“智慧政务”中的文本挖掘应用

摘要

本文旨在基于收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，通过进行文本分类，文本数据挖掘，建立关于留言内容的一级分类标签模型，挖掘热点问题，并给出答复意见的质量评价方案，从而提升政府的管理水平和施政效率。

针对问题 1，首先对文本数据进行预处理，得到文本向量表示，然后构建 FastText 模型，进行关于留言内容的一级分类模型训练，最后通过 F-Score 对分类方法进行模型评价，使得模型优化。

针对问题 2，首先采用余弦法对留言文本向量之间的相似度计算，识别出在众多留言中的相似留言，然后用 excel 进行文本处理，按照特定地点、特定时间、特定问题将数据归并，即把相似的留言归为同一问题，得出对应表 2 的结果。最后依据所定义的一级和二级热度评价指标，对留言问题进行指标排序，得到对应表 1 的结果。

针对问题 3，从相关性、完整性、可解释性三个角度评判相关部门答复意见的质量，确定基于三个角度的评价标准图表 1，得到一套完善的评价方案，其 1.0，2.0，3.0 分别对应“差、中、好”的答复意见质量。

关键词：Fasttext 模型；热度评价指标；Reddit；答复意见质量评价标准

目录

一、问题分析

1.问题一的分析

问题1要求根据附件2给出的数据，建立关于留言内容的一级标签分类模型，并对分类方法进行评价。通过观察附件2的数据以及分析题目1的要求，可发现问题1存在文本语义交叉、长文本无意义表达太多等难点，因此在文本预处理环节，要注意解决以上难点。在取得文本向量之后，构建关于留言内容的FastText模型，并进行一级标签分类模型训练，最后对模型用F-Score进行评价，使模型得到进一步优化。

2.问题二的分析

问题2要求根据附件3将某一时段内反映特定地点或者特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，即排名前5的热点问题以及相应热点问题对应的留言信息。由此，可确定问题2需要解决的有三点，一是相似问题识别，二是留言问题归类，三是热度评价。其中，热点问题是指某一时段内群众集中反映的某一问题，依据这一定义，得出其关键词是一段时间集中爆发的问题，多人反映同一问题。因此，热度评价指标中的一级指标根据关键词定义，二级指标则进一步定义为时间长度、反对数以及点赞数，最后根据定义的指标，得出评价结果。

3.问题三的分析

问题3要求针对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给出一套评价方案，并尝试实现。首先定义相关性、完整性、可解释性的具体意义，然后确定基于这三个角度的评判标准，使得答复意见的质量可以得到评判。

二、模型假设

1.模型假设

（1）假设所给的文本数据真实，可靠。

三、数据准备

1.剔除重复文本数据

在群众留言的内容中，存在同一用户多次留言同一内容的情况，对于这部分重复文本数据，为了不影响留言分类、归类操作以及相关部门答复工作效率，故删去。

2.提取关键句

群众反映的问题时，为了表述清楚或者表达敬意等，会留言过长文本，在这些长文本中无意义的表达太多，对此，需要考虑转为短文本，提取其关键句为文本挖掘作准备。

四、问题求解

1.一级分类标签模型

1.1 文本预处理

（1）数据清洗

为了保证数据的完整性，使得后续对于文本数据进行分析的结果更为准确，需要通过数据清洗将残缺的数据、错误的数据、重复的数据删除及纠正。对于附件 2 的数据，一些文字，例如“已经”、“请问”、“了解”等等对留言内容的语义没有特殊的影响，在数据清洗过程，需要进行特殊字符处理以及留言时间格式纠正等。用正则表达式对特殊字符进行匹配。

（2）分词和词性，去停用词

采用 jieba 对训练集和测试集文字进行分词，建立去停用词字典，过滤掉一些无意义的词，并将一些必须的词汇、近义词等引入词库，在 jieba 分词完之后设置参数获取词性，进行词性标注。

1.2 文本特征选择

采用信息增益（Intermation Gain,IG）的方法选取最具有分类信息的特征，其计算公式为：

$$IG(t_j) = -\sum_{i=1}^{|c|} P(c_i)\log_2 P(c_j) + P(t_j)\sum_{i=1}^{|c|} P(c_i|t_j)\log_2 P(c_i|t_j) + P(\bar{t}_j)\sum_{i=1}^{|c|} P\left(c_i\middle|\bar{t}_j\right)\log_2 P\left(c_i\middle|\bar{t}_j\right)$$

其中，IG(t)表示特征词 t 的信息增益值，c 表示文本类变量，C 表示文本类的集合，有 C=(c_1,c_2…c_i)。[1]

1.3 文本表示

为了后续更好地进行文本分类、文本聚类等操作，构建词的模型，即词向量。使用 gensim 工具包的 word2vec 训练数据，保存词向量，再通过 TF-IDF 加权平均得到文本向量表示。

1.4 构建模型

构建 FastText 模型。

1.5 模型评价、调优

对建立的一级标签分类模型，用 F-Score 对分类方法进行评价：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

同时进行分词优化、去停用词优化以及特征工程优化，最后优化算法，完善一级标签分类模型。

2. 热点问题挖掘

为了更好地对群众集中反映的热点问题进行深入细致地分析，对于热点问题的挖掘，首先对附件 3 的问题进行识别，识别出相似的留言，然后按照附件 3 所给的留言用户、留言主题、留言时间、反对数、点赞数等维度对群众留言的问题进行归类，最后根据所定义的合理的热度评价指标，全面地给出评价结果。

2.1 问题识别

对于附件 3 的留言内容，同样进行文本分词、词性标注、去停用词等预处理，特征选择之后，通过向量空间模型，将留言问题用向量表示出来，从而把众多留言中的相似留言文本运算变成文本向量运算。采用余弦法[2]进行文本向量之间的相似度计算：

$$Sim(D_1, D_2) = \frac{D_1 \bullet D_2}{\|D_1\| \times \|D_2\|} = \frac{\sum_i (w_{1i} \times w_{2i})}{\sqrt{\sum_i w_{1i}^2 \times \sum_i w_{2i}^2}}$$

其中，$Sim(D_1,D_2)$意为任意两个留言文本D_1和D_2向量之间的相关联系数，指的是两个留言文本在表达上的主题相关联度。

由此，从众多留言中识别出相似的留言。

2.2 问题归类

将识别出的相似留言归为同一问题，并用 excel 进行文本处理，按照特定地点或者人群的数据归并起来，得出结果表 2（热点问题留言明细表）。

2.3 热度评价

所谓热点问题是指某一时段内群众集中反映的某一问题，基于群众留言文本数据的时效性以及指标体系构建的系统性和科学性，故问题二根据附件 3 收集自互联网公开来源的群众问政留言记录，构建合理的热度评价指标。其中，一级指标包括留言时段和相似留言用户数。

一级指标	二级指标
留言时段	时段长度
相似留言用户数	反对数
	点赞数

根据以上一级、二级指标建立 Reddit 计算公式：

$$S = \log_5 z + \frac{yt}{3600}$$

其中，x=点赞数-反对数

$$y = \begin{cases} 1, x > 0 \\ 0, x = 0 \\ -1, x < 0 \end{cases}$$

根据以上热度评价指标的定义和按计算方法计算之后，对指标排名后得出对应结果表 1（热点问题表），其排名前五的热点如下表所示。

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述

1	34	4313	2019/07/02 至 2019/09/01	A 市武广新城广铁集团	A 市伊景园滨河苑车位捆绑销售
2	35	4312	2019/6/25 至 2020/01/26	A 市丽发新城小区	小区附近建搅拌站噪音扰民、环境污染
3	226	4311	2019/01/01 至 2019/07/08	西地省 A 市 A4 区	58 车贷案件进展情况
4	1347	4310	2019/05/05 至 2019/9/19	A5 区五矿万境 K9 县	A5 区五矿万境 K9 县交房后仍存在诸多问题
5	156	4309	2019/01/06 至 2019/05/22	A1 区辉煌国际城二期商铺	辉煌国际城二期物业提供虚假场地证明，居民楼下商铺非法取证

表 1　排名前五热度问题

3.答复意见的评价

3.1 评价方案

（1）评价目的

为了切实保证群众的留言得到相关部门比较全面的，解决到疑问的答复意见，故对答复意见的质量进行评价，使得一方面可以积极引导群众提出疑问和建议，有效监督政务工作；另一方面可以使得相关部门有效解决群众和吸取群众建议，并提升其管理水平和施政效率。

（2）评价角度

相关性角度：相关部门是否对留言的核心内容进行核实并解答，答复意见的内容是否与问题相关。

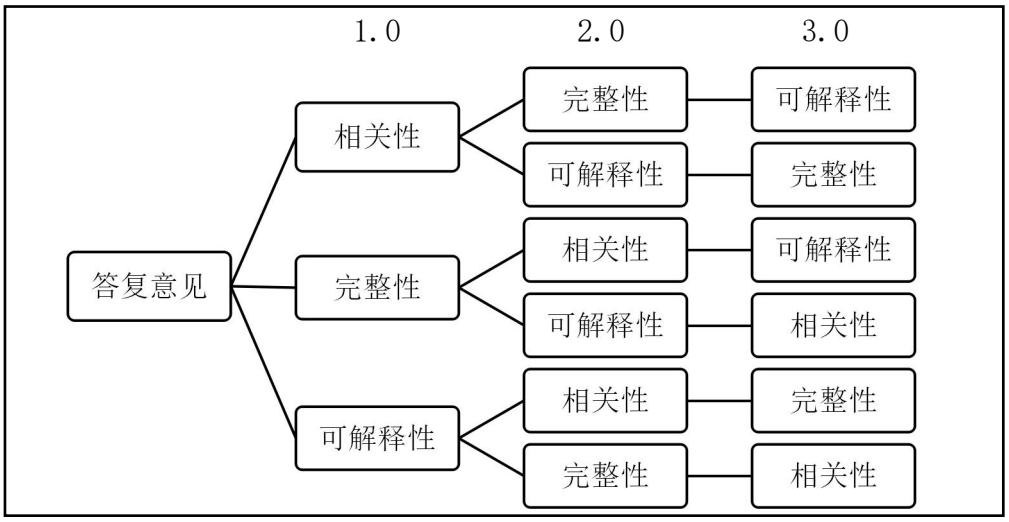
完整性角度：当政府知悉问题来源时，要从问题根源出发，要对问题进行回应，给出问题的结果，答复意见是否符合某种规范。

可解释性角度：答复意见中内容的相关解释，涉及的一些专业名词等是否解释清楚。

（3）评价标准

一个质量好的意见回复应该是相关性、完整性、可解释性，三者缺一不可。若是存在缺失情况，则以下面图表的标准对答复意见的质量进行评价。其中 1.0 为“质量差的答复意见”，2.0 为“质量中等的答复意见”，3.0 为“质量好的

答复意见”。其相关关系如下表：



五、参考文献

[1] 李海瑞. 基于信息增益和信息熵的特征词权重计算研究[D].重庆大学,2012.

[2] 张金鹏. 基于语义的文本相似度算法研究及应用[D].重庆理工大学,2014.