“智慧政务”中文本信息的分析与挖掘

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，利用 jieba 分词对留言内容信息进行分词处理，得到分词数据之后基于贝叶斯算法建立 LDA 模型。

对于问题 2，首先通过 python 对留言进行分词停词处理，得到相关的关键词后再运用 Excel 的文本筛选将留言筛选分类，经过去假处理即重新分类留言内容与留言主题不符的留言后将最后汇总的数据统计出热度前 5 的留言并制定热度评价指标。

对于问题 3，根据数据挖掘与分析的文本特征，运用 LDA 主题模型，统计学知识，从相关部门对留言答复的相关性、完整性、可解释性等角度制定出一套合理的部门答复评价方案。

关键词：jieba 分词 TF-IDF 算法 LDA 建模 python Excel 筛选

Analysis and mining of text information in "intelligent government affairs"

Summary

In recent years, as online platforms such as Wechat, Weibo, mayor's mailbox and sunshine hotline have gradually become an important channel for the government to understand public opinions, pool people's wisdom and gather people's morale, the amount of text data related to various social sentiments and public opinions has been increasing, it brings great challenge to the work of the department which used to rely on manual to divide the message and arrange the hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, 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 can promote the management level and administration efficiency of the government greatly.

For Question 1, using the word segmentation of Jieba to process the message content information, after getting the word segmentation data, LDA model was established based on Bayesian Algorithm.

For Question 2, first use python to do a word segmentation

and stop processing on the message, and then use the Excel to filter the message into categories, after the treatment of the past is to re-classify the message content and message topics do not match the message will be the final data collected out of the top 5 hot messages and the development of heat evaluation indicators.

For Question 3, based on text features of data mining and analysis, using LDA subject model, statistical knowledge, from the relevant departments of the response to the message relevance, integrity, interpretability and other aspects of a reasonable set of departmental response evaluation program.

Keywords: Jieba segmentation TF-IDF Algorithm LDA modeling Python Excel screening

目录

1、挖掘目标

本次数据挖掘是利用网络问政平台发布的群众问政留言记录数据以及相关部门对部分群众的留言意见数据，对群众留言进行中文分词、停词过滤，运用方法及算法，达到以下三个目标：

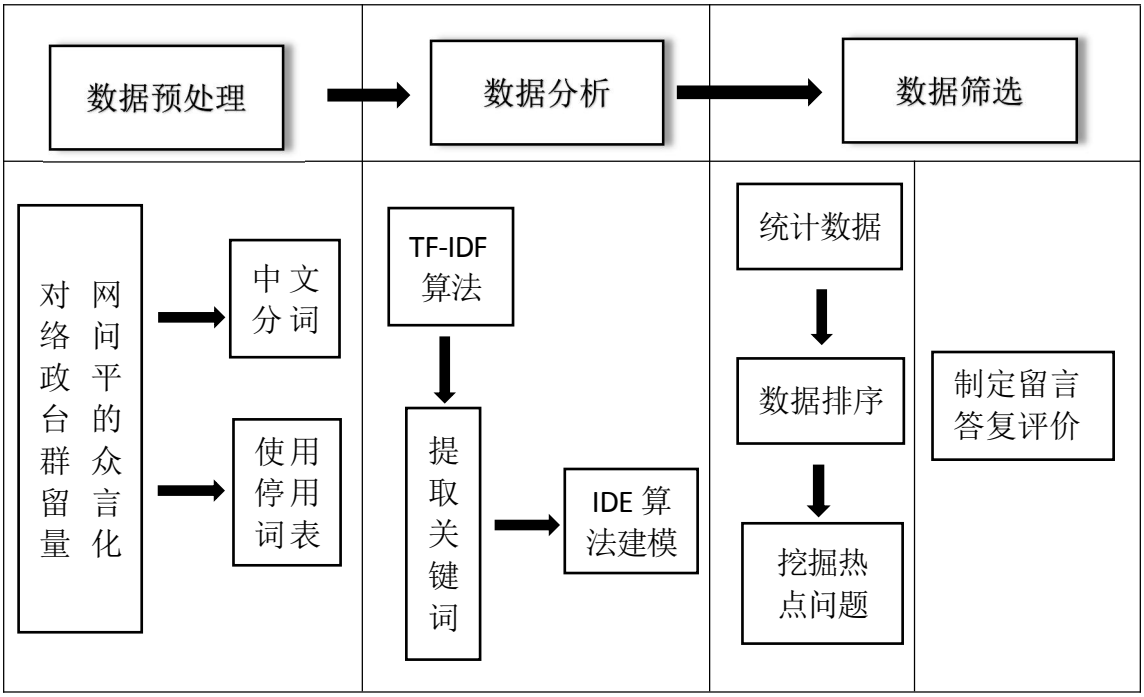
1）利用文本分词和文本类聚的方法对非结构化的数据进行文本挖掘，根据类聚结果，按照一定的划分体系，建立关于留言内容的一级分类标签，以便后续将群众留言分派至相应的职能部门。

2）根据某一时间段内反映特定地区或特定人群问题的留言进行分类，定义合理的热度评价指标，从而挖掘出群众所反映的热点问题。

3）根据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制定出一套评价方案。

2、分析方法与过程

总体流程图



2. 1 问题一分析方法与过程

2.1.1 数据预处理

2.1.1.1 留言信息的去重、去空、去假

在题目给出的留言数据中，出现了留言主题与内容重复且为同一用户的留言数据，在处理留言数据时，我们运用 Excel 对留言进行筛选重复项并进行删除。同时对留言主题或者留言详情为空的数据进行去空处理，并筛选出留言主题与留言内容不符合的留言进行去假处理。进行数据预处理后的留言数据保存为 csv 格式。

2.1.1.2 对留言主题和留言详情进行中文分词

利用 jieba 中文分词将连续的字序列按照一定的规范切分成一个一个单独的词。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。【1】

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）。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。但是，并没有一个明确的停用词表能够适用于所有的工具。甚至有一些工具是明确地避免使用停用词来支持短语搜索的。使用停用词表可以将无用词语筛选掉，主要包括英文字符、数字、数学字符、标点符号及使用频率特高的单汉字等。【2】

2.1.2 基于贝叶斯算法利用 LDA 建模

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA 是一种非监督机器学习技术，可以用来识别大规模文档集

（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

2.2 问题二分析方法与过程

2.2.1 数据筛选

（1）对附件 3 的留言数据进行去重处理，考虑到同一用户的重复留言将会对留言热度造成影响，故运用 Excel 对留言用户、留言主题以及留言详情同时进行重复项筛选并将其删除。

（2）使用 jieba 分词将留言主题进行分词处理然后通过 TF-IDF 算法提取出前 5 个关键词。

2.2.2 数据统计

（1）根据分词后的关键词对去重后的留言进行文本筛选，针对留言主题一列输入关键词将其筛选分类。

（2）留言分类完毕后，将重点筛选留言主题与留言详情内容不符的留言，并最终根据留言详情内容将其正确归类，汇总整理成正确的热度留言。

（3）对汇总后的留言进行数目统计，若某两类留言数目相同则综合留言的点赞数选择点赞数目更多的那类留言，选出留言数目前 5 的留言，并根据特定地点特定人群定义合理的热度评价指标。

2.2.3 数据整理

根据热度前 5 的留言分别制作热点问题表和热点问题留言明细表。

2.3 问题三分析方法与过程

2.3.1 数据筛选

运用 LDA 主题模型分析部门答复与群众留言之间的相关性，初步筛选出符合条件的答复。下面是详细介绍：

潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）是由 Blei 等人在 2003 年提出的生成式主题模型。生成模型，即认为每一篇文档的每一个分词都是通过“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”。LDA 模型也被称为三层贝叶斯概率模型。包含文档（d）、主题（z）、词（w）三层结构，能够有效对文本进行建模，传统的空间向量模型（VSM）相比，增加了概率的信息。通过 LDA 主题模型，能够挖掘数据集中的潜在主题，进而分析数据集的集中关注点及其相关特征词。

LDA 模型采用词袋模型（Bag Of Words，BOW）将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。

定义词大小为 V，一个 V 维向量（1,0,0...,0,0）表示一个词。由 N 个词构成的答复记为 项分布的参数估计$\theta_{j,s}$：

$$\phi_{e\,i}=(n_{s,i}+\beta_i)/(\sum^{V} n_{s,i}+\beta_i)$$

$$\theta_{i,s}=(n_{i,s}+\alpha_s)/(\sum^{k} n_{i,s}+\alpha_s)$$

其中，ns,i 表示词 wi 在主题 zs 中出现的次数，nj,s 表示文档 dj 中包含主题 zs 的个数。

LDA 主题模型在文本类聚，主题挖掘，相似度计算等方面都有广泛的应用，相对于其他主题模型，其引入了狄利克雷先验知识，因此，模型的泛化能力较强，不易出现拟合现象。其次，它是一种无监督的模式，只需要提供训练文档，它就可以自动训练出各种概率，无需任何人工标注过程，节省大量人力及时间。[3]

2.3.2 数据分析

在答复内容的研究中，即对答复中的潜在主题进行挖掘，答复中的特征词是模型中的可观测变量。一般来说，每则答复中都存在一个中心思想，即主题。如果某个答复是相应留言中的的主题，则这一主题与留言的相关性越高。

首先，为提高主题分析在不同分类中的精确度，本文在语义网络

的基础上，对不同分类下的主题分别进行挖掘分析，从而得到不同分类下部门答复的具体情况。接着，分别统计整个答复语料库中的主题分布情况，对不同分类下，各主题出现的次数从高到低进行排序，根据分析需要，选取排在前若干位的主题作为热门关注点。

本文运用 Python 软件编写 LDA 主题模型的算法，并采用 Gibbs 抽样方法对 LDA 模型的参数进行近似估计。由上文的模型介绍可知，模型中存在 3 个可变量需要确定最佳取值，分别是狄利克雷函数的先验函数α和β、主题个数 k。【4】

3、结果分析

3.1 问题 1 结果分析

3.1.1 对留言内容进行分词并建立模型

通过去重后对留言内容进行 jieba 分词处理，得到各留言分词结果，部分留言内容分词如下图。

得到分词数据后运用 LDA 建模和贝叶斯分类器。

	详情_clean	label
9200	[\n, \n, 艾滋病, 血液, 传播, 传播方式, 日常, 接触, 拥抱, 接吻, 传播...	卫生计生
9201	[\n, \n, 我开, 小型, 餐馆, 食品药品, 监督局, 办个, 卫生, 许可证, 3...	卫生计生
9202	[\n, \n, 尊敬, 领导, 您好, 一名, 妇科病, 患者, 号, 网上, 预约, 中...	卫生计生
9203	[\n, \n, 张, 厅长, 基层, 卫生, 提些, 建议, 抓, 落实, 现有, 政策方...	卫生计生
9204	[\n, \n, 强烈建议, 医院, 手术室, 装上, 高清, 摄像机, 远程, 传输, 卫...	卫生计生
9205	[\n, \n, 夫妻, 农村户口, 女, 岁, 岁, 斤, 治疗, 两年, 一级, 脑瘫,...	卫生计生
9206	[\n, \n, 号, B, 市中心, 医院, 做, 无痛, 人流, 手术, 手术, 怀孕,...	卫生计生
9207	[\n, \n, 再婚, 想, 小孩, 不知, 我省, 二胎, 新, 政策, 先, 怀孕, ...	卫生计生
9208	[\n, \n, K8, 县惊现, 奇葩, 证明, 西地省, K8, 县人, 想, 生二孩,...	卫生计生
9209	[\n, \n, 领导, 你好, 未, 婚生子, 2013, 接受, 处罚, 小孩, 上户,...	卫生计生

3.1.2 运用 F-Score 对分类方法进行评价

综合考虑 Precision 和 Recall 的调和值。

$$F\text{-}Score=\frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall}$$

F−Score=(1+β2)· Precision· Recallβ2· Precision+Recall

（1）Precision(精确率)

关注预测为正样本的数据(可能包含负样本)中,真实正样本的比例

计算公式

$$Precision = \frac{}{TP + FP} \tag{2}$$

（2）Recall(召回率)

关注真实正样本的数据(不包含任何负样本)中,正确预测的比例

计算公式

$$Recall = \frac{}{TP + FN} \tag{3}$$

F-score 中β值的介绍：β是用来平衡 Precision,Recall 在 F-score 计算中的权重,取值情况有以下三种:

如果取 1,表示 Precision 与 Recall 一样重要；如果取小于 1,表示 Precision 比 Recall 重要；如果取大于 1,表示 Recall 比 Precision 重要

一般情况下,β取 1，认为两个指标一样重要.此时 F-score 的计算公式为:

$$F1 - Score = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * (Precision + Recall)} \tag{4}$$

（3）其他考虑

预测模型无非就是两个结果

1.准确预测(不管是正样子预测为正样本,还是负样本预测为负样本)

2.错误预测

在运用 F-Score 后计算的结果,得到

Fscore=(2*0.8571*0.9231)/(0.8571+0.9231)=88.89%，我们发现分类方法合理。

3.2 问题 2 的结果分析

3.2.1 对留言的热度评价指标的分析

对留言详情内容进行分词处理完后再运用 TF-IDF 算法提炼出热度为前 5 的关键词，之后通过 Excel 的筛选功能利用文本筛选将与关键词相关的留言进行汇总。分词结果存为附件 1。

3.2.2 对热度留言的热度概括

针对筛选出来的热度前 5 的留言，根据留言反映的特定地点和特定人群进行提炼总结，概括热点留言评价指标，并建立热点问题表和热点问题留言明细表。

汇总的热度前 5 留言的特定地点或特定人群分别为：

A 市伊景园滨河苑

A 市 A2 区丽发新城小区居民

A 市 A5 区魅力之城小区居民

A 市购房人才

西地省展星投资有限公司设立的 58 车贷 p2p 平台受害人

3.3 问题 3 的结果分析

3.3.1 基于答复的相关性分析

基于语义网络的评论分析初步数据感知后，通过从统计学的角度，对答复内容的特征词出现频率进行量化表示。运用 LDA 主题模型，用以挖掘答复中的更多信息，主题模型在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。直观上来说，传统判断两个文档相似性的方法是通过查看两个文档共同出现的单词的多少，如 TF、TF-IDF 等，这种方法没有考虑到文字背后的语义关联，可能在两个文档共同出现的单词很少甚至没有，但两个文档是具有关联性的，应用语义挖掘，而语义挖掘的有效工具即为主题模型。

如果一篇文档中有多个为主题，则一些特定的可代表不同主题的词语会反复出现，此时，运用主题模型，能够发现文本中使用词语的

规律，并把规律相似的文本联系到一起。以寻求非结构化的文本集中的有用信息。运用主题模型将与特征相关的描述性词语，同相应的特征词语联系起来，从而深入了解部门答复与群众留言的相关。LDA 模型作为其中一种主题模型，属于无监督的生成式主题概率模型。

通过分析群众留言问题与部门留言答复之间线性相关程度的强弱，并用适当的统计指标表示出来的过程成为相关分析。一般用以下几个指标评价相关性：Person 相关系数:要求连续变量的取值服从正态分布，一般用于分析连续性变量之间的关系。Spearman 秩相关系数：一般用于分析不服从正态分布的变量、分类或等级变量之间的关联性。不同的留言类型会有不同的答复，部门留言答复的相关性主要用上述两种系数进行衡量，满足条件的可认为是与留言问题有一定相关性的答复。这类答复是具有一定的质量和价值并值得被群众采纳的。省网络信息办公室将回复内容及时在网上发布，并协助做好留言板舆情动态分析研判工作，加大回复工作的宣传力度。[5]

3.3.2 基于答复的完整性分析

根据相关政策，各部门对收到的群众你留言应当及时答复。一般告知性答复应当在 7 天内完成；结果性答复应当在自受理日起 60 天内完成；情况复杂的，经本行政机关负责人批准，可以是适当延长答复期限，延长答复期限最长不得超过 30 天。留言人同意公开，并有关部门审核同意的，可以公开回复，不适宜公开回复的，有关部门应当在规定的回复期限内，以其他适当方式予以回复。针对不同类型的留言要采取不同的方式加以解决，将网上反映问题与网下解决问题很好的衔接起来，有些留言，是关于咨询、建议类的，应第一时间回复；而有些留言如投诉类的问题立即回复有困难，需要经过实际调查后才能回复的，也应该先解释说明，同时网下加快实际调查速度，尽可能在最短的时间内拿到调查结果及时反馈给网民；对于关系群众切身利益的民生问题，经调查属于合理利益诉求的，应在网下第一时间予以解决。[6]

部门对于群众留言的答复内容应该是详细而完整的，部门应该耐

心详尽的答复群众留言的问题，给他们最完整的答复，这样才不会出现遗漏问题未回答情况。并且各部门要诚心诚意接受群众合理的意见和建议，并及时推进和推动工作。做好接待工作，认真解答有关问题，对反映的问题要及时调查处理，并通过适当的方式告知调查处理结果。各州政府可参照省政府办公厅的做法，指派专人从事市州网民留言回复工作。省政府政务督察室加强指导，适时举办业务培训，召开相关会议交流经验。

3.3.3 基于答复的可解释性分析

对于群众留言的答复内容应是实事求是，符合法律法规、规章或者其他有关规定的支持的，这样以便于群众在收到答复后能够很好的理解部门所传达的意思。并且回复中要注意称呼、方式、措辞、语句的表达，不断提高群众对回复内容的满意程度，听取网民对留言工作的意见和建议。

判断答复的可解释性我们通常用判定系数来衡量回归方程对 y 的解释程度。留言回复内容要严格依据现行的方针政策和工作实际，做到实事求是。对需调查核实的留言，有关部门要坚持到现场了解真实情况，不得采用利益相关单位的答复。对回复内容不符合相关要求的，将退回重办。回复要做到语言朴实、简洁、亲切、既要表现出个性化，更要展示出人性化。

参考文献

[1]黄文妍.社会化网络环境下餐饮创业企业的品牌塑造——基于对茶颜悦色的网络文本分析[J].财富时代,2020(03):140-141.

[2]文彤,刘璐.博物馆文化展演与城市记忆活化传承——基于旅游留言档案的文本分析[J].热带地理,2019,39(02):267-277.

[3]刘磊,柳旭东.外国观众对中国电影在线评论的文本分析——基于IMDb网站的样本调查[J].当代电影,2020(03):144-149.

[4]孟天广,李锋.网络空间的政治互动:公民诉求与政府回应性——基于全国性网络问政平台的大数据分析[J].清华大学学报(哲学社会科学版),2015,30(03):17-29.

[5]苏忠林，翁彬，王亚飞.“互联网 + 政务服务”标准化建设优化路径［J］.学习与实践，2019（7）.

[6]费军，贾慧真.智慧政府视角下政务 APP 提供公共服务平台路径选择［J］.电子政务，2015（9）.