# 第八届“泰迪杯”
# 全国数学挖掘挑战赛

作品名称：“智慧政务”中的文本挖掘应用

# “智慧政务”中的文本挖掘应用

## 摘要

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。

对于问题 1，首先将多分类问题转换为多个二分类问题，利用组合的知识，7 个标签两两之间任意组合，生成 21 个分类器，通过分类器之间的投票，选出最适合数据的一级标签；利用 EXCEL 的“自动筛选”功能中的“筛选重复项”功能，得到不重复的留言。其次利用 python 进行数据预处理，导入 pandas 包对数据进行清洗、导入 jieba 包利用 jieba 中文分词工具对留言详情进行分词、根据停用词表去停用词；再次对处理好的数据进行抽取（80%用来建立模型，20%用来测试模型），计算每条留言的 TF-IDF 权值得到每条留言的权值向量生成 495*10000+的向量空间。为了提高运行速度和效率利用主成分分析进行 PCA 降维，其次利用高斯贝叶斯公式建立模型，根据上述方法建立模型的精度为 80%，最后再根据题目中给的调精公式，对模型进行优化。

对于问题 2，首先通过 EXCEL 中的替换功能将留言内容中的“*”替换成“x”，其次利用 python 的 pandas 包、jiaba 包进行数据预处理（数据清洗、分词、去停用词）。再次根据 python 的 TfidfTransformer 函数计算每个词语的 TF-IDF 权值建立向量空间，并调用 cosine 函数利用数据之间的夹角余弦公式计算每个数据与其他数据之间的相似度，利用 argsort 函数得到每个数据的索印进行排序，选出与每个数据相似度最高的前五个数据，再次用 EXCEL 进行操作，根据索引将

与每个数据相似度最高的前五个数据的详细内容移动到每个单独的 EXCEL 表中。最后利用 re 模块的正则表达式中的”findall”函数，根据前置词（如：在、位于等）以及后置词（如：区、小区等）找出特定的时间或者人群。最后将热点问题首先根据时间范围、其次根据点赞数与反对数给出合理的热度评价指标。

对于问题 3，将留言详情和回复内容进行数据预处理，其次进行相似度计算，根据相似度的匹配结果来判断。

关键词：TF-IDF　PCA　余弦度量　相似度　去重　中文分词 正则表达式

# Text mining application in "smart government"

**abstract**

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that mainly rely on manual work to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

For problem 1, firstly, the multi classification problem is transformed into a multi two classification problem, and 21 classifiers are generated by combining 7 labels randomly. The most suitable primary labels are selected by voting among classifiers. The non repeated messages are obtained by using the "filter duplicates" function in the "automatic selection" function of excel. Secondly, we use Python to preprocess the data, import pandas package to clean the data, import Jieba package to use Chinese word segmentation tool of Jieba to segment the details of the message, and then use the stoppage words according to the stoppage word list; thirdly, we extract the processed data (80% to build the model, 20% to test the model), calculate the TF-IDF weight of each message to get the weight vector of each message It's a vector space of 495 * 10000 +. In order to improve the running speed and efficiency, PCA is used to reduce the dimension of PCA. Secondly, Gaussian Bayes formula is used to build the model. The precision of the model is 80%. Finally, the model is optimized according to the precision formula given in the topic.

For problem 2, firstly, the "*" in the message content is replaced by "X" through the replacement function in Excel, and then the data preprocessing (data cleaning, word segmentation, de stop words) is carried out by using pandas package and Jiaba package of Python. Again, according to Python's tfiftransformer function, calculate the TF-IDF weight of each word to establish a vector space, and call cosine function to calculate the similarity between each data and other data by using the included angle cosine formula between data, and use argsort function to get the index print of each data for sorting, select the first five data with the highest similarity with each data, and then use Excel to do the same again According to the index, the details of the first five data with the highest similarity to each data are moved to each separate Excel table. Finally, the "findall" function in the regular expression of re module is used to find out the specific time or person according to the preposition (such as in, in, etc.) and the postposition (such as area, cell, etc.). At last, according to the time range, and then according to the number of likes and dislikes, a reasonable heat evaluation index is given.

For question 3, we preprocess the message details and reply contents, then calculate the similarity, and judge according to the matching results of similarity.

Key words: TF IDF PCA cosine measurement similarity de duplication of Chinese word segmentation

# 1、挖掘目标

本次建模目标是利用收集的群众留言信息数据，；利用 jieba 中文分词工具对留言内容详情进行分词，达到以下三个目标：

1）利用文本分词和文本聚类的方法对非结构化的数据进行文本挖掘；建立分类模型。

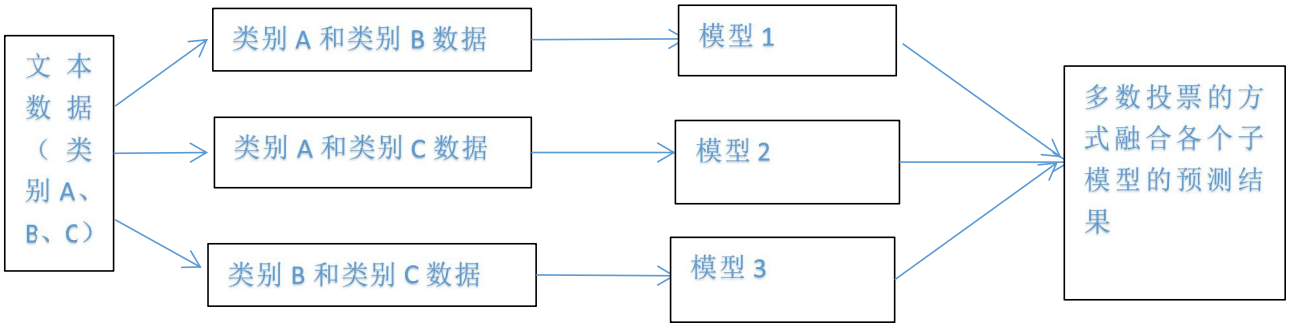
2）根据热点问题的定义，利用余弦相似度从众多留言中筛选出热点问题，并利用正则表达式找出特定的地点或者人群，统计赞成数和反对数给出热度评价指标。

3）根据留言详情和答复内容，从答复的相关性、完整性、可解释性、及时性对答复意见的质量给出一套可行的评价方案。

# 2.1 问题 1 分析方法与过程

### 2.1.1 多分类问题转换成多个二分类问题

问题一是多分类问题，根据附件二的一级标签，通过 EXCEL 的自动筛选功能，可以得出一共有 7 个一级标签，这里针对多分类问题，我们转换成多个二分类问题。利用排列组合的知识，训练 21 个二分类模型，也就是说训练阶段无序从 7 个类别中抽取 2 个类别训练一个分类器，预测阶段，输入一个样本实例，统计 21 个分类器的预测类别，结果中出现次数最多的类别，就是最终的预测结果。




### 2.1.2 数据预处理

#### 2.1.2.1 留言内容的去重和数据抽取

在附件二所给的留言内容中，出现了重复的数据，利用 EXCEL 中的查重功能，发现有两条一模一样的数据，选择其中的一条保留下来。因为要对最后建立的模型进行测试，故对数据进行抽取，一部分用来实验，一部分进行测试。对于每个一级标签的数据，选取 85%作为实验数据，15%作为测试数据。

#### 2.1.2.2 对留言内容进行中文分词操作

在对文本数据进行分析之前，要先把非结构话数据转换成结构化数据，之后进行一系列的分析操作。题目中所给的数据是文本数据，为了便于转换成结构化数据，先要对去重后的留言内容进行中文分词操作，这里选择的 python 的中文分词包 jieba 进行分词。结巴分词其实通过词典分词，然后对不在词典的词使用 HMM 算法识别新词，使得能更好的实现中文分词效果。

#### 2.1.2.3 去停用词操作

分完词后的文本数据中还有很多表达无意义以及对寻找关键词无用的词语，这些词称为“停用词”为了提高搜索效率和节省空间，还要进行去停用词操作，经过一系列操作过滤掉一些无用的词语。停用词又两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本表达无意义。

在根据词性选择停用词时，选择出现频率高并对文本表达无意义的动词、副词、语气词、非文本词语、单词等（选择以名词为特征）。

#### 2.1.2.4 建立向量空间

因为计算机不能够直接处理文本数据，因此需要把文本数据转换成非文本数据，也就是文本数字化。这里选择的根据 TF-IDF 建立向量空间。

TF-IDF 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。

1、TF（词频）

TF 表示词条在文档 d 中出现的频率

2、IDF（逆文档频率）

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，再将得到的商取以 10 为底的对数得到。

因此，每一个词语都有一个对应的 TF-IDF 权值，故可以根据 TF-IDF 建立一个权值向量空间。由此，非结构化数据数据转换成了结构化数据。

#### 2.1.2.5 降维（主成分分析）

主成分分析由卡尔·皮尔逊于 1901 年发明，用于分析数据及建立数理模型。其方法主要是通过对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）与它们的权值（即特征值）。PCA 能从冗余特征中提取主要成分，在不太损失模型质量的情况下，提升了模型训练速度 PCA(Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。PCA 的主要思想是将 n 维特征映射到 k 维上，这 k 维是全新的正交特征也被称为主成分，是在原有 n 维特征的基础上重新构造出来的 k 维特征。PCA 的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第 1,2 个轴正交的平面中方差最大的。依次类推，可以得到 n 个这样的坐标轴。通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面 k 个坐标轴中，后面的坐标轴所含的方差几乎为 0。于是，我们可以忽略余下的坐标轴，只保留前面 k 个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为 0 的特征维度，实现对数据特征的降维处理。

PCA 的算法步骤：

设有 m 条 n 维数据。

1）将原始数据按列组成 n 行 m 列矩阵 X

2）将 X 的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值

3）求出协方差矩阵

4）求出协方差矩阵的特征值及对应的特征向量

5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P

6）Y=PX 即为降维到 k 维后的数据

#### 2.1.2.5 模型建立

这里利用高斯贝叶斯生成分类器建立模型。

## 2.2 问题 2 分析方法与流程

### 2.2.1 数据预处理

在对留言详情进行挖掘分析之前，要先把非结构化的文本信息转化为计算机能够识别的结构化信息。对于附件的留言详情，以中文文本的方式给出了数据。为了便于转换，先要对其进行中文分词。利用 python 的中文分词包 jieba 进行分词，将一段话变成数个词语，为之后转换成结构化数据以级计算相似度做基础。

### 2.2.2 生成向量空间

进行完分词后，为了计算相似度，需要将非结构化数据转换为结构化数据，将文本数字化，使两两之前能够进行计算。这里采用的是利用 TF-IDF 权值生成向量空间。利用词频以级逆文本频率，计算每个词的权重，生成矩阵，其中行数是数据的个数，列数是所有词的个数。生成 TF-IDF 的具体步骤如下：

（1）使用 TF-IDF 算法，找出描述每个留言的前 15 个关键词

（2）对每个留言内容提取的前 15 个关键词，合并成一个集合，计算每个留言内容对于这个集合中词的词频，如果没有记为 0

（3）生成每个留言的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频*逆文档频率

至此，利用 python 将文本数据转换成了结构化数据，以便于进行下一步的计算。

### 2.2.3 相似度计算

为了找出相似的留言需要根据距离进行相似度计算。传统的推荐算法（如协同过滤、基于物品的推荐等）采用的相似度计算公式主要有：余弦夹角、欧氏距离、杰卡德系数和皮尔森相关系数等，这里采用了余弦夹角进行相似度的计算。

余弦夹角公式对向量进行了归一化处理，解决了向量个体间存在度量标准不统一问题产生的计算偏差；余弦夹角的值域区间为[-1,1]，相对于欧式距离的值域范围[0,正无穷大],能够很好的对向量间的相似度值进行了量化。 余弦相似度衡量的是维度间取值方向的一致性，注重维度之间的差异，不注重数值上的差异。是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的度量。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"。

利用余弦相似度，根据权值矩阵计算矩阵中行与行之间的相似度，找出与每一行相似度最高的前五个数据，导出到 EXCEL 表中，再进行一系列操作。

### 2.2.4 利用正则表达式找出相同的留言

正则表达式(regular expression)描述了一种字符串匹配的模式（pattern），可以用来检查一个串是否含有某种子串、将匹配的子串替换或者从某个串中取出符合某个条件的子串等。

因为特定的地点或者人群表达的方式多种多样，利用正则表达式中的“findall”函数找出特定的地点或者人群，将相似的留言归结成相同的留言。xxx 小区，或者在 xxx 小区，地点有前置词（在、位于等一些方位词或介词）和后置词（省、市、路、街道、小区等），两个词中间加上地点，利用正则表达式将其筛选出来。

在特定的地点或者人群的范围内的相似留言归结成了相同的留言，由此筛选出了热点问题。

### 2.2.5 定义热度指标

热点问题的指标首先根据热点问题的时间范围来确定，时间范围越短热点问题的排名越高，如若时间差不多，再根据反对数以及点赞数，将反对数以及点赞数相加，比较相加后的和，在相同的时间范围内，它们之间的和越大，热点问题的排名越高。得出排名后，再利用 EXCEL，通过人工操作，将具体的留言编号、留言用户、留言主题、留言时间、留言详情、反对数、点赞数呈现出来。

## 参考文献

[1] 泰迪杯《第四届 C 题 2_网络招聘信息的数据挖掘与综合分析》
[2] 吴军《数学之美》
[3] 李英《基于词性选择的文本预处理方法研究》
[4] 王丹、樊兴华《面向短文本的命名实体识别》