“智慧政务”中的文本挖掘应用

摘　要： 近年来，微信、微博等网络问政平台逐步成为政府了解民意的重要渠道，各类社情民意相关的文本数量攀升。与此同时，大数据、人工智能等技术飞速发展，建立基于自然语言处理技术的智慧政务系统是社会治理创新发展的新趋势。本文将利用自然语言处理和文本挖掘的方法，结合互联网上收集到的群众问政留言记录及部分群众留言的答复意见，建立模型解决问题。

针对问题一：通过 jieba 对附件 2 中的文本数据进行删除符号、停用词过滤、中文分词等数据预处理。划分训练集、测试集，留作模型评估使用。对数据做分类情况平衡性检验，结果显示数据分类差距较小，无需做平衡处理。利用 TF-IDF 模型将文本信息转化为向量化数字信息后，分别使用多项式朴素贝叶斯、逻辑回归、线性支持向量机三个分类模型进行留言一级标签分类，最后以 F-Score 方法对分类模型作出评价，得出基于逻辑回归或线性支持向量机的一级标签分类模型效果更佳。

针对问题二：同样通过 jieba 对附件 3 的文本信息进行停用词过滤、名词提取、分词处理等数据预处理，再利用 gensim 算法筛选出相似样本，并将相似样本汇总归类；然后以问题的发布时间、提出频数和社会关注度作为热点评价指标，使用 max-min 法对原始指标数据进行处理后，建立综合评价模型对问题进行热度评价，从而选出热点前五的留言问题，分别是：“丽发新城小区搅拌站引起噪音，粉尘问题”，“A 市人才购房补贴”，“A5 区魅力之城小区临街餐饮店油烟噪音扰民”，“伊景园滨河项目车位捆绑销售”和“A 市城市建设需求”。

针对问题三：同样对附件 4 的文本信息使用停用词过滤、分词处理等数据预处理技术，再将答复划分为三种情况。定义并量化各项答复评价指标，其中通过 SimHash 模型和机器学习识别分类来对相关性作两步判断，通过 TF-IDF 筛选关键词，建立对应关键词库将完整性和可解释性作等级划分。最后针对不同情况类别的答复设定差异化的指标要求，并按照其对答复进行评价，留言答复评价等级分为：“优秀”，“不完整”，“缺乏解释”和“不相关”。

关键词： TF-IDF；Jieba 分词；逻辑回归；线性支持向量机；genism；综合评价模型；SimHash

Text Mining Application in "Smart Government Affairs"

Abstract:

In recent years, online questioning platforms such as WeChat and Weibo have gradually become an important channel for the government to understand public opinion, and the number of texts related to various social conditions and public opinion has risen. At the same time, the rapid development of big data, artificial intelligence and other technologies, the establishment of smart government systems based on natural language processing technology is a new trend of social governance innovation and development. This article will use natural language processing and text mining methods, combined with the public's question and answer message records collected on the Internet and some of the people's comments, to establish a model to solve the problem.

Aiming at problem one:. Data preprocessing such as deleting symbols, stop word filtering, Chinese word segmentation, etc. are performed on the text data in Annex 2 through jieba. Divide the training set and test set for the model evaluation. The balance test of the classification of the data shows that the data classification gap is small and no smooth processing is required. After the text information is converted into vectorized digital information using the TF-IDF model, three classification models of polynomial naive Bayes, logistic regression, and linear support vector machine are used for message first-level label classification, and finally the classification model is made with F-Score Evaluation, it is concluded that the first-level label classification model based on logistic regression or linear support vector machine is better

Aiming at problem two: Similarly, through jieba, the text information of Annex 3 is subjected to preprocessing such as stopword filtering, noun extraction, and word segmentation, and then using gensim algorithm to filter out similar samples and summarize the similar samples into categories; The release time frequency and social attention are used as hotspot evaluation indicators. After processing the original index data using the max-min method, a comprehensive evaluation model is established to evaluate the problems, and the top five hotspot questions are selected, namely: "Lifa Xincheng Community The noise and dust problems caused by the mixing plant "," Subsidy for talent purchase in City A "," Fume noise from street restaurants in the charming city community of A5 District, disturbing people "," Bundled sales of parking spaces in Yijingyuan Riverfront Project "," A city construction demand " ..

Aiming at problem three: Similarly, data preprocessing techniques such as stopword filtering and word segmentation processing are used for the text information in Annex 4, and then divide the response into three cases. Define and quantify various response evaluation indicators, including two-step judgment on relevance through SimHash model and machine learning recognition classification, screening keywords through TF-IDF, establishing corresponding keyword database to rank completeness and interpretability as grades Divide. Finally, set differentiated index requirements for the responses of different types of situations, and evaluate the responses according to them. The evaluation level of the message reply is subdivided into: "excellent", "incomplete", "lack of explanation", "irrelevant".

Key words: TF-IDF; Jieba word segmentation; genism; logistic regression; linear support vector machine; comprehensive evaluation model; SimHash model

目　录

一、 挖掘目标

网络问政平台凭借其信息传播速度快，空间距离短，方便快捷等优点，已然成为市民反馈社情民意的重要渠道。而原始的人工处理在留言数据量的渐增下，更会出现效率低，错误率高的情况。本次建模的目标是利用 jiieba 分词工具，TF-IDF 算法，线性支持向量机模型，gensim 算法，SimHash 模型等，对留言进行处理和分析，解决以下三个问题：

问题一：留言分类问题。即根据附件一提供的划分体系以及附件二提供的已做好一级标签分类的留言数据，建立关于留言内容的一级标签分类模型，以便后续的群众留言能通过建立的模型完成分类，派送至相应的职能部门处理。

问题二：热点问题挖掘，包括热点问题提取和热点指数评定两部分。根据留言主题的相似度，对相似留言进行归类，对热点指数进行综合评价建模，挖掘出群众集中反映的热点问题，利于工作人员和相关部门及时了解群众的聚焦点，尽快处理该问题。

问题三：对答复意见的评价。根据对应留言实际，构建答复评价指标，建立答复评价模型，检验答复质量，输出结果显示具体问题，便于工作人员进行答复勘误。

二、 分析方法与过程

2.1 问题一分析方法与过程

2.1.1 流程图

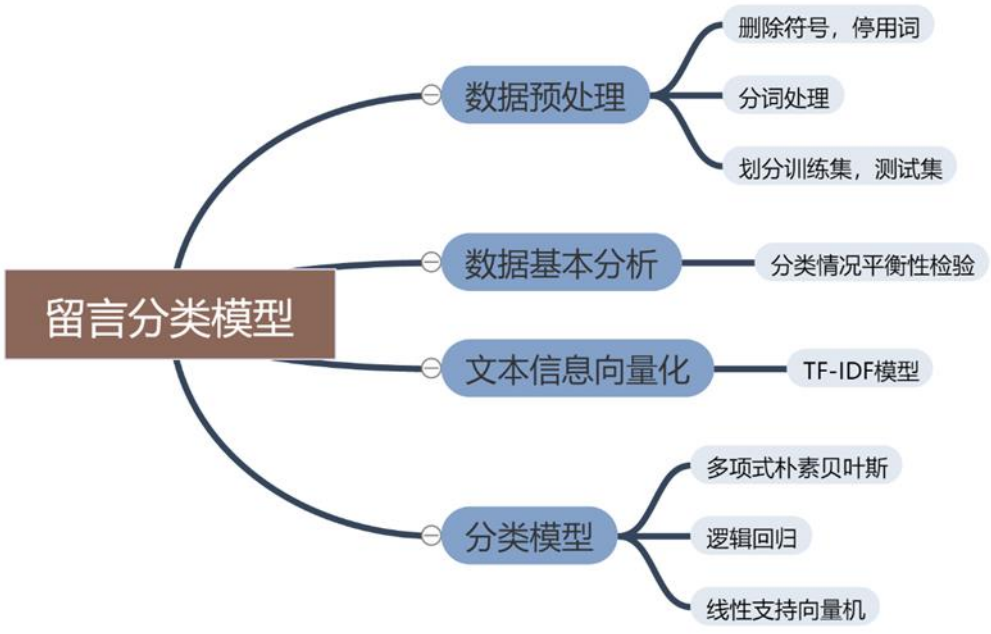

图 1 留言分类模型流程图

2.1.2 数据预处理

通过观察附件二的数据，可以得出标签分类的主要依据是留言主题和留言详情。由于留言详情多为较长的中文文本，存在不少累赘的表达，会给后续的模型建立带来不少的干扰信息。而留言主题是留言详情内容的基本概括，且留言主题文本短而关键，所以本文仅选择留言主题作为一级标签分类的依据来建立模型。

（一）删除符号

留言主题数据中存在的标点符号、特殊符号对分析以及预测留言内容无意义，应将其去除，从而能够减少无用信息对模型的干扰，减少计算的复杂度。

（二）过滤停用词

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之前或之后自动过滤掉某些字或词。停用词的主要特征为无实际意义或使用十分广泛。因此在建立模型前需要将停用词过滤掉。过滤停用词需要停用词表，但没有一个明确的停用词表能够使用于所有的模型，因此需要在后续的建模过程中不断调整、修改停用词表。

（三）进行 jieba 分词

中文分词是文本挖掘的基础，是将连续的字序列按照一定的规范重新组合成词序列的过程。jieba 分词主要是基于统计词典，构造一个前缀词典；然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置，构造一个有向无环图；通过动态规划算法，计算得到最大概率路径，得到最终的切分形式。本文选择 jieba 分词的精确模式，将文本句子精确地切开，得到词序列。

（四）训练集与测试集划分

若使用训练模型的数据对模型进行测试，其结果会过于乐观，因此需要将数据集划分为训练集和测试集。为了能够更好地对建立好的模型进行测试与评估，本文将附件二提供的留言数据集按照 8:2 的比例划分为训练集和测试集。

2.1.3 数据基本分析

若数据集中类别极其不平衡，容易造成稀有类的错判。在建立分类模型前应了解数据集中各类样本的数量。因此统计了附件二中各类别留言数据的数量，结果如下图所示：

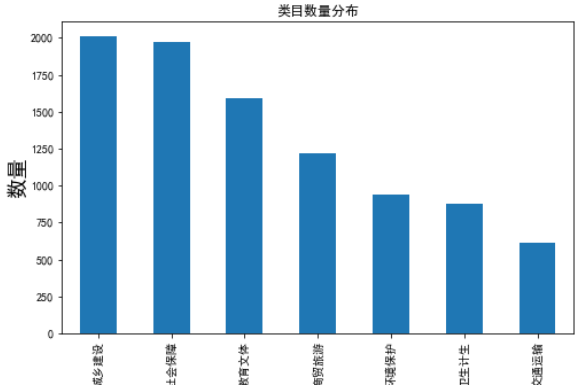

图 2　各留言类别数量分布

可以看出本数据集并非一个类的分布极其不平衡的数据集，因此不对数据集做平衡处理。

2.1.4 将文本信息转换为向量信息

自然语言文本信息的分类需要将文本信息转化成数字向量信息[1]，这一过程称为文本表征。为了能够建立文本多分类模型，本文选择使用 TF-IDF 模型，对经过预处理的留言主题信息进行数字化。

TF-IDF 是一种用于信息检索与文本挖掘的常用加权技术。 TF-IDF 是一种统计方法，用以评估一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。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。所以可以在训练集中借助 TF-IDF 技术选出同一标签文本中具有重要意义的关键字词，可认为该词具有很好的类别区分能力，从而利用该词进行文本分类。

（1）TF 词频（term frequency）

TF 表示词条在文本中出现的频率：

$$TF_m = \frac{\text{在某一级标签分类中词条 m 出现的次数}}{\text{该类中所有的词条数目}}$$

（2）IDF 逆向文件频率(Inverse Document Frequency)

$$\text{IDF} = \log\left(\frac{\text{语料库中的文本总数}}{\text{包含词条 m 的文本数} + 1}\right)$$

（3）TF-IDF

$$\text{TF} - \text{IDF} = \text{TF} * \text{IDF}$$

本文首先通过调用 Python 中 sklearn.feature_extraction.text 里的 CountVectorizer，来训练文本数据中各词在各类文本下的词频，形成词频矩阵。再使用 Python 中 sklearn.feature_extraction.text 里的 TfidfTransformer，统计出每个词语的 tf-idf 权值。经过这两个步骤，便完成了将文本信息转换成数字向量信息，从而可以进行分类模型的训练。

2.1.5 分类模型

（一）多项式朴素贝叶斯[2]

朴素贝叶斯算法是一种基于贝叶斯定理和特征条件独立假设的分类方法。贝叶斯分类器所需估计的参数很少，对缺失数据不太敏感，算法也比较简单，可解释性强。

贝叶斯分类器训练阶段的公式为：

$$\text{先验概率：}\mathrm{P(C=c)}=\frac{\text{属于类 c 的文档数}}{\text{训练集文档总数}}$$

$$\text{条件概率：}\mathrm{P}(w_i|\mathrm{c})=\frac{\text{词}\mathrm{w_i}\text{在属于类 c 的所有文档中出现的次数}}{\text{属于类 c 的所有文档中的词语总数}}$$

贝叶斯分类器预测阶段的公式为：

$$\arg\max_{c\in C} P(c|w_1,w_2,\dots,w_3)=arg\max_{c\in C}[logP(c)+logP(w_1|c)+\cdots+logP(w_n|c)]$$

（二）逻辑回归

逻辑回归是经典的二分类算法。选用 Sigmoid 函数，将输入映射为概率值，实现预测功能，通过设置概率阈值实现分类功能。逻辑回归也能实现多类别分类，对于 k 个可能的结果，将运行 k-1 个独立二元逻辑回归模型。

逻辑回归的运算公式为：

$$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}}$$

$$\mathrm{P(y|x;\theta)}=(h_\theta(x))^y(1-h_\theta(x))^{1-y}$$

（三）线性支持向量机[3]

SVM 是在解决小样本、非线性、高维的分类和回归问题时具有特有优势的机器学习方法,在 SVM 基础上发展的线性支持向量机已成为处理文本分类等海量高维稀疏数据的一种有效机器学习方法。

线性支持向量机可以表示为求解下式的无约束优化问题：

$$\min_w f(w)=\frac{1}{2}w^T w+C\sum_{i=1}^{l}\xi(w,x_i,y_i)\,(\mathrm{C}>0)$$

2.2 问题二分析方法与过程

2.2.1 流程图

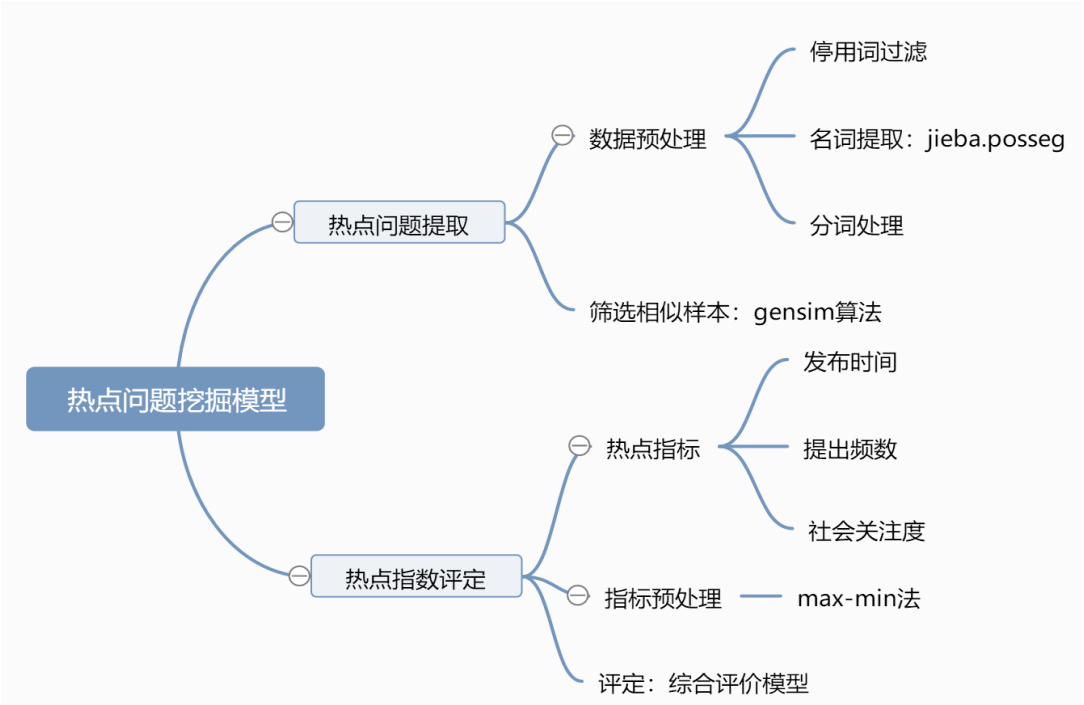

图 3 热点问题挖掘模型流程图

2.2.2 数据预处理

由于同类热点问题的表达具有多样性，在相似度检验中会出现区分不明显的情况，本文将重点聚焦在留言主题提出的面向主体之中，利用已处理过停用词的留言主题，运用 jieba.posseg 进行标注后，对名词进行提取。

2.2.3 筛选相似样本

当两条留言主题进行预处理后，二者的相似度越高，两条留言提出的问题为同一事件的可能性越高。因此下面先通过 python 中的 gensim 库[4]，对样本进行筛选，并将相似样本归为一类。gensim 库是一个用于从文档中自动提取语义主题的 python 库，可处理原生非结构化的数值化文本。

本文运用的是其中的 Latent Semantic Analysis 算法(潜在语义分析 LSA) 。LSA 算法通过在语料库的训练下检验词的统计共生模式(statistical co-occurrence patterns)来发现文档的语义结构。

其中本文运用到的 gensim 核心概念包括：

（1）Corpus：数字化文档的集合，被用于自动推断文档的结构，主题等。此处为预处理后全体留言。

（2）Vector：在向量空间模型中，每个文本被表示成了一组特征。后续需要利用该特征计算相似度。

（3）Model：基于 Corpus 对文本的表示进行转换。此处采用的是 LSA 模型。

2.2.4 热点指数衡量模型

（一）热点问题的定义

某一时间段群众集中反映的某一问题可称为热点问题。

（二）综合评价模型

通过建立合适的综合评价数学模型将多个评价指标综合成为一个整体的综合评价指标，即得到相应的综合评价结果。

假设 n 个被评价对象的 m 个数据指标向量，$x = (x_1, x_2, \dots, x_m)^T$,$w = (w_1, w_2, \dots, w_m)^T$，则构造综合评价函数$y = f(w, x)$

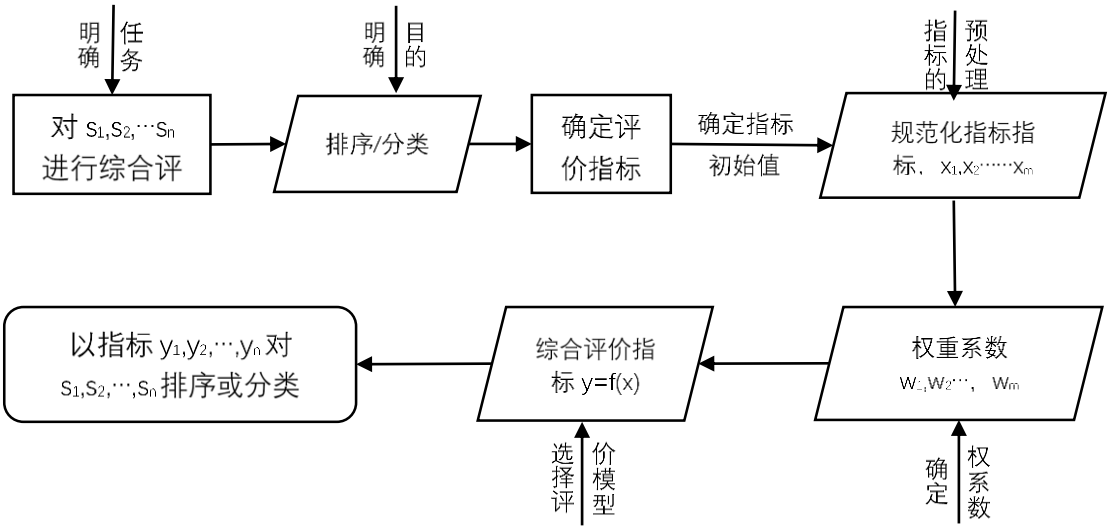

图 4 综合评价流程图

（三） 热点指标的定义和预处理

根据附件三中给出的数据，结合现实情境分析得出衡量热点的指标应有以下三点：

（1）发布时间：该问题被反映的最新时间越接近当下，热点越高。

$$T_i = t_{new} - t_i$$

t_{new}：目前最新留言的日期，t_i：该问题最新的留言日期

（2）提出频数：提出该问题的次数和人数越多，热点越高。

$$R_i = (Q_i - P_i) \times 0.6 + P_i$$

Q_i：该问题留言总数，P_i：提出该问题的人数

（3）社会关注度：除发布者以外的群众对该问题的关注程度越高，热点越高。

$$S_i = A_i + D_i$$

A_i：该热点问题的所有留言点赞总数，D_i：该热点问题的所有留言反对总数

按照上述对热点问题的提取结果，统计热点问题对应指标的极值，具体如下：

由于各热点指标性质，量级不一，数值较高的指标在综合分析中的作用会削弱数值水平较低的

作用。因此为保证评价结果的可靠性，需要对原始指标数据进行数据标准化处理，运用 max-min 法进行预处理。利用上述的数据特征，对序列 t,Q,S 中的每个数据进行变换：

$$t'_i = \frac{t_i - min\{t\}}{max\{t\} - min\{t\}}$$

$$R'_i = \frac{R_i - min\{R\}}{max\{R\} - min\{R\}}$$

$$S'_i = \frac{S_i - min\{S\}}{max\{S\} - min\{S\}}$$

（四）热点指数综合评价模型

（1）在权值系数方面，为获取相对科学的权值系数，通过问卷调查的方式，获得 200 份大众对三个指标重要性的排序结果，所获得到的统计结果如下：

热点问题指标重要性

第 1 题 请给以下三个选项按照对问题热点影响的重要性排序 [排序题]

选项	平均综合得分
发布时间:发布时间越接近当下，热点越高	2.64
提出频数:提出该问题的人和次数越多，热点越高。	2.25
社会关注度:对该留言的点赞数反对数越高，热点越高	1.11

图 5 热度评价指标重要性排序

可见从发布时间，提出频数到社会关注度的综合评分为 2.64，2.25，1.11，对应对问题热点指数的重要程度依次降低。按照问卷结果给三热度指标赋予权重值：

$$w_i = \frac{\textit{该指标平均综合得分}}{\textit{三指标综合得分加和}}$$

其赋值分别为 0.44，0.375，0.185。

（2）评价函数的选择

由于三个热点指标之间的关联性不强，相对独立，故采用线性加权综合法（定义热点满值为 100）：

$$H_i = (w_1 \times t'_i + w_2 \times R'_i + w_3 \times S'_i) \times 100$$

2.3 问题三分析方法与过程

2.3.1 流程图

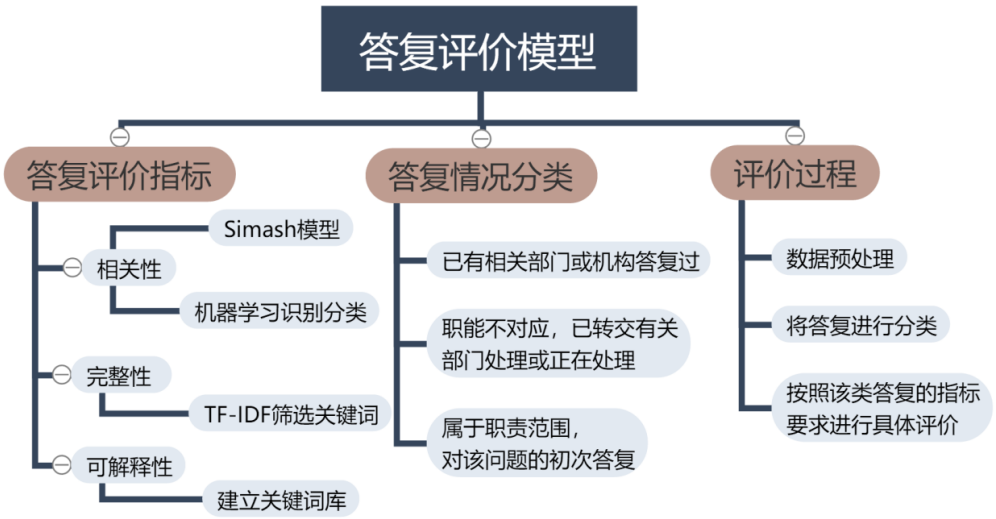

图 6 答复评价模型流程图

2.3.2 数据预处理

（一）对留言和答复信息的停用词处理

在对留言和答复信息进行挖掘分析前，由于文本中存在大量的如“的”，“了”，“和”等对文本大意影响较弱的词和符号，为了提高对文本搜索效率和效果，将该类词语归为“停用词”，并在正式分析前将符号和停用词去除。

（二）对留言和答复信息进行分词处理

在根据留言内容对答复进行评价之前，需要先把非结构化的文本信息转化为计算机能识别的结构化信息。在附件四中，需要运用到的留言详情和答复意见均以中文文本的形式给出数据。因此这里采用 python 的中文分词包 jieba 对以上文本信息进行分词处理。jiieba 基于 Trie 树结构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），并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，而对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法，是目前做最好的 Python 中文分词组件，从而使信息的中文分词效果较好。

2.3.3 答复情况的分类

针对情况下的答复，其评价要求应有所不同。故在评价前应对答复进行分类。本文考虑到问政平台的职能和问题是否曾被提出，将答复分为以下三种情况：

（1）已有相关部门做出过答复：

此类答复常具有“已经答复”,“曾答复”等标志字眼，具有篇幅小（60 字以内）的特点。

（2） 职能不对应，已转交有关部门处理或正在处理：

此类答复常具有“已转”,“已将…转”等标志字眼，具有篇幅小（60 字以内）的特点。

（3） 属于职责范围，对该问题的初次答复：

筛选完以上两种情况的留言后，将剩余答复归入此类。此类答复普遍具有篇幅长（60 字以上）的特点，并需要有提供相关渠道或信息的功能和有完整格式的规范。

2.3.4 评价指标的定义和量化方法

（一）评价指标的定义

针对附件四中相关部门对留言的答复意见，本文从答复的相关性，完整性，可解释性三个指标对答复进行评价，以下为指标的具体定义：

①相关性：答复意见的内容分是否与问题相关。

②完整性：是否满足某种规范。

③可解释性：答复意见中是否有对留言问题的相关解释。

（二）评价指标的量化方法

（1）相关性：

第一步判断：SimHash 模型[5]

采用 SimHash 模型，初步检验得出各条留言详情与答复意见的相关程度，dis 为两条文本的汉明距离，dis 越大，文本相关性越小。

SimHash 算法的原理是降维，将高维的特征向量映射成低维的特征向量，通过两个向量的汉明距离来确定文章是否重复或者高度近似。其中，两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数。即将一个字符串变换成另一个字符串所需要替换的字符个数。例如：1011101 与 1001001 之间的汉明距离是 2。至于我们常说的字符串编辑距离则是一般形式的汉明距离。

SimHash 生成文本指纹可分为以下 5 个步骤：分词，Hash 权重计算，加权，合并和降维。以下为 SimHash 流程图：

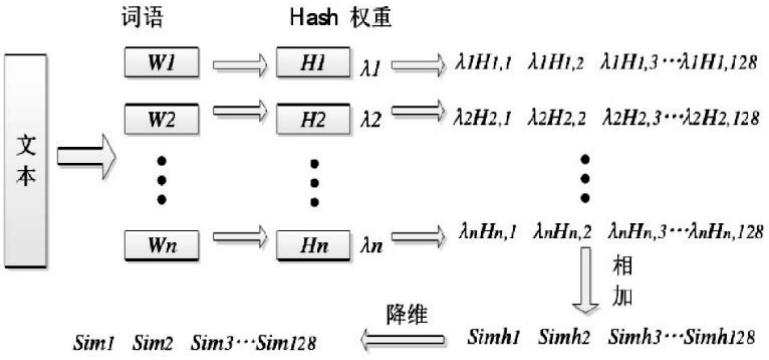

图 7 SimHash 流程图

操作步骤：

① 计算留言详情与答复意见的汉明距离[5]

取每个文本中 TF-IDF 权重最高的前 20 个词，对这些词进行普通的 Hash 算法处理后，得到每个词一个 64 位二进制长度为 20 的集合，在二进制数对应位置中分别取正负权重，从而得到对应列表，从而获取文本的 SimHash 值，对 SimHash 取异或，看其中相异的个数。由于检验目标针对的留言详情和答复意见篇幅较短。将超过 3 定义为不相关，小于等于 3 则定义为相关。

② 相关性初步划分

将通过第三步检验的答复相关性记为 2，剩余答复记为 0。

第二步判断：机器学习分类

由于 SimHash 模型对短文本检测的准确率难以保证，故对未能通过 SimHash 检验的文本将继续检验。运用第一题构建的留言分类模型，对留言详情和答复意见进行分类模型构建，将未通过第一步的答复进行再判断，若留言详情和答复意见的分类结果相同，则答复相关性记为 1，否则记为 0。当相关性为 0 时，该答复被判定为不相关。

（2）完整性：

首先，按照答复情况特性，将所有答复以 60 字为界限划分为两类。其中少于 60 字的样本数量较少。通过观察，定义对应完整性规则（在情况分类要求下将详细说明）。而大于 60 字的部分，大多归类于“属于职责范围，对该问题的初次答复”。我们有理由相信，当答复意见对应字数较多时，其完整性高的可能性大，故在 excel 中筛选出字数大于 200 的答复意见，对答复进行停用词去除后，根据 TF-IDF 算法，根据词频，初步挑选出完整性较高答复中的关键词，然后对答复意见再进行检验，对完整性词库人工进行分类和补充。将以下格式视为第三类答复意见中衡量完整性应包含的内容：

①称呼，问好（如：尊敬的网友，您好）

②指出并回应留言中提出的问题（如：针对 xx 问题，回应如下）

③对民众表示感谢或抱歉（如：感谢您对 xx 的关注）

④标注时间（如：20xx 年 x 月 x 日）

当程序检索到文本含有对应部分关键词时，该部分完整性计分 1，若无则为 0。满分 4 分，分值越高，完整性越高。

（3）可解释性：

指答复意见中有使用法律，条例，方案标准等具有官方权威的材料对答复进行支撑。量化过程和完整性大致相同，但需事前基于大量答复意见，人工建立可解释性的关键词库。再对答复意见进行检索，包含关键词则认为答复具有可解释性，用 1 表示；否则认为该答复缺乏解释，用 0 表示。

2.3.5 对应答复分类下的评价指标要求

（一）已做出过答复：

由于其文本长度较短，叙述内容并非与留言直接相关，此处不对此类进行相关性检验。

具体指标要求如下：

（1）完整性：需具有称呼和答复日期两部分

（2）可解释性：提供对应问题的答复链接。

根据以上两点要求将对该类答复的评价分为“不完整”，“未附上链接”和“优秀”三类。

（二）职能不对应，已转交有关部门处理或正在处理

同样由于其文本长度较短，叙述内容并非与留言直接相关，不对此类进行相关性检验。因为不属于职能范畴，故对此类答复的可解释性不作要求。

具体指标要求如下：

（1）完整性：需具有称呼，答复日期，致谢三部分

按此要求将该类答复的评价划分为“不完整”和“优秀”两类。

（三）属于职责范围，对该问题的初次答复

筛选完以上两种情况的留言后，将剩余答复归入此类。由于是初次答复，其要求需相对严格，对各项评价指标均有要求。

对于属于职责范围，针对该问题的初次答复各指标要求如下：

（1）相关性：答复意见的内容部分是否与问题相关

（2）完整性：称呼，答复问题，致谢，日期

（3）可解释性：答复意见具有相关解释

按照以上要求，将该类答复的评价分为“不完整”，“缺乏解释”，“不相关”和“优秀”四类。

三、 结果分析

3.1 问题一结果分析

在用模型进行分类后，我们将借助 F-score 方法对模型进行评价。

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

其中 P 为精确率(Precision)，R 为召回率(Recall)，n 为类型数量。

由于分别使用了多项式朴素贝叶斯，逻辑回归，线性支持向量机三种分类模型，需分别计算三种模型的 F_1 值。F_1 值最高的模型将被认为分类效果最好。

（一）多项式朴素贝叶斯

表 1 多项式朴素贝叶斯模型 F1 值计算

	precision	recall	fscore	support
城乡建设	0.69551777	0.88932806	0.78057242	506
劳动和社会保障	0.94086022	0.71721311	0.81395349	244
教育文体	0.98863636	0.58	0.73109244	150
商贸旅游	0.8516624	0.85384615	0.85275288	390
环境保护	0.76666667	0.92555332	0.83865087	497
卫生计生	0.87148594	0.68454259	0.76678445	317
交通运输	0.92253521	0.65829146	0.76832845	199

通过计算上述数据,得出使用多项式朴素贝叶斯分类模型的整体 F1 值为:0.7931621414408682。

（二） 逻辑回归

表 2 逻辑回归模型 F1 值计算

	precision	recall	fscore	support
城乡建设	0.74471831	0.83596838	0.7877095	506
劳动和社会保障	0.85714286	0.81147541	0.83368421	244
教育文体	0.92592593	0.83333333	0.87719298	150
商贸旅游	0.87631579	0.85384615	0.86493506	390
环境保护	0.82089552	0.88531187	0.85188771	497
卫生计生	0.86415094	0.72239748	0.78694158	317
交通运输	0.85638298	0.80904523	0.83204134	199

通过计算上述数据，得出使用逻辑回归分类模型的整体 F1 值为：0.8334846264658584。

（三） 线性支持向量机

表 3 线性支持向量机模型 F1 值计算

	precision	recall	fscore	support
城乡建设	0.74471831	0.83596838	0.7877095	506
劳动和社会保障	0.85714286	0.81147541	0.83368421	244
教育文体	0.92592593	0.83333333	0.87719298	150
商贸旅游	0.87631579	0.85384615	0.86493506	390
环境保护	0.82089552	0.88531187	0.85188771	497
卫生计生	0.86415094	0.72239748	0.78694158	317
交通运输	0.85638298	0.80904523	0.83204134	199

通过计算上述数据，得出使用线性支持向量机分类模型的整体 F1 值为：0.8334846264658584。

比较三个分类器计算出来的 F1 值，可以看出多项式朴素贝叶斯 F1 值最低，而逻辑回归和线性支持向量机的 F1 值一样高。因此推断出应选择逻辑回归和线性支持向量机作为一级标签分类模型的分类器。

3.2 问题二结果分析

通过热点综合评价模型，统计得出热点指数为前五的问题如下：

表 4 热点排名前五的问题

热度排名	问题 ID	热点指数	时间范围	地点/对象	问题描述
1	1	78.75	2019/11/15 至 2020/1/9	丽发新城小区	搅拌站造成噪音和粉尘问题
2	2	77.71	2018/11/15 至 2019/12/2	A 市人才	购房补贴通知问题
3	3	50.55	2019/7/21 至 2019/9/25	A5 区劳动东路魅力之城小区	小区临街烧烤夜宵摊油烟噪音扰民
4	4	45.42	2019/7/7 至 2019/9/1	景园滨河苑项目	非法绑定车位出售，引起众怒
5	5	39.37	2019/1/8 至 2019/9/9	A 市	加快现代服务业中心城市建设

对应的热点问题留言明细部分结果如下，详情见附件中的热点留言明细表：

表 5 热点问题留言明细部分结果

问题ID	留言编号	留言用户	留言主题	留言时间	点赞数	反对数
1	288673	A909216	举报A2区	2019/11/15	0	0
1	288673	A909201	丽发新城	2019/11/15	2	0
1	288720	A909208	投诉小区	2019/11/18	0	0
1	288720	A909207	投诉小区	2019/11/23	9	2
1	288673	A0005147	丽发新城	2019/11/29	1	0
1	288673	A909242	丽发新城	2019/12/4	1	0
1	288720	A0009627	投诉小区	2019/12/4	0	0

3.3 问题三结果分析

3.3.1 评价模型结果展示

对全部数据进行分类和评价的部分结果如下：（详情见附件中的分类评价结果）

表 6 数据分类和评价的部分结果

留言编号	留言详情	答复意见	答复分类	具体评价
2549	2019年4月以	现将网友在	属于职责范围的初次答复	优秀
2554	潇楚南路从2(	网友“A0002	属于职责范围的初次答复	优秀
2555	地处省会A市[	市民同志：(	属于职责范围的初次答复	不完整
2557	尊敬的书记：	网友“A0001	属于职责范围的初次答复	优秀
2574	建议将“白竹	网友“A0009	属于职责范围的初次答复	缺乏解释
2759	欢迎领导来A(	网友“A0007	属于职责范围的初次答复	缺乏解释
2849	尊敬的胡书记	网友“A0001	属于职责范围的初次答复	优秀
3681	我做为一东濂	网友“UU008	属于职责范围的初次答复	优秀
3683	我是美麓阳光	网友“UU008	属于职责范围的初次答复	优秀

其中展示的内容包括：留言编号，留言详情，答复意见，答复分类以及具体评价。通过 excel 表格的筛选功能，可选择出存在问题的答复，方便同时对应留言详情和答复意见直接进行修改。

3.3.2 评价模型结果分析

基于分类评价结果利用 excel 软件绘制关于答复分类和具体评价的数据透视表。根据图表显示，在总共 2816 个样本中，属于职责范围的初次答复共有 2605 次，已答复的答复 9 次，已转交的答复 202 次。根据评价模型的输出结果，检验出各类答复所存在的问题以及对应数量。

表 7 各类答复存在的问题

计数项:留言详情	答复分类			
具体评价	属于职责范围的初次答复	已答复	已转交	总计
不完整	1179	8	175	1362
不相关	104			104
缺乏解释	90			90
未附上链接		1		1
优秀	1232		27	1259
总计	**2605**	**9**	**202**	**2816**

从数量最少的已有相关部门做出过答复的类别开始，通过展开答复意见的具体内容开始分析。其中被评价为不完整类别的部分，对于称呼，日期方面均不两者皆具；而被评价为未附上链接的答复可见的确未附上对应的答复链接，证明对该类别下的评价均为正确。

表 8 不完整和未附上链接问题的评价分析

留言编号	留言详情	答复意见	答复分类	具体评价
51521	老工人的苦	网友：您好！您所反映的问题，已进行过回复。	已答复	不完整
125614	田县长你好，	“UU0081255”您反映的问题已有相关单位作出回复，请您耐心等待。（详见感谢您的留言。	已答复	不完整
126023	我们在2017	网友“UU0081524”您好！您反映的问题相关部门已经答复。链接：https://baidu.com/	已答复	不完整
126048	尊敬的领导：	“UU008929”：您好！您反映的问题相关部门已经答复。链接：	已答复	不完整
126049	通道第一民	“UU008995”：您好！您反映的问题教育局已经答复。链接：	已答复	不完整
126058	团头村转朝	UU008936：您好！您反映的问题县交通运输局已经答复。链接：	已答复	不完整
126060	关于通K6县	“UU008732”：您好！您反映的问题县交通运输局已经答复。链接：	已答复	不完整
130205	自L4县	“UU008984”：　　您好，您反映的问题县教育局已回复，链接如下：　　2017年3月6日	已答复	未附上链接
135204	投诉L10县隆	“UU008604”您反映的问题，相关单位已作出回复。（详见：感谢您的留言。	已答复	不完整

对职能不对应，已转交有关部门处理或正在处理类别的答复，按照具体评价类别所占的比例，分别随机抽取 17 条被评价为不完整和 3 条被评价为优秀的答复对评价准确性进行检验，通过实际对照答复意见内容，整理出其对应的实际情况。可见实际情况和具体评价结果一一对应，证明在抽取样本中，此类评价的准确率极高。

表 9 职能不对应问题的评价分析

留言编号	留言详情	答复意见	答复分类	具体评价	实际情况
119730	你好，我	您好，请您到市交警队咨询相关事宜。感谢您对房产工作的支持和理解。	已转交	优秀	具备称呼，致谢和日期
122126	华夏湖天	网友：您好，您反映的问题已转至L1区，感谢您的问政。2016年11月24日	已转交	优秀	具备称呼，致谢和日期
122349	情况如下：	网友：　　您好，您反映的问题已转至L5县，感谢您的问政。2016年6月	已转交	优秀	具备称呼，致谢和日期
76363	F市南湖新区	您的留言已收悉。关于您反映的问题，已转南湖新区管委会调查处理。	已转交	不完整	缺乏日期
82452	由于乡下偏远	网友：您好，留言已收悉，已转交相关单位办理。2019年2月1日	已转交	不完整	缺乏致谢
106505	你好，我是	网友：你好！你的帖文已收悉并转至相关单位。2019年10月9日	已转交	不完整	缺乏致谢
121228	地处L市核心	网友：您反映的问题已转至市城管局调查核处。2018年4月17日	已转交	不完整	缺乏致谢
130546	我们岩山脚	“UU0081950”您好！你反映的问题我办已转交给相关单位。	已转交	不完整	缺乏日期
74186	我们是F市阿	您的留言已收悉，关于您反映的问题，已转市住建局调查处理。	已转交	不完整	缺乏称呼，致谢和日期
74240	F市吉顺安商	您的留言已收悉，关于您反映的问题，已转相关部门调查处理。	已转交	不完整	缺乏称呼，致谢和日期
74462	国家供电部	您的留言已收悉，关于您反映的问题，已转市城管局调查处理。	已转交	不完整	缺乏称呼，致谢和日期
119869	12月10日，H	网友：您好！您反映的问题已转高速公路管理部门调查、核处。	已转交	不完整	缺乏致谢和日期
123928	本人家里用	网友：您好！您反映的问题已转L市水务投公司调查、核处。	已转交	不完整	缺乏致谢和日期
135699	彭书记	亲爱的网友：　　您反映的问题已转至县农业局核实、办理。	已转交	不完整	缺乏致谢和日期
133902	L6县邮政局	已转办至L6县邮政局L6县网宣办2016年2月22日	已转交	不完整	缺乏称呼和致谢
135695	L10县县	亲爱的网友：　　您反映的问题转至县民宗局核实、办理。	已转交	不完整	缺乏致谢和日期
36464	敬启者	网友：　　您好！您所反映的问题已移交B4区法院强制执行。感谢留言！	已转交	不完整	缺乏日期
58143	南门湖延期	网友：您好！您反映的问题，我们已转交相关单位。2019年12月9日	已转交	不完整	缺乏致谢
117473	我是潇楚佳	您反映的问题不属于我单位职责范围，感谢您的留言。	已转交	不完整	缺乏称呼和日期
120022	你好！L市木	网友：您好！您反映的问题已转市人社部门调查、核处。	已转交	不完整	缺乏致谢和日期

对属于职责范围，对问题的初次答复，由于已知 Simash 模型对于识别文本长度差异较大时，对文本相关性检验准度较差，故此处特针对被评价为不相关的部分进行检验，抽取了 20 条被评价为不相关的数据，通过对比实际问题和对应答复意见的真实相关性归纳出下表。下表显示的被误判的共有 8 条，真实不相关的共有 12 条。可见即使经过第二步判断，对相关性检验这方面还是没有达到很好的效果。而关于完整性和可解释性，由于是利用关键词搜索衡量，通过更新关键词词库和规范工

作人员用词，二者均能达到较好的效果。

表 10 相关性检验分析

留言编号	留言详情	答复意见	答复分类	具体评价	实际情况
7582	我于201	网友“UU0081151” 您好！您的留言已收悉	属于职责范围	不相关	相关
8567	由于公	网友“UU0081760” 您好！您的留言已收悉	属于职责范围	不相关	相关
10029	书记您好	感谢您对我们工作的关心、监督与支持。	属于职责范围	不相关	并无回应问题，不相关
10646	我是A7县	网友“UU0081200” 您好！您的留言已收悉	属于职责范围	不相关	相关
10924	尊敬的易书记	网友：您好！留言已收悉	属于职责范围	不相关	并无回应问题，不相关
11924	A市是西	网友：您好！留言已收悉	属于职责范围	不相关	并无回应问题，不相关
19602	尊敬的领	网友：您好！留言已收悉	属于职责范围	不相关	并无回应问题，不相关
19697	开慧镇的	网友“UU008267” 您好！来信收悉。现将	属于职责范围	不相关	相关
12181	易书记领	网友：您好！留言已收悉	属于职责范围	不相关	并无回应问题，不相关
20765	拆迁人口	网友： 您好。您的留言已收悉，根据长县	属于职责范围	不相关	相关
33695	无良驾校、训	尊敬的问政网友0000-00000000您好：您反映的	属于职责范围	不相关	相关
37346	现在因为阳	网友：2019年7月9日	属于职责范围	不相关	并无回应问题，不相关
37459	请问，带小孩	2019年1月14日	属于职责范围	不相关	并无回应问题，不相关
37482	现在B9市的	2018年12月12日	属于职责范围	不相关	并无回应问题，不相关
39494	醴娄高速（沙	网友：您好！您反映的“醴娄高速进度如何”-	属于职责范围	不相关	相关
94701	现在用户	国网西地省J9县供电公司	属于职责范围	不相关	并无回应问题，不相关
140568	套牌车类	网友： 您好！您的留言转由我单位办理，现	属于职责范围	不相关	相关
144850	坚决反对破坏	你好！2019年6月13日	属于职责范围	不相关	并无回应问题，不相关
119660	尊敬的冯书记	网友：您好！留言已收悉	属于职责范围	不相关	并无回应问题，不相关
125462	我是华都	L1区网宣办 2015年9月18日	属于职责范围	不相关	并无回应问题，不相关

四、 结论与建议

对“智慧政务”平台的留言信息进行分析研究，及时获悉社情民意，对功能机关解决问题，缓解纠纷起到重要的作用。传统的人工处理已经不能满足数据量越发庞大的留言信息。本文主要运用逻辑回归，线性支持向量机，gensim 库和 SimHash 模型，建立留言分类模型，热点问题评定模型和答复评价模型。

对于答复评价模型中所存在的缺陷，为减轻工作人员查看并修改答复的负担，提高留言评价模型的准确率，下面是对“智慧政务”平台和工作人员的一些建议：

（1）“智慧政务”平台在留言详情区域应作出提示，引导市民较详细的叙述问题或需求，从而减少模型由于文本长度相差过大导致的相关性检验误判。

（2）工作人员对留言进行答复时，尽量运用留言中所提及的主体进行答复，避免由于说法不同导致的相关性检验误差；充分运用每阶段的正确答复，定期进行完整性和可解释性词库的更新，从而保证完整性和可解释检验的正确率维持在较高的水平。

五、 参考文献

[1] CSDN博主「-派神-」. 使用python和sklearn的中文文本多分类实战开发，2019. https://blog.csdn.net/weixin_42608414/article/details/88046380

[2] 梁柯，李健，陈颖雪，等. 基于朴素贝叶斯的文本情感分类及实现[J]. 智能计算机与应用，2019（05）：150-153+15.

[3] 杨锋. 基于线性支持向量机的文本分类应用研究[J]. 信息技术与信息化，2020（03）：146-148.

[4] CHENYIBAI. 初识gensim，博客园，2019. https://www.cnblogs.com/chenyibai/p/10735417.

html
[5] 王添男，冯锋．基于SimHash的文本相似检测算法研究[J]．电子测试，2019（15）：87-89．