“智慧政务”中的文本挖掘应用

摘 要

随着自然语言处理技术的发展，在生活中的应用也越来越广泛。本文主要构建了一个基于朴素贝叶斯和 k-means 的智慧政务模型。

在数据预处理阶段，我们从原表格中进行分词去除停用词等操作。在第一题的朴素贝叶斯模型中，我们利用前 600 个分词，结果形成词语列表，获取特征向量。利用朴素贝叶斯的多项式分类器来检测数据，并对他们进行逐一输出。最后计算 F-score 的值来进行评价。在给出的测试集中，F1-score 为 0.91101 ，验证了本文模型的有效性。

对于热点问题的挖掘我们利用 TF-IDF 构建语料库并生成词频向量。利用 k-means 这种无监督的聚类算法得出热点问题。

对于答复意见的评价，我们主要从两个方面来分析。我们先利用生成的 TF-IDF 词频矩阵，计算每一个留言详情和答复意见的余弦相似度。然后利用答复意见在留言详情中的分词占比，来计算答复意见的完整度。最终给出一个整体的评价。

关键词： 词向量，朴素贝叶斯，TF-IDF

目　录

简介

1.1 挖掘目标

本次建模目标是收集自互联网公开来源的群众问政留言记录数据，我们要构建一个智能的文本挖掘模型，模型可以起到整理互联网公开来源的群众问政留言记录数据的作用，帮助相关部门减轻工作负担，显著提高解决问题的效率。

具体到使用情景上，对于用户输入的问题与文档，模型可以定位到文档中能帮我们找到关键词，直接给出明确的分类作为答案输出。

1.2 挖掘流程

首先构建分类器的第一步涉及原始数据的处理和分析，文本数据的准备，一方面对留言主题和留言详情中的的字符串进行分词，进行去除数字和标点符号；另一方面还要利用停用词表对无意义词语和单个字组成的词进行去除，将数据进行预处理。再采用数据挖掘技术，朴素贝叶斯模型按留言主题对留言进行分类，使用 F-Score 对分类方法进行评价。利用 K-means 算法聚类生成热点问题表和热点问题留言明细表。解决了之后评价答复意见，方案设计，并进行相似度分析完整度分析最后整合数据形成最终评价表格。

1.3 挖掘意义

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，关于写建议信，相信大家都不陌生，在写信方面，我们接受了传统语文教育中的书信写法；在生活中，我们也常常会遇到阅读投诉信，建议信，感谢信等应用场景。除去欣赏优美的语言，更多情况下，我们只是需要从文本中查找某一些片段来 明白书信的主要内容。比如，查看一封投诉信主要投诉什么，这时并不需要看其中的强烈个人色彩的语言表达，我们只是想知道重点是投诉什么，关于哪一方面。但是，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，其中内容往往很难在短时间内为人了解，就算是略读也不容易，更不必说从浩繁卷帙中准确的定位答案。因此，我们希望自然语言处理能够在这方面提供一些帮助。为了构建智能文本挖掘模型，学界对于机器阅读理解的研究从未止步。机器阅读理解作为目前热门的自然语言处理任务，目标是使机器在能够理解原文的基础上，正确回答与原文相关的问题。提高机器对语言的理解能力。机器阅读理解技术的发展对信息检索、问答系统、机器翻译等自然语言处理研究任务有积极作用，同时也能够直接改善搜索引擎、智能助手等产品的用户体验，因此，以文本理解、文本挖掘为契机研究机器理解语言的技术，具有重要的研究与应用价值。

预处理

2.1 去除无意义字符

导入 re 模块，对留言主题，留言详情中的每一句进行中文分词。re 分词用到的算法为最短，利用 re.sub（）函数，用来实现通过正则表达式，实现比普通字符串的 replace 更加强大的替换功能，利用正则表达式的函数来去除导入文本中出现的“\t\n、空格、数字、字母”等字符。

2.2 分词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，本文采用 Python 开发的一个中文分词模块———jieba，对问题和回答中的每一句话进行分词进行中文分词。jieba 分词用到的算法为最短，路径匹配算法该算法首先利用词典找到字符串中所有可能的词条，然后构造一个有向无环图。其中，每个词条对应图中的一条有向边，并可利用统计的方法赋予对应的边长一个权值，然后找到从起点到终点的最短路径，该路径上所包含的词条就是该句子的切分结果。

2.3 去停用词

在文本处理中，停用词是指那些功能极其普遍，与其他词相比没有什么实际含义的词，它们通常是一些单字，单字母以及高频的单词，比如中文中的“我、

的、了、地、吗”等，英文中的“the、this、an、a、of” 等。对于停用词一般在预处理阶段就将其删除，避免对文本，特别是短文本，造成负面影响。

建立训练数据和测试数据集

3.1 用有监督的机器学习方法对留言主题进行分类

由于对数据已经进行好了预处理，准备用朴素贝叶斯模型按留言主题对留言进行分类。

3.1.1 朴素贝叶斯模型简介

朴素贝叶斯算法（Naive Bayesian algorithm）是应用最为广泛的分类算法之一。

朴素贝叶斯方法是在贝叶斯算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立。也就是说没有哪个属性变量对于决策结果来说占有着较大的比重，也没有哪个属性变量对于决策结果占有着较小的比重。虽然这个简化方式在一定程度上降低了贝叶斯分类算法的分类效果，但是在实际的应用场景中，极大地简化了贝叶斯方法的复杂性。

3.1.2 朴素贝叶斯模型算法原理

朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y。

设有样本数据集 $D=\{d_1,d_2\cdots d_n\}$ 对应样本数据的特征属性集为的类变量为 $Y=\{y_1,y_2\cdots,y_m\}$，即 D 可以分为 y_m 类别。其中 $x_1,x_2,\cdots,x_d$ 相互独立且随机，则

Y 的先验概率 $P_{\text{prior}} = p(Y)$，Y 的后验概率 $P_{\text{post}} = P（Y \mid X）$，由朴素贝叶斯算法可得，后验概率可以由先验概率 $P_{\text{prior}} = p(Y)$ 、证据 $P(X)$、类条件概 $P（X \mid Y)$ 计算出：$P(Y \mid X) = \dfrac{P(Y)P(X \mid Y)}{P(X)}$

朴素贝叶斯基于各特征之间相互独立，在给定类别为 **y** 的情况下，上式可以进一步表示为下式：$P(X \mid Y = \mathrm{y}) = \prod_{i=1}^{\mathrm{d}} \mathrm{P}(x_{\mathrm{i}} \mid Y = y)$

由以上两式可以计算出后验概率为：$P_{post} = P(Y \mid X) = \dfrac{P(Y)\prod_{i=1}^{d} P(x_i \mid Y)}{P(X)}$

由于 $P(X)$ 的大小是固定不变的，因此在比较后验概率时，只比较上式的分子部分即可。因此可以得到一个样本数据属于类别 y_i 的朴素贝叶斯计算如下图所示

$$P(y_i \mid x_1, x_2, \cdots, x_d) = \frac{P(y_i)\prod_{j=1}^{d} P(x_j \mid y_i)}{\prod_{j=1}^{d} P(x_j)}$$

3.1.3 朴素贝叶斯模型的具体应用

1.首先明确一级分类标签。

2. 将“留言主题”作为训练模型输入，对应的“一级分类”作为模型输出。

3. 查看每条留言的信息，用 allword 存放所有留言的分词信息。

4. 利用 counter（）和 chain（）函数获取每个分词语的频次。

5. 获取 topword600 的词语列表，利用词语列表，统计每个词语在 allword 列表中是否出现作为本条留言的特征向量，由此生成所有留言的特征向量 v 这个列表。

6. 模型输入的特征向量，统计得到每个词在 top600 的词语列表中出现的个数生成列表 v。

7. 自定义函数，对模型输出进行格式转换编码，将交通运输转换编码为 1，卫生计生编码为 2，劳动合同社会保障编码为 3，商贸旅游编码为 4，城乡建设编码为 5，教育文体编码为 6，环境保护编码为 7，方便接下来的训练、查找。然后查看每一条留言的转换形式。

8. 我们采用 sklearn 包中的朴素贝叶斯算法实现，选择适用于文本分类的多项式模型，从 sklearn.naive_bayes 中导入 MultinomialNB 模型，并创建该多项式分类器。 MultinomialNB 假设特征的先验概率为多项式分布，即如下式：

$$P(X_j = X_{jl} \mid Y = C_k) = \frac{X_{jl} + \lambda}{m_k + n\lambda}$$

9.首先，用 MultinomialNB 模型检测数据，并根据已知结果与模型输出相比较，看是否一致，用某一数据来检测模型。其次自己写个文本，即投诉主题，提取文本的特征向量，用模型进行分类检测。另外对原始数据，进行逐一输出分类，检验模型的有效性与准确率，将'原来分类'与'模型分类输出'归为二列，整合成 dataframe，增加一列 dat，用来输出原来分类与模型分类输出相等与否。

10. 使用 F-Score 对分类方法进行评价：自定义函数用来查准率，p 为查准率，自定义函数用来查全率，R 为查全率，用模型检测数据，查看已知结果与模型输出是否一致检测。

结果如下：

```
p
```

```
[0.9607843137254902,
 0.9482758620689655,
 0.8547008547008547,
 1.0,
 0.8942307692307693,
 0.912621359223301,
 1.0]
```

```
R
```

```
[0.8909090909090909,
 0.9482758620689655,
 0.9615384615384616,
 0.7083333333333334,
 0.9207920792079208,
 0.9791666666666666,
 0.8484848484848485]
```

```
f1
```

```
0.9110176157643893
```

热点问题挖掘

4.1 留言数据预处理

将留言主题留言详情用 re 正则表达式进行去\n、\t 格式控制符，去字母，数字和空格，利用 jieba 库进行分词。由于留言详情中存在'市长'，'局长'，'领导'，'干部'，'书记'，'您好'，'你好'，'尊敬'等无用词，将这些词加入到 stop_words 中，进行去除停用词并把无意义的单个字符删除的操作。

4.2 K-means 算法聚类

4.2.1 构建词矩阵

在进行文本聚类之前，我们需要将词进行向量化，这里向量化的方式选用计算 TF-IDF 矩阵（词频-逆文件频率）。TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。

词频：指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化(一般是词频除以文章总词数)， 以防止它偏向长的文件。

逆向文件频率：IDF（InversDocument Frequency）表示计算倒文本频率。如果包含词条 t 的文档越少， IDF 越大，则说明词条具有很好的类别区分能力。文本频率是指某个关键词在整个语料所有文章中出现的次数。倒文档频率又称为逆文档频率，它是文档频率的倒数，主要用于降低所有文档中一些常见却对文档

影响不大的词语的作用。

构建语料库（语料库就是把平常我们说话的时候的句子、一些文学作品的语句段落、报刊杂志道上出现过的语句段落等等在现实生活中真实出现过的语言材料整理在一起，形成一个语料库，以便做科学研究的时候能够从中取材或者得到数据佐证。）

TfidfVectorizer 模型建立后，可通过 fit_transform()函数进行训练，将文本中的词语转换为词的 TF－IDF 矩阵，并计算文档的 TF－IDF 矩阵，TF－IDF 以稀疏矩阵的形式存储，将 TF－IDF 转化为数组的形式，输出为矩阵形式，通过 toarray()函数可看到 TF－IDF 矩阵的结果。

4.2.2K-means 聚类

聚类算法是根据相似性原则，将具有较高相似度的数据对象划分至同一类簇，将具有较高相异度的数据对象划分至不同类簇。 K-Means 算法是无监督的聚类算法。目的是使得每个点都属于离它最近的均值（此即聚类中心）对应的簇 Ai 中。这里使用 sklearn 库中的 K-means 聚类算法对数据进行聚类分析，得到每一条留言所属的簇。

对 word_vectors 进行 k 均值聚类，将他要分成 500 类，聚类得到留言详情的类别，将类别赋给 cluster，以 cluster 的方式展现出来。

4.3 对问题进行分析

4.3.1 制作热点问题表

首先对之前的 k 均值聚类结果 cluster500 类进行统计数量，并将数量按照降序排序，输出前七个数量最多的 cluster。

Out[99]:

cluster	留言详情
403	113
116	71
169	39
63	36
206	29
47	29
279	29

由表中数据可以看出，第 403 和 116 聚类的数量远大于其他聚类，可判断该聚类为无关聚类(即各种杂七杂八的无分类的留言)。

运行结果除去第一，二个聚类，

剩下 5 类即为题目所要求的 5 个热点聚类。

最终结果显示最高频率的聚类为 169，63，206，47，279。

读取附件给 data_new，并将 kmean_labels['cluster'] 加入分类标识 data_new['聚类']中，由于最高频率的聚类是 169，筛选出 data new 中'聚类'等于 169 的数据赋给 data_1st，读取 data_1st 留言时间'给时间排序，便于找出时间范围，可以了解到热度第一的问题的时间范围为 2019-01-06 至 2020-01-06，接下来制作热点问题表，总结出热度第一的问题描述为物业问题解决不够妥当，存在乱收停车费的现象，填充在第一行的位置，填充具体的时间范围为 2019-01-06 至 2020-01-06，选取第一列并去除不需要的列数如'留言编号'，'留言用户'，'留言主题'，'点赞数'，'反对数'，'聚类'，插入'地点人群"问题 ID"热度指数"热度排名'，总结出地点人群为 A 市各区和 L 市以及热度指数为 5，问题 ID 为 1，热度排名为 1，并将留言时间替换成时间范围，留言详情替换成问题描述，得到 top1

的数据。

筛选出 data_new 中'聚类'等于 63 的数据赋给 data_2nd, data_2nd 如 data1_st 方法一样，可以了解到热度第二的问题的时间范围为 2019-01-07 至 2020-01-06，总结出热度第二的问题描述为公司工资拖欠，存在诈骗行为，填充在第一行的位置，填充的具体时间范围为 2019-01-07 至 2020-01-06，同样去除不需要的列数，插入列，给出地点人群是 A 市各区居民，热度指数是 4，问题 ID 是 2，热度排名是 2，填充三列热度问题表所需要的信息，同样替换列名称，得到 top2 的数据。

筛选出 data_new 中'聚类'等于 206 的数据赋给 data_3rd'，data-3rd 如 data_1st 方法一样，可以了解到热度第三的问题的时间范围为 2017-06-08 至 2019-12-24，总结出热度第三的问题描述为 A 市强制学习情况多，填充在第一行的位置，填充的具体时间范围为 2017-06-08 至 2019-12-24，同样去除不需要的列数，插入列，给出地点人群是 A 市学生，热度指数是 3，问题 ID 是 3，热度排名是 3，填充三列热度问题表所需要的信息，同样替换列名称，得到 top3 的数据。

筛选出 data_new 中'聚类'等于 47 的数据赋给 data_4th, data_4th 如 data_1st 方法一样，可以了解到热度第四的问题的时间范围为 2019-01-05 至 2019-12-24，总结出热度第三的问题描述为医院难以就诊，途中经常堵车，填充在第一行的位置，填充的具体时间范围为 2019-01-05 至 2019-12-24，同样

去除不需要的列数，插入列，给出地点人群是 A 市和 M 市居民，热度指数是 2，问题 ID 是 4，热度排名是 4，填充三列热度问题表所需要的信息，同样替换列名称，得到 top4 的数据。

筛选出 data_new 中'聚类'等于 279 的数据赋给 data_5th，data_5th 如 data_1st 方法一样，可以了解到热度第五的问题的时间范围为 2019-10-14 至 2019-10-20，总结出热度第五的问题描述为 A7 县住房条件差，土地污染问题严重，民生问题解决不妥，填充在第一行的位置，填充的具体时间范围为 2019-10-14 至 2019-10-20，同样去除不需要的列数，插入列，给出地点人群是 A 市各区居民，热度指数是 1，问题 ID 是 5，热度排名是 5，填充三列热度问题表所需要的信息。

得到了 top5 的列表之后，用 append()将列表拼接得到 top_final，将 top_final 导出成 Excel 表格，并删除前面的索引。

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	5	2019-01-(	A市各区和	物业问题解决不够妥当，存在乱收停车费的现象

热点问题表（部分，具体见 Excel 文件）

通过 top_final 导出的表格，我们可以看到热度排名在第一的是物业问题，第二的是公司工资问题，第三的是 A 市强制学习问题，第四的是 A 市和 M 市医院就诊和马路拥堵问题，第五的是 A7 县民生问题。其次通过对时间的分析，其中历时最长的是排在第三的 A 市学生反映的 A 市强制学习问题持续了两年六个月十六天，其他问题几乎持续了一年。通过对地址人群的分析，我们发现前五个

热点问题都在 A 市出现过，大部分都属于民生问题。

4.3.2 制作热点问题留言明细表

读取附件 excel 赋给 data ，读取附件给 data, 并将 kmean_labels['cluster'] 加入分类标识 data['聚类']中，并将留言详情进行去除'\t\n、空格、数字、字母'等字符的操作。

由于最高频率的聚类为 169，63，206，47，279，当 data_['聚类'] 等于 169，加入一列'问题 ID'并填入 1，当 data_['聚类'] 等于 63，加入一列'问题 ID'并填入 2，当 data_['聚类'] 等于 206，加入一列'问题 ID'并填入 3，当 data_['聚类'] 等于 47，加入一列'问题 ID'并填入 4，当 data_['聚类'] 等于 279，加入一列'问题 ID'并填入 5，最终将 data_1, data_2, data_3, data_4, data_5 连接起来得到 data_final，并将 data_final 的'聚类'这一列去掉。

将 data_final 导出成 excel 表为"热点问题留言明细表.xlsx"。

热点问题留言明细表（部分，具体见 Excel 文件）

	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	相关性	完整性	最终评价
0	2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映"A2区景蓉花苑物业管理有问题"的调查核...	2019/5/10 14:56:53	0.259	0.159184	1.627051
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友"A00023583"：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	0.044	0.018072	2.434667
2	2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的"请加快提高民营幼儿园教师的待遇"的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	0.371	0.138122	2.686040
3	2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友"A000110735"：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	0.329	0.127273	2.585000
4	2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将"白竹坡路口"更名为"马坡岭小学"，原...	网友"A0009233"，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议"白竹坡...	2019/5/9 9:51:30	0.726	0.139535	5.203000
...	...	...	...	...	...	...	...	...	...	...
2811	181267	UU008766	汽车北站进站口附近居民强烈反对建设I市平康肾病医院!	2018/12/12 15:20:46	\n\t\t\t\t\t\n\t\t\t\t\t我们是I市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转I1区委、区人民政府调查处理。	2019/1/8 16:54:53	0.000	0.050000	0.000000
2812	181603	UU008194	强烈反对I市9路公交车改线路	2018/6/12 8:51:03	\n\t\t\t\t\t\n\t\t\t\t\t强烈反对I市9路公交车改线路获悉从7月1日起...	"UU008194"您的留言已收悉。关于您反映的问题，已转市交通运输局调查处理。	2018/7/4 16:55:53	0.000	0.000000	NaN
2813	184423	UU0082115	对G7县文盛小学特色班的一点质疑	2018/10/11 20:02:52	\n\t\t\t\t\t\n\t\t\t\t\tG7县文盛小学引入特色班，每个学生必须参加，...	"UU0082115"您好！获悉关于"对G7县文盛小学特色班的质疑"的网帖后，我局领导高度重...	2018/10/24 9:22:07	0.211	0.110465	1.910105
					\n\t\t\t\t\t\n\t\t\t\t\t[illegible]	西地省平台《问政西地省》栏目组:				

通过对“热点问题留言明细表.xlsx”进行解读，我们可以发现在热度第一的问题中，“A 市楚江财富金融中心未达标就交房了”这条留言主题受到的非议比较大，受到了 1 个反对 2 个点赞，“关于对 A 市物业服务等级的确定依据不理解”和“A4 区辰北三角洲 C2 区业主申请重新选聘物业”这两条评论收到了相对较多的点赞数。 热度第二的问题中，“投诉 A 市轨道 3 号线保洁服务项目招标”这条留言主题受到的非议比较大，“A 市“江千锦园”团购户维权”和“请调查西地省建望集团及西地省辉 K4 县建工程有限公司的违法行为”这两条评论收到了相对较多的点赞数。热度第三的问题中，“A 市高新区白马学校收费贵”这条留言主题受到的非议比较大，“A 市经济学院强制学生外出实习”和“A 市经济学院体育学院变相强制实习”收到了较多的反对数，在热度第四的问题中，“A7 县妇幼医院连儿童常见病都无法接诊”这条留言主题受到的非议比较大，“西地省人民医院的停车费贵的太离谱”“请 A7 县星沙派出所民警合理执法”“A7 县星沙八医院对病人置之不理”收到了较多的点赞数。热度第五的问题中留言主题和详情受到的点赞数和反对数较少。

答复意见的评价

5.1 导入并处理数据

读取附件 4.xlsx 赋给 data，将 data 中的留言详情和答复意见进行利用正则表达式去除‘\t\n、空格、数字、字母‘等字符的操作。去除完之后对留言详情和答复意见进行 jieba 分词，由于在留言详情以及答复意见中存在'市长'，'局长'，'领导'，'干部'，'书记'，'您好'，'你好'，'尊敬'，'网友'等无关词，所以把这些无关词

也加入到导入停用词文件的 stop_word 中，然后将留言详情和答复意见进行去除停用词和长度为 1 的词的操作。

5.2 评价方案设计

评价主要从两个方面入手。一是答复意见与留言详情的相关性，这一点可以看出答复建议是否与问题有关，是否能很好的解决民生问题，二是看答复意见的完整性，从这一分析可以看意见是否全面，减少懒政怠政问题。

5.3 相似度分析

5.3.1 统计词频

将留言详情和答复意见分词后的结果用 append（）加到 allwords 中，以留言详情和答复意见的所有分词后的文本内容来构建一个语料库，然后利用 TF-IDF 方法形成一个词频矩阵。

5.3.2 计算余弦相似度

每一个文本内容都变成了一个词向量，利用 cosine_similarity()函数可以计算每一行留言详情与答复意见的余弦值，余弦值月接近于 1，相似度越大，余弦值越接近于 0，相似度越小。将余弦值加入到空集合 cos 中。

5.3.3 生成新列

df 为新生成的余弦相似度的列，通过 DataFrame 函数将生成的余弦相似度添加到原有 data 中。

5.4 完整度分析

5.4.1．导入并分析数据

读取附件 4.xlsx 赋给 data_2 将 data_2 中的留言详情和答复意见进行利用正则表达式去除‘\t\n、空格、数字、字母‘等字符的操作。去除完进行分词。由于在留言详情以及答复意见中存在’市长'，'局长'，'领导'，'干部'，'书记'，'您好'，'你好'，'尊敬'，'网友'等无关词，所以把这些无关词也加入到导入停用词文件的 stop_word 中，然后将留言详情和答复意见进行去除停用词和长度为 1 的词的操作。

5.4.2 进行完整度分析

本次分词利用词语所占比例来进行分析，算出答复意见分词在留言详情中所占的比例，用以当做完整度的评判标准。

首先对答复意见分词数进行统计，统计数填充在‘答复意见分次数’中。

计算大幅意见和留言详情共有的分词数。

利用 for 语句循环叠加出答复意见和留言详情共有的分词数，填充在‘答复意见 in 留言详情分词数’中。

最后计算完整度，完整度及用答复意见和留言详情共有的分词数去除以答复意见的分词数，得到的数字填充到‘完整度’中。

5.5 总结评价

5.5.1 整合数据

将 data_2 中的‘完整度’中的数据赋给 data 中的‘完整性’。

最终评价等于余弦值的一半除以完整性的一半，得出的结果填充到 data 表的‘最终评价‘中。

5.5.2 形成最终评价表格

读取附件 4.xlsx 赋给 data_final，将刚刚得到的 data['余弦值']赋给 data_final[‘相关性’]，data['完整性']赋给 data_final[‘完整性’]，data[‘最终评价’]赋给 data_final[‘最终评价’]，最后得到最终评价表格 data_final。

如图为最终评价表(部分)

	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	相关性	完整性	最终评价
0	2549	A00045581	A2区景荟华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\n\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景荟花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	0.259	0.159184	1.627051
1	2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	0.044	0.018072	2.434667
2	2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\n\t\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	0.371	0.138122	2.686040
3	2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\n\t\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	0.329	0.127273	2.585000
4	2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\n\t\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于采信人建议“白竹坡...	2019/5/9 9:51:30	0.726	0.139535	5.203000
...	...	...	...	...	...	...	...	...	...	...
2811	181267	UU008766	汽车北站进站口附近居民强烈反对建设|市平康肾病医院！	2018/12/12 15:20:46	\n\t\t\t\t\n\t\t\t\t\t\t我们是|市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转|1区委、区人民政府调查处理。	2019/1/8 16:54:53	0.000	0.050000	0.000000
2812	181603	UU008194	强烈反对|市9路公交车改线路	2018/6/12 8:51:03	\n\t\t\t\t\n\t\t\t\t\t\t强烈反对|市9路公交车改线路获悉从7月1日起...	“UU008194“您的留言已收悉。关于您反映的问题，已转市交通运输局调查处理。	2018/7/4 16:55:53	0.000	0.000000	NaN
2813	184423	UU0082115	对G7县文盛小学特色班的一点质疑	2018/10/11 20:02:52	\n\t\t\t\t\n\t\t\t\t\t\tG7县文盛小学引入特色班，每个学生必须参加，...	“UU0082115“您好！获悉关于“对G7县文盛小学特色班的质疑”的网帖后，我局领导高度重...	2018/10/24 9:22:07	0.211	0.110465	1.910105