关于自然语言处理在“智慧政务”中的文本挖掘应用

摘要

本文利用群众留言记录及相关部门的答复意见数据，采用算法对数据进行文本内在信息的挖掘与分析，建立关于留言内容的一级标签分类，分析并提炼出留言中相关的热点问题，建立有效数学模型进行评价，为“智慧政务”提炼出有价值的留言信息，也为进一步提升政府的管理水平和施政效率提供有用依据。

针对问题一，利用 Python 语言对附件 2 中评论数据进行数据预处理、利用 jieba 进行中文文本分词、内容过滤、数据清洗、tf－idf 算法进行词频统计与文本向量化、再利用随机森林算法建立模型以实现文本分类和数据优化，提升数学模型的可建模度。

针对问题二，利用 Python 调用 Hanlp 做命名实体识别并提取关键字进行文本聚类，后通过 LDA 主题模型思想对热点问题的评论数据进行统计和排名最后整合出结果。

针对问题三，通过 tf－idf 算法和 jieba 建立文本相似度匹配模型，并利用 Excel 进行数据处理，然后从完整性、相关性、可解释性角度给出关于答复意见质量的评价方案，并实现所建立的评价方案。

关键词：标签分类　评论数据　关键字提取　LDA

目录

一、 挖掘目标

本次建模目标是通过研究群众问政留言记录及相关部门对部分群众留言的意见答复文本数据，借助 Python、Excel 等软件对文本数据进行基本的预处理、利用 jieba 进行中文文本分词、tf – idf 算法进行词频统计与文本向量化、再利用随机森林算法建立模型以实现文本分类，通过对数据的深入研究挖掘与分析，建立 LDA 模型等多种数据挖掘模型，实现对文本数据的倾向性判断和内在信息的挖掘与分析，获得有价值的内容。

二、 分析方法与过程

2.1 总体流程

本文对群众留言记录及相关部门答复意见进行深入的分析和研究，通过数据处理文本评论的详细情况，建立相关数学模型，解决问题。具体建模步骤如下所示：

第一步：分析问题和原始数据，确定基本思路；

第二步：采用合适的方法对数据进行预处理；

第三步：建立数学模型，对文本数据进行分析和挖掘；

第四步：从分析结果中获取相关结论，并建立一套评价体系。

2.2 数据预处理

根据题目要求读取附件 2 数据，提取留言详细与一级标签的列表。提取文本后，我们首先对文本评论数据进行预处理。去除两列表内容中的所有非中文信息（包括英文字符和数字）、进行数据清洗（去除无用、重复数据），因为如果将这些评论数据也列入接下来的分词、数据分析等步骤，会导致分析结果的准确性下降，使得结果存在问题。然后对列表中的内容运用的是 jieba 库进行分词处理。接着用 TF-IDF 统计方法进行文本向量化，并建立随机森林模型进行测试。

2.2 1 文本向量化

tf－idf 是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。[1]字词的重要性随着它在文件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。tf－idf 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。其中：

$$\text{tf} = \frac{\text{一个词在某个文件中出现的次数}}{\text{该文件中所有词出现的次数总和}}$$

$$\text{idf} = \frac{\log \text{文档总数}}{\text{包含这个词的所有文档数之和+1}}$$

idf 表示一个词能将当前文件和其他文件区分开的能力，越大越好。

tf 表示某个词出现在这段文本的频率，越大越好。

在文本分类中，我们把每个类别解释为一个文件，重新定义 tf，idf：

$$\mathrm{tf}=\frac{\text{一个词在某个分类的所有样本中出现的次数}}{\text{这个分类中所有样本包含的所有词出现的次数总和}}$$

$$\mathrm{idf}=\frac{\log \text{所有分类包含的样本总数}}{\text{包含这个词的所有分类的所有样本数之和}}_{\mathrm{i}}$$

2.2.2 随机森林模型

假设训练集中 n 个样本，每个样本有 d 个特征，需要训练一个包含 T 棵数的随机森林[2]，具体的算法流程如下所示：

（1）对于 T 棵决策树，分别重复如下操作：

A、使用 Bootstrap 抽样，从训练集 D 获得大小为 n 的训练集 D；

B、从 d 个特征中随机选取 m。

（2）根据投票原则，确定最终的类别。

每棵树的生成都是随机的，至于随机选取的特征数，决定随机选取的特征数的大小主要有两种方法：

A、交叉验证；

B、经验性设置 $\mathrm{m}=\log_2 d+1$。

2.3 针对热点问题的分析与挖掘过程

2.3.1 分析思路

由于热点问题是针对某一件问题的多频次反映，而一个实例问题的影响范围和影响人群都有实质边界，所以我们将从下面几个步骤来建立数学模型：

（1） 先采用 HanLP 命名实体识别技术来提取所有特定地点或特定

组织；

（2） 然后将特定地点或特定组织的名字作为关键字进行文本聚类，以获得各地点所存在的问题集合；

（3） 将各问题集合通过 LDA 主题生成模型获得各问题集合中反映最多的问题，并通过文本相似度计算将各问题集合中与主题问题最相似的问题提取出来，从而产生各个热点问题的集合；

（4） 通过计算各热点问题的集合的元素个数，以此来反映一个热点问题的热度并作为评价标准来获得前 5 个热点问题的热度指数与热度排名；

（5） 计算各热点问题的时间跨度并整合数据得到最终结果。

2.3.2 HanLP 命名实体识别

HanLP 是由一系列模型与算法组成的工具包，目标是普及自然语言处理（NLP）在生产环境中的应用。[3]HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点；能提供词法分析（中文分词、词性标注、命名实体识别）、句法分析、文本分类和情感分析等功能。HanLP 内部的算法不依赖第三方类库，可以自主控制文本所有细节。

HanLP 采取算法与语料库分离的模式，节省人力成本，同时运用 HanLP 还能使文本上的准确率大幅提升。

其主要运行步骤如下图 1：

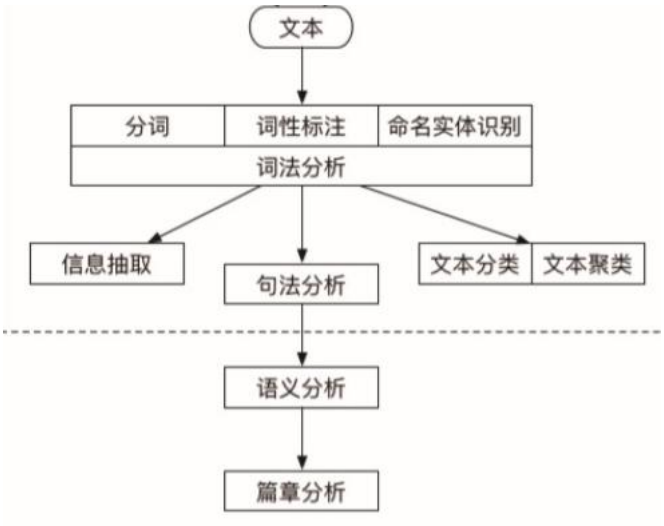

图 1　HanLP 运行步骤

2.3.3　LDA 主题模型

2.3.3.1 LDA 主题模型介绍

LDA 模型是一种经典的贝叶斯模型，分为文档集层、主题层及特征词层，每层均有相应的随机变量或参数控制。[4]所谓生成模型，就是每一篇文档都是由其隐含的主题随机组合生成，每一个主题对应特定的特征词分布，不断抽取隐含主题及其特征词，直到遍历完文档中的全部单词。因此把各个主题在文档中出现的概率分布称为主题分布，把各个词语在某个主题中出现的概率分布称为词分布。

利用 LDA 模型生成文档，首先需要生成该文档的一个主题分布，再生成词的集合；根据文档的主题分布随机选择一个主题，然后根据主题中词分布随机选择一个词，重复这个过程直至形成一篇文档。同时 LDA 主题模型对文档集进行内部语义信息分析，从而得出 3 层贝叶斯模型：文本-主题-特征词模型。假设所有文档存在 K 个隐含主题，

LDA 的图模型结构如图 2 所示。

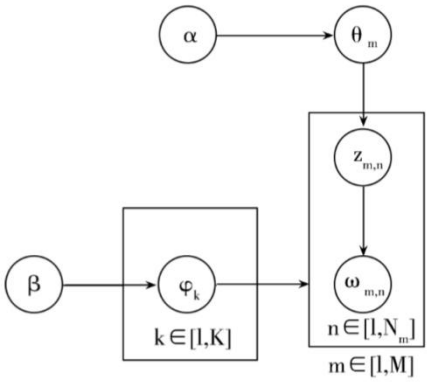

图 2 LDA 图模型结构

给定一个文档集合 $D = \{d_1, d_2, \ldots, d_M\}$，$M$ 为文档总数，N_m 为第 m 篇文档的单词总数。假设主题数为 K 个，α和β表示语料级别的参数，取先验分布为 Dirichlet 分布，其中α是每篇文档下主题的多项分布的 Dirichlet 先验参数，β是每个主题下特征词的多项分布的 Dirichlet 先验参数。主题向量θ_m 为第 m 篇文档能生成主题的概率分布变量，每个文档对应一个θ，φ_k代表第 k 个主题下的特征词分布变量。对于第 m 篇文档中的第 n 个特征词 $\omega_{m,n}$生成第 m 篇文档中第 n 个词的主题$z_{m,n}$。$\omega_{m,n}$是可以观测到的已知变量，α和β是根据经验给定的先验参数，把θ_m，φ_k和 $z_{m,n}$当作隐藏变量，根据观察到的已知变量估计预测。

根据 LDA 的图模型结构，可以推理出所有变量的联合分布：对于第 m 篇文档中第 n 个特征词$\omega_{m,n}(1 \le n \le N_m, 1 \le m \ll M)$，生成一个主题$z_n$服从隐含变量θ的多项式分布；根据先验参数β，对特征词ω_n生成 $P(\omega_n | z_n, \beta)$。

则每个文档的概率密度函数为：

$$\mathrm{P}(\omega|\alpha,\ \beta)=\int P(\theta|\alpha)\,(\prod_{n=1}^{N}\sum_{z_n}P(z_n|\theta_d)\,P(\omega_n|z_n,\ \beta))d\theta$$

2.3.3.2 LDA 主题模型估计

LDA 主题模型中常用的估计参数方法有变分推理、Laplace 近似、Gibbs Sampling 抽样算法和期望-扩散。[5]本文使用马尔科夫链蒙特卡洛（Markov Chain Monte Carlo，MCMC）算法中的 Gibbs Sampling 抽样算法。利用 Gibbs Sampling 算法对 LDA 模型进行参数估计的过程如下:

由吉布斯更新规则：

$$\mathrm{P}\left(z_i=k\middle|z_{\neg i},\ \omega\right)=\frac{P(\omega|z)}{P\left(\omega_{\neg i}\middle|z_{\neg i}\right)P(\omega_i)}\cdot\frac{P(z)}{P\left(z_{\neg i}\right)}$$

$$\propto\frac{\Delta(n_z+\beta)}{\Delta\left(n_{z,\ \neg i}+\beta\right)}\cdot\frac{\Delta(n_m+\alpha)}{\Delta\left(n_{m,\ \neg i}+\alpha\right)}$$

其中$z_i=k$ 表示特征词ω_i属于第 k 个主题的概率，$z_{\neg i}$表示其他所有词的概率，$n_{z,\ \neg i}$表示不包含当前特征词ω_i被分配到第 k 个主题下的个数。

当 Gibbs 结果收敛后，可以推导出特征词ω_i在第 k 个主题中的分布参数估计$\varphi_{k,\ l}$和多项式分布参数估计$\theta_{m,\ k}$：

$$\varphi_{k,\ l}=\frac{n_k^{(l)}+\beta_l}{\sum_{l=1}^{V}n_k^{(l)}+\beta_l}$$

$$\theta_{m,\ k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k}$$

LDA 主题模型在文本聚类、主题挖掘、相似度计算等方面都有广泛的应用，并引用了狄利克雷函数知识，该模型的泛化能力较强，不容易出现拟合现象。其次，LDA 主题模型是一种无监督模式，只需提供文本即可得出概率，节省大量时间精力。而且，LDA 主题模型也可以解决多种代指问题，减少有价值文本的缺失度。

2.4 评价方案模型的建立

根据题目要求，需要建立一套对答复意见的质量的评价方案。结合数据以及实际情况分析，本文将给出从答复的相关性、完整性、可解释性角度对答复意见的质量建立的一套评价方案。

2.4.1 文本相似度匹配模型

利用 jieba 模块的分词后，提取关键字，并使用自定义词典等方法，通过 tf－idf 算法与 jieba 建立文本相似度匹配模型，。[6]由此模型联系答复的相关性问题。具体步骤如下：

（1）读取文档；

（2）对要计算的文档进行分词；

（3）对文档进行整理成指定格式，方便后续进行计算；

（4）计算出词语的频率；

（5）通过语料库建立词典；

（6）加载要对比的数据；

（7）将要对比的文档通过 doc2bow 转化为稀疏向量；

（8）对稀疏向量进行进一步处理，得到新语料库；

（9）将新语料库通过 tf－idf model 进行处理，得到 TF－IDF；

（10）通过 token2id 得到特征数；

（11）稀疏矩阵相似度，从而建立索引；

（12）得到最终相似度结果。（其中越接近 1，相似度速度越高，即越匹配。）

2.4.3 对答复意见质量的评价方案

设置满分为 100 分，其中完整性、相关性、可解释性这三项分值分别为 40 分，40 分，20 分（每一处答复内容的完整性、相关性、可解释性分数比值也是 2:2:1，最后由这三项的中位数的总和代表整体的分数。）。

（1）完整性（总分为 40 分）

主要是文本的完整性，体现为文本格式是否准确，解释是否清楚，回复内容是否详细等方面。

① 通过内容体现（总分 30 分）

内容体现分为格式和内容详细程度两个方面，字数在 358 字到 1000 字为宜，大于或小于此区间则扣除 5 分。

② 通过格式体现（总分 10 分）

格式：有……反映，……答复如下：您好/你好，感谢……，……的情况已收悉，……（回复日期）。

（2）相关性（总分为 40 分）

根据所建立的文本相似度匹配模型进行评分，打印出来的相似度都在 0-1 之间，越接近 1 表示相似度速度越高。

（3）可解释性（总分为 20 分）

参考《书记市长网上留言回复工作规范》[7]，申诉类、投诉类要在 10 个工作日内办理回复。对于无法在 10 个工作日内办理完毕的投诉类信件，要将信件所反映问题的最新办理进展情况及时回复来信人，有了最终办理结果后再进行二次回复。但最长时限不得超过 30 日。逾期未按规定结案，也未说明情况，造成群众重信重访出现的，将视情况给所在单位进行通报批评。

并有以下设定：若答复天数少于 10 天则为满分，若答复时间大于 10 天小于 20 天则为 15 分，若答复时间大于 20 小于 30，则为 10 分；若大于 30 天则为 0 分。

三、 结果分析

3.1 针对问题一的结果分析

3.1.1 利用 F－Score 对分类方法进行评价

F－Score 公式为：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

利用 F－Score 对分类方法进行评价可得出以下结果：

对一级标签分类后所得评价结果如图 3 所示：

一级分类标签	查准率	查全率
交通运输	0.93	0.65
劳动和社会保障	0.85	0.95
卫生计生	0.94	0.77
商贸旅游	0.86	0.72
城乡建设	0.78	0.91
教育文体	0.92	0.90
环境保护	0.92	0.87

图 3 F－score 评价结果图

所得F_1－Score 值为 0.848135865。由此可见，针对问题一所建立的随机森林模型是合理的。

3.2 针对问题二的结果分析

3.2.1 热度评价指标

对于热度评价指标，我们同时考虑到反映次数、点赞数与反对数等因素对于热点问题的影响，并使用以下算法进行热评指数计算：

$$y=(1+\alpha)x$$

其中，y 为热度指数，α为点赞数+反对数/100，x 为非重复性的反映次数。

3.2.2 热度评价结果

根据热度评价指标，我们可以求出前五个热点问题，内容如下图4所示：

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	51.3	2019/11/02至2020/1/26	A市A2区丽发新城小区	小区附近搅拌站噪音扰民，污染环境
2	2	22.68	2019/7/28至2020/12/04	A市A5区魅力之城小区	小区临街宵夜摊油烟噪音扰民
3	3	13.25	2019/3/26至2020/4/12	A市A6区月亮岛路	高压电线未埋地
4	4	10.26	2017/6/08至2020/11/27	A市经济学院学生	学校强制学生定点实习
5	5	5.25	2019/1/26至2020/6/04	A市A7县中茂城商铺租客	店铺老板拖欠费用

图4 热度评价结果

3.3 针对问题三的结果分析

3.3.1 由完整性角度给出的评价方案

完整性（总分为40分，本套评价方案可得35分）。

（1）通过内容体现（总分30分，本套评价方案可得30分）

通过LEN函数算得每条答复的字数，并通过进一步处理得到中位数为358.0540156。由于规定字数在358字到1000字不扣分，所以此步骤得5分。再通过IF函数得到最终字数得分，算得总体中位数为25。所以此步骤得25分，共30分。

（2）通过格式体现（总分10分，本套评价方案可得5分）

通过IF函数、ISNUMBER函数、FIND函数筛选得到不包含以上关键字的，并进行扣分，最终此项总体为5分。

3.3.2 由相关性角度给出的评价方案

相关性（总分为 40 分，本套评价方案可得 28 分）。

模型运行结果为： [0.6437381 0. 0. ...0. 0.73923959]

可知相似度较高，总体相似度由模型计算得 0.7，所以此时总体相似度得分为 28 分。

3.3.3 由可解释性给出的评价方案

可解释性（总分为 20 分，本套评价方案可得 15 分）。通过对比留言时间和答复时间，得出答复速度。通过 DATEDIF 函数算出留言日期和答复日期相差的天数，再通过 IF 函数每条留言的得分，最终通过 MEDIAN 函数得到总体的可解释性的中位数为 15。所以最终总体数据通过 Excel 计算得出结果为 15 分。

综合完整性、相关性、可解释性最终本套评价方案得分为 68 分。

四、 结论

本文通过对群众问政留言记录及相关部门对部分群众留言的答复意见数据的文本评论处理后，利用 Python 等多种工具建立了随机森林模型、LDA 主题模型、文本相似度匹配模型等其他数据挖掘模型，得到了具有一定价值的结果，实现了对群众问政留言记录及相关回复

的深入分析和挖掘，并给出一套从答复的相关性角度对答复意见的质量建立的一套评价方案，相信能为 “智慧政务”提供一定程度上的帮助。

当然，从我们的分析结果中也可以看出总体来讲效果还不是特别的好，比如我们所建立的由完整性、相关性、可解释性角度给出的评价方案，最终得分并不是特别高，我们也在反思是否可从其他角度进行评价方案的建立？从而得到分数更高的评价方案。这将是我们后期可以进一步对由不同角度给出的不同评价方案的实现价值继续研究并深入探讨的地方。

五、 参考文献

[1]今天也要笑笑鸭. 文本分类任务中 tf－idf 的理解[G/K]. https://blog.csdn.net/silent_crown/article/details/84657863, 2018-11-30.

[2]无味之味. RandomRorest 随机森林算法 [G].https://www.jianshu.com/p/d2c07986c819 ,2019-01-29.

[3]张贝贝. HanLP：一触即发 叩响自主创新之门[J]. 软件和集成电路, 2019.

[4]李霄野, 李春生, 李龙,等. 基于 LDA 模型的文本聚类检索[J]. 计算机与现代化, 2018, No.274(06):11-15.

[5]李湘东, 张娇, 袁满. 基于 LDA 模型的科技期刊主题演化研究 [J]. 情报杂志, 2014(7).

[6]温柔易谈. 3 大数据挖掘系列之文本相似度匹配 [G].https://www.cnblogs.com/liaojiafa/p/6287314.html,2017-01-15.

[7]南阳市人民政府. 《书记市长网上留言回复工作规范》 [B].http://www.henan.gov.cn/10ztzl/system/2012/08/14/010326166.shtml,2012-08-14.