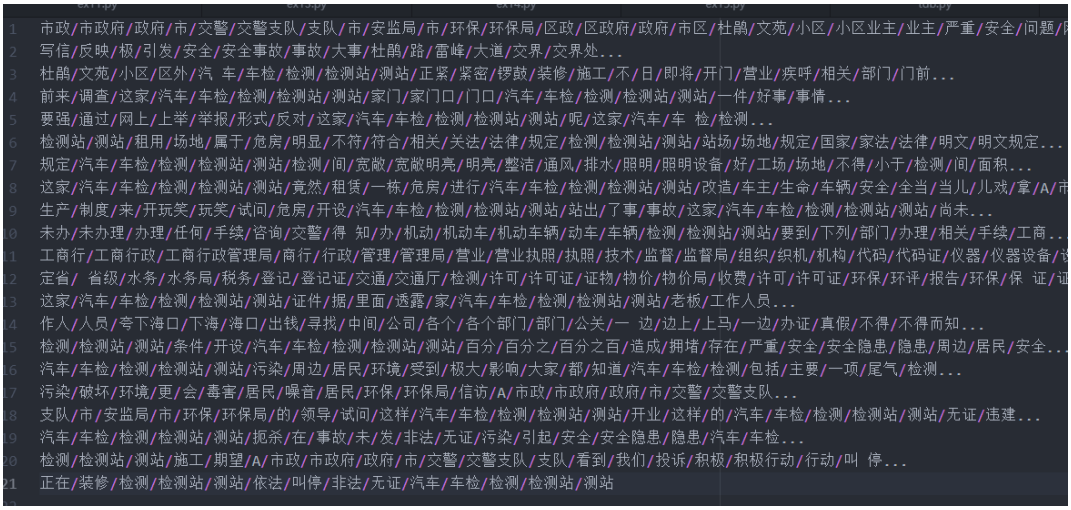
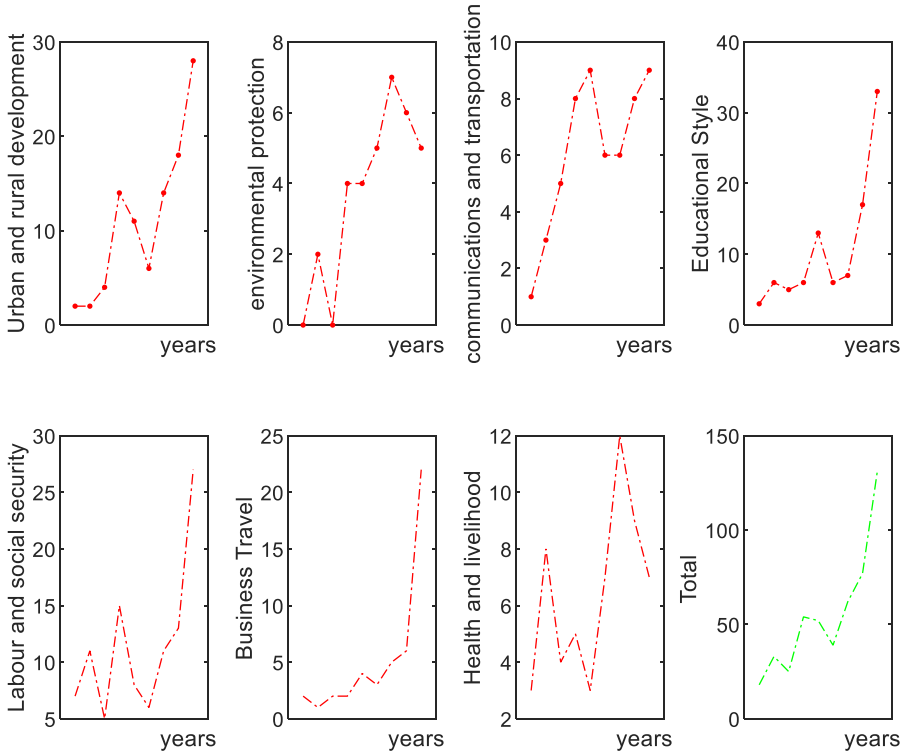


图:停用词过滤后分词结果

5.2数据信息的整理

在此只选用了所给附件二的留言信息，针对一级标签所有类别中（即城乡建设、环境保护、交通运输、教育文体、劳动与社会保障、商贸旅游、卫生计生），在2011~2019年的留言信息进行了整理与记录（鉴于2020年并不包括全月份）。具体如下图所示：



图：不同标签的留言信息随时间变化

由图可知：①图一至图七中，分别是一级标签类别下城乡建设、环境保护、交通运输、教育文体、劳动与社会保障、商贸旅游、卫生计生在2011~2019年期间，留言信息容量呈渐增式趋势。②图八描述的是2011~2019年期间，每年所有留言信息的变化趋势。③在这9年期间留言容量的渐增，一方面可能是社会背景的缘故，群众通过网络问政平台留下信息因得到普及，另一方面也反映了群众所反馈的问题是否实际得到解决的现状。

六、问题一模型的建立与求解

6.1模型的建立

本题要求根据样本留言，对留言进行一级标签分类，并评价分类方法。构建数据集并划分，剔除文本标记和特殊字符，进行中文分词并去停用词。基于 $TF-IDF$ 对文本特征提取，以向量空间模型表示留言文本，用 χ^2 统计选择特征。为进一步对词向量降维，再结合基于文本余弦相似度的一对一 SVM 多分类模型，建立留言的一级标签分类模型，实现按一级标签分类。

平衡考虑准确率和召回率的影响，引入 $F-Score$ 作为综合指标，对以上分类模型进行评估。

6.1.1文本特征提取及表示

（1）***TF-IDF* 权重**

为将文本转化为结构化的词向量，考虑到：**词的重要性随着它在文本中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降**。故此

处采用$TF-IDF$权重，以评估单个词对训练集中的一个文本的重要程度。

①计算单个不重复词的词频（TF），如下：

$$tf_{td}=\frac{\text{特征项}t\text{在文本}d\text{中的出现次数}}{\text{文本}d\text{总词数}}$$

词频反映特征项在文本中出现的频率，是作为权重最直观的一种方法。

②引入词频的权重调整系数——逆文档频率（IDF），反映了特征项在文本数据集中的分布情况。计算如下：

$$idf_{td}=\ln\frac{D}{df_t+\beta}$$

D为训练集中的文本总数，df_t为训练集中包含特征项t的文档数。同时实验中我们为上述公式添加一个经验因子β，β常取值为0.01、0.1、1。

③综上2项，计算词频-逆文档频率（$TF-IDF$），作为文本d中特征项t的权重w_{td}，如下：

$$w_{td}=tf_{td}-idf_{td}=tf_{td}\ln\frac{D}{df_t+\beta}$$

TF对于高频、区分度不强的特征项对分类性能影响较大的缺点可以通过IDF来改善。同时，IDF对于对文档类别判断具有决定作用的稀有特征项，又能增加它们的重要性。故$TF-IDF$能较好表示文本特征权重。

（2）文本的向量空间模型

基于$TF-IDF$对文本特征提取，将文本词向量化。考虑到文本数据集的规模较大，这里选用向量空间模型（VSM），如下：

文本d被看成由二元特征组组成的特征向量，表示如下：

$$V(d)=\{(t_1,w_1(d)),(t_2,w_2(d)),\cdots,(t_l,w_l(d))\}$$

t_k为d中的特征项，w_k为t_k对应的权重，$k=1,2,\cdots,l$。l为特征空间的大小。

当数据集中的特征空间确定时，每个文本对应特征空间中的一个点，则上式简化为：

$$V(d)=W=(w_1(d),w_2(d),\cdots,w_l(d))$$

上式即为文本d的特征权重向量，实现了对文本d的特征提取。

上述文本特征抽取将全部特征项筛选若干个候选特征项，这时需要构建一个词袋，根据留言文本的特征项对应词袋中的位置，组成统一维数的向量：

$$C=(t_1,t_2,\cdots,t_i,\cdots,t_l)$$

其中C为词袋集合，t_i是每个词在向量中对应的位置。

这样留言文本信息根据词袋组成了同一维数的词向量，再通过$TF-IDF$将它们向量化得到一个词汇-文本矩阵：

$$\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_l \end{array}\begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{l1} & w_{l2} & \cdots & w_{\ln} \end{pmatrix} \end{array}$$

6.1.2特征选择

为降低文本 d 特征向量的维度，减轻分类器的学习负担，提高分类准确率，进行文本特征选择，此处采用 χ^2 统计。

（1）单类的 *CHI*

假设特征项 t_k 与类别 c_j 满足一阶自由度的 χ^2 分布。 χ^2 统计量的计算公式如下：

$$CHI(t_k,c_j)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$

其中 A 表示类别 c_j 中出现特征项 t_k 的文本数，B 表示除了类别 c_j 的其他类别中出现特征项 t_k 的文本数，C 表示类别 c_j 中没有出现特征项 t_k 的文本数，D 表示类别 c_j 之外的其他类别中没有出现特征项 t_k 的文本数， N 为训练集的文本数，且 $N=A+B+C+D$ 。

（2）多分类的 CHI

基于以上，对多分类采用加权平均法。先计算特征项和每个类别之间的 χ^2 统计值，然后以每类出现概率 $P(c_j)$ 为权对所有类别的 χ^2 统计值求加权平均值，作为该特征项 t_k 的最终 χ^2 统计值，如下：

$$P(c_j)=\frac{N_j}{N}$$

$$CHI_{avg}(t_k)=\sum_{j=1}^{m}P(c_j)CHI(t_k,c_j)$$

其中 N_j 为类别 c_j 中的文本数。

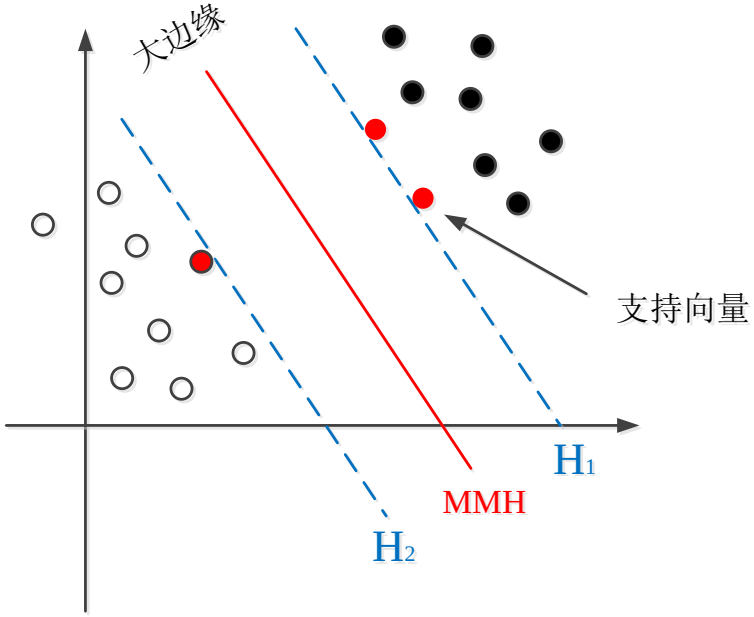
通过计算出的 $CHI_{avg}(t_k)$ 对特征项进行选择。其值越大，说明该特征项含有的类别信息越多，那么该特征就有更大概率被选择。

6.1.3基于支持向量机的文本分类器

由于多分类问题不易直接分类，考虑采取将多分类转化为多个二分类的分解策略，并选取分类策略，判定最终结果。

（1）线性可分支持向量机

对一个线性可分的二类分类问题，可找到一个将两个类别区分开的超平面，这样的超平面有很多个，下面要找到最优的那一个。具体如图所示：



图：线性可分的支持向量机示意图

直观上，为两类数据找到最优分类超平面（*MMH*），它尽可能区分属于不同类的数据。即使这个平面分别与两类数据的最近的数据点之间的距离之和最大。与 *MMH* 平行，且正好经过两类数据中距离 *MMH* 最近的数据点的超平面，记为几何间隔（H_1、H_2）。

假设存在一个线性可分样本集$\{x_i, y_i\}, i=1,2,\cdots,n, x\in R^c, y\in\{-1,1\}$表示两个类别，通过间隔最大化或者求解对应的凸二次规划问题得到的分离超平面是$WX+b=0$，对应的决策函数为：

$$f(x)=\mathrm{sgn}(WX+b)$$

当线性判别函数$|g(x)|=1$时，样本与分类超平面的距离最近，分类间隔为$\frac{2}{\|W\|}$。为使分类间隔最大，即使$\frac{\|W\|}{2}$最小。转化为二次规划，如下：

$$\min\phi(W)=\frac{1}{2}\|W\|^2$$

$$s.t.\ \ y_i(Wx_i+b)\geq 1, i=1,2,\cdots,n$$

引入拉格朗日因子a_i，构造拉格朗日公式，上式转化为对偶问题：

$$\max w(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}y_iy_ja_ia_jX_i^TX_j$$

$$s.t.\ \ \sum_{i=1}^{n}y_ia_i=0, a_i\geq 0, i=1,2,\cdots,n$$

求解上式的最优化问题，即可得到最优化问题的最优解。

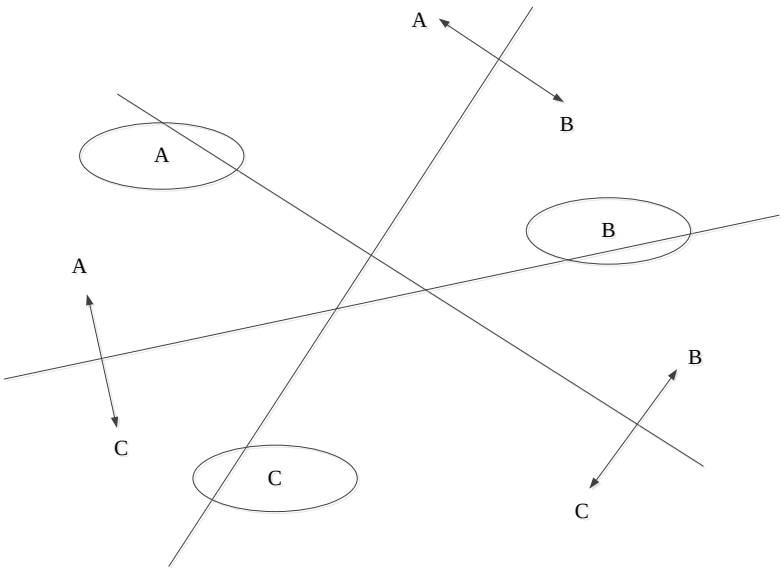
把判决函数解释成$\hat{y}=1$的对数几率，则 *SVM* 分类器的概率输出为：

$$P(\hat{y}=1|x)=\sigma(Af(x)+B)$$

得到此二分类器的输出概率，做为后续多分类投票策略的依据。

（2）基于余弦相似度的一对一 *SVM* 多分类模型

在任意两类样本之间设计一个二分类器，因此 m 个类别的样本就需要设计 $\frac{m(m-1)}{2}$ 个二分类器。任意两个分类器之间，关系如图：



图：分类器间两两关系图

将每个二分类器的输出构建一个矩阵 R，如下：

$$R=\begin{pmatrix} - & r_{21}\cdots r_{1m} \\ r_{21} & -\cdots r_{2m} \\ \vdots & \vdots \\ r_{m1} & r_{m2}\cdots - \end{pmatrix}$$

其中，$r_{ij}\in[0,1]$ 表示分类器区分类别 i 和 j 时偏向类别 i 的概率。

①余弦相似度

为进一步缩减矩阵 R 的维数，计算测试集文本 x 与每个类别 y 的余弦相似度：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

$$\cos(x,y)=1-\frac{1}{2}d^2(x,y)$$

将待分类文本 d 的每个 $\cos(x,y)$ 存储在向量 $d=(d_1,d_1,\cdots,d_m)$ 中。去除与文本余弦相似度最小的类别，**将 R 矩阵的维数从 m 维缩减到 p 维**（$p\le m$）。

②分类策略

以下根据分类策略，判定最终结果，此处采用权重投票策略：

$$Class=\arg\max_{i=1,\cdots,p}\sum_{1\le j\ne i\le p}r_{ij}$$

以上对每个类别参与的所有分类器的结果进行投票，结果最大的类别即作为待分类文本 d 最后的判决类别 $Class$。

6.1.4一级标签分类模型的建立

综上，基于 $TF-IDF$ 的特征提取，用向量空间模型表示留言文本，用 χ^2 统

计选择特征，再结合基于文本余弦相似度的一对一 *SVM* 多分类模型，建立文本留言的一级标签分类模型，如下：

$$
\begin{cases}
tf_{td}=\dfrac{\text{特征项}t\text{在文本}d\text{中的出现次数}}{\text{文本}d\text{总词数}} \\
idf_{td}=\ln\dfrac{D}{df_t+\beta} \\
w_{td}=tf_{td}-idf_{td}=tf_{td}\ln\dfrac{D}{df_t+\beta} \\
d=W=(w_1,w_2,\cdots,w_l) \\
CHI(t_k,c_j)=\dfrac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \\
P(c_j)=\dfrac{N_j}{N},CHI_{avg}(t_k)=\sum_{j=1}^{m}P(c_j)CHI(t_k,c_j) \\
\max w(a)=\sum_{i=1}^{n}a_i-\dfrac{1}{2}\sum_{i,j=1}^{n}y_iy_ja_ia_jX_i^TX_j \\
\quad s.t.\ \sum_{i=1}^{n}y_ia_i=0,a_i\geq 0,i=1,2,\cdots,n \\
P(\hat{y}=1|x)=\sigma(Af(x)+B) \\
R=\begin{pmatrix} - & r_{21}\cdots r_{1m} \\ r_{21} & -\ \cdots r_{2m} \\ \vdots & \vdots \\ r_{m1} & r_{m2}\cdots - \end{pmatrix} \\
d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2},\cos(x,y)=1-\dfrac{1}{2}d^2(x,y) \\
Class=\arg\max\limits_{i=1,\cdots,p}\sum\limits_{1\leq j\neq i\leq p}r_{ij}
\end{cases}
$$

通过训练集，训练以上模型，对测试集的均匀留言文本进行分类，得到待分类文本 *d* 的 *Class* 即为最后的判决类别。

6.2模型的求解

（1）基于 *Python*，对数据集划分中，采用 *sklearn* 提供的函数，实现分层抽样，以保证训练集、验证集、测试集3部分数据的一级标签分布均匀性。

（2）分词采用 *Python* 开发的一个中文分词模块—— *jieba* 分词器，算法如下：

a.使用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生成词情况所构成的有向无环图。

b.再采用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。

c.对于未登录词，采用了基于汉字成词能力的 *HMM* 模型，并使用了 *Viterbi* 算法求解，来预测分词。

（3）下载并建立停用词字典。通过维护一个停用词表，在分词后将停用词去除。

（4）基于 $TF-IDF$ 对文本特征提取，以向量空间模型（ SVM ）表示留言。

$Word2vec$ 是一个 $Estimator$ ，它采用一系列代表文档的词语来训练 $word2vec\ model$。该模型将每个词语映射到一个固定大小的词向量，将文本结构化。

（5）用 χ^2 统计选择特征。先计算特征项和每个类别之间的 χ^2 统计值，然后以每类出现概率 $P(c_j)$ 为权对所有类别的 χ^2 统计值求加权平均值，作为该特征项 t_k 的最终 χ^2 统计值，对特征向量进行降维。

（6）为进一步对词向量降维，再计算文本余弦相似度，去除与文本余弦相似度最小的类别，缩减矩阵 R 的维数。

（7）用 $Python$ 求解一对一 SVM 多分类模型，以下根据分类策略，判定最终结果，此处采用权重投票策略：

$$Class=\arg\max_{i=1,\cdots,p}\sum_{1\le j\ne i\le p} r_{ij}$$

以上对每个类别参与的所有分类器的结果进行投票，结果最大的类别即作为待分类文本 d 最后的判决类别 $Class$ 。

6.3模型的评估

为对以上分类模型进行评估，首先考虑准确率：

$$P_j=\frac{n_z(c_j)}{n_c(c_j)}$$

其中 $n_c(c_j)$ 为通过分类模型标注为 c_j 类的留言数，$n_z(c_j)$ 为在 $n_c(c_j)$ 中确实属于 c_j 类的留言数。 P_j 即反映了模型分类的准确度。

再考虑模型分类的全面性，即召回率，如下：

$$R_j=\frac{n_z(c_j)}{n_q(c_j)}$$

其中 $n_q(c_j)$ 为本属于 c_j 类的全部留言。

我们自然是希望精确率和召回率都越高越好。但通常情况下，一般来说准确率和召回率呈负相关，二者之间是矛盾的，不可兼得，所以我们引入调和平均数 $F-Score$ 作为综合指标，如下：

$$F_1=\frac{1}{n}\sum_{j=1}^{m}\frac{2P_jR_j}{P_j+R_j}$$

上式 F_1 平衡了准确率和召回率的影响，能较为全面地评价一个分类模型，F_1 的计算值越大，模型分类效果越好。

七、问题二模型的建立与求解

7.1模型的建立

本题要求请根据附件留言，挖掘热点问题，定义热度指标，并对热点排名。对留言预处理后，通过词汇-文本矩阵的奇异值分解对向量语义化，即采用潜在语义分析 LSA 对文本向量进行语义空间降维，再计算文本余弦相似度，建立基于 $K\text{-}means$ 聚类的文本聚类模型，实现热点问题挖掘，并按热点汇编。

对已挖掘的热点进行主成分分析，分析样本，将每个热点包含留言数、留言时间密集度、点赞数、反对数作为评价指标。基于每个热点对应的权重计算主成分综合得分，以此为热度指数，对热点问题进行排名，并制作热点排名表。

7.1.1文本预处理

对附件3的留言经过文本标记处理、特殊字符处理、中文分词、去停用词等步骤，得到纯文字文本，方法与算法与问题一处理一致。

7.1.2 语义空间降维

按照理论上，当得出文本向量后就可以直接比较两向量的夹角的余弦值进行相似度的计算。但可以发现现在构造的词汇-文本矩阵是一个巨大矩阵，计算起来较困难。

另外，附件3的留言文本信息中存在同义词和近义词等词语，即使通过特征抽取转化得到的文本向量可能达不到自然语言属性本质的要求。

因此，这里需要借用潜在语义分析 （*Latent Semantic Analysis*，*LSA*）理论将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。而 *LSA* 对特征空间进行处理时用的关键技术就是奇异值分解（*Singular Value Decomposition*，*SVD*），在统计学上，它是针对矩阵中的特征向量进行分解和压缩的技术。

（1） 一般奇异值分解

奇异值分解可以将网页文本通过向量转换后的非完全正交的多维特征投影到较少的一个潜在语义空间中，同时保持原空间的语义特征，从而可以实现对特征空间的降噪和降维处理。

在线性代数中，奇异值分解是一类矩阵分解，是正规矩阵酉对角化的一种推广，对于任意的矩阵 A，如由招聘文本信息组成的词汇-文本矩阵。它的奇异值分解表达式如下：

$$A=U\sum V^T$$

其中 A 是一个 $m\times n$ 矩阵。U 是一个 $m\times n$ 阶的酉矩阵，$\sum$ 是一个 $m\times n$ 阶的对角矩阵，V^T 是一个 $n\times n$ 阶的酉矩阵，即 V 的共扼转置。在 $\sum$ 矩阵中对角线上的元素就是 A 的奇异值。奇异值分解适用于任何矩阵。A 的奇异值是由 A 矩阵唯一确定的，它的表达如下：

$$\begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}$$

其中 σ_r 为奇异值，$\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r\geq 0$。

由3.3.1奇异值分解定理 [10]：设 $A\in R^{m\times n}$，且 $Rank(A)-r\leq\min(m,n)$，存在正交矩阵 $U\in R^{m\times n}$ 和 $V\in R^{m\times n}$，也称为矩阵 A 的左右奇异向量，其中对角矩阵 $\sum\in R^{m\times n}$，$\sum=diag(\sigma_1,\cdots,\sigma_r),\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r\geq 0, UU^T=I_m, VV^T=I_n$。

定理3.3.2：在奇异值分解 $Rank(A)-r\leq\min(m,n)$，有 A 的 k 阶截距阵即

$$A_k = \sum_i^k u_i \sigma_i v_i = U_k \sum_k V_k^T$$

$$\|A - A_k\|_F = \min_{rank(B)\le k} \|A - B\|_F = \sqrt{\sum_{i=k+1}^n \sigma_i^2}$$

$$\|A - A_k\|_2 = \min_{rank(B)\le k} \|A - B\|_2 = \sigma_{k+1}$$

上述可以知道，在 $F-$范数中， A_k 是和 A 相似度最高的 k 秩矩阵，这将用于矩阵降维。

（2）词汇-文本矩阵的奇异值分解

通过上面对奇异值分解理论的介绍，可以知道经过奇异值分解能够得出三个矩阵，而每个矩阵对于原矩阵有着不同的含义和作用，下面就分析下词汇-文档矩阵分解后的三个矩阵的作用。对于矩阵词汇-文档矩阵 $A_{m\times n}$ 的奇异值分解可表示为：

$$A_{m\times n} = U_{m\times m} \sum_{m\times n} V_{n\times n}^T$$

$$U = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mm} \end{pmatrix}_{m\times n} \qquad V^T = \begin{pmatrix} d_{11} & d_{21} & \cdots & d_{n1} \\ d_{12} & d_{22} & \cdots & d_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1n} & d_{2n} & \cdots & d_{nn} \end{pmatrix}_{n\times n}$$

$U_{m\times m}$ 称为词汇矩阵中的每一行主要承载着文本关键词信息，其中的每个非零元索表示该类词的重要性，数值越大表示越重要。而 $V_{n\times n}^T$ 称为文档矩阵，它的每一列都表示留言信息中同一主题的文本，其中的每个元素代表这类文本中每条文本的相关性。

Σ 矩阵表示的是某类词与留言文本之间的相关性。在生成的这个“语义空间”中，大的奇异值对应的维度更具词的共性，而小的奇异值所对应的维度更具有词的个性。

$$A_{m\times n} = \begin{pmatrix} U_1 S V_1^T & U_1 S V_2^T & \cdots & U_1 S V_n^T \\ U_2 S V_1^T & U_2 S V_2^T & \cdots & U_2 S V_n^T \\ \vdots & \vdots & \ddots & \vdots \\ U_m S V_1^T & U_m S V_2^T & \cdots & U_m S V_n^T \end{pmatrix}$$

由上式可以看出，在 A 矩阵中，每一行 i 的信息都是由 U_i 和 Σ 决定的，而每一列 j 的信息则是由 V_j^T 和 Σ 决定。在对角矩阵 Σ 的信息主要由奇异值大小决定，奇异值越大，对 Σ 的影响也越大。对整个矩阵的影响也越大。因此，可以通过保留较大的奇异值，删去较小的奇异值，从而对矩阵进行行与列的降维处理。

另一方面，Σ 矩阵中的奇异值 $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$ 中，如果 $\sigma_i (i = 1, 2, \cdots, r)$ 的值比较小，则它对整个词汇-文本矩阵 A 的影响也小，所以就可以删除对矩阵 A 影响较小的 σ 以及对应的 U 和 V^T 的信息，保留影响较大主要信息，得到 $A_{m\times n}$ 的近似矩阵 A_k。

$$A_{m\times n}=U_{m\times m}\sum_{m\times n}V^{T}_{n\times n}$$

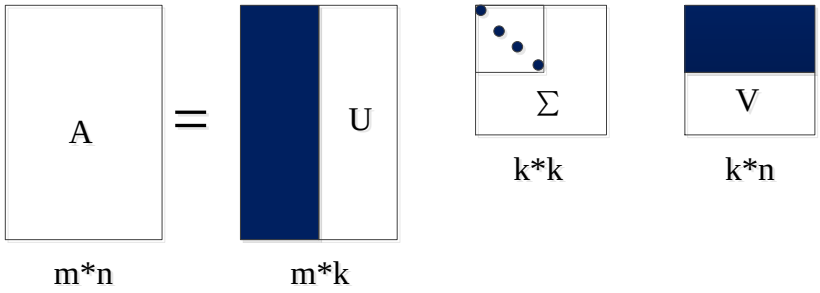
$$=\begin{pmatrix}t_{11}&t_{12}&\cdots&t_{1m}\\t_{21}&t_{22}&\cdots&t_{2m}\\\vdots&\vdots&\ddots&\vdots\\t_{m1}&t_{m2}&\cdots&t_{mm}\end{pmatrix}_{m\times n}\begin{pmatrix}\sigma_1&0&\cdots&0&\cdots&0\\0&\sigma_2&\cdots&0&\cdots&0\\\vdots&\vdots&\ddots&\vdots&\cdots&\vdots\\0&0&\cdots&\sigma_r&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&0&\cdots&0\end{pmatrix}_{m\times n}\begin{pmatrix}d_{11}&d_{21}&\cdots&d_{n1}\\d_{12}&d_{22}&\cdots&d_{n2}\\\vdots&\vdots&\ddots&\vdots\\d_{1n}&d_{2n}&\cdots&d_{nn}\end{pmatrix}_{n\times n}$$

$$\approx\begin{pmatrix}t_{11}&t_{12}&\cdots&t_{1k}\\t_{21}&t_{22}&\cdots&t_{2k}\\\vdots&\vdots&\ddots&\vdots\\t_{m1}&t_{m2}&\cdots&t_{mk}\end{pmatrix}_{m\times k}\begin{pmatrix}\sigma_1&0&\cdots&0\\0&\sigma_2&\cdots&\vdots\\\vdots&\vdots&\ddots&0\\0&\cdots&0&\sigma_k\end{pmatrix}_{k\times k}\begin{pmatrix}d_{11}&d_{21}&\cdots&d_{n1}\\d_{12}&d_{22}&\cdots&d_{n2}\\\vdots&\vdots&\ddots&\vdots\\d_{1k}&d_{2k}&\cdots&d_{nk}\end{pmatrix}_{k\times n}$$

$$\approx A_k=U_{m\times k}\sum_{k\times k}V^{T}_{k\times n}$$

这样在不影响对留言文本分析的结果的同时对矩阵进行降维处理，简化了运算的复杂度。

以上用一个形象的图表示就是：



图：语义空间降维处理图

在很多情况下，前10%甚至1%的奇异值的和占了整个奇异值之和的99%。k值的选取影响了近似矩阵的相似性，k值取越大，包含的的主要的信息越多，对次要信息的删除能力越小，而且减弱了降维的效果，而取值越大，将删除更多的信息，以至于剩下的信息缺乏区分能力。

由上式可以得出，由于在∑矩阵中只取非零的奇异值，所以近似矩阵中的三个矩阵的元素个数为$m\times k+n\times k+k\times k$，只要满足$m\times n\geq m\times k+n\times k+k\times k$的时候，即满足$k\leq\frac{mn}{m+n+1}$，可以去掉次要的信息，保留主要信息，达到降维的目的，降低计算机对存储的要求，从而保证聚类的准确性。

（3）向量语义化

对某一特征项为n的文本向量t进行奇异值分解得到[11]：

$$t=t'\sum U$$

得出t在进行k维映射后得到的向量t'为：

$$t'=tU_k^{T}\sum_k^{-1}$$

进行语义压缩后的向量被认为投影在同一空间里，然后方可进行文本聚类。

7.1.3文本聚类

所谓文本聚类就是将无类别标记的文本信息根据不同的特征，将有着各自特征的文本进行分类，使用相似度计算将具有相同属性或者相似属性的文本聚类在一起。这样就可以通过文本聚类的方法根据留言内容，对不同留言进行分类。通过聚类方法，可以快速挖掘热点问题。

（1）留言文本相似度计算

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体间差异大小，也即对文本信息进行相似度计算，将根据相似特性的信息进行归类。本文先计算基于距离度量的欧几里得距离，再转化为余弦相似度，以此表示计算留言间的差异。

令 $i=(x_1,x_2,\cdots,x_p)$ 和 $j=(y_1,y_2,\cdots,y_p)$ 是两个被 p 个数值属性标记的对象，则对象 i 和 j 之间的欧氏距离为：

$$d(i,j)=\sqrt{\sum_{k=1}^{p}(x_k-y_k)^2}$$

根据余弦相似度和欧氏距离的关系[..]，留言文本间的余弦相似度可表示为：

$$\cos(i,j)=1-\frac{1}{2}d^2(i,j)$$

（2）基于 $K-means$ 聚类的文本聚类

①$K-means$ 聚类原理

$K-means$ 算法是很典型的基于划分的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似性就越大。

$K-means$ 算法的基本思想是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

假设要把样本集分为 k 个类别，算法描述如下：

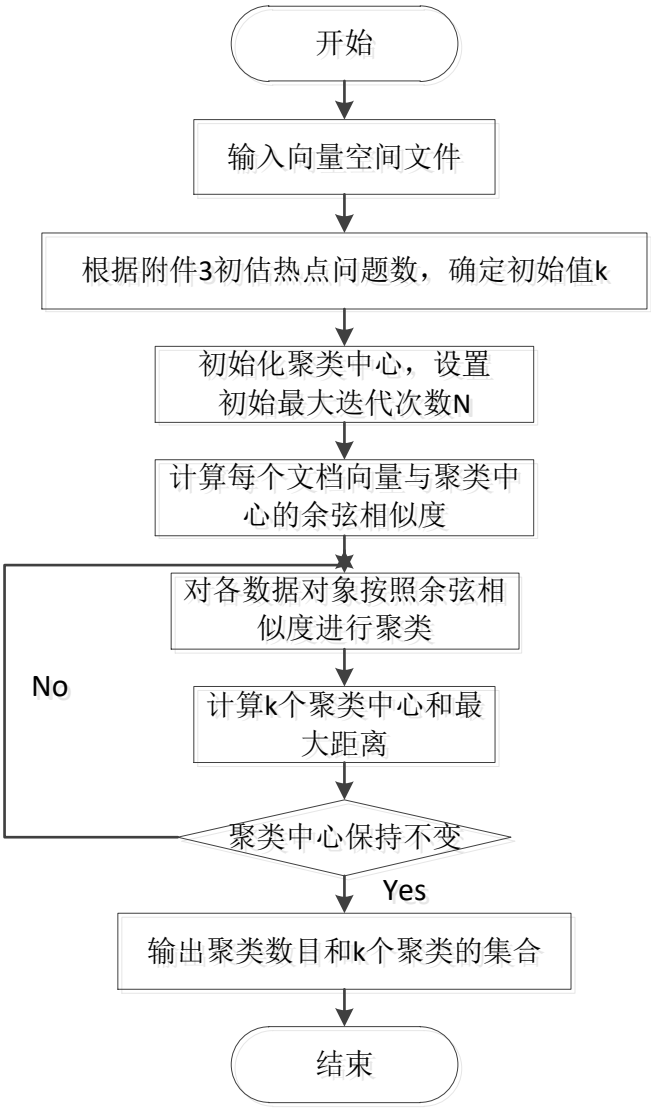
a.适当选择 k 个类的初始中心。

b.在第 k 次迭代中，对任意一个样本，求其到 k 个中心的距离，将该样本归到距离最短的中心所在的类。

c.利用均值等方法更新该类的中心值。

d.对于所有的 k 个聚类中心，如果利用②③的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

②$K-means$ 聚类流程图，如下：



图：*K-means* 聚类流程图

该算法要求在计算之前给定 k 值。本文通过初步估计附件3给出留言的热点问题数，并以此作 k 的值，这里令 $k=7$ 为初值，根据后续的热度值大小，进行适当增减 k 的值，也即热点问题数量调控。

③$K-means$ 聚类结果

按题目表2的格式按热点类别汇编，分别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”，详见附件。

7.1.4矩阵和特征量的计算

主成分分析又称为主分量分析或主轴分析，将多指标转化为少数几个综合指标的一种同计分方法。在实际问题中，研究很多变量的问题是经常遇到的，并且彼此之间有一定的相关性，因而使得所观测到的数据在一定程度上反映的信息有所重叠。而且在变量较多时，在高维空间中研究样本的分布规律比较复杂，势必增加分析问题的复杂性。

考虑到影响热点问题之间额差异性，将每个热点包含留言数、留言时间密集度、点赞数、反对数等作为评价指标。

我们自然希望用较少的综合变量来代替原来较多的变量，而这几个综合变量

又能尽可能多的反映原来变量的信息，并且彼此之间互不相关。

选取以上m_1个指标。

标准化指标变量

$$x_j^* = \frac{x_j - \mu_j}{s_j}, j = 1,2,...,m_1$$

式中

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}; s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}, j = 1,2,...,m_1$$

相关系数矩阵$R = (r_{ij})_{m_1 \times m_1}$

其中

$$r_{ij} = \frac{\sum_{k=1}^{n} x_{ki}^* x_{kj}^*}{n-1}, i, j = 1,2,...,m_1$$

式中：$r_{ii} = 1$，$r_{ij} = r_{ji}$，r_{ij}为第i个评价对象的第j项指标的相关系数。

特征值和特征向量

计算相关系数矩阵R的特征值$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{m_1} \geq 0$，及对应的特征向量$a_1, a_2, ..., a_{m_1}$，其中$a_j = \left[a_{1j}, a_{2j}, ..., a_{m_1 j}\right]^T$，由特征向量组成$m_1$个新的指标变量：

$$F_1 = a_{11}x_1^* + a_{21}x_2^* + ... + a_{m_1 1}x_{m_1}^*$$

$$F_2 = a_{12}x_1^* + a_{22}x_2^* + ... + a_{m_1 2}x_{m_1}^*$$

$$\vdots$$

$$F_{m_1} = a_{1m_1}x_1^* + a_{2m_1}x_2^* + ... + a_{m_1 m_1}x_{m_1}^*$$

上式中共m_1个主成分，$F_j (i = 1,2,...,m_1)$为第j主成分。

7.1.5主成分的选择

为达到降维目的，计算各主成分的信息贡献率和累计贡献率，以其值来选取部分更具代表性的主成分。

主成分F_j的信息贡献率：

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m_1}\lambda_k}, j = 1,2,...,m_1$$

主成分$F_1, F_2, ..., F_p$的累计贡献率：

$$\alpha_p = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m_1}\lambda_k}$$

当α_p接近于1（取$\alpha_p > 0.95$）时，则选择前p个指标变量$F_1, F_2, ..., F_p$作为p个主成分，代替原来m_1个指标变量，从而可对p个主成分进行综合分析。

7.1.6主成分分析的综合评价

（1）主要指标分析

①筛选出 P 个主成分；

②通过标准化指标前的特征向量的数值相对大小，分析各主成分主要反映的对应指标。

（2）综合评价

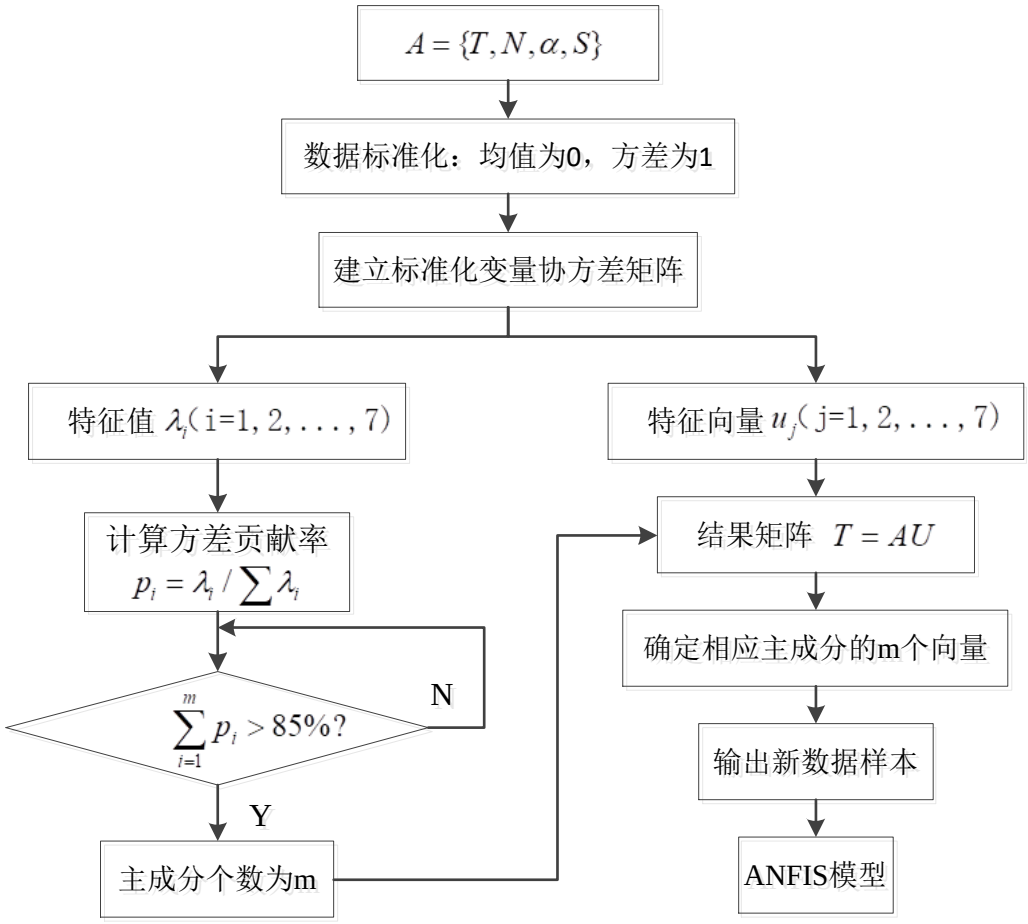
以 P 个主成分的信息贡献率为权重，构建主成分综合评价模型：

$$Z=\sum_{j=1}^{p} b_j F_j$$

根据以上综合得分可进行评价，求出 Z 值，排序。

（3）主成分分析法

基本步骤如图：



图：主成分基本步骤

7.2模型的求解

（1）与问题一采用相同步骤进行文本预处理，对附件3的留言经过文本标记处理、特殊字符处理、中文分词、去停用词等步骤，得到纯文字文本。

（2）调用 *Python* 的库函数，根据语义分析 *LSA* 的奇异值分解 *SVD* 技术和 $K-means$ 算法，实现留言语义空间降维，将相似问题聚类并实现热点挖掘。按题目表2的格式按热点类别汇编，分别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”，详见附件。

（3）经统计，附件3中共有4326条留言，本文利用 *Python* 求得每个热点的留言信息，首先筛选出发布信息量在前10名的热点占总招聘信息数的98.86%，因此其余留言可以忽略不计，进而构造上述指标，利用 *SPSS* 对其进行综合排名。

①对筛选得到的数据导入 *SPSS* 进行标准化处理。

②得到解释的总方差表和成分矩阵。

（4）由上表可知，上述数据进行主成分分析后得到了7个主成分，其方差累计贡献率达到了92%（>85%）符合主成分分析方差提取原则。

（5）用表中成分矩阵除以表中载入合计的特征值的开方得到7个主成分所对应的特征向量矩阵，将得到的特征向量与标准化后的数据相乘，可以得出各个主成分得分值。

对筛选出的4个主成分，将各标准化指标前的特征向量的系数做表，如下表二所示：

表二：各标准化指标前的特征向量

序号	1	2	3	4	5	6	7	8	9	10
特征值	2.60	2.39	1.81	1.10	0.94	0.42	0.34	0.23	0.09	0.08
信息贡献	25.9	23.8	18.1	11.0	9.45	4.20	3.42	2.25	0.90	0.76
累计贡献	25.9	49.8	68.0	79.0	88.4	92.6	96.0	98.3	99.2	100.0

从上4个主成分的系数可看出，第一主成分反映了第3、7、9个指标，第二主成分反映了第2、5、8个指标，第三主成分反映了第1、10个指标，第四主成分反映了第4、6个指标。

（6）以每个主成分所对应的特征值占总特征值的比例作为权重计算主成分综合得分：

$$F=\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}F_1+\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3}F_2+\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}F_3$$

其中 λ_i 表示第 i 主成分因子的特征值。

（7）得到各个热点的综合排名，按题目表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”，详见附件。

八、问题三模型的建立与求解

8.1模型的建立

本题要求从相关性、完整性、可解释性等对留言的答复意见的质量给出评价。对留言和答复预处理后，选取每条答复包含的问候、对点回答、感谢与致歉，答复紧凑度作为评价指标。用极差化法对指标进行归一，再用熵值法确定评价指标的权重系数。

构建一个偏离理想范围的加权距离，采用逼近理想点 *TOPSIS* 法的改进方法，定义并计算质量劣值 Y_i，以其正负对答复意见好坏进行区分，负值的大小可清晰区别质量较差的答复意见质量不好的程度。

8.1.1文本预处理

对附件4的留言进过文本标记处理、特殊字符处理、中文分词、去停用词等

步骤，得到纯文字文本，方法与算法与问题一、二处理一致。

8.1.2评价指标的标准化处理

考虑到附件4中每条留言对应的答复意见的问候、对点回答、感谢与致歉，答复紧凑度等不同，即存在不可公度性。在应用前需要对这些数据做一定的预处理，以便在综合评价中做相应的运算、比较和分析。

由于答复意见的评价指标间差别较大。鉴于此，本模型采用无量纲化处理，选用极差化法，处理如下。

（1）极差化方法

考虑到影响答复意见之间额差异性，将每条答复包含的答复内容中的问候、对点回答、感谢与致歉，答复紧凑度（答复时间-留言时间）作为评价指标。

用 *Excel* 整理出附件4中共 s 条答复意见，每条答复对应 u 个评价指标。

则第 i 条答复的第 j 项指标为 $x_{ij}(i=1,2,...,s;\ j=1,2,...,u)$。

s 条答复的第 j 项指标的最大值：

$$M_j=\max\left\{x_{1j},x_{2j},...,x_{sj}\right\}$$

s 条答复的第 j 项指标的最小值：

$$m_j=\min\left\{x_{1j},x_{2j},...,x_{sj}\right\}$$

所以新的标准化指标：

$$x_{ij}^{*}=\frac{x_{ij}-m_j}{M_j-m_j}\in[0,1],(i=1,2,...,s;\ j=1,2,...,u)$$

上式中 x_{ij}^{*} 为已消除单位、数量级从差异影响的各对象各项指标值，作为各项指标值分析的初值，进行后续评价。

8.1.3评价指标权重系数的确定

考虑到各项指标对于答复的重要性和其对于评价对象取值的差异性没有必然联系，所以要同时根据各项指标所提供信息大小和差异程度来决定相应的权重系数，即基于“指标差异”赋权方法，此处选用熵值法，如下：

（1）熵值法

第 j 项指标的熵值：

$$I_j=-k\sum_{i=1}^{s}p_{ij}\ln(p_{ij})(j=1,2,...,u)$$

其中常数

$$k=\frac{1}{\ln(s)}$$

第 i 条答复的第 j 项指标的特征比重：

$$p_{ij}=\frac{x_{ij}^{*}}{\sum_{i=1}^{s}x_{ij}^{*}}(i=1,2,...,s;\ j=1,2,...,u)$$

得到第 j 项指标的差异系数：

$$r_j = 1 - I_j (j = 1,2,...,u)$$

差异系数是反映综合评价指标作用大小的一个量，其值越大，指标的作用就越大，反之亦然。

最后，第 j 项指标的权重系数：

$$w_j = \frac{r_j}{\sum_{k=1}^{u} r_k}, j = 1,2,...,u.$$

上式中权重系数 w_j 刻画了评价指标之间的相对重要性的大小，是构建综合评价模型的一个重要参数，确保总评价结果的合理性和可行度。

8.1.4综合评价数学模型的建立

搜集答复意见中各项指标正常的范围，以此作为对被评价对象的评价指标设定的一个理想范围 $(l_j, u_j), j = 1,2,...,m$，然后对每一个评价对象的评价指标值与理想范围进行比较。考虑到参考值的误差，将各项指标参考范围上下限扩大5%，$L_j = 0.95 l_j, U_j = 1.05 u_j$ 即如果各项评价指标 $x_{ij}^* \in (L_j, U_j)$，则答复意见为质量好；若 $x_{ij}^* \notin (L_j, U_j)$，则答复意见为质量不好，类比逼近理想点 $(TOPSIS)$ 法，进一步构建一个偏离理想范围的加权距离，对质量较差的答复意见的质量进行区别。

（1）逼近理想点 $(TOPSIS)$ 法

对被评价对象的评价指标假定一个理想点 $(x_1^*, x_2^*, ..., x_m^*)$，对每一个评价对象的评价指标值 $(x_{i1}, x_{i2}, ..., x_{im})$，定义二者之间的加权距离：

$$y_i = \sum_{j=1}^{m} w_j (x_{ij} - x_j^*)^2, i = 1,2,...,n_2$$

上式作为综合评价函数，即反映第 i 条答复的指标值与理想点的差异程度，按照这种方法考察各评价对象的指标值与理想点接近的程度对答复意见的态度好坏程度依次排序。

（2）逼近理想点 $(TOPSIS)$ 法的改进方法

类比上模型的理想点 $(x_1^*, x_2^*, ..., x_m^*)$，不妨将答复意见中各项指标正常的范围，以此作为对被答复意见的评价指标设定的一个理想范围 $(L_j, U_j), j = 1,2,...,m$，类比逼近理想点 $(TOPSIS)$ 法，进一步构建一个偏离理想范围的加权距离，对质量较差的答复意见的质量进行区别。将其定义为质量劣值 Y_i。

首先定义第 i 条答复意见的第 j 项指标的分质量劣值 y_{ij}：

$$y_{ij} = \begin{cases} w_j (L_j - x_{ij})^2, i = 1,2,...,n_2, x_{ij} < L_j \\ 0, i = 1,2,...,n_2, L_j \le x_{ij} \le U_j \\ w_j (x_{ij} - U_j)^2, i = 1,2,...,n_2, x_{ij} > U_j \end{cases}$$

上式中 y_{ij} 是构成质量劣值 Y_i 的各指标分量，体现了每一项指标对质量劣值 Y_i 的分贡献值，以便求和。

由上述，对第i条答复意见各指标的y_{ij}求和，得质量劣值Y_i：

$$Y_i=\sum_{j=1}^{m}y_{ij},i=1,2,...,n_2$$

上式作为改进后的综合评价函数，即反映第i条答复意见的指标值与理想范围的偏差程度，按照质量劣值Y_i考察各评价对象的指标值与理想范围接近的程度对对答复意见的质量好坏程度依次排序。

当某个$Y_i=0$时，即所有指标达理想范围，为质量好。当$Y_i\neq0$时，为质量不好，Y_i值越大，质量不好的程度越大。此处可清晰区别质量较差答复意见的质量不好的程度。

8.2模型的求解

（1）与问题一、二采用相同步骤进行文本预处理，对附件4的留言和答复意见经过文本标记处理、特殊字符处理、中文分词、去停用词等步骤，得到纯文字文本。

（2）根据建立的综合评价模型，再利用*Matlab*软件求解函数值问题，计算各留言对应答复意见的质量劣值Y_i，根据正负及数值对各留言的答复意见的好坏进行区分，负值的大小可清晰区别质量较差的答复意见质量不好的程度。

（3）将以上求解得到的附件4的所有留言的答复意见的质量劣值Y_i从上到下，从大到小依次排列。

九、模型的评价

优点	缺点
➢ 模型中对于相关留言信息进行了充分挖掘，使得所结合多种的模型对于留言分类的分类、归类效果较好，适用于大量文本数据的情况。 ➢ 利用主成分分析法可消除评价指标之间的相关影响，减少指标选择的工作量，且便于实现； ➢ K-means算法原理比较简单，可解释极强，容易实现且算法运行结果具有可靠性和有效性； ➢ TOPSIS 法简单实用，适用性比较强。	➢ 基于留言文本进行降维方式匹配筛选的综合模型算法复杂，时间复杂度高，最后的精度很低。 ➢ 模型中仅针对中文字符特点进行了算法设计，缺乏对其他语言以及表情图像等一般情况的考虑。 ➢ K-means算法局部最优，对异常点很敏感，对非凸数据集或类别规模差异太大的数据效果不好且限定数据种类。

十一、参考文献

[1]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.

[2]胡学钢, 董学春, 谢飞. 基于词向量空间模型的中文文本分类方法[J]. 合肥工业大学学报:自然科学版, 2007, 30(10):1261-1264.

[3]徐明, 高翔, 许志刚,等. 基于改进卡方统计的微博特征提取方法[J]. 计算机工程与应用, 2014, 50(19):113-117.

[4]郑翠翠. 面向领域文本的潜在语义分析研究[D]. 南京理工大学, 2010.

[5]张振亚, 王进, 程红梅,等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, 32(9):160-163.

[6]张跃, 李葆青, 胡玲,等. 基于 K-Mean 文本聚类的研究[J]. 中国教育技术装备,2014(18):50-52.