基于词频和 word2vec“智慧政务”中的文本挖掘应用

摘要

伴随着云计算、大数据、物联网、人工智能等信息技术的快速发展和政府的逐渐数字化转型，各类社情民意相关的文本数据量不断攀升。网络民情反馈等数据量呈现几何级增长。原本靠人工进行文本数据处理的工作形式不再适用，为解决群众留言分类、热点信息整理、留言答复评估三个问题，本文构建了基于补集朴素贝叶斯的分类模型和基于支持向量机的分类模型、基于 DBSCAN 的文本聚类模型、基于命名实体识别的地点人群提取模型、基于文本相似度的留言答复评价模型。

在数据预处理阶段，我们对文本进行 jieba 分词和去停用词处理。通过 TFIDF 对文本表征，并降低通用词的权重，最终得到文档的向量表征。

在分类问题上,我们用补集分布的朴素贝叶斯作为分类器,并通过 Platt 的 Sigmoid 模型来进行校准，从而得到提高模型的表现效果。另外，我们还应用了支持向量机作为分类器，为了找到一个可划分的超平面，我们增加了线性核函数，通过画学习曲线调线性核函数，返回效果最好的参数 C。两个模型的 F1-Score 都达到 90%以上，相比之下，支持向量机效果稍微好一些，但贝叶斯运行时间上快很多。

对于热点信息的整理，我们用 PCA 算法对样本数据进行降维处理，以更好地进行 DBSCAN 聚类，然后自定义调参，根据 KNN 分布与数学统计分析自适应计算出最优全局参数 Eps 与 MinPts，在样本维度少的部分数据实验中可以不需要人工干预，实现聚类过程的全自动化。但由于全部数据样本维度过高，最优参数有所偏差，我们在返回的参数附件调参，得到较好的聚类结果。针对每一类留言，我们制定了一套合理的热度指标计算方法,并利用 bi-LSTM-CRF 命名实体识别技术,提取出每一类留言中的地点与人群信息，组成热点问题表。

在留言答复评估上，为了抽取文本的语义信息，我们通过 word2vec 获得初步的词向量，用 Synonyms 相似度算法计算留言跟答复之间的文本相似性，根据文本相似度表现，划分等级。

关键字：智慧政务； 文本挖掘； TFIDF； word2vec； 贝叶斯； 支持向量机； dbscan； 命名实体识别

Abstract

With the rapid development of cloud computing, big data, Internet of things, artificial intelligence and other information technologies, as well as the gradual digital transformation of the government, the amount of text data related to various social situations and public opinions keeps increasing. The amount of data, such as the network people's situation feedback, is increasing geometrically. Original text data processing by artificial forms of work no longer applies, in order to solve the mass message classification, hot spot information, message reply to evaluate three problems, this paper constructs the complement of naive bayesian classification model and the classification model based on support vector machine (SVM), the text based on dbscan clustering model, based on the location of the named entity recognition crowd extraction model, message response evaluation model based on text similarity.

In the data preprocessing stage, we carry out jieba word segmentation and stop word processing on the text. TFIDF is used to represent the text and reduce the weight of common words, so as to obtain the vector representation of the document.
On the classification problem, we use the naive bayes of complement distribution as the classifier and calibrate through the Sigmoid model of Platt, so as to improve the performance of the model. In addition, we also use support vector machines as classifiers. In order to find a divisible hyperplane, we add a linear kernel function, which returns the best parameter C by drawing the learning curve. The f1-score of both models is above 90%. In comparison, support vector machines are slightly better, but the running time of bayes is much faster.

For the finishing of the information of the hot spots, we use PCA algorithm to dimension samples, in order to better to carry on the DBSCAN clustering, and then the custom, according to the distribution of KNN and mathematical statistical analysis of adaptive Eps and MinPts, calculate the optimal global parameters in the sample dimension less part of the data of experiments could not need human intervention, realize full automation in the process of clustering. However, due to the high dimension of all data samples and the deviation of the optimal parameters, we adjusted the parameters in the returned parameters to obtain a better clustering result. For each type of message, we developed a set of reasonable heat index calculation method, and used BiLstm-crf named entity recognition technology to extract the location and population information in each type of message, to form a hot issues table.

In the evaluation of message reply, in order to extract the semantic information of the text, we obtain the preliminary word vector by word2vec, and calculate the textual similarity between message and reply by Synonyms similarity algorithm, and divide the level according to the representation of text similarity.

Keywords: smart government affairs; text mining; TFIDF; word2vec; Bayesian; support vector machine; dbscan; named entity recognition

目录

1. 文本挖掘背景与问题重述分析 5

1.1 挖掘背景 5

1.2 问题重述 5

1.3 问题分析 5

1.4 挖掘流程 6

2. 文本预处理 6

2.1 分词 6

2.2 去停用词 6

2.3 基于词频统计的 TF-IDF 文本表征 6

3. 分类模型 7

3.1 基于补集朴素贝叶斯的分类模型 7

3.1.1 朴素贝叶斯 7

3.1.2 多项式贝叶斯 7

3.1.3 补集朴素贝叶斯 8

3.1.4 概率校准 9

3.2 基于支持向量机的分类模型 9

3.2.1 核函数 10

4. 基于 dbscan 的文本聚类模型 10

4.1 PCA 降维 10

4.2 DBSCAN 11

4.1 DBSCAN 自适应调参 12

5. 基于 BiLSTM-CRF 的命名实体识别模型 13

5.1 NER 与 CRF 13

5.2 RNN 与 LSTM 13

5.3 BiLSTM-CRF 15

5.3.1 BiLSTM 15

5.3.2 BiLSTM-CRF 16

6. 基于文本相似度的留言答复评价模型 17

6.1 相关性 17

6.2 完整性与可解释性 17

6.3 Word2vec 17
6.4 Synonyms 相似度算法 18
7. 实验评估 **18**
7.1 群众留言一级分类 18
7.1.1 实验数据 18
7.1.2 实验评估指标 18
7.1.3 调线性核函数 19
7.1.4 相关结果 19
7.2 热点整理 20
7.2.1 实验数据 20
7.2.2 热度指标定义 20
7.2.3 相关结果 20
7.3 答复评价 21
7.3.1 实验数据 21
7.3.2 评价方案 21
8. 参考文献 **21**

1. 文本挖掘背景与问题重述分析

1.1. 挖掘背景

近年来，微信、微博、市长信箱、阳光热线等网络问政平台逐步成为了政府了解民意、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。面对不断实时更新的海量数据，工作人员需要对群众留言进行分类，以便后续将留言分派至相应的职能部门处理；还需要进行热点排序，发现某一时段内群众集中反映的问题，以便相关部门进行有针对性的处理，提高服务效率。有时，为检查对群众留言问题的解答情况，工作人员需要浏览群众问题和对应的答复意见，根据经验做出评价。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大，效率低，且差错率高等问题。

这种主要依靠人工来进行留言划分、热点整理、答复评价的工作形式不仅耗人力，更是耗时间，而且不能保证获取信息的实时性，评价的客观性。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. 问题重述

1） 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 对模型进行评价。

2） 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，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

3） 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，并尝试实现。

1.3. 问题分析

1) 一级分类标签共有七个，而且附件 2 中的每条数据已经打好标签，所以需要构建一个有监督的中文文本多分类模型。数据中有留言主题和留言详情两部分文本信息，留言主题文本很短，噪声较少，留言详情大部分比较长，但还是属于短文本，噪声较多，所以可以尝试只用留言主题或两者结合作为训练集，对比两者的分类效果。主要存在的困难是样本不平衡。

2) 附件 3 的数据没有标签，需要用聚类算法，地点人群的提取则可以使用命名实体识别技术，再进行提取，难点在于聚类算法的调参和提取的信息表达式多样化，很难直接是我们想要的。

3) 附件 4 是留言详情及回复详情，相关性可以理解为答复意见的内容是否与问题相关，

完整性可以理解为是否符合规范，可解释性可以理解为答复意见中的相关解释。我们可以基于文本相似度做文本挖掘。难点在于文本的语义分析。

1.4. 挖掘流程

文本挖掘分为三个过程，预处理、特征工程、构建模型。其中，预处理包括分词，去停用词，基于词频统计的 TF-IDF 文本表征。由于文本的向量表示是高维的稀疏矩阵，所以用 PCA 进行降维。最后是针对三个问题构建不同的模型：基于补集朴素贝叶斯的分类模型和基于支持向量机的分类模型、基于 dbscan 的文本聚类模型、基于命名实体识别的地点人群提取模型、基于文本相似度的留言答复评价评价模型。

2. 文本预处理

2.1. 分词

中文文本的词与词之间没有像英文词间的空格一样明显的界限，因此，需要有专门的分词工具，本文采用 Python 中文分词组件——jieba 分词[1]。

jieba 分词是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[2]。

2.2. 去停用词

对于本文的“智慧政务”，停用词是指人类语言中包含的功能词，这些词在语言表达中应用极其普遍，与其他词相比，没有什么实际的含义。停用词主要包括英文字符、数字、数学字符、标点符号，使用频率特高的单汉字等。为过滤分词结果中的噪声，提高模型的效果，我们进行了停用词过滤。

本文采用将 github 上包括哈工大停用词、四川大学机器智能实验室停用词、百度停用词、中文停用以及最全中文停用词表（1893）进行整合得到新的中文停用词表[3]，总共 2462 个。

2.3. 基于词频统计的 TF-IDF 文本表征

词频－逆文档频率（TermFrequency-InverseDocu- mentFrequency，TF-IDF），TF-IDF 分为两部分，TF 和 IDF，TF 是统计文档中各个词出现的频率，突出该词在文档中的重要程度，但可能是没有意义的高频词；IDF 的大小与一个词的常见程度成反比，这个词越常见，编码后为它设置的权重会倾向于越小，与词频相乘，可以来压制频繁出现的一些无意义的词的重要程度。

根据 TF-IDF=TF*IDF 计算得出每篇文档所有词的 TF-IDF 值。

设文本库共有|D|篇文档，将所有文档中的不同词语构成一个词库 M，词数为|M|，则每篇文档对应一个 M 维向量，文档向量中每个维度的值为 TF-IDF 值或 0。整个文本库可以构成一个|D|×|M|矩阵，得到文本向量表征。

3. 分类模型

3.1. 基于补集朴素贝叶斯的分类模型

3.1.1. 朴素贝叶斯

首先了解一下贝叶斯理论等式：

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}$$

这个式子，就是一切贝叶斯算法的根源理论。可以把特征 X 当成是条件事件，而要求解的标签 Y 当成是我们被满足条件后会被影响的结果，而两者之间的概率关系就是 $P(X/Y)$，这个概率在机器学习中，称为是标签的后验概率（posterior probability），即是说我们先知道了条件，再去求解结果。而标签 Y 在没有任何条件限制下取值为某个值的概率，被我们写作 $P(Y)$，与后验概率相反，这是完全没有任何条件限制的，标签的先验概率（prior probability）。而我们的 $P(X/Y)$ 被称为“类的条件概率”，表示当 Y 的取值固定的时候，X 为某个值的概率。

朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y 。[4]

朴素贝叶斯基于各特征之间相互独立，在给定类别为 y 的情况下，可以进一步表示下式：

$$P(X|Y = y) = \prod_{i=1}^{d} P(x_i|Y = y)$$

由以上两式可以计算出后验概率为：

$$P_{post} = P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(x_i|Y)}{P(X)}$$

由于 P(X)的大小是固定不变的，因此在比较后验概率时，只比较上式的分子部分即可。因此可以得到一个样本数据属于类别 y_i 的朴素贝叶斯计算公式：

$$P(y_i|x_1, x_2, \cdots, x_d) = \frac{P(y_i)\prod_{j=1}^{d} P(x_j|y_i)}{\prod_{j=1}^{d} P(xj)}$$

3.1.2. 多项式朴素贝叶斯

多项式朴素贝叶斯（multinomial naive Bayes，MNB）也是基于原始的贝叶斯理论，但假设概率分布是服从一个简单多项式分布。

在一种标签类别 $Y = c$ 下，我们有一组分别对应特征的参数向量 $\theta_c =$

$(\theta_{c1},\theta_{c2},\cdots,\theta_{cn})$，其 中 n 表示特征的总数。一个$\theta_{cn}$表示这个标签类别下的第 i 个特征所对应的参数。这个参数被我们定义为：

$$\theta_{ci}=\frac{特征X_i在Y=c这个分类下的所有样本的取值总和}{所有特征在Y=c这个分类下所有样本的取值总和}$$

对于一个在标签类别 Y=c 下，结构为(m, n)的特征矩阵来说，我们有：

$$X_y=\begin{bmatrix}x_{11} & \cdots & x_{1n}\\ \vdots & \ddots & \vdots\\ x_{m1} & \cdots & x_{mn}\end{bmatrix}$$

其中每个 x_{ji} 都是特征 X_i 发生的次数。基于这些理解，我们通过平滑后的最大似然估计来求解参数(α被称为平滑系数)：

$$\theta_{ci}=\frac{\sum_{y_j=c}x_{ji}+\alpha}{\sum_{i=1}^{n}\sum_{y_j=c}x_{ij}+\alpha n}$$

3.1.3. 补集朴素贝叶斯

补集朴素贝叶斯（complement naive Bayes，CNB）算法[5]是标准多项式朴素贝叶斯算法的改进。多项式分布擅长的是分类型变量，在其原理假设中，$P((x_i|Y)$的概率是离散的，并且不同下的相互独立，互不影响。CNB 一定程度上忽略朴素假设，一定程度上解决贝叶斯中的“朴素”假设带来的各种问题，算法能够不去关心所有特征之间是否是条件独立的并且能够解决样本不平衡问题。[5]

CNB 使用来自每个标签类别的补集的概率，并以此来计算每个特征的权重：

$$\hat{\theta}_{i,y\neq c}=\frac{\alpha_i+\sum_{y_j\neq c}x_{ij}}{\alpha_i n+\sum_{i,y\neq c}\sum_{i=1}^{n}x_{ij}}$$

其实就是多项式分布的逆向思路。

$$\omega_{ci}=\frac{\log\hat{\theta}_{i,y\neq c}}{\sum_j\left|\log\hat{\theta}_{i,y\neq c}\right|}$$

基于这个权重，补充朴素贝叶斯中一个样本的预测规则为:

$$P(Y\neq c|X)=arg\ \min_c\sum_i x_i\omega_{ci}$$

即求解出的最小补集概率所对应的标签就是样本的标签，因为$Y\neq c$的概率越小，则意

味着$Y = c$的概率越大，所以样本属于标签类别 c。

3.1.4. 概率校准

执行分类时，我们希望不仅可以预测类标签，还要获得相应标签的概率。这个概率增加一些预测的信心。校准模块可以更好地校准给定模型的概率，或添加对概率预测的支持，从而提高模型的表现效果。

本文对补集朴素贝叶斯进行概率校准，使用 sklearn 中的概率校正类 CalibratedClassfierCV，输入"sigmoid"和"isotonic" 。输入'sigmoid'，使用基于 Platt 的 Sigmoid 模型来进行校准；输入'isotonic'，使用等渗回归来进行校准。

3.2. 基于支持向量机的分类模型

支持向量机（support vector machines，SVM）[7]的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，$\omega \cdot x + b = 0$ 即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。[6]

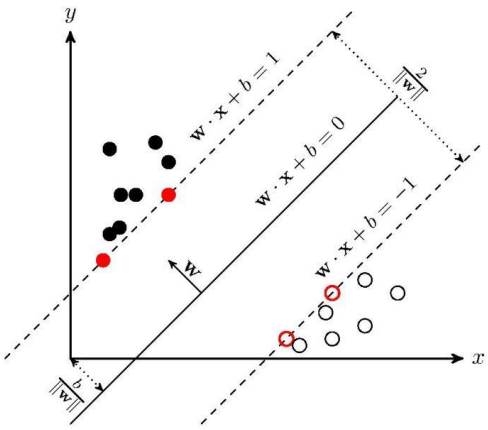

图 1 支持向量机

在样本空间中，用线性方程来表示划分超平面：$\omega^T x + b = 0$ ；其中ω = (ω1;ω2; … ; ωd)为法向量，决定超平面内的方向；b 为位移项，决定超平面与原点之间的距离。

则样本空间中任意点 x 到超平面的距离为：

$$\mathrm{r} = \frac{|\omega^T x + b|}{\|\omega\|}$$

假设超平面(ω, b)能够正确分类样本，即对$(x_i , y_i) \in$D，若y_i= +1，有$\omega^T x_i + b$ >0 ；若 y_i= -1，有$\omega^T x_i + b$ < 0. 令：

$$\begin{cases} \omega^T x_i + b \geq +1, & y_i = +1 \\ \omega^T x_i + b \leq -1, & y_i = -1 \end{cases}$$

两个异类支持向量到超平面的距离之和（也成为间隔 margin）为：

$$d = \frac{2}{\|\omega\|}$$

想要找到最大间隔(maximum margin)的划分超平面，也就是找使 d 最大时满足的约束参

数ω和b；也就是要最大化 ‖ω‖$^{-1}$，等价于最小化‖ω‖^{2} 。

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2$$
$$subject\ \ to\ \ y_i(\omega\cdot x_i+b)\geq 1,\qquad i=1,2,\cdots,N$$

3.2.4. 核函数（kernel functions）

如果不存在可以正确划分两类样本的超平面，我们可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。必须先对数据进行升维度，即将原始的x转换成$\Phi(x)$,则式（3.2.4）可写为：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2$$
$$subject\ \ to\ \ y_i(\omega\cdot\Phi(x_i)+b)\geq 1,\qquad i=1,2,\cdots,N$$

核函数是一种能够使用数据原始空间中的向量计算来表示升维后的空间中的点积结果的数学方式。具体表现为，$\mathrm{K}(u,\ v)=\Phi(u)\cdot\Phi(v)$ 。而这个原始空间中的点积函数$\mathrm{K}(u,\ v)$ 就是核函数。

有了核函数之后，我们无需去担心Φ究竟应该是什么样，因为非线性 SVM 中的核函数都是正定核函数 (positive definite kernel functions)，他们都满足美世定律(Mercer's theorem)，确保了高维空间中任意两个向量的点积一定可以被低维空间中的这两个向量的某种计算来表示（多数时候是点积的某种变换）。

本文使用线性核函数，降低了计算的复杂度，在原始空间中进行，避免了维度诅咒的问题。并且通过画学习曲线，调线性核函数，设置重要参数 C（参数 C 用于权衡”训练样本的正确分类“与”决策函数的边际最大化“两个不可同时完成的目标，希望找出一个平衡点来让模型的效果最佳)。

4. 基于 dbscan 的文本聚类模型

4.1. PCA 降维（Principal Component Analysis）

在高维数据中，必然有一些特征是不带有有效的信息的（比如噪音），或者有一些特征带有的信息和其他一些特征是重复的（比如一些特征可能会线性相关）。我们希望能够找出一种办法来帮助 我们衡量特征上所带的信息量，让我们在降维的过程中，能够减少特征的数量，又保留大部分有效信息——将那些带有重复信息的特征合并，并删除那些带无效信息的特征等等——逐渐创造出能够代表原特征矩阵大部分信息的，特征更少的新特征矩阵。

如果一个特征的方差很小，则意味着这个特征上很可能有大量取值都相同，那这一个特征的取值对样本而言就没有区分度，这种特征就不带有有效信息。从方差的这种应用就可以推断出，如果一个特征的方差很大，则说明这个特征上带有大量的信息。因此，在降维中，PCA 使用的信息量衡量指标，就是样本方差，又称可解释性方差，方差越大，特征所带的信息量越多。

$$\mathrm{Var} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Var 代表一个特征的方差，n 代表样本量，x_i代表一个特征中的每个样本取值，$\hat{x}$代表这一列样本的均值。

PCA 算法步骤[8]

输入：样本集 X={x1,x2,...,xm}；

过程：

1. 对所有样本进行中心化：*xi←xi−1m∑mixixi←xi−1m∑imxi;*
2. 计算样本的协方差矩阵XX^T
3. 对协方差矩阵XX^T做特征分解；
4. 取最大的 *d* 个特征值所对应的特征向量 *w1,w2,...,wd.*

输出：投影矩阵 W=(w1,w2,...,wd).

在本文中，我们对需要做聚类的文本向量利用 PCA 进行降维处理，在保留其 95%的信息的前提下，减少了将近 60%的特征。

4.2. DBSCAN

DBSCAN（Density-Based Spatial Clustering of Applications with Noise） 聚类算法是一种经典的密度聚类算法[7]，其将一个高密度的区域分成多个簇，簇即是密度相连的点的最大集合，以该集合内的单位数目为标准，根据事先设置的阈值，将区域内的点划分为噪声点和核心点， 同时将核心点区域内的单位划分为边界点从而实现聚类。其中，DBSCAN 聚类算法的有关术语定义如下：

1) *Eps* 临域：对象 *Eps* 所包含的区域；

2) 核心点对象：在 *Eps* 邻域内含有不少于 MinPts 个数据点的对象称为核心点对象；

3) 边界对象：边界对象在 *Eps* 邻域内含有小于 MinPts 个数据点，但其落在其他核心点对象的 *Eps* 邻域内；

4) 密度可达：若某点 p 在 q 的 邻域内，且 q 是核心点则 p-q 直接密度可达。若有一个带你的序列 q0、q1、…、qk，对任意 qi-qi-1 是直接密度可达的，责成从 q0 到 qk 密度可达，这实际上是直接密度可达的“传播”。

举例说明，如图 2：A 表示核心对象、B 和 C 表示边界点以及 N 表示离群点。

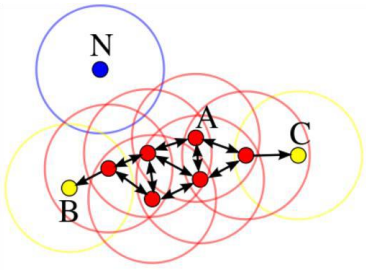

图 2

DBSCAN 聚类算法首先随机选择一个对象 p ，参照 *Eps* 和 MinPts 的值，提取从对象 p 密度可达的所有对象。若对象 p 为核心点对象，那么从对象 p 密度可达的所有对象被划归为当前类， 由它们进行下一步扩展；若对象 p 为边界对象，则将 p 看作噪声并忽略其存在，算法寻找下一个对象重复以上操作，直到生成一个完整的聚类。接着，算法再重新选择新的对象开始下一轮聚类，当所有的对象都被测试过后，算法终止。

4.3. DBSCAN 的自适应调参[9]

由于 DBSCAN 算法的全局参数 Eps 和 MinPts 的选取依赖于人工干预，对密度分布均匀的数据根据 k-dist 曲线升序排列后，人为选择曲线变化幅度开始陡升的点作为 Eps 参数，并且确定 MinPts 参数为固定常量 4，实施过程繁琐，依赖于人工干预。本文根据数据的距离空间的统计分布特性，统计出 k-dist 值的分布情况，再进行曲线拟合，拟合出分布曲线，通过计算拟合曲线的拐点处对应的值，自适应确定出 Eps 参数，并根据数据中每个点 Eps 领域内点数的分布情况，计算出参数 MinPts。以下是本文在部分数据上的测试效果。

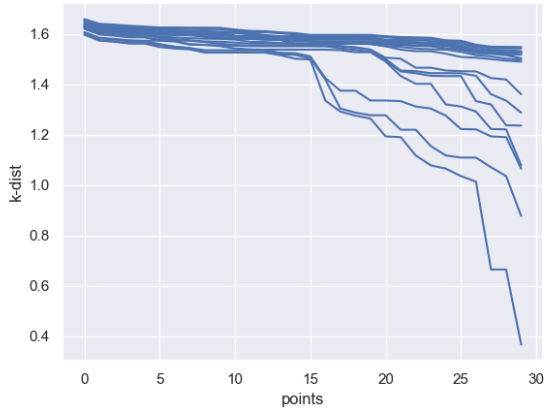

图 3　KNN 图

根据 k-dist 分布曲线可以看出，k=4 的 $dist_4$ 曲线可以反映出其他 $dist_k$ 曲线的形状。本文选择 k=4 的 $dist_k$（k-最近邻距离）的数据进行统计分析，统计 $dist_4$ 的概率分布：

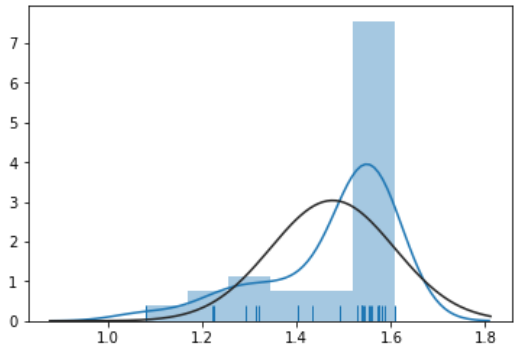

图 4　4-dist 概率分步和高斯拟合

由图 3 可以发现任何一条曲线都是在平缓变化后下降，Distk 中大部分值落在一个比较密集的区域内，因此可以判断 Distk 中大部分值应落在一个比较密集的区域内(曲线平缓段)，如果可以通过数学方法找出 distk 中平缓变化到急剧上升处的点或者 distk 概率分布最为密集的区域，可确定扫描半径参数 Eps，所以本文选择图 3 中 distk 拐点处的值为 Eps。

根据 KNN 升序排列曲线确定 Eps，对 $dist_4$ 曲线进行拟合，对于升序排列得到 KNN 分布数据，采用高斯拟合方法进行曲线拟合。然后计算数据对象的数学期望，即$MinPts$的值，如下式所示：

$$MinPts = \frac{1}{n}\sum_{i=1}^{n} p_i$$

其中，p_i 表示在点 i 的 Eps 邻域的点数。

5. 基于 BiLSTM-CRF 的命名实体识别模型

5.1. NER 与 CRF

命名实体识别，又称作专名识别，简称 NER，是自然语言处理中的一项基础任务，应用范围非常广泛。命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。NER 系统就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体，比如产品名称、型号、价格等。因此实体这个概念可以很广，只要是业务需要的特殊文本片段都可以称为实体。而在本文任务中，将命名实体识别技术应用于提取中文文本中的地点与人群。

在基于机器学习的方法中，NER 被当作序列标注问题。利用大规模语料来学习出标注模型，从而对句子的各个位置进行标注。NER 任务中的常用模型包括隐马模型 HMM、最大熵隐马模型 MEMM、条件随机场 CRF 等。HMM 不能考虑上下文的特征，限制了特征的选择，而 MEMM 由于在每一节点都要进行归一化，从而只能找到局部最优值并带来了标记偏见的问题。而 CRF 很好地解决了这些问题，它并不在每一个节点进行归一化，而是利用所有特征进行全局归一化，因此可以求得全局的最优值。CRF 是 NER 目前的主流模型。

另一方面，与 DL-softmax 模型相比，DL-CRF 模型将输出层面的关联分离了出来，显式地考虑了上下文关联，使得命名实体识别的效果大幅提高。因而在本文中，我们利用 CRF 作为我们的预测算法。

5.2. RNN 与 LSTM

循环神经网络（Recurrent Neural Network，RNN[10]）是一种特殊的神经网络结构，它是根据“人的认知是基于过往的经验和记忆”这一观点提出的。RNN 与深度神经网络（Deep Neural Network,DNN）的不同点在于，它不仅考虑前一时刻的输入，且赋予了网络对前面的内容的一种记忆功能，使 RNN 能够较好地表征上下文的语义。

相对于 DNN，RNN 在隐层上增加了一个反馈，使得一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，

即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一个时间步隐藏层的输出。标准 RNN 由简单的神经网络模块按时序展开成链式。这个重复模块往往结构简单且单一，如一个 tanh 层。

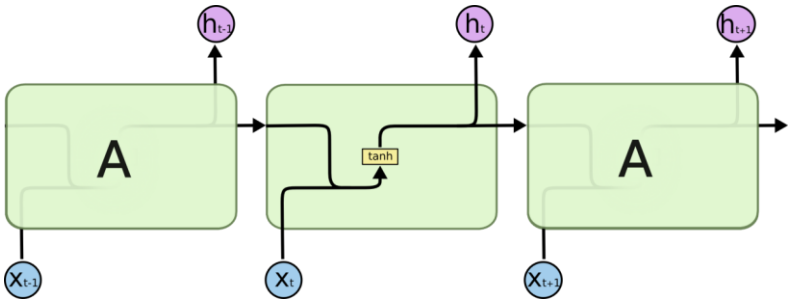

图 5　RNN 结构图

RNN 的优点在于能够将先前的信息连接到当前的任务，却也存在一些严重的缺点。首先是梯度消失和梯度爆炸，当激活函数 tanh 函数的导数在 0 到 1 之间，且参数 W 初始化为小于 1 的数时，多个 tanh′ *W 将导致求得的偏导极小，从而导致梯度消失；而当参数初始化足够大时，使得 tanh′ *W 大于 1，则将导致偏导极大从而导致梯度爆炸。其次是 RNN 对较长的问句有严重语义遗忘问题，当相关信息和当前预测位置之间的间隔不断增大时，RNN 有限的记忆能力会因为信息不断增多而耗尽，因此当信息越来越多，一部分初始的记忆信息就会被遗忘，这些长期依赖的遗忘和缺失最终会导致问句语义丢失。因而 RNN 在大型预料下进行训练时表现较差。为了避免以上问题，我们选择使用 LSTM。

长短期记忆网络（Long Short-Term Memory，LSTM[11]）是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸与严重语义遗忘问题，相比起普通的 RNN，LSTM 能够在更长的序列中有更好的表现。

在 RNN 的基础上，LSTM 引入了细胞状态，并使用输入门，遗忘门、输出门三种门来保持和控制信息。输入信息由当前序列的输入以及上一序列的输出组成，经由遗忘门层，符合算法的信息将会被留下，成为当前序列输出的一部分，而不符合算法的信息将会被遗忘。

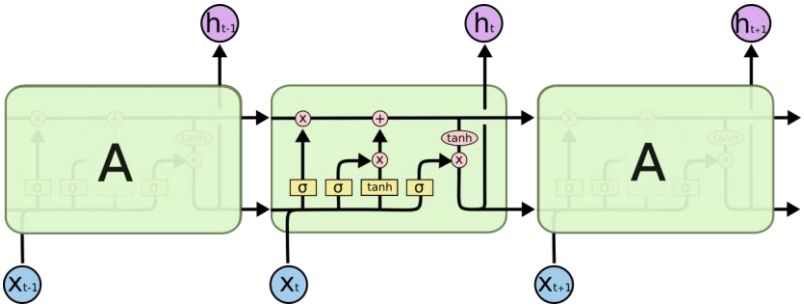

图 6　LSTM 结构图

遗忘门层：遗忘门结合上一隐藏层状态值h_{t-1}和当前输入x_t，通过 sigmoid 函数（σ），决定舍弃哪些旧信息。sigmoid 值域为(0, 1)，1 表示完全记住，0 表示完全忘

记。其公式为:

$$f_t = \delta\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

sigmoid 函数曲线如下图所示:

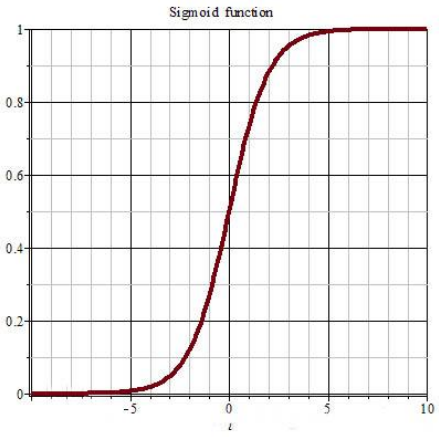

图 7　Sigmoid function

输入门层：输入门和 tanh 决定从上一时刻隐藏层激活值h_{t-1}和当前输入值x_t中保存哪些新信息，并得到候选值$\overline{c_t}$　。接下来，结合遗忘门和输入门进行信息的舍弃和保存，得到当前时刻的细胞状态c_t。

其公式为：

$$i_t = \sigma(W \cdot [h_{t-1}, x_t] + b_i)$$

$$\overline{c_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$c_t = f_t * c_{t-1} + i_t * \bar{c}_t$$

输出门层：输出门结合 tanh 决定h_{t-1}、x_t、c_t中哪些信息输出为本时刻的隐藏层状态h_t，并作为当前时间步的输出而输入到下一层。其公式为：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(c_t)$$

5.3. BiLSTM-CRF

5.3.1. BiLSTM

由于自然语言的语义复杂性，语句的重要信息可能出现在语句的任何地方，并且信息之间具有复杂的相互关联性。普通的 LSTM 可以解决 RNN 的遗忘问题，但它只能做到根据前文理解后文，而无法做到根据后文理解前文，这样在某些情况下，LSTM 仍然可能丢失部分重要语义。因而本文引进双向 LSTM（BiLSTM），BiLSTM 结合了从序列起点开始移动的 LSTM 和另一个从序列末端开始移动的 LSTM，其中正向和逆向的循环网络都由一个个 LSTM 单元组合而成。通过 BiLSTM 可以更好地捕捉双向的语义依赖。

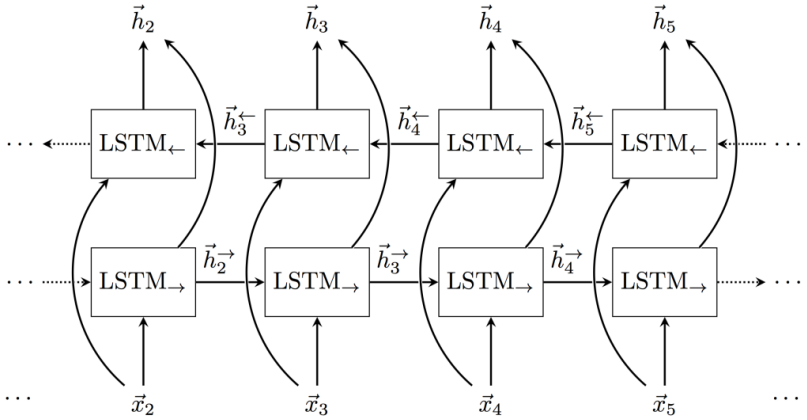

图 8　BiLSTM 结构图

5.3.2. BiLSTM-CRF

BILSTM-CRF 模型[12]主要包括两部分，即 BiLSTM 层和 CRF 损失层。如下图所示。对于一个输入句子 X，首先经过 Embedding 层将每个词汇或者字符映射为一个词向量或者字符向量，然后传入 BiLSTM 层，获得句子的前向和后向向量，接着将前向和后向向量进行拼接作为当前词汇或字符的隐藏状态向量，而该向量将作为 CRF 层的输入。CRF 层有一个状态转换矩阵参数，通过 CRF 层，不仅可以考虑当前输入的信息，还可以考虑前后标签的信息，从而使其在命名实体识别任务上表现较好。

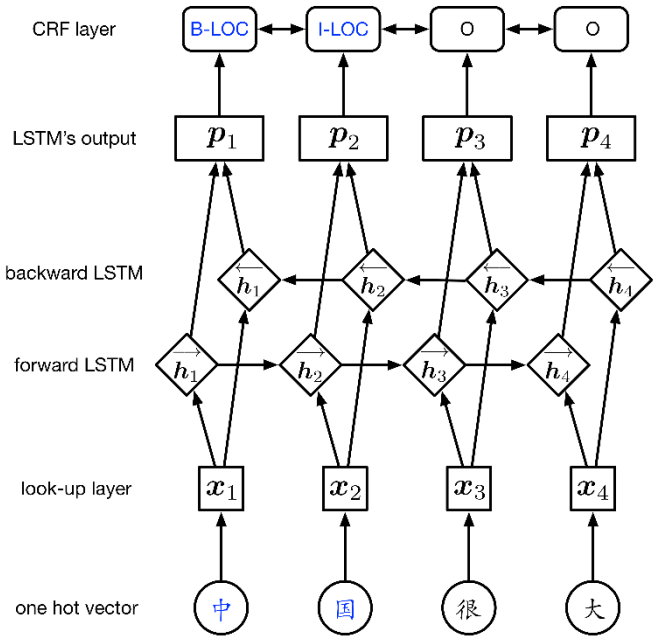

图 9　BiLSTM-CRF 结构图

FoolNLTK[13]是一个使用双向 LSTM 构建的中文处理工具包。在中文命名实体识别任务上，作者使用了字向量+BiLSTM+CRF 模型，不仅能够识别出实际存在的地名等实体，且对于本文任务中的如“A 市”、“G5 区”等指代地点也有较高的识别率。因而本文引入

FoolNLTK 工具包，作为提取数据中留言主题中的地点与人群的工具。

6. 基于文本相似度的留言答复评价模型

6.1. 相关性

计算文本相关性，即度量文本的相似程度，又称作文本相似度问题。在传统的统计学方法中，计算文本相似度就是对句子向量间距离的一种度量，如计算欧氏距离，曼哈顿距离，余弦距离等。距离越小，文本间的相似度就越大，反之距离越大，则文本间的相似度越小。此外，在句子做向量化的过程中，句子中关键词会占有该句子向量比较大的权重，因而文本相似度问题又可以解释为关键词匹配问题。即假设存在句子 A 与 B，当 A 中的大部分关键词在 B 中存在相同词或近义词时，我们可以认为句子 A 与 B 是相似的。

基于以上思想，在此前的实验中，我们实现了多种计量文本相似度的算法，如 BM25 算法、SimHash 算法、WMD 算法等，但以上算法均表现不佳。一方面是由于文本的相似度不仅取决于其关键词，与文本的语义也有较大关联，而以上算法没有对文本语义进行表征；另一方面，针对于本文的长文本任务，过多的噪声在算法对文本进行向量化时具有较大影响，导致其不能很好地表征句子向量，甚至在计算短句子与长句子的文本相似度时起到了反效果。

Word2Vec 对句子向量进行表征时，能够保留一定程度上的语义信息，带有语义上相近的词汇对应的向量更接近，从而得到我们所需要的句子向量。因而在本文中，我们利用 word2vec 训练句子向量，并计算其编辑距离与余弦距离，赋予不同距离不同的权重，最终得到在[0，1]之间的相似度分数。越接近 0 则越不相关，越接近 1 则越相关。通过测试，与 BM25 算法、SimHash 算法、WMD 算法等算法相比，在本文任务上该算法在效果上有了巨大的提升。

6.2. 完整性与可解释性

针对本文任务中的“答复意见”，我们将其完整性理解为在“答复意见”中，是否有对“留言详情”中提出的每个问题做出解答；同理，我们将其可解释性理解为“答复意见”是否对“留言详情”中提出的问题做出符合实情的详细解答。因而，不论是完整性或是可解释性，都能够在带有语义的文本相似性计量中有所体现。通过测试我们发现，“答复意见”中没有对“留言详情”中问题做出正面回答的，以及“答复意见”中缺少正文的信息，在算法计算所得的相似度分数中，其分数均很低，这正验证了我们的假设。

6.3. Word2Vec 原理

当前较为常用的词向量表示方法是 One-hot Representation，即独热编码，这种方法所表示的向量会过于稀疏，并造成维度灾难，且利用独热编码表示的任意词之间是相互独立的，无法表示语义层面上词汇之间的相关信息。词嵌入（Word Embeddings）是对独热编码方式的一种改进技术，在对词汇进行向量化时，能够将一些信息嵌入到词汇向量里，使其能表现出更多的信息。Word2Vec[14]是当今较为主流的词嵌入技术，Word2Vec 包含两种训练模型：

C-BOW 模型：即 Continuous Bag of Words Model，是利用已知的上下文词汇来预测当前词汇出现概率的模型；

Skip-Gram 模型：与 C-BOW 模型相反，是利用已知当前词汇来预测上下文词汇的模型。D-BOW 与 Skip-Gram 模型框架如下图所示。C-BOW 对小型数据库比较合适，而 Skip-Gram 在大型语料中表现更好。

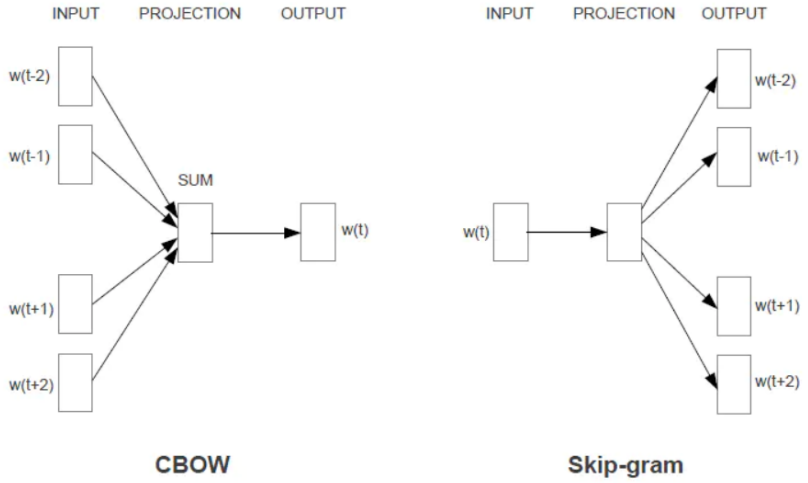

图 10　CBOW 和 Skip-gram

6.4.　Synonyms 相似度算法

Synonyms[15]是一个开源的中文近义词工具包，可用于如文本对齐、推荐算法、相似度计算、关键字提取等多种 NLP 任务。该中文近义词工具包采用的基本技术是 Word2Vec，在计算文本相似度时，其结合了编辑距离与余弦距离，通过画出 PR 曲线与 ROC 曲线两种图像来确定编辑距离与余弦距离的最优加权组合，从而赋予了编辑距离 5 倍的权重，赋予了余弦距离 0.8 倍的权重。

编辑距离，又称 Levenshtein 距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。

余弦距离是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。

7. 实验评估

7.1.　群众留言一级分类

7.1.1. 实验数据：

出题方 C 题全部数据附件 2.xlsx，顺序打乱，训练集：测试集=7:3。

7.1.2. 实验评估指标

对于本文中的模型，我们采用 F1-Score、准确率（accuracy）来评价我们模型的表现效果。详细定义如下：为了方便后面符号的说明定义一个混淆矩阵，如表 2 所示：

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP Type I
	Negative	FN Type II	TN

表 1 混淆矩阵

- TP(True Positive)：正类项目被判定为正类
- FP(False Positive)：负类项目被判定为负类
- FN(False Negative)：正类项目被判断为负类
- TN(True Negative)：正类项目被判断为负类

（1） 准确率(Accuracy)：准确率是指预测正确的样本所占比例

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

（2） F1-Score：准确率和召回率的调和平均值

$$F1 = \frac{2PR}{P + R}$$

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}$$

7.1.3. 调线性核函数

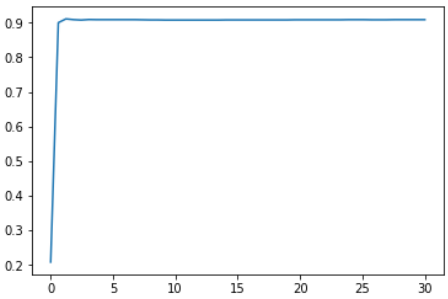

图 11 调线性核函数学习曲线

返回效果最好的重要参数 C。

7.1.4. 相关结果

	Complement+Isotonic	SVM+kernel linear	Bert
Accuracy	0.908	0.911	0.888
F1-Score	0.906	0.908	0.888

表 2 不同模型的分类效果

由于 TF-IDF 不考虑词义和语义，为了更好地理解上下文，我们还尝试了基于 attention 机制的 bert 模型，直接将文本做为训练数据，但 bert 表现效果没有很好。由表 2，基于支持向量机的分类模型效果最好，准确度达到 91.1%，但贝叶斯分类速度快很多，是 ms 级的，而支持向量机是 min 级的，尤其是调线性核函数时。所以总体来说，贝叶斯是个不错的选择。

7.2. 热点整理

7.2.1. 实验数据：

出题方 C 题全部数据附件 3.xlsx

7.2.2. 热度指标定义

某一时段内群众集中反映的某一问题可称为热点问题，及时发现热点问题，有助于相关部门进行有针对性的处理，提升服务效率。群众对某问题的关注程度称为热度值，且热度值与留言条数、点赞数、反对数等因素具有正相关性。考虑到时间跨度越大，说明存在的问题迟迟未得到解决，需要有关部门增加关注，我们也将留言时间跨度视为热度指标的正相关因素。本文制定了一条热度值的计算公式，且因每一因素对热度值的贡献度不同，赋予不同因素不同的计算权重。热度值计算公式如下：

假设热度值为 heat_values,留言条数为 users,点赞数为 agree,反对数为 oppose，时间跨度以 1 个月为单位，记为 time，我们将 heat_values 等同于留言条数,将其他因素折算成留言条数，则

$$heat\_values = users + agree * 0.01 + oppose * 0.01 + time$$

考虑到同类留言中有大于一条的留言可能是由相同用户在不同时间下所留，故赋予重复留言在原有权重下的一半权重。

7.2.3. 相关结果

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	40.58	2019/11/2至2020/1/26	丽发新城小区	A市万家丽南路丽发新城居民区附近搅拌站扰民
2	2	36.98	2019/7/7至2019/9/1	伊景园滨河苑	关于伊景园滨河苑捆绑销售车位的维权投诉
3	3	31.51	2018/11/15至2019/12/25	A市	咨询A市人才购房补贴政策
4	4	21.91	2019/8/19至2019/8/19	A市A5区汇金路五矿万境K9县	A市A5区汇金路五矿万境K9县存在一系列问题
5	5	18.61	2019/4/11至2019/4/11	反映A市金毛湾配套入学	反映A市金毛湾配套入学的问题

表 3　热度前五表单

7.3. 答复评价

7.3.1. 实验数据：

出题方 C 题全部数据附件 4.xlsx

7.3.2. 评价方案

通过 Synonyms 的相似度算法，能够得到在[0,1]之间的文本相似度分数，记为 score。因为该算法基于 Word2Vec 技术，在对文本进行向量化时保留有一定程度上的语义信息，因而我们同时将此分数作为文本完整性与可解释性的衡量分数。

通过测试，我们观察到，当 score<=0.25 时，文本不相关，且“答复意见”没有对“留言详情”中的问题做出正面回答，或者缺少回复正文，我们将其界定为等级“差”；当 0.25<score<=0.4 时，文本部分相关，“答复意见”对“留言详情”中的问题做出部分详细回答，我们将其界定为等级“良”；当 score>=0.4 时，文本完全相关，且“答复意见”对“留言详情”中的问题做出详细且全面的回答，我们将其界定为等级“优”。

	评价等级
Score<=0.25	差
0.25<score<=0.4	良
score>=0.4	优

表 4 评价方案

8. 参考文献

[1] https://github.com/fxsjy/jieba

[2] https://www.cnblogs.com/cyandn/p/10891608.html

[3] https://blog.csdn.net/qq_33772192/article/details/91886847

[4] 基于朴素贝叶斯的中文文本分类及 Python 实现[D]. 张航.山东师范大学 2018

[5] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

[6] https://zhuanlan.zhihu.com/p/31886934

[7]周志华. 《机器学习》[M].

[8]https://blog.csdn.net/ShiZhixin/article/details/51181379

[9]一种改进的自适应快速 AF-DBSCAN 聚类算法[J]. 周治平,王杰锋,朱书伟,孙子文. 智能系统学报. 2016(01)

[10] Mikolov T, Karafiat M,Burget L,et al.Recurrent neural network based language model[C].Interspeech.2010,2:3.

[11] Seep Hochreiter,Jurgen Schmidhuber.LONG SHORT-TERM MEMORY.Neural Computation.1997

[12] Zhiheng Huang,Wei Xu,Kai Yu.Bidirectional LSTM-CRF Models for Sequence

Tagging.Computation and Language.2015.
[13] https://github.com/rockyzhengwu/FoolNLTK
[14] Tomas Mikolov,Kai Chen,Greg Corrado,Jeffrey Dean.Efficient Estimation of Word Representations in Vector Space.Computation and Language.2013