第八届“泰迪杯”全国数据挖掘挑战赛论文报告

所选题目：C 题 ： “智慧政务”中的文本挖掘应用

基于“智慧政务”群众留言信息分析与建模

综合评定成绩：_____

评委评语：

评委签名：

基于“智慧政务”群众留言信息分析与建模

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。因此，我们设计了一套算法流程，用于完成群众留言提取、结构相似，模型建立，完成群众留言分类以及有效答复，提高效率，建立智慧政务，对于问题一，通过利用 excel 分类将群众留言进行分类筛选，根据计数的内容去分类到相应的部门系统，并且我们通过 DOM 树建立风格树，对群众留言进行分解，得到了较好的留言正文区域结构。。对于问题二，通过 PositionId 对招聘信息表、职位描述表进行去重，得到不重复的招聘职位信息。然后用 jieba 中文分词工具对群众留言信息进行分词，并通过 TF-IDF 算法提取每个群众留言的 5 个关键词。再利用 TF-IDF 算法得到每个群众留言的 TF-IDF 权重向量，然后进行热点聚集。对于问题三，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系列评价。

关键词：风格树，DOM 树，jieba，TF-IDF

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. Therefore, we have designed a set of algorithm flow, which is used to complete the extraction, structure similarity, model establishment, classification and effective response of the mass message, improve efficiency, and establish intelligent government.For the first problem, we use Excel to classify and filter the mass message, and classify it to the corresponding department system according to the content of the count. Moreover, we build a style tree through DOM tree to decompose the mass message, and get a better message text area structure.. For the second problem, position ID is used to de duplicate the recruitment information table and position description table to get the non repetitive recruitment position information. Then we use the Chinese word segmentation tool of Jieba to segment the message of the masses, and extract 5 keywords of each message of the masses by TF-IDF algorithm. Then, TF-IDF algorithm is used to get TF-IDF weight vector of each message, and then hot spots are gathered. As for question 3, in view of the reply opinions of the relevant departments in Annex 4 to the message, a series of comments are given on the quality of the reply opinions from the perspectives of relevance, integrity and interpretability

Keywords: Style tree, DOM tree, Jieba, TF IDF

目录

1．挖掘目标

本次建模目标是利用附件中数据，利用 jieba 中文分词工具对群众留言进行分词、K-means 聚类的方法及 KNN 算法，达到以下三个目标：

1）利用文本分词和文本聚类的方法对非结构化的数据进行文本挖掘，结果，结合群众留言的特点分析目前热门问题。

2）根据分析出来的的数据，筛选出热点问题。留给相关部门并给出专业的建议和有效的解决方案

2. 分析方法与过程

2.1 问题分析

2.1.1 问题一

通过利用 excel 分类将群众留言进行分类筛选，根据计数的内容去分类到相应的部门系统，并且我们通过 DOM 树建立风格树，对群众留言进行分解，得到了较好的留言正文区域结构。

2.1.2 问题二

通过 PositionId 对招聘信息表、职位描述表进行去重，得到不重复的招聘职位信息。然后用 jieba 中文分词工具对群众留言信息进行分词，并通过 TF-IDF 算法提取每个群众留言的 5 个关键词。再利用 TF-IDF 算法得到每个群众留言的 TF-IDF 权重向量，然后进行热点聚集。

2.1.3 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系列评价。

2.2 总体流程

本用例主要包括如下步骤：

步骤一：数据预处理，在题目给出的数据中，出现了很多重复的留言数据，在原始的数据上进行去重处理，在此基础上进行中文分词。

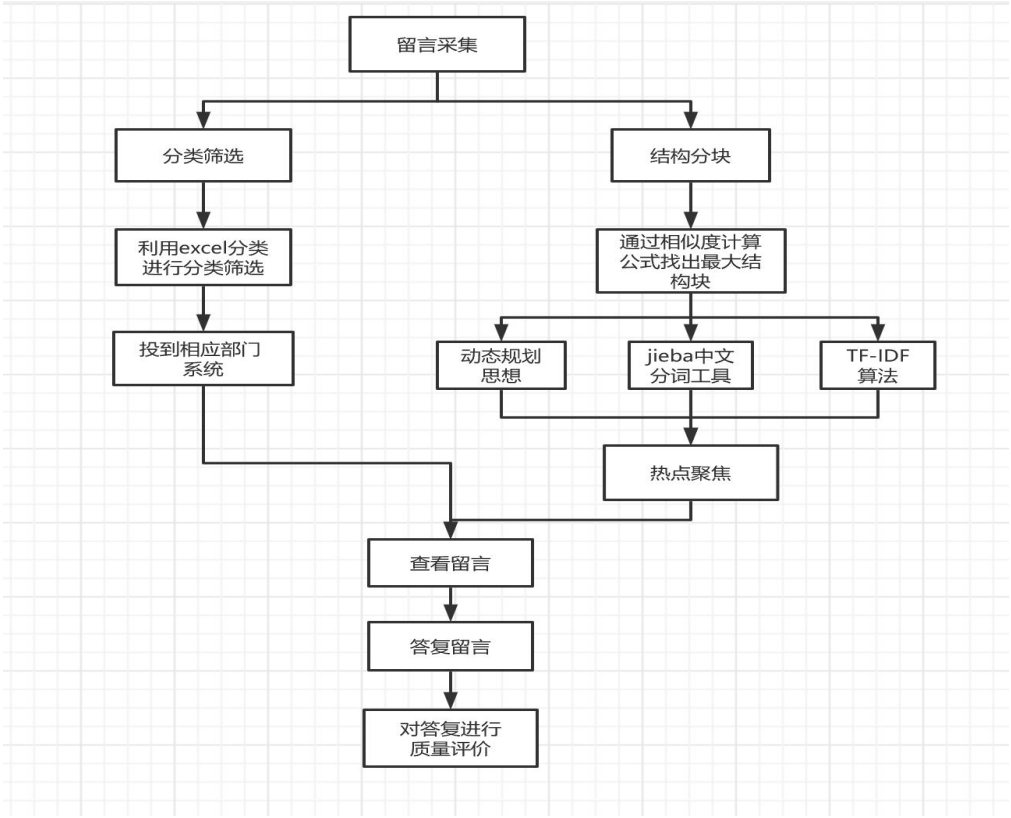
步骤二：数据分析，在对描述信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，找出每个热点问题描述的关键词，把描述信息转换为权重向量。采用 K-means 算法对热点问题进行分类，利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别。

步骤三：数据筛选，统计相关数据，分类筛选汇总，预测热门的问题、大众需求走向和相关问题的答复情况等。

步骤四：利用步骤一的结果构建专业语义库，通过计算与语义库的距离，找出对应的 IT 职业的 ID，统计留言分布情况。

2.3 问题 1 分析方法与过程

2.3.1 流程图



2.3.2 流程

对留言信息表相同留言去重（positionld 相同的取最新更新状态）统计留言类型种类 K 统计聚类结果中各个类别留言类型的数量，以数量最多的为该类类型原始数据

1. 留言信息表
2. 留言描述表利用 TF-IDF 与余弦相似性进行 k-means 聚类，对留言描述表相同留言去重（positionld 相同的取最新更新状态）jieba 中文分词

2.3.2 数据预处理

1.招聘信息的去重、去空

在题目给出的数据中，出现了很多重复的留言数据。例如留言信息表跟留言描述表中出现了很多重复的留言信息。考虑到各个留言平台可能每天都会对要留言信息进行更新，因此在去重的时候应该取更新时间最晚的记录，去掉历史记录。考虑到python 中的字典在保存数据时，key 相同的内容，value 取值为最后更新的值。因此在读取数据时，按时间升序把留言信息的 PositionId 作为 key，把整个留言信息作为 value 保存在 value 中。最后再将字典中的内容写入文本即可。同时在留言描述表中出现了留言描述为空的记录，干扰了问题的分析，采取直接滤过方法，从文本中删除

2.对留言信息表进行中文分词

词频（TF）=某个词在文本中的出现次数

该文本出现次数最多的词的出现次数

3.计算 IDF 权重，即逆文档频率（InverseDocumentFrequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

逆文档频率（IDF） log(语料库的文本总数)包含该词的文本数+1

4.计算 TF-IDF 值（TermFrequencyDocumentFrequency）。

TF-IDF=词频（TF）×逆文档频率（IDF）

实际分析得出 TF-IDF 值与一个词在留言描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的留言描述表中文本的关键词。

5..生成 TF-IDF 向量

生成 TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找出每个留言描述的前 5 个关键词；

（2）对每个留言描述提取的 5 对于这个集合中词的词频，如果没有则记为 0；

（3）生成各个留言平台描述的 TF-IDFTF-IDF=词频（TF）×逆文档频率（IDF）

6.留言类型的分类

生成留言描述的 TF-IDFTF-IDF 权重向量，对 7 类。

X={x1,x2, ,xi, ,xn}，其中 xK 个划分 C={ck,i=1,2, ,K}。 μi。选取欧式距离作为相似μi 的距离平方和（1）

K 聚类目标是使各类总的距离平方和 J(C)=KK2J(ck)最小， ∑k=1

J(C)=J(ck)=∑∑∑kkxC=1=1xi-μi=i∈idkixi∑∑ki=1=1Kn2-μi（2）

⎧ 1,若 xi∈ci 其中，dki=⎨ ,所以根据最小二乘法和拉格朗日原理，聚类中心μk⎩ 0,若 xi∉ci

应该取为类别 ck 类各数据点的平均值。

K-mean 聚类的算法步骤如下：

1、从 X 中随机取 K 个元素，作为 K 个簇的各自的中心。

2、分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

4、将 X 中全部元素按照新的中心重新聚类。5、重复第 4 步，直到聚类结果不再变化。6、将结果输出。

K-mean 聚类的算法流程如下：

1）选定数据空间中 K 个对象作为初始聚类中心，每个对象代表一个类别的中心

2）对于样品中的数据对象，则根据它们与这些聚类中心的欧氏距离，按距离最近的准则分别将它们分配给与其最相似的聚类中心所代表的类

3）计算每个类别中所有对象的均值作为该类别的新聚类中心，计算所有样本到其所在类别聚类中心的距离平方和，即 J(C)值

是

4）聚类中心和 J(C)值发生改变？

否

聚类结束

聚类算法流程

由于留言描述表给出了 539216 条记录，去重后还有 402727 条记录，如果把所有的留言信息都用来挖掘分析，会占用很大的机器性能跟时间。为了节省机器性能跟时间，获得结果。从 402727 条记录数据中随机抽取 40000 条记录，然后利用抽样样本进行分词、求 TF-IDF 向量，并利用 K-mean 聚类，把样本分成 7 类

7..Knn 最邻近分类算法

由 K-Means 分类得到聚类中心，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。根据向量空间模型，将每一类别文本训练后得到该类别的中心向量记为 Cj(W1,W2, ,Wn)，将待分类文本 T 表示成 n 维向量的形式 T(W1,W2, ,Wn)，则文本内容被形式化为特征空间中的加权特征向量，即 D=D(T1,W1;T2,W2; Tn,Wn)文本的相似度，找出 K 个最相似的文本。

具体算法步骤如下：

（1）对于一个测试文本，根据特征词形成测试文本向量。

（2）Sim(di,dj)=Wik∑M× Wjk 式中，di 为测试文本的特征向量，djM 为特征向量维数；Wk 为向量的第 k 维。k 然后根据实验测试 K 的结果来调整 K 值。

（3）k 个文本。

（4）在测试文本的 k(di)y(di,Cj)-b≥0d∉ P(X,0,其他 Sim(x,di)为相似度计算公式；b 为阙值，y(di,Cj)的值为 1 或 0，如果 di 属于 Cj，则函数值为 1，否则为 0。

（5）比较类的权重，将文本分到权重最大的那个类别中。

8..分析留言类型和初步定义热点问题

对附件 2，3，4 根据 K-means 聚类方法和 Knn 最邻近分类得出 7 个点和每个点周围 100 个 id，根据这些 id 对照附件 2，3，4 所属的留言 PositionFirstType，包括留言时间，留言问题，留言详情三大分类，统计数量出现最多的关键词即为目前热点问题

2.4 问题 2 分析方法与过程

2.4.1 数据筛选

（1）根据留言信息表对不同问题进行分类筛选，得到物业问题、交通问题、上学问题、环境问题等 21 个不同领域问题。

（2）根据留言信息表对不同问题进行分类，得到 2909 个不同的问题。

（3）根据留言信息表对不同的留言所属大类分类，分为学业、住宿、环境、工作、外地人落户、网络信号、办理手续七大类。

2.4.2 数据统计

定义热点问题

对各个热点进行分类计数

2.5 评价方案

2.5.1 评价宗旨

以为人民服务为宗旨，建设服务型政府。进一步深化政府与人民的沟通，促进政府更好地了解民意，解答民惑，接受人民监督。

2.5.2 指导思想

政府及时地处理人民留言，汇聚民智、解答民惑，对提升政府的管理水平和施政效率具有极大的推动作用。政府如何对民意作出答复是其中极为重要的环节。为此，我们将从答复的相关性、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。

2.5.3 评价指标

答复意见的评价从答复的相关性、完整性、及时性、可解释性 4 个角度出发，并按不同比例计入等级。

2.6 评价说明：考核采用等级制

满分为 100 分，总分 85 分以上评价等级为 A,总分 70 分以上评价等级为 B，总分 70 分以下为 C 等级

2.6.1 答复的相关性（25%）

评价内容： 是否找准留言问题关键，明确留言意图与目的，围绕主题进行解答。
评价方法：根据关键词的重复与相关名词进行判断

2.6.2 答复的完整性（25%）

评价内容：是否逻辑连贯并完全地表述出所需表达的内容
评价方法：根据语言前后的逻辑关系判断是否完整答复留言

2.6.3 答复的及时性（25%）

评价内容：是否及时回复网友留言
评价方法：根据网友留言时间与答复时间的间长进行判断。
3 天之内：25 分
一周之内：[20,25)分
14 天之内：[18,20)分
14 天以后：15 分及以下

2.6.4 答复的可解释性（25%）

评价内容：回答内容是否符合科学发展观，是否与实际情况相符合，是否具有正确理由作为依据
评价方法：根据答复的合理性以及符合现实的程度进行判断

2.6.5 结论

根据此评价方案，对答复意见质量进行评价（数据来自于附件 4 结论见附件 1）

3.结论

对留言信息进行分析研究，了解人民群众的需求和意见，对相关部门工作有重大意义，同时也是文本分析的一个课题、本解读已经不能满足数据量庞大的留言信息。本文采用根据法和 Knn 深入分析各个平台的留言信息现状。

由分析结果可以看出，进行计数分类各个板块的留言，从而各个平台用户留言状况，序得出各大平台的留言意见情况，并定义热点问题，联网相关答复意见的评价。

4.参考文献

[1]赵琳瑛..西安电子科技大学.2007
[2]K_means 文本..2014
[3]电子科技大学.硕士学位论文
[4]K-means 聚类算法研究综述.2012
[5]应用特征聚合进行中文文本分类的改进 KNN 算法
[6]卜凡军.KNN 算法的改进及其在文本分类中的应用.江南大学.硕士学位论文.2009
[7]曹卫峰.中文分词关键技术研究.南京理工大学.硕士学位论文.2009
[8]杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007