“智慧政务”中的文本挖掘应用

摘要：

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。

针对问题 1，我们现将数据导入并进行预处理操作。预处理操作包括了文本去重、去除正则符号、利用 jieba 对文本进行分词处理的处理和去除停用词。为了将不同地名能够很好的识别出来，在 jieba 分词中添加了地名为自定义词典。然后采用朴素贝叶斯算法进行留言分类，分别使用了三种特征：留言主题，留言内容以及两者的结合。最后发现两者的结合分类效果是最好的。

针对问题 2，核心问题是如何判别不同留言的内容是否相同，我们采用的办法是将不同文本留言内容之间进行两两的相似度比较，并设置阈值为 0.4，当相似程度大于 yuzhi 时，则认为为两者留言内容是相同的。对于如何提炼出问题所涉及的特地地点和人群，暂时没有较好的思路完成。

针对问题 3，我们通过了将回复分为三个指标来评论，用完全相关、部分相关、不相关来评论回复的相关性;用 100 分制的评分来评价其完整性和可解释性。实现方法是前期先通过人为为回复从上述三个角度评分，后续通过机器学习算法完成。由于人工评价工作量较大，暂时没有进行代码实现。

关键词：朴素贝叶斯算法；自然语言处理；

一、数据描述

C 题的数据共有四个附件，附件 1 群众回复内容分类三级标签体系。

附件 2 内容为每个留言的主题、对应的留言内容以及标签分类。群众留言一级分类是城乡建设、党务政务、国土资源、环境保护、纪检监察、交通运输、经济管理、科技与信息产业、民政、农村农业、商贸旅游、卫生计生、政法、教育文体、劳动和社会保障共计 15 种分类，涵盖社会各面的问题。

附件 3 内容为每个留言内容以及对应问题的点赞数。

附件 4 则包含了相关部门对留言的回复。

二、数据预处理

我们将数据全部导入 Python。导入后可发现数据中有很多换行、空格等字符。这些换行字符是对我们分类没有帮助的，会影响建立模型。（如下图所示）所以要进行的预处理第一步就是去除这些的正则符号。

```
b['留言详情']
```

```
0       \n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...
1       \n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...
2       \n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...
3       \n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...
4       \n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...
5       \n\t\t\t\t\t\n\t\t\t\t\t我在2015年购买了盛世耀凯小区17栋3楼，...
6       \n\t\t\t\t\t\n\t\t\t\t\t由于西地省地区常年阴冷潮湿的气候，加之近年气...
7       \n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：您好！家住A市A3区桐梓坡西路...
8       \n\t\t\t\t\t\n\t\t\t\t\t我们是梅家田社区辖区内的小区居民，我们每年都...
9       \n\t\t\t\t\t\n\t\t\t\t\t尊敬的A市政府领导：您们好！我是A市A3区魏...
10      \n\t\t\t\t\t\n\t\t\t\t\t尊敬的A市政府领导：您们好！我是A市A3区魏...
11      \n\t\t\t\t\t\n\t\t\t\t\t请求依法监督泰华一村小区第四届非法业主委员会...
12      \n\t\t\t\t\t\n\t\t\t\t\t我住在梅溪湖壹号御湾4楼，自2019年8月份...
13      \n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：你们好！我是A市A4区捞刀河镇彭...
14      \n\t\t\t\t\t\n\t\t\t\t\t地铁5号线施工，导致万家丽路锦楚国际星城小区...
15      \n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：你好！我是A6区润和紫郡的业主，...
16      \n\t\t\t\t\t\n\t\t\t\t\tA市A5区朝晖路锦楚国际新城三区从6月份开始...
17      \n\t\t\t\t\t\n\t\t\t\t\t肯定是选择在A9市西南角，支持A9市西南角设...
18      \n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A6区这几年发展突飞猛进，但城市...
19      \n\t\t\t\t\t\n\t\t\t\t\tA5区楚府线，包括森林雅苑、楚府十城、天际山...
20      \n\t\t\t\t\t\n\t\t\t\t\t我叫涂愈，是一名普通的建筑业从业者，在这里求...
21      \n\t\t\t\t\t\n\t\t\t\t\tA市A2区黄谷路368号山水嘉园1栋三单元7...
22      \n\t\t\t\t\t\n\t\t\t\t\tA市政府、市交警支队、市安监局、市环保局、A...
23      \n\t\t\t\t\t\n\t\t\t\t\tA市、B市、C市要融城，交通是基础。目前长株...
24      \n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：现有一个民生问题跟您反映：近两...
25      \n\t\t\t\t\t\n\t\t\t\t\tA市地铁2号线西延二期建议暂缓修建，一是造价...
26      \n\t\t\t\t\t\n\t\t\t\t\t修建A市火车站(城铁站)经A市火车南站到A市...
27      \n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：现有两个民生问题跟您反映：1、...
28      \n\t\t\t\t\t\n\t\t\t\t\t胡书记，这冬天又到了，A市这湿冷的冬天真是受...
29      \n\t\t\t\t\t\n\t\t\t\t\tA3区银盆南路华韵城市海岸群租房泛滥小区居民...
```

将留言中换行字符和空格字符去除并进行文本去重操作之后，我们再次观察数据。从中可以发现，留言主题和留言内容数据里面出现很多自定义地方名字，如 A 市、A3 区、K9 县等。总结得到数据中的地名规律可分为 3 步：首先是选定大写字母 A-Z 中的一个；第二步共有两种情况，不选择添加数字或者在数组 0-9 中选定一个数字；最后一步是加上市或县或区。于是我们将上面所有添加到 jieba 分词的自定义词典。这样可以保证在后面的操作中地名能很好的划分出来，而不是作为单一的字母出现。

三、问题一

附件 2 提供的数据类别分别是留言编号、留言用户、留言主题、留言时间、留言详情与一级分类。本题的目的是建立关于留言内容的一级标签分类模型，因此重要数据为留言主题、留言内容和一级分类。

分析题目可知，C 题是一个非常明显的文本分类问题，即给定文档 p（可能含有标题 t），将文档分类为 n 个类别中的一个或多个。因此本题采用的解题方法是文本分类中的分类模型构建。文本分类与非文本分类之间的区别，最重要的一点是文本分类需要用到分词。而中文与英文不同，其基本文法具有特殊性：（1）词语之间没有分隔；（2）词与词组之间界限模糊。分析数据可知，群众留言使用语言为中文，大部分句子较长，所含词语较多。

综上所述，我们使用中文分词的 jieba 分词组件的精准模式来进行分割。以下举例分析其分割效果。例句："A 市高新区麓谷街道办事处东塘居委会 3 组雷庆和的堂客周某 65 年出生家中 1 女，2003 年一师拆迁户。"分割后变成 A 市/高新区/麓谷/街道/办事处/东塘/居委会/3/组/雷庆和/的/堂客/周某/65 年/出生/家中/1/女/，2003 年/一师/拆迁户。

由于在留言当中会有很多无意义的词汇，如语气词和问候词等，大多是因为书面礼仪等原因出现，因此其出现的频次较高。这些词语与分类无关，对我们进行分类并没有太多的帮助，所以我们要将这些词汇除去。实现方法是导入一个停用词表（stoplist.txt），当词汇不在这个词表的时候，我们就能继续下面的操作。

上面已经提到，我们将使用文本分类方法进行运算，文本分类在机器学习当中我们比较常用的算法时贝叶斯算法。贝叶斯算法共有朴素贝叶斯算法、TAN 算法等多种分类。朴素贝叶斯算法是贝叶斯分类算法中较为基础和简便的一种分类算法，根据此方法，对一个未知类别的样本 X，可以先分别计算出 X 属于每一个类别 Ci 的概率 P(X|Ci)P(Ci)，然后选择其中概率最大的类别作为其类别。朴素贝叶斯算法的优点包括稳定性高，简便，高效和理论基础强等。朴素贝叶斯算法的

分类质量在很大的程度上取决于构造方法的选择,以及待分类数据的特性和数量。

决定使用什么算法之后，我们便需要选择合适的部分作为训练特征，我们有三个选项，一是以留言主题，二是留言内容，三则是上述的结合。若选用留言主题作为分类特征的话，好处是我们可以比较迅速的完成分类，训练速度较快，但由于留言主题是比较精简的内容，那么一旦留言主题有一点相似的，便会有较大概率会将这些留言进行错误的分类，也就是其查准率会比较低。若是只采用的留言内容作为特征的话，虽然查准率会有一定的提升，但是其训练时间则大幅度增加，训练时间已经比第一种长很多了，所以完全可以将留言主题也放入训练的特征，时间并不会有太大的变化，但是效果会相较于只使用留言内容稍好一点。

最后经过运算可得，该程序实现的正确率为 70%。

四、问题二

第二题为热点问题挖掘，题目的要求是“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”，即得到排名前 5 的热点问题与相应热点问题对应的留言信息。本题的重点在于如何将类似的留言筛选出来将其进行归类。

首先，将每一份群众留言视为一份文件，这样我们可以通过比较任意两两文件的相似度来体现他们的留言是否相同。相似度越高，证明留言相同可能性越大。由于留言较多，如果将所有文件两两进行对比，比较运算量较大，所以完整的将所有文件进行一次运行所花费的时间较长，大约需要 1.5 小时。因此我们认为，缩短运算时间是未来代码可改进的方向之一。

在第一题中，我们所编写的代码已经包含了制作停用词表以及与 jieba 分词有关的部分的操作，接下来将默认已有以上步骤而不再赘述。参考多份资料后，我们最终编写的代码流程如以下三步所示。

第一步，分词处理。在对目标文档的数据进行 jieba 分词处理之后，我们可以把分词后形成的列表展示出来。与此同时，我们把测试文档也进行 jieba 分词操作，并保存在列表中。

第二步，制作语料库。制作语料库的步骤为制作词袋，词袋中的每个词语将会得到一个编号，方便我们的引用。这些编号就是组成二元向量的重要部分。我们将编号和频次数组合起来得到一个二元组，这些二元组作为元素可以组成向量，而向量组合则得到了我们所需要的语料库。

第三步，进行相似度分析。要进行相似度分析，我们首先需要确定合适的相关程度参数。相关程度参数设置是本题解题过程中非常重要的的一步。如果相关程度过高，则能保证编写的程序的查准率较高，但不能其保证查全率；过低则相反，能保证编写的程序的查全率较高，但不能其保证查准率。当相关程度设为大于 0.1 则认为它们是相似的数据时，发现会有很多问题不同但是地名相同的留言

会被错误归为一类。所以我们需要上调参数。当设置相关程度大于 0.5 则认为它们是相似的数据时，可发现每个热点问题中所包含的留言过少，粗略估计每个热点问题只有大约 10 条左右的留言，那么此时得到的热点问题也失去它存在的意义。由此可见相关度设置不能过高也不能过低。经过不断调试，我们认为 0.4 是一个比较好的阈值。当相关度设置为 0.4 时，既保证了一定的查全率，查准率也不会太低。

进行以上三步操作，本小题所需的程序也大致编写完成。需要注意的是，判断问题的热度可以通过每个分类后的问题包含的数量来表示。包含数量越多，则热度越高。得到所有问题的热度排行情况之后，我们将热度前五名的问题数据提取出来分别命名为 ID1-5，即得到热度前五的热度问题及其对应的热度留言信息。

对于如何将各个留言所针对的特定人群或地点，并没有较好的思路如何提取出来，所以暂未完成这部分的内容。

五、问题三

评价相关部门对留言的答复意见，可以从以下几个指标去评定：

1. 相关性，即答复内容与问题的相关度。相关度的大小评定在前期可以由人工完成，评定方法主要是检索与**留言问题的关键词**有关的关键词，如留言提到“扰民”，则在答复内容里寻找有无与“扰民”词义相近的关键词；反映的地方为 A 市，则检索答复内容有无提到 A 市。前期人工对相关性的审核可以给出三类评定：完全相关、部分相关、不相关，然后根据检索有无关键词的方法对模型进行训练。

2. 完整性，即答复内容是否规范，语法是否正确，回答是否完整等。一个完整的答复应包含问题重述、事由解释和解决方案。可以从这三个部分入手作为完整性的主要检测维度，问题重述部分与相关性检测同理；事由解释部分可以检测“因为”“原因”“经调查”等字眼来识别；解决方案部分可以检测“今后”“已经”等字眼识别。另外还有标点符号运用是否准确、是否有错别字等的细节可以判断。前期通过人工进行评定，给出 0-100 的连续变量作为评分，并筛选出各个关键词添加进识别词库，再训练模型。

3. 可解释性，即答复内容的相关解释是否到位。这部分的主观色彩比较强，答复内容既然面对群众，那么用通俗易懂的语言比用专业术语要更优。可以检索答复内容里是否有一些不常用的词语来判断。其次，对每一个问题，可以人工添加几种不同的释义作为参考模板，然后将答复内容的解释部分与之相比对，作为是否解释到位的一个判断维度。前期通过人工评定，给出 0-100 的连续变量作为评分，根据一定量的评定数据去训练模型。

根据这三个指标，加上的对应问题标签，用人工评定好的数据作为训练集训练模型，便可得出答复内容评价系统。

由于数据量庞大，小组人员有限，人工评价答复内容比较麻烦，仅根据三人便对结果进行评价也不够客观，因此仅提供思路，不作具体的实现。

六、总结与展望

在问题一中我们采用的是朴素贝叶斯算法，这个算法在文本分类当中是具有较好的效果的，但是还是有许多不足的，比如无法对其进行调参使得模型达到最好的效果。未来可以用深度学习中 CNN 或 RNN 的算法使得分类达到更好的效果

在问题二中两两相似文本进行比较，随着数据量的增大，所花费的时间也是会急剧增加的，当数据量大量增加后，本文所采用的办法将不再具有较好的可行性。所以未来可以通过数据筛选的方式减少所需要进行比较的数量达到减少花费时间的目的。

在问题三中未来可以想办法将所提出的评价方案进行具体的实现。

七、参考文献

[1]郭勋诚.朴素贝叶斯分类算法应用研究[J].通讯世界,2019,26(01):241-242.

[2]喻凯西. 朴素贝叶斯分类算法的改进及其应用[D].北京林业大学,2016.