“智慧政务”中的文本信息挖掘的探究

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。而在这种艰难的大背景下，除了提高对民意的服务质量和增加人手力度外，了解更多群众的心声对于政府来说也变得越来越重要，其中非常重要的方式就是对群众的文本留言数据和部门的答复意见进行内在信息的数据挖掘分析。而得到这些信息，也会有利于政府及时地了解民意，更好地为人民服务。本文将基于数据挖掘技术对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进行内在信息的挖掘和分析。

对于问题 1，首先对附件 2 给出的数据的一级标签进行分类和统计并且将分类转换为 id 值，然后根据留言详情和留言主题针对一些专用名词进行提取，之后对留言详情和主题等文本信息进行去空去重，加入专用名词词库进行中文分词，停用词过滤等数据预处理，然后将留言主题数据进行向量化处理之后进行 LSTM 的建模工作，将训练集和测试集数据进行拆分，并定义一个 LSTM 的序列模型（Sequential 模型）来完成数据训练，最后使用 F-score 对模型的分类效果进行评价。

对于问题 2，首先还是对附件 3 进行文本分词，停用词过滤等数据预处理，利用 gensim 来进行文本相似度计算，并通过比较相似度完成分类工作，最后结合各版本热度算法综合数据参考了点赞数，反对数，相似留言数量和时间区域的因素，总结出热度值计算公式完成热度排名工作。

对于问题三，对附件 4 进行文本分词，停用词过滤等数据预处理，结合相关数据进行质量评价指标的定义。最后通过专家权重确立方法得到指标的权重数值，再通过求和得出评价分数。

关键词：中文分词；文本向量化；LSTM 模型；gensim 模块；re 模块

目录

1.挖掘目标

本次数据挖掘建模是利用附件给出了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的文本数据，利用 jieba 分词，LSTM 模型，gensim 模块，re 模块以及相关指标数值计算公式达到以下三个目标。

(1)利用 jieba 分词和文本向量化的方法，通过建立 LSTM 模型和定义一个序列模型（Sequential 模型）对附件 2 的数据进行文本挖掘，完成留言的一级标签分类工作并结合 F-score 评价模型效果。

(2)根据附件 3 给出的数据，将某一时段内反映特定地点或特定人群问题的留言进行归类，并且定义出合理的热度评价指标，给出评价结果。

(3)根据附件 4 中的答复意见质量，构建一套评价方案，并从答复意见的相关性、完整性、可解释性、时效性等角度完成实现。

2.问题分析与求解

2.1 问题 1 分析与求解

2.1.1 问题求解的流程

先对数据进行预处理，对附件 2 标签数值化处理，文本去除重复项及空行、中文文本分词、停用词过滤，以便后续分析；后进行模型构建，文本向量化，构建 LSTM 模型，进行模型训练；最后模型预测与分析，使用 F-score 对模型进行评价；具体流程如图 2.1 所示。

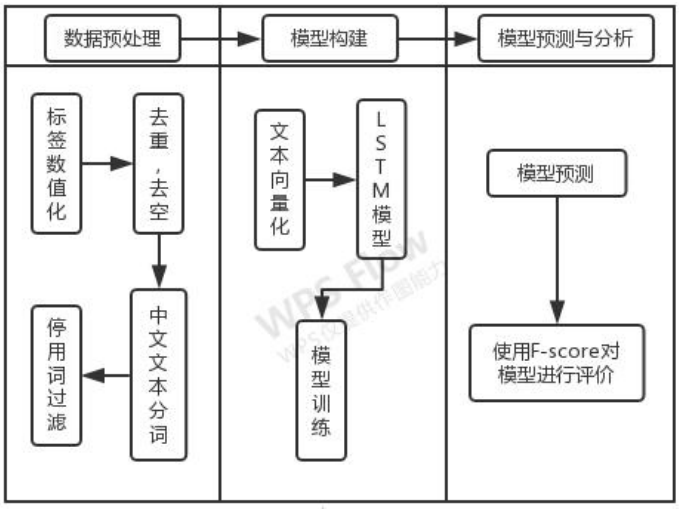

图 2.1 问题 1 求解流程

2.1.2 数据预处理

（1）数据描述

通过观察所给数据，可以发现附件 2 的数据量比较大(共 9210 条记录)，且附件 1 提供了包含所有的附件 2 分类的级别信息。由于留言主题大多为文本格式，需要将其量化成数值形式才能对其进行分析。而附件 2 的留言主题的内容中也有空行以及重复的情况，如果不做处理会对后续分析造成影响，并且留言主题的内容存在着大量垃圾内容的特征，如果把这些数据也引入进行分词、词频统计乃至文本分类模型等，则必然会对分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。

（2）文本的预处理

首先对附件 2 给出的数据进行分析，附件 2 的数据的关键字段为“一级标签”，“留言主题”，先对“一级标签”进行分类和统计，由于中文不好进行数据处理，因此将“一级标签”转换为 id 值，方便以后的分类模型训练。之后对留言主题进行观察分析，提取专用名词词库，保存为 newdic1.txt，之后下载停用词词库命名为 stopword1.txt。结合两个文件对留言主题进行中文分词，去重去空以及停用词过滤的数据预处理，并定义预处理函数名为 data_process。

2.1.3 LSTM 建模

完成预处理工作后，开始选择分类模型。传统的分类模型需要考虑到文本的特征选择，而深度学习可以自动学习一些特征，这样就不需要人工去进行文本特征选择，从一定程度上节约了时间和人力成本。因此模型上决定选择 LSTM 模型。而 Keras 搭建 LSTM.Keras 封装了一些优秀的深度学习框架的底层实现，可以很方便的进行模型搭建。先开始 LSTM 建模的预备工作：设置最频繁使用的词数和每条留言主题最大的词语数，将留言主题信息进行向量化处理,将每条留言主题转换成一个整数序列的向量。由于 keras 只能接受长度相同的序列输入，因此需要进行序列填充，之后定义序列模型 Sequential，第一层（嵌入层）Embedding 负责预训练词向量，SpatialDropout1D 层可以在训练中每次更新时防止过拟合，LSTM 层可以存储学习记忆，输出层负责输出向量。模型建立完毕后即可开始训

练模型。将训练集和测试集进行拆分，通过不断比较训练集的效果值和测试集的效果值来提高模型的准确率。

2.1.4 模型评价

对于分类模型的评估方法使用 F 分数(F-Score)。F-score 需要四个参数，分别是 TP，FP，TN，FN，如图 2.2 所示。

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

图 2.2 F-Score 方法

接下来结合相关数值进行 precision 和 recall 的计算。

(1)Accuracy 的计算公式：$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$

(2)precision 的计算公式：$Precision = \frac{TP}{TP+FP}$

(3)recall 的计算公式：$Recall = \frac{TP}{TP+FN}$

Accuracy 为精确度，precision 为准确率，又称为查准率，表示正确预测为正样本的概率，recall 为召回率，又称为查全率，表示最后被正确预测为正样本的概率。最后的评价公式为:

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

Pi 为第 i 类的查准率，Ri 为第 i 类的查全率。最后可得到相关数据结果，如图 2.3 所示。

模型预测分析报告:

	precision	recall	f1-score	support
城乡建设	0.85	0.80	0.83	438
环境保护	0.78	0.89	0.83	150
交通运输	0.75	0.86	0.80	109
教育文体	0.85	0.86	0.86	288
劳动和社会保障	0.88	0.86	0.87	394
商贸旅游	0.72	0.74	0.73	229
卫生计生	0.82	0.76	0.79	173
accuracy			0.82	1781
macro avg	0.81	0.82	0.81	1781

图 2.3 数据结果

2.2 问题 2 分析与求解

2.2.1 问题求解的流程

先对数据进行预处理，对附件 3 文本分词、停用词过滤，整理预处理过的数据；然后进行数据分析，将数据文本向量，构建 TfidfModel 模型，通过模型相似度计算，结合相似度进行热点点指数计算；最后进行数据筛选，根据热度公式进行热度值计算和排序，提取热点指数前 5 的数据，生成 2 个 excel 文档；具体流程如图 2.4 所示。

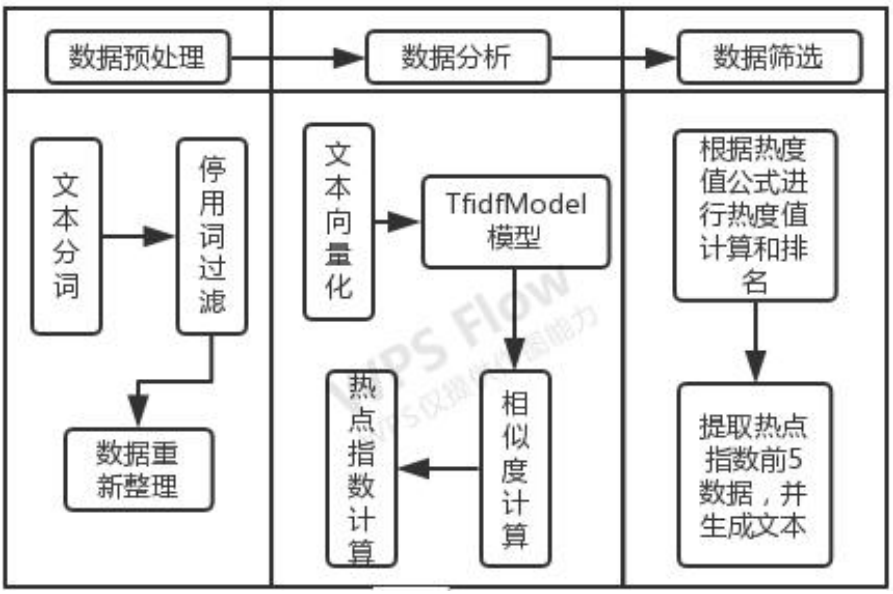

图 2.4 问题 2 求解流程

2.2.2 数据预处理

首先对附件 3 给出的数据进行分析，附件 3 的数据的关键字段为“留言主题”，“留言详情”，结合专用词汇和停用词的两个文件对“留言主题”进行中文分词，去重去空以及停用词过滤的数据预处理，并对数据进行重组。

2.2.3 利用 gensim 进行文本相似度计算

gensim 是一个 python 的自然语言处理库，里面包括 TF-IDF，LSA，LDA，和 word2vec 在内的多种主题模型算法。它也有它的四个基本概念：

(1)语料（Corpus）：文档集的表现形式。

(2)向量（Vector）：由一组文本特征构成的列表。

(3)稀疏向量（SparseVector）：通常，我们可以略去向量中多余的 0 元素。此时，向量中的每一个元素是一个(key, value)的元组。

(4)模型（Model）：可以将文本的一种向量表达变换为另一种向量表达，在 models 中可以使用多种模型。例如 tf-idf 模型，lsi 模型，lda 模型等。

因此首先采用 tf-idf 模型来进行词频，通过词频将相关文本信息向量化，最后采用余弦相似度算法实现相似度计算。TF-IDF 模型有两个指标：

词频(TF)：指的是某词语在该文本中出现的次数。

$$TF = \frac{\text{在某一类词条}\omega\text{出现的次数}}{\text{该类中所有的词条数目}}$$

逆向文件频率（IDF）：如果包含某词语的文本越少，IDF 越大。

$IDF = \log(\frac{\text{语料库的文档数}}{\text{包含词条}\omega\text{的文档数}+1})$，分母之所以要加 1，是为避免分母为 0。

最后 TF-IDF 计算公式为 $\mathrm{TF}\text{-}\mathrm{IDF} = \mathrm{TF} * \mathrm{IDF}$。

文本相似度计算将采用余弦相似度算法，因此需要将文本信息根据 TF-IDF 值进行文本向量化处理。由公式：

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

即可计算出文本相似度的值。

2.2.4 计算热度指数并排序

首先先根据相似度标准将相似留言进行归类，考虑相似文本的特征，先定标准为相似度大于 0.5 的为相似留言记录。标准会根据之后得到的数据结果进行微调。之后结合各版本的热度计算公式，最终将热度公式决定为：

$$H_T = \log(\sum_{i=1}^{n}((n\_W_j)^2 + n) * w_j)$$

n 代表相似文本同类的记录数，n_Wj 代表认同值（点赞数和反对数的差值）。

考虑留言的热度还与文本相似度以及时间因素有关，这里定义热度时间范围为 14 天。根据这些因素，总结出 wj 的计算公式：

$$wj = \frac{\sigma}{\frac{tj - t0}{14} + 1} + \delta$$

σ 和 δ 作为调节因子，经过测试，得到 σ 为 0.6， δ 为 0.4，t0 为相似文本最先留言时间，tj 为第 j 篇相似文本留言时间。

根据热度值计算公式，将数据进行排名，将排名前五的热点问题提取出来，最后将题目所需结果数据放在热点问题表.xls 和热点问题留言明细表.xls。结果如图 2.5 和 2.6 所示。

热度排名	问题ID	热度指数	时间范围	地点/人物	问题描述
1	1	14.94273	2019-04-1	A市金毛湾	反映A市金毛湾配套入学的问题
2	2	14.86575	2019-05-0	A市五矿万	A市五矿万境K9县房屋出现质量问题
3	3	14.00723	2019-02-2	A市A4区5	请书记关注A市A4区58车贷案
4	4	13.34407	2019-02-2	A市58车贷	严惩A市58车贷特大集资诈骗案保护伞
5	5	13.01157	2019-09-0	A4区绿地	A4区绿地海外滩小区距长赣高铁最近只有30米不到，合理吗？

图 2.5 热点问题

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	223297	A0008752	反映A市金	2019/4/11	书记先生：	1762	5
2	208069	A0009443	A5区五矿	2019/5/5 1	本人是A5[	2	0
2	208636	A0007717	A市A5区汇	2019/8/19	我是A市A5	2097	0
2	262599	A0001004	A市五矿万	2019/9/19	我是西地省	0	0
3	194343	A0001061	承办A市58	2019/3/1 2	胡书记：您	733	0
3	220711	A0003168	请书记关注	2019/2/21	尊敬的胡书	821	0
4	217032	A0005654	严惩A市58	2019/2/25	胡市长：您	790	0
5	263672	A0004144	A4区绿地海	2019/9/5 1	您好，近日	669	0

图 2.6 热点问题留言明细

2.3 问题 3 分析与求解

2.3.1 问题求解的流程

先对附件 4 数据进行预处理，文本去除重复项及空行、中文文本分词、停用词过滤；然后进行数据分析，将文本数据向量化，构建 TfidfModel 模型，通过模型相似度计算，可解释性计算、完整性计算、时效性计算，通过权重进行质量评价计算；具体流程如图 2.7 所示。

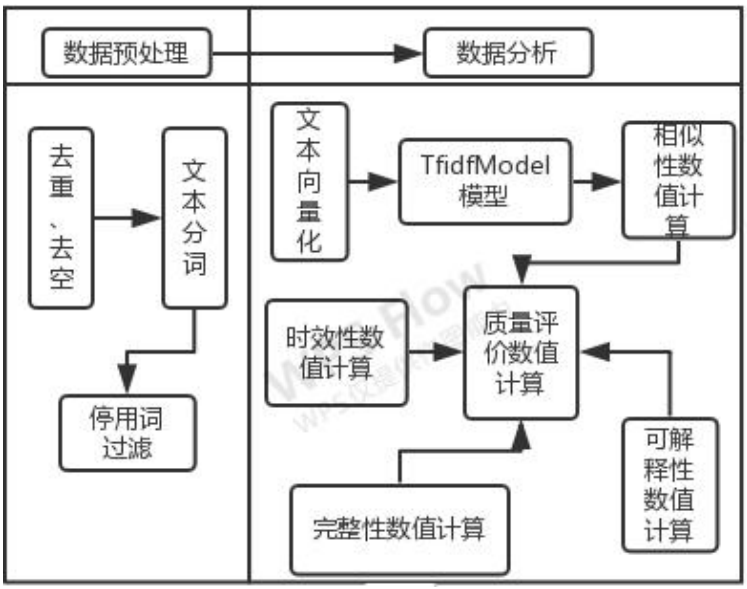

图 2.7 问题 3 求解流程

2.3.2 数据预处理

首先对附件 4 给出的数据进行分析，附件 3 的数据的关键字段为“留言主题”，“留言详情”，“答复意见”结合专用词汇和停用词的两个文件对“留言详情”和“答复意见”进行中文分词，去重去空以及停用词过滤的数据预处理，并对数据进行重组。

2.3.3 各个质量评价指标的计算

(1)相关度

衡量答复意见和留言详情的相关程度，看两者内容是否相关。计算方法步骤如下：

①首先还是将文本向量化。

②根据 TF-IDF 模型计算出文本的 TF-IDF，并进行相似度计算。

③将得到的相似度数值作为计算相关度的标准值。

(2)可解释性

衡量答复意见中可解释性句子成分比重。看是否有相关依据回复民众留言。计算方法步骤如下：

①提取可解释性句子的文本信息，用 re 模块可以实现。

②计算可解释性句子的字数占答复意见总字数的比例。

③确立比例值和可解释性的关系，大于 0.5 即判断可解释性高。

④将计算出的比例值作为计算可解释性的标准值。

(3)完整性

判断答复意见是否完整。通过观察附件 4 的数据特点，最后得出两个依据，有无感谢话语，结尾有无时间信息。两者都有完整度高，记为 1，只有感谢话语为半完整，记为 0.5，两者都无记完整度为 0。

(4)时效性

衡量答复意见的及时回复程度。计算步骤如下：

①通过公式: $s = \frac{1}{e^{t2 - t1}}$ ，其中 t1 为留言时间，t2 为答复时间。

②将公式得到的数值作为时效性的标准值。

(5)最终质量评分

综合相关度，可解释性，完整性和时效性的指标数值，由专家权重确立方法，确定相关度基本分为 30 分，可解释性为 40 分，完整性为 10 分，时效性为 20 分。最后计算公式为: 30*相关度+40*可解释性+10*完整性+20*时效性。最后根据最终质量评分公式计算各答复意见的分数，如图 2.8 所示。

	留言详情	答复意见	相似度	可解释性	完整性	时效性	意见质量排
2	2019 年 4 月 以来 位于 A 市	现将 网友 在 平台	0.24258101	0.936123	10	3.05902E-07	54.72237
3	潇楚 南路 从 2018 年 开始 修	网友 A00023583：	0.034068212	0.977049	10	8.31529E-07	50.10403
4	地处 省会 A 市 民营 幼儿园	市民 同志： 你好	0.310558766	0.945378	5	8.31529E-07	52.13191
5	尊敬 书记： 您好 我 研究生	网友 A000110735：	0.248590559	0.975806	5	8.31529E-07	51.48999
6	建议 将 白竹坡 路口 更名 为	网友 A0009233 您好	0.668207228	0.959627	10	3.05902E-07	68.43132
7	欢迎 领导 来 A 市 泥泞不堪	网友 A00077538：	0.073121324	0.969828	10	3.44248E-14	50.98674
8	尊敬 胡书记： 您好 过去 在	网友 A000100804：	0.187188134	0.969388	10	4.24835E-18	54.39115
9	我 做 为 一东澜湾 社区 居民	网友 UU00812 您好	0.177043438	0.982372	10	6.9144E-13	54.60617
10	我 是 美麓 阳光 a 栋 803 业主	网友 UU008792 您好	0.159412548	0.977228	10	1.12535E-07	53.87149
11	胡书记 好 根据 规划 洋湖 新	网友 UU008687 您好	0.228721604	0.948661	5	1.12535E-07	49.80808
12	我 家住 在 A 市 A2 区大托 街	网友 UU0082204 您	0.102948666	0.97546	10	3.97545E-31	52.10686
13	胡书记： 您好 我 想 请问 一	网友 UU008829 您好	0.408083558	0.973068	10	9.35762E-14	61.16522
14	尊敬 书记： 我 是 一名 居住	网友 UU00877 您好	0.17453751	0.920863	10	1.12535E-07	52.07066
15	尊敬 领导： 我们 是 贵省 A	网友 UU0081480 您	0.026122713	0.881188	10	0.006737947	46.16597
16	建议 增开 A 市 261 路 公交车	网友 UU0081227 您	0.129622698	0.942857	10	4.13994E-08	51.60297

图 2.8 答复意见的评分

3. 结果分析

3.1 F-score 分析结果

从上文得到的数据结果来看，各分类得到的数据结果的 F 分数值不高，推测面对 9200 条留言记录，专用词汇提取效率不高，容易造成分词歧义的现象出现，需要更加有效率的提取专用词汇的模型。

3.2 热点问题分析结果

从上文得到的两个数据文件来看，热度指数主要取决于点赞数，不过考虑到点赞确实能体现出大多数人的心声，因此将点赞数作为热点问题的主要评判标准是合理的。

3.3 质量评分结果分析

从意见质量评分的数值上看，结合了各个指标数值，发现得到的分数还是合理的。但是相关度的值容易发生偏差， 猜测是因为留言详情有一些文本信息过短，意见回复过长，导致相关度下调。需要考虑更多的因素去确立相关度的指标数值。

4. 结论

对各类社情民意相关的留言文本进行分析研究，可以很大程度促进了解群众心声的效率。这对于相关政府的决策也有重大意义。不过有效的挖掘留言信息是一个大难题。但通过对民众留言信息的不断挖掘和处理，可以让政府及时地了解民意，更好地为人民服务。对标签的文本分类模型的探索也可以极大的减少人工分类带来的时间成分，提高了民众心声的反馈效率。但是模型的探索并不容易，本文中的模型，对于专用词汇的提取就是一大难关，如何有效提取出文本的专用词汇，减少分词歧义现象的出现是现在我们需要攻克的一大难题。而对热点问题的提取，有利于让政府及时的了解关于当时民众的热点事件，从而优先去解决。计算答复意见的质量分数，也是更好的发现相关部门的内在问题，及时处理，有利于更好的解决民生问题。

参考文献

[1]王美方，刘培玉，朱振方．基于 TFIDF 的特征选择方法[J]．计算机工程与设计，2007，28(23):5795-5796.

[2]胡学钢，董学春，谢飞．基于词向量空间模型的中文文本分类方法[J]．合肥工业大学学报:自然科学版，2007，30(10):1261-1264.

[3]杨震，段立娟，赖英旭.基于字符串相似性聚类的网络短文本舆情热点发现技术[J].北京工业大学学报,2010,36(05):669-673.

[4]聂卉.基于内容分析的用户评论质量的评价与预测[J],图书情报工作,2014,58(13):83-89.

[5]-派神-.基于 LSTM 中文文本多分类实战[EB]/[OL].https://blog.csdn.net/weixin_42608414/article/details/89856566