基于 fasttext 模型与 AHP 的
“智慧政务”文本挖掘应用技术研究

摘要：近年来，随着互联网技术的广泛应用和社会群众意识的提高增长，众多网络问政平台逐步成为政府了解群众、服务群众的重要渠道。因此，运用网络文本分析和数据挖掘技术对网络问政平台的群众留言的研究具有重大的意义。

针对问题一，本文首先将附件 2 中的‘留言主题’和‘留言详情’进行去空、停用词过滤、中文分词等数据预处理，然后基于 fastText 方法构建分类模型，对群众留言信息进行分类并计算 F1-Score。

针对问题二，首先将附件 3 中的‘留言主题’和‘留言详情’进行去空、停用词过滤、中文分词等数据预处理，接着通过 TF-IDF 算法提取关键词构造词汇-文本矩阵。再通过层次聚类和 k-means 聚类算法进行聚类对每个分类进行热度分析，用 Reddit 算法分析时间间隔、点赞反对数，得出热度指数，最终给出排名前五的热点问题。

针对问题三，利用网络层次分析法(ANP)构建留言答复构建综合指标体系模型，主要从留言答复的相关性、完整性、可解释性三个角度进行分析，提取出回应问题、时效性、专业用语、礼貌用语、字数标准、提出意见、工作反馈 7 个特征，客观设计判断矩阵，最后算出各个指标的权重赋值。

关键词：fastText；TF-IDF 算法；K-means 聚类；ANP 评价模型

Abstract

In recent years, with the wide application of Internet technology and the increase of social mass consciousness, many online political platforms have gradually become an important channel for the government to understand and serve the masses. Therefore, the use of network text analysis and data mining technology is of great significance to the research of the mass message of network politics platform.

To solve the first problem, this paper first processes the data of "message subject" and "message details" in Annex 2, such as de empting, stop word filtering and Chinese word segmentation, and then constructs a classification model based on fasttext method to classify the message information of the masses and calculate F1 score.

In order to solve the second problem, firstly, the "message subject" and "message details" in Annex 3 are preprocessed with data such as de empting, stop words filtering, Chinese word segmentation, and then the TF-IDF algorithm is used to extract keywords to construct the vocabulary text matrix. Then through the hierarchical clustering and K-means clustering algorithm to analyze the heat degree of each classification, using reddit algorithm to analyze the time interval, the number of likes and objections, get the heat degree index, and finally give the top five hot issues.

Aiming at question 3, this paper uses ANP to build a comprehensive index system model of message reply, which mainly analyzes from three angles of relevance, integrity and explainability of message reply, extracts seven characteristics of response question, timeliness, professional language, polite language, word number standard, putting forward opinions and work feedback, designs judgment matrix objectively, and finally calculates the judgment matrix Weight assignment of each indicator.

Key words: FastText; Jieba segmentation; TF-IDF algorithm; K-means clustering; ANP evaluation model

目录

1.挖掘目标

1.1 问题描述

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。我们需要根据附件 2 给出的群众留言的文本数据，建立关于留言内容的一级标签分类模型。并且用 F1-Score 对分类方法进行评价。

某一时段内群众集中反映的某一问题可称为热点问题，及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。我们需要根据附件 3 的‘留言主题’、‘ 留言详情’、‘留言时间’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，按照规定的格式给出排名前 5 的热点问题及相应热点问题对应的留言信息。

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、可解释性三个角度对答复意见的质量给出一套评价方案及实现。

1.2 总体流程与步骤

本文的总体架构及思路如下：

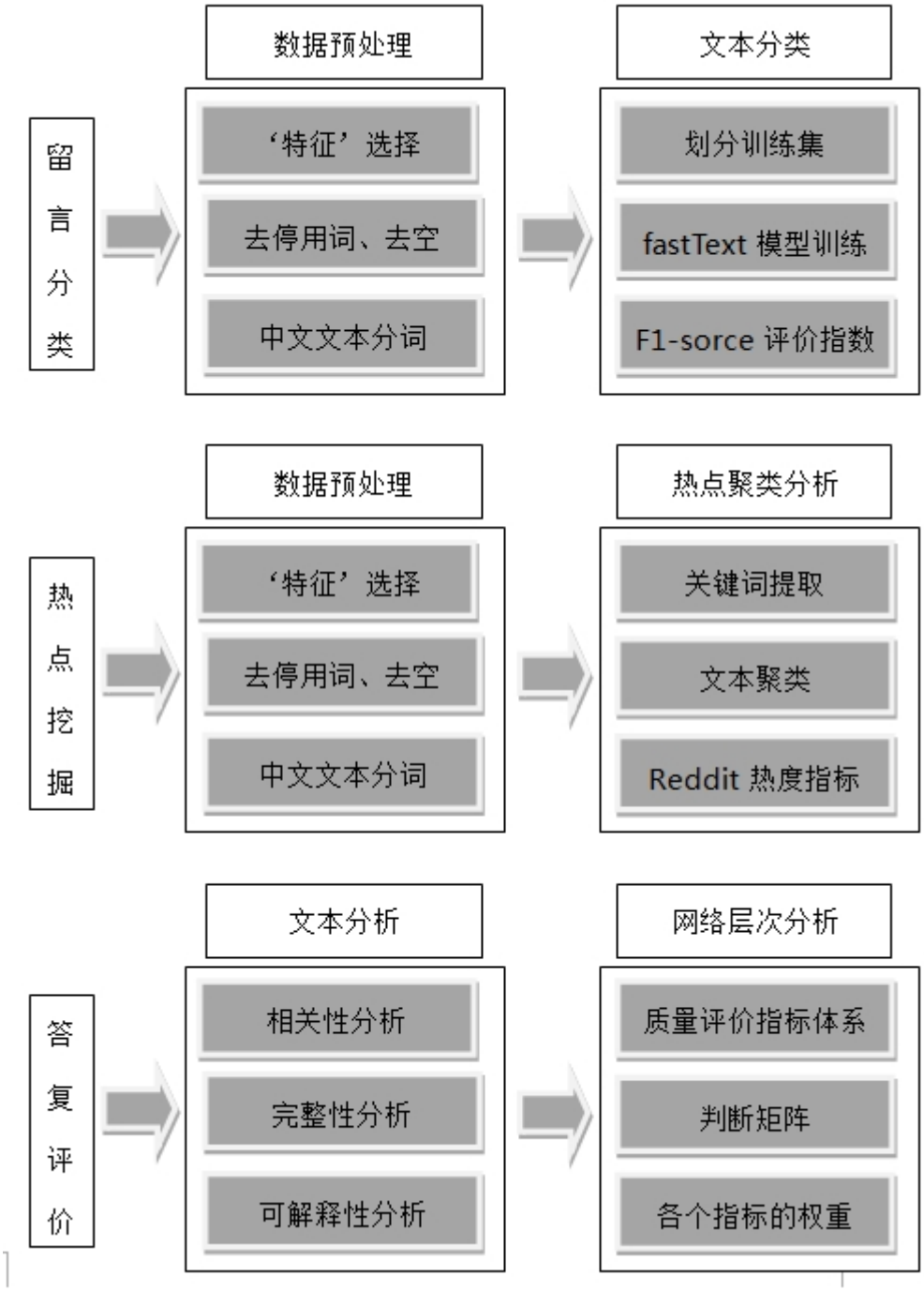

图 1.2 总体流程

步骤一：分类数据预处理，在附件 2 的‘留言主题’、‘留言详情’中，需要先提取文本的‘特征’，即选择需要处理的对象。而‘特征’中存在空格、换行

和停用词，需要对原始数据进行去空及去停用词，从而达到减少文本赘余的特征，降低文本向量的维度，最后再进行中文分词。

步骤二：文本分类，以 0.9/0.1 划分训练集和测试集，90%的文本作为 fastText 模型的训练指标，最后预测文本并计算 F1-Sorce。

步骤三：聚类数据预处理，对附件 3 的‘留言主题’、‘留言详情’提取文本的‘特征’，‘ 特征’中有空格、换行符和停用词等冗余信息，必须将这些部分删除以达到减少文本向量维度的目的，最后使用中文分词。

步骤四：热点聚类分析，分词后的文本需要进行关键词的提取（地名‘ns’、名词‘n’），提取完关键词后需要将中文信息转换成数字信息，即词向量。将处理后的词向量矩阵放入层次聚类模型中得出其最大的聚类数，再把最大聚类数做为 K 值放入 kmeans 再次进行迭代聚类，得出最优的聚类结果。对于聚类后得到的文本信息需要通过 Reddit 计算其热度指数。

步骤五：文本分析，即通过分析文本答复信息与留言信息的匹配度以及相关度和答复信息的完整性，可以得出答复的相关性分析、答复的完整性分析、答复的可解释性性分析。

步骤六：网络层次分析（AHP），对相关性、完整性、可解释性构建质量评价指标体系，算出各自的判断矩阵，最后用网络层次分析的软件得到各个指标的赋值权重。

2.群众留言分类

2.1 分类模型

2.1.1 fastText 模型

FastText 是 Facebook 于 2016 年开源的一个词向量计算和文本分类工具，在文本分类任务中，FastText（浅层网络）往往能取得和深度网络相媲美的精度，却在训练时间上比深度网络快许多数量级。在标准的多核 CPU 上， 能够训练 10 亿词级别语料库的词向量在 10 分钟之内，能够分类有着 30 万多类别的 50 多万句子在 1 分钟之内。

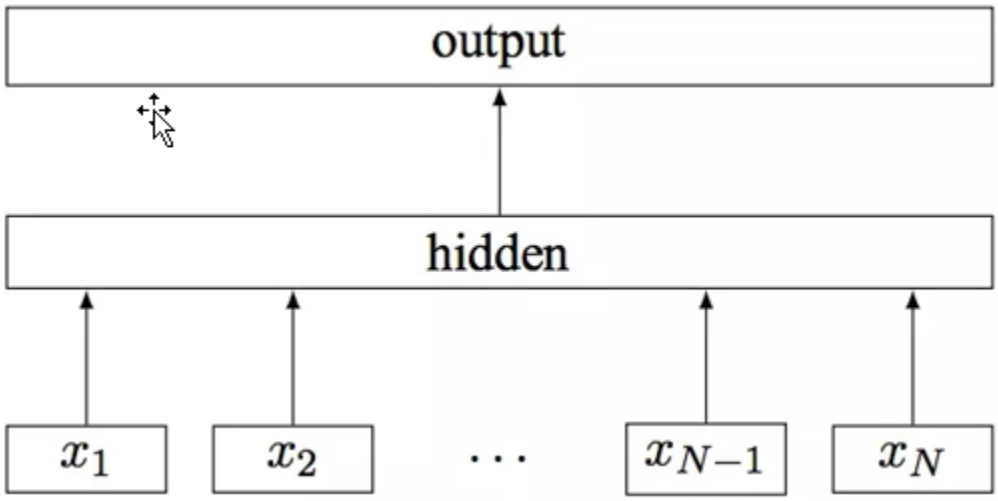

图 2.1 FastText 模型架构

其中 $x_1,x_2,\cdots,x_{N-1},x_N$ 表示一个文本中的 n-gram 向量，每个特征是词向量的平均值。CBOW 用上下文去预测中心词，而此处用全部的 n-gram 去预测指定类别。，fastText 模型有三层：输入层、隐含层、输出层（Hierarchical Softmax），输入都是多个经向量表示的单词，输出都是一个特定的 target，隐含层都是对多个词向量的叠加平均。

fastText 在输入时，将单词的字符级别的 n-gram 向量作为额外的特征；在输出时，fastText 采用了分层 Softmax，大大降低了模型训练时间。

2.1.2 word2vec 模型

word2vec 是 Google 团队发表的工具。word2vec 工具主要包含两个模型：跳字模型（skip-gram）和连续词袋模型（continuous bag of words，简称 CBOW），以及两种高效训练的方法：负采样（negative sampling）和层序 softmax（hierarchical softmax）。

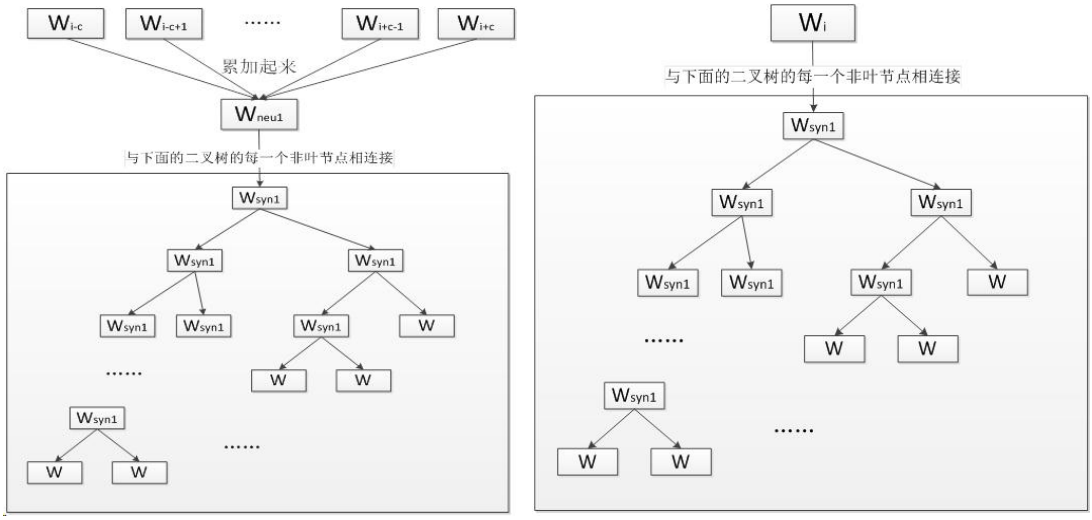

图 2.1 wor2vec 模型架构

word2vec 又可称为 word embedding，即词嵌入，其表示方法是在神经概率模型中的嵌入层使用，作为神经网络模型能够处理的符号数据使用。

每一个词在 word2vec 中同样是一个向量，但该向量的长度不固定，即非语料库的长度，该长度 M 可有算法的使用者而定，而向量的每个维度也并非直接使用 0 或 1 这种简单的方式表示，一般是任意的实数。由此表示形式，可以得出，这种词的表示方法既可以表示当前词的身份信息（即区别于其他词），又可以计算当前词与其他词之间的语义上的关系。

我们可以思考一下这种词之间的语义关联性。首先，对于每一个词而言，因其是多维向量，因此可以映射到高维空间中。其次，在映射到高维空间后，可以联系向量之间的关系，通过计算两个向量之间的余弦值而得出两个向量之间的关联程度。更甚之，可以通过词向量之间的加减，计算出其他词。

对于如何获取词向量而言，对于神经概率模型中，词向量（word embedding）即为神经网络中的参数，因其作为嵌入层的参数，输入形式为实数向量。因此获取 word embedding，即是神经网络中的参数的学习。

这种训练通过给定一个训练序列数据集，而神经网络模型的训练目标则是为找到一组参数 θ，使得对数似然函数最大。其中 θ 表示网络中的所有参数，包括嵌入矩阵(M×N)以及神经网络的权重和偏置。

2.1.3 fastText 与 word2vec 异同点

（1）FastText 与 Word2Vec 的相同之处：

图模型结构很像，都是采用 embedding 向量的形式，得到 word 的隐向量表达。采用很多相似的优化方法，比如使用 Hierarchical softmax 优化训练和预测中的打分速度。

训练词向量时，两者都是无监督算法。输入层是 context window 内的 term。输出层对应的是每一个 term，计算某 term 的概率最大；

在使用层次 softmax 的时候，huffman 树叶子节点处是训练语料里所有词的向量。

（2）FastText 与 Word2Vec 的不同之处：

模型的输出层：word2vec 的输出层，对应的是每一个 term，计算某 term 的概率最大；而 fastText 的输出层对应的是分类的 label。不过不管输出层对应的是什么内容，其对应的 vector 都不会被保留和使用；

模型的输入层：word2vec 的输入层，是 context window 内的 term；而 fastText 对应的整个 sentence 的内容，包括 term，也包括 n-gram 的内容；

两者本质的不同，体现在 Hierarchical softmax 的使用。Wordvec 的目的是得到词向量，该词向量最终是在输入层得到，输出层对应的 Hierarchical softmax 也会生成一系列的向量，但最终都被抛弃，不会使用。

fastText 则充分利用了 Hierarchical softmax 的分类功能，遍历分类树的所有叶节点，找到概率最大的 label（一个或者 N 个）

FastText 的优点是适合大型数据+高效的训练速度、支持多语言表达 fastText 专注于文本分类，在许多标准问题上实现当下最好的表现（例如文本倾向性分析或标签预测）。比 word2vec 更考虑了相似性，比如 fastText 的词嵌入学习能够考虑 english-born 和 british-born 之间有相同的后缀，但 word2vec 却不能。

2.2 fastText 模型搭建

2.2.1 模型搭建步骤

模型搭建遵循以下步骤：

（1）添加输入层（embedding 层）。Embedding 层的输入是一批文档，每个文档由一个词汇索引序列构成。Embedding 层将每个单词映射成 EMBEDDING_DIM 维的向量，每个词被分布式表示的向量的维度，这里设置为 100。比如对于“油烟”这个词，会被一个长度为 100 的类似于[0.97860014, 5.93589592, 0.22342691, -3.83102846, -0.23053935, …]的实值向量来表示。

（2）添加隐含层（投影层）。投影层对一个文档中所有单词的向量进行叠加平均。keras 提供的 GlobalAveragePooling1D 类可以帮我们实现这个功能。

（3）添加输出层（softmax 层）。fastText 这层是 Hierarchical Softmax，因为 keras 原生并没有支持 Hierarchical Softmax，所以这里用 Softmax 代替。这层指定了 CLASS_NUM，对于一篇文档，输出层会产生 CLASS_NUM 个概率值，分别表示此文档属于当前类的可能性。

（4）指定损失函数、优化器类型、评价指标，编译模型。损失函数我们设置为 categorical_crossentropy，它就是我们上面所说的 softmax 回归的损失函数；优化器我们设置为 SGD，表示随机梯度下降优化器；评价指标选择 accuracy，表示精度。

2.2.2 fastText 算法过程

（1）通过 embedding 层，我们将词汇映射成 EMBEDDING_DIM 维向量

（2）通过 GloalAveragePoolingID，平均了文档中所有词的 embedding

（3）通过输出层 Softmax 分类，得到类别概率分布

（4）定义损失函数、优化器、分类度量指标

（2）计算准确率，召回率，F1-Score 值

2.3 F1-Sorce

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

本文用 fastText 在全部数据上根据训练集与测试集比例为 9:1 进行预测，得到的 F1-Score 值为 0.91，准确度比 word2vec 算法高，所以认为 fastText 对政务文本信息的处理较其他算法好。

2.4 结果

最终通过模型分类得到分类结果的 F1-Sorce 为 91.6%

划分训练集和测试集

```
s_data = StratifiedShuffleSplit(n_splits=10, test_size=0.1, train_size=0.9, random_state=0)
for train_index, test_index in s_data.split(data, data['一级标签']):
    data_train = data.loc[train_index]
    data_test = data.loc[test_index]
print(data_train['一级标签'].value_counts()/len(data_train))
#一级分类占比
pre_data_process(data_train,'train_data')
pre_data_process(data_test,'test_data')
```

```
城乡建设        0.218120
劳动和社会保障     0.213777
教育文体        0.172518
商贸旅游        0.131982
环境保护        0.101822
卫生计生        0.095186
交通运输        0.066594
Name: 一级标签, dtype: float64
```

```
model = fasttext.train_supervised('train_data.txt', lr=0.1, dim=100, bucket=2000000, epoch=100, word_ngrams=1,verbose = 1)
model.save_model('classifier.bin')

# 加载已经训练好的文本分类器
classifier = fasttext.load_model('classifier.bin')
# 使用测试数据集评估模型,数据格式与训练数据集一样
sorce = classifier.test('test_data.txt',k = 1)
print(sorce)
print("P查准率：",sorce[1])
print("P查全率：",sorce[2])
print('F1 Score：%0.3f'%((2*sorce[1]*sorce[2])/(sorce[1]+sorce[2])))
```

```
(921, 0.9163952225841476, 0.9163952225841476)
P查准率： 0.9163952225841476
P查全率： 0.9163952225841476
F1 Score：0.916
```

图 2.2 模型分类结果

3．文本聚类与热度指数

3.1. 数据预处理

3.1.1 数据描述

通过观察所给数据，附件 3 中需要处理的文本格式字段是‘留言详情’和‘留言主题’，而需要处理的数值字段是‘留言时间’、‘点赞数’、‘反对数’。对于分析和聚类文本数据，需要将其量化成数值形式，而且文本中存在大量空行以及停用词的情况，如果不做处理会对后续分析造成影响。关于热度评价指标，需要根据附件 3 给出的‘留言时间’、‘点赞数’、‘反对数’，需要预处理的数据量主要是点赞数与反对数的差值，以及留言时长。于是本文首先要对文本数据以及数值进行预处理。

3.1.2 文本预处理

(1)‘特征’选择

附件 3 文本中‘留言详情’和‘留言主题’是此次文本聚类的主要内容，进行分析数据前需要将‘留言详情’和‘留言主题’合并成为新的‘特征’。

(2) 去停用词、去空

‘特征’中存在大量空行和停用词。为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words（停用词）。停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。

(3)中文分词

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 分词[1]，对附件 3 中每一个岗位描述进行中文分词，jieba 分词用到的算法：

①基于 Trie 树结构[2]实现高效的词图扫描，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)

②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

③对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

jieba 分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。

	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	特征	data_jieba_cut
2906	257384	A00098443	现由A7县规划的山河智能到锦泰广场的公交什么时候能开通?	2019/3/7 18:32:24	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的书记：你好？2018年4月6号公布的山...	0	5	现由A7县规划的山河智能到锦泰广场的公交什么时候能开通? \n\t\t\t\t\t\t\n\t\t...	现由 A7 县 规划 的 山河 智能 到 锦泰 广场 的 公交 什么 时候 能 开通 ？ \...
1217	217186	A00053545	A市民政职业技术学院凌晨施工	2019/9/6 15:06:49	\n\t\t\t\t\t\t\n\t\t\t\t\t\t2019年9月6日凌晨学校说要搞装修，两个月...	0	0	A市民政职业技术学院凌晨施工 \n\t\t\t\t\t\t\n\t\t\t\t\t\t2019年9月6...	A 市 民政 职业 技术 学院 凌晨 施工 \n \t \t \t \t \t \n \t ...
1333	219646	A00093252	A市网约车司机报名网约车驾驶员考试却迟迟收不到考试通知	2019/3/20 10:33:40	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我是一名普通的网约车司机，现在正在报名参加网...	0	0	A市网约车司机报名网约车驾驶员考试却迟迟收不到考试通知 \n\t\t\t\t\t\t\n\t\t\...	A 市网 约 车 司机 报名 网约车 驾驶员 考试 却 迟迟 收不到 考试通知 \n \t ...
4004	283220	A00020612	A3区楷林楼下的便利店光污染严重	2019/9/5 17:56:23	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导：现反应一处离小区很近的楷林一楼今...	0	0	A3区楷林楼下的便利店光污染严重 \n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导：...	A3 区 楷林 楼下 的 便利店 光污染 严重 \n \t \t \t \t \t \n \...
3276	265551	A00076292	咨询A市公积金贷款买房的问题	2019/5/23 15:06:05	\n\t\t\t\t\t\t\n\t\t\t\t\t\t去年购买了首套房，但是当时刚参加工作，还未缴...	0	0	咨询A市公积金贷款买房的问题 \n\t\t\t\t\t\t\n\t\t\t\t\t\t去年购买了首套房...	咨询 A 市 公积金 贷款 买房 的 问题 \n \t \t \t \t \t \n \t ...

图 3.1 部分分词结果

	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	特征	data_jieba_cut	data_key_word
2906	257384	A00098443	现由A7县规划的山河智能到锦泰广场的公交什么时候能开通?	2019/3/7 18:32:24	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的书记：你好？2018年4月6号公布的山...	0	5	现由A7县规划的山河智能到锦泰广场的公交什么时候能开通? \n\t\t\t\t\t\t\n\t\t...	现由 A7 县 规划 的 山河 智能 到 锦泰 广场 的 公交 什么 时候 能 开通 ？ \...	A7 锦泰 山河 开通 智能 广场 公交线 现由 2018 规划 交通局 县内 时候 公交 ...
1217	217186	A00053545	A市民政职业技术学院凌晨施工	2019/9/6 15:06:49	\n\t\t\t\t\t\t\n\t\t\t\t\t\t2019年9月6日凌晨学校说要搞装修，两个月...	0	0	A市民政职业技术学院凌晨施工 \n\t\t\t\t\t\t\n\t\t\t\t\t\t2019年9月6...	A 市 民政 职业 技术 学院 凌晨 施工 \n \t \t \t \t \t \n \t ...	凌晨 2019 我门 12 点来 钻机 暑假 开学 民政 装修 时间 宿舍 偏偏 早上 施工...
1333	219646	A00093252	A市网约车司机报名网约车驾驶员考试却迟迟收不到考试通知	2019/3/20 10:33:40	\n\t\t\t\t\t\t\n\t\t\t\t\t\t我是一名普通的网约车司机，现在正在报名参加网...	0	0	A市网约车司机报名网约车驾驶员考试却迟迟收不到考试通知 \n\t\t\t\t\t\t\n\t\t\...	A 市网 约 车 司机 报名 网约车 驾驶员 考试 却 迟迟 收不到 考试通知 \n \t ...	考试 网约车 报名 迟迟 驾驶员 请问 司机 驾校 市网 考试通知 局驾 培处 要开 收不到...
4004	283220	A00020612	A3区楷林楼下的便利店光污染严重	2019/9/5 17:56:23	\n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导：现反应一处离小区很近的楷林一楼今...	0	0	A3区楷林楼下的便利店光污染严重 \n\t\t\t\t\t\t\n\t\t\t\t\t\t尊敬的领导：...	A3 区 楷林 楼下 的 便利店 光污染 严重 \n \t \t \t \t \t \n \...	便利店 光污染 关灯 30 小区 灯光 店里 投诉 公德心 我们 邻居 楷林 根本 夜间 部...
3276	265551	A00076292	咨询A市公积金贷款买房的问题	2019/5/23 15:06:05	\n\t\t\t\t\t\t\n\t\t\t\t\t\t去年购买了首套房，但是当时刚参加工作，还未缴...	0	0	咨询A市公积金贷款买房的问题 \n\t\t\t\t\t\t\n\t\t\t\t\t\t去年购买了首套房...	咨询 A 市 公积金 贷款 买房 的 问题 \n \t \t \t \t \t \n \t ...	公积金 商业贷款 缴纳 结婚 婚房 购买 贷款 办手续 当时 使用 名下 但是 套房 问题 ...

图 3.2 部分关键词结果

3.2 文本特征抽取方法

经过上述文本预处理后，虽然已经去掉部分停用词，但还是包含大量词语，给文本向量化过程带来困难，所以特征抽取的主要目的是在不改变文本原有核心信息的情况下尽量减少要处理的词数，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。常用的方法有词频-逆向文档频率(TF-IDF)、X^2统计、命名实体识别等。

3.2.1 词频-逆向文档频率(TF-IDF)

词频(TF)是词语在文本中出现的频率，如果某一个词在一个文本中出现的越多，它的权重就越高。它的基本公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

逆向文档频率(IDF)是指在少数文本中出现的词的权重比在多数文本中出现的词的权重高，因为在聚类中这些词更具有区分能力。它的基本公式如下：

$$idf_i = log\frac{N}{|\{j:t_i \in d_j\}|}$$

最后可以得出：

$$w_{ij} = tf_{i,j} \times idf_i$$

3.2.2 X^2统计

X^2统计方法度量词条 t 和文档类别之间的相关程度，词条对于某类的X^2统计值越高，它与该类之间的相关性越大，携带的类别信息也较多。假设 t 和 c 之间符合具有一阶自由度的X^2分布。令 N 表示训练语料中的文本总数， c 为某一特定类别，t 表示特定的词条，A 表示属于 c 类且包含 t 的文档频数，B 表示不属于 c 类但是包含 t 的文档频数，C 表示属于 c 类但是不包含 t 的文档频数，D 是既不属于 c 也不包含 t 的文档频数，则 t 对于 c 的X^2值由下式计算：

$$X^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

3.2.3 命名实体识别

命名实体识别（NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。简单的讲，就是识别自然文本中的实体指称的边界和类别。

早期的命名实体识别方法基本都是基于规则的。之后由于基于大规模的语料库的统计方法在自然语言处理各个方面取得不错的效果之后，一大批机器学习的方法也出现在命名实体类识别任务。宗成庆老师在统计自然语言处理一书粗略的将这些基于机器学习的命名实体识别方法划分为以下几类：

有监督的学习方法：这一类方法需要利用大规模的已标注语料对模型进行参数训练。目前常用的模型或方法包括隐马尔可夫模型、语言模型、最大熵模型、支持向量机、决策树和条件随机场等。值得一提的是，基于条件随机场的方法是命名实体识别中最成功的方法。

半监督的学习方法：这一类方法利用标注的小数据集（种子数据）自举学习。

无监督的学习方法：这一类方法利用词汇资源（如 WordNet）等进行上下文聚类。

混合方法：几种模型相结合或利用统计方法和人工总结的知识库。

值得一提的是，由于深度学习在自然语言的广泛应用，基于深度学习的命名实体识别方法也展现出不错的效果，此类方法基本还是把命名实体识别当做序列标注任务来做，比较经典的方法是 LSTM+CRF、BiLSTM+CRF。

3.2.4 方法选取

X^2统计方法基于X^2分布，如果这种分布被打破，则对低频词不可靠。

TF-IDF 算法可以将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重。

因此本文采用比较有效的 TF-IDF 算法，提取出关键词的 TF-IDF 矩阵，并求出不同文本之间相似度矩阵。

```
#计算相似性
from sklearn.metrics.pairwise import cosine_similarity
dist = cosine_similarity(tfidf)
print(dist)
```

```
[[1.         0.01318232 0.01827928 ... 0.0109344  0.00264904 0.02083455]
 [0.01318232 1.         0.04963971 ... 0.03195812 0.00576159 0.00200906]
 [0.01827928 0.04963971 1.         ... 0.03400813 0.         0.00330946]
 ...
 [0.0109344  0.03195812 0.03400813 ... 1.         0.16002035 0.14348556]
 [0.00264904 0.00576159 0.         ... 0.16002035 1.         0.11172844]
 [0.02083455 0.00200906 0.00330946 ... 0.14348556 0.11172844 1.        ]]
```

图 3.3 TF-IDF 矩阵

3.3 文本聚类

聚类方法作为无监督学习的一个重要方法，聚类的思想就是把属性相似的样本归到一类。对于每一个数据点，我们可以把它归到一个特定的类，同时每个类之间的所有数据点在某种程度上有着共性，比如空间位置接近等特性。目前流行的聚类方法包括 K-means 聚类、Affinity Propagation(AP)聚类、层次聚类。

文本聚类是将无类别标记的文本信息根据不同的特征，将有着各自特征的文本进行分类。通过文本聚类的方法对具有近似相同的‘留言主题’、‘ 留言详情’进行聚类，可以及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率分析居民。

3.3.1 K-means 聚类

K-means 算法是基于划分的聚类算法，采用欧氏距离作为相似性的评价指标。K-means 算法的基本流程是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。假设要把样本集分为 k 个类别，算法描述如下：

（1）在样本中随机选取 K 个点，作为每一类的中心点。

（2）计算剩下 n-K 个样本点到每个聚类中心的欧氏距离，对于每一个样本点，将它归到和他距离最近的聚类中心所属的类；

（3）更新每个聚类中心的位置。步骤 2 中得到的结果是 n 个点都有自己所属的类，将每一个类内的所有点取平均值，计算出新的聚类中心。

（4）重复步骤 2 和 3 的操作，直到所有的聚类中心不再改变。

K-means 算法中 K 值的选择十分关键，这里可以引入类内距离 J，每一类都会对应一个 J 值，其计算就是把类内所有点之间的距离累加起来。我们肯定希望 J 越小越好，因为小的类内间距代表这一类样本的相似程度更高（离得更近）。如果 K 很小，则聚类可能不彻底，即隔着很远的两波点也被聚为一类，会使 J 变得很大；相反的，过大的 K 虽然会降低类内间距 J，但有时候分得过细会对数据的泛化性造成损害。

3.3.2 AP 聚类

将全部样本看作网络的节点，然后通过网络中各条边的消息传递计算出各样本的聚类中心。聚类过程中，共有两种消息在各节点间传递，分别是吸引度(responsibility)和归属度(availability) 。AP 算法通过迭代过程不断更新每一个点的吸引度和归属度值，直到产生 m 个高质量的 Exemplar（类似于质心），同时将其余的数据点分配到相应的聚类中。AP 算法流程：

（1）算法初始，将吸引度矩阵 R 和归属度矩阵初始化为 0 矩阵。

（2）更新吸引度矩阵。

$$\boldsymbol{T_{t+1}(i,k)=\begin{cases}S(i,k)-max_{j\neq k}\{a_t(i,j)+r_t(i,j)\},i\neq k\\ S(i,k)-max_{j\neq k}\{S(i,j)\},i=k\end{cases}}$$

（3）更新归属度矩阵。

$$\boldsymbol{a_{t+1}(i,k)=\begin{cases}min\left\{0,r_{t+1}(k,k)+\sum_{j\neq i,k}max\{r_{t+1}(j,k),0\}\right\},i\neq k\\ \sum_{j\neq k}max\{r_{t+1}(j,k),0\},i=k\end{cases}}$$

（4）根据衰减系数 λ 对两个公式进行衰减。

$$\boldsymbol{r_{t+1}(i,k)=\lambda*r_t(i,k)+(1-\lambda)*r_{t+1}(i,k)}$$

$$\boldsymbol{a_{t+1}(i,k)=\lambda*a_t(i,k)+(1-\lambda)*a_{t+1}(i,k)}$$

（5）重复步骤 2,3,4 直至矩阵稳定或者达到最大迭代次数，算法结束。最终取 a+r 最大的 k 作为聚类中心。

3.3.3 层次聚类

层次聚类通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法。

层次聚类流程：

（1）把每个样本归为一类，计算每两个类之间的距离，也就是样本与样本之间的相似度。

（2）寻找各个类之间最近的两个类，把他们归为一类（这样类的总数就少了一个）。

（3）重新计算新生成的这个类与各个旧类之间的相似度。

（4）重复 2 和 3 直到所有样本点都归为一类。

3.3.4 聚类方法的选择

（1）AP 聚类方法基于网络进行分类，其特点是处理速度快，而且与目标数据库中记录的个数无关，只与把数据空间分为多少个单元有关。当数据量比较庞大时，无法准确把控 preference 的数值，从而无法精准分类。

（2）层次聚类方法基于层次进行分类，其特点是较小的计算开销。通过层次聚类可以得出 K-means 中的 K 值，为下一步聚类计算提供了较大的准确度。

（3）K-means 聚类方法基于划分进行分类，其特点是计算量大。很适合发现中小规模的数据库中小规模的数据库中的球状簇。因此本文采用 K-means 算法进行文本的聚类。

```
#AP算法
from sklearn.cluster import AffinityPropagation
from sklearn import metrics
from sklearn.datasets.samples_generator import make_blobs
import pylab as pl
from itertools import cycle

# Compute Affinity Propagation
X=tfidf
af = AffinityPropagation(preference= -2).fit(X)

#层次聚类
from scipy.cluster.hierarchy import ward, dendrogram, linkage
import scipy.cluster.hierarchy as sch
from scipy.cluster.vq import vq,kmeans,whiten
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

linkage_matrix = sch.linkage(dist, method='ward', metric='euclidean', optimal_ordering=False)
cluster= sch.fcluster(linkage_matrix,t=1,criterion='inconsistent')
X=tfidf
clf = KMeans(n_clusters=max(cluster),init='k-means++')#n_init选择质心的次数，返回最好的结果
s = clf.fit(X)
```

图 3.4 聚类方法

模型结果对比：

```
print("聚类结果：")
print("ap:",af.labels_)
print("层次聚类：",cluster)
print("层次+kmeans：",clf.labels_)
print("")
print("聚类评价：")
print("ap:",metrics.silhouette_score(X, af.labels_, metric='euclidean'))
print("层次聚类:",metrics.silhouette_score(dist, cluster, metric='euclidean'))
print("层次+kmeans:",metrics.silhouette_score(X, clf.labels_, metric='euclidean'))
```

```
聚类结果：
ap: [372 632 671 ... 719 719 719]
层次聚类： [1512 1188 1254 ... 1585 1577 1577]
层次+kmeans： [ 844 1536 1041 ...  445  445  445]

聚类评价：
ap: 0.6465310656346491
层次聚类: 0.8303106877940972
层次+kmeans: 0.9298630636012019
```

图 3.5 模型结果对比

由此可见，通过层次聚类后再将聚类数作为 k 值放入 kmeans 中得到的聚类效果是最好的。

3.4 热度评价指标

3.4.1 Wilson score interval 算法

信任排序使用 Wilson score interval 算法，它的数学表达式是这样的：

$$\frac{p+\frac{1}{2n}z_{1-\alpha/2}^2 \pm z_{1-\alpha/2}^2\sqrt{\frac{p(1-p)}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\alpha/2}^2}$$

在上面的公式中，各个参数的定义如下：

- p 是支持票的百分比
- n 总票数
- $z_{1-\alpha/2}^2$是正态分布$1-\alpha/2$分位数

我们对上面的介绍做一些总结：

- 信任排序是把票数看作一次全体读者的抽样调查
- 信任排序会给一条评论一个临时评级，认为它有 85%的可信度

- 票数越多，可信度越高
- Wilson’s interval 算法能很好的处理票数很少和低端概率情况

3.4.2 Reddit 算法

$$f(t_s,y,z) = log_{10}z + \frac{yt_s}{45000}$$

Reddit 是美国最大的网上社区，它的每个留言前面都有向上和向下的箭头，分别表示"赞成"和"反对"。用户点击进行投票，Reddit 根据投票结果，计算出最新的"热点文章排行榜"。需要考虑这样几个因素:

（1）留言的新旧程度t_s

$$t_s = A - B$$

A = 最新的留言时间　B = 最早的留言时间

（2）赞成票与反对票的差 x

x = U − D　　U = ‘点赞数’　　D = ‘反对数’

（3）投票方向 y

$$y = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{cases}$$

y 是一个符号变量，表示对文章的总体看法。如果赞成票居多，y 就是+1；如果反对票居多，y 就是-1；如果赞成票和反对票相等，y 就是 0。

（4）留言的受肯定（否定）的程度 z

$$z = \begin{cases} |x| & if\ |x| \geq 1 \\ 1 & if\ |x| < 1 \end{cases}$$

z 表示赞成票与反对票之间差额的绝对值。如果对某个留言的评价，越是一边倒，z 就越大。如果赞成票等于反对票，z 就等于 1。

（5）Reddit 的最终得分计算公式如下：

$$\text{Score} = f(t_s,y,z) = log_e z + \frac{yt_s}{45000}$$

这个部分表示，赞成票与反对票的差额 z 越大，得分越高。需要注意的是，这里用的是以 e 为底的对数，意味着 z=e 可以得到 1 分，z=e^2 可以得到 2 分。也就是说，前 2 个投票人与后 6 个投票人的权重是一样的，即如果一个留言特别受到欢迎，那么越到后面投赞成票，对得分越不会产生影响。

当赞成票等于反对票，z=1，因此这个部分等于 0，也就是不产生得分。

这个部分表示，t_s越大，得分越高，即新留言的得分会高于老留言。它起到自动将老留言的排名往下拉的作用。分母的 45000 秒，等于 12.5 个小时，也就是说，后一天的留言会比前一天的留言多得 2 分。结合前一部分，可以得到结论，如果前一天的留言在第二天还想保持原先的排名，在这一天里面，它的 z 值必须增加 100 倍（净赞成票增加 100 倍）。

y 的作用是产生加分或减分。当赞成票超过反对票时，这一部分为正，起到加分作用；当赞成票少于反对票时，这一部分为负，起到减分作用；当两者相等，这一部分为 0。这就保证了得到大量净赞成票的文章，会排在前列；赞成票与反对票接近或相等的文章，会排在后面；得到净反对票的文章，会排在最后（因为得分是负值）。

	A	B	C	D	E	F
1	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
2	1	1	79.50221801	2019/7/11 至 2019/9/1	A市伊景园滨河苑	捆绑车位销售
3	2	2	77.86788212	2019/11/2 至 2020/1/26	A市丽发新城小区	混泥土搅拌站，粉尘和噪音污染严重
4	3	3	66.69311697	2018/11/15 至 2019/10/29	A市大学生	人才购房补贴申请
5	4	4	63.12963588	2019/7/21 至 2019/12/4	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
6	5	5	55.36512875	2019/1/6 至 2019/9/12	A市A3区茶场村五组村民	咨询拆迁规划

图 3.6 最终结果：排名热度前五的问题

4. 答复意见的评价

4.1 综合评价指标体系的构建

4.1.1 答复的相关性分析

答复的相关性与时效性和回应问题有关。

根据附件 4 的‘留言时间’、‘答复时间’可以得出时效性这个属性的值。

$$\text{时效性} = \text{答复时间} - \text{留言时间}$$

根据附件 4 的‘答复意见’、‘留言详情’可以知道答复的意见是否回应了用户反映的问题。

首先我们对‘留言详情’进行中文分词以及关键词提取，如果‘留言详情’中的关键词出现在‘答复意见’中，我们可以认为该‘答复意见’回应了问题。

4.1.2 答复的完整性分析

答复的完整性与专业用语、礼貌用语、字数标准有关。

根据附件 4 的‘答复意见’，可以提取出专业用语（‘反映’、‘收悉’、‘答复如下’）、礼貌用语（‘尊敬’、‘你好’、‘感谢’、‘谢谢’）。通过计算‘答复意见’字数的均值可以得出字数标准。

专业用语是回复意见的标准模板，通过评判句子是否存在专业用语可以知道句子是否完整，且符合标准。礼貌用语会给予留言用户一个尊重、友好的第一印象。礼貌是文明交谈的首要前提，在交谈中要体现出敬爱、友善、得体的气度和风范。‘答复意见’的字数也是评判句子完整性的一个首要标准，当答复意见字数大于均值时可以认为该答复意见完整性较好。

4.1.3 答复的可解释性分析

答复的可解释性与提出意见、工作反馈有关。

根据附件 4 的‘答复意见’，可以知道回复问题需要先解释问题发生的情况以及提出解决此问题的意见或建议，甚至需要先处理好问题再向用户进行反馈。其中‘《》’、‘1、2、3、一、二、三、第一、第二、第三’、‘根据’、‘咨询’、‘致电’、‘电话’可以知道‘答复意见’中是否对问题提出相应的意见或建议。‘已安排’、‘已建议’、‘已要求’可以知道是否有工作人员对问题展开进一步的工作。

4.1.4 答复意见质量评价指标

根据以上的分析建立基于答复相关性的 2 种指标（回应问题、回复时长）；基于答复完整性的 3 种指标（专业用语、礼貌用语、字数标准）；基于答复可解释性的 2 种指标（解释问题、已做工作），建立如图 4.1 所示的答复意见质量评价指标集：

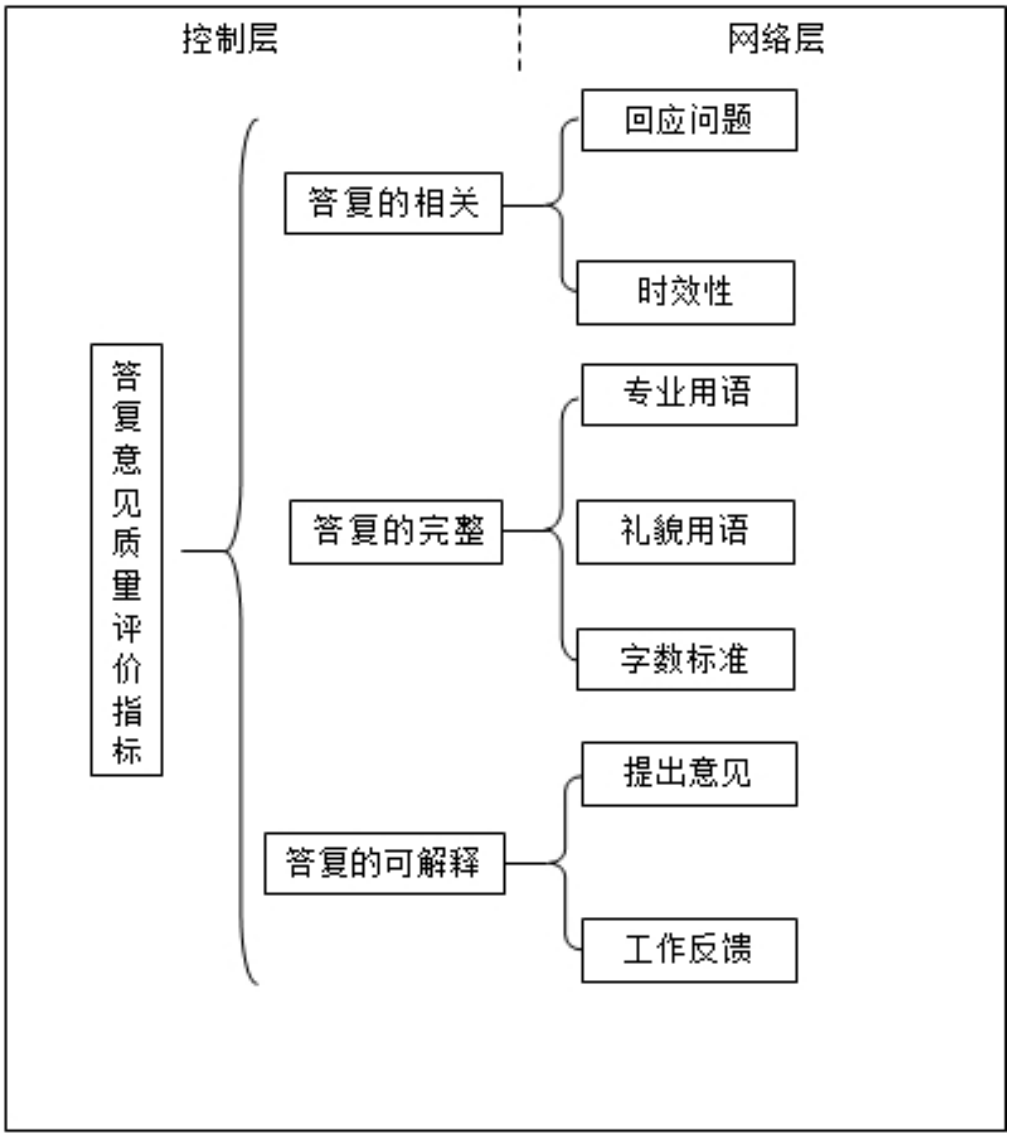

图 4.1 答复意见质量评价指标集

4.2 综合评价模型构建

为了更真实准确地反应出答复意见质量评价指标的相互反馈关系以及本研究中评价指标所体现的分布特点，采用网络层次分析法对答复意见质量评价指标权重赋值进行主观性评价、客观权重赋值。

4.2.1 网络层次分析

网络层次分析（ANP）考虑到一般层次分析各层次结构内部存在相互关系，将元素划分为两大部分，控制层和网络层。控制层包括研究的问题目标和决策准则，认为各准则之间是相互独立的；网络层由控制层准则之下的各特征元素组成，考虑这些元素之间的相互关系，所以每个控制层准则之下都是一个具体的特征元素网络[9]。

我们这里通过网络层次分析法，根据答复的相关性、答复的完整性、答复的可解释性，综合考虑答复意见质量评价体系。主要分为两个步骤：

①根据本题数据，构造 ANP 典型结构；

②构造 ANP 超级矩阵计算特征权重。

（1）构造 ANP 结构

a)给出答复意见质量评价的指标体系，如表 5.1：

答复的相关性（A）	回应问题（A1）
	时效性（A2）
答复的完整性（B）	专业用语（B1）
	礼貌用语（B2）
	字数标准（B3）
答复的可解释性（C）	提出意见（C1）
	工作反馈（C2）

图 4.2 答复意见质量评价的指标体系

b)构建元素间相互关系

控制层各准则的权重可以根据 AHP 方法决定。网络层次分析还需要确定各特征元素之间的相互关系，这部分由专家以一个二联表的形式给出，本题我们只能根据主观判断给出各指标间是否相关，结果如表 5.2（说明：顶部元素为被影响的特征因素，左列元素为引起顶部特征因素的因素，打√的表明左列的因素会影响顶部对应的特征）：

被影响因素 影响因素		Project			相关性(A)		完整性(B)			可解释性(C)	
		P1	P2	P3	A1	A2	B1	B2	B3	C1	C2
Project	P1				√	√	√	√	√	√	√
	P2				√	√	√	√	√	√	√
	P3				√	√	√	√	√	√	√
相关性(A)	A1	√	√	√			√	√	√	√	√
	A2	√	√	√						√	√
完整性(B)	B1	√	√	√				√		√	
	B2	√	√	√							√
	B3	√	√	√	√					√	√
可解释性(C)	C1	√	√	√	√				√		√
	C2	√	√	√	√	√				√	

表 4.1 各特征元素之间的相互关系

（2）对于方案层各元素，给出其相对重要性，构建判断矩阵。

a)准则层元素对目标层的相对重要性判断矩阵，结果如下式：

$$\boldsymbol{A}=\begin{bmatrix}\mathbf{1} & \mathbf{2} & \mathbf{1/3}\\ \mathbf{1/2} & \mathbf{1} & \mathbf{1/4}\\ \mathbf{3} & \mathbf{4} & \mathbf{1}\end{bmatrix}$$

b)方案层相关元素对答复的相关性判断矩阵

$$\boldsymbol{B_1}=\begin{bmatrix}\mathbf{1} & \mathbf{1/3}\\ \mathbf{3} & \mathbf{1}\end{bmatrix}$$

c)方案层相关元素对答复的完整性判断矩阵

$$\boldsymbol{B_2}=\begin{bmatrix}\mathbf{1} & \mathbf{3} & \mathbf{1/3}\\ \mathbf{1/3} & \mathbf{1} & \mathbf{1/4}\\ \mathbf{3} & \mathbf{4} & \mathbf{1}\end{bmatrix}$$

d)方案层相关元素对答复的可解释性判断矩阵

$$\boldsymbol{B_3}=\begin{bmatrix}\mathbf{1} & \mathbf{2}\\ \mathbf{1/2} & \mathbf{1}\end{bmatrix}$$

（2）计算指标的全局权重和综合权重

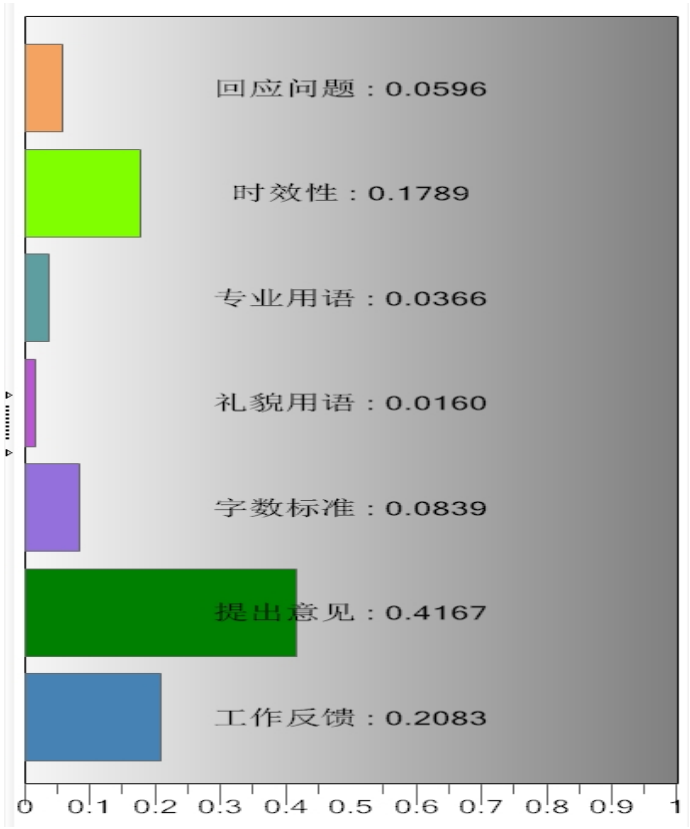

图 4.3 全局权重

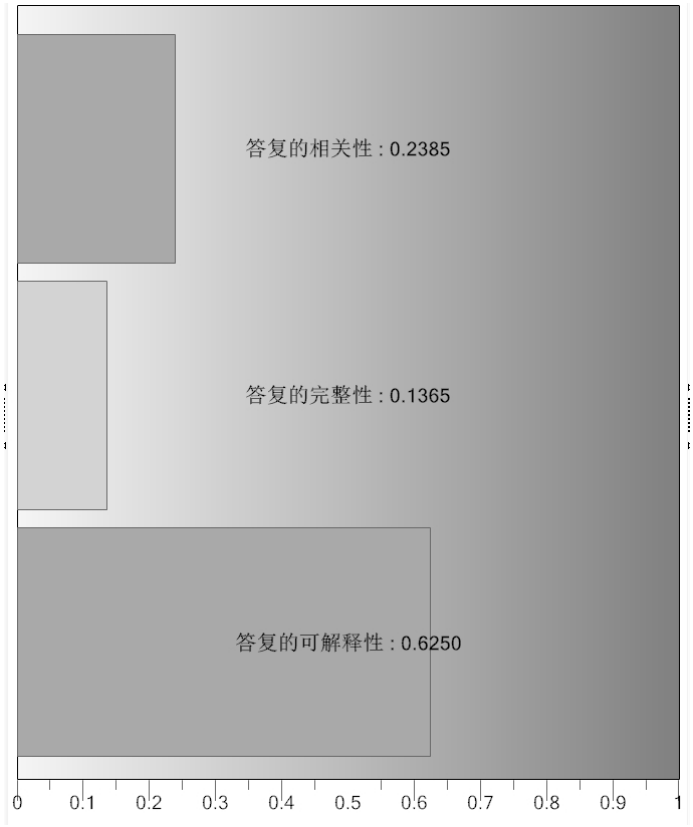

图 4.4 综合权重

4.2.2 结果分析

本章我们根据答复的相关性、答复的完整性、答复的可解释性的数据，提取出回应问题、时效性、专业用语、礼貌用语、字数标准、提出意见、工作反馈 7 个特征，给出了答复意见质量评价指标体系。

根据网络层次分析的结果，由指标的全局权重可以看出，提出意见、工作反馈、时效性是评价指标体系中占权重最大的三个指标，这也与我们的主观感觉相符。而对于相关性、完整性、可解释性这三个方面，根据指标的综合权重结果可以看出，我们给出的这个综合评价指标体系，关注的程度最大的还是可解释性，然后是相关性，最后才是完整性。

5.结论

总结本文，我们基于 fastText 对群众留言内容构建了分类模型，并且比较了 fastText 算法与 word2vec 算法之间的优劣，最终通过 fastText 模型对群众留言内容分类得到 F1-Score=0.916。构造词汇-文本矩阵，进一步运用基于 TF-IDF 对词汇-文本矩阵进行空间语义降维，通过 k-means 聚类算法对热点问题进行了聚类，再通过 Reddit 算法分析热点问题的热度指标。根据答复的相关性、答复的完整性、答复的可解释性，提取出回应问题、时效性、专业用语、礼貌用语、字数标准、提出意见、工作反馈 7 个特征，用网络层次分析法（AHP）构建了答复意见质量评价指标体系。

但是我们最后得到的聚类结果准确度不是特别的好，主要原因是通过层次聚类无法确定准确的 K 值。后期我们会进一步对文本挖掘进行深入探讨。

6.参考文献

[1]古倩. 基于特征向量构建的文本分类方法研究[D].西安理工大学,2019.

[2]余传明,王曼怡,林虹君,朱星宇,黄婷婷,安璐.基于深度学习的词汇表示模型对比研究[J/OL].数据分析与知识发现:1-19[2020-05-07].

[3]王光慈,汪洋.基于 FastText 的短文本分类[J].电子设计工程,2020,28(03):98-101.

[4]毛郁欣,邱智学.基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J].中国信息技术教育,2020(08):99-101.

[5]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[6]王俊丰,贾晓霞,李志强.基于 K-means 算法改进的短文本聚类研究与实现[J].信息技术,2019,43(12):76-80.

[7]黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.

[8]但宇豪,黄继风,杨琳,高海.基于 TF-IDF 与 word2vec 的台词文本分类研究[J].上海师范大学学报(自然科学版),2020,49(01):89-95.

[9]杨俊峰,尹光花.基于 word2vec 和 CNN 的短文本聚类研究[J].信息与电脑(理论版),2019,31(24):20-22.

[10]李海磊,杨文忠,李东昊,温杰彬,钱芸芸.基于特征融合的 K-means 微博话题发现模型[J].电子技术应用,2020,46(04):24-28+33.

[11]李锋刚. 基于词向量和 AP 聚类的短文本主题演化分析[C]. 中国管理现代化研究会、复旦管理学奖励基金会.第十三届（2018）中国管理学年会论文集.中国管理现代化研究会、复旦管理学奖励基金会:中国管理现代化研究会,2018:526-533.

[12]杨玉娟,冯霞,王永利.QH-K:面向新闻文本主题抽取的改进 H-K 聚类算法[J/OL].南京邮电大学学报(自然科学版),2020(01):1-7[2020-05-07].

[13]曹春萍,黄伟.基于用户权威度与热度分配聚类的微博热点发现[J].计算机工程与设计,2020,41(03):664-669.

[14]郭银灵. 基于文本分析的在线评论质量评价模型研究[D].内蒙古大学,2017.

[15]李倩. 甘肃省网络问政制度化完善对策研究[D].西北师范大学,2017.

[16]史文雷,阮平南,徐蕾,魏云凤.创造共享价值的企业战略实施绩效评价——基于改进 BSC 和 DEMATEL-ANP 方法的模糊综合评价模型[J].技术经济与管理研究,2019(11):3-9.

[17]Joulin A , Grave E , Bojanowski P , et al. Bag of Tricks for Efficient Text Classification[J]. 2016.

[18]Qaiser S , Ali R . Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents[J]. International Journal of Computer Applications, 2018, 181(1).

[19]Soucy P , Mineau G W . Beyond TFIDF Weighting for Text Categorization in the Vector Space Model[C]// International Joint Conference on Ijcai. Morgan Kaufmann Publishers Inc. 2005.

[20]Cao X . Improved Online Wilson Score Interval Method for Community Answer Quality Ranking[J]. 2018.

[21]Jing L P , Huang H K , Shi H B . Improved Feature Selection Approach TFIDF in Text Mining[M]// Improved feature selection approach TFIDF in text mining. 2002.

[22]Saaty T . DECISION MAKING - THE ANALYTIC HIERARCHY ANDNETWORK PROCESSES (AHP/ANP)[J]. 系统科学与系统工程学报:英文版, 2004, 13(1):1-35.