摘要

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。

针对问题一，对附件 1 的留言详情和留言主题共同进行数据预处理和模型构建，对留言详情进行分词操作，将特殊字符和影响语义的常用词删去，再对其进行特征选择，将具有代表性的特征抽取，并用向量文本表示，最后将数据集划分为训练集和预测集使用有监督学习的分类器，将留言划分为附件 1 的一级分类标签。采用了使用最广泛、分词速度最快 python 里的 jieba 库对文本分词，jieba 库分词深度最深，对地名、人名的分词效果很差，但由于本题地名等对分类影响较小，则结巴分词对分词效果影响不大。

针对问题二：对附件 2 的留言主题和留言详情进行命名实体识别，使用 HanLP 分词获取句子词性标注，将留言内容的时间、地点、人物、事件提取出来。将提取出来的句子进行相似度计算，距离相近的归为一类。最后根据点赞数和反对数对热点问题赋予热度指标。命名实体识别是自然语言处理应用中的重要步骤，它可以检测出实体边界，还检测出命名实体的类型，是文本意义的基础，本题我们采用其中 HanLP 的 python 接口方法将地名、人名等提取出来。

针对问题三：对附件 4 相关部门对留言的答复，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。在本题中 TF-IDF 算法将文本特征词项转换为向量后，为空间中的点，因此使用余弦相似度，它对绝对的数值不敏感，更多的可以用于留言详情中的关注点来区分相似度和差异，同时修正了投诉用户间可能存在的度量标准不统一的问题。

关键词：TF-IDF 算法 命名实体识别 Hanlp 分词 热点问题评价指标

英文摘要

With WeChat,Microblog,Mayor's Mailbox,Sunshine Hotline and other Internet platforms gradually becoming government platforms for understanding public opinion,As an important channel for gathering people's wisdom and spirit, the volume of text data related to various social situations and public opinions has been continuously increasing, giving the past the main reasons.The work of the relevant departments,which rely on manpower to divide messages and sort out hot spots,has brought great challenges.

In response to question 1, data preprocessing and model construction are jointly carried out on the message details and message topics of Annex 1,word segmentation is carried out on the message details, special characters and common words affecting semantics are deleted,then feature selection is carried out on the message details,representative features are extracted and expressed by vector text,finally the data set is divided into training sets and prediction sets, and supervised learning classifiers are used to divide the message into primary classification labels of Annex 1. Jieba library in python, which is widely used and has the fastest segmentation speed, is used for text segmentation. jieba library has the deepest segmentation depth and has a poor segmentation effect on place names and personal names. However, due to the fact that place names and the like have little impact on classification, stuttering segmentation has little impact on segmentation effect.

In response to question 2, name entity identification is carried out on the message subject and message details in attachment 2,sentence part-of-speech tagging is obtained using Hanlp word segmentation,and the time, place,person and event of the message content are extracted.Calculate the similarity of the extracted sentences,and classify the sentences with similar distance into one class.Finally,according to the number of likes and dislikes, heat index is given to hot issues.Named entity recognition is an important step in the application of natural language processing.It can detect entity boundaries and types of named entities.It is the basis of text meaning.In this topic, we use python interface method of Hanlp to extract place names and names.

In response to question 3: to the reply of the relevan departments in annex 4 to the message,a set of evaluation plans are given for the quality of the reply comments from the angles of relevance, completeness and interpretability.In this topic, TF-IDF algorithm converts text feature words into vectors, which are points in space. Therefore, cosine similarity is used, which is not sensitive to absolute values.It can be used for more attention points in message details to distinguish similarities and differences, and at the same time,it corrects the problem that measurement standards may not be uniform among complaining users.

Key words: TF-IDF algorithm named entity recognition Hanlp segmentation hot topic eva luation index

目录

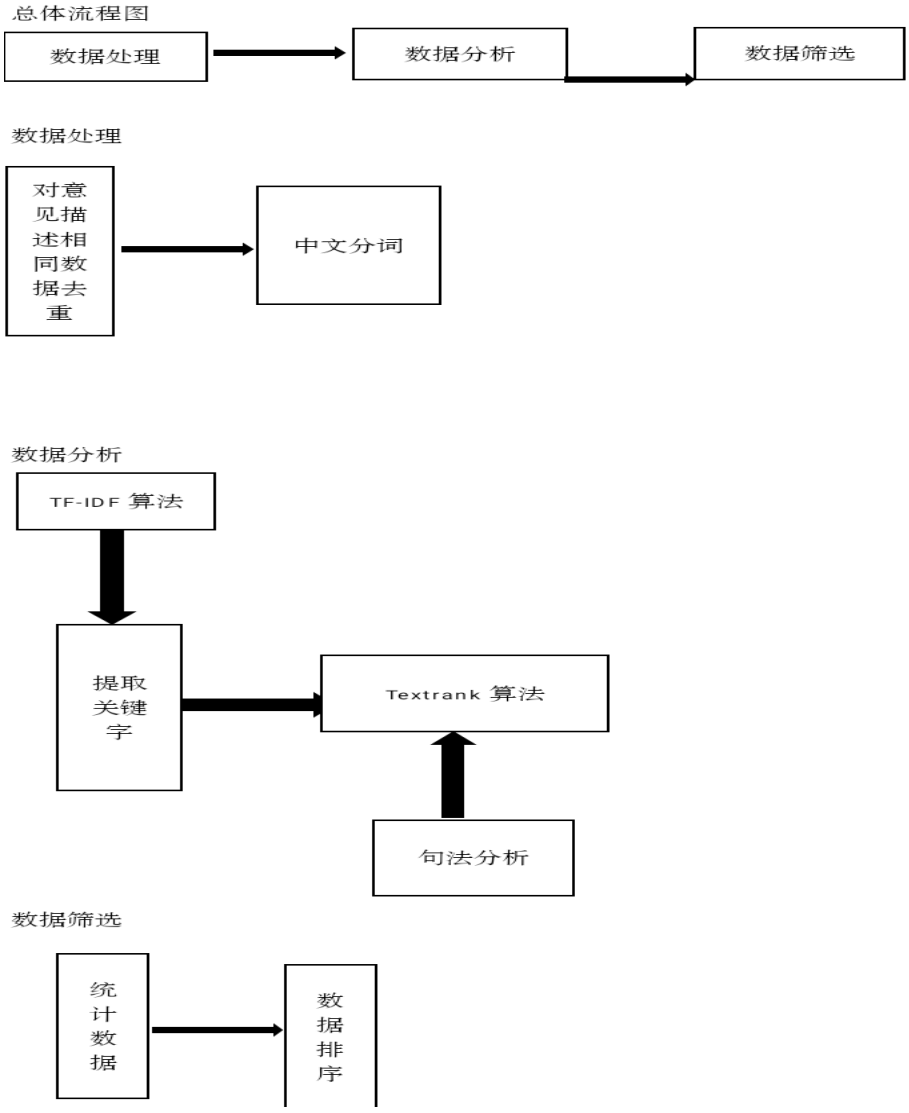
一、挖掘目标

大数据时代背景下，加快推动智慧政务系统建设已经是社会发展的新趋势，对提升政府办事效率以及紧密联系政府与民众之间的关系都有极大助力作用。本文针对给出的群众留言记录以及相关部门对部分群众留言的答复意见。采用 Jieba 中文分词工具、线性支持向量机以及 K-Means 聚类等方法，来完成以下三个目标：

（1）利用 Jieba 中文分词和 TF-IDF 方法来对留言详情进行文本预处理，再运用线性支持向量机建立关于留言内容的一级标签分类模型，最后使用 F-Score 对模型进行评价。

（2）在对附件 3 进行数据预处理之后，通过 k-Means 聚类以及 Topsis 方法得到热点问题，同时给出评价结果。

二、总体思路与流程



三、群众留言分类

3.1 文本数据预处理

3.1.1 Jieba 分词

中文分词是中文文本处理的基本步骤，也是中文人机自然语言交互的基础模块，在进行中文自然语言处理时，通常需要进行分析。Jieba 分词算法使用了前缀此前高效的词图扫描，生成句子中汉字所有可能生成词情况所构成的有向无环图（DAG），再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。

（1）在进行 Jieba 分词之前，首先将所有数据一级标签与留言详情相对应，将一级标签用具体的数字代替，存在 category_id_df（如图 3-2 所示）中，便于之后的处理计算。针对附件 2 中的所有留言内容，先考虑到留言内容中存在非中文字符，故先进行去标点、去空和去英文数字的操作。

一级标签	category_id
城乡建设	0
环境保护	1
交通运输	2
教育文体	3
劳动和社会保障	4
商贸旅游	5
卫生计生	6

图 3-2 一级标签分类

（2）在完成以上的基础上，调用 Python 的中问问分词包 Jieba 来进行分词。为节省存储空间和提高搜索效率，对留言内容进行去停用词（出现频率高且对信息含量的词），例如我、的、了、呢等，最终得到分词的结果并存储在 data 中。随后利用 TF-IDF 算法来提取关键词。

本文利用 Jieba 分词得到的留言一级分类的结果，保存在附件。

3.1.2 TF-IDF 算法

TF-IDF（term frequency-inverse document frequency）是一种用于咨询检索与咨询探勘的常用加权技术，其中 TF 为词频、IDF 为逆向文档频率。其主要思想是：如果某个词或短语在一篇文档中出现的频率 TF 高，并且在其他文档中很少出现，则认为此词或短语具有很好的类别区分能力，适合用来分类；换言之，如果包含某个词的文档越少，IDF 越大，这说明该词具有很好的类别分区能力。

假设某个词为ω，则

$$TF_{\omega} = \frac{\text{在某一类中词}\omega\text{出现的次数}}{\text{该类中所有的词数目}} \qquad (1)$$

$$IDF_{\omega} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词}\omega\text{的文档数}+1}\right) \qquad (2)$$

分母之所以加 1，是为了避免分母为 0。

某一特定文档内的高词频率，以及该词语在整个文档集合中的低文档频率，会产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF - IDF = TF * IDF \qquad (3)$$

由于担心存在数据排序的影响，故打乱了数据的顺序，然后对数据按照 4:1 的比例划分训练集与测试集。本文通过 TF-IDF 方法对数据进行向量化，将每个词在该类文本的 TF-IDF 值来作为量化的结果，分别对训练集和测试集采取相同的操作，提取得到各个一级分类的得分前 100 个关键词，如图 3-3 所示。明显可以看到卫生计生的得分较高的关键词有：医院、医生、患者等；而旅游商贸的关键词有：旅游、游客、食品等；劳动和社会保障的关键词有：劳动、单位、工资等等；其他类别不再一一赘述。

图 3-3 一级分类关键词

3.2 一级标签分类模型

运用线性支持向量机对分类进行预测，随后与文件给出的类别进行对比，计算分类查准率与查全率，最后利用 F-Score 进行分类方法评价。除此之外本文还利用不同的预测模型（Bayes 预测、逻辑回归、多层感知机分类器）进行分类并进行分类评价，经过对此线性支持向量机模型的 F-Score 值最大，约为 0.9057。故本文只对线性支持向量机模型预测展开详细描述，其他方法将简单提及，代码详情请见附录。

支持向量机（support vector machine，SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 还包括核技巧，这使它称为实质上的非线性分类器。SVM 学习的基本思想上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。

线性支持向量机处理的是线性不可分的数据集。对于线性支持向量机的优化问题，就是在线性可分支持向量机的基础上加了一个松弛变量ξ。

分类超平面：

$$\omega^{*} \cdot x + b^{*} = 0 \qquad (4)$$

相应的决策函数为：

$$f(x) = \mathrm{sign}\left(\omega^{*} \cdot x + b^{*}\right) \quad (5)$$

优化问题：

$$\min \frac{1}{2}\|\omega\|^{2} + C\sum_{i}\xi_{i} \quad (6)$$

$$y_{i}\left(\omega^{t}\phi\left(x_{i}\right)\right) \geq 1 - \xi_{i} \quad (7)$$

$$\xi_{i} \geq 0, i = 1, \ldots, n \quad (8)$$

如图 3-4 是不同分类器运行时间曲线图，明显可以看到逻辑回归所需时间最长，而其余三种时间相差不大。由此可知线性支持向量不但 F-Score 值最大而且运行所需时间也相对较短。故选用线性支持向量机预测模型是可行且合适的。

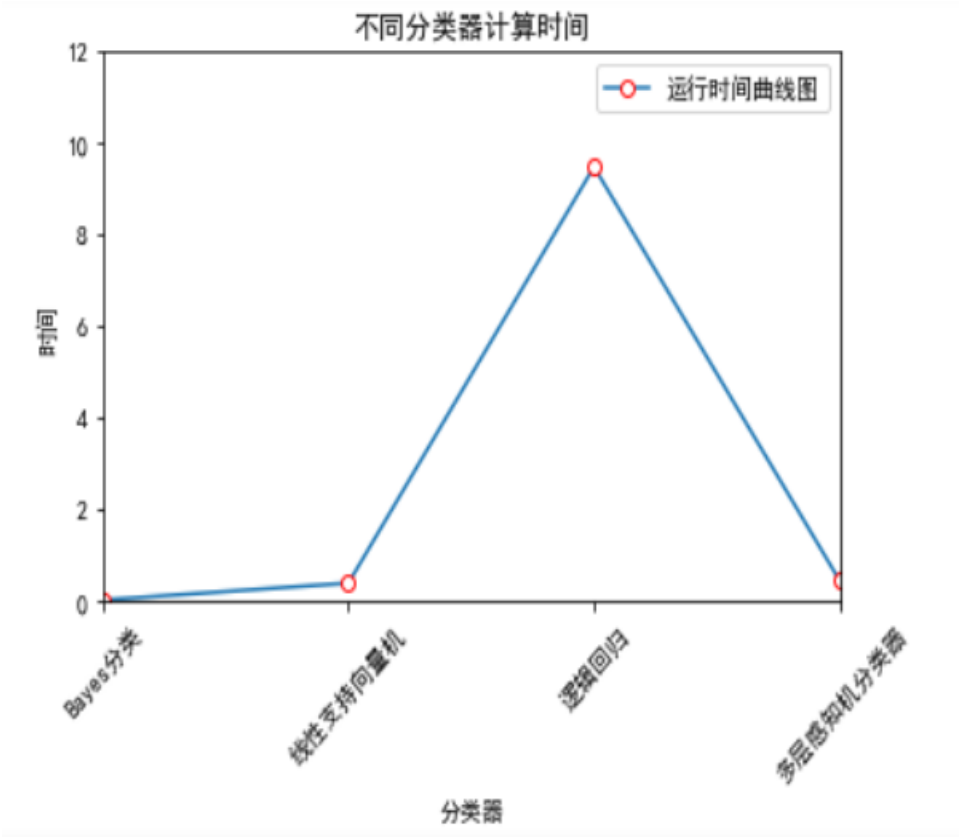


图 3-4 不同预测方法计算时间

3.3 分类结果评价

F-Score 是一种用于评价分类模型分类好坏的方法，其中其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。查准率是针对预测结果而言，计算的是所有“正确被检索的 item（TP）”占所有“实际被检索到的 item（TP+FP）”的比例。查全率是针对原样本而言，计算的是“正确被检索的 item（TP）”占所有“应该检索到的 item（TP+FN）”的比例。从某一类别发角度考虑，F1 值就是查准率和查全率的调和平均：$\frac{2}{F1}=\frac{1}{P}+\frac{1}{R}$。而从所有类别考虑，能写成 $F1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$ 的形式。

一般来说，查准率和查全率之间是矛盾的，引入 F-Score 作为综合指标，就是为了平衡准确率和查全率的影响，较为全面地评价一个分类器。

本文连续调用多个库模型，利用训练集数据来训练模型，并对测试集做一个预测。期间利用到了 Bayes 模型、线性支持向量机、逻辑回归模型和多层感知机分类器，并计算各个模型的 F-Score 值，最后得出线性支持向量机的 F-Score 值最大，值约为 0.9057。如下表 3-1 所示即为不同预测模型的 F-Score 值。而图 3-5 是不同模型的 F-Score 值的曲线图。

表 3-1 不同预测方法的 F-Score 值

预测方法	查准率	查全率	F-Score
Bayes 预测	0.8446	0.8426	0.8399
线性支持向量机	0.9076	0.9061	0.9057
逻辑回归	0.8794	0.8713	0.8706
多层感知机分类器	0.5107	0.5434	0.5070

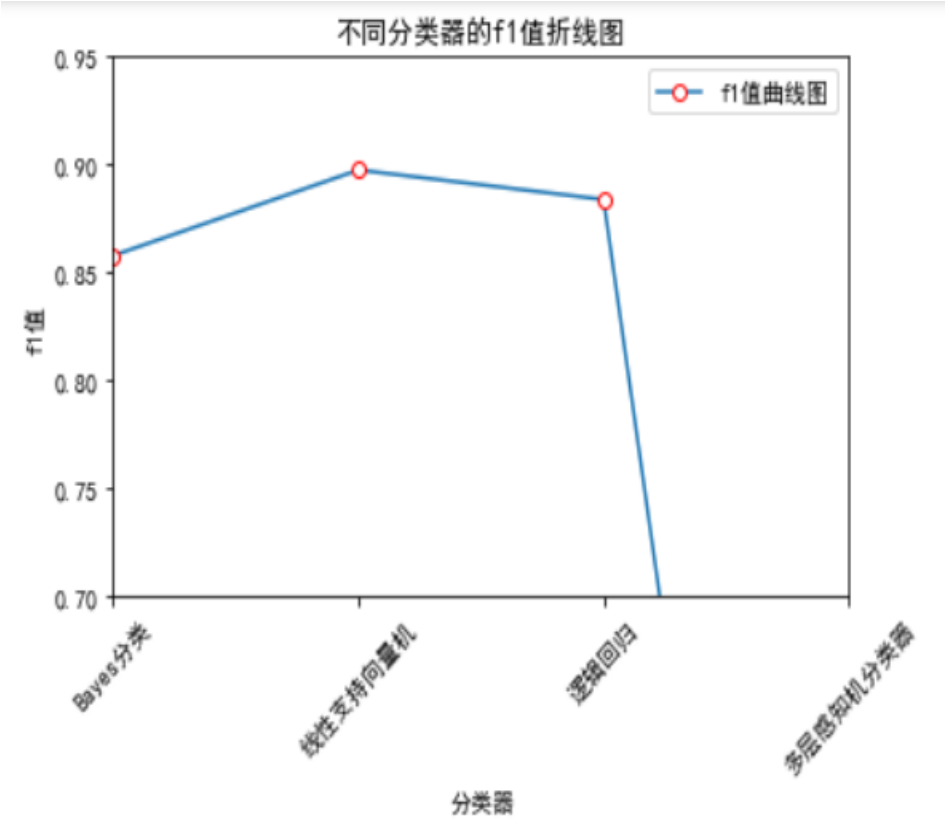


图 3-5 不同模型的 F-Score 值

四、热点问题挖掘

4.1 数据预处理

针对群众的留言详情,运用 Jieba 分词以及 TF-IDF 算法对数据进行向量化,具体原理与 3.1 数据预处理相同。但由于该题所涉及的数据量较大为方便计算对数据进行 TSNE 降维。

TSNE 是由 T 和 SNE 组成，也就是 T 分布和随机邻近嵌入，它是一种用于探索高位数据的非线性降维算法。它将多维数据映射到适合于人类观察的两个或多个维度。TSNE 主要步骤包括两个：第一，TSNE 构建一个高维对象之间的概率分布，是的相似的对象有更高的概率被选择，而不相似的对象有较低的概率被选择。第二，TSNE 在低维空间里在构建这些点的概率分布，使得这两个概率分

布之间尽可能的相似。

高位数据用 x 表示，x_i表示第 i 个样本，低维数据用 Y 来表示，则高维中的分布概率矩阵 P 定义如下：

$$p_{j|i}=\frac{\exp(-\|x_i-x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i-x_j\|^2/2\sigma_i^2)} \qquad (9)$$

低维中的分布概率矩阵计算如下：

$$q_{j|i}=\frac{\exp(-\|y_i-y_j\|^2)}{\sum_{k\neq i}\exp(-\|x_i-x_j\|^2)} \qquad (10)$$

随机给定一个初始化的 Y，进行优化，使得 Y 的分布矩阵逼近 x 的分布矩阵。给定目标函数，用 KL 散度来定义两个不同分布之间的差距：

$$C=\sum_i KL(P_i||Q_i)=\sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}} \qquad (11)$$

则可以计算梯度为：

$$\frac{\partial C}{\partial y_i}=2\sum_j(p_{j|i}-q_{j|i}+p_{i|j}-q_{i|j})(y_i-y_j) \qquad (12)$$

TSNE 对高维中的分布采用对称 SNE 的做法，低维中的分布则采用更一般的 T 分布，也是对称的，具体算法流程如下：

Algorithm1：Simple version of t-Distributed Stochastic Neighbor Embedding

---

Data: data set $x=\{x_1,x_2,...,x_n\}$

cost function parameters: perplexity Perp

optimization parameters: number of iterations T, learning rate η ,momentum $\alpha(t)$.

Result: low-dimensional data representation $Y^{(T)}=\{y_1,y_2,...,y_n\}$

begin

compute pairwise af nities $P_{j|i}$ with perplexities Perp (using Equation 9)

set $P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n}$

sample initial solution $Y^{(0)} = \{y_1, y_2, ..., y_n\}$ from $N(0, 10^{-4}I)$

for $t = 1$ to T do

compute low-dimensional af nities q_{ij} (using Equation 10)

compute gradient $\frac{\partial C}{\partial Y}$ (using Equation 12)

set $Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial C}{\partial Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$

end

end

经过 TSNE 降维数据维度降到？维

4.2 热点问题归类——K-Means 聚类[基于 K-means 算法的电力监控信息聚类研究]

K-Means 聚类算法属于无监督机器学习，是典型的基于原型的目标函数聚类方法的代表，它以数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则，其采用距离作为相似性评价的指标，即认为两个对象的距离越近，其相似度就越大。

K-Means 算法流程如下：

1.首先确定一个 K 值，即希望将数据集经过聚类得到 K 个集合；

2.从数据集中随机选择 K 个数据点作为质心；

3.对数据集中每个点，计算其余每一个质心的距离（如欧氏距离），离哪个

质心近，就划分到哪个质心所属的集合；

4.把所有数据归好集合后，一共有 K 个集合，然后重新计算每个集合的质心。

5.如果新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），可以认为聚类已经达到期望的结果，算法终止。

6.如果新质心和原质心距离变化很大，需要迭代 3~5 步。

K-Means 数学原理：

假设原始数据集合为$(x_1,x_2,...,x_n)$，并且每个人x_i为 d 维的向量，K-Means 聚类算法要把这 n 个数据对象划分到 k 个簇中$(k \le n)$，即找到使得下式满足的S_i：

$$\underset{S}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x-u_i\|^2 \qquad (9)$$

其中u_i是S_i中所有点的平均值。

$$u_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_i \qquad (10)$$

本文通过 K-Means 聚类把留言内容进行分类，由于不知道聚成多少类合适，因此从 2-2500 类分别计算得分。由于类别众多，在聚类前进行了 TSNE 数据降维，最后发现 1525 类的得分最高，故最后选择 1525 类对数据聚类。

4.3 热点评价指标——Topsis

Topsis 法亦称为理想解法，是一种有效的多指标评价方法。这种方法通过构造评价问题的正理想解和负理想解，即各指标的最优解和最劣解，通过计算每个方案到理想方案的相对贴近度，即靠近正理想解和远离负理想解的程度，来对方案进行排序，从而选出最优方案。

Topsis 方法和原理：

设多属性决策方案集为 $D=\{d_1,d_2,...,d_m\}$，衡量方案优劣的属性变量为，$x_1,x_2,...,x_n$，这时方案集 D 中的每个方案$d_i(i=1,...,m)$的 n 个属性值构成的向量是

$[a_{i1},...,a_{in}]$，它作为 n 维空间中的一个点，能唯一地表征方案d_i。

正理想解C^*是一个方案集 D 中并不存在的虚拟的最佳方案，它的每个属性值都是决策矩阵中该属性的最好值；而负理想解C^0则是虚拟的最差方案，它的每个属性值都是决策矩阵中该属性的最差值。在 n 维空间中，将方案集 D 中的各备选方案d_i与正理想解C^*和负理想解C^0的距离进行比较，既靠近正理想解又远离负理想解的方案集 D 中的最佳方案；并可以据此排定方案集 D 中各备选方案的优先序。

Topsis 法所用的是欧式距离。至于既用正理想解又用负理想解是因为在仅仅使用正理想解时有时会出现某两个备选方案与正理想解的距离相同的情况，为了区分这两个方案的优劣，引入负理想解并计算这两个方案与负理想解的距离，与正理想解的距离相同的方案离负理想解远者为优。

Topsis 法的具体算法如下：

（1）用向量规划的方法求得规范决策矩阵

设多属性决策问题的决策矩阵 $A=(a_{ij})_{m\times n}$，规范化决策矩阵 $B=(b_{ij})_{m\times n}$，其中$b_{ij}=a_{ij}\Big/\sqrt{\sum_{i=1}^{m}a_{ij}^2}$，$i=1,2,...,m$，$j=1,2,...,m$　　（11）。

（2）构成加权规范阵 $C=(c_{ij})_{m\times n}$

设由决策人给定各属性的权重向量为 $w=[w_1,w_2,...,w_n]^T$，则$c_{ij}=w_j\cdot b_{ij}$，$i=1,2,...,m$，$j=1,2,...,n$　　(12)

（3）确定正理想解C^*和负理想解C^0

设正理想解C^*的第 j 个属性值为c_j^*,负理想解C^0第 j 个属性值c_j^0，则

$$\text{正理想解}c_j^*=\begin{cases}\max\limits_i c_{ij}，\ j\text{为效益型属性},\\ \min\limits_i c_{ij}，\ j\text{为成本型属性},\end{cases}\quad j=1,2,...,n \qquad (13)$$

$$\text{负理想解}c_j^0=\begin{cases}\min\limits_i c_{ij}，\ j\text{为效益型属性},\\ \max\limits_i c_{ij}，\ j\text{为成本型属性},\end{cases}\quad j=1,2,...,n \qquad (14)$$

（4）计算各方案到正理想解与负理想解的距离

备选方案d_i到正理想解的距离为

$$s_i^*=\sqrt{\sum_{j=1}^{n}(c_{ij}-c_j^*)^2}，\ i=1,2...,m \quad (15)$$

备选方案d_i到负理想解的距离为

$$s_i^0 = \sqrt{\sum_{j=1}^{n}(c_{ij}-c_j^0)^2},\ i=1,2...,m \quad (16)$$

（5）计算各方案的排队指标值（即综合评阶指数）

$$f_i^* = s_i^0 / (s_i^0 + s_i^*),\ i=1,2,...,m \quad (17)$$

（6）按f_i^*由大到小排列方案的优劣次序。

本文选定的属性是留言用户数量、时间跨度、反对数、点赞数，但由于部分点赞数过高导致不同热点问题之间得分差异过大，故通过调节不同属性的比例来缓解误差。本文对于留言用户数量、时间跨度、反对数、点赞数的比例为 6:2:1:1，由此降低点赞数带来的影响，最终得到一个问题得分的排名。

4.4 评价结果

对于本文中的模型，我们采用精确率(Precision)、召回率(Recall). 两者的调和平均数 F1-Score.准确率(accuracy)、MRR、MAP 来评价我们模型的表现效果。以上指标的详细定义如下:为了放便后面符号的说明定义一个混淆矩阵，如表 3 所示:

	相关	不相关
被检测到的	TP	FP
未被检测到的	FN	TN

●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)精准率(Preisin)

是衡量某-检索系统的信号噪声比的种指标，即检出的相关 文献与检出的全部文献的百分比。

$$P = \frac{TP}{(TP + EP)} \quad (18)$$

(2)召回率(Recall)

召回率(Recall)是检索出的相关文档数和文档库中所有的相关文档数的比率，

衡量的是检索系统的查全率。

$$R = \frac{TP}{(TP + FN)} \tag{19}$$

(3)准确率(Accuracy)

正确率(Accuracy) 是指是指正确检索出的有关无关文档数，占文档库中所有文档数的比率。

$$A = \frac{TP + TN}{TP + FP + FN + TN} \tag{20}$$

(4)F1-Score

分类的 F1 值就是准确率和召回率的调和平均值，具体计算公式为:

$$F1 = \frac{2PR}{P + R} \tag{21}$$

(5)MRR

最后我们的系统会对回答的每-一个句子打一个分，然后设置阈值返回含有多个句子的集合，同时将得分高的结果考前返回，用 MRR 评判这个系统系统好坏就是看第一个正确答案的位置，第一个正确答案越靠前，MRR 评分越高，即:

$$MMR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{ranki} \tag{22}$$

其中,Q 为样本 query 集合,|Q| 表示 Q 中 query 个数, rank; 表示在第 i 个 query 中,第一个正确答案的排名。

(6)MAP

定义单个问题检测出的相关回答的准确率为 AP

$$AP = \frac{\sum_{k}^{|R|}\frac{rank_{k}}{ranl_{j}}}{|A|} \tag{23}$$

其中 A, R 分别为系统返回的相关回答和，其中正类项目的集合，AI,|R|表示 A 和 R 集合的大小，rank;和 rank;表示该条回答在 A 和 R 中的排名。MAP 则为对去一个查询集合中每条查询的 AP 值的平均，即

$$MAP = \frac{\sum_{k=1}^{|Q|} AP(K)}{|Q|} \quad (24)$$

其中，Q 为样本 query 集合，|Q| 表示 Q 中 query 个数。

4.4 实验设置及结果分析

4.4.1 相关参数

我们实现了基于 Attention-over- Attention 机制下的双向 LSTM 智能阅读模型，系统训练时相关参数设置如表 4 所示:

4.4.2 阈值确定

模型经过最后的 attention-over-attention 机制得到的是对应位置答句与问句的匹配程度 P,我们需要设置一个阈值 δ ,当 P> δ 时，判断该答句为有关回答，即输出为 1;反之，则判断该答句为无关回答，即输出为 0。同时实验来选取阈值 δ 。

由图 8 可以看出，权衡精准率和召回率，我们取 δ =0.5 作为最后的阈值。

表 4:系统训练时相关参数

参数	参数值
学习速率	0.4
学习速率减缓因子	0.99
最大梯度范数	5.0
批次大小	8
层大小	128

五、答复意见的评价

在处理相关部门对留言的答复意见时，评价其质量是十分重要的。我们可以从答复的相关性、完整性和可解释性等角度对答复意见进行评价，并建立一个评价方案，以确保答复的质量满足需求。

5.1 对问句语义提取

可以将知识库理解为一一个三元组组成的集合实体，实体关系,实体。智能的阅读系统就算是可以看作给定一一个实体 和一个联系，查看候选答句中是否包另一个实体。

因此可以首先基于语义分析的方法通过将实体和联系进行识别和标注将自然语言形式的问句转换为 lambda 表达式或依存组合语义树等逻辑表达形式。通过对这种逻辑表达形式，进行向量化表示，从而进一步加强，对语义的聚焦。

5.2 文本蕴含

文本蕴含(textual entaiment)是指文本对之间的指向关系，用符号 T 表示被蕴含的文本，用 H 表示被蕴含的文本，也就是假设。如果 H 意思能够从 T 中推导出来，那么就认为 T 蕴含 H。

我们也可以将文本蕴含的相关思想运用到智能阅读系统中,对候选答案进行重排序。因为可以假设在一-个已经存 在的阅读系统中返回的最佳的候选答案，虽然最佳候选答案可能不是正确答案,但是在大多数情况下正确答案位于返回的候选答案集合中。

因此可能创建一个蕴含对的集合,其中将系统返回的候选答案合并作为蕴含对的文本，将转换的问题作为蕴含对的假设。文本蕴含系统接下来依次应用在蕴含对上，那些能蕴含转换文本的候选蕴含对移到列表的顶端，非蕴含的则移动到底部。Harabagiu 和 Hickl!15]在问答系统上对文本蕴含计算进行了测试，系统的准确率得到提高。

5.3文本相似度计算

Ｗｏｒｄ２Ｖｅｃ 是一种用于生成词向量（ｗｏｒｄｅｍｂｅｄｄｉｎｇｓ ）的常用技术，它能够将单词映射到连续向量空间中的向量表示，使得语义上相似的单词在向量空间中距离较近。这种表示方式有助于计算机理解和处理自然语言的语义信息。

Ｗｏｒｄ２Ｖｅｃ 模型是由Ｇｏｏｇｌｅ 的Ｔｏｍａｓ Ｍｉｋｏｌｏｖ 等人在２０１３ 年提出的，它主要包括两种架构：连续词袋和Ｓｋｉｐ－ｇｒａｍ。

ＣＢＯＷ模型：ＣＢＯＷ模型的目标是根据上下文预测目标词汇。它将一个词的上下文作为输入，然后尝试预测该词。这种方法适用于大型语料库，并且对罕见词有很好的泛化能力。

Ｓｋｉｐ－ｇｒａｍ模型：Ｓｋｉｐ－ｇｒａｍ模型与ＣＢＯＷ相反，它从一个词汇中生成上下文。它的目标是根据给定的词汇生成周围的上下文。Ｓｋｉｐ－ｇｒａｍ在小型语料库中表现较好，并且能够更好地处理频率较低的词汇。

训练过程：

Ｗｏｒｄ２Ｖｅｃ 模型的训练过程通常基于大量的文本语料库。训练过程基本上是一个监督学习的任务，它尝试通过最小化预测词与实际词之间的距离来学习词汇的嵌入。

构建词汇表：首先，将语料库中的所有单词收集起来构建一个词汇表，每个单词都会有一个唯一的索引。

生成训练样本：对于ＣＢＯＷ模型，训练样本由上下文单词和目标单词组成；对于Ｓｋｉｐ－ｇｒａｍ模型，训练样本由目标单词和其上下文单词组成。

定义损失函数： 常用的损失函数是负对数似然损失函数，它用来衡量模型预测词与实际词之间的差距。

优化模型参数： 使用梯度下降等优化算法，调整模型的参数，使得损失函数最小化。

5.3 评价体系

相关性

答复的相关性指答复内容与留言主题之间的关联程度。一个高质量的答复应当与留言内容直接相关，能够准确解决提出的问题或反映的诉求。相关性评价可以通过比对答复内容与留言主题的关键词、关键信息等来进行。如果答复能够涉及留言提出的关键问题，并给出清晰、准确的解释或回答，那么可以认为答复具有较高的相关性。我们通过以下方法评价答复的相关性：

- 语义相似度计算：使用自然语言处理技术，如词向量模型或文本相似度算法，计算答复意见与留言内容之间的语义相似度。
- 关键词匹配：检查答复意见中是否包含留言中提到的关键词或关键短语以判断答复是否与留言内容相关。
- 相关性得分。基于语义相似度计算结果和关键词匹配情况，给出答复意见与留言内容相关性的评分，范围从0 到1 ，1 表示完全相关，0 表示完全不相关。

完整性

答复的完整性指答复内容是否全面、详尽地覆盖了留言提出的问题或诉求。一个高质量的答复应当包含对所有相关问题的回答或解决方案，并且应当提供足够的信息以满足留言者的需求。完整性评价可以通过检查答复内容是否涵盖了留言中提及的所有关键点和问题，并且是否提供了必要的背景信息和解决方案来进行。我们通过以下方法评价答复的完整性：

- 信息覆盖：检查答复中是否包含了解决问题所需的所有关键信息，包括解释问题原因、提供解决方案或建议、说明相关政策或法规等。
- 问题回答：确保答复中明确回答了留言中提出的问题或疑问，避免遗漏或回避关键问题。
- 完整性得分： 根据答复意见包含的信息数量和质量，给出答复的完整性评分，范围从0 到1 ，1 表示完全完整，0 表示不完整。

可解释性

答复的可解释性指答复内容是否清晰易懂，能够让留言者理解相关部门的立场、政策或行动。一个高质量的答复应当以简洁清晰的语言表达，避免使用过多专业术语或复杂的理论概念，以确保留言者能够理解并接受答复内容。可解释性评价可以通过留言者反馈或针对答复内容进行的用户调查来进行。我们通过以下方法评价答复的可解释性：

语言简洁度：检查答复中是否使用了清晰简洁的语言表达，避免使用过于专业化或复杂的术语，以确保群众易于理解。

背景信息提供：如有必要，提供相关背景信息或解释，帮助群众理解答复内容，并理解相关政策或法规的背景和影响。

相关性评价方案：

- 可解释性得分： 根据答复意见的语言简洁度和提供的背景信息，给出答复的可解释性评分，范围从0 到1 ，1 表示非常易懂，0 表示不易懂。

5.3 未来改进

由以上论述可知，对于输人的问题，我们的模型可以出色地完成将答案定位到所在句子的任务。下一 步，我们希望继续改进模型，实现更细粒度的答案提取。如果需要将答案定位到词语，我们可以去掉嵌人句向量的步骤，用 Bi-LSTM 得到的蕴含上下文信息的词向量进行 Attention- over-Attention 的计算。按列计算的注意力代表了问题中每个词选择答案的重要性，按行计算的注意力代表了文档中的每个词响应问题的重要性。正反注意力经过点乘，最终获取的注意力向量每个分量值含义为文档中每个词与问题的匹配分数。统计每个词对应分量值之和，即以该词作为答案时与问题的匹配分数之和,获得所有和值并排序，分数越高则该词是正确答案的可能性越大。

参考文献

[1]周靖力.基于文本数据和陆空通话数据处理的空管运行风险识别和分析,中国民用航空飞行学院 2017。
[2]姚尹雄，贺尚红.“文本型”数据处理方法及其实现研究,长沙交通学院 1995。
[3]田军霞.基于短文本处理算法优化的文本信息推荐系统的设计与实现,北京交通大学 2017。
[4]刘懿霆.基于维基百科的文本样本扩展方法及其应用研究,上海大学 2018。