“智慧政务”中的文本挖掘应用

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。因此，将数据挖掘和自然语言处理技术应用于政务系统中具有重要意义。

对于问题 1，首先使用 python 中的 jieba 库对附件 2 的留言内容进行分词并去除停用词以减少无关信息对后续操作的干扰。根据 SVM 算法，训练得出模型。

对于问题 2，比较文本相似度，统计出现频率最高的五个问题，人工提取关键人群、地点信息，归纳出问题描述。

对于问题 3，设计出的方案中共包含四个指标：时效性，相关性，完整性，可解释性。时效性：以三十天为界限，第三十天回复记为 40 分，少于 30 天每提前一天加 2 分，满分 100 分，多于三十天每多出一天减少 1 分，七十天以上均记为零分，视为无效答复。相关性：在答复中检索留言主题中的关键群体和地点是否出现。可解释性：是否使用与逻辑连接和引用有关的连接词、介词。完整性：回复是否满足完整格式。

关键词：**文本预处理 SVM 文本相似度比较 TF-IDF TextRank**

目录

1．研究目标

本次研究目标是利用网络问政平台中的信息数据，利用 jieba 中文分词工具对留言内容和回复进行分词处理，使用 Python 和 Excel 等工具进行分析处理已解决如下三个问题：

1)群众留言分类:使用文本分词和生成词向量等方法进行文本挖掘，使用 SVM 进行分类模型的训练。

2）热点问题挖掘：通过对关键人群及地点的提取对留言内容进行分类归并，按留言条数有多至少排列，取排在前五位的留言主题，人工归纳出问题的描述。

3）答案意见的评价：从完整性、可解释性、相关性、时效性等角度对答复的质量作出评价，制定合适的分类标准。

2．分析方法与过程

2．1 问题 1 分析方法与过程

首先用 pandas 库导入附件 2 中的数据，提取出其中的七个一级分类，分别对应设置一个 id。

	一级分类	一级分类_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

接下来自定义两个函数，一个用于保留文本中的中文、英文和数字，去除标点符号，便于后面简化对数据的处理；另一个用于去除文本中的停用词，如：你、我、他、嗯、呢等对于整个文本意义不大的字词。

	留言详情	一级分类	一级分类_id	清洗后的留言详情
2971	\n\t\t\t\t\t\n\t\t\t\t\t 请想办法取消机场高速公路的收费！ \n\...	交通运输	2	请想办法取消机场高速公路的收费
5437	\n \n 尊敬的彭厅长： 你好！由于近四个月正在和西地省新易迅科技有限公...	劳动和社会保障	4	尊敬的彭厅长你好由于近四个月正在和西地省新易迅科技有限公司进行关于未签劳动合同未缴社保和非法...
5555	\n \n 7月1日起，新修订的《劳动合同法》将正式实施。新法最大的亮点...	劳动和社会保障	4	7月1日起新修订的劳动合同法将正式实施新法最大的亮点就是明确规定了被派遣劳动者享有与用工单位...
4691	\n\n7月4日，这周的星期三晚上。频频出现丑闻的D市D2区皓源实验中学，又一次在教育局的严...	教育文体	3	7月4日这周的星期三晚上频频出现丑闻的D市D2区皓源实验中学又一次在教育局的严令下爆出了暑假...
8397	\n \n 尊敬的张厅长： 你好！ 现在医疗政策很好，但涉及到丙...	卫生计生	6	尊敬的张厅长你好现在医疗政策很好但涉及到丙型肝炎患者还是一个盲区我们这里市级医院没有治疗丙肝...
472	\n\t\t\t\t\t\n\t\t\t\t\t 在C市"五桥头"环境整治和交通优化规划项...	城乡建设	0	在C市五桥头环境整治和交通优化规划项目审查会上五桥头环境整治初步规划方案公布计划通过用座主题...
6369	\n \n 领导大大，我想请问下，报到证到底还有用么？在事业单位还有干部...	劳动和社会保障	4	领导大大我想请问下报到证到底还有用么在事业单位还有干部编制的待遇吗如果没有为什么还搞这个东西
6888	\n \n 您好，我是一名职业教育工作从事者，我想咨询一下贵单位，如何办一个...	劳动和社会保障	4	您好我是一名职业教育工作从事者我想咨询一下贵单位如何办一个职业培训学校我看申请材料里需要申报...
2509	\n\t\t\t\t\t\n\t\t\t\t\t局长好： K2区河东梅湾路通联公司的机房是...	环境保护	1	局长好K2区河东梅湾路通联公司的机房是一个很大的污染源该机房的机械很多机械24小时运转不停没...
6355	\n \n 尊敬的贺厅长： 您好！我是C3县易俗河镇转城户人员，现住花...	劳动和社会保障	4	尊敬的贺厅长您好我是C3县易俗河镇转城户人员现住花果山社区因1992年县址移址易俗河镇征收农...

去停用词的同时，我们引用 jieba 库来对每一条留言进行分词，便于后面提取每个字词的向量。我们采用的是精确模式。

jieba介绍：

一、支持三种分词模式：

- 精确模式，试图将句子最精确地切开，适合文本分析；
- 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；
- 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

二、jieba自带了一个叫做dict.txt的词典，里面有2万多条词，包含了词条出现的次数(这个次数是于作者自己基于人民日报语料等资源训练得出来的)和词性. 这个第一条的trie树结构的词图扫描，说的就是把这2万多条词语，放到一个trie树中，而trie树是有名的前缀树，也就是说一个词语的前面几个字一样，就表示他们具有相同的前缀，就可以使用trie树来存储，具有查找速度快的优势。

三、jieba分词应该属于概率语言模型分词

概率语言模型分词的任务是：在全切分所得的所有结果中求某个切分方案S，使得P(S)最大。

效果图：

一级分类	一级分类_id	清洗后的留言详情	分词后的留言详情
城乡建设	0	A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工...	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
城乡建设	0	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 后 烂尾 多年 ...
城乡建设	0	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 A 市程明 物业管理 有...
城乡建设	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 ...
城乡建设	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 ...

接下来我们导入 sklearn.feature_extraction.text 中 TfidfVectorizer 模块对文本数据完成向量化和 TF-IDF 的预处理。在 ngram_range 参数中，我们设置可以抽取单个词语以外，还可以抽取每个词相邻的词来组成词语对，这样能丰富特征集的数量，有利于提高分类文本的准确度。

对数据预处理完以后，我们开始选用模型进行训练，我们挑选了几个不同的模型进行尝试，最后发现 SVM 中的 LinearSVC 模型准确率最高。

最后的得出模型各个参数，F-score 的得分都在 0.9 左右。

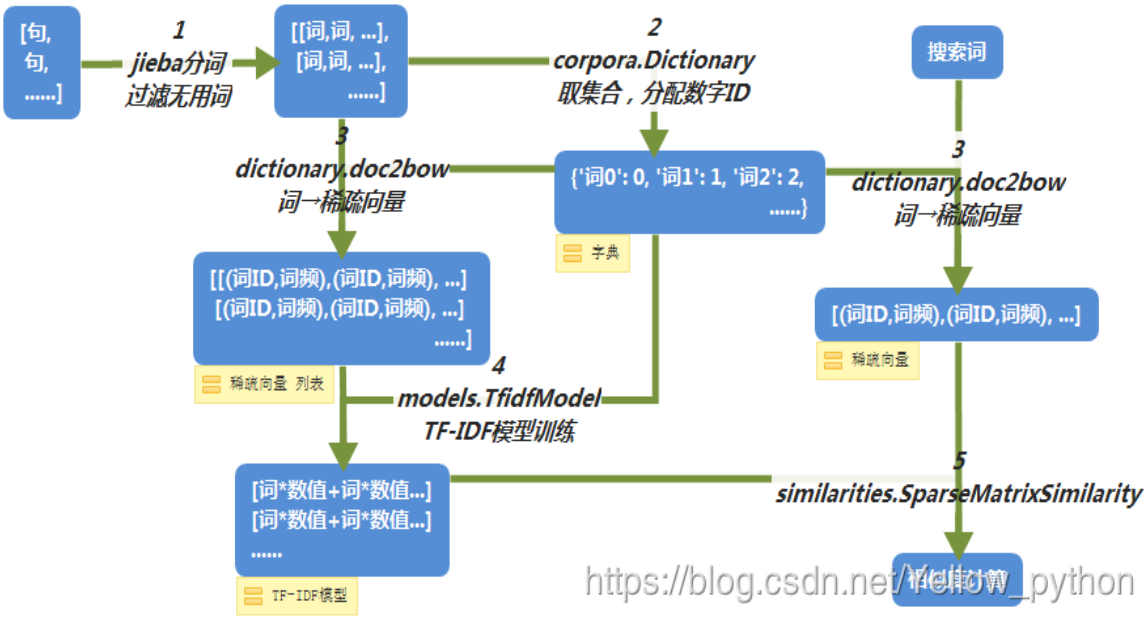
```
accuracy 0.9016447368421052
              precision    recall  f1-score   support

        城乡建设       0.83      0.95      0.88       663
        环境保护       0.95      0.94      0.94       310
        交通运输       0.95      0.70      0.81       202
        教育文体       0.94      0.94      0.94       525
    劳动和社会保障       0.90      0.95      0.92       650
        商贸旅游       0.91      0.83      0.87       401
        卫生计生       0.95      0.84      0.89       289

    accuracy                           0.90      3040
   macro avg       0.92      0.88      0.89      3040
weighted avg       0.91      0.90      0.90      3040
```

2．2 问题 2 分析方法与过程

对数据预处理和第一问相似，去标点符号，去停用词，分词。



第二题我们主要的思想是，在不知道主题数量的时候，我们将附件 3 中的留言依次进行两两对比，相似度近的可以看做是一个主题，从第一条留言开始遍历整个附件 3，如果已经分好主题的留言便无需再进行两两比较，减少了运行的时间。

原理的话首先将其中一条留言作为关键文本，对剩下的留言作为比较文本建立词典，获得辞典特征数，然后用 doc2bow 再转化成稀疏向量。关键文本也转化为稀疏向量，然后创建 TF-IDF 模型，传入刚刚的比较文本，最后一一和关键文本进行比较，得出相似度。经过多次试验，发现相似度>0.1 大多数的文本内容较接近，最后进行归类，以主题内留言数量多少和点赞数多少为判断以及排序出前五的热点问题。

接下来通过人工识别出每个主题的主要内容。

2．3 问题 3 分析方法与过程

数据预处理与前面相似，无需建立 TF-IDF 模型，但首先要通过 Excel 自带的函数来统计每一条回复的时间差。便于后面对时效性得分的计算。相关性得分是基于 jieba 库里的关键字提取模块进行提取五个关键词，用的是 TextRank 方法。提取留言中的五个关键词，然后根据回复之中出现关键词的数量来判断相关性得分。

TextRank算法

TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同, TextRank不需要事先对多篇文档进行学习训练, 因其简洁有效而得到广泛应用。

TextRank 一般模型可以表示为一个有向有权图 G =(V, E), 由点集合 V和边集合 E 组成, E 是V ×V的子集。图中任两点 Vi , Vj 之间边的权重为 wji , 对于一个给定的点 Vi, In(Vi) 为指向该点的点集合 , Out(Vi) 为点 Vi 指向的点集合。点 Vi 的得分定义如下:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS$$

其中, d 为阻尼系数, 取值范围为 0 到 1, 代表从图中某一特定点指向其他任意点的概率, 一般取值为 0.85。使用TextRank 算法计算图中各点的得分时, 需要给图中的点指定任意的初值, 并递归计算直到收敛, 即图中任意一点的误差率小于给定的极限值时就可以达到收敛, 一般该极限值取 0.0001。

3．参考文献

https://www.cnblogs.com/echo-cheng/p/7967221.html

https://blog.csdn.net/weixin_42608414/article/details/88046380

https://blog.csdn.net/laobai1015/article/details/77747702

https://blog.csdn.net/Yellow_python/article/details/81021142

https://zhuanlan.zhihu.com/p/57162092