“智慧政务”中的文本挖掘应用

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。

针对问题 1 的一级标签分类模型，我们利用 ***Doc2Vec*** 以及 ***Logistic*** **回归**，利用 ***python*** 编写程序，得到了相关的一级标签分类模型。根据 F-Score 评价方法，得出构建的一级分类标签模型的评分为 $F_1 = 70.38$ 。

针对问题 2 的热点问题、热点问题的指标，我们利用了 ***LDA*** **模型**对留言进行了分类处理，对留言主题相似的留言进行聚类处理；我们构建了相应的**指标评价体系**：以留言主题、点赞数和反对数作为一级指标，构建指标评价体系。最终我们得出前五个热点问题的热度指数分别为：0.127、0.065、0.063、0.060、0.055.

针对问题三，我们制定了一套**评价方案**来评价答复意见的质量，以相关性、完整性、可解释性、时效性四个方面来评价答复意见，并将答复意见的质量分为优、中、差三个级别，在提供中文本数据中，这三个级别的答复意见数理分别为 1270、1071、475。

关键词：*Doc2Vec*　***Logistic*** 回归　***python***　***LDA*** 模型　指标评价体系　评价方案

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on human to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

For the first level label classification model of question 1, we use doc2vec and logistic regression, and use Python to write programs to get the related first level label classification model. According to the F-score evaluation method, the score of the first level classification label model is.

In view of the hot issues and indicators of the hot issues in question 2, we use LDA model to classify the messages and cluster the messages with similar topics; we build the corresponding index evaluation system: take the topic of the message, the number of likes and objections as the first level indicators, and build the index evaluation system. Finally, we get the heat index of the first five hot issues: 0.127, 0.065, 0.063, 0.060, 0.055.

In response to question 3, we have formulated a set of evaluation scheme to evaluate the quality of the reply, and evaluated the reply in four aspects: relevance, integrity, interpretability and timeliness. The quality of the reply is divided into three levels: excellent, medium and poor. In the provision of Chinese data, the three levels of reply are 1270, 1071 and 475.

Key words: doc2vec logistic regression Python LDA model index evaluation system evaluation scheme

目录

一、问题重述

1.1 问题背景

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。

1.2 要解决的问题

- 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。并使用 F-Score 对分类方法进行评价。
- 根据附件 3 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，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题。
- 针对附件 4 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。

二、符号说明

符　号	意　义
x_w	当前内部节点的词向量
θ	训练样本中得到的参数
θ_j	第 j 个非叶子节点对应的向量
P_i	第 i 类的查准率
R_i	第 i 类的查全率
θ_d	文档的主题分布
β_k	主题中的词分布

$W_{d,n}$	文档 d 中的第 n 个词
$z_{d,n}$	文档 d 中的第 n 个词的主题

三、问题分析

基于本文需要解决的三个问题，本文从如下角度进行分析，并依此构建策略模型：

3.1 问题一的分析

对于问题 1 的据，我们利用 *Doc2Vec* 进行机器学习分类，具体流程图如下：一级标签分类模型，根据附件 1 所提供的标签体系和附件 2 所给出的数据，我们利用 *Doc2Vec* 进行机器学习分类，具体流程图如下：

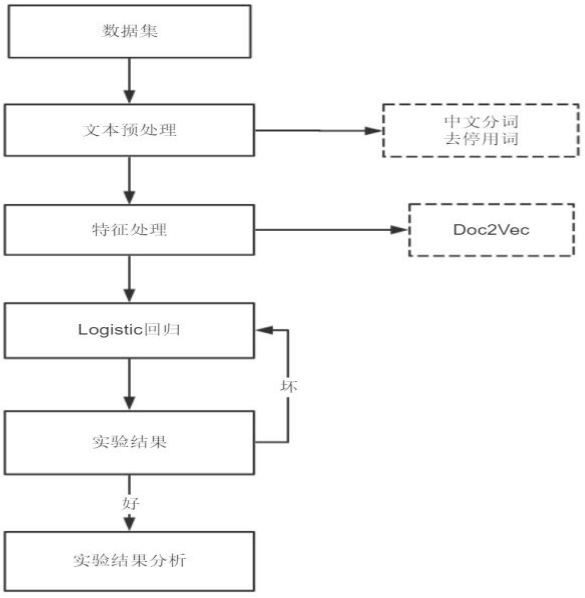

图 1. 模型流程图

3.2 问题二的分析

热点问题的存在反映了当前群众希望迫切解决的问题，及时找出热点问题并解决热点问题，提升政府的管理水平和施政效率。对于找出热点问题，我们利用 *LDA* 模型，找出在同一时间段群众集中反映的问题，定义合适的指标，确定相应指标的权重，构建一个热点问题指标体系。

3.3 问题三的分析

对于问题 3 的答复意见评价方案，我们对留言内容进行三级标签分类处理，找出其标签，根据答复意见的主题，判断答复意见与留言内容的相关性，并且根据答复意见的完整性、可解释行等角度，构建一个完整的评价方案。

四、模型的建立与求解

4.1 文本的预处理

文本预处理是文本分类中至关重要的一步，中文分词的结果以及停用词的存在都会直接影响特征提取的结果，进而影响文本分类的效果。

4.1.1 中文分词

中文分词技术是自然语言处理领域中很多关键技术的基础，在进行文本分类时具有很强的实用性。现如今的中文分词方法有很多，例如字典分词方法、理解分词方法和统计分词方法。在对附件 2 的文本数据进行中方分词处理时，调用了 python 的 jieba 库，对群众留言内容进行了分词处理。

4.1.2 引入停用词

在对群众留言内容进行中文分词处理后，由于一起停用词的存在，会导致处理后的数据维度过高，不利于文本分类处理。在此引入停用词去除分词结果中的无效词，例如：“的”、“啊”、“吧”等。

4.2 问题一：建立关于留言内容的一级标签分类模型

文本经过预处理后，如果直接使用这些数据作为特征向量，会存在维度过高的情况，针对这一情况，我们利用 *Doc2Vec* [1]来构建词向量模型。

Doc2Vec 是一个无监督框架，学习文本段落的连续分布向量表示，文本可以是可变长度的。通过 *Doc2Vec* 将文本进行数字化表示，构建词向量模式，度量词与词之间的相似度。 *Doc2Vec* 有两种模型： *CBOW* （连续词袋模型）[2]和 $Skip-gram$ 模型。在本文中，我们采用了 *CBOW* （连续词袋模型）。

CBOW 模型通过给定上下文 $context(w_t)$ 来预测当前词 w_t ，其模型图如下所示：

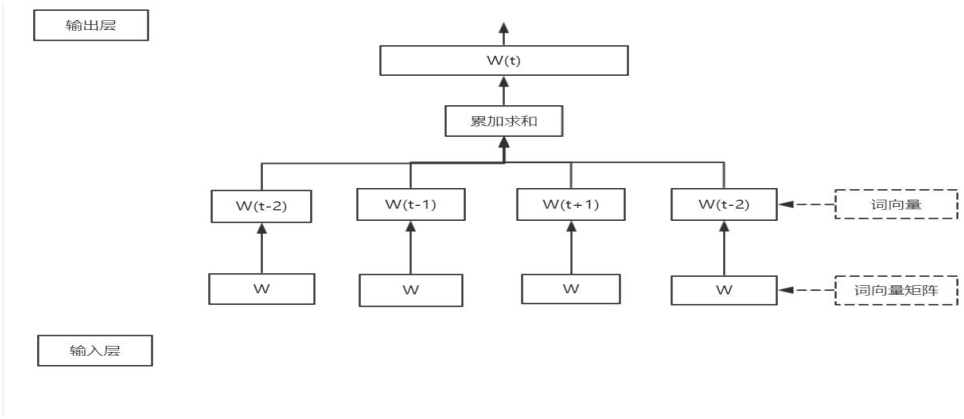

图 2：*CBOW* 模型示意图

如上图所示，我们利用 *Doc2Vec* 的 *CBOW* 模型，对我们经过文本预处理的留言内容进行构建词向量模型。*CBOW* 模型只有输入层，投影层、输出层，而且其中间层对所有的输入词向量进行累加求和。其具体的训练过程为：对给定文本中待预测的单词位置，设定窗口获取其前 k 个单词和后 k 个单词，即 $context(w_t)$ 的规模大小为 $2k$ ，再将这 $2k$ 个单词按照设定好的维度进行随机初始化，然后进行简单的向量求和，最后交由输出层进行归一化运算从而获得预测位置上出现单词 w_t 的概率。模型的训练目标是：

$$L = \sum_{W_t \in C} \log p(w_t \mid w_{t-k}, w_{t-k+1}, \cdots w_{t+k-1}, w_{t+k}) \tag{1}$$

式中的 c 为 w_t 所在的语料库，在模型训练过程中利用神经网络的反向传播算法求出模型参数的同时，也更新了输入单词的词向量。

Doc2Vec 虽然去除了隐藏层的计算，大大缩减了模型计算的复杂度，但是在实践训练过程中，模型需要把词汇表中的所有词汇进行相似度计算和归一化，其计算复杂度仍然较高。因此，在 *CBOW* 模型计算的过程中，我们引进 *logistic* 回归[3]，将复杂的归一化概率分解为一组条件概率的乘积。在输出层的哈夫曼二叉树中，从根节点开始，左孩子节点为负类（编码为 1），右孩子节点为正类（编码为 0）。根据 *logistic* 回归，任意一个节点被分为正类的概率为[4]：

$$\sigma(x_w{}^T\theta) = \frac{1}{1+e^{-x_w{}^T\theta}} \tag{2}$$

则，节点被分为负类的概率为：

$$1-\sigma(x_w{}^T\theta) \tag{3}$$

式中 x_w 为当前内部节点的词向量，而 θ 是从训练样本中得到的参数。

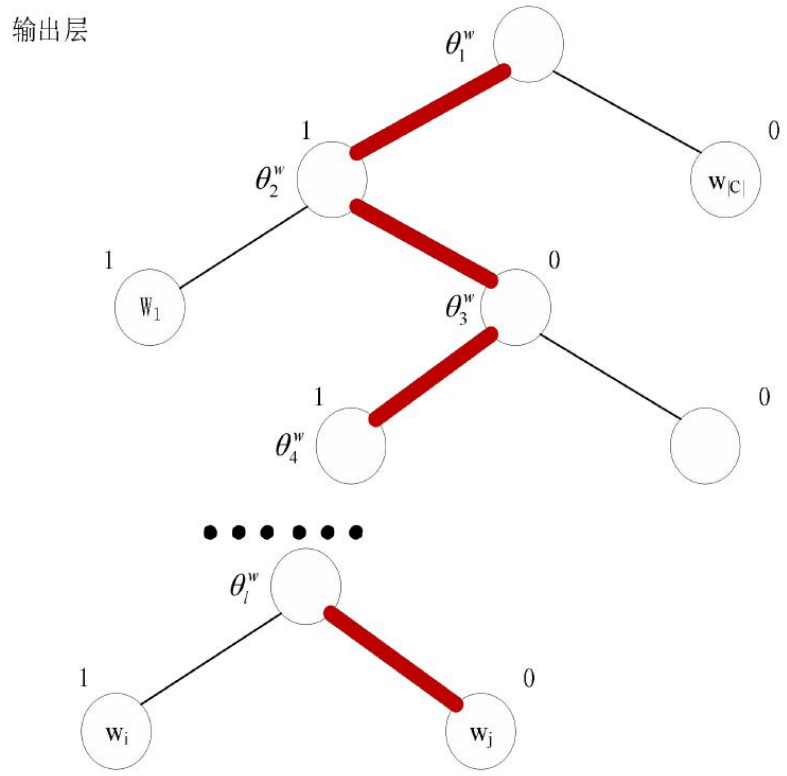

图 3：基于 *CBOW* 模型输出层的哈夫曼二叉树

由于 *CBOW* 模型中输出层的哈夫曼二叉树，叶子节点与语料词典中 C 中的词一一对应。所以，对每个单词 $w \in C$，在二叉树中一定有唯一一条从根节点指向词 w 所指代叶子节点的路径。设改路径中非叶子节点有 $l-1$，相应地产生 $l-1$ 个节点概率，那么 w 的预测概率就是将这 $l-1$ 个节点概率相乘而得。即：

$$p(w_t \mid context(w_t)) = \prod_{j=2}^{l} p(d_j \mid x_w, \theta_{j-1}) \tag{4}$$

$$p(d_j \mid x_w, \theta_{j-1}) = [\sigma(x_w{}^T \theta_{j-1}]^{1-d_j} \bullet [1-\sigma(x_w{}^T \theta_{j-1})]^{d_j} \tag{5}$$

其中 θ_j 表示第 j 个非叶子节点对应的向量，d_j 表示从根节点到叶子节点的路径中的第 j 个非叶子节点的编码，且

$$d_j \in \{0,1\} \tag{6}$$

通过引进 *logistic* 回归，模型的训练目标已经转化为：

$$L = \sum_{W_t \in C} \log \prod_{j=2}^{l} \{[\sigma(x_w{}^T \theta_{j-1})]^{1-d_j} \bullet [1-\sigma(x_w{}^T \theta_{j-1})]\} \tag{7}$$

根据 *CBOW* 模型，我们用 python 编写相应的代码，运行出模型结果。根据附件 2 提供的数据，我们运行的模型结果为：

表 1:运行结果明细表

	查全率	查准率
交通运输	69.71	60.56
劳动和社会保障	73.20	82.49
卫生计生	74.47	66.66
商贸旅游	63.81	59.38
城乡建设	69.59	64.17
教育文体	83.07	81.76
环境保护	67.00	71.66

根据 $F-Score$ 评价方法，即：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{8}$$

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。通过计算我们最终得出计算结果：

$$F_1 = 70.38 \tag{9}$$

通过表 1，我们可以看出，我们的分类模型在一些一级标签可以有很好的查全率和查准率，这说明我们的分类模型在机器学习的过程中有优异的表现，可以用于劳动与卫生保障、教育文体这些相关的群众留言问题上。但是在一些一级标签上没有体现出来，主要是在商贸旅游这个一级标签上，在这个一级标签上，我们的分类模型没有得到很好的表现。

4.3 问题二：对留言进行归类，定义合理的热度评价指标

热点问题的存在反映了当前群众的迫切需求，及时发现并解决热点问题，能够提高政府的施政效率。在群众留言中找出大部分群众强烈反映的留言问题，并且根据他们的描述划分给相应负责的部门进行处理。

LDA 模型[5]是一种概率生成模型，能够对语料库进行建模，达到文档降维的效果。它通过构建“文档—主题—词”三层的贝叶斯结构，将文档集中，且将文档的主题以概率分布的形式给出，从而根据文档主题进行主题分类。

假设 α 是每篇文档下主题的迪利克雷先验输入，η 是每个主题下的特征词的迪利克雷先验输入。训练语料中有 D 篇文档，K 个主题，文档中词的总个数为 N 个，θ_d 是文档的主题分布，β_k 是主题中的词分布，$W_{d,n}$ 是文档 d 中的第 n 个词，

$z_{d,n}$是文档d中的第n个词的主题。所以，LDA 主题的联合概率公式为[6]：

$$p(\theta_d, z_{d,n}, W_{d,n}, \beta_k \mid \alpha, \eta) = \prod_{k=1}^{K} p(\beta_k \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{d=1}^{N} p(z_{d,n} \mid \theta_d) p(W_{d,n} \mid z_{d,n}, \beta_k) \quad (10)$$

LDA 模型示意图如下：

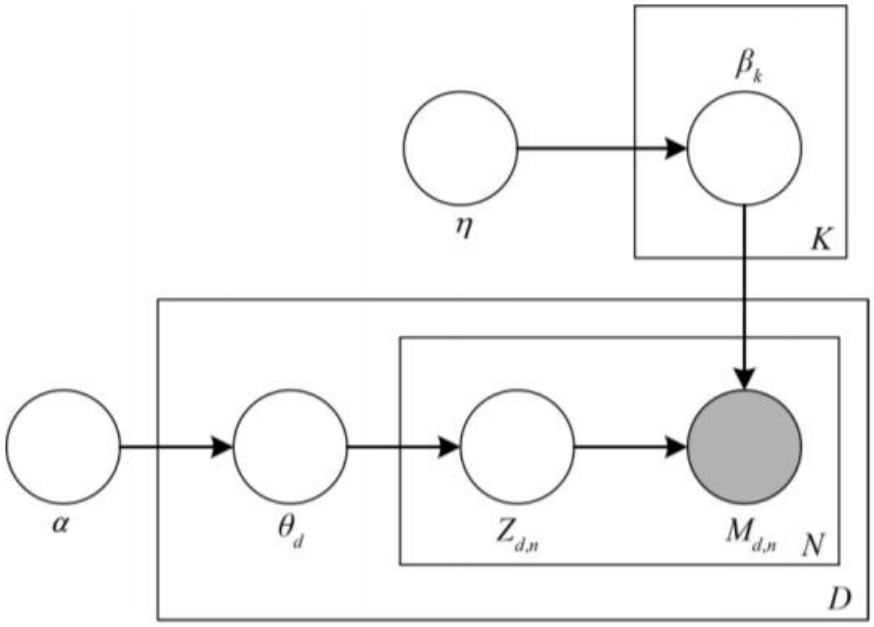

图 4. *LDA* 模型示意图

在*LDA* 主题模型中，文档主题的先验分布式迪利克雷分布，对于文档集合D中的任意一文档d，其主题概率分布是：

$$\theta_d = Dirichlet(\alpha) \quad (11)$$

主题词语的先验分布是迪利克雷分布，对于主题集合K的任意一主题k，其主题的词语概率分布是：

$$\beta_k = Dirichlet(\eta) \quad (12)$$

对于文档d中的第n个词，从主题分布θ_d中采样生成其主题的概率分布为：

$$z_{d,k} = multi(\theta_d) \quad (13)$$

对于主题$z_{d,n}$，从β_k中采样生成的概率分布为：

$$W_{d,n} = multi(\beta_k) \quad (14)$$

对于*LDA* 模型中存在的参数θ_d、$z_{d,n}$、β_k不能直接在*LDA* 模型中获取，我们采用随机抽样算法通过不断迭代采样推理主题的参数值。

通过*LDA* 模型，我们对附件 3 的群众留言进行分类。为了更好地定义热点问题，我们定义了一些指标，构建一个指标评价体系来定义热点问题。对于一个

热点问题，在某一段时间里群众大多数反映的问题，我们称之为热点问题，通过*LDA*模型，我们已经按照留言主题进行了分类，按照留言主题相似度，我们对主题进行聚类，找出在同一时间段里群众留言最多的问题留言。同时，根据群众留言的点赞数和反对数，在当今时代，大多数人会对与自身有关的事件进行点赞或反对，这同时反映了一个问题是不是热点问题的表现。点赞数越多，表明该类问题也是一个群众关心的问题，希望该问题可以得到好的解决。

所以我们先通过*LDA*模型对留言主题进行分类，通过聚类，并且根据点赞数和反对数来完善我们的指标评价体系。即：

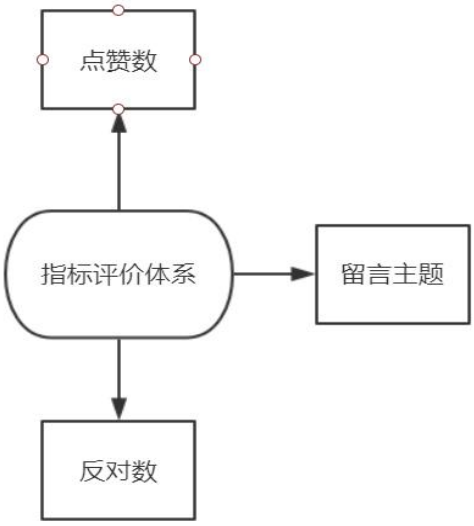

图 5.指标评价体系

指标评价体系由留言主题、点赞数、反对数构成，其中留言主题的权重相较于点赞数的权重要高，反对数由于是负面的评价指标，其权重是负的。

根据*LDA*模型和指标评价体系，我们通过 python 编写相应的程序，最终得出排名前五的热点问题，具体如下：

表 2：热点问题明细表

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	0.127	2019-01-01 至 2020-01-26	A 区市小区 A3A7	物业存在问题，影响业主居民休息
2	2	0.065	2017-06-08 至 2020-01-08	A 市区西地省县 A7 学院	西地省学院存在交通堵塞情况

3	3	0.063	2018-11-15 至 2020-01-08	A 市区 A7 县 A3 小区	开发商存在欺诈现象，购房者无法维权
4	4	0.060	2018-05-17 至 2020-01-07	A 市 A7 县小区 A3	车位捆绑违规销售
5	5	0.055	2019-08-18 至 2019-9-4	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民

4.4 问题三：对答复意见的质量给出一套评价方案

在完成群众留言的分类并交由相应负责的部门进行调查了解，解决群众留言问题，政府会给出相应的答复意见。对于政府给出的答复意见，我们希望看到政府对于群众留言问题的解决方法，所以我们希望对政府给出的答复意见做一个评价方案，借此方案来推动政府关于群众留言工作的工作效率。

对于答复意见，我们通过答复意见的相关性、完整性、可解释性、时效性这四个方面进行评价。即关于答复意见的评价方案，我们有四个指标：相关性、完整性、可解释性、时效性。

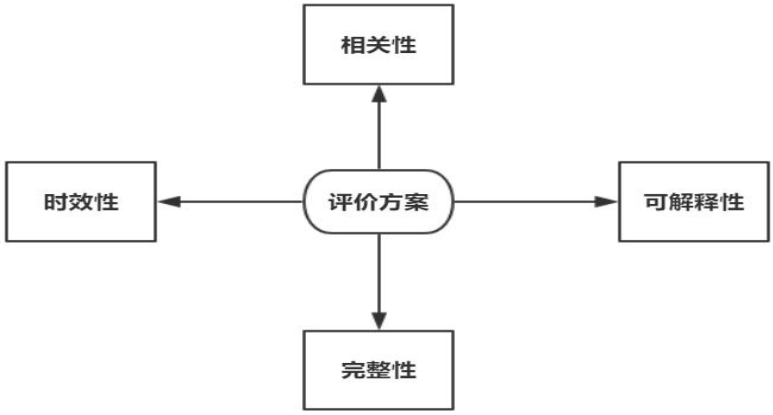

图 6：评价方案指标

相关性：答复意见与留言内容之间的相关性；

完整性：答复意见明确回答群众留言问题；

可解释性：答复意见中对于群众留言的解释，明确回答群众留言的问题；

时效性：答复意见在一定时间内给出。

根据附件 4 中的文本数据，根据附件 1 中的三级分类标签，我们对群众留言内容进行三级分类标签，利用 *Doc2Vec* ，我们对群众留言内容进行分类标签，同

时，我们对答复意见进行分类，确定答复意见的主题，通过判断两者主题的相似度，从而判断他们之间是否具有相关性。

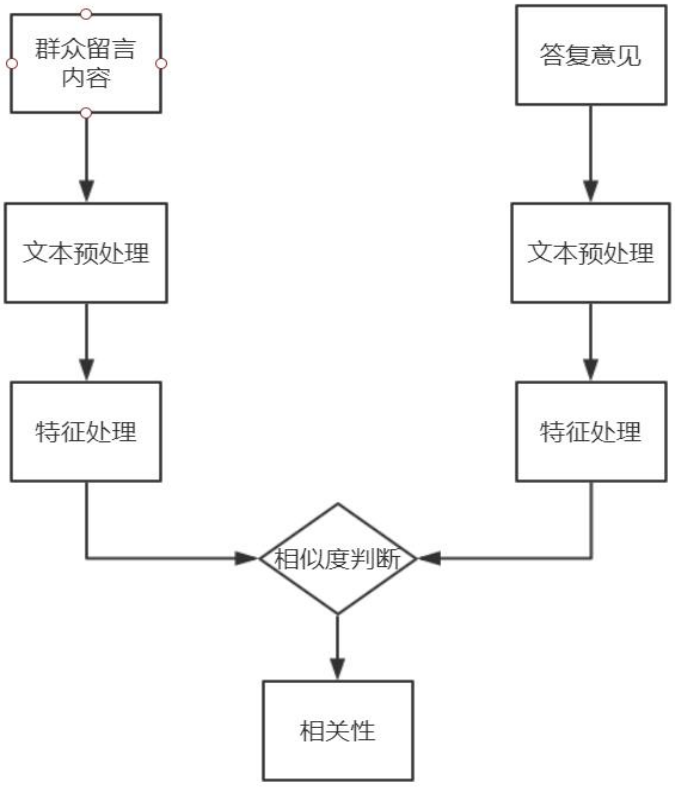

图 7. 相关性判断示意图

对于答复意见的完整性，我们利用文本的对象属性是否完整，文本 S 的任意对象属性属性 OP，属性 $S[OP_i]$ 的完整性记作 $T(S[OP_i]$，可以表示为 $\gamma(S[OP_i])$，其中 γ 是一个抽象的度量函数。若一个主体 S 的 OP 不存在，则认为该属性缺失，否则认为该属性完整。[7]函数 γ 定义:

$$\gamma(S[OP_i]) = \begin{cases} 1, S[OP_i]\text{不为空} \\ 0, S[OP_i]\text{为空} \end{cases} \quad (15)$$

利用函数 γ，我们来判断答复意见的完整性。

至于答复意见的可解释性，我们对于给出答复的一下属性进行说明，答复是否明确回答了群众留言问题，对于答复内容是否符合程序，对于群众留言内容能够达到很好的解释度。

至于时效性，当群众在网络问政平台上留言时，肯定会希望自己的留言问题能够得到政府的回应。因此在对答复意见的评价方案上，我们由时效性这样一个指标，通过政府答复意见日期与群众留言日期之间的差值，通过这个差值，我们来判断政府答复意见的时效性。

对此，我们对答复意见已经制定了一套评价方案，按照评价方案，我们对附件 4 中 2816 条留言分为优、中、差三个级别。我们通过 python 编写相关的程序，并对附件 4 中的答复意见进行评价。评价结果如下：

表 3:答复意见评价表

评价质量	数量
优	1270
中	1071
差	475

根据上表，我们对附件 4 中的答复意见质量进行了评价，按照我们制定的评价方案，评价为优的答复意见一共有 1270 条，评价为中的答复意见有 1071，评价为差的答复意见有 475 条。从评价结果可以看出，政府对于群众留言内容具有很大的重视，政府在收到群众留言后，会立刻进行调查，并及时将调查结果反馈给群众，从而提高政府的施政效率，也树立了政府了威信，让地方群众对政府有很高的认可度。

五、模型的评价

5.1 模型的优点

1.本文构建的 *Doc2Vec* 一级标签分类模型在对文本进行分类有优异的表现，可以很好地运用在机器学习、文本分类等问题上，具有很好的科学性和使用性。

2.本文构建的 *LDA* 模型是一种针对语料库进行建模的概率生成模型，对主题分类具有很好的表现情况，使用于文本聚类问题。

3.本文对答复意见质量构建了相应的评价方案，该方案对于答复意见质量评价进行指标评价，具有现实应用意义。

5.2 模型的缺点

1.本文构建的 *Doc2Vec* 一级标签分类模型在处理文本数据时不可避免会造成文本信息的缺失，这会导致分类模型会有误差。

六、参考文献

[1]李峰,柯伟扬,盛磊,陈雯,陈丙赛,罗韵晴.Doc2vec 在政策文本分类中的应用研究[J].软件,2019,40(08):76-78.
[2]贺益侗.基于 doc2vec 和 TF-IDF 的相似文本识别[J].电子制作,2018(18):37-39.
[3]李莹莹,李英.基于分类算法的移动支付系统商户采纳行为预测模型[J/OL].工业工程与管理:1-11[2020-05-08].http://kns.cnki.net/kcms/detail/31.1738.T.20200506.1516.004.html.
[4]桑菁. 基于 Doc2vec 和深度学习的文本情感分析研究[D].华北电力大学,2019.
[5]刘惠,赵海清.基于 TF-IDF 和 LDA 主题模型的电影短评文本情感分析——以《少年的你》为例[J].现代电影技术,2020(03):42-46.
[2]王瑞,龙华,邵玉斌,杜庆治.基于 Labeled-LDA 模型的文本特征提取方法[J].电子测量技术,2020,43(01):141-146.
[7]袁满,胡超,仇婷婷.基于 Linked Data 的数据完整性评估新方法*[J/OL].吉林大学学报(工版):1-8[2020-05-08].https://doi.org/10.13229/j.cnki.jdxbgxb20190546.