摘要

近年来，随着互联网的急速发展，基本上每个人都用上了智能手机，也就是说基本所有人都可以用 4G 网进行沟通与联系。因此，为了加强政府与百姓的沟通，许多软件例如微信、微博、市长信箱、阳光热线都开启了群众留言的功能，使群众可以向政府或相关部门及时发表提出自己希望得到解决的问题或自己对周边的一些意见。但是，也因为近年网络的发展与普及，各类民众留言数量不断攀升.那对用户留言内容进行分类整理，从而更好地解决用户的问题成了一个极大的难题。中文分词是整个挖掘过程的第一个环节，因此会直接影响后续所有环节的效果。我们在这里采用的 python 的中文分词包 jieba 分词。当前，我们对文本分类办法效率进行评价的话，常用的方法是 F-Score，$F_1= \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$。这个评价方法中的评价标准选择了查准率和查全率两个，这两个条件是反映检索分类结果的重要指标。

对于问题一以及问题二，我们都需要用到上文提到的 jieba 分词进行分词处理。

关键词：中文分词 文本分类 KNN

Abstract

In recent years, with the rapid growth of the Internet, almost everyone of China has used smart phones, that is to say, almost everyone can use 4G network for communication and contact. Therefore, in order to strengthen the communication between the government and the people who can not communicate with government directly, many software such as wechat, Weibo, mayor's mailbox and sunshine hotline have opened the function of the masses' message, which is that the masses can express their opinions to the government in time. But also because of the development and popularization of the network in recent years, the number of all kinds of public messages is increasing. It has become a great problem to sort out the user message content, so as to better solve the user's problem. Chinese word segmentation is the first link of the whole mining process, so it will directly affect the effect of all subsequent links. The Chinese word segmentation package of python that we use here is Jieba. At present, we often use F-score to evaluate the classification methods.，$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$. The evaluation methods only choose the precision rate and recall rate, which are important indicators to reflect the retrieval effect.

Key words: Chinese words segmentation； text classification； KNN

目录

1、 挖掘目标

本次建模目标是针对用户在各个平台上的留言，利用 jieba 中文分词工具对大家留言的具体内容进行分词，以达到以下两个目标：

1）在这里我们将运用中文文本分词的方法对那些散乱的、非结构化的数据进行文本挖掘，之后再根据题目内划分的三级体系对用户的留言进行分类整理，以便于后面更方便地将群众留言分派到相应的部门。

2）利用 jieba 中文分词工具将用户留言分词，对分词后的关键词进行计数，根据权重比例高的关键词将留言分类，并定义合理的热度评价指标，得出排名前五的问题。

2、 分析方法与过程

总体流程图

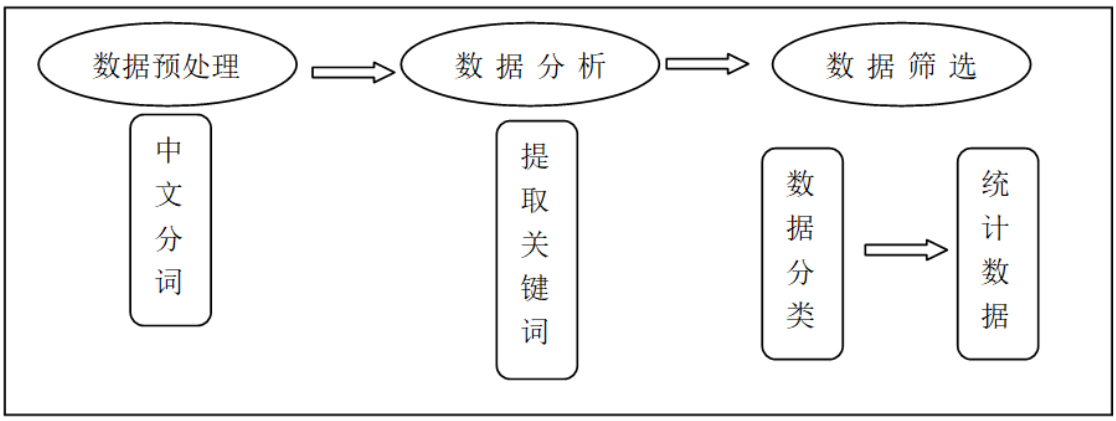


图 1 总体流程图

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，文字过多，过于冗杂，我们在原始的数据上进行 jieba 中文分词。

步骤二：进行中文分词后，对停用词进行去除，以免最后面的 TF-IDF 算法造成误差。

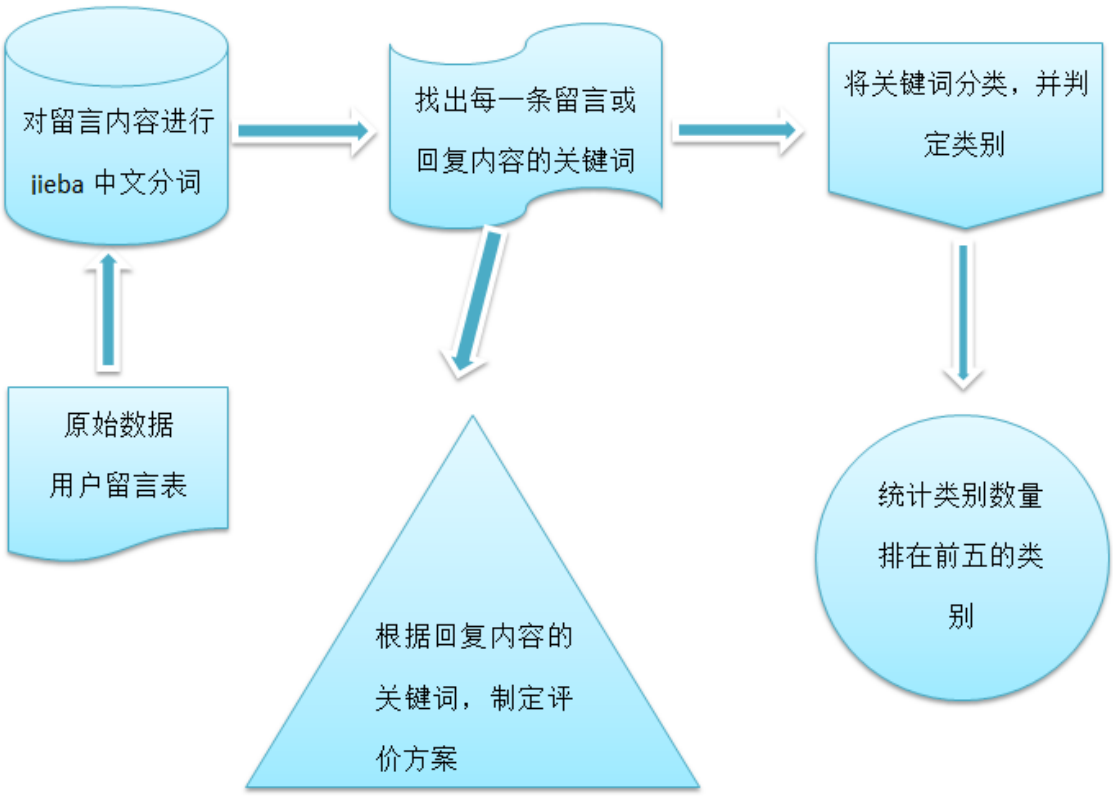
步骤三：数据分析，在对居民的留言内容及工作人员给予的回复内容进行分词并去掉停用词后，我们需要将这些保留下来的词语都转化为向量，以便我们在之后的题目中对这些数据进行挖掘、分析、使用。在这一个步骤之后，我们采用的是 TF-IDF 这一算法，以此来找到每一个留言内容及回复内容的关键词，然后再对它们进行分类。

步骤四：利用 K-means 算法对职业进行分类，利用 Knn 算法找出每一条留言及回复内容中相似的关键词，最后判定它的类别。统计相关的数据，从中选出留言数量前五的类别，将其汇成热点问题表。

步骤五：根据工作人员回复内容的关键词，从与居民提问的相关度、完整度、可解释性来制定一套评价方案。

2.1 问题 1 分析方法与过程

2.1.1 流程图



问题 1 流程图

2.1.2 数据预处理

2.1.2.1 留言内容的中文分词

在题目给出的留言数据中，用户的留言总数量高达近万条，在我们对用户留言的内容进行数据挖掘分析之前，我们先要把那些散乱的、非结构化的文本信息内容转换为计算机能够识别的，并且能运用到编程手段里的结构化信息。首先，我们要对这些留言内容信息以及工作人员的回复内容信息进行中文分词。

目前常用的分词算法有很多，例如 hanlp 分词、jieba 分词、清华大学 THULAC 等，这里采用了 python 的中文分词包 jieba 进行分词。Jieba 可以对中文文本进行分词、词性标注、关键词取出等十分实用且便捷的功能。Jieba 分词采用了以前缀词典为基础以便于实现的高校此图扫描，然后就可以生成留言及回复句子中的汉字所有大概率可以组成词语情况所构成的有向无环图，同时该分词法又采

用了动态规划的方法，让这个方法可以查找到最大的、最靠谱的概率路径，同时再找出那些以词频为基础的最大切分组合，而对于那些未登录词，该分词法又采用了以汉字成词能力为基本的 HMM 模型，这个 HMM 模型使得 jieba 能更好地、更快捷地实现中文分词的效果。

分词结束后，将采用 TF-IDF 算法，抽取每个留言的关键词。

2.1.2.2 TF-IDF 算法

在我们对用户的留言内容以及工作人员对此回复的内容进行了分词处理之后，这些词需要被我们转化为向量，这些被转化成向量之后的数据可以更方便地被我们在之后的挖掘分析中来使用。这里，我们采用的是 TF-IDF 算法，TF-IDF 是一种统计学方法，能够估评一个词在整个文件或整个文档库中的重要性。用户留言信息及工作人员回复信息如何被转换为权重向量。具体原理及步骤如下：

第一步，词频计算，公式为 $TF=\frac{某个词在该文档中出现的次数}{该文档的总词数}$。数据代入到该公式之后，可以算出每一个中文分词分出来的词在当前文本中的一个关于出现频次数据，称为量化系数。

第二步，计算逆文档频率，我们面对冗长的留言内容信息时，一个最直接的思路就是找到出次次数最多的词，如果这个词很重要，那它应该会在留言中反复出现，这里我们会用“词频”（TF）统计。以此让我们更加快速地知道这个留言的核心内容是什么，但有一点很麻烦的是，往往在大家的留言中出现次数最多的是“的”，“是”这一类遣词造句必不可少的字眼，我们将它们叫做“停用词”。它们出现的次数很多，但它们是我们想要的关键字吗？

显然不是这样的，所以这里就是我们这一步骤的重点，我们需要一个重要性调整系数，衡量一个词是不是常见词。如果某个词比较少见，但它却在留言中多次出现，那么它很可能就是这个留言中的关键词。

用统计学的语言来描述，就是在计算出来的词频的基础之上，我们再对每个文本分词分配一个“重要性”权重。例如“的”、“是”、“在”“吗”这一类十分常见的字词，我们将对其分配最小的权重系数。而一些较少见的词，例如“环境”、

“生育”、“卫生”等等，将被我们给予较大的权重。这个权重被称作做“逆文档频率”（IDF），逆文档频率的大小通常与一个字词的普遍程度成反比。

第三步，计算 TF-IDF，TF-IDF=词频（TF）*逆文档频率（IDF）是计算 TF-IDF 的公式，根据我们实际分析得出结论，TF-IDF 值与一个词在留言内容信息文本以及回复内容信息文本中出现的频率成正比，某个字词的出现频率越高，TF-IDF 值就会成比例相应增大。

2.1.2.3 生成 TF-IDF

具体步骤：

（1） 使用 TF-IDF 算法，找出每一条留言内容的前五个关键词；

（2） 对于每条留言内容法提取出的五个关键词，将它们合并成一个集合，计算集合中每一个关键词的词频；

（3） 生成各条留言的权重向量，计算公式如下：

TP-IDF=词频*逆文档频率（TF*IDF）

2.1.3 分类留言

留言内容及回复内容描述的 TF-IDF 权重向量经过 TP-IDF 计算被生成后，我们将根据每条留言及回复生成的 TF-IDF 权重向量，将对每一条留言根据题目所提供的三级标签体系分类。

我们常见的分类算法的思路有以下四种

（1） 朴素贝叶斯分类器

（2） 支持向量机分类器

（3） KNN 方法

（4） 决策树方法

在这里，我们采取的分类方法是 CNN 方法，它在一些方面与其它的分类方法比起来有很大的优势，以 CNN 为基本来做文本分类，可以利用到此的顺序包含的信息。CNN 模型的一个实现，共分为四层：

第一层，词向量组，每个关键词将被分别映射到词向量空间，假设词向量为 a

维，则 s 个词映射后，一张 a*s 维的图像将被自动生成；

第二层，卷积层，词向量层同时被多个滤波器作用，不同滤波器将会生成各种的 feature map;

第三层，pooling 层，每个 feature map 的最大值都将被自动收取，这一层的关键作用主要是依赖于滤波器的个数；

第四层，全连接的 softmax 层，每个类目的概率将被依次输出，我们一般会在中间加个 dropout，防止概率数据过拟合。

2.2 问题二分析方法及过程

2.2.1 数据预处理

2.2.1.1.数据清洗，去停用词

在题目给出的数据中，众多留言中出现了很多特殊字符，例如有群众的留言是“十分的感谢您！致电~~~国投滨江印象老城***，哈哈哈 0.0”等等，考虑到统计数据的方便，数据清洗必不可少。其次，应该过滤无意义的词。

2.2.1.2 分词和调性

在对留言信息进行数据挖掘之前，先要把非结构化的文本信息转化为计算机能够识别的结构化信息，在附件留言表中，以中文文本的方式给出了数据。为了便于转换，先要对这些留言信息进行中文分词。这里采用 python 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。

在分词的同时，采用了 TF-IDF 算法，抽取每个留言信息描述中的前 5 个关键词，这里采用 jieba 自带的语义库。

2.2.1.3 TF-IDF 算法

在对留言描述信息分词后，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，把留言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）

考虑到文章有长短之分，为了便于不同文章的比较，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：

词频（TE）=某个词在文本中的出现次数/文本的总词数　(2)
或
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）

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency)，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)

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-IDE 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职位描述表中文本的关键词。

2.2.1.4 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 个关键词，合并成一个集合，计算每个留言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)

2.2.2 问题归类

生成热点问题留言描述的 TF-IDF 权重向量后，根据每条留言的 TF-IDF 权重向量，对热点问题进行分类。这里采用 K-means 算法把热点问题分成几类。K-means 聚类的原理如下：

原始的 k-means 算法首先随机选取 k 个点作为初始聚类中心，然后计算各个数据对 象到各聚类中心的距离，把数据对象归到离它最近的那个聚类中心所在的类； 调整后的新类计算新的聚类中心，如果相邻两次的聚类中心没有任何变化，说明 数据对象调整结束，聚类准则函数 f 已经收敛。在每次迭代中都要考察每个样本的分类是否正确，若不正确，就要调整。在全部数据调整完后，再修改聚类中心，进入下一次迭代。如果在一次迭代算法中，所有的数据 对象被正确分类，则不会有调整，聚类中心也不会有任何变化，这标志着 f 已 经收敛，算法结束。其实这跟普通的前馈神经网络使用逆向传播算法训练模型的原理类似，分析误差，修改模型直至达到要求的误差范围。

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。

2.2.3 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 C_j（w_1，w_2，…，w_n），将待分类文本了表示成 n 维向量的形式，T(w_1,w_2,…,w_n),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(T_1，W_1；T_2，W_2；…T_n，W_n)。对于一个测试文本，计算它与训练样本集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：
$Sim(di,dj)= \sum W_{ik}*W_{jk}/\sqrt{\sum W_{ik}^2}*\sqrt{\sum W_{jk}^2}$
式中，di 为测试文本的特征向量，dj 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。
（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：
$P(X, Cj) = \sum Sim(x, di)y(di, Cj) -b \geq 0$
P(X，Cj) =0,其他，
式中，x 为测试文本的特征向量；Sim（x，di）为相似度计算公式：b 为阙值，有待于优化选择：而 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。

2.2.4 分析定义热度评价

对热点问题留言评价根据 K-means 聚类方法和 Knn 最邻近分类根据把相似留言归于同一类问题，统计数量最多的即为目前热度评价指标最高的热点问题，并定义相关热度评价指标。

2.3 问题三分析方法及过程

2.3.1 评价目的

（1）每一个公司为了公平、有效、客观、公正地评价工作人员的工作业绩、工作能力和工作态度，并且希望能够及时纠正偏差，改进工作人员的工作方法，同时激励工作人员之间争先创优。以此来进行整体团队的优化提高，从而提高工作质量，特制定本评价方案。

（2）对工作人员进行的评价结果将作为进行人员调整的决策依据。

（3）由评价方案生成的评价结果将会被自动融入公司管理过程，在评价中成为员工和公司的双向沟通的渠道，增进公司内部的管理效率，推动公司更加良好地运作。

2.3.2 评价指标

（1） 回复率：被工作人员做出回复的客户数量与工作人员接待的总客户数量之间的一个比例，如果一个工作人员对所有接待的客户都能予以回复，则该工作人员的回复率为100%。

指标	标准	分值	权重
回复率	≥99.0	100	10%
	≥98.0	80	
	≥95.0	60	
	≥90.0	40	

（2）首次响应时间：用户在等待工作人员回复的等待时间中，用户在咨询工作人员的第一句被作出回应的时间差；该数值越小越好。

指标	标准	分值	权重

首次响应时间	≤130.0	40	10%
	≤100.0	60	
	≤80.0	80	
	≤60.0	100	

（2） 平均响应时间：在用户等待工作人员回复的等待时间中，用户咨询到工作人员作出回应的每一次的时间差的均值被称作为平均响应时间，该数值越小越好，主要看工作人员面对客户的最大服务压力，还有接待人数为多少的时候会影响响应时间。

指标	标准	分值	权重
平均响应时间	≤130.0	40	10%
	≤100.0	60	
	≤80.0	80	
	≤60.0	100	

（3） 客户满意度：在完成每一次咨询过后，将会让客户评价此次工作人员的整体表现，细分为相关性、完整性、可解释性。

指标	分类	标准			权重
客户满意度	相关性	非常满意（100）	一般（75）	不满意（50）	20%
	完整性	非常满意（100）	一般（75）	不满意（50）	20%
	可解释性	非常满意	一般	不满意	30%

	（100）	（75）	（50）

3.结果分析

3.1 问题 1 结果分析

3.1.1 分类结果

在应用了 python 里的 jieba 分词之后。并且去掉我们所谓的“停用词”，我们可以得到关于一级体系分类的分类关键词。

根据上文中的步骤，所得到的一级分类体系的前五关键词为

关键词	出现次数
卫生	1689
学校	1379
居民	1286
小区	1193
安全	1100

3.2 问题 2 结果分析

3.2.1 聚类中心分类结果

通过去停用词后对文本进行分词，提取关键词后由 K-means 分类得到聚类中心，利用 Knn 算法找出离各个聚类中心最近的元素，根据“少数服从多数”判定聚类中心所属类别。KNN 算法的大致步骤如下：

1、算距离：给定聚类中心，计算它与样本中的每个 TF-IDF 权重向量的距离

2、找邻居：圈定距离最近的样本，作为聚类中心的近邻

3、做分类：根据这近邻归属的主要类别，来对聚类中心进行分类

结合抽样样本，分别找出聚类中心的近邻样本点所属的热点问题类型

结果如下表所示为：

3.2.2 热点问题分类

某一时段内群众集中反映的某一问题称为热点问题，讲某一时段内反映特定地点或特定人群的留言进行归类，定义合理的热度评价指标，评价结果如下表：

表-1 热点问题表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1		2019/09/05 至 2020/01/02	梅溪湖	物业问题
2	2		2019/08/16 至 2019/09/05	丽发新城	搅拌站扰民
3	3		2019/01/22 至 2020/01/03	滨河苑	强制购车位
4	4		2019/08/10 至 2020/01/01	松雅	传销及交通问题
5	5		2019/06/12 至 2020/01/02	保利	扰民及物业问题

表-2 热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	
1	191596	A00051936	A 市中海梅溪湖壹号业主私搭乱建	2019/9/22 22:50	中海梅溪湖壹号业主私搭 章搭建房屋，面积约 20 在不久之前，我家就已来 示这个违章建筑确实存在 这样的违法行为，不但得
1	192337	A00042313	A3 区梅溪湖看云路一师润芳园小区临街门面油烟扰民	2019/9/5 12:23:38	《张傅师烧烤》《津市米

1	194652	A00046589	A 市梅溪湖城市热点垄断市场，装机不及时	2019/9/30 18:25:05	我是进驻梅

3.3 问题 3 结果分析

我们从工作人员回复用户的第一次时间差、平均时间差以及回答内容的相关性、完整性、可解释性来分配不同的权重，以此来衡量一个工作人员在回复用户时的工作质量以及工作效率。

会在每次用户的问题被解决之后，弹出一个服务小调查的弹窗，里面主要是希望顾客从工作人员回复内容的相关性、完整性、可解释性来进行评分。

4. 结论

对于大量文本要按关键词分类而言，若人工一条一条地去分类无疑是一个工程浩大的工作，但若是运用程序编码对大量的数据进行操作，那便不用耗费太多的人力、时间、金钱。

我们根据 Python 的 jieba 分词对留言内容信息及回复内容信息进行了数据的预处理，以及挑选出了具有代表性的关键词，然后再用 CNN 进行分类处理。

5. 参考文献

【1】秦赞.中文分词算法的研究.吉林大学.2016

【2】王强.基于 CNN 的字符识别方法研究.天津师范大学.2014

【3】曾小芹.基于 Python 的中文结巴分词技术实现.信息与电脑.2019

【4】牛萍.TF-IDF 与规则结合的中文关键词自动抽取研究.大连理工大学.2015

【5】王千，王成，冯振元，叶金凤.K-means 聚类算法研究综述.2012

【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