题目：“智慧政务”中的文本挖掘应用

摘要

近年来，随着网络的快速发展，技术不断提高，逐步进入人们生活中发挥着积极作用，在许多行业提供了更加优质的服务和工作效率。一方面，政府大力支持科技发展，重视大数据、云计算、人工智能等技术的发展；另一方面，当前政府部门通过网络平台进行信息采集，再利用 MATLAB，Excel 等计算机数据挖掘技术对采集回来的群众反应数据进行统计，答复。这将会对提升政府的管理水平和施政效率具有极大的推动作用。

对于问题一，我们在处理网络问政平台的群众留言时，基本上都是依靠人工操纵电子政务系统根据个人的经验进行处理，存在工作量大、效率低、而且差错率高等问题。我们要根据收集到的一些热点问题，建立 Excel 表格，并进行数据分析，建立关于留言内容的一级标签分类模型。

对于问题二，我们在收集到的热点问题的基础上，进行数据的细化，包括对热点问题反映特定地点或特定人群问题的留言进行归类，并且进行数据分析处理，运用自然语言处理方法，定义合理的热度评价标准，得出评价结果。

对于问题三，我们利用大量政府等相关部门对留言的答复意见，运用文本挖掘等技术，结合答复的相关性、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，并且尝试实现。

关键词：建立分类模型　　文本聚类

目录

一、引言

1.1 研究背景

近年来，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。

1.2 研究目标

本次研究目标是在针对各大网络问政平台的群众留言的基础上，首先建立关于留言内容的一级标签分类模型，对附件 2 给出的数据进行一级分类，其次对群众留言进行识别和相关无用，无意义的词进行剔除（又称去除停用词），再采用数据挖掘技术，利用词频（DF)和逆文档频率（IDF)从大量的留言当中提取出比较热点的词语，从而找出相关的热点问题，最后针对附件 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。

1.3 挖掘目标

利用 jieba 中文分词工具对附件内容进行分词,K-means 聚类的方法及 KNN 算法，达到目标：利用文本分词和文本聚类的方法对非结构化数据（附件文档内容）进行文本挖掘，根据聚类结果，分类出众多留言中的热点问题。

二、总体流程

本操作主要包括以下步骤：
1、步骤一：使用 pandas 库读取附件二的“留言主题”，获得初始数据
2、步骤二：使用 jieba 分词对初始数据进行分词，将整个句子切成独立的词块
3、步骤三：用 word2vec 将已经切碎的词块转化成词向量
4、步骤四：通过词频和逆文档频率找出词块中出现次数较为频繁的词语，标记为热点词语
5、步骤五：将各种热点词语的属性归为一类，从而找出留言主题中的热点问题
6、步骤六：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

三、分析方法与过程

1、数据预处理——数据分析——数据筛选
2、TF-IDF 算法——提取关键词——K-means 算法
3、KNN 算法——K-means 算法
4、构建专业语义库——计算与语义库长度
具体分析：
(1) 数据预处理：在题目给出的数据里面，出现了很多重复的数据，在原始的数据上进行去重处理，在此基础上进行中文分词。
(2) 数据分析：在对上述数据进行分词了之后，需要把这些词语转换成向量，以供挖掘分析使用，这里采用 TF-IDF 算法，找出分词后的关键词，并转化为权重向量，采用 K-means 算法对此分类，利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别。
(3) 数据筛选：统计相关数据，分类筛选汇总。

四、TF-IDF 算法

计算词频，即 TF 权重（Term Frequency)
词频=某个词在文本中出现的次数　或　词频=某个词在文本中出现的次数/文本的总词数　或词频=某个词在文本中出现的次数/该文本出现次数最多的词的出现次数
逆文档频率（IDF)=log(语料库的文本总数/包含该词的文本数+1）
TF-IDF=TF*IDF
意义：某个词的文本的重要性越高，TF-IDF 值越大，计算文本中每个词的 TF-IDF 值，进行排序，就能提取出文档中较为关键的词。

五、问题分析

5.1 问题一

在如今的信息化时代，在处理网络问政平台的群众留言时，基本上都还是依靠人工操纵电子政务系统根据个人的经验进行处理，存在工作量大、效率低、而且差错率高等问题。对于提升政府的管理水平和施政效率，重视和发展大数据、云计算、人工智能便会成为必要的事情。

先利用收集到的数据建立 Excel 表格，并对数据进行分析，根据现代化建设的需求和发展，分为一二三级，由高到低，并进行分类。

5.2 问题二

对问题二研究的意义的分析：

问题二属于一个文本挖掘问题，针对这种问题，采用数据挖掘技术，通过词频和逆文档频率找出词块中出现次数较为频繁的词语，标记为热点词语，将各种热点词语的属性归为一类，根据聚类结果，从而找出相关的热点问题。

5.3 问题三

对问题三研究分析。

问题三属于统计数学问题，对于问题三这一类数学问题的分析，对附件 4 中可以依据它具有的一些数据特点和特性，按所给定的问题三要求进行有效的数据分析。由于以上的种种原因，我们首先要建立一个以 KNN 算法和 TF-IDF 算法为基础针对问题三和附件 4 的数学模型 5，如何再根据数据本身具有的类别特性建立一个数据信息分类模型 6，然后通过以上述两种算法为基础，再以 IBM SPSS 具有的数据分析功能得知各关键词，然后再以一个较为简单的 Python 编程建立起来以多次出现的关键词引点，进行不同种类数据的分类，最后进行结合实际的分析，对结果进行预测，并将所得的各个结果进行比较得出最终分析。

六、结论

本文根据附件提供的数据，去除了大量异常的数据，再利用 Excel 表格等工具进行数据挖掘，通过得出的结果，建立关于留言内容的一级标签分类模型。

我们在收集到的热点问题的进行一级标签分类后，包括对热点问题反映特定地点或特定人群问题的留言进行归类，并且进行数据分析处理，运用自然语言处理方法，数据挖掘技术，利用词频（DF)和逆文档频率（IDF)定义合理的热度评价标准，得出评价结果。

我们完善热度评价标准后，利用大量政府等相关部门对留言的答复意见，运用文本挖掘，数据分析和挖掘等技术，结合答复的相关性、完整性、可解释性等角度对答复意见的质量设计出一套完整的评价方案。

参考文献

[1] 百度网站.<泰迪杯数模优秀论文>

[2] https://www.baidu.com/link?url=K9ypFWHUCAyh6QaNiyelyQElmrb6n2oXe6hTm5GcgG6vqw_efzt1dTOFsMa8yhpmu6f4zS94ZeB_CgTsb1SKLK&wd=&eqid=f0320aca000005ed000000065eb4fb1e（百度网站.泰迪杯全国大学生数据挖掘竞赛优秀作品）