“智慧政务”中的文本挖掘应用

摘要：

智慧政务，核心要义是要实现管理智能化、服务智慧化，打破各个部门的信息孤岛，实现数据共享。利用这些信息能辅助政府机关有更好的决策，、所以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。

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 TF-IDF 权重法来对特征词进行权重计算，形成词袋，构造词汇-文本矩阵，由于这种方法具有高维度，高稀疏度以及同义词影响的缺点，因此， 本文进一步利用基于潜在语义（LSA）分析的奇异值分解算法（SVD）对词汇-文本矩阵进行空间语义降维，语义压缩后的文本向量被认为投影在了同一空间里，再通过 k-means 文本聚类算法群众的留言进行归类。

针对问题二：利用第一题做了的文本分词后，再用 TF-IDF 来计算文本相似度，通过 LSI 自然语言处理模型，通过大量的测试，在问题描述这种找出词汇之间的关系，得出相关词汇构成的一个潜在的主题，给词汇聚类，达到降维的目的，实现对热点问题的归类与排序。

关键词：TFIDF；奇异值分解；朴素贝叶斯分类器；LSI 模型

Application of text Mining in Intelligent Government Affairs.

Abstract: the core of intelligent government affairs is to achieve intelligent management and intelligent service, break the information islands of various departments, and realize data sharing. so the establishment of an intelligent government system based on natural language processing technology has become a new trend in the innovative development of social governance. it plays a great role in improving the management level and administration efficiency of the government.

Aiming at problem 1: this paper first preprocesses the unstructured data in Annex 2, such as re-emptying, Chinese word segmentation and stop word filtering, and then calculates the weight of feature words based on TF-IDF weight method to form a word bag and construct a vocabulary-text matrix. because this method has the shortcomings of high dimension, high sparsity and synonym influence, This paper further uses the singular value decomposition algorithm (SVD) based on latent semantic (LSA) analysis to reduce the spatial semantic dimension of the vocabulary-text matrix, and the semantically compressed text vector is considered to be projected in the same space, and then classified by the messages of the masses of the k-means text clustering algorithm.

Aiming at problem 2: use the text segmentation done in the first question, and then use TF-IDF to calculate the text similarity, through the LSI natural language processing model, establish the word bag model, through a large number of statistics, describe the relationship between words, get a potential topic of related words, cluster words, and achieve the purpose of dimensionality reduction.

Keywords: TFIDF; singular value decomposition;Naive Bayesian classifier; LSI model;

目录

1. 挖掘目标

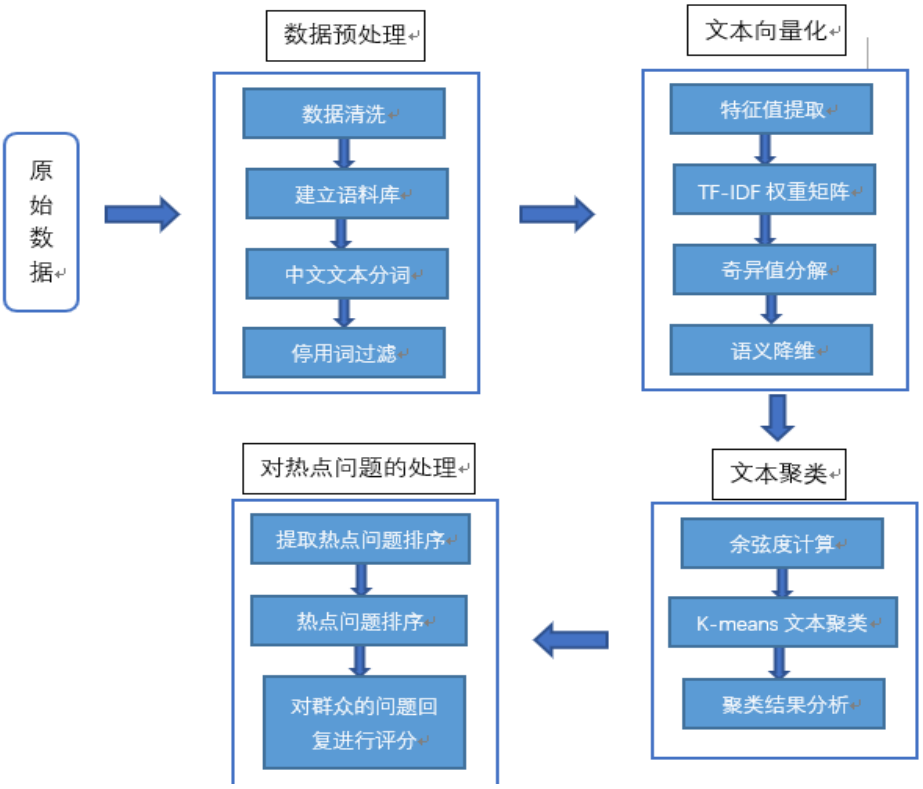
智慧政务，核心要义是要实现管理智能化、服务智慧化，打破各个部门的信息孤岛，实现数据共享。利用这些信息能辅助政府机关有更好的决策，所以建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

本次建模目标是利用文本挖掘的方法，对群众留言问题数据进行基本的预 处理、中文分词、停用词过滤后，一方面根据附件一的留言分类标签进行细分，采用朴素贝叶斯分类器对群众的留言进行分类。对群众留言问题进行分类进行聚类；另一方面对热点问题进行归类、排序，采用 LSI 自然语言处理模型，建立词袋模型，通过大量的统计，在问题描述这种找出词汇之间的关系，得出相关词汇构成的一个潜在的主题，给词汇聚类，达到降维的目的，实现对热点问题的归类与排序。

2. 总体流程与步骤

2.1. 总体流程

本文的总体架构及思路如下：



步骤一：数据预处理，由于数据文本量很大，要对附件进行数据清洗、建立语料库、中文文本分词、停用词过滤，以便后续分析；

步骤二：文本向量化，基于 TFIDF 权重法提取关键词，构造词汇-文本矩阵，进而利用奇异值分解算法进行语义空间降维，去除同义词的影响，简化计算。

步骤三：文本聚类，根据文本向量，计算文档间的欧式距离，再基于 k-means 聚类算法对各个岗位描述进行聚类。

步骤四：对热点问题的处理，基于对热点问题的影响，提取热点问题，在此基础上对热点问题排序，对群众的问题回复进行评分。

3. 分析方法与过程

3.1 数据预处理

3.1.1 数据描述

通过观察所给数据，可以发现数据量比较大，附件 1:三级分类表格，将问题分为三级，附件 2 将问题归纳为附件 1 中的三级分类，附件 3 将针对问题进行点赞或反对，来决定对市民的影响成程度，而附件 4：对于问题已影响到市民生活，如果不做处理会对后续分析造成影响，所以要对留言问题进行回复。如果把这些数据也引入进行分词、词频统计乃至文本聚类等，则必然会对聚类结果的质量造成很大的影响，于是本文首先要对数据进行预处理。

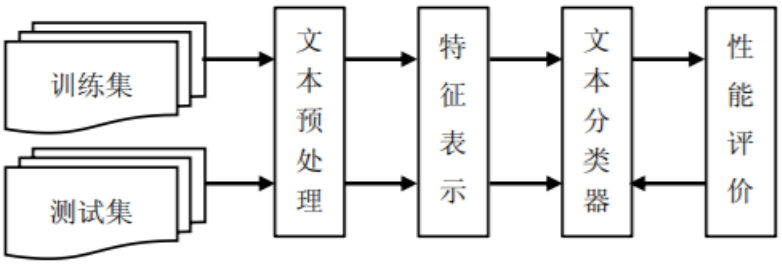


图 3.1.1 文本分类流程图

3.1.2 文本预处理

我们把这些文本数据的预处理分为四个部分：

（一）**数据清洗**

由于附件数据文本量很大，存在大量的无用字符文本，因此去除不需要的字段很有必要。

（二）建立语料库

需要建立一个语料库（corpus），用来模拟语言的使用环境。这里用了两个文本.txt（train_corpus 和 test_corpus）来进行建立语料库，

corpusl_1.txt：从 excel 中读取，去空格制表符换行符后，每单元格内数据成一行储存，用于观察分析句子特征，后期可删，读写会影响速度。

train_corpus.txt：储存 text_corpus.txt 内数据的分词结果，用于观察分词效果与词特征。

（三）中文文本分词

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，对附件中每一个问题描述进行中文分词，jieba 分词用到的算法：

基于 Trie 树结构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 jieba 分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。 部分分词结果示例如图：

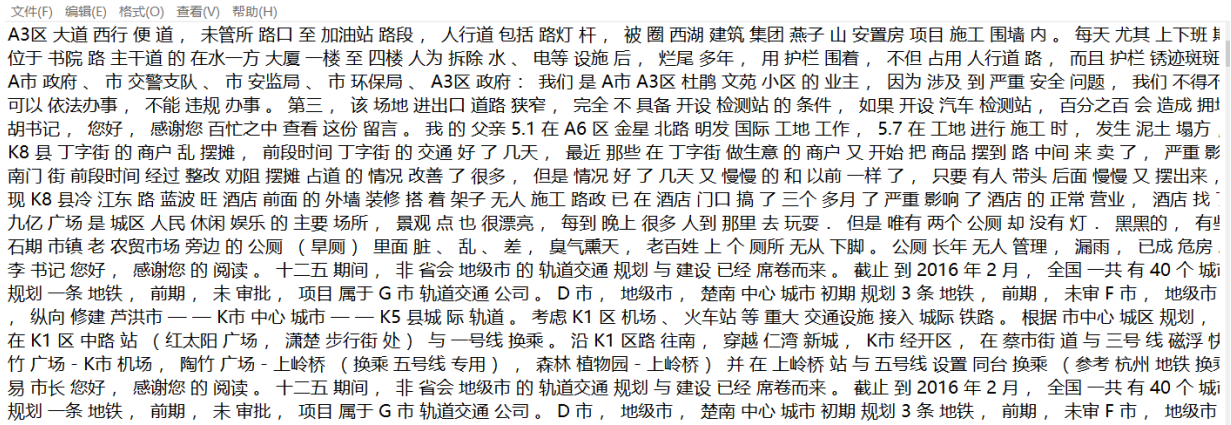

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
A3区 大道 西行 便道， 未管所 路口 至 加油站 路段， 人行道 包括 路灯 杆， 被 圈 西湖 建筑 集团 燕子 山 安置房 项目 施工 围墙 内 。 每天 尤其 上下班 其
位于 书院 路 主干道 的 在水一方 大厦 一楼 至 四楼 人为 拆除 水 、 电等 设施 后， 烂尾 多年， 用 护栏 围着， 不但 占用 人行道 路， 而且 护栏 锈迹斑斑
A市 政府 、 市 交警支队 、 市 安监局 、 市 环保局 、 A3区 政府： 我们 是 A市 A3区 杜鹃 文苑 小区 的 业主， 因为 涉及 到 严重 安全 问题， 我们 不得不
可以 依法办事， 不能 违规 办事 。 第三， 该 场地 进出口 道路 狭窄， 完全 不 具备 开设 检测站 的 条件， 如果 开设 汽车 检测站， 百分之百 会 造成 拥堵
胡书记， 您好， 感谢您 百忙之中 查看 这份 留言 。 我 的 父亲 5.1 在 A6 区 金星 北路 明发 国际 工地 工作， 5.7 在 工地 进行 施工 时， 发生 泥土 塌方，
K8 县 丁字街 的 商户 乱 摆摊， 前段时间 丁字街 的 交通 好 了 几天， 最近 那些 在 丁字街 做生意 的 商户 又 开始 把 商品 摆到 路 中间 来 卖 了， 严重 影
南门 街 前段时间 经过 整改 劝阻 摆摊 占道 的 情况 改善 了 很多， 但是 情况 好 了 几天 又 慢慢 的 和 以前 一样 了， 只要 有人 带头 后面 慢慢 又 摆出来，
现 K8 县冷 江东 路 蓝波 旺 酒店 前面 的 外墙 装修 搭 着 架子 无人 施工 路政 已 在 酒店 门口 搞 了 三个 多月 了 严重 影响 了 酒店 的 正常 营业， 酒店 找
九亿 广场 是 城区 人民 休闲 娱乐 的 主要 场所， 景观 点 也 很漂亮， 每到 晚上 很多 人到 那里 去 玩耍 . 但是 唯有 两个 公厕 却 没有 灯 . 黑黑的， 有
石期 市镇 老 农贸市场 旁边 的 公厕 （ 早厕 ） 里面 脏 、 乱 、 差， 臭气熏天， 老百姓 上 个 厕所 无从 下脚 。 公厕 长年 无人 管理， 漏雨， 已成 危房
李 书记 您好， 感谢您 的 阅读 。 十二五 期间， 非 省会 地级市 的 轨道交通 规划 与 建设 已经 席卷而来 。 截止 到 2016 年 2 月， 全国 一共 有 40 个 城
规划 一条 地铁， 前期， 未 审批， 项目 属于 G 市 轨道交通 公司 。 D 市， 地级市， 楚南 中心 城市 初期 规划 3 条 地铁， 前期， 未审 F 市， 地级市
， 纵向 修建 芦洪市 — — K市 中心 城市 — — K5 县城 际 轨道 。 考虑 K1 区 机场 、 火车站 等 重大 交通设施 接入 城际 铁路 。 根据 市中心 城区 规划，
在 K1 区 中路 站 （ 红太阳 广场， 潇楚 步行街 处 ） 与 一号线 换乘 。 沿 K1 区路 往南， 穿越 仁湾 新城， K市 经开区， 在 蔡市街 道 与 三号 线 磁浮 快
竹 广场 - K市 机场， 陶竹 广场 - 上岭桥 （ 换乘 五号线 专用 ） ， 森林 植物园 - 上岭桥 ） 并 在 上岭桥 站 与 五号线 设置 同台 换乘 （ 参考 杭州 地铁 换乘
易 市长 您好， 感谢您 的 阅读 。 十二五 期间， 非 省会 地级市 的 轨道交通 规划 与 建设 已经 席卷而来 。 截止 到 2016 年 2 月， 全国 一共 有 40 个 城
规划 一条 地铁， 前期， 未 审批， 项目 属于 G 市 轨道交通 公司 。 D 市， 地级市， 楚南 中心 城市 初期 规划 3 条 地铁， 前期， 未审 F 市， 地级市

图 3.1.1 部分中文分词结果

如图所示的分词结果是没有停用词过滤的结果，可以看到，其中有大量标点及表达无意义的字词，对后续分析会造成很大影响，因此接下来需要进行停用词过滤。

（四）停用词过滤

为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词）。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。词条的文档频率低于某个阈值是低频词，低频词不含或含有较少的类别信息，一般会从文档特征的空间中移除这类词，可以降低文档特征空间的维数，也有可能提高分类的精度，尤其是当低频词是噪音词条时。而高于阈值的词称为中频词和高频词，这类词含有较多类别信息，对分类结果影响较大，分类时应该保留。

文档频率特征提取简单，其时间复杂度和文本个数呈线性计算复杂度关系，因此常被用于大规模和超大规模文本数据统计和处理中。

1、词频(TF)

TF 是一种简单的评估函数，其值为训练集合中此单词发生的词频数。 TF 评估函数的 理论假设是当一个词在大量出现时，通常被认为是噪声词。

2、文档频数(DF)

DF 同样是一种简单的评估函数，其值为训练集合中包含此单词的文本数。DF 评估函 数的理论假设是当一个词在大量文档中出现时，这个词通常被认为是嗓声词。

本文选用 DF 方法筛选出如下停用词：我，的，了，是等。将筛选出的停用词加入停用词表，再利用停用词表过滤停用词，将分词结果与停用词表中的词语进行匹配， 若匹配成功，则进行删除处理。去除停用词后的部分结果示例如图：

A3区 大道 西行 道 未管所 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑 集团 燕子 山 安置房 项目 施工 围墙 上下班 期间 条 路上 人流 车流 安全隐患 请
位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水 电等 设施 烂尾 多年 护栏 围着 占用 人行道 路 护栏 锈迹斑斑 倒塌 危机 过往行人 车辆 请求 部门
A市 政府 市 交警支队 市 安监局 市 环保局 A3区 政府 A市 A3区 杜鹃 文苑 小区 业主 涉及 网上 写信 方式 一件 引发 安全事故 杜鹃 路 雷峰 大道 交界处 杜鹃
胡书记 您好 感谢您 百忙之中 查看 这份 留言 父亲 5.1 A6 区 金星 北路 明发 国际 工地 工作 5.7 工地 施工 时 发生 泥土 塌方 受伤 治疗 期间 工地 拒绝 支付
K8 县 丁字街 商户 乱 摆摊 前段时间 丁字街 交通 几天 丁字街 做生意 商户 商品 摆到 路 卖 影响 这条 街 交通 摩托车 城管局 领导 制定 措施 制止 形为
南门 街 前段时间 整改 劝阻 摆摊 占道 情况 改善 情况 几天 慢慢 有人 带头 慢慢 摆出来 商户 干脆 钩子 货物 挂 门口 屋檐下 电线 上有政策 对策 城管 检查 稍
现 K8 县冷 江东 路 蓝波 旺 酒店 外墙 装修 搭 架子 无人 施工 路政 酒店 门口 搞 三个 多月 影响 酒店 营业 酒店 找 施工队 人员 情况 时间 搞好 营业 施工人员
九亿 广场 城区 休闲 娱乐 场所 景观 点 很漂亮 每到 晚上 人到 玩耍 两个 公厕 灯 黑黑的 外面 大小便 影响 不好 如果说 灯 不好 管理 景观灯 并网 开关 希望 解
石期 市镇 农贸市场 旁边 公厕 早厕 脏 乱 差 臭气熏天 老百姓 厕所 无从 下脚 公厕 长年 无人 管理 漏雨 已成 危房 这座 早厕 气味 难闻 夏天 蚊蝇 乱飞 安全卫
李 书记 您好 感谢您 阅读 十二五 期间 非 省会 地级市 轨道交通 规划 建设 席卷而来 截止 2016 年 月 全国 一共 40 城市 获批 轨道交通 建设 未含 有轨电车 规
株潭 城际 铁路 待建 https baidu com 公里 总 投资 152.8 亿元 J 市 磁浮 快线 郴 资永线 投资 156.6 亿元 涵盖 景区 交通枢纽 兼顾 城区 出行 客流 K市 潇楚
式 地铁 跨座式 轻轨 号线 蓝线 马坪 两 中心 预留 北延线 芦洪 市镇 规划 号线 北起 马坪 经开区 M9 县路 穿越 M9 县园 串联 K1 区 中路 商圈 K1 区 中路 站
城 K市 机场 K市 火车站 二号线 换乘 客流 引入 K1 区 中路 商圈 M9 县园 滨江 新城 一号线 换乘 客流 引入 滨江 新城 中心 汽车站 K1 区 城区 四号线 换乘 K3
易 市长 您好 感谢您 阅读 十二五 期间 非 省会 地级市 轨道交通 规划 建设 席卷而来 截止 2016 年 月 全国 一共 40 城市 获批 轨道交通 建设 未含 有轨电车 规
株潭 城际 铁路 待建 https baidu com 公里 总 投资 152.8 亿元 J 市 磁浮 快线 郴 资永线 投资 156.6 亿元 涵盖 景区 交通枢纽 兼顾 城区 出行 客流 K市 潇楚
式 地铁 跨座式 轻轨 号线 蓝线 马坪 两 中心 预留 北延线 芦洪 市镇 规划 号线 北起 马坪 经开区 M9 县路 穿越 M9 县园 串联 K1 区 中路 商圈 K1 区 中路 站
城 K市 机场 K市 火车站 二号线 换乘 客流 引入 K1 区 中路 商圈 M9 县园 滨江 新城 一号线 换乘 客流 引入 滨江 新城 中心 汽车站 K1 区 城区 四号线 换乘 K3

图 3.1.2 部分停用词过滤后结果

3.2 文本向量化

3.2.1 特征值提取

经过上述文本预处理后，虽然已经去掉部分停用词，但还是包含大量无用词语，给文本向量化过程带来困难，所以特征抽取的主要目的是在不改变文本原有核心信息的情况下 尽量减少要处理的词数，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。本文利用的方法是词频-逆向文档频率(TF-IDF)。

3.2.2 TF-IDF 权重矩阵

TF-IDF 算法 ：是一种用于信息检索与数据挖掘的常用加权技术。TF 意思是词频（是一词语出现的次数除以该文件的总词语数。），IDF 意思是逆文本频率指数。

在对信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。

这里采用 TF-IDF 算法，把信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

词频（TF）= 某个词在文本中出现的次数

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频（TF）} = \frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}}$$

或

$$\text{词频（TF）} = \frac{\text{某个词在文本中的出现次数}}{\text{该文本出现次数最多的词的出现次数}}$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（train_corpus），用来模拟语言的使用环

境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF）

实际分析得出 TF-IDF 值与一个词在表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的表中文本的关键词。

3.2.3 Word 2 vec 词向量模型

在附件中的留言，首先按照一定的划分体系（附件 2）对留言进行分类，以便后续将群众留言分派来自相应的级别处理，目前大部分电子政务系统还是依靠人工根据经验处理存在工作量大，效率低且差错率高等问题。所以要根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型，实际上就是文本分类，也就是多分类的问题，一般常用的就是 Word 2 vec 工具，在 NLP 领域具有非常重要的意义。

Word2vec 输出的词向量可以被用来做很多 NLP 相关的工作，，那么 Word2vec 就可以把特征映射到 K 维向量空间，可以为文本数据寻求更加深层

	V1	V2
1	(0, 6057)	.167428371664318
2	(0, 6025)	.177348382089821
3	(0, 287)	.092144754358712
4	(0, 6058)	.191329851513102
5	(0, 9427)	.131012986273248
6	(0, 4132)	.150686825855881
7	(0, 9694)	.191329851513102
8	(0, 1487)	.201249861938605
9	(0, 9646)	.177348382089821
10	(0, 6497)	.137601057349003
11	(0, 684)	.215231331361886
12	(0, 3504)	.183635293640394
13	(0, 6126)	.143526891815534
14	(0, 10430)	.135832333942826
15	(0, 4148)	.215231331361886
16	(0, 7435)	.215231331361886
17	(0, 10358)	.183635293640394
18	(0, 4861)	.139465432817756
19	(0, 9177)	.215231331361886
20	(0, 9656)	.183635293640394
21	(0, 2629)	.143526891815534
22	(0, 1496)	.177348382089821
23	(0, 9655)	.323949598646362
24	(0, 2571)	.215231331361886
25	(0, 9650)	.183635293640394
26	(376, 9442)	.049816050837817
27	(376, 2803)	.111411588613371

次的特征表示，如图：

图 3.2.3 词袋向量的对应 TF-IDF 值

如图所示的词袋向量，括号里的是行数和词编号，右边的值为该词的权重。

3.2.5 朴素贝叶斯分类器

贝叶斯分类法可以通过预测类成员关系的可能性，进而进行分类。贝叶斯分类算法基于贝叶斯定理。其基本思想是假设留言中的特征项是相互独立的，在此前提下利用贝叶斯公式计算文档属于各个类别的概率，选择概率值最大的类别为最终结果，这一假定可以简化所需要的计算。

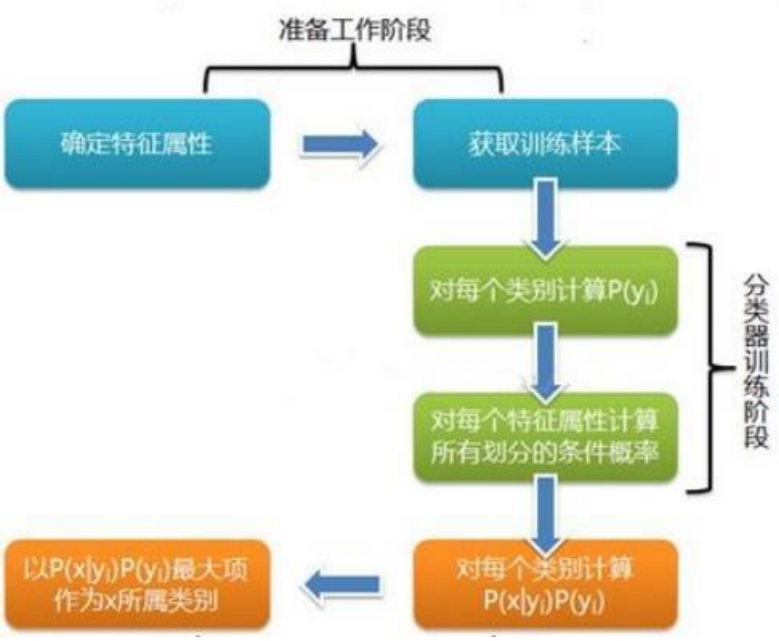


图 3.2.6 使用朴素贝叶斯分类器的准备工作阶段

朴素贝叶斯分类器是各种分类器中分类错误概率最小或者在预先给定代价的情况下平均风险最小的分类器。它的设计方法是一种最基本的统计分类方法。其分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。贝叶斯分类器是用于分类的贝叶斯网络。

文本 d_j 与类 c_i 的条件概率，使用贝叶斯公式估计如公式（2-14）：

$$P(c_i \mid d_j) = \frac{P(c_i)P(d_j \mid c_i)}{P(d_j)}$$

其中 $P(d_j)$ 对计算结果无影响，可以忽略不计算。因此根据单词间的独立性假设，各个 特征值独立地给出类标号，极大地简化计算：

$$P(d_j \mid c_i) = \prod_{k=1}^{|V|} P(w_{kj} \mid c_i)$$

假设 $N(w_t, d_i)$ 为特征 w_t 在文本 d_i 中出现的频率数，而 $Pr(c_j \mid d_i) \in \{0,1\}$ 表示文本 d_i 是否在类 c_i 中出现，为 $Pr(c_j \mid d_i)=1$ 时，代表文本 d_i 在类 c_j 出现，否则 $Pr(c_j \mid d_i)=0$ 代表未出现，而 $P(c_i)$和 $P(W_{kj} \mid c_i)$可以用下式来进行估计：

$$\hat{P}(c_j) = \frac{\sum_{i=1}^{|D|} P(c_j \mid d_i)}{|D|}$$

$$\hat{P}(w_t \mid c_j) = \frac{\sum_{i=1}^{|D|} N(w_t, d_i) P(c_j \mid d_i)}{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(w_s, d_i) P(c_j \mid d_i)}$$

下图为介绍朴素贝叶斯分类器实现后的效果

```
训练数据读入....
1       2       3       4       5       6       7       8       9       10
11      12      13      14      15      16      17      18      19      20
21      22      23      24      25      26      27      28      29      30
31      32      33      34      35      36      37      38      39      40
41      42      43      44      45      46      47      48      49      50
51      52      53      54      55      56      57      58      59      60
61      62      63      64      65      66      67      68      69      70
71      72      73      74      75      76      77      78      79      80
81      82      83      84      85      86      87      88      89      90
91      92      93      94      95      96      97      98      99      100
101     102     103     104     105     106     107     108     109     110
111     112     113     114     115     116     117     118     119     120
121     122     123     124     125     126     127     128     129     130
131     132     133     134     135     136     137     138     139     140
141     142     143     144     145     146     147     148     149     150
151     152     153     154     155     156     157     158     159     160
161     162     163     164     165     166     167     168     169     170
```

图 3.2.5 构建朴素贝叶斯分类器

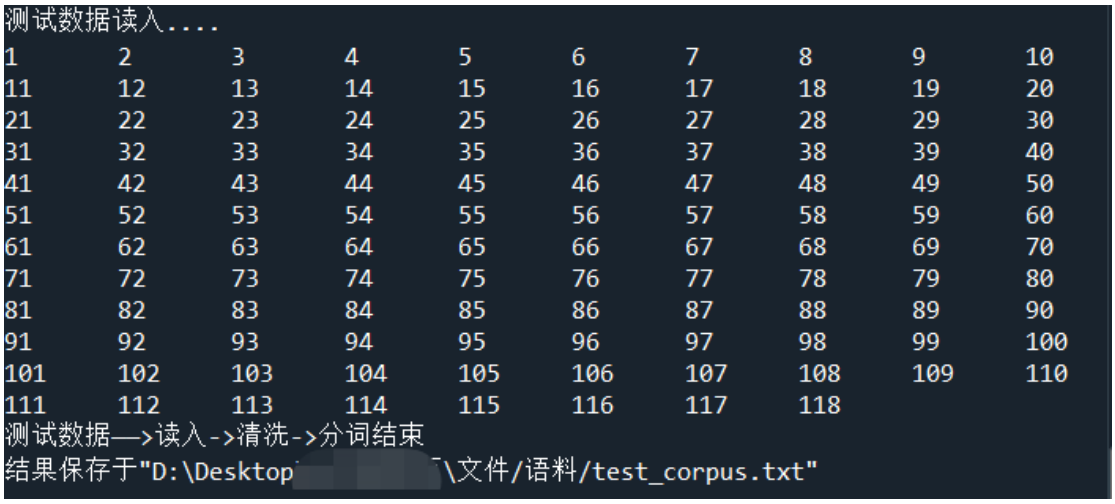


图 3.2.6 构建朴素贝叶斯分类器

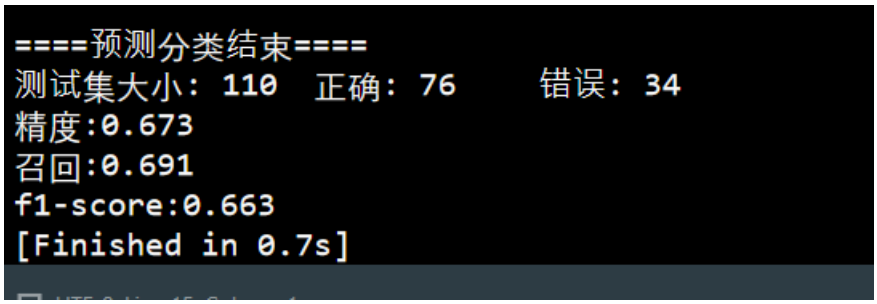


图 3.2.7 测试分类效果

3.2.6 向量化语义

对某一特征项为n 的文本向量t 进行奇异值分解得到：

$$t = t^{'} \sum U$$

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了同一空间里。

3.3 文本聚类

3.3.1 文本聚类

（一）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把相同的留言问题进行合并，再累加。即可获取热点问题。

由于计算机不能够直接处理文本信息，我们需要对文本进行处理，将文本表示成为计算机能够直接处理的形式，即文本数字化。文本表示(Test Expression)也称为文本特征表达，它不仅要求能够真实准确的反映文档的内容，而且要对不同的文档具有区分能力。目前常用的文本表示模型有向量空间模型、布尔模型和概率模型等。

而向量空间模型[8](Vector Space Model, VSM)最早是由 Salton 和 McGill于20 世纪60年代末提出的，是目前在文本挖掘技术中最常用的表示模型。

其主要思想：将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项 (词条)对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重，这里的权重可以由 TF-IDF 等算法得到。向量空间模型就是将文本表示成为一个特征向量：

$$V(d)=(t_1,w_1(d),t_2,w_2(d),\ldots,t_n,w_n(d))$$

3.3.2 K-means 文本聚类

K-means 算法[13]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。

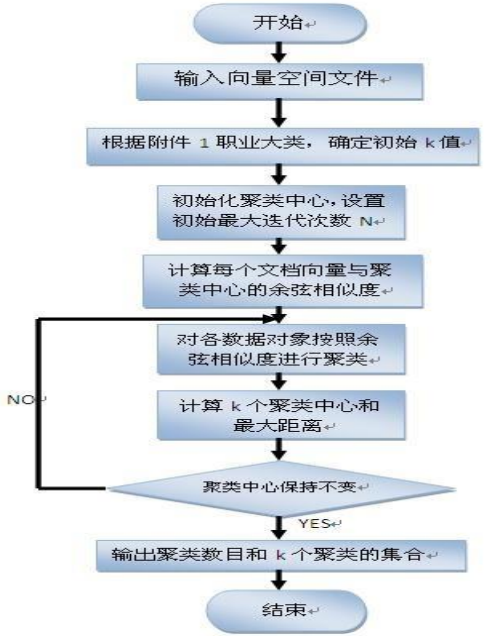
假设要把样本集分为k 个类别，算法描述如下：

（1）适当选择k 个类的初始中心；

（2）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；

（3）利用均值等方法更新该类的中心值；

（4）对于所有的k 个聚类中心，如果利用(2)(3)的迭代法更新后，值保持不变，则 迭代结束，否则继续迭代。



图为 3.3.2 文本聚类流程图

该算法要求在计算之前给定k 值。由附件1 及附件 2 得出。已有的分类数据有 7 个一级分类，因此在本文中暂时使用 7 个分类，以保证能快速分类的功能，在使用的一级分类中确定出k 的值，这里令$k=7$，即″城乡建设″，″环境保护″，″交通运输″，″教育文体″，″劳动和社会保障″，″商贸旅游″，″卫生计生″。

```
catelist = ["城乡建设", "环境保护", "交通运输", "教育文体", "劳动和社会保障", "商贸旅游", "卫生计生"]
nums = []
contents = []
```

图为 3.3.3 一级分类的数据训练

4. 结论

总结本次比赛，我们基于 TFIDF 权重法提取特征词，建立词袋模型，进一步运用 Word2vec 来对文本进行向量化，构造朴素贝叶斯分类器对向量化的数据进行处理，用训练数据进行训练模型。对测试数据进行测试，得出了为 90.6%的匹配准确率。

但是我们最后得到的匹配效果还可以更好，对词袋模型的处理，没有对其结构化，这可能造成后面对数据的处理不够好，这也涉及到当今中文文本挖掘模型的不足，我们后期也会进一步对文本挖掘进行深入探讨。

5. 参考文献

[1]刘健，张维明. 基于互信息的文本特征选择方法研究与改进[J]. 计算机工程与应用，2008，(10)：135．137．

[2]刘海峰，姚泽清，苏展．基于词频的优化互信息文本特征选择方法[J]．计算机工程，20 1 4，40(07)：1 79—1 82．

[3]石慧，贾代平，苗培．基于词频信息的改进信息增益文本特征选择算法[J]-计算机应用，2014，34(1 1)：3279—3282．

[4]黄章树，叶志龙．基于改进的 CHI 统计方法在文本分类中的应用[J]．计算机系统应用，2016，25(1 1)：136—140．

[5]阮光册，夏磊. 基于关联规则的文本主题深度挖掘应用研究[J]. 现代图书情报技术，2016，12：50—56．

[6]王鹏，高铖，陈晓美．基于 LDA 模型的文本聚类研究[J]-情报科学，2015，01：63—68．

[7]张俊妮．统计模型在中文文本挖掘中的应用[J]．数理统计与管理，2017，02：1—18．

[8]周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 中国科学院研究生院(计算技术研究