“智慧政务”中的文本挖掘应用

摘要

近年来，随着网络的开发，网络问政已是常见的事，因此，对于网上信息的整合逐渐变成重要并且有难度的事。而此次的问题，是对问题的分类，选择重要的问题，和对留言回复做一套评估分析。

本次建模，我们首先使用 SPSS Modeler 来实现问题的分类，再配合 python 实现后面两个问题，实现以下的目标：

1.利用所给的数据，使用 SPSS Modeler 软件，通过已给的信息和标签，训练出一套可以自主分类信息的模型。

2.先通过点赞数选出热门问题，再对关键词进行筛选，通过关键词筛选的模型，对选出来词语进行编号。再通过杰卡德系数来比较两句话之间的相似性，得到与热门问题类似的问题。

3.对答复意见的相关性、完整性和可解释性做权重分析，并设定标准，对每个答复意见都输出答复意见质量指数值（QRI），以此作为评价标准。

目录

1.问题 1 的分析方法与过程

1.1.问题分析

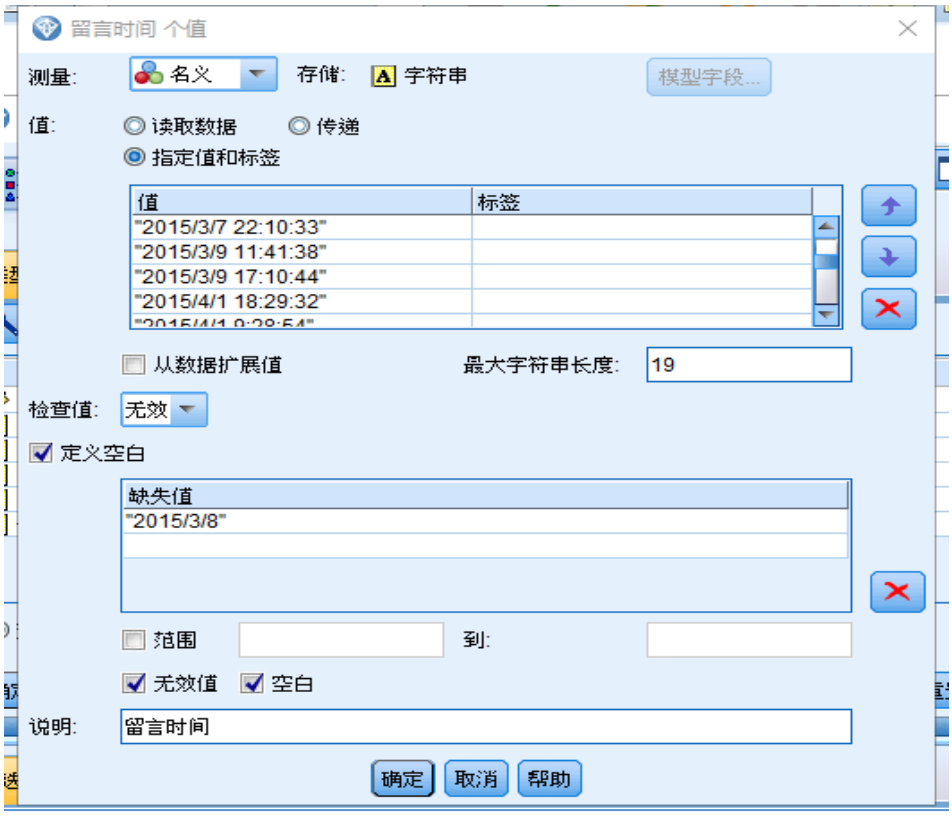
我们组从附件 1 当中已经知道了各级标签的分类，由于利用人工对每条留言进行分类太浪费时间，也容易造成工作人员的疲劳工作，所以问题要求我们利用附件 2 给出的数据，建立一个关于留言内容的一级标签分类模型。我们组对于问题的理解是我们需要利用附件 2 给出的数据，对数据进行训练和测试，得出一个准确率较高的一级标签分类模型，之后我们就能够利用这个模型进行留言的一级分类，减轻工作人员的工作量。

1.2.软件

对于问题 1 的做法我们组想了很久，也讨论了很多种方法，最后我们选择用 SPSS Modeler 软件来帮助我们完成第一问。主要的原因是因为我们最近正在学习这个数据挖掘软件，运用会比较熟练；另一方面，我们觉得 SPSS Modeler 的操作界面简单易懂，而且不需要进行太多复杂的代码编写，可以直接利用 SPSS Modeler 中的模型帮助我们建模，能够节省不少时间。所以综合以上的因素，我们选择了 SPSS Modeler 来进行第一问的解答。[2]

1.3.数据预处理

确定了我们要用的软件和解题方向之后，我们就可以将要用到的附件 2 数据导入到 SPSS Modeler 中了。这里有一个小插曲，由于我们使用的是 SPSS Modeler 14.1，所以如果我们要将 Excel 文件导入 SPSS Modeler 中的 Excel 源节点中，我们就要先将附件 2 保存为 1997-2003 版本的 xls 文件，然后才能正常导入和读取数据。在导入数据之后，我们发现很多数据有着各种各样的问题，都是不能直接被采用的，所以我们进行了数据的预处理。由于在源节点后面加入类别读取数据之后，我们发现有一个时间数据出现了问题，缺少了时间，所以我们决定在类别节点里面直接对其进行处理，将出问题的数据定义为无效，如下图：



图一

另外我们在后面运行模型时，发现另一个巨大的问题，就是留言详情的部分，出现了大量的空格，于是我们决定在附件 2 中直接进行更改，将空格的部分进行替换，想以此来将所有的空格去掉，使得后面的建模能够更加顺利。

在进行完这一系列的操作之后，我们的数据预处理部分算是大致上完成了。

1.4.数据的划分

由于我们是利用附件 2 的数据建立一个一级标签分类模型，所以我们需要对附件 2 的数据进行划分，分为训练集和测试集。在 SPSS Modeler 中我们可以利用分区节点对所有数据进行划分。我们组在讨论过后决定将数据集划分为 80%的训练集和 20%的测试集，随机数种子设定为 1234567，设定如下图：

图二

1.5.模型的运用

在这里我们组产生了较大的分歧，最后我们决定了用神经网络模型来建立一级标签分类模型。我们将留言主题作为输入数据，将一级标签作为输出的目标，然后代入 SPSS Modeler 提前给出的神经网络模型中进行训练。

但是在经过真正的模型训练之后，我们发现利用这个神经网络训练一级标签分类模型存在着几个严重的问题。首先，训练时间是在太久了。我们将留言主题一个数据当作输入数据，SPSS Modeler 的神经网络都要运行接近一个小时。而当我们尝试将留言主题和留言详情两个数据作为输入，运行模型发现用了将近三个小时都无法得出一个模型。另一个主要的问题，也是最重要的问题就是用神经网络训练出来的模型，准确率实在是不堪入目，只有 20%多的准确率实在是拿不出手，下图是神经网络训练的模型结果：

图三

一级标签 的分类

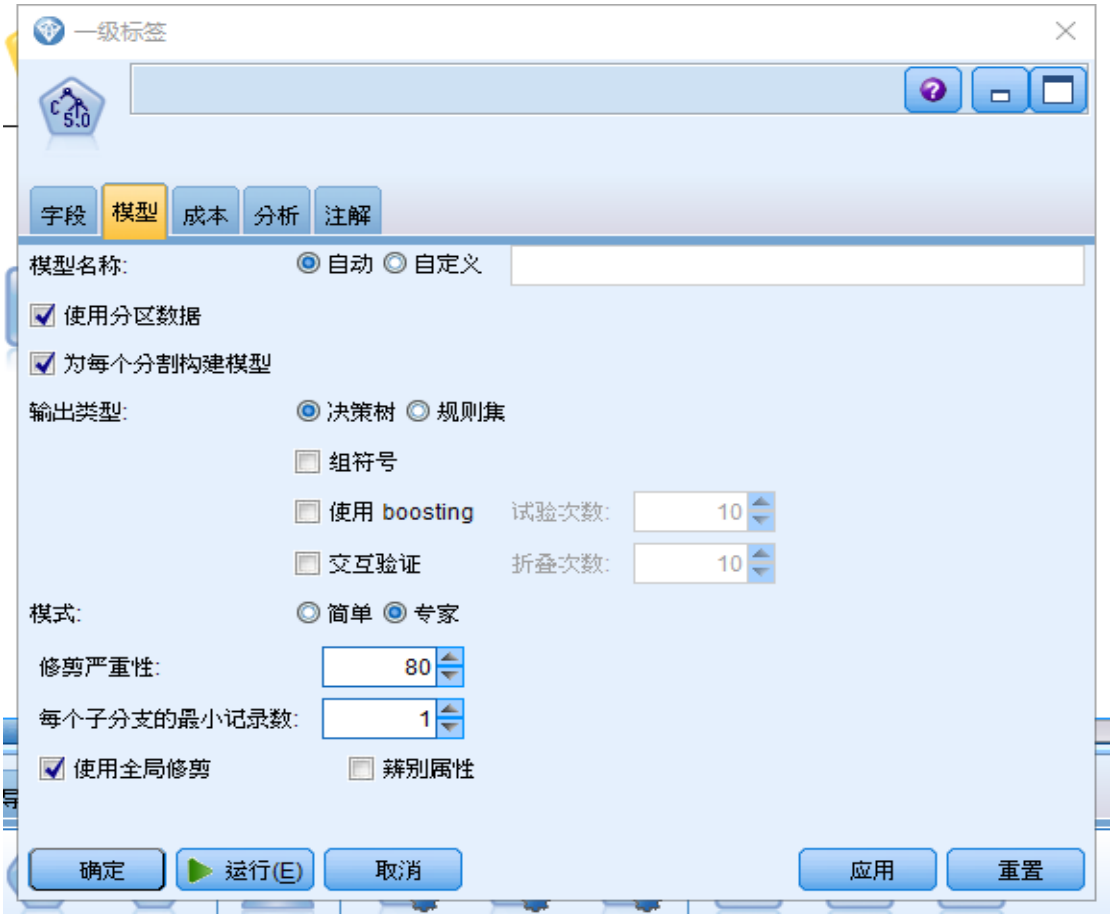
总体百分比校正 = 20.7%

观察	预测						
	交通运输	劳动和社会保障	卫生计生	商贸旅游	城乡建设	教育文体	环境保护
交通运输	0.0%	4.8%	0.0%	16.3%	34.1%	44.7%	0.0%
劳动和社会保障	0.0%	3.6%	0.0%	13.4%	50.9%	32.1%	0.0%
卫生计生	0.0%	5.1%	0.0%	18.1%	46.7%	30.0%	0.0%
商贸旅游	0.0%	3.1%	0.0%	15.9%	48.5%	32.4%	0.0%
城乡建设	0.0%	3.6%	0.0%	14.1%	52.0%	30.2%	0.0%
教育文体	0.0%	3.2%	0.0%	11.6%	47.8%	37.4%	0.0%
环境保护	0.0%	4.2%	0.0%	16.9%	48.4%	30.5%	0.0%

行百分比
100.00
80.00
60.00
40.00
20.00
0.00

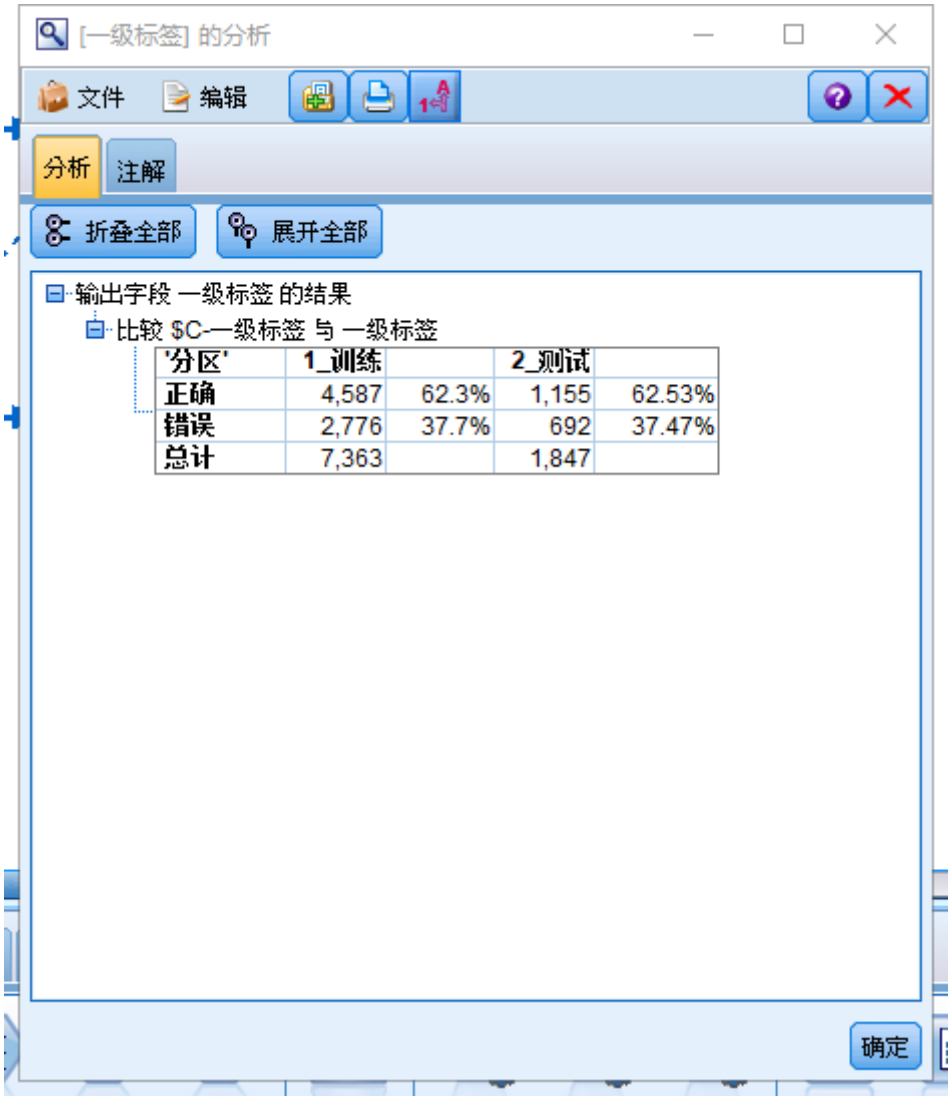
图四

在经过又一轮的思考之后，我们决定舍弃神经网络训练出来的模型，采用另一个 C5.0 模型来对数据进行训练和测试，具体设置如下：



图五

使用 C5.0 模型训练，输入数据我们选择了留言主题，留言详情等，输出同样为一级标签。这次训练和测试所使用的时间大大减少，大概只要一分钟就能输出结果，而且不会占太多的内存。模型的准确率，以及部分数据如下图所示：



图六

一级标签	分区	$C-一级标签	$CC-一级标签
商贸旅游	1_训练	商贸旅游	0.934
商贸旅游	1_训练	商贸旅游	0.934
商贸旅游	1_训练	商贸旅游	0.934
商贸旅游	1_训练	商贸旅游	0.400
商贸旅游	1_训练	商贸旅游	0.400
商贸旅游	2_测试	商贸旅游	0.400
商贸旅游	1_训练	商贸旅游	0.250
商贸旅游	1_训练	商贸旅游	0.794
商贸旅游	2_测试	商贸旅游	0.794

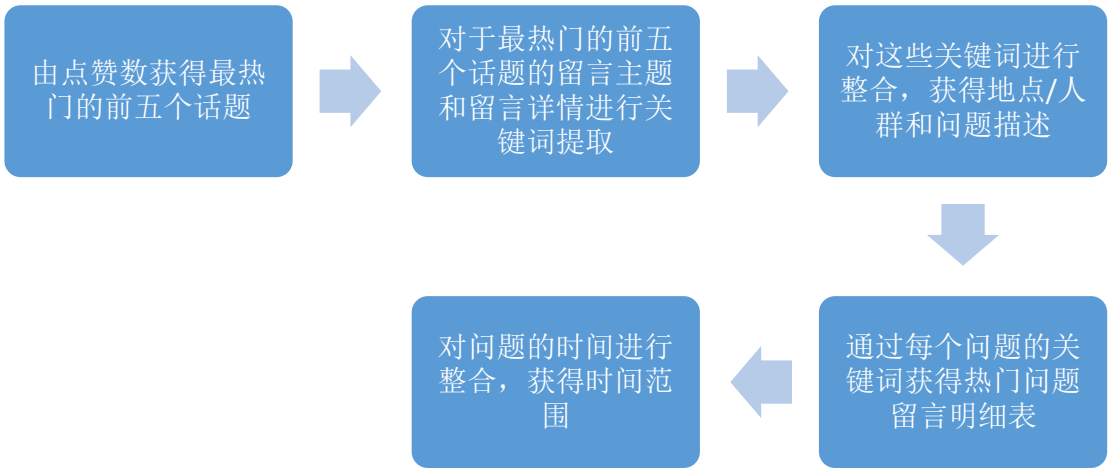
图七

1.6. 模型的不足

模型的不足其实有很多方面。首先也是最重要的是，由于时间不足够以及我们思考得不够全面，所以导致模型准确率不算高。从上表就可以看出来，模型的训练和测试准确率都只有 62%左右，实在算不上高。另一个很大的不足是我们的模型并不算全面，有很多方面都没有考虑到。这些缺陷与不足让这个模型并不算是一个非常好的模型，但是同时也给了我们足够多的提升空间。

2 问题二的分析方法与过程

2.1 流程图



图八

2.2 构造表格的想法

通过对数据的观察，可以得知有赞成票和反对票，而反对票相对来说是比较少的，很多的问题基本是 0 反对票。但是赞成票很多都是有票的，甚至有的很多，这就是说人们会对自己感兴趣的问题来点赞，所以我们认为点赞数是一个很好衡量问题是否是热门问题的标准。因此我们将热门的问题变成寻找点赞数最多的问题。至于第二个表格，是需要根据第一个表格的 5 个热门问题得出它们类似的问题。我们是打算找出每个句子的关键词，然后通过同样的方法找出每个问题的关键词，最后做出比较，找出与热门问题最接近的问题，这些问题就可以归类到第二个表格。

2.3 数据预处理

对于寻找关键词这一步，我们认为是将文字转变成一个个词语，并且把标点符号去除[1]。在网上有类似的模型，能够将文字转变成一个个文字，并且把一些重要的文字提取出来，不重要的文字不采取。比如说一些词如“最大”,“一个”这种大多数都有的词语就会被舍弃。一个简单的例子：亚太地区影响最大的议会间组织——亚太议会论坛 28 日召在符拉迪沃斯托克召开年会，['亚太地区', '影响', '议会', '间', '组织', '亚太', '议会', '论坛', '日召']就会这样子将关键词排序。

然后我们将生成的每个词语进行标号，通过 word2vec 生成词向

量，方便后面的匹配。

2.4 具体步骤

首先，选出点赞数最多的 5 个热门问题，把这几个问题当成最热门的问题。

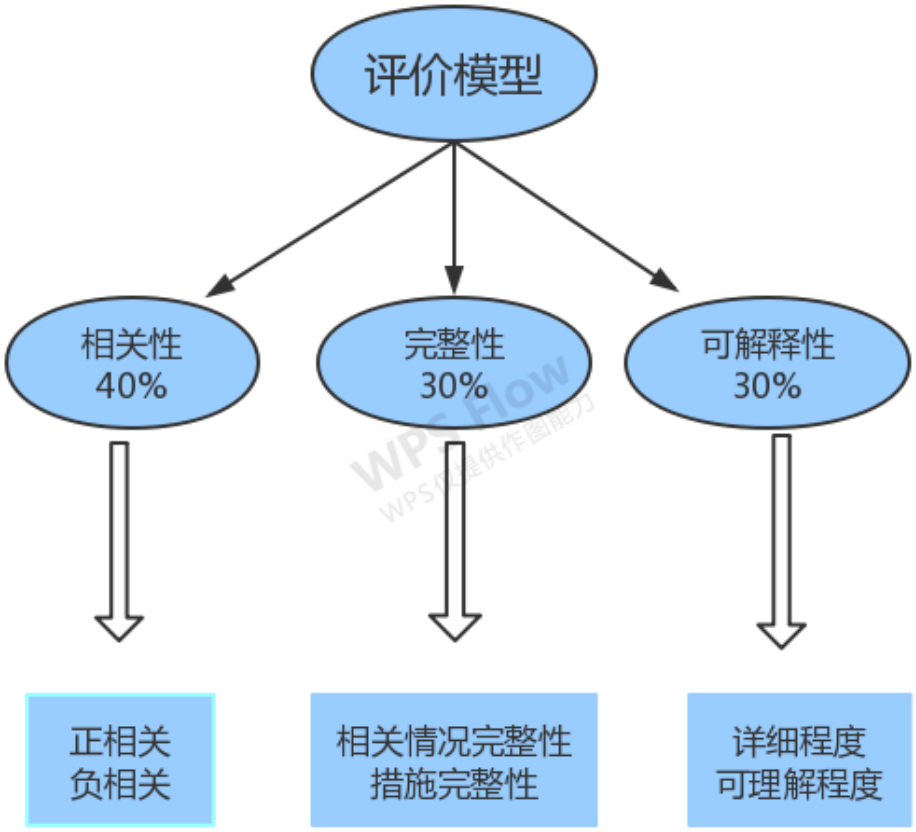
1	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	列1
251	208636	A00077171	汇金路五矿万境K9县存在一系	2019/8/19 11:34:04	狗咬人，请问有人对	0	2097
845	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	纳入配套入学，A3	5	1762
1214	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	案情消息总是失望，	0	821
1385	217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	股东、苏纳弟弟苏	0	790
1485	194343	A000106161	A市58车贷案警官应跟进关注	2019/3/1 22:12:30	侦并没有跟进市领	0	733

之后开始提取关键词，并且对普通问题的关键词也进行提取。

提取到关键词之后，我们通过杰卡德系数来计算句子之间的相似性，用于比较有限样本集之间的相似性与差异性。杰卡德系数值越大，样本相似度越高。杰卡德系数的计算方法是就是两个样本的交集除以并集得到的数值，当两个样本完全一致时，结果为 1，当两个样本完全不同时，结果为 0。所以说，两句话关键词相同的越多 ，那这两句话的关系就越接近。这样，只要我们设定合理的数值，若一个问题与热门问题的杰卡德系数高过这个数值，我们就可以认为这个问题是与热门问题类似的。

3.问题 3 的分析方法与过程

3.1 流程图



图九

3.2 构建答复质量的评价模型

在第三问中，我们建立了一个评价模型来分析答复意见的质量。该模型主要由三大评价指标构成，分别为答复意见的相关性、完整性和可解释性。该模型输出一个答复意见质量指数值（QRI），该值由三大指标的得分相加得到。该值越大，说明答复意见的质量越好；该值

越小，说明答复意见的质量越差。

记相关性指数值为 REI（relativity）、完整性指数值为 IT（integrity）、可解释性指数值为 IP（interpretability）。根据实际情况，我们认为在三大指标中，相关性相对于另外两个指标较重要，所以我们分别为三大指标赋予的权重为 4、3、3。则

$$QRI = \frac{4}{10}REI + \frac{3}{10}IT + \frac{3}{10}IP$$

①相关性：在答复意见中，相关性指的是有关部门给出的答复意见与群众留言的关联程度，通过提取留言和答复的关键字进行判断，将相关性分为正相关、负相关。将留言、答复意见都出现的关键字数认为是正相关，只在留言出现的关键字数认为是负相关。

记正相关指数值为 PC、负相关指数值为 NC。则

$$REI = PC - NC$$

②完整性：在答复意见中，完整性指的是答复意见与留言问题是否对应完整，可分为相关情况完整性、措施完整性。通过提取答复内容的关键字及内容，来判断答复中是否存在留言所提问题的相关情况介绍以及措施处理。

记相关情况完整性指数值为 RSI、措施完整性指数值为 MI。则

$$IT = RSI + MI$$

③可解释性：在答复意见中，可解释性指的是答复相关情况的详细程度、可理解程度。

记详细程度指数值为 DDI、可理解程度指数值为 UDI。则

$$IP = DDI + UDI$$

4.参考文献

[1] 陈文实,刘心惠,鲁明羽.面向多标签文本分类的深度主题特征提取[J].模式识别与人工智能,2019,32(09):785-792.

[2] 万磊,张立霞,时宏伟.基于 CNN 的多标签文本分类与研究[J].现代计算机,2020(08):56-59+95.