第八届“泰迪杯”

全国数据挖掘挑战赛

C 题“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。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。

本文将基于数据挖掘技术对附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见的信息数据进行内在的信息挖掘，提取我们需要进行分析的部分进行深度挖掘和分析。

针对问题一：本文首先对附件 2 中的非结构数据进行了数据的预处理，对数据进行了去除停用词、中文分词、去空等，然后基于 TFIDF 权重法提取 50000 个候选特征词，形成词袋，构造词汇-文本矩阵，由于这种方法具有高维度，高稀疏度以及同义词影响的缺点，因此，本文进一步利用基于潜在语义（LSA）分析的奇异值分解算法（SVD）对词汇-文本矩阵进行空间语义降维，语义压缩后的文本向量被认为投影在了同一空间里，再通过 k-means 聚类算法对职位的职业类型和专业领域进行划分。

针对问题二：本文将附件 2 里的结构化数据对其进行了处理后，对附件 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。及时发现问题，并进行集中整治。

针对问题三：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。

关键词：数据挖掘，文本处理，Python 数据处理

Absrtact: in recent years, with Wechat, Weibo, mayors mailbox , sunshine hotline and other network political platforms gradually become an important channel for the government to understand public opinion, pool the wisdom of the people and rally the public spirit, the amount of text data related to all kinds of social conditions and public opinion continues to climb, which brings great challenges to the work of the relevant departments that mainly rely on manual message division and hot spot sorting in the past. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government. Based on the data mining technology, this paper gives the records of the public political messages collected from the open sources of the Internet, and the information data of the responses and opinions of some of the masses messages by the relevant departments, and abstracts the parts that we need to analyze for in-depth mining and analysis.

In order to solve problem 1: in this paper, the unstructured data in Annex 2 is preprocessed, and the data are removed from stop words, Chinese word segmentation, emptiness and so on. Then, 50000 candidate feature words are extracted based on TFIDF weight method to form word bag and construct vocabulary-text matrix. Because of the shortcomings of this method, such as high dimension, high sparsity and synonym influence, this method has the disadvantages of high dimension, high sparsity and synonym influence. In this paper, the singular value decomposition algorithm (SVD) based on latent semantic (LSA) analysis is further used to reduce the spatial semantic dimension of the vocabulary-text matrix. The compressed text vector is considered to be projected in the same space, and then the occupation type and professional domain of the position are divided by k-means clustering algorithm. In view of problem 2: after dealing with the structured data in Annex 2, this paper classifies the messages reflecting specific location or population problems in Annex 3, defines the reasonable thermal evaluation index, and gives the evaluation results. Identify problems in a timely manner and carry out centralized rectification. In response to question 3: in response to the responses of the relevant departments i

n Annex 4 to the messages, a set of evaluation programmes are given for the quality of the responses from the perspectives of relevance, completeness and interpretability of the responses.