基于语义的短文分类留言处理

摘要

21 世纪以来，信息技术发展迅速，微信、微博、市长信箱、阳光热线等网络问政平台逐渐成为政府了解民意、汇集明智、汇聚民气的重要渠道，各类社情民意相关的文本数据不断攀升，目前只是靠人工来将留言进行分类和将热点进行整理。因此利用自然语言处理和文本挖掘来系统分类处理大量的留言将会大大提高政府的工作效率，也对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题一：在处理问题前，我们先进行数据预处理，获取停用词表；读取文本文件进入代码中；文本分词处理并去除停用词表；备注文本信息；对备注信息提供分类；获得文本分类的七张照片。建立类别体系，将附件二中的留言编号、留言用户、留言主题和留言详情分别标记为n_1,n_2,c_1,c_2。然后根据文本特征建立文本模型。在建立文本模型的基础上，采用 F-score 对第一节建立的分类函数进行性能能评价以及分类函数的准确率的情况。

针对问题二：运用 KNN 算法，根据距离所选择邻居已经进行了正确的分类。具体算法可以描述如下： 确定最近邻的数量，也就是 K 值大小。K 值的选取对于分类算法的精确度有很大的影响，需要反复试验。 采用向量空间模型表示文本，形成文本特征向量。 计算测试文本与训练集中每个文本的相似度，通常使用余弦定理来计算相似度。

针对问题三：根据群众留言和工作人员的回复的相关性、完整性和可解释性的层次性，运用层次分析法建立了一个多级综合评价指标体系，并在此基础上用熵权确定指标权重，然后运用模糊综合评判模型对评价对象进行综合评价，得出一个答复中回复的相关性最重要，假如回复很完整和有解释性，但是跟群众所提问题无关的话，那么这个回复也是没有意义的；其次到可解释性，回复的内容除了跟所提内容相关之外还要具有一定程度的解释性，不然回复太过刻板、官方，群众无法理解，所作出的回复意义也不大，最后到完整性。但是三者的比重虽然有差异，但是差异比不大，说明在回复当中三者皆有才是回复最完美的时候。

关键词：层次分析法；F-score；KNN 算法；文本模型；模糊综合评判模型；

Abstract

Since the 21st century, information technology is developing rapidly, WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing gradually become the government to understand public opinion, pooling wisdom, we have important channel of bull, all kinds of public opinion text data related to rising, now just to leave a message by artificial classification and hot spots.Therefore, the use of natural language processing and text mining to systematically classify a large number of comments will greatly improve the working efficiency of the government, as well as the management level and efficiency of the government.

For problem 1: before dealing with the problem, we preprocessed the data to obtain the stop word list;Read the text file into the code;Text word segmentation process and remove stop word list;Note text information;Provide classification of remarks;Get seven photos of text categorization.Establish a category system and mark the message number, message user, message subject and message details in annex ii as.Then the text model is built according to the text feature.On the basis of establishing the text model, f-score is used to evaluate the performance and the accuracy of the classification function established in the first section.

For problem two, KNN algorithm is used to classify the neighbors according to the distance.The specific algorithm can be described as follows: determine the number of nearest neighbors, that is, the size of K value.The selection of K value has a great influence on the accuracy of classification algorithm, which needs to be tested repeatedly.Text is represented by vector space model to form text feature vector.Calculate the similarity between the test text and each text in the training set, usually using the law of cosines to calculate the similarity.

To question three: according to the correlation of the masses and staff reply message, integrity, and interpretability of gradation, using the analytic hierarchy process (ahp) to establish a multi-level comprehensive evaluation index system, and on the basis of the use of entropy to determine the index weight, then use fuzzy comprehensive evaluation model for integrated evaluation of the evaluation objects, get a reply to reply to the correlation of the most important, if the reply is complete and explanatory, but has nothing to do with the issues, so the reply is meaningless;Secondly, to the interpretability, the content of the reply should be interpreted to a certain extent in addition to the content mentioned. Otherwise, the reply is too rigid and official, which cannot be understood by the public, and the meaning of the reply is not big. Finally, it is complete.However, although there is a difference in the proportion of the three, the difference is not big, indicating that in the reply of all three is the most perfect time to reply.

Key words: analytic hierarchy process; F- score; KNN algorithm.Text model; Fuzzy comprehensive evaluation model;

一、研究背景与意义

21 世纪以来，信息技术发展迅速，微信、微博、市长信箱、阳光热线等网络问政平台逐渐成为政府了解民意、汇集明智、汇聚民气的重要渠道，各类社情民意相关的文本数据不断攀升，目前只是靠人工来将留言进行分类和将热点进行整理。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。 因此利用自然语言处理和文本挖掘来系统分类处理大量的留言将会大大提高政府的工作效率，也对提升政府的管理水平和施政效率具有极大的推动作用。

二、问题分析

1. 对群众留言分类，首先按照一定的划分体系对流言进行分类，以便后续将群众留言分派至相应的职能部门，我们首先采用 bert 模型进行分类，然后再采用 F-Score 来对所分类的情况进行评分。

2. 识别问题，在众多留言中识别出相似的留言；把特定地点或者人群的数据进行归并为一类问题，结果对应表 2；建立热度评价体系的定义和计算方法，对指标排名之后得出对应的表 1.

3. 针对相关部门的答复意见，从答复的相关性，完整性、可解释性等角度意见的质量给出一套评价方案，并尝试实现。

三、符号说明及定义

表一：符号说明

符号	定义
TPR	正样本中被喻为正的
FNR	正样本中被喻为错的
FPR	负样本中被喻为错的
TNR	负样本中被喻为对的
$Accuracy$	正确率
n_1	留言编号
n_2	留言用户
c_1	留言主题

c_2	留言详情
CI	一致性指标
CR	一致性比率

四、数据预处理

1. 获取停用词表；
2. 读取文本文件；
3. 文本分词处理并去除停用词表；
4. 备注信息；
5. 对备注信息提供分类；
6. 获得分类的文本照片。

五、模型建立

5.1 类别体系的建立

文本分类的实质是文本到类别的一个映射关系。本文将建立一个有效的、具备一定完备性的类别集合-微博类别体系。

留言编号是唯一的，是作为区别其他留言用户的唯一标识，在用户使用过程当中是不会发生改变的。因此本文将附件二中的留言编号、留言用户、留言主题和留言详情分别标记为n_1, n_2, c_1, c_2。

预处理阶段，首先对留言详情进行分类之前需要留言详情四个要素n_1, n_2, c_1, c_2分别进行分词以及停用词过滤处理，然后分别对四个文本要素提出出来的特征词用文本模型：

$$text(m) = \{m_1, m_2, ..., m_n\}$$

表示。

编写代码，在 pycharm 中运行得出的结果如下:

图 1 卫生计生

图 2 商贸旅游

图 3 劳动和社会保障

图 4 教育文体

图 5 交通运输

图 6 环境保护

图 7 城乡建设

5.2 基于 F-Score 对分类性能进行评价

第二节主要利用附件 2 的数据对第一节建立的分类函数进行性能能评价以及分类函数的准确率的情况。

F-Score 具体知识点：

1.准确率：

$$准确率 = \frac{正确分类的样本数}{总样本数}$$

2. 混淆矩阵

表二

	预测为正	预测为正
真实为正	TP	FN
真实为负	FP	TN

(1) $TPR = \frac{TP}{TP + FN}$

(2) $FNR = \frac{FN}{TP + TN}$

(3) $FPR = \frac{FP}{FP + TN}$

(4) $TNR \frac{TN}{FP + TN}$

(5) $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

(6) 精准查准率:$\frac{TP}{TP+FP}$

(7) 召回率/查重率：$\frac{TP}{TP+FN}$

5.3 基于语义短文分类的留言处理

本文在传统的 *KNN* 算法的基础上，继续深入 *KNN* 算法，从而在这个基础上提出一种基于语义的短文分类算法。

1. 传统的 *KNN* 算法

在算法中，根据距离所选择邻居已经进行了正确的分类。具体算法可以描述如下：

（1） 确定最近邻的数量，也就是 K 值大小。K 值的选取对于分类算法的精确度有很大的影响，需要反复试验。

（2） 采用向量空间模型表示文本，形成文本特征向量。

（3） 计算测试文本与训练集中每个文本的相似度，通常使用余弦定理来计算相似度。

$$dist(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

（4） 在测试文本的 K 个邻居中，应用下面的公式依次计算每类的权重。

5.4 评价方案

（1）评价体系的建立

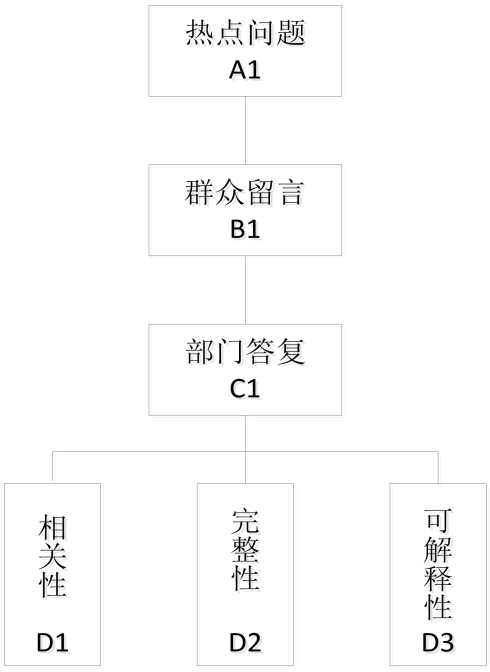

图 8 评价体系的构建

（1）构造比较判断矩阵并求权重

表 3：标度法

重要性标度	含义
1	表示两个元素相比，具有同等重要性
3	表示两个元素相比，前者比后者稍重要
5	表示两个元素相比，前者比后者明显重要
7	表示两个元素相比，前者比后者强烈重要
9	表示两个元素相比，前者比后者极端重要
2，4，6，8	表示上述判断的中间值
倒数	若元素i与j的重要性之比为 a_{ij}，则元素j与元素i的重要性之比为 $a_{ji}=1/a_{ij}$

采用 1-9 标度法，通过对同一级别的评价指标重要程度相互比较，构建出对应的比较矩阵。因为第一级评价指标相比具有同等重要性，故二者权重各占化 0.5，同理到第二、三级。得到以下比较判断矩阵：

$$A_1=\begin{bmatrix} 1 & 1/4 & 1/3 & /15 & 1/2 \\ 4 & 1 & 4/3 & 4/5 & 4/2 \\ 3 & 3/4 & 1 & 3/5 & 3/2 \\ 5 & 5/4 & 5/3 & 1 & 5/2 \\ 2/1 & 2/4 & 2/3 & 2/5 & 1 \end{bmatrix}$$

$$A_2=\begin{bmatrix} 1 & 2 & 2/5 & 2/4 & 1 \\ 1/2 & 1 & 1/5 & 1/4 & 1/2 \\ 5/2 & 5 & 1 & 5/4 & 5/2 \\ 4/2 & 4 & 4/5 & 1 & 4/2 \\ 1 & 2 & 2/5 & 1/2 & 1 \end{bmatrix}$$

（4）特征向量的提取

假设影响回复的因素个数 n，由此我们利用一个 n 阶正交反阵来解决问题。假设 n 阶正互反阵是一致矩阵

则 $$a_{ij}=a_{i1}\cdot a_{j1=}a_{j1/a_{i1}}$$

故第i行为故第 i 行为 $\frac{a_{11}}{a_{i1}}$ $\frac{a_{12}}{a_{i1}}$ … $\frac{a_{1n}}{a_{i1}}$(i=2, 3, ……, n)，与第一行成比例。

A的秩：rank=1 所以A只有一个特征根非零

设特征值为c_1，

则 $$n = randkA = \sum_{i=1}^{n} c_i = c_1.$$

（5）求最大特征值

设A的最大特征值$c_{max}=n$，相应的特征向量为ω ＝（ω_1，$\omega_2, \cdots, \omega_n$）.

由引理 $$\omega > 0.$$

由于 $$c_{max} \cdot \omega = A_{\omega}$$

所以 $$c_{max}\omega\, \imath = \sum_{j=1}^{n} a_{ij\omega_j} \quad (i=1, 2, \cdots, n)$$

设A的最大特征值$c_{\max} = n$，相应的特征向量为$\omega = (\omega_1, \omega_2, \cdots, \omega_n)^T$。

由引理$\omega > 0$.　由于$c_{\max} \cdot \omega = A_{\omega}$：

$$c_{max} = \sum_{j=1}^{n} a_{ij\omega_j} \quad (i=1, 2, \cdots, n)$$

$$c_{max} = \frac{1}{n}\sum_{i=1}^{n} \omega_i^{-1} \sum_{j=1}^{n} a_{ij\omega_j} \frac{1}{n}\sum_{i=1}^{n} \omega_i^{-1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \omega_i^{-1} \sum_{j=1}^{n} a_{ij\omega_j} \frac{\omega_i}{j_i}$$

$$= \frac{1}{n}\left[\sum_{i \neq j} a_{ij} \frac{\omega_i}{\omega_j} + n\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(a_{ij} \frac{\omega_i}{\omega_j} + a_{ij} \frac{\omega_j}{\omega_i}\right) + n\right]$$

$$\gg \frac{1}{n}\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 + n\right]$$

（6）结果分析

得到特征向量w_1、w_2：

$$w_1 = [0.1429 \quad 0.0714 \quad 0.3571 \quad 0.2857 \quad 0.1429]^T$$

$$CI = 2.2204e-16$$
$$CR = 1.7623e-16$$

$$w_2 = \begin{bmatrix} 0.0667 & 0.2667 & 0.2000 & 0.3333 & 0.1333 \end{bmatrix}^T$$

$$CI = 0.2204e-16$$
$$CR = 1.7623e-16$$

由于$CR < 0.1$,所以所给的主观评分是可以接受的结果。所求出的权重也是相对比较准确地。从各项指标的权重大小中可以知道，在留言答复的相关性、完整性和可解释性来说，相关性的比重最大，因此，一个答复中回复的相关性最重要，假如回复很完整和有解释性，但是跟群众所提问题无关的话，那么这个回复也是没有意义的；其次到可解释性，回复的内容除了跟所提内容相关之外还要具有一定程度的解释性，不然回复太过刻板、官方，群众无法理解，所作出的回复意义也不大，最后到完整性。但是三者的比重虽然有差异，但是差异比不大，说明在回复当中三者皆有才是回复最完美的时候。

六、模型评价

优点：KNN 方法基于类比学习，是一种非参数的分类技术，在基于统计的模式识别中非常有效，对于未知和非正态分布可以取得较高的分类准确率。KNN 方法虽然从原理上也依赖于极限定理，但在类别决策时，只与周围有限的邻近样本有关，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。

缺点：时间复杂度和空间杂度都比较高。

七、参考文献

[1] 刘婧姣，基于语义的短文分类算法研究，郑州，2013 年。
[2]樊兴华，王鹏. 基于两步策略的中文短文本分类研究[J]. 大连海事大学学报，2008，
11(2)：201—206.
[3] 江斌，微博自动分类方法研究及应用，哈尔滨大学，2012 年 6 月。

八、附件

```
import jieba
import pandas as pd
from wordcloud import WordCloud
```

```
def stopwordslist(stopwordspath):
    """
    获取停用词表
    :param stopwordspath: 停用词文件路径
    :return: 停用词列表
    """
    stopwordlist = [word for word in open(stopwordspath, "r", encoding="utf-8").readlines()]
    return stopwordlist


def read_excel(fpath):
    """
    读取文本文件
    :param fpath: 文件路径
    :return: 要分析的文本内容
    """
    df = pd.read_excel(fpath)
    df = df.astype(str)
    return df


def seg_depart(df, stopwordlist):
    """
    文本分词处理并去除停用词
    :param df: 要处理的文本
    :param stopwordlist: 停用词表
    :return: 新生成的文本
    """
    remarks = df["留言主题"]
    list_traget = []
    list_num = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
    for s in remarks:
        wordlist = jieba.lcut(s.replace("\n", ""))
        new_wordlist = [word for word in wordlist if word not in (stopwordlist and list_num)]
        word = " ".join(new_wordlist)
        list_traget.append(word)
    df["处理后留言主题"] = list_traget
    return df


def information_classifying(df):
    """
    对备注信息提供分类
```

```
    :param df:  要处理的文本
    :return:
    """
    build = open("build.txt", "w", encoding="utf-8")
    environment = open("environment.txt", "w", encoding="utf-8")
    transport = open("transport.txt", "w", encoding="utf-8")
    education = open("education.txt", "w", encoding="utf-8")
    labor = open("labor.txt", "w", encoding="utf-8")
    travel = open("travel.txt", "w", encoding="utf-8")
    sanitation = open("sanitation.txt", "w", encoding="utf-8")
    for i in range(df.shape[0] - 1):
        t = df["处理后留言主题"][i].replace("\n", "")
        if df["一级标签"][i] == "城乡建设":
            build.write(t)
        elif df["一级标签"][i] == "环境保护":
            environment.write(t)
        elif df["一级标签"][i] == "交通运输":
            transport.write(t)
        elif df["一级标签"][i] == "教育文体":
            education.write(t)
        elif df["一级标签"][i] == "劳动和社会保障":
            labor.write(t)
        elif df["一级标签"][i] == "商贸旅游":
            travel.write(t)
        elif df["一级标签"][i] == "卫生计生":
            sanitation.write(t)
    build.close()
    environment.close()
    transport.close()
    education.close()
    labor.close()
    travel.close()
    sanitation.close()


def generated_cloud_image(stopwordlist, fpath, save_path):
    """
    生成词云图
    :param stopwordlist:停用词列表
    :param fpath:处理文本路径
    :param save_path:图片保存路径
    :return:
    """
    stopwords = set(stopwordlist)
```

```
    text = open(fpath, encoding="utf-8").read()
    wordcloud = WordCloud(font_path="C:/Windows/Fonts/simfang.ttf",  # 设置字体
                          background_color="white",  # 设置背景颜色
                          ).generate(text)
    wordcloud.to_file(save_path)
    image = wordcloud.to_image()
    image.show()


stopwordlist = stopwordslist("D:\software\stoplist.txt")
remark = read_excel("D:\software\C 题全部数据\C 题全部数据\附件 2.xlsx")
new_df = seg_depart(remark, stopwordlist)
information_classifying(new_df)
generated_cloud_image(stopwordlist, "build.txt", "build.jpg")
generated_cloud_image(stopwordlist, "environment.txt", "environment.jpg")
generated_cloud_image(stopwordlist, "transport.txt", "transport.jpg")
generated_cloud_image(stopwordlist, "education.txt", "education.jpg")
generated_cloud_image(stopwordlist, "labor.txt", "labor.jpg")
generated_cloud_image(stopwordlist, "travel.txt", "travel.jpg")
generated_cloud_image(stopwordlist, "sanitation.txt", "sanitation.jpg")
```