# “智慧政务”中的文本挖掘应用

## 一、前言

**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。

**关键词** TF/IDF 算法 DBSCAN 密度聚类 支持向量机（SVM）热点问题分析

**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 used to rely mainly on human 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.

**KEY WORDS** TF/IDF DBSCAN SVM news hotspots analysis

## 目录

# 二、预处理

## 2.1 分词

在汉语中，词以字为基本单位的，但是一篇文章的语义表达却仍然是以词来划分的。因此，在处理中文文本时，需要进行分词处理，将句子转化为词的表示，这个过程就是中文分词。我们选用中文分词工具——jieba

Jieba 分词结合了基于规则和基于统计这两类方法。首先基于前缀词典进行词图扫描，前缀词典是指词典中的词按照前缀包含的顺序排列，从而形成一种层级包含结构。如果将词看作节点，词和词之间的分词符看作边，那么一种分词方案则对应着从第一个字到最后一个字的一条分词路径。因此，基于前缀词典可以快速构建包含全部可能分词结果的有向无环图，这个图中包含多条分词路径，有向是指全部的路径都始于第一个字、止于最后一个字，无环是指节点之间不构成环。基于标注语料，使用动态规划的方法可以找出最大概率路径，并将其作为最终的分词结果。

本文关键词：**手机号 通话 短信 县政府**

## 2.2 去停用词

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。人类语言包含很多功能词。与其他词相比，功能词没有什么实际含义。最普遍的功能词是限定词（“the”、“a”、“an”、“that”、和“those”），这

些词帮助在文本中描述名词和表达概念，如地点或数量。介词如："over"，"under"，"above" 等表示两个词的相对位置。对于停用词一般在预处理阶段就将其删除，避免对文本，特别是短文本，造成负面影响。

本文停用词：**请 请求 请问 希望 是不是**

## 三、解决问题

### 3.1 群众留言分类

为了提高工作效率，降低差错率，需要对网络问政平台的群众留言进行体系划分。在这里用到的是分类的思想，并用 F-Score 对分类方法进行评价。

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

#### 3.1.1 TF/IDF 算法

为了降低分类时的压力，这里首先提取关键字作为分类的依据。关键词是代表文章重要内容的一组词。对文本聚类、分类、自动摘要等起重要的作用。在这里我们采用的是 TF/IDF 算法。

TF-IDF 算法由两部分组成：TF 算法以及 IDF 算法。TF 算法是统计一个词在一篇文档中出现的频次，其基本思想是，一个词在文档中出现的次数越多，则其对文档的表达能力也就越强。而 IDF 算法则是统计一个词在文档集的多少个文档中出现，其基本的思想是，如果一个词在越少的文档中出现，则其对文档的区分能力也就越强。

在本次解决方案中，TF 的计算常用式如下

$$\mathrm{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

其中 nij 表示词 i 在文档 j 中的出现频次，但是仅用频次来表示，长文本中的词出现频次高的概率会更大，这一点会影响到不同文档之间关键词权值的比较。所以在计算的过程中一般会对词频进行归一化。分母部分就是统计文档中每个词出现次数的总和，也就是文档的总词数。

IDF 的计算常用式如下

$$\mathrm{idf}_i = \log\left(\frac{|D|}{1+|D_i|}\right)$$

|D|为文档集中总文档数，|Di|为文档集中出现词 i 的文档数量。分母加 1 是采用了拉普拉斯平滑，避免有部分新的词没有在语料库中出现过而导致分母为零的情况出现，增强算法的健壮性。

TF-IDF 算法就是 TF 算法与 IDF 算法的综合使用，本次计算所用计算式如下

$$\mathrm{tf}\times\mathrm{idf}(\mathrm{i},\mathrm{j}) = \mathrm{tf}_{ij}\times\mathrm{idf}_i = \frac{n_{ij}}{\sum_k n_{kj}}\times\log\left(\frac{|D|}{1+|D_i|}\right)$$

根据 TF/IDF 所得的值由大到小排序取前 n 个作为关键字。

### 3.1.2 朴素贝叶斯

朴素贝叶斯方法是基于贝叶斯定理与特征条件独立假设的分类方法，对于给定的训练集合，首先基于特征条件独立（所以叫朴素版的贝叶斯）学习输入、输出的联合概率分布；然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。

设输入空间 X⊆Rn 为 n 维向量的集合，输出空间为类标记集合 Y={c1，c2，…，ck}.输入为特征向量，输出为类的标记，训练集合为：

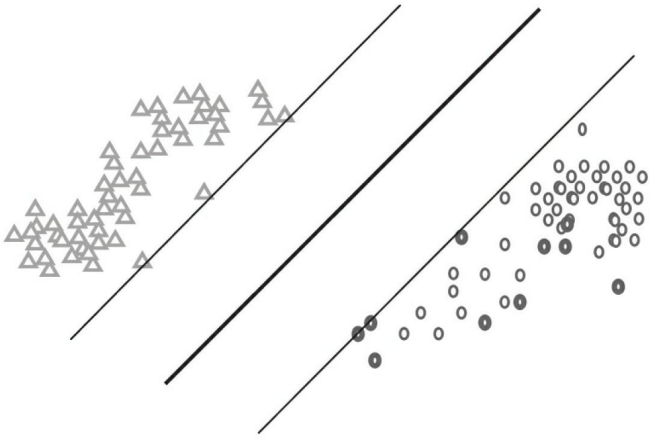
$$T=\{(x1，y1)，(x2，y2)，\ldots(xN，yN)\}$$

假设 P(X，Y)独立分布。可以表示为：

$$y=\arg\max_{c_k} P\left(Y=c_k\right)\Pi_j P\left(X^{(j)}=x^{(j)}\mid Y=c_k\right)$$

### 3.1.3 支持向量机（SVM）

通俗地说，支持向量机（SVM）的最终目的是在特征空间中寻找到一个尽可能将两个数据集合分开的超级平面（hyper-plane）。我们希望这个超级平面能够尽可能大地将两类数据分开，如下图所示。



推出的优化方程如下所示：

$$\min_{w,b}\frac{1}{2}w^{\mathrm{T}}w$$

$$s.t.y_i\left(w^{\mathrm{T}}.x_i+b\right)\geqslant 1,i=1,2\ldots n$$

通过拉格朗日法以及求导之后，原有的方程可以转化为如下的对偶问题：

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \left(x_i \cdot x_j\right)\alpha_i \alpha_j - \sum_{j=1}^{n}\alpha_j$$

$$s.t.\sum_{i=1}^{n} y_i a_i = 0, 0 \leqslant a_i \leqslant C, i = 1,\ldots,n$$

其中αi∈Rm 为拉格朗日乘子，而 C 为惩罚因子。

### 3.2 热点问题挖掘

#### 3.2.1 基于 DBSCAN 聚类算法的问题聚类

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。对于热点问题的及时发现有助于有关部门进行针对性的处理，提升服务效率。

所以在使用 TF/IDF 进行新闻内容文本特征提取之后，使用了基于密度的 DBSCAN 算法进行新闻聚类是最为准确的做法，这种算法的特点可以发现任意形状的簇，同时可以过滤离群点，而不必把离群点分在某一簇种，增加聚类的偏差。

DBSCAN 聚类过程如下所示：

```
(1) 标记所有对象为unvisited;
(2) do
(3)     随机选择一个unvisited对象p;
(4)     标记p为visited;
(5)     if p的ε-邻域至少有MinPts个对象
(6)         创建一个新簇C，并把p添加到C;
(7)         令N为p的ε-邻域中的对象的集合;
(8)         for N中每个点p1
(9)             if p1是unvisited
(10)                标记p1为visited;
(11)                if p1的ε-邻域至少有MinPts个点，把这些点添加到N;
(12)            if p1还不是任何簇的成员，把p1添加到C;
(13)        end for
(14)        输出C;
(15)    else 标记p为噪声;
(16) until 没有标记为unvisited的对象;
```

所以聚类通过获取密度核心，由核心往密度较高的地方延展，将相近的问题合并为同一个簇。在聚类过程需要输入三个参数：

1）D：一个包含 n 个对象的数据集

2）ε:半径参数

3）MinPts：邻域密度阈值

在计算过程中，使用“余弦相似度”来计算距离，所以本文在设置ε参数时，一般都设置为 0.3-0.5 之间，因为超过 0.5 的半径值会使不属于同类新闻聚在一起，小于 0.3 则无法识别相同事件的新闻。设置 MinPts 参数时，可以人为的假定一个阈值，假设有 5 条新闻报道同一事件就认为它就是一个热点，那么可以设置 MinPts 值为 5。

### 3.2.2 热点问题聚类结果

通过热点问题 DBSCAN 聚类的结果如下：

| 留言主题 | 留言详情 | 留言主题1 | 留言详情1 | labels_1 | rank |
|---|---|---|---|---|---|
| A3 区业振城 组团 开发商 欺诈… | 闵 信 A市 A3 业 振 城 组团 业主 现在 临近 近交 交房 发 现 重大 设计 施工 进过 开发 开发商 五轮 谈判 毫无 进展 针对 关键 要点 … | A3区业振城三期a组团开发商欺诈… | … | 13 | 14 |
| A3 区峰景 项目正常 工期 施工… | 荣 盛 A3 峰 景 小区 业主 小区 现在 小区 临近 近交 交房 之日 毫无 质量 可言 赶工 赶工期 工期 业主 沟通 无果 情况 求助 政府 帮… | A3区峰景项目不按正常工期施工… | … | 13 | 14 |
| A市 晟 福兴 汝金城 二期 涉嫌… | 胡书记 A市 晟 福兴 汝 金城 小区 二期 购房 购房者 小区 日安 安排 业主 收 房 业主 收 房 过程 中发 发现 建造 效果 当初 宣传 设计… | A市晟福兴汝金城二期涉嫌虚假宣… | … | 13 | 14 |
| 投诉 A市 晟 福兴 汝 金城 二… | A市 晟 福兴 汝 金城 小区 二期 购房 购房者 小区 日安 安排 业主 收 房 业主 收 房 过程 中发 发现 建造 效果 当初 宣传 设计 存在 差… | 投诉A市晟福兴汝金城二期涉嫌虚… | … | 13 | 14 |
| A市 转业 士官异地 安置 | 转业 士官 异地 安置 A市 需要 条件 谢谢 | 咨询A市转业士官异地安置问题 | … | 12 | 12 |
| A市 士官 转业政策 | 本人 妻子 结婚 婚前 已 A市 工作 多年 社保 缴纳 超过 购房 落户 A市 落户 办理 相关 结婚 手续 这种 情况 符合 士官 转业 相关 政策 … | 咨询A市士官转业政策 | … | 12 | 12 |
| A市 军人 转业异地 安置 | 今年 转业 媳妇 A市 买 房 今年 九月 才 户口 迁到 A市 今年 今年年底 年底 转业 回来 异地 安置 A市 A市 买 房 | 咨询A市军人三期转业异地安置的… | … | 12 | 12 |
| 退伍兵 转业 异地 安置 入户 … | 今年 转业 回 西地 军人 解下 媳妇 A市 A5 买 房 今年 户口 迁入 A市 结 婚 已 满 现在 面临 期 转业 问下 这种 情况 媳妇 户籍 异地 … | 咨询退伍兵转业异地安置入户A… | … | 12 | 12 |
| A市 易地 转业相关 政策 | 下关 A市 市易 易地 转业 转业费 政策 第一 西地 C4 媳妇 西地 I市 领 结婚 结婚证 媳妇 购房 落户 A市 入伍 满足 转业 条件 … | 咨询A市易地转业相关政策等问题 | … | 12 | 12 |
| 新天地 豹悍 武装 押运 有限公… | 黄 异 智 书 读 少 岁 参军 参军入伍 入伍 入党 嘉奖 优秀 士兵 岁 退伍 进 西地 新天地 天地 豹 悍 武装 押运 有限 有限公司 公司 至… | 西地省新天地豹悍武装押运有限… | … | 11 | 11 |
| A2 区富惠 天下公司 涉嫌 非法… | 西地 富 惠 天下 商务 商务信息 信息 服务 有限 有限公司 公司 A2 注册 成立 富 惠 天下 互联 互联网 联网 平台 发布 项目 社会 吸引 … | A2区富惠天下公司涉嫌非法集资… | … | 11 | 11 |
| 聚利人 普惠 投资 有限公司 涉… | 胡书记 百姓 傲主 帮助 千名 百姓 亿 血汗 血汗钱 A市 p2p 平台 聚 利 网 骗 截止 月份 平台 彻底 停止 兑付 疑似 主要 控制… | 西地省聚利人普惠投资有限公司… | … | 11 | 11 |
| 西地 省利 聚人普惠 投资 有限… | 百姓 傲主 帮助 千名 百姓 亿 血汗 血汗钱 A市 p2p 平台 利 聚 网 骗 截止 月份 平台 彻底 停止 兑付 疑似 主要 控制 帅 跑 路 现 平… | 西地省利聚人普惠投资有限公司… | … | 11 | 11 |
| 西地 省津楚 投资 有限公司 法… | A市 公安 公安局 西地 津 楚 投资 责任 任有 有限 有限公司 公司 实际 控制 董事 董事长 兼 法人 范 可风 至今 处于 无法 联系 状态 法… | 西地省津楚投资有限公司法人失… | … | 11 | 11 |
| 投诉 A市 A6 区费 之步 大地 … | 相关 部门 负责 负责人 责人 怀着 痛心 痛心疾首 心疾 首 心情 汇报 情况 以下 材料 句句 属实 孩子 A市 A6 丁字湾 街道 湖外 外路 … | 投诉A市A6区责之步大地艺术培训… | … | 10 | 9 |
| A市 科技 工程学校 涉嫌 欺骗… | 年级 学生 家长 马上 孩子 中考 考试 A市 科技 技工 工程 工程学 学校 学校 职业 学校 学校 招生 班主 班主任 强烈 强烈推荐 推荐 成绩 般的… | A市科技工程学校涉嫌欺骗学生 | … | 10 | 9 |
| 楚江 工贸 技工学校 涉嫌 招生… | 孩子 今年 A市 西地 楚江 工贸 技工 技工学校 工学 学校 学 校 录取 班主 班主任 推荐 学校 A市 实地 学校 民办 学校 现在 为止 学校 … | 西地省楚江工贸技工学校涉嫌招… | … | 10 | 9 |
| A市 卫生 职业学院 没搬 A8 … | 单 招来 学校 学生 现在 当时 学校 开学 学校 搬迁 迁至 A8 新 校区 很多 学校 被骗 现在 都还没 搬到 新 校区 学校 招生 招生简章 简… | A市卫生职业学院为何还没搬至A8… | … | 10 | 9 |
| 北京师范大学 A市 附属 学校 … | 北京 北京师范大学 京师 师范 师范大学 大学 A市 附属 学校 小 年级 新生 家长 现 小孩 入校 前 遇到 一系 系列 求助 应 学校 邀约 参… | 反映北京师范大学A市附属学校… | … | 10 | 9 |

由图可见，程序运行后将聚类产生的中间结果用 csv 文件保存下来，每条新闻都得到了一个标签，标签一样的新闻即为同一热点。

## 3.3 答复意见的评价

### 3.3.1 对用户留言和回复意见提取中文

先将附件 4 读取，并将用户留言列入表 list_data1,将回复意见列入 list_data2，

然后使用正则表达式讲读取的字符串提取中文

### 3.3.2 对用户留言和回复意见进行中文分词

本题通过对附件 4 的分析，附件 4 使用中文文本的方式给出了数据，为了便于转化，先对这些答复意见进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。Jieba 采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有生成词情况所构成的有向无环图（DAG），同时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，使得能更好的实现中文分词效果。

用正则表达式提取中文后，再使用 jieba 进行分词并按（词语，词频）的元组格式建立列表， 同时去除部分无用词。

### 3.3.3 使用 F1-score 进行评价

F1 分数（F1-score）是分类问题的一个衡量指标。一些多分类问题的机器学习竞赛，常常将 F1-score 作为最终测评的方法。它是精确率和召回率的调和平均数，最大为 1，最小为 0。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

此外还有 F2 分数和 F0.5 分数。F1 分数认为召回率和精确率同等重要，F2 分数认为召回率的重要程度是精确率的 2 倍，而 F0.5 分数认为召回率的重要程度是精确率的一半。计算公式为：

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

G 分数是另一种统一精确率和的召回率系统性能评估标准，G 分数被定义为召回率和精确率的几何平均数。

$$G = \sqrt{precision \cdot recall}$$

此时，将建立的用户留言列表按词频进行排序，然后将用户留言列表词频前十的词作为搜索词对留言意见列表进行查找，并将用户列表的词频作为 y_true，将留言意见列表所查找到的词频作为 y_pred，查找不到的以 0 代替，然后进行 F1 函数的评测。

### 3.3.4 使用词云查看部分高频词


## 四、总结

### 4.1 结论

本文进行了群众留言分类、热点问题挖掘、答复意见的评价三个问题的分析，

实现了热点问题的发现，关联事务的分类获取。

主要完成的工作包括以下几个方面：

第一、利用 jieba 分词、TF/IDF 算法、DBSCAN 聚类算法分别对提供的政务数据进行预处理、特征提取以及聚类分析。

第二、在解决第一问之余，利用贝叶斯算法以及 SVM 的应用，将二问热点问题的分析进行聚类，成功提取出了热点问题。

第三、

分析结果解决了各类社情民意相关的文本数据不断攀升而人工处理不及时的问题。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。

**4.2 未来展望**

人工智能技术将使社情民意调查更加高效便捷，不仅能够充分抓取民众最为关心关注的重大自然灾害、重大事故、住房房价、社会问题、领导班子等问题，也能出具统计体系最为关注的正负面统计、趋势分析实时报告；同时对意见领袖分析、对重点事件、突发事件进行预警及跟踪也设定了合理预警规则，做了重点预警报告。通过人工智能技术可以把大量复杂的数据进行整合，帮助政府更科学地制定具体政策。

## 五、参考文献

[1] https://github.com/fxsjy/jieba

[2] 涂铭,刘祥,刘树春.Python 自然语言处理实战核心技术与算法[M].机械工业出版社,2018.

[3] Steven Bird,Ewan Klein,Edward Loper;陈涛 译.Python 自然语言处理[M].东南大学出版社,2010.

[4]叶建成.利用文本挖掘技术进行新闻热点关注问题分析[D].广州大学,2018.