1.背景介绍

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。

目标挖掘

（1）对中文信息进行中文分词的处理提取关键词，将文本信息利用 IF-ILD 算法转化为向量，建立关于留言内容的一级标签分类模型，使用 F-Score 对分类方法评价。

（2）请根据附件 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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”对热点问题的挖掘，词语的出现频率以及时间。

（3）根据热点问题，对热点问题进行可靠的回复、完整的评价方案。

2.1 问题 1 分析方法与过程

2.1.1 数据处理

1.对留言信息进行中文分词和去除停用词

在对信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。给出附录 2 中的中文信息，为了便于转换，先要对这些职位描述信息进行中文分词。这里采用 pycharm 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)，同时采用了动态规划 查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。

去除停用词，在中文分词后有部分词语是无用信息，这时我们要建立一个停用词表“stopwords”除去形容词、英文和无用标点符号。

在分词的同时，采用了 TF-IDF 算法，在附录 2 中抽取每种一级标签（城市建设、交通运输...）下留言内容中的前 5 个关键词，这里采用 jieba 自带的语义库。

2.1.2 信息转化为向量（TF-IDF 算法）

在对留言信息分词后，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，把职位描述信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

词频（TF）= 某个词在文本中出现的次数 （1）

考虑到留言有长短之分，为了便于不同留言的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频（TF）} = \frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}} \quad （2）$$

或

$$\text{词频（TF）} = \frac{\text{某个词在文本中的出现次数}}{\text{该文本出现次数最多的词的出现次数}} \quad （3）$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需建立一个语料库（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）

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF） （5）

实际分析得出 TF-IDF 值与一个词在留言文本中出现的次数成正比,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职位描述表中文本的关键词。

生成 TF-IDF 向量

（1）使用 TF-IDF 算法，找出每个留言一级分类描述的前 5 个关键词；

（2）对每个留言一级分类描述提取的 5 个关键词，合并成一个集合，计算每个留言一级分类描述对于这个集合中词的词频，如果没有则记为 0；

（3）生成各个留言一级分类描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF） （6）

2.1.3 模型的建立

根据附录 2 的七种一级分类,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。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)，对于一个测试文本，计算它与样本集（附录 2）

中每个文本的相似度，找出 K 个最相似的文本，根据加权距离和判断测试文本所属的类别。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） 计算该测试文本与训练集中每个文本的文本相似度，计算公式为：

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^{2}} \sqrt{\sum_{k=1}^{M} W_{jk}^{2}}}$$

d_i 为测试文本的特征向量,d_j 为 j 类的中心向量；M 为特征向量维数；W_k 为向量的第 k 维。k 值的确定一般先采用一个初始值，然后根据实验测试 K 的结果来调整 K 值。

因为在二分类的问题下，对于一个新样本 x，可以用下式求解该样本的分类 y：

$$y = \arg\max_{c_j} \sum_{(x_i, y_i) \in N_K(x)} f_{c_j}(y_i)$$

其中 Nk(x)表示距离样本 x 最近的 k 个样本的集合,f 为关于 yi 的指示函数：

$$f_{c_j}(y_i) = \begin{cases} 1, & y_i = c_j \\ 0, & y_i \neq c_j \end{cases}$$ 其中 Cj 是中心向量。

本题则是七分类的问题，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，将每个分类下的中心向量运用迭代的方法逐一确立 k 值。

（3） 按照文本相似度，在训练文本集中选出与测试文本最相似的 k 个文本。

（4） 在测试文本的 k 个近邻中，以此计算每类的权重，计算公式如下：

$$P(X, Cj) = \begin{cases} 1, \text{若} \sum \mathrm{Sim}(x, d_i)\, y(d_i, C_j) - b \geqslant 0 \\ 0, \text{其他} \end{cases}$$

最后用 F-Score 对分类方法评价： $$F_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2P_i R_i}{P_i + R_i},$$

其中 Pi 为第 i 类的查准率， Ri 为第 i 类的查全率。

2.2 问题 2 分析方法与过程

2.2.1 数据筛选

（1）根据附件三将某一时间段内反应特定地点或特定热播群问题的留言进行分类得到排名前五的热点问题

（2）根据问题本身的热度，通过留言问题点赞数热度量化，再加上与提出这一问题的时间段挂钩。

（3）在上面的基础上，找到社会共性热点，对留言进行主题抽取，在所有留言中针对抽取的关键词之类的做关联，把关联起来的留言同样量化热度到具体某个相关的里头，形成了同个主题的焦点问题。

2.2.2 数据统计

（1）通过数据筛选排序选出排名前五的留言定义为热点问题

（2）这五大热点问题分别是关于 A 市 A5 区魅力之城小区，A 市经济学院学生，A 市公民，A 市南塘城轨和梅溪湖 CBD，A 市公民提出的。

2.2.3 排名指标的确定

通过 NDCG 归一化折损累积增益确定。NDCG 允许以实数形式进行相关性打分。这种关系类似分类和回归的关系。NDGG 的两个思想如下：

1、高关联度的结果比一般关联度的结果更影响最终的指标得分；

2、有高关联度的结果出现在更靠前的位置的时候，指标会越高。

累计增益（CG）

CG，cumulative gain，是 DCG 的前身，只考虑到了相关性的关联程度，没有考虑到位置的因素。它是一个搜素结果相关性分数的总和。指定位置 p 上的 CG 为：

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

reli 代表 i 这个位置上的相关度。

举例：假设搜索“篮球”结果，最理想的结果是：B1、B2、 B3。而出现的结果是 B3、B1、B2 的话，CG 的值是没有变化的，因此需要下面的 DCG。

折损累计增益（DCG）

DCG， Discounted 的 CG，就是在每一个 CG 的结果上处以一个折损值，为什么要这么做呢？目的就是为了让排名越靠前的结果越能影响最后的结果。假设排序越往后，价值越低。到第 i 个位置的时候，它的价值是 1/log2(i+1)，那么第 i 个结果产生的效益就是 reli * 1/log2(i+1)，所以：

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

当然还有一种比较常用的公式，用来增加相关度影响比重的 DCG 计算方式是：

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

百科中写到后一种更多用于工业。当然相关性值为二进制时，即 reli 在{0,1}，二者结果是一样的。当然 CG 相关性不止是两个，可以是实数的形式。

归一化折损累计增益（NDCG）

NDCG， Normalized 的 DCG，由于搜索结果随着检索词的不同，返回的数量是不一致的，而 DCG 是一个累加的值，没法针对两个不同的搜索结果进行比较，因此需要归一化处理，这里是处以 IDCG。

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

IDCG 为理想情况下最大的 DCG 值。

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

其中 |REL| 表示，结果按照相关性从大到小的顺序排序，取前 p 个结果组成的集合。也就是按照最优的方式对结果进行排序。

2.3 问题三分析过程与方案

2.3.1 评价体系的构建

对相关部门给出的答复意见的质量进行综合评价，并制定一套能全面反映答复的相关性、完整性、可解释性的指标体系，对高效率的解决问题有很大帮助，能够进一步改善和提高居民的生活质量。而通过因子分析方法，对相关部门给出的答复意见进行降维处理，需要检测答复意见与留言的相关性、完整性、可解释性。其中主要的是设置因子分析模型（KMO 和 Bartlett 的球形度检验）：

KMO 检验统计量是用于比较变量间简单相关系数和偏相关系数的指标。主要应用于多元统计的因子分析。KMO 统计量是取值在 0 和 1 之间。Kaiser 给出了常用的 kmo 度量标准： 0.9 以上表示非常适合；0.8 表示适合；0.7 表示一般；0.6 表示不太适合；0.5 以下表示极不适合。KMO 统计量是取值在 0 和 1 之间。当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时，KMO 值接近 1.KMO 值越接近于 1,意味着变量间的相关性越强，原有变量越适合作因子分析；当所有变量间的简单相关系数平方和接近 0 时,KMO 值接近 0.KMO 值越接近于 0,意味着变量间的相关性越弱，原有变量越不适合作因子分析。

Bartlett’s 球形检验用于检验相关阵中各变量间的相关性，是否为单位阵，即检验各个变量是否各自独立。如果变量间彼此独立，则无法从中提取公因子，也就无法应用因子分析法。Bartlett 球形检验判断如果相关阵是单位阵，则各变

量独立，因子分析法无效。由 SPSS 检验结果显示 Sig.<0.05（即 p 值<0.05）时，说明各变量间具有相关性，因子分析有效。

2.3.2 给相关部门的意见

有关部门也可成立答复情况质量评价委员会，由委员会对答复情况进行收集汇总，定期广泛征求社会各界对答复情况的意见和建议，对群众不满意的答复事项由委员会论证后，提出新的处理意见。

3. 结论

对群众的意见进行一个系统性的分类，有助于相关部门提高办事效率，对城市的发展有很大的帮助。在数字信息化时代，传统的人工已经不能够对留言起到很好的处理了，本文采用的是 KNN 分类模型（K—近邻模型）建立一个分类模型就能够很快的将信息分类规划到相关职能部门，去解决这些问题是个非常好的解决问题的方法。

热点问题的提取和整理能够让人们察觉到当下城市发展面临着哪些方面的问题，我觉得是更加有利于管理者更好的去把握城市发展的方向和进程。从分析结果来看，城市建设、交通运输和卫生卫计是当下居民比较关注的。

4. 参考文献

[1]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
[2] 张明旺.基于内容的垃圾短信分类技术研究.四川警察学院.2015
[3]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
[4]赵琳瑛.基于隐马尔科夫模型的中文命名实体识别研究.西安电子科技大学.2007
[5]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
[6]施聪莺，徐朝军，杨晓江.TF-IDF 算法研究综述.南京师范大学.2009