# “智慧政务”中的文本挖掘应用

## 摘要

近年来，我国各地政府均在建设和推进智慧政务，智慧政务在推动政府廉洁高效运转、政府政策制定更加精准、服务大众更加便捷、信息化透明化程度更加高等方面发挥了积极的作用。留言信息遵循一定的规律在产生，每一条信息都有它的价值。对留言大数据的利用可进行对已经发生的事情进行精确的判断处理，对未发生的事情进行预测等操作。随着智慧政务的建设，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。本文主要通过这个问题进行研究，其研究结果如下：

1.针对群众留言分类问题，依照原有的群众留言信息，综合附件一的分类三级标签体系。运用 Python 中文分词组件 jieba 分词模型进行分词，并去除停用词提取关键词，通过各分类选取前 $N_i$ 个高频词语，形成各个分类的高频关键词词表，再进行人工干预等数据预处理方法；随后计算 $N_i$ 个高频关键词的 TF-IDF 数值，再提取部分已知所在分类的留言并转换为向量数据，作为 BP 神经网络算法的训练数据，构建模型，最后将需分类的留言导入 BP 神经算法的模型计算，得到留言所在的分类。

2.针对群众热点问题的挖掘，需要的是热度指标，且进行热度值排名。故利用 BERT 模型来进行对某特定人群或特定地区事件的汇总，将相同的留言进行整合，并汇总其点赞数、反对数、提出同一问题的

用户数、留言间的时间间隔、关键词词频等数据。利用设定的热度评价模型进行对信息热度值判断、排名。

3.针对政府答复意见评价的问题，该题题意在于对政府给于市民留言咨询、投诉的问题所答复的内容进行评价，从答复的相关性、完整性、时效性等方面进行建立评价指标。

**关键词：**Python， jieba 分词，BP 神经网络算法，BERT 模型

# Application of Text Mining in Intelligent Government

## Abstract

In recent years, local governments in China are building and promoting smart government, smart government has played a positive role in promoting the clean and efficient operation of the government, making government policies more accurate, serving the public more convenient, and the degree of information transparency is higher. Message information follows a certain rule, and every message has its value. The use of message big data can accurately judge and process what has happened and predict what has not happened. With the construction of smart government, the amount of text data related to all kinds of social conditions and public opinions is rising, which brings great challenges to the work of relevant departments which used to rely mainly on manual message division and hot spot collation. This paper mainly studies this problem, and the results are as follows:

1. In view of the problem of mass message classification, according to the original mass message information, the three-level label system of classification in Annex I is synthesized. Using the Chinese word

segmentation component of Python jieba word segmentation word segmentation, and extract key words, through the selection of the first Ni words, the formation of each category of high-frequency key words table, and then manually remove stop words and other data preprocessing method; Then calculate the TF-IDF values of Ni high-frequency keywords, and then extract part of the vector data of the known classification, as the training data of BP neural network algorithm, build the model, and finally import the message to be classified is imported into the model calculation of BP neural algorithm to get the message classification.

2. In view of the hot issues of the masses, what we need is the heat index and the ranking of heat value. Therefore, BERT model is needed to collect the events of a specific group of people or a specific area, integrate the same messages, and collect the data of the number of likes, objections, users who raise this issue, the time interval between messages, the number of words left and the frequency of keywords. The heat value of information is judged and ranked by using the set heat evaluation model.

3. In view of the evaluation of the government's reply, the purpose of this question is to evaluate the content of the government's reply to the public's message consultation and complaint, and to establish evaluation indicators from the aspects of relevance, completeness and timeliness of the reply.

**Keywords**: Python, Jieba participle, Back Propagation, BERT model

# 目录

# 1.绪论

## 1.1 背景

近年来， 随着电子技术、信息技术和网络技术的发展， 物联网、移动互联网、云计算、移动互联网、人工智能、数据挖掘和知识管理等技术日益发展，并推动了大数据的发展， 这为以提供公共服务为核心的智慧政务带来了发展的技术基础和推动力量。当前， 学术界和理论界并未对智慧政务形成一个统一的概念， 但从当前的研究现状来看， 一般认为， 智慧政务是在信息化时代背景下，综合运用互联网和信息网络技术， 以大数据为核心， 通过信息化手段为公众提供高效服务的政务模式，促进政务资源的整合和信息孤岛现象的消减，盘活数据资产，推动政务大数据全面应用，提升政府公共服务能力等自身建设；还可以促进政府和公众互动，让政务透明，帮助政府进行社会管理和解决社会难题，建立公众与政府间的沟通渠道，推进政府信息资源进一步开放共享，开发利用效率倍增，提升为民服务能力，促进经济社会快速发展。

## 1.2 问题再述

问题一：群众问题分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。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。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

问题二：热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 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”。

问题三：答复意见的评价

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。

# 2. 预处理

## 2.1 问题一

### 2.1.1 分词和提取关键词

人类语言都是词语加句式结构所形成，中文文本需要通过分词获得单个的词语。因此先对留言文本进行分词，由于我们使用 python 进行编程实现，故我们选用了 jieba 中文分词组件进行分词并抽取留言文本中的关键词[1]。jieba 分词包整体的工作流程如下图 2.1.1 所示

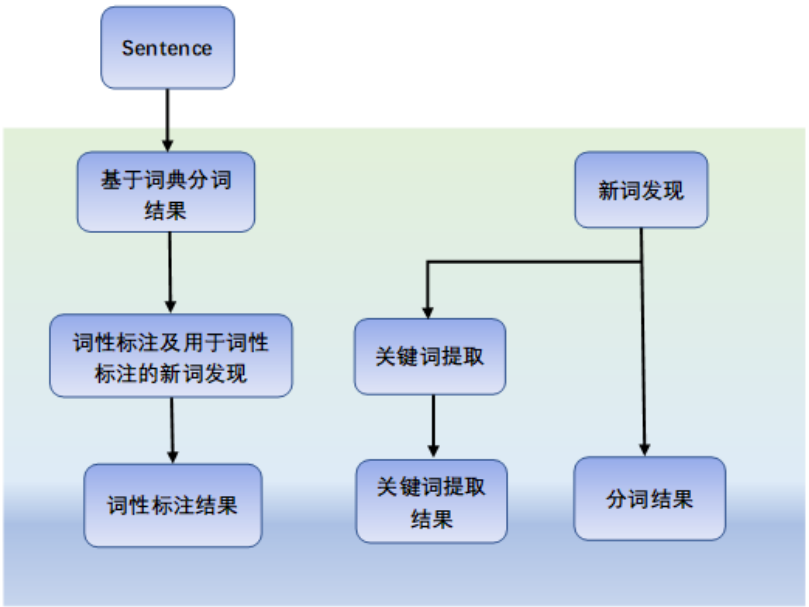



图 2.1.1 整体流程图

（1）分词提取

jieba 分词中，首先通过对照典生成句子的有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取。对于未登陆词（不在词典中的词）使用 HMM 进行新词发现[1]。

我们从附录 2 中的居民建议中选取其中的一条建议进行分词，从而可以的到如图所示：

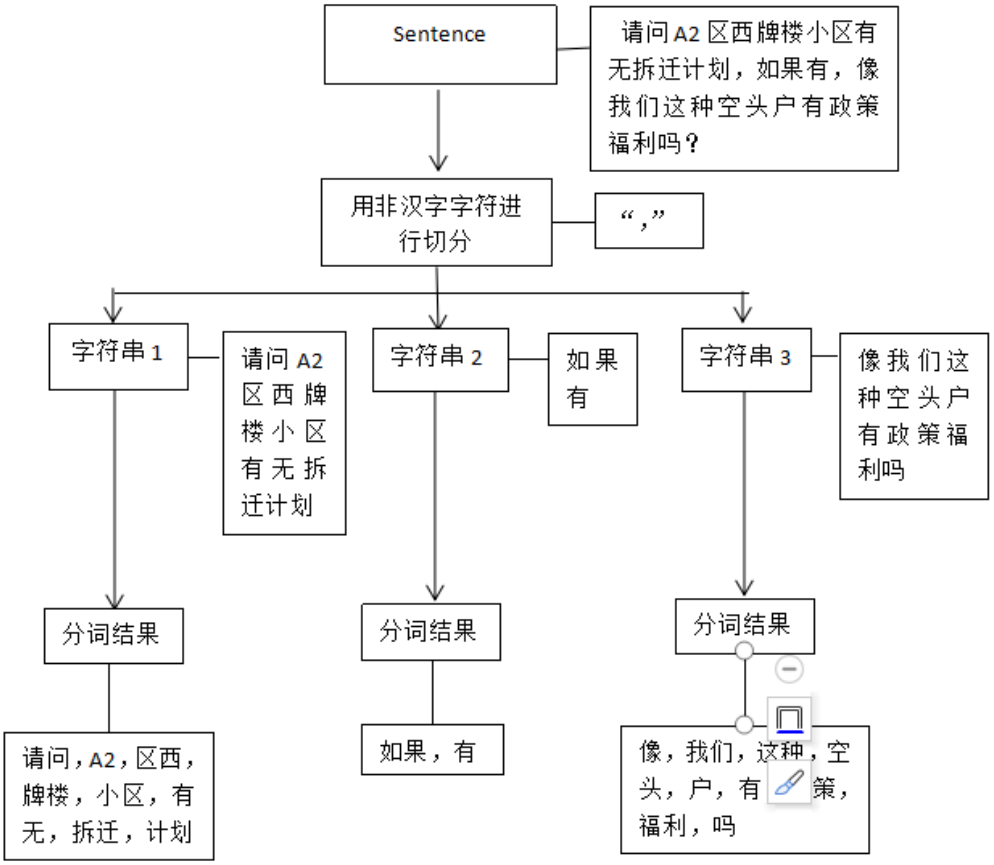



图 2.1.1（1）分词流程图

上图主要演示了分词的流程，但是其中只演示了被切分出来的子字符串的分词操作结果，在实际操作流程中，将对每一个子字符串都分别进行更为复杂的处理，生成句子的有向无环图（DAG），再根据词典概率，在 DAG 中生成最优路径，根据最优路径截取词语，最后将切分的分词结果与非汉字部分依次连接起来，作为最终的分词结果。

如果开启了 HMM，那么将会连起来不在词典中出现的连续单字进行新词发现。例子中的“有，无”，词典中没有这个词，所以会拿去 HMM 模型中进行新词发现，最后形成分词“有无”。

（2）抽取关键词

ieba 分词中有两种不同的用于关键词抽取的算法，分别为 TextRank 和 TF-IDF。实现流程比较简单，其核心在于算法本身。这里我们使用 TF-IDF 进行实现。

在此，我们继续使用上次的例子。

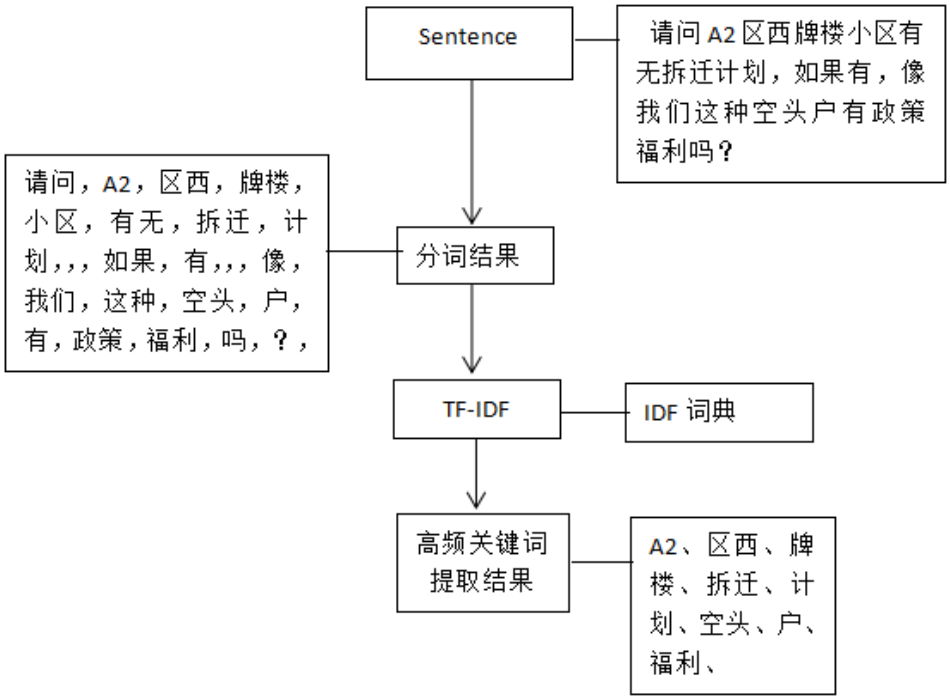



图 2.1.1（2）抽取关键词流程图

（3）去停用词，通用词

在抽取高频关键词前，在自然语言中，我们会选取一些停用词。所谓的停用词，就是一些不包含什么信息的词语，以及一些特别高频的词。比如的、是、这个、这样等，由于分析的群众给予政府部门的留言，故相对于“政府”、“回复”、“建议”、“市长”、“希望”等词语，也该列入本次的停用词中。

通过这种方法，就可以增加相对偏重所在分类内容的高频关键词的权重，并统计该词在该分类数据中出现的次数，选取前 $N_i$ 个(每个分类根据能描绘出该分类具体特征的词语而定)词语，形成各个分类的高频关键词词表。

在词表中，进行人工干预、去除部分不必要的词语，比如人名，符号等停用词无去除词语，随后计算 $N_i$ 个高频关键词的在该分类文本出现总个数，随后计算各个词语在此表中的权重，计入表中，形成词表。

载入数据集，逐一读取留言句子，对该句子进行分词、抽取关键词、去除停

用词与通用词，形成关键词集合（包含关键词与数量），由于需要反复的搜索，对于词语多的留言，需要较大的计算量，故设定一个阈值，限制留言提取出的词的数量，避免过多的词语导致计算量过大与重点语意偏移。且使用留言中统计的高频词进行计算，以此提高计算速度且对结果影响不大。

## 2.2 问题二

### 2.2.1 聚类汇总处理

问题二与问题一不同在于，问题一有已知固定的分类类别进行分类，而题二的问题则需要聚类，将相同的问题进行汇总。

问题二主要由某一人群、地区。如“xx 学院强制学生实习”，“xx 小区垃圾成堆”“xx 公司拖欠工资”等等；都有存在地词、机构团体名、或是工作相关名词；名词成立关键的分类标准，由于本次数据的地名几乎采用的字母代词。而对于实际运用于某个地区的政务留言进行处理而言，可先设立一该地区含有的地区名、机构名、各个企业的名字、街道的名字的汇总表，由大到小一级一级的分，利用该表对留言所发表的问题所涉及的地区与人群进行提取；从而可以更好的进行分类。

根据“附件 3”中数据，将其概述为“西地省 A 市政府的留言”，将留言信息分类为如图 2.2.1 所示结构，

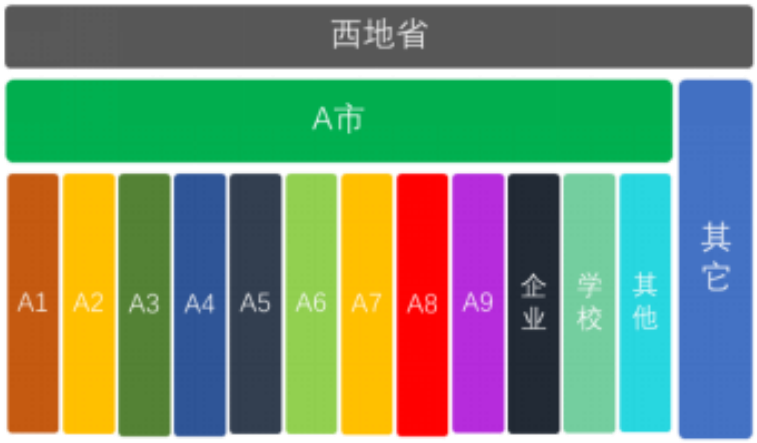



图 2.2.2（1）留言信息分类图

由于本次数据的地名几乎采用的字母代词，故分类时只需提取留言信息中的响应英语字符，而对于企业，学校等内容分类，由于没有名单、故只能运用

分词汇总后，根据经验进行划分。

而对于“其他”分类中的留言，因没标注所在区（县），但部分也属于各区（县）所有，故将其留言主题中的地区名称提取一一匹配，归类其所在区县的分类中。

但是，这样的分类留言还是存在着不足，就如下表所示的内容

| 留言主题 |
| --- |
| A 市商贸旅游职业技术学院强制学生实习 |
| A 市涉外经济学院强制学生实习 |
| A 市经济学院强制学生实习 |
| A 市经济学院强制学生外出实习 |
| A 市经济学院体育学院变相强制实习 |
| …… |

**表 2.2.1（1）部分留言数据表**

其中，反映内容与情况大致相同，但所涉及的学校却不相同，而且相隔时间较长，如将其分为同一问题进行热度分析，有失公允。

因此，通过分类、将各个地点事件分开处理，可以更高效的对群众所反映的问题进行汇总处理。

随后对各个地点的问题进行独立处理：如，涉及“万科魅力之城小区”中有四个问题：

问题 1：A 市万科魅力之城小区近百户楼板开裂墙面开裂

问题 2：A 市万科魅力之城未交房业主孩子不能上小区配套小学

问题 3：万科魅力之城小区底层门店深夜经营，各种噪音扰民

问题 4：A 市万科魅力之城楼板和墙面开裂，请政府管一管吧

首先将该地点所有问题出去关于该地点名词的字符，即

问题 1：近百户楼板开裂墙面开裂

问题 2：未交房业主孩子不能上小区配套小学

问题 3：底层门店深夜经营，各种噪音扰民

问题 4：楼板和墙面开裂，请政府管一管吧

此次我们运用 bert-as-service 。bert-as-service 是腾讯 AI Lab 开源

的一个 BERT 服务，它让用户可以以调用服务的方式使用 BERT 模型而不需要关注 BERT 的实现细节。基于 bert-serving-server 和 bert-serving-client 可以部署一个 bert 词向量 server，方便的调用 client API 进行词或句子的 embedding。

通过 BERT 模型进行相似度计算如下表所示；由此可见，问题 1 与问题 4 间相似度最高，且二者反映的问题相同，故将其分为一类。

| | 问题 1 | 问题 2 | 问题 3 | 问题 4 |
|---|---|---|---|---|
| 问题 1 | 0.99999976 | 0.90185297 | 0.9003415 | 0.96119356 |
| 问题 2 | 0.90185297 | 0.9999999 | 0.8793235 | 0.84974766 |
| 问题 3 | 0.9003415 | 0.8793235 | 1.0000005 | 0.87991023 |
| 问题 4 | 0.96119356 | 0.84974766 | 0.87991023 | 0.9999998 |

**表 2.2.1（2）问题相似度表**

# 3.研究方案问题的解决

## 3.1 问题一

### 3.1.1 TF-IDF

（1）原理[2]

TF-IDF（term frequency－inverse document frequency，词频-逆向文件频率)是一种用于信息检索(information retrieval)与文本挖掘(text mining)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。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用

来分类。

①词频 TF

词频（TF）表示词条（关键字）在文本中出现的频率，其公式为：

$$tf_{ij}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$ ，即：$$TF_w=\frac{\text{在某一类中的词条w出现的次数}}{\text{该类中所有的词条数目}}$$

其中 $n_{i,j}$ 是该词在文件 dj 中出现的次数，分母则是文件 dj 中所有词汇出现的次数总和。

②IDF 是逆向文件频率

逆向文件频率（IDF）：某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。

如果包含词条 t 的文档越少，IDF 越大，则说明词条具有很好的类别区分能力。其公式为：

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

其中，|D| 是语料库中的文件总数。 $|\{j:t_i\in dj\}|$ 表示包含词语 $t_i$ 的文件数目（即 $n_{i,j}\neq 0$ 的文件数目）。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用 $1+|\{j:t_i\in dj\}|$

（2）应用

留言词语集合的每个词都分别在各个词表中搜索匹配，，累计各个分类的 TF·TDF 值，如下图所示

| | | | | | |
|---|---|---|---|---|---|
| 1 | | | | 25077 | |
| 2 | | 词出现数量 | 词频TF | 逆文本相关度：IDF | TF-IDF |
| 3 | 医院 | 1792 | 0.07146 | 0.880492726 | 0.06292 |
| 4 | 医生 | 970 | 0.038681 | 1.256185676 | 0.04859 |
| 5 | 患者 | 450 | 0.017945 | 2.024056403 | 0.036321 |
| 6 | 生育 | 417 | 0.016629 | 1.659652781 | 0.027598 |
| 7 | 病人 | 342 | 0.013638 | 2.006961969 | 0.027371 |
| 8 | 手术 | 336 | 0.013399 | 2.123686243 | 0.028455 |
| 9 | 小孩 | 327 | 0.01304 | 1.78043432 | 0.023217 |
| 10 | 治疗 | 310 | 0.012362 | 1.864991708 | 0.023055 |
| 11 | 检查 | 306 | 0.012202 | 1.957365028 | 0.023885 |
| 12 | 医疗 | 301 | 0.012003 | 1.449770425 | 0.017402 |
| 13 | 孩子 | 276 | 0.011006 | 2.006961969 | 0.022089 |
| 14 | 家属 | 262 | 0.010448 | 2.408198741 | 0.02516 |
| 15 | 家庭 | 249 | 0.009929 | 2.15267378 | 0.021375 |
| 16 | 住院 | 224 | 0.008932 | 2.1142075 | 0.018885 |
| 17 | 主任 | 210 | 0.008374 | 1.828886703 | 0.015315 |
| 18 | 独生子女 | 205 | 0.008175 | 2.358805986 | 0.019283 |
| 19 | 办理 | 203 | 0.008095 | 2.032714465 | 0.016455 |
| 20 | 计划生育 | 198 | 0.007896 | 2.289010224 | 0.018073 |
| 21 | 卫生院 | 197 | 0.007856 | 2.420937767 | 0.019018 |
| 22 | 医师 | 194 | 0.007736 | 2.358805986 | 0.018248 |
| 23 | 证明 | 182 | 0.007258 | 2.446913253 | 0.017759 |
| 24 | 户口 | 179 | 0.007138 | 2.266787087 | 0.01618 |
| 25 | 护士 | 179 | 0.007138 | 2.734595326 | 0.01952 |
| 26 | 药品 | 178 | 0.007098 | 2.514966717 | 0.017852 |
| 27 | 时间 | 176 | 0.007018 | 1.7670113 | 0.012402 |
| 28 | 卫生 | 164 | 0.00654 | 1.256185676 | 0.008215 |
| 29 | 子女 | 162 | 0.00646 | 2.050258775 | 0.013245 |
| 30 | 卫生局 | 161 | 0.00642 | 2.487187152 | 0.015968 |
| 31 | 健康 | 161 | 0.00642 | 2.358805986 | 0.015144 |

城乡建设 | 环境保护 | 交通运输 | 教育文体 | 劳动和社会保障 | 商贸旅游 | 卫生计生

就绪 滚动

**图 3.1.1（1）TF.IDF 数值图**

而我们通过累计各个分类的 TF·TDF 值，最终形成一个七维向量，如下图所示

```
0.0000000000000000  0.2247360534388290  0.0593873246172781  0.0000000000000000  0.0060764651904741  0.1339009059601600  0.0000000000000000
0.0578672543056421  0.1930913410442320  0.0663608186109288  0.0335142839584175  0.0414450463910037  0.0596565171943163  0.0279372637062533
0.0000000000000000  0.0313851770838205  0.0000000000000000  0.0000000000000000  0.0000000000000000  0.0232520590296930  0.0000000000000000
0.0294297591751832  0.2316605284563080  0.1035776662863620  0.0416807854716909  0.0332687322125691  0.0630634742133088  0.0064646083081546
0.0389995789527891  0.1460643634067480  0.0782161684819097  0.0857236723257844  0.0630264299961892  0.0313654599419497  0.0621635095588422
0.0565216240068791  0.1603228699346600  0.0763171295555835  0.0719681628993487  0.1030998024807430  0.0776200639629381  0.0554102242968700
0.0164430067705797  0.0872856668841559  0.0046299275606478  0.0375898543073214  0.0221338938225549  0.0093652511876162  0.0000000000000000
0.0435813179739616  0.1031727365475670  0.0249920008027423  0.0391355801542167  0.0345709150948138  0.0915447333887898  0.0292291675602498
0.0212179319587248  0.0422733718320713  0.0368657641954999  0.0139462032756673  0.0114590025628287  0.0255966818631985  0.0059294319295668
0.0813763695577279  0.2570699240492700  0.0999315835084981  0.1396300459109430  0.0955013991052404  0.0670087413566364  0.0359251753138679
0.0418105839323997  0.2155294376754230  0.1059272092409600  0.0746255851031914  0.0464407520880486  0.0544121897884871  0.0069221171121764
```

**图 3.1.1（2）七维向量图**

## 3.1.2 基于 BP 神经算法

（1）算法内容原理[3]

BP 算法（backpropagation,反向传播算法）的学习过程，由信息的正向传播和误差的反向传播两个过程组成。输入层各神经元介绍来自外界的输入信息，并传递给中间层各神经元。中间层是内部信息处理层，负责信息变换，根据信息变化能力的需求，中间层可以设计为单隐层或多隐层结构。最后一个隐

层传递到输出层各神经元的信息，经过进一步处理后，完成一次正向的传播过程，由输出层向外界输出信息处理的结果。当实际输出与期望输出不同时，进入误差的反向传播阶段。误差通过输出层。按误差梯度下降的方式修正各层权值，向隐层、输入层、逐层反传。周而复始的信息正向传播和误差反向传播过程，是各层权值不断调整的过程，也是神经网络学习训练的过程，此过程一直进行到网络输出的误差减少到可以接受的程度或者预先设定的学习次数为止。

（2）理论推导[3]

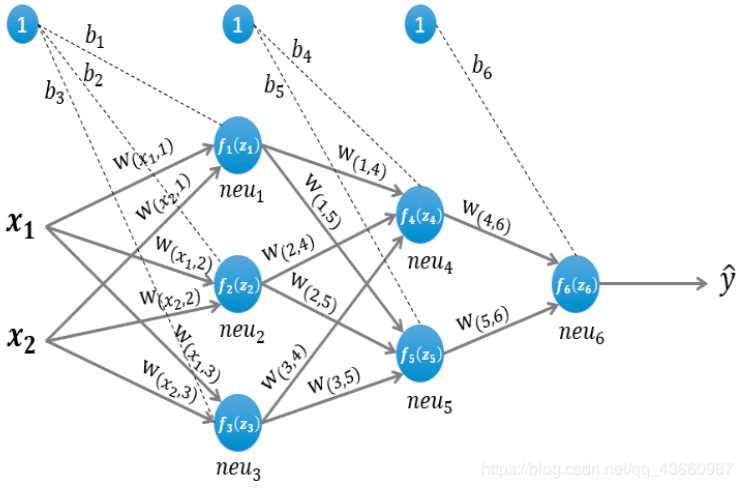



图 3.1.1 简单的神经网络结构图

如图 3.1.1 所示为一个简单的三层（两个隐藏层和一个输出层）神经网络结构图，假设输入样本为$\vec{a} = （x_1，x_2）$。

第一层网络的参数为：

$$W^{(1)} = \begin{bmatrix} w_{(x_1,1)} & w_{(x_2,1)} \\ w_{(x_1,2)} & w_{(x_2,2)} \\ w_{(x_1,3)} & w_{(x_2,3)} \end{bmatrix}, \quad b^{(1)} = [b_1, b_2, b_3]$$

第二层网络参数为：

$$W^{(2)} = \begin{bmatrix} w_{(1,4)} & w_{(2,4)} & w_{(3,4)} \\ w_{(1,5)} & w_{(2,5)} & w_{(3,5)} \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} b_4 & b_5 \end{bmatrix}$$

第三层网络的参数为：

$$W^{(3)} = \begin{bmatrix} w_{(4,6)} & w_{(5,6)} \end{bmatrix}, \quad b^{(3)} = \begin{bmatrix} b_6 \end{bmatrix}$$

①第一层隐藏层的计算：

第一层隐藏层有三个神经元（$nue_1$，$nue_2$，$nue_3$），该层的输入为

$$z_1 = w(x_1,1) * x_1 + w(x_2,1) * x_2 + b_1$$

同理，有：

$$z_2 = w(x_1,2) * x_1 + w(x_2,2) * x_2 + b_2$$

$$z_3 = w(x_1,3) * x_1 + w(x_2,3) * x_2 + b_3$$

假设我们用 f（x）做该层的激活函数，那么该层的输出层为 $f_1$（$z_1$），$f_2$（$z_2$）和 $f_3$（$z_3$）。

②第二层隐藏层的计算

第二层隐藏层有两个神经元（$neu_4$，$nue_5$），该层的输入为：

$$z^{(2)} = W^{(2)} * \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix}^T + (b^{(2)})^T$$

即第二层的输入层是第一层的输出层乘以第二层的权重，再加上第二层的偏置，因此得到 $neu_4$，$nue_5$ 的输入分别是：

$$z_4 = w_{(1,4)} * z_1 + w_{(2,4)} * z_2 + w_{(3,4)} * z_3 + b_4$$

$$z_5 = w_{(1,5)} * z_1 + w_{(2,5)} * z_2 + w_{(3,5)} * z_3 + b_5$$

该层的输出分别为 $f_4$（$z_4$）和 $f_5$（$z_5$）

③输出层的计算：

输出层只有一个神经元 $neu_6$，故该层输出为

$$z^{(3)} = W^{(3)} * \begin{bmatrix} z_4 & z_5 \end{bmatrix}^T + (b^{(3)})^T$$

即

$$z_6 = w_{(4,6)} * z_4 + w_{(5,6)} * z_5 + b_6$$

因为该网络要解决二分类问题，所以输出层的激活函数也可以用一个 Sigmoid 型函数，神经网络最后的输出为 $f_6$（$z_6$）

（3）反向传播[3]

反向传播的计算过程。假设我们使用随机梯度下降的方式来学习神经网络的参数，损失函数定义为$L(y,\hat{y})$，其中 y 是该样本的真实类标。使用度下降进行参数的学习，我们必须计算出损失函数关于神经网络中各层参数(权重 w 和偏置 b)的偏导数。

假设我们要对第 k 层隐藏层的参数 $W^{(k)}$ 和 $b^{(k)}$ 求偏导数，即求$\frac{\partial L(y,\hat{y})}{\partial W^{(k)}}$和$\frac{\partial L(y,\hat{y})}{\partial b^{(k)}}$；假设 $z^{(k)}$ 代表第 k 层神经元的输入，即 $z^{(k)}=W^{(k)}*n^{(k-1)}+b^{(k)}$，其中 $n^{(k-1)}$ 为前一层神经元的输入，根据链式法则，有：

$$\frac{\partial L(y,\hat{y})}{\partial W^{(k)}}=\frac{\partial L(y,\hat{y})}{\partial b^{(k)}}*\frac{\partial z^{(k)}}{\partial W^{(k)}}$$

$$\frac{\partial L(y,\hat{y})}{\partial b^{(k)}}=\frac{\partial L(y,\hat{y})}{\partial z^{(k)}}*\frac{\partial z^{(k)}}{\partial b^{(k)}}$$

因此，我们只需计算偏导数$\frac{\partial L(y,\hat{y})}{\partial z^{(k)}}$，$\frac{\partial z^{(k)}}{\partial W^{(k)}}$，$\frac{\partial z^{(k)}}{\partial b^{(k)}}$

①计算偏导数

因为偏置 b 是一个常数项，因此其偏导数的计算为：

$$\frac{\partial z^{(k)}}{\partial b^{(k)}}=\begin{bmatrix}\frac{\partial(W_{1:}^{(k)}*n^{(k-1)}+b_1)}{\partial b_1} & \cdots & \frac{\partial(W_{1:}^{(k)}*n^{(k-1)}+b_1)}{\partial b_m}\\ \cdots & \cdots & \cdots\\ \frac{\partial(W_{m:}^{(k)}*n^{(k-1)}+b_m)}{\partial b_1} & \cdots & \frac{\partial(W_{1:}^{(k)}*n^{(k-1)}+b_1)}{\partial b_m}\end{bmatrix}$$

依然以第一层的隐藏层为例，则有：

$$\frac{\partial z^{(1)}}{\partial b^{(1)}}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

偏导数$\frac{\partial L(y,\hat{y})}{\partial z^{(k)}}$又称误差项，也称灵敏度，一般用$\delta$ 表示，例如 是第一层

神经元的误差项，其值的大小代表了第一层神经元对于总误差的影响大小。根据前面的前向计算，我们知道第 k+1 层的输入与第 k 层输入的关系为：

$$z^{(k+1)} = W^{(k+1)} * n^{(k)} + b^{k+1}$$

又因为 $n^{(k)} = f_k(z^{(k)})$ ，根据链式法则我们可以得到 $\delta^{(k)}$

$$\delta^{(k)} = \frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{z}^{(k)}} = \frac{\partial n^{(k)}}{\partial \mathrm{z}^{(k)}} * \frac{\partial \mathrm{z}^{(k+1)}}{\partial n^{(k)}} * \frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{z}^{(k+1)}}$$

$$= \frac{\partial n^{(k)}}{\partial \mathrm{z}^{(k)}} * \frac{\partial \mathrm{z}^{(k+1)}}{\partial n^{(k)}} * \delta^{(k+1)}$$

$$= f_k^{'}(z^{(k)}) * ((W^{(k+1)})^T * \delta^{(k+1)})$$

由上式我们可以看到，第 k 层的神经元的误差项 $\delta^{(k)}$ 是由第 k+1 层的误差项乘以第 k+1 层的权重，再乘以第 k 层激活函数的导数得到，则 $\frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{W}^{(\mathrm{k})}}$ , $\frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{b}^{(\mathrm{k})}}$ 可分别表示：

$$\frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{W}^{(\mathrm{k})}} = \frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{z}^{(k)}} * \frac{\partial \mathrm{z}^{(k)}}{\partial \mathrm{W}^{(\mathrm{k})}} = \delta^{(k)} * (n^{(k-1)})^T$$

$$\frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{b}^{(\mathrm{k})}} = \frac{\partial \mathrm{L}(\mathrm{y},\hat{\mathrm{y}})}{\partial \mathrm{z}^{(k)}} * \frac{\partial \mathrm{z}^{(k)}}{\partial \mathrm{b}^{(\mathrm{k})}} = \delta^{(k)}$$

（4）导入模型

提取一部分已知所在分类的向量数据，作为 BP 神经网络算法的训练数据，通过 BP 的原理，并且建立 BP 神经网络模型，得出分类数据。

由 BP 模型计算，分类结果准确度接近 85%。

### 3.1.3 F-Score

根据评价公式：

$$F1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2P_i R_i}{P_i + R_i}$$

其中， $P_i$ 为第 $i$ 类的查准率， $R_i$ 为第 $i$ 类的查全率

表 3.1.1 测试样本分类结果

样本总数：954

| 样本 / 分类 | 交通 | 商贸 | 卫生 | 城乡 | 劳动 | 教育 | 环境 |
|---|---|---|---|---|---|---|---|
| **交通运输** | 116 | 5 | 2 | 1 | 0 | 9 | 0 |
| **商贸旅游** | 14 | 121 | 8 | 25 | 0 | 6 | 14 |
| **卫生计生** | 4 | 4 | 76 | 2 | 2 | 9 | 4 |
| **城乡建设** | 1 | 1 | 0 | 48 | 0 | 0 | 3 |
| **劳动和社会保障** | 15 | 7 | 12 | 13 | 116 | 15 | 3 |
| **教育文体** | 1 | 0 | 6 | 12 | 0 | 158 | 2 |
| **环境保护** | 1 | 2 | 0 | 1 | 2 | 0 | 83 |
| **准确率** | 76.3% | 86.4% | 73% | 65.2% | 97.5% | 80.2% | 77.6% |
| F-Score | | | | 0.7790 | | | |

由分类结果数据可知，整体分类效果较好，但整体略有偏向于商贸旅游与劳动和社会保障，由此可知，该二者词表构建不完善，存在部分不适当词语，故可适当进人工干预。且当选择词表的高频词个数时，需根据整体留言数据来定，无需按照统一标准进行提取，从而能提高分类准确度。

## 3.2 问题二

### 3.2.1 BERT 模型

（1）模型的主体结构

BERT 模型沿袭了 GPT 模型的结构，采用 Transfomer 的编码器作为主体模型结构。Transformer 舍弃了 RNN 的循环式网络结构，完全基于注意力机制来对一段文本进行建模

Transformer 所使用的注意力机制的核心思想是去计算一句话中的每个词对于这句话中所有词的相互关系，然后认为这些词与词之间的相互关系在一定程度上反应了这句话中不同词之间的关联性以及重要程度。因此再利用这些相互关系来调整每个词的重要性（权重）就可以获得每个词新的表达。这个新的表征不但蕴含了该词本身，还蕴含了其他词与这个词的关系，因此和单纯的词向量相比

是一个更加全局的表达。

Transformer 通过对输入的文本不断进行这样的注意力机制层和普通的非线性层交叠来得到最终的文本表达[5]。

（2）模型的输入输出[4]

BERT 模型的全称是：**BidirectionalEncoder Representations from Transformer**。从名字中可以看出，BERT 模型的目标是利用大规模无标注语料训练、获得文本的包含丰富语义信息的 Representation，即：文本的语义表示，然后将文本的语义表示在特定 NLP 任务中作微调，最终应用于该 NLP 任务。

在基于深度神经网络的 NLP 方法中，文本中的字/词通常都用一维向量来表示（一般称之为“词向量”）；在此基础上，神经网络会将文本中各个字或词的一维词向量作为输入，经过一系列复杂的转换后，输出一个一维词向量作为文本的语义表示。特别地，我们通常希望语义相近的字/词在特征向量空间上的距离也比较接近，如此一来，由字/词向量转换而来的文本向量也能够包含更为准确的语义信息。因此，BERT 模型的主要输入是文本中各个字/词的原始词向量，该向量既可以随机初始化，也可以利用 Word2Vector 等算法进行预训练以作为初始值；输出是文本中各个字/词融合了全文语义信息后的向量表示，如下图所示（为方便描述且与 BERT 模型的当前中文版本保持一致，本文统一以字向量作为输入）：

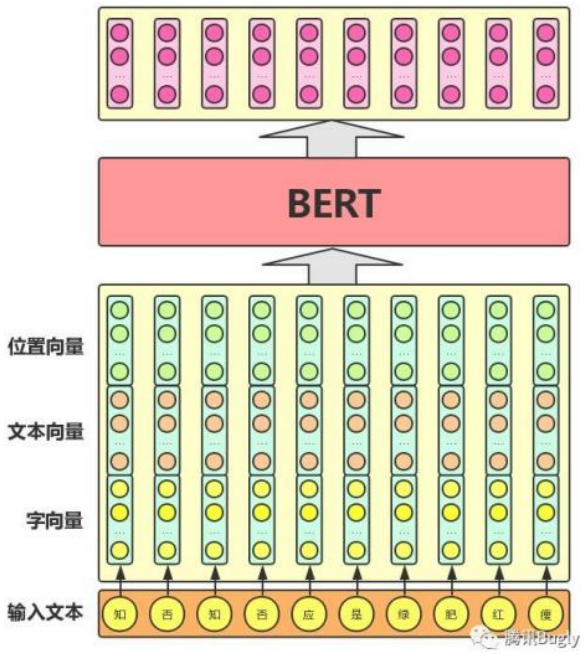



图 3.2.1BERT 模型原理图

从上图中可以看出，在计算中文向量时，可以直接输入整个句子不需要提前分词。因为 Chinese-BERT 中，语料是以字为单位处理的，因此对于中文语料来说输出的是字向量。BERT 模型通过查询字向量表将文本中的每个字转换为一维向量，作为模型输入；模型输出则是输入各字对应的融合全文语义信息后的向量表示。此外，模型输入除了字向量，还包含另外两个部分：

1. 文本向量：该向量的取值在模型训练过程中自动学习，用于刻画文本的全局语义信息，并与单字/词的语义信息相融合。

2. 位置向量：由于出现在文本不同位置的字/词所携带的语义信息存在差异（比如：“我爱你”和“你爱我”），因此，BERT 模型对不同位置的字/词分别附加一个不同的向量以作区分。

（3）BERT 模型

在本文中，我们聚焦文本分类任务，建立 BERT 模型，通过导入模型输入量，由于利用 BERT，输入每个词句得到的是（1，728）的向量，如利用该向量进行聚类，该数值的计算过大，需要更为专业的计算机或服务器来实现。

独立对关于各个地点的 n 个问题留言两两间依次进行相似度计算，并设定一个阈值，当二者相似度大于阈值时，对二者的问题编号进行标记记录，最后统计该地点下记录的问题。如 M 个留言中存在 N 个留言间两两均有标记，即 N 个留言归为一类，剩下的留言当与其中一分类中 N 个留言中的αN 个留言存在对应标记时，该留言归为该分类，否则独立为一分类。即如图 2.2.1（3）所示，留言与留言 B 应归类于分类一，留言 C、E、F、D 应归类于分类二，而留言 G 虽和留言 F 相似度较高，但分类二整体相似度较低，故不归类于分类二，独立为分类三。

聚类后归为一类的留言，赋予留言 ID 标记，便于辨识与热度指标计算

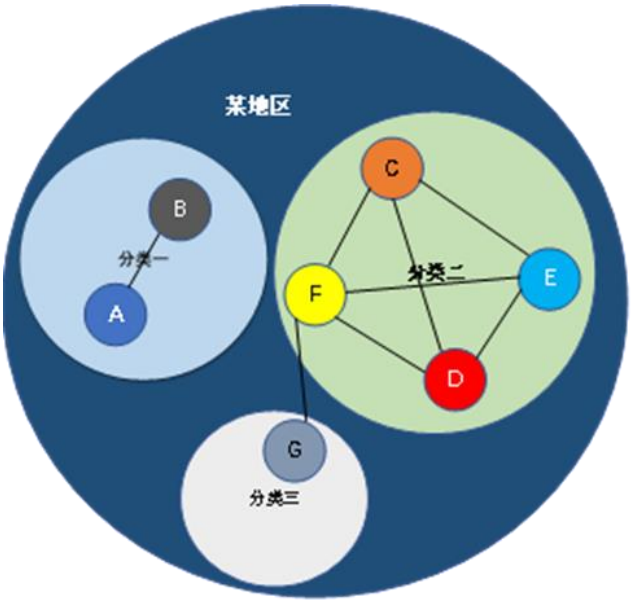



图 2.2.1（3）留言分类图

### 3.2.2 建立热度公式

通过 BERT 模型的聚类汇总处理，从附件 3 的数据中，可定义一个热度评价指标，如下表 3.2.2 所示

表 3.2.2 热度指标表

| | 指标 | | | |
|---|---|---|---|---|
| **热度指数** | 用户发表数 | 点赞数 | 反对数 | 间隔月数 |

因为考虑到热度会随着时间发生衰减，所以我们采用牛顿冷却定律作为时间衰减函数来计算热度指数：

$$H = H_0 \bullet e^{0.1\Delta t}$$

$H_0$ 表示这个事件的初始热度指数，$\Delta t$ 表示这个事件的时间范围（间隔月数）。对于初始的热度，我们用三个指标来衡量：用户发表数、点赞数、反对数。

事件的初始热度指数公式如下

$$H_0 = 0.675N_1 + 0.175N_2 + 0.15N_3$$

其中，$N_1$ 表示用户发表数（用户发表同一事件总数），$N_2$ 表示这一事件的点赞数，$N_3$ 表示这一事件的反对数。其权重系数根据现有内容规律推断。

该公式的特点在于热度指数不是固定的，会随着时间的增长而发生改变，这

样做的好处是能真正反映出当前人们关心的或迫切需要解决的热点问题。

具体热点信息如附件所示。

## 3.3 问题三

### 3.3.1 建立模型评价指标

根据附录 4 的数据，我们通过总结归纳，可以得到如下表和图所示


图 3.3.1 数据指标

表 3.3.2 数据指标

| 指标 | 相关性 | 完整性 | | | 时效性 | 语句简洁性 | |
|---|---|---|---|---|---|---|---|
| 具体方面 | BERT 模型计算 | 礼貌用语 | 引用规定（法律法规） | 结尾时间标注 | 答复间隔时间 | 关键词数占比 | 二者关键词数对比 |
| 数值 | （0-1） | （0，1） | （0，1） | （0，1） | 天数 | （0-1） | （0- 1） |

由 3.3.2 表可见只有天数值与留言与答复关键词词数对比起伏波动比较大，其他的数值均在于 0 与 1 之间，对此，使得这些指标之间存在着不可公度性，这就为综合评价带来了困难，尤其是为综合评价指标建立和依据这个指标的大小排序产生不合理性。

因此，需将答复时间间隔进行数据标准化（也称为无量纲化）。

采用了直线型无量纲化方法，其是指指标原始值与无量纲化后的指标值之间呈现线性关系，常用的线性量化方法有阈值法。阈值也称临界值，是指衡量事物

发展变化的一些特殊指标值，如极大值、极小值等，而阈值法就是通过实际值与阈值对比得到无量纲化指标值的方法。

由于时间指标对于评价值来说是逆相关的，由此使应采用模型：

$$Y_i = \frac{\max X_i + \min X_i - X_i}{\max X_i}$$

评价值范围：$[\frac{\min X_i}{\max X_i}, 1]$

影响评价值因素：$X_i > 0; \max X_i; \min X_i$

该模型特点在于评价值随指标增大而减小，适合对逆指标进行无量纲化处理，即无量纲化和指标转换同时进行。

附件 4 数据中，由于出现回复间隔天数最大值与其他值相差较多，达 1160 天与 893 天，故认为该二值为奇异值，将其剔除。无量纲化后，时间指标数值在于（0-1）之间。时间间隔最小值接近与 1；而最大值接近于 0，符合计算要求。

同理，留言与答复关键词词数对比：

$$n_i = \frac{\text{答复关键词数}}{\text{留言关键词数} + \text{答复关键词数}}$$

对于多指标（或多因素）的综合评价问题,就是要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,作为综合评价的依据，从而得到相应的评价结果。

假设 n 个被评价对象的 m 个评价指标向量为 $x = (x_1, x_2, \cdots, x_m)^T$ ，指标权重向量为 $w = (w_1, w_2, \cdots, w_m)^T$ ，由此构造综合评价函数为 $y = f(w, x)$

对于权重值的确定,根据各个指标值的具体分布情况,以评价总分为(0-100)计算，即权重值如下表所示：

**表 3.3.3 权重值表**

| **指标** | **相关性** | **完整性** | | | **时效性** | **语句简洁性** | |
|---|---|---|---|---|---|---|---|
| **具体方面** | BERT 模型计算 | 礼貌用语 | 引用规定（法律法规） | 结尾时间标注 | 答复间隔时间 | 关键词数占比 | 二者关键词数对比 |
| **权重** | 30 | 20 | | | 30 | 10 | 10 |

且评定五个等级：即：A（81-100）、B（61-80）、C（41-60）、D（21-40）、E（1-20）；根据该评价模型计算，得大致结果如下图 3.3.4 所示：

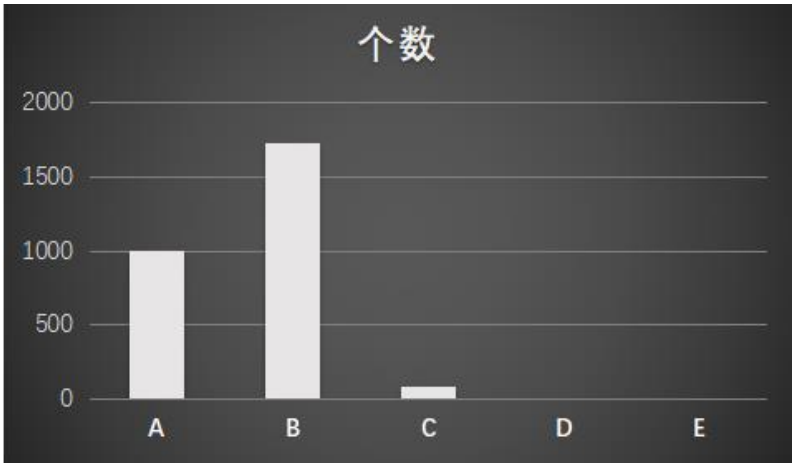



图 3.3.4 各个等级评价数量的柱状图

根据附件 C 数据中可发现，等级 A、B 的答复内容具体、详细且与留言具有一定相关性、且大部分具有开头与结尾的格式，且运用了相关的规定（法律法规）进行解答。而等级 C、D 缺乏部分只是简单的答复，如“已知悉”、“已通知相关部门了解”等，没有一定的答复有效性，故分数较低。因此该评价模型整体可行性较高，符合对该留言答复机制的评分要求。

# 4. 总结

## 4.1 结论

本文首先对数据进行预处理，对群众的留言等问题进行分类处理，从而使数据更加的整体、整洁。

为了给群众留言贴上标签，我们使用了 Python 中的 jieba 分词，提取群众留言的高频关键词，从而使数据更加的明了。再运用 BP 神经网络建立模型，得出群众留言的分类。

但分词效果相对来说较差，为提高准确率，可通过自定义 jieba 模型的词典进行分词，更精准的将各名称信息进行提取，避免出现地点名称与其他名词混淆。导致分词不明确，同一词分词后有不同的结构，从而匹配不上的结果。分词时进行词性标注也更加的精准，利用词语的词性与表达意思多重匹配，且采用庞大的

数据进行处理分析，可以很精准的对留言信息进行分类处理。

对于群众热度问题，我们需要的是热度指标，并且进行热度排行。为此，我们需要对某特定地点或特定人群的留言进行归类汇总，故我们采用建立 BERT 模型进行整合。并且我们根据设定的热度指标（用户发表数，点赞数，反对数，间隔月数）来建立热度公式，依据热度评价模型进行热度值的判断排行。

对于答复意见的评价，我们整合数据，从答复的相关性，完整性，时效性等方面来建立答复评价指标（A，B，C，D，E 五个等级）。

## 4.2 展望

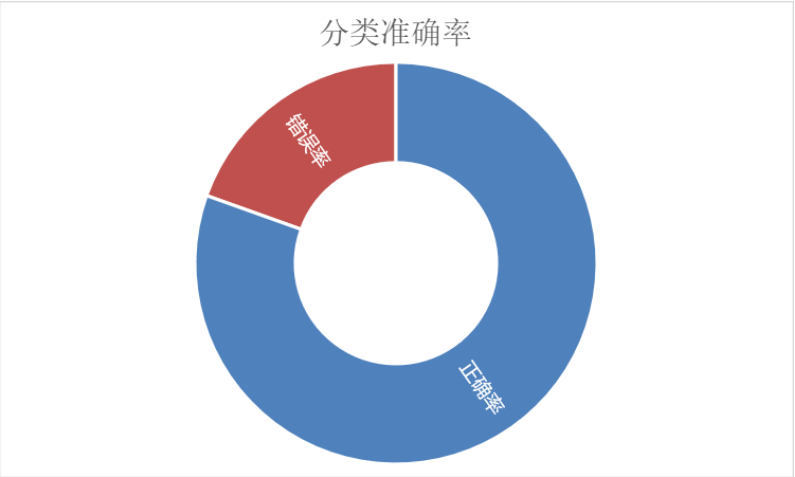
本小组通过老师和学长知道，网上视频，查找网页等学习这些技术算法，能力得到了锻炼。数据挖掘是一种很有意思的工作，其中思路和方法是很重要的。我们小组分工合作，虽然想到了许多的思路，但由于各种的原因没能实现，这让我们认识到自身的局限性。所以，我们希望能够更加深入学习交流，体验大数据所带来的对生活生产的便利和对万物之间规律的进行探索。

# 5.参考文献

[1] 尘嚣看客. jieba 分词详解-[EB/OL]-
https://www.jianshu.com/p/2cccb07d9a4e-2019.07.16
[2] Asia-Lee. TF-IDF 算法介绍及实现-[EB/OL]-
https://blog.csdn.net/asialee_bird/article/details/81486700 -
2018.08.07
[3]BP 算法基础及实现-[EB/OL]-
https://www.csdn.net/gather_22/MtTacg4sMjM4MS1ibG9n.html-2019.06.23
[4] 腾讯 Bugly.图解 BERT 模型：从零开始构建 BERT-[EB/OL]
https://cloud.tencent.com/developer/article/1389555 2019.01.30
[5] 追一科技. BERT 模型 [EB/OL]
https://www.zhihu.com/question/298203515/answer/512838775.2018.10.18
[6] 赵志勇.Python 机器学习算法[J].电子工业出版社，2017.7;114-136.

# 附录

## 附录一 题一分类结果




## 附录二 题二热点信息

| 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 | 828.92 | 2019/1/11至2019/7/8 | A市58车贷 | 58车贷诈骗案没有公开案件处理进展 |
| 2 | 2 | 598 | 2019/5/5至2019/9/19 | A市五矿万境K9县 | 房屋存在质量问题 |
| 3 | 3 | 319.37 | 2019/4/1至2020/1/26 | A市A2区丽发新城小区 | 混凝土搅拌站粉尘和噪音污染严重 |
| 4 | 4 | 309.78 | 2019/4/11 | A市金毛湾 | 配套入学文件至今没有发布 |
| 5 | 5 | 124.11 | 2019/8/23至2019/9/6 | A4区绿地外滩小区 | 小区离长赣高铁太近，噪音影响严重 |