基于“智慧政务”中文本挖掘应用探索

摘 要

近年来，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。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。

首先，我们对留言详情进行文本分类，主要方法是运用 Python 的程序代码，将留言去重、分词。参照高频词统计，制作相应的停用词表，并去除停用词。然后，基于附件 1 的 15 个一级类型建立关于留言内容的一级标签分类模型，再测试模型。

其次，我们再对热点问题进行挖掘，具体是指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，需要我们得到两个 Excel 表格，分别是热点问题表和热点问题留言明细表。并且要求是按照题目所给表格样式统计出前 5 的热点问题。

最后，运用我们所建立的模型对留言的答复，从答复的相关性、完整性、可解释性等角度进行分析，给出一套评价方案。

总结，对于本次案例，我们的理解还不够深刻，由于经验不足，出现许多疏漏，感谢各位老师参评！

关键词：文本分类；文本挖掘；NLP；机器学习

Research on text mining application based on “smart government”

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that mainly rely on manual work to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

First of all, we classify the message details by text. The main method is to use Python program code to de duplicate and segment the message. According to the statistics of high frequency words, make the corresponding stop words list and remove the stop words. Then, based on the 15 first level types of attachment 1, the first level label classification model of message content is established, and then the model is tested.

Secondly, we mine the hot issues, specifically, we classify the messages that reflect the problems of specific places or specific groups of people in a certain period of time, define a reasonable heat evaluation index, and give the evaluation results. We need to get two excel tables, namely the hot issues table and the hot issues message list. And it is required to count the top 5 hot issues according to the table style given by the title.

Finally, we use our model to analyze the response of the message from the perspective of relevance, integrity and interpretability, and give a set of evaluation scheme.

In conclusion, for this case, our understanding is not deep enough. Due to lack of experience, there are many omissions. Thank you for your participation!

Key words：Text classification；Text mining；NLP；machine learning

目录

1 前期准备

前期准备我们队伍四月份实现进行自主学习，主要是观看泰迪云课堂里的赛前指导进阶课程 A 题、B 题和 C 题都有去观看，然后最终选定 C 题。

2 问题 1

2.1 问题分析

C 题的第 1 问群众留言分类，我们先进行问题分析。问题 1 的目的是要我们根据所给数据中的附件 2 给出的数据，建立关于留言内容的一级标签分类模型；以及在建立好分类模型之后，需要对使用 F-Score 对分类模型进行评价。要求是该分类模型是需要参考附件 1 提供的内容分类三级标签体系来对留言进行分类。

2.2 解题思路

根据我们观看观察的赛前指导视频，我们的解题思路是附件 1 中所给的数据，我们可以得到给出了 15 个一级标签，但附件 2 的数据中，出现的一级标签只有 7 个类型，所以依据附件 1 建立分词表时，只参考’城乡建设’，’劳动和社会保障’，’教育文体’，’商贸旅游’，’环境保护’，’卫生计生’和’交通运输’对应的二级标签和三级标签。

2.3 解体流程

根据上述分析，我们得到解题流程为：

1、对附件 2 进行处理，将留言去重。

2、绘制标签留言分布的饼图，如下图图 1 所示。

3、根据附件 2 制作分词表，并进行分词

4、参照高频词统计，制作相应的停用词表，并去除停用词。

5、制作词条矩阵的文档。

6、构建分词模型。

7、依据题目给的F_1的计算式，对构建出来的分词模型进行评估和优化。

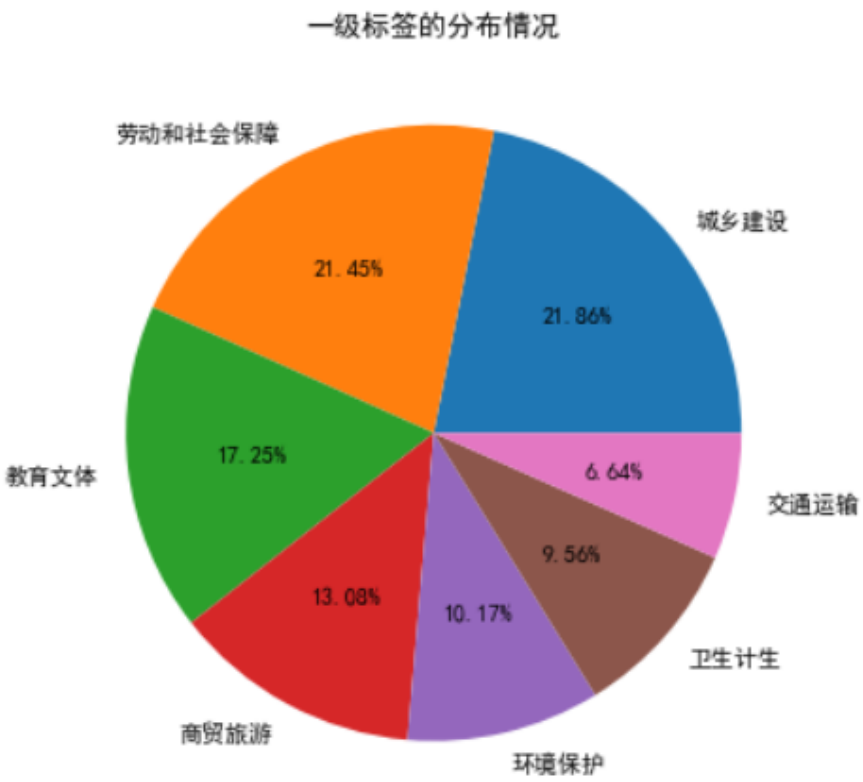

图 1 分布饼图

2.4 解题步骤

具体的实现步骤为：

一、运用 pandas 库实现留言去重，将相同留言去掉。

二、运用 Excel，将一级标签下’城乡建设’，’劳动和社会保障’，’教育文体’，’商贸旅游’，’环境保护’，’卫生计生’和’交通运输’运用替换，分别替换为数字 1、2、3、4、5、6、7，保存并命名为“附件二”，方便后续饼图绘制。

三、运用 jieba 库，对“留言详情”进行分词、去除停用词。

四、定义函数，进行词频统计。

五、运用 sklearn 库下 CountVectorizer，从而建立词条矩阵。

六、将数据划分测试集、训练集（八二分），测试集和训练集的具体属性见下图图 2 所示。

七、分别构建 MultinomialNB 模型和 SVC 模型。

八、通过评判F_1值进行后续改进模型。

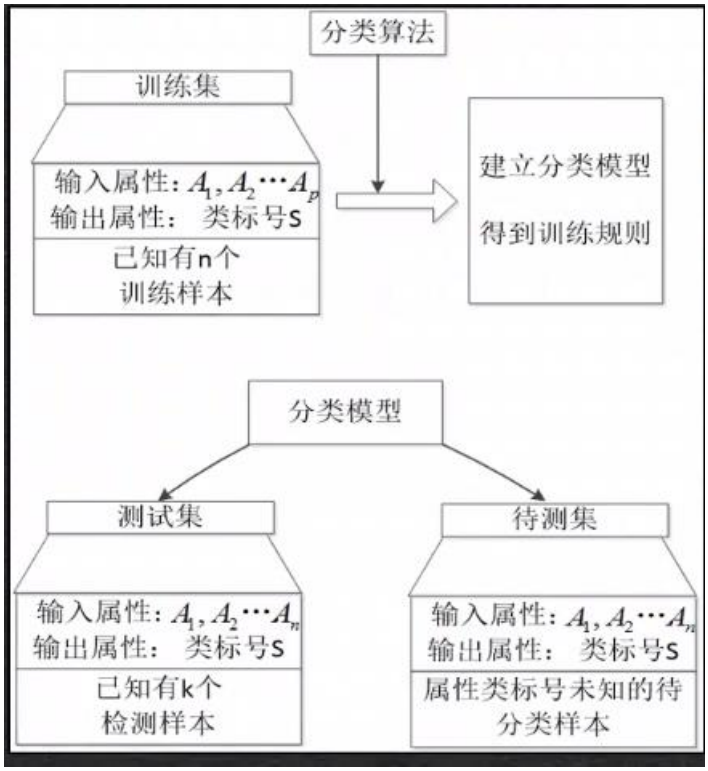

图 2 测试集和训练集

2.5 模型改进

利用F_1值进行评判之后进行模型的修改。开始使用 jieba 库自带分词、停用词表，后在词频统计中发现，仍有较多词语，标签特征不够明显，从而影响到标签分类准确率。而后，根据附件 1 及词频统计，重新制定分词表、停用词表后，模型精度有所提高。

2.6 改进后进行比较

将模型改进之后，我们进行结果对比分析，观察是否能使结果更加精准。通过分析 MultinomialNB 模型的准确率约为 86.44%，SVC 模型的准确率约为 87.04%，两种分类模型准确率较为接近，SVC 模型略优。

比较 MultinomialNB 模型 SVC 模型的F_1值，如下图图 3、图 4 所示。

	precision	recall	f1-score	support
交通运输	0.89	0.57	0.70	126
劳动和社会保障	0.85	0.90	0.87	399
卫生计生	0.90	0.83	0.87	170
商贸旅游	0.87	0.80	0.83	244
城乡建设	0.82	0.89	0.86	403
教育文体	0.88	0.93	0.90	299
环境保护	0.92	0.93	0.93	195

图 3 MultinomialNB 模型的F_1值

	precision	recall	f1-score	support
交通运输	0.85	0.75	0.80	126
劳动和社会保障	0.89	0.89	0.89	399
卫生计生	0.88	0.88	0.88	170
商贸旅游	0.84	0.80	0.82	244
城乡建设	0.81	0.87	0.84	403
教育文体	0.92	0.92	0.92	299
环境保护	0.93	0.91	0.92	195

图 4 SVC 模型的F_1值

对比发现’交通运输’标签的F_1值，在两种分类模型中均为最低，可能原因为相关留言数据过少，学习样本不足，建议后续，增多相关留言，进行学习，相关精确率会有所提升。

其余标签的F_1值对比，SVC 模型略优于 MultinomialNB 模型，所以，在追求准确率的情况下，更建议使用 SVC 模型。

3 问题 2

3.1 问题分析

问题 2 的目的是将附件 3 中进行热点问题的提取，具体是指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，需要我们得到两个 Excel 表格，分别是热点问题表和热点问题留言明细表。并且要求是按照题目所给表格样式统计出前 5 的热点问题。

3.2 解题思路

3.2.1 初期思路：

对“留言主题”、“留言详情”分词，去停用词，做词频统计，筛选出出现次数前五的“地点/人群”词语，再对这五个特定词语下，分别去筛选出出现次数前五的“问题”词语，合并后得出综合出现次数前五的词语，最后得出前五的热点问题。

未能实现原因：不同留言对同一热点问题表述不同，导致在筛选“地点/人群”词语时出现巨大疏漏。例如，“A 市 A5 区魅力之城小区临街餐饮店油烟噪音扰民”，地点应为“魅力之城小区”，但可能有些相同问题留言，并未出现“魅力之城小区”，相应用“魅力城小区”、“魅城小区”、“魅力城”等代替，导致大量相同问题留言未正确区分，因此用上述初期思路，未能实现。

3.2.2 改正思路：

对“留言主题”、“留言详情”分词，去停用词，（去除大量与问题本身无关的背景、影响、建议等），转换 TF-IDF 词向量，两两计算留言之间相似度，确定相似度阈值，导出相似留言（即热点问题）

3.3 Python 程序的实现

目前只进行到去停用词阶段，后续计算相似度方面尚未掌握。

4 问题 3

4.1 问题分析

问题 3 的目的是针对附件 4 相关部门对留言的答复意见，要求是从答复的相关性、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。

4.2 解题思路

具体思路见下图图 5。

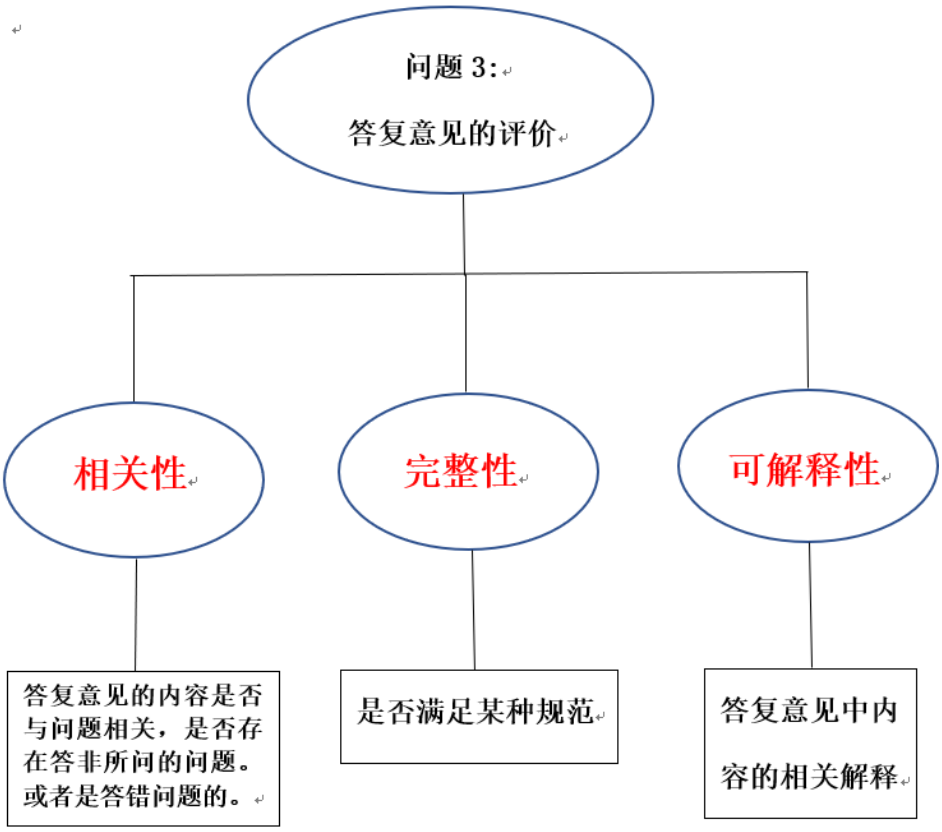

图 5 问题 3 解题思路

1. 对比留言答复与留言详情，分析出词语的相似度，来评判相关性。
2. 答复是需要有一套标准的模式，判断留言答复是否满足标准，有开头有结尾等等。
3. 答复的内容里是否有对问题的解释或理论支撑，引经据典，法律法规等等解释。若问题无法解决，也需要根据某种法律来说明，为什么问题无法解决，给出解释。

4.3 Python 程序的实现

后续步骤由于运行代码问题，还未解决。

5 结束语

经过了一个多月的努力，我们团队完成从前期确认选题，到中期不断讨论、构思还有一些争议，以及后期收尾整理，最后得到了这篇不是很好的论文。虽然我们的论文并没有完成题目的要求，但是这些都是我们队伍三个人努力思考出来的，是独立完成出来的。我们想在这里总结一些，这次的比赛，是我们第一次参加数据挖掘类型的比赛，能看出我们队伍一些问题，例如时间上面的安排、分工上面的安排等等，这些问题是需要我们后续去想办法解决的。我们认识到我们队伍是有很多知识上面的不足的，Python 上面的运用还不是很收敛，并且我们这次主要是在程序代码上面花费很多的时间，也修改了很多次。

对于本次案例，理解不够深刻，经验不足，出现许多疏漏。我们会在后续继续努力，并且去请教老师，将该篇论文继续完成。感谢各位老师参评！