基于 NLP 技术智慧政务系统的建立

摘要:近年来，随着互联网的发展，网络问政平台逐渐成为政府了解民意、汇聚民智的重要渠道，各类社情民意相关的文本数量不断攀升，给依靠人工来进行留言分类和热点整理的相关部门带来了极大的工作挑战。因此建立基于自然语言处理技术的智慧政务系统已成为发展的新趋势。

本文主要围绕留言问题分类、热点问题挖掘、答复意见评价三个方面，运用自然语言处理技术建立智慧政务系统，通过合理运用算法，针对每一个问题给出一个分类或者评价的标准。

本文第一章阐明建立该智慧政务系统的挖掘目标；第二章建立解决问题的脉络图；第三章将介绍本文中用到的一些算法。

第四章重点介绍如何建立智慧政务系统，在 4.1 中首先阐明三个问题的解题思路；4.2 中解决留言分类问题，主要分为文本预处理、特征提取、建立学习与知识模式、知识模式评价四个大的步骤，首先将文本集用 jieba 库进行分词处理，删除停用词以减少训练模型、判断分类的时间，构建词袋模型，然后用 TfidfVectorizer 完成向量化，用 TF-IDF 算法计算关键词权重以构造词向量，将每一个文本都用一个词向量表示。随后建立学习与知识模型，按照 4：1 的比例随机划分训练集和测试集，调用 k-means，决策树、朴素贝叶斯法和 XGBoost 四种方法进行模型训练，最后得出四种方法各自的 F-Score 值，以此为评价标准选择出最优的模型完成留言分类问题。4.3 解决热点挖掘问题，首先用正则表达式进行地名的提取，将提取后的结果用 AgglomerativeClustering 模型进行地点聚类，将文本集分为 40 类，将分成的 40 类问题在每一类中计算余弦相似度进行文本相似度比较，相似度>0.15 的文本分为一类，相似度<0.15 的分为另一类，将 40 类文本进一步分为 80 类，将热度值定义为关键词出现次数和时间跨度的表达式，合理赋予权重，通过计算 40 类文本相似度>0.15 的文本，获得各自的热度值，将热度值由大到小进行排序，排名前五的问题即为所要找五个热点问题。4.4 主要建立答复意见评价方案，通过考虑关键词覆盖率、时间间隔、文本相似度、文本情感四个方面，将每一项的量纲压缩在[-1,1]之间，并将每一方面合理赋予权重，构造答复意见质量关于这四个方面的表达式通过计算每一条文本的答复意见质量值进行质量评价。

第五章进行部分实验效果的展示，得出本次实验的结论，用 XGBoost 方法进行留言问题分类，展示五个热点问题，以及部分答复意见质量的评价值。第六章为参考文献。

关键词：自然语言处理　jieba 算法 TF-IDF 法 文本相似度　XGBoost 法

Abstract: In recent years, with the development of the Internet, the online political platform has gradually become an important channel for the government to understand the public opinion and gather people's wisdom. The number of texts related to various social situations and public opinions has been increasing, which has brought great challenges to the relevant departments relying on human to classify messages and sort out hot spots. Therefore, the establishment of intelligent government system based on natural language processing technology has become a new trend.

This paper mainly focuses on three aspects: Message question classification, hot question mining and reply opinion evaluation. It uses natural language processing technology to establish intelligent government system. Through the reasonable use of algorithm, it gives a classification or evaluation standard for each question.

The first chapter clarifies the mining goal of establishing the intelligent government system; the second chapter establishes the venation chart of solving the problem; the third chapter introduces some algorithms used in this paper.

The fourth chapter focuses on how to build a smart government system. In 4.1, it first clarifies the solutions to the three problems; 4.2, to solve the problem of message classification, it is mainly divided into four steps: text preprocessing, feature extraction, building learning and knowledge model, and knowledge model evaluation. First, the text set is segmented with the Jieba database, and the stop words are deleted to reduce the training model and judgment In the time of classification, the word bag model is constructed, then the keyword weight is calculated with TF-IDF algorithm to construct the word vector, and each text is represented by a word vector. Then, the learning and knowledge model is established, and the training set and test set are randomly divided according to the proportion of 4:1. K-means, decision tree, naive Bayes and XGBoost are used to train the model. Finally, the F-score values of each of the four methods are obtained, and the optimal model is selected as the evaluation criteria to complete the message classification. 4.3 to solve the problem of hot spot mining. firstly, we use regular expression to extract the place names, and use the Agglomerative Clustering model to cluster the extracted results. We divide the text set into 40 categories, and calculate cosine similarity in each category to compare the text similarity. The definition of the heat value is the expression of the number of keyword occurrence and time span, and the weight is given reasonably. The heat value of each category is obtained, and the heat value is sorted from large to small. The top five questions are the five hot questions to be found. 4.4 the main purpose of this paper is to establish an evaluation scheme of reply opinions. By considering four aspects of keyword coverage, time interval, text similarity and text emotion, the dimension of each item is compressed between [- 1,1], and each aspect is given a reasonable weight. The expression of reply opinions quality on these four aspects is constructed to evaluate the quality by calculating the reply opinions quality value of each text.

In Chapter 5, we will show the results of some experiments, and get the conclusion of this experiment.

Key words: natural language processing, Jieba algorithm, TF-IDF method, text similarity, XGBoost method

目录

1. 挖掘目标

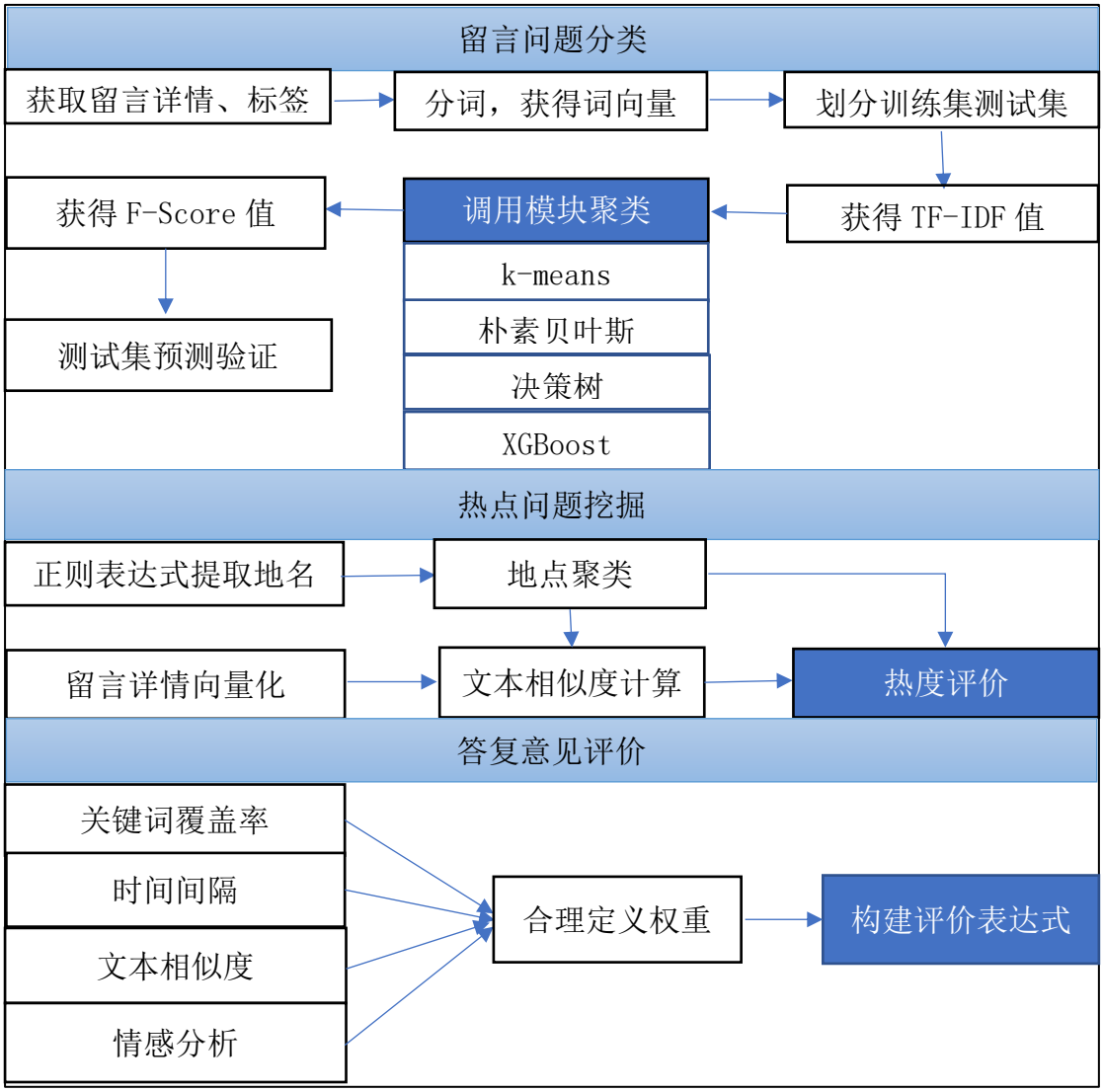
近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。

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。

本文旨在通过比较 k-临近、决策树、朴素贝叶斯以及 XGBoost 这四种方法，选出最优的方法，建立留言划分体系，形成关于留言内容的一级分类标签，从而解决人工处理中存在的工作量大、效率低、差错率高的问题。其次利用给定文本中的时段、地点、人群进行问题的留言归类，用关键词出现的频率以及时间跨度定义热度评价指标，然后为留言热度进行排序，找出某一时间段内群众集中反映的热点问题，帮助相关部门进行针对性地处理，提升服务效率。最后对相关部门的留言答复意见，从关键词覆盖率、时间间隔、文本相似度、文本情感四个方面体现答复的相关性、完整性、可解释性，形成答复意见质量的评价方案，帮助相关部门进行问题处理。

主要围绕留言问题分类、热点问题挖掘、答复意见评价三个方面，运用自然语言处理技术建立智慧政务系统，通过合理运用算法，针对每一个问题给出一个分类或者评价的标准。

2. 解题脉络图


留言问题分类
获取留言详情、标签
分词，获得词向量
划分训练集测试集
获得 F-Score 值
调用模块聚类
获得 TF-IDF 值
k-means
朴素贝叶斯
决策树
XGBoost
测试集预测验证
热点问题挖掘
正则表达式提取地名
地点聚类
留言详情向量化
文本相似度计算
热度评价
答复意见评价
关键词覆盖率
时间间隔
文本相似度
情感分析
合理定义权重
构建评价表达式

3.数据分类介绍

3.1 K-邻近算法

K-邻近算法首先给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最邻近的 K 个实例,观察这 K 个实例大多数属于某个类别，就把该输入实例分类到这个类中。

3.2 决策树

决策树是一种典型分类方法，类似于流程图的树结构，每个内部节点表示在一个属性上的测试,每个分支代表一个测试输出,每个树叶节点存放一个类编号。决策树本质上是通过一系列规则对数据进行分类的过程,分类时给定一个类标号未知的元组 X，测试元组的属性值，跟踪一条由根到叶节点的路径，叶节点存放该元组的类测试。其生成由两个阶段构成—决策树构建与树剪枝。算法种类多，有 CART、ID3、C4.5 等。

3.3 朴素贝叶斯算法

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法,假定给定目标值时属性之间相互条件独立,每一特征同等重要。模型所需估计的参数很少，对缺失数据不太敏感，算法较简单。主要的应用是文本分类，如网络信息过滤、信息检索和信息推荐等。

3.4 XGBoost 算法

XGBoost 是一个优化的分布式梯度增强库，旨在实现高效，灵活和便携。粗略可由模型、参数、目标函数和优化算法构成。模型由 CART 树组成，通过将多个分类准确率较低的树模型组合起来，成为一个准确率很高的预测模型，模型不断迭代，每次迭代就会生成一颗新的树。目标函数由梯度提升算法损失与正则化项组成。

4.具体步骤

4.1 解题思路

4.1.1 问题一

该问题要求建立关于群众留言内容的一级标签分类模型,由于每个文本的类编号是未知的，要学习的类集合也是未知的，所以通过一系列的度量、观察来进行无监督学习，主要分为文本预处理、特征提取、建立学习与知识模式、知识模式评价四个大的步骤[1]，下面是实验思路。

1.获取附件 2 的数据，主题、内容、标签。
2.分词，删除停用词，获得词向量。
3.随机划分训练集测试集，比例为 4：1。
4.获得训练集测试集的 TFIDF 值。

5.调用模块进行训练。

6.用测试集进行预测。

7.获得评价标准 F-Score 值。

由此思路我们进行详细的实验操作，具体操作以及所用算法将在 4.2 中详细介绍。

4.1.2 问题二

该问题要求定义合理的热度评价指标，某一时段内群众集中反映的某一问题可以称为是热点问题，因此时间、地点以及某一问题在单位时间内出现的次数，都是非常重要的，我们的实验思路如下：

1.首先应找出“留言主题”中的特定地点，把特定地点的留言归为一个类，所以应该有一个命名识别的程序对“留言主题”中的地点识别。

2.其次对每个留言主题的“留言详情"做文本相似度比较，对"留言详情"向量化处理。

3.最后进行热度评价:留言主题在-定时间内的频次，或者密度应该是重要的评价指标。

由此思路我们进行详细的实验操作，具体操作以及所用算法将在 4.3 中详细介绍。

4.1.3 问题三

题目要求从答复的相关性、完整性、可解释性等角度对相关部门的答复意见进行质量评价，我们从四个方面来体现。

1.关键词覆盖率：留言详情中的关键词在答复意见中包含了多少。覆盖率越高，则相关性越高。

2.时间间隔：从出现留言详情到出现答复意见的时间间隔是多少。时间间隔越短，说明该问题能够很快的得到处理，相关部门重视度高，也可作为答复意见质量的一个评价标准。

3.文本相似度：观察留言详情和答复意见的余弦相似度，相似度越高，说明该答复紧紧围绕留言详情进行回答，答复的相关性、完整性越高。

4.情感分析：该答复意见的情感多为正向，则说明该答复意见平易近人，能够更好地被市民所接受，也能够更好地解决问题，答复意见的可解释性越高，质量越高。

给每一个方面赋予不同权重以体现各方面对于答复意见质量评价的重要程度，综合以上四个方面，为每一条答复意见计算出一个分值，分值越高，则答复意见质量越高。

由此思路我们进行详细的实验操作，具体操作以及所用算法将在 4.4 中详细介绍。

4.2 群众留言分类

4.2.1 文本预处理

与传统数据库中的结构化数据相比，文档具有有限结构（即使有也只是着重于格式，而非文档内容），甚至没有结构，因此需要对附件 2 中的文本数据进行

相应的标准化预处理；其次由于文本内容使用自然语言进行描述，计算机难以直接处理其语义，所以还需要进行文本数据的信息预处理，抽取代表其特征的元数据或特征词条将每一条文本用一个词向量表示出来。对于文本的信息预处理，我们主要分为词条切分处理以及对已有标签进行编号处理。

首先对已有的 7 类标签进行编号处理，变成 0，1，2，3 等数字，以便之后进行预测和模型学习。

分词：对于词条的切分处理，我们选用 Jieba 分词算法[2]。

首先通过附件 2 获取数据，包括获取留言主题、留言详情和标签。通过对留言详情用 Jieba 库进行词条切分处理，将所有的文本放在一起，去掉停用词之后，构建一个词袋模型。示例如下：

采用 Jieba 库中的精确模式进行分词：

```
ty = open('呆萌的停用词表.txt', encoding='utf-8')
texts = ['\n', '\t', '\xa0', ' ']
labels = []
corpus = []
for word in ty:
    word = word.strip('\n')
    texts.append(word)
for i in range(0, len(content)):
    labels.append(topic[i])
    data = jieba.cut(content[i])
    data_adj = ''
    delete_word = []
    for item in data:
        if item not in texts:
            data_adj += item + ' '
        else:
            delete_word.append(item)
    print(data_adj)
    corpus.append(data_adj)
```

得到结果：

```
M2 县 住房 公积金 中心 取 公积金 难 2011 年购 房 贷款 公积金 说 教师 公积金 太低 贷款 只好 建设银行 贷 款 搞 房屋 装修 提 公
工程 M2 县 安监局 老爷 老百姓 衣食父母 老百姓 办实事 镇 南路 官老爷 负责 质量 镇 南路 改 造 工程质量 差 偷工减料 部门 负责 责
尊敬 杨 市长 您好 M5 市 桥头镇 大范村 村民 九月份 发给您 报告 桥头镇 大范村 邓子山 扶贫 工路 硬化 实属 豆腐渣工程 事情 人来 这
M2 县征 拆 中心 原住 建局 县城 中 河 棚户 改造 项目 工程 潜心 运作 招标 阶段 硬 万立方 土石方 场内 大部分 回填土 造价 502782
M 市邦宁 特色 政府 引进 外企 香港 房子 漏水 强制 交房 想 问 法制 社会 新房 交房 已近 疤 污水 管装 大门口 中建 五局 政府
咨询 M 市 公积金 领导 M 市 公积金 标准 何时能 月份
同大沁园 项目 M 市 职业 技术 学院 团 购房 年初 正式 起动 全线 竣工 下半年 公司 资金 链断裂 无故 拖欠 工程款 农民工 工资 数千万
西地省 M 市 高新技术 经济 开发区 新村 原西 地省 M 市 M5 市 石马 山镇 木灵村 2008 世界 五百强 企业 落户 村 我市 村划 经济 开发
我国 经济 发展 惠民 政策 层出不穷 确确实实 低收入者 实惠 侧面 我国 惠民 工程 廉租房 公 租房 实行 租 购并 举 政策 解决 低收入者
尊敬 杨 市长 碧桂园 豪园 购买 住房 申请 公积金 贷款 告知 办理 询问 得知 碧桂园 豪园 出尔反尔 公积金 缴纳 对象 抗议 市政府 施加
M1 区 黄 泥塘 办事处 东来 村关 塘 组涟 钢汉 大三 六零 移民 基地 黄 泥塘 建筑 公司 承建 质量 属 豆腐渣工程 钢筋 裸露 外面 水泥
尊敬 上级领导 您好 M1 区景屏 街 业主 因景屏 街 老街区 门面 不带 卫生间 景屏 街 位置 一间 公共 卫生间 供 门面 店主 顾客 如厕 现
国家 政策 几年 不知 困难 家庭 住 房 过上 温饱 生活 依然 困难 没 享受 国家 大好 政策 市 房产局 公 租房 门面 绝大多数 关系户 租
M1 区楚阳 街 香山 红叶 住宅小区 栋 新房 交房 阳台 横梁 发现 贯穿性 裂纹 开发商 建房 施工 电梯 螺栓 安装 孔 打断 根 钢筋 主筋
几年 前 买 M 市耳九 置业 有限公司 世家 房产 房子 漏雨 下雨 墙 渗水 楼上 天天 漏水 找 物业 未 办两证 钱 早 M 市耳九 置业 代收
市长 住房 公积金 中心 贷过 款 需去 贷款 住房 为啥 强制 交 住房 公积金 上辈子 欠 公积金 中心 钱 这辈子 尝 请 领导 取消 不合理
偶们 M 市越 建越 鳞次栉比 新房子 新 高楼 50% 暂时 人去 住 空房子 大型 青少年 活动中心 每到 双休日 寒暑假 逛逛 公园 百去 不厌
尊敬 领导 您们好 M3 县 富康 家园 业主 开发商 M3 县星燎 房地产 开发 有限公司 建成 交房 至今已有 四年 未 应有 不动产 登记证 现特
M3 县荣 花乡 国家 湿地 公园 核心 景区 荣花村 华天 村 龙湾 村 鹊桥 村 公厕 服务 好不好 旅游 公厕 指标 西地省 各大 景区 加大 旅
```

去停用词：目前停用词字典有 2000 个左右，停用词主要包括一些副词、形容词及其一些连接词。经过删去停用词，可减少训练模型、判断分类的时间。
由此建立词袋模型。

4.2.2 特征提取

基于词袋模型的特征表示，对于特征提取，我们采用关键词权重计算算法：TF-IDF[3]。

Jieba 库进行分词，建立词袋模型后，由于要考虑词语出现的频率和词语的重要性，所以采取 TF-IDF 算法，对文本中的每个分好的词语赋予权重，即 TF-IDF 值，获得词向量，将每一个文本都用一个词向量表示出来。

在进行这一步前，首先用 TfidfVectorizer 完成向量化与 TF-IDF 预处理。代码如下：

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf2 = TfidfVectorizer()
re = tfidf2.fit_transform(corpus)
print re
```

经过这一步，完成向量化，TF-IDF 与标准化。

然后采用 scikit-learn 包进行 TF-IDF 分词权重计算，将已经利用 jieba 库分好的词放入代码中得出结果，程序如下：

```
    vectorizer=CountVectorizer()#该类会将文本中的词语转换为词频矩阵，矩阵元素a[i][j]
    transformer=TfidfTransformer()#该类会统计每个词语的tf-idf权值
    tfidf=transformer.fit_transform(vectorizer.fit_transform(corpus))#第一个fit_tr
    word=vectorizer.get_feature_names()#获取词袋模型中的所有词语
    weight=tfidf.toarray()#将tf-idf矩阵抽取出来，元素a[i][j]表示j词在i类文本中的tf-id
```

由此获得词向量，将每一个文本都表示成一个词向量。

4.2.3 建立学习与知识模型

划分训练集和测试集：在附件 2 中，已有数据已被划分为 7 类一级标签，所以将这些已有标签的文本按照 4：1 比例，随机划分训练集和测试集，进行模型的训练与测试。训练集用于训练模型，测试集用于测试训练结果。并计算出训练集的 TF-IDF 值。

调用模块进行训练：将训练集的 TF-IDF 值代入模块进行模型的拟合和训练。在调用模块这里，我们用了四种模块来进行拟合，看哪一个的拟合效果最好。这四种模块分别是 k-means，决策树、朴素贝叶斯法和 XGBoost 方法，效果如下：

k-means:

GaussianNB:

	precision	recall	f1-score	support
0	0.63	0.72	0.67	427
1	0.82	0.71	0.76	186
2	0.77	0.23	0.35	106
3	0.59	0.80	0.68	279
4	0.73	0.71	0.72	417
5	0.64	0.62	0.63	244
6	0.78	0.60	0.67	183
accuracy			0.68	1842
macro avg	0.71	0.63	0.64	1842
weighted avg	0.69	0.68	0.67	1842

可知，调用 k-means 方法的 f-score 为 0.68。

朴素贝叶斯法：

KNeighboursClassifier:

	precision	recall	f1-score	support
0	0.63	0.72	0.67	427
1	0.82	0.71	0.76	186
2	0.77	0.23	0.35	106
3	0.59	0.80	0.68	279
4	0.73	0.71	0.72	417
5	0.64	0.62	0.63	244
6	0.78	0.60	0.67	183
accuracy			0.68	1842
macro avg	0.71	0.63	0.64	1842
weighted avg	0.69	0.68	0.67	1842

可知，朴素贝叶斯法的 f-score 为 0.68。

决策树法：

DecisionTreeClassifier:

	precision	recall	f1-score	support
0	0.71	0.75	0.73	427
1	0.75	0.75	0.75	186
2	0.66	0.59	0.62	106
3	0.85	0.89	0.87	279
4	0.83	0.79	0.81	417
5	0.75	0.73	0.74	244
6	0.72	0.70	0.71	183
accuracy			0.77	1842
macro avg	0.75	0.75	0.75	1842
weighted avg	0.77	0.77	0.77	1842

可知，决策树法的 f-score 为 0.77。

XGBoost:

	precision	recall	f1-score	support
0	0.79	0.87	0.83	406
1	0.91	0.86	0.88	176
2	0.85	0.79	0.82	111
3	0.94	0.92	0.93	322
4	0.90	0.93	0.92	408
5	0.88	0.80	0.84	253
6	0.93	0.86	0.89	167
accuracy			0.88	1843
macro avg	0.89	0.86	0.87	1843
weighted avg	0.88	0.88	0.88	1843

可知，XGBoost 的 f-score 为 0.88。

由四种方法比较可知， 所以我们选择调用 XGBoost 模块来进行分类。

4.2.4 模型评价

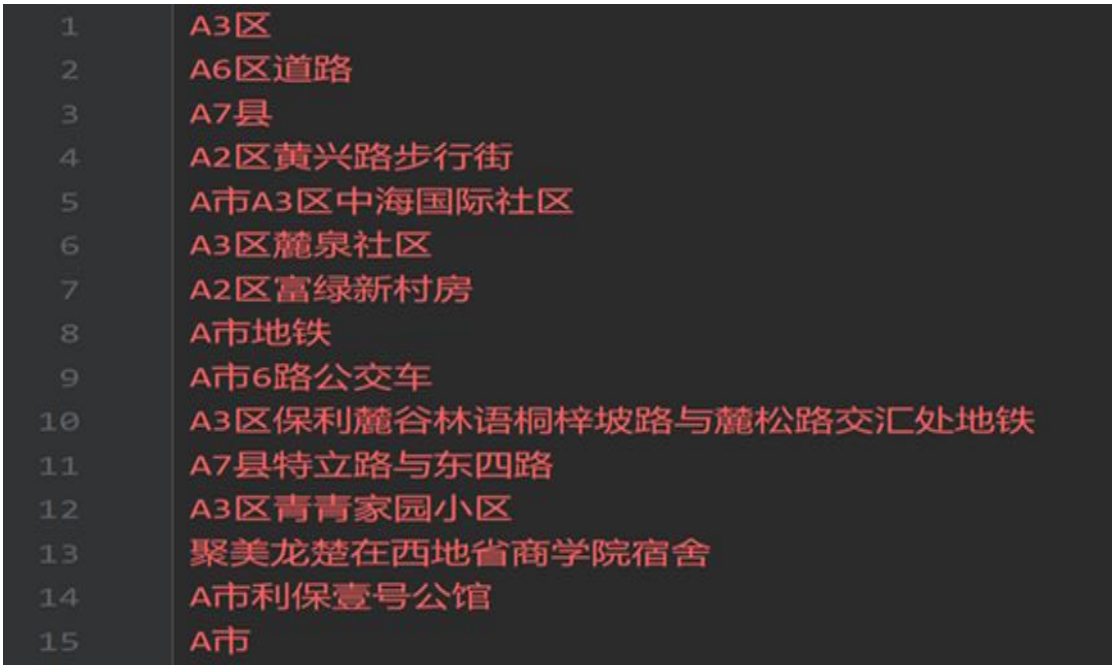
我们用已经分出的测试集进行预测，并检验模型的学习能力。用 F-Score 进行模型评价可知，用 XGBoost 进行分类的分类效果最好，因此采用 XGBoost 分类方法。

由此，我们得到结论：用 XGBoost 进行分类，获得的评价标准 F-Score 为 0.88。

4.3 热点问题挖掘

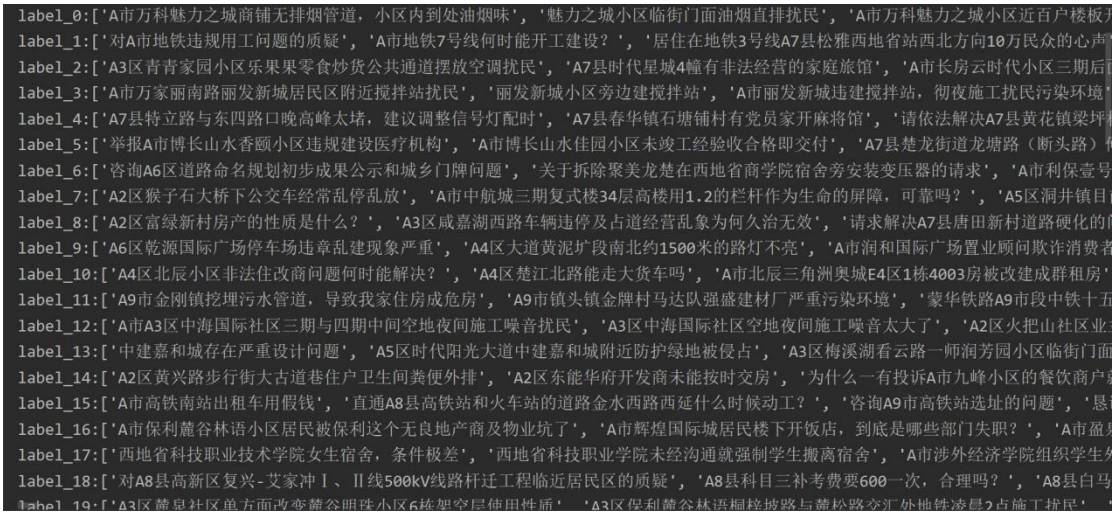
4.3.1 地名提取与分类

首先用正则表达式进行地名的提取，结果如下：



将提取后的结果用 AgglomerativeClustering 模型进行聚类，使地点相同的文本归在一起。结果如下：

经过该方法，我们将附件 3 中的所有文本分为了 40 类。



4.3.2 文本相似度比较

运用第一问中的方法，进行文本的向量化处理，使每一个文本用一个词向量表示出来，以便后面进行文本的相似度比较。

将我们前面按照地点分出的 40 类文本中，采用余弦相似度法，在每一类中进行文本的相似度比较，提取出相似度高的文本。我们将相似指标定为 0.15，即两个词向量的余弦如果达到 0.15，则归为一类，如果没有达到 0.15，则归为另一类，这样我们将 40 类文本中文本相似度没有达到 0.15 的文本剔除。

4.3.3 热度评价

我们将 40 类文本相似度大于 0.15 的文本进行热度评价，对热度评价建立表达式，计算每一类文本的热度值然后进行由大到小进行排序，取前五名，即我们所找出的热度最高的五个问题。

计算每一类问题的热度值公式如下：

$$热度值 = 0.65 * x1 + 0.35 * x2$$

其中 x1 为关键词出现次数，x2 为时间跨度。

我们认为，关键词在单位时间内出现的次数，可以很大程度上反映一个问题的热度，在相同时间内，关键词出现的频率越高，说明问题越热，所以我们赋予 0.65 的权重。

对于该问题中关键词从第一次出现最后一次出现，也一定程度上反映了问题的热度，如果该问题的时间跨度非常短，而出现频率又非常高，说明该问题是热点；如果该问题的时间跨度非常大，但是关键词出现的频数并不高，说明该问题只是偶尔有人反应而没有引起重视，若此时时间跨度的权重很大的话，整体的热度值也会变大，导致该问题的热度值虚高，所以我们给时间跨度赋予 0.35 的权重，以减小时间跨度大对问题热度带来的冲击。

经过运行程序，得出结果：

```
0 : 0.4350141031149674
1 : 0.6660822973919222
2 : 0.7356233211262176
3 : 0.6396967359909457
4 : 0.7324001956517046
5 : 0.600011369176586
6 : 0.7168874744559152
7 : 0.571412442181793
8 : 0.45881405944913345
9 : 0.43833568633233505
10 : 0.665916655870876
11 : 0.40769130745776316
12 : 0.5059244118528443
13 : 0.4047780639154618
14 : 0.6582223249471728
15 : 0.39315478158707284
16 : 0.4952819352073205
17 : 0.5143472487182524
18 : 0.41857415519622665
19 : 0.7348032550451488
20 : 0.44541961674817837
21 : 0.6343346446560617
22 : 0.36743094689539263
23 : 0.48733354997045175
24 : 0.5170662395205364
25 : 0.4163268239643495
26 : 0.6924027194258336
27 : 0.4237079216878293
28 : 0.555160841822786
29 : 0.4377396632206117
30 : 0.4669553928329351
31 : 0.46793406121487746
32 : 0.36823403974476626
33 : 0.4192191967296528
34 : 0.422263037937364
35 : 0.419136607451053
36 : 0.3651650486662768
37 : 0.3955095464880862
38 : 0.37190235029212626
39 : 0.37980676395922763
```

经排序得出结论：

A2 去丽发新城、伊景园滨河苑、楚龙街道、A 市万科魅力之城、经济学院，这五个地点中的热度最高的五个问题为热点问题。

4.4 答复意见评价

对于答复意见的评价，建立以下表达式进行答复意见质量计算：

$$答复意见质量 = 0.3 * 关键词覆盖率 + 0.05 * 时间间隔 + 0.5 * 文本相似度 + 0.15 * 文本情感$$

4.4.1 关键词覆盖率

这一步分为两小步：即：确定关键词覆盖率、确定关键词覆盖率所占权重。

确定关键词覆盖率：如果在留言详情中出现了 n 个关键词，在答复意见里面出现了 m 个（m<n），则该文本的关键词覆盖率为 m/n,这样处理即可保证答复意见的关键词覆盖率在 0-1 之间，便于之后表达式中统一量纲进行计算。

关键词调用 jieba.analyse.extract_tags 提取。

确定关键词覆盖率所占比重：关键词覆盖率体现了答复意见的相关性和完整性，我们认为在此公式中，关键词覆盖率的重要性仅次于文本相似度，因此给关键词覆盖率赋予 0.3 的权重。

4.4.2 时间间隔

这里的时间间隔指的是从一条留言详情出现时间到收到答复意见的间隔。间隔时间越短，说明相关部门处理问题的效率越高，能够及时解决群众反映的问题。

设定四个时间间隔，并对四个时间间隔打分如下：

时间间隔	0-7 天	8-31 天	32-100 天	>100 天
所获分值	0.8	0.4	0.2	0.1

即如果在留言详情出现后的七天内回复，则该项获得 0.8 的分值，若在留言详情出现的 100 天以后再回复，则该答复意见在该项所获分值为 0.1。

但是也不排除相关部门的回复意见与留言详情没有太大的相关性，不能够正确解决群众反映的问题。为了避免答复意见只注重量而不注重质的问题，我们将该项的权重设置为 0.05，这样既能够保证答复意见的时间间隔在该表达式中起作用，又能避免上述问题的出现。

4.4.3 文本相似度

对于答复意见质量的评价，应该重点考虑留言详情以及答复意见的文本相似度，文本相似度越高，说明答复意见对该问题的描述越详细，答复意见的相关性、完整性、可解释性越高。

首先将留言详情与答复意见均通过 jieba 分词处理构建词袋模型，然后计算词袋模型的每个词语出现在留言详情与答复意见中各自的个数获得词向量，计算出每一条文本中留言详情和答复意见的余弦相似度。

$$similarity(V_i, U) = \cos(V_i, U) = \frac{V_i \times U}{V_i \circ U}$$

$$= \frac{\sum_{k=1}^{m} w_{ik} \circ w_i}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \sum_{k=1}^{m} w_k^2}}$$

4.4.4 答复意见情感

通过对答复意见的情感分析，能够提取出问题回复者的态度。如果在文本中出现了正向的情感词，例如“请”、“您”、“谢谢”等词语，并判断在这个词是否存在前一个词，如果有的话判断是否为否定词或者程度副词;如果出现否定的消极词语，就给这一项加 1 分，如果出现的是程度副词的消极情感词，则减 2 分，如果只有否定词，则减 1 分；同理如果出现否定的积极情感词则减 1 分，出现程度副词的积极情感词加 2 分，如果只有积极词语则加 1 分；否定词减 0.5 分。这样计算出每一条答复意见的情感值，然后将每一个情感值除以所有文本的情感值中绝对值最大的那个，以保证将答复意见的情感值压缩在[-1,1]内。

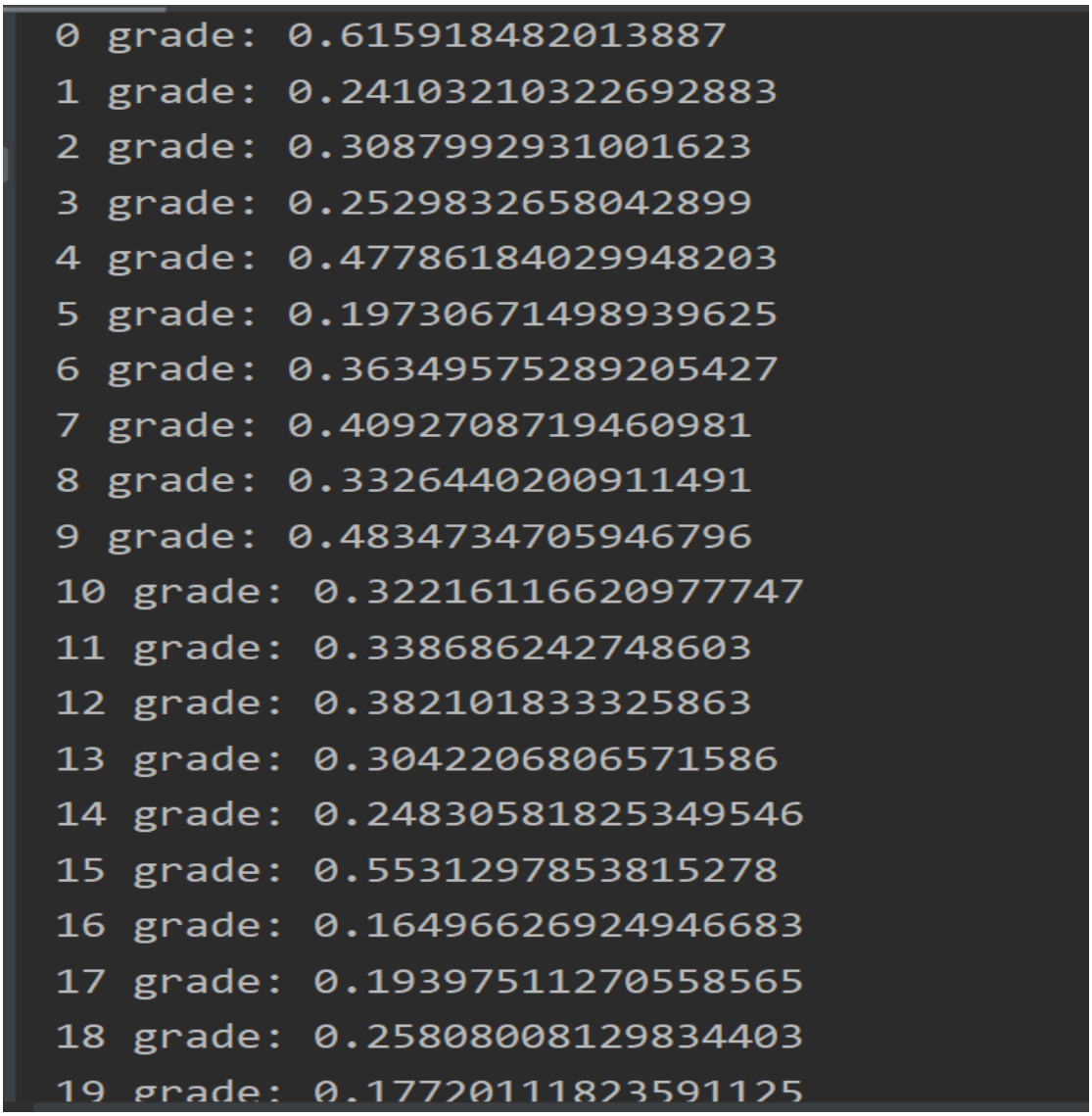
并且规定文本情感的权重为 0.15。

4.4.5 计算答复意见质量值

答复意见质量计算公式：

$$答复意见质量 = 0.3 * 关键词覆盖率 + 0.05 * 时间间隔 + 0.5 * 文本相似度 + 0.15 * 文本情感$$

经过程序运算，我们得到每一条答复意见的质量值如下：



答复意见质量值越大，则证明答复意见质量越高。部分结果如下：

5. 效果展示

5.1 群众留言分类

方法	k-means	朴素贝叶斯	决策树	XGBoost
F-Score	0.68	0.68	0.77	0.88

因此，最终选用 XGBoost 方法进行留言问题的一级分类。

5.2 热点问题挖掘

5.2.1 热点问题表

A	B	C	D	E	F	G	H
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	0.784	2019/4/10至2020/1/26	A市A2区丽发新城居民	A市A2区丽发新城小区附近存在非法搅拌站，噪音、灰尘、污水等扰乱居民生活。		
2	2	0.736	2019/7/7至2019/9/1	A市伊景园滨河苑业主	A市伊景园滨河苑捆绑销售无产权车位，定向限价商品房违规涨价		
3	3	0.716	2019/1/7至2019/12/11	A7县楚龙街道人民	A7县楚龙街道长期交通拥堵，人行道围挡影响出行，拆迁问题未得到解决，施工扰民		
4	4	0.636	2019/1/8至2019/12/4	A市万科魅力之城业主	A市万科魅力之城小区近百户楼板墙面开裂		
5	5	0.628	2017/6/8至2019/11/27	A市经济学院学生	A市涉外经济学院强制学生实习		

经过程序运算，得到前五个热点问题为：

1. A 市 A2 区丽发新城小区附近存在非法搅拌站，噪音、灰尘、污水等扰乱居民生活。
2. A 市伊景园滨河苑捆绑销售无产权车位，定向限价商品房违规涨价。
3. A7 县楚龙街道长期交通拥堵，人行道围挡影响出行，拆迁问题未得到解决，施工扰民。
4. A 市万科魅力之城小区近百户楼板墙面开裂。
5. A 市涉外经济学院强制学生实习。

5.2.2 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	留言详情	反对数	点赞数
1	188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019/11/19 18:07	处建搅拌站，运渣车	0	1
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/11/13 11:20	方建搅拌站。可想而	0	0
1	190108	A909240	丽发新城小区旁边建搅拌站	2019/12/21 15:11	影响几千名学生的健康	0	1
1	190523	A00072847	A市丽发新城违建搅拌站，彻夜施工扰民污染环境	2019/12/26 13:55	量；3、搅拌站几百米	0	0
1	191943	A00038563	A市A2区丽发新城道路坑坑洼洼	2019/7/3 12:03	道路坑坑洼洼，下雨	0	1
1	193091	A00097965	A市富绿物业丽发新城强行断业主家水	2019/6/19 23:28	供地摊上买的收据，	0	242
1	199379	A00092242	A2区丽发新城附近修建搅拌厂，严重污染环境	2019/11/25 10:17	还得了疾病住院，该	0	0
1	203393	A00053065	A市丽发新城小区侧面建设混泥土搅拌站，粉尘和	2019/11/19 14:51	的粉尘，严重影响居	0	2
1	208714	A00042015	A2区丽发新城附近修建搅拌站，污染环境，影响	2020/1/2 0:00	量和声环境质量急剧	0	4
1	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019/12/10 12:34	觉拌站。该搅拌站的	0	0
1	213930	A909218	A2区丽发新城附近违规乱建混凝土搅拌站谁来监	2019/12/27 23:34	强烈呼吁政府和有关	0	0
1	214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和污染环境	2020/1/25 9:07	尺吵，烦死了不仅吵还	0	0
1	215563	A909231	A2区丽发新城小区旁边的搅拌厂是否合法经营	2019/12/6 12:21	了噪音和灰尘。这给	0	0
1	215842	A909210	A2区丽发新城小区附近太吵了	2020/1/26 19:47	觉拌厂是怎么回事！	0	0
1	216824	A909214	搅拌站大量加工砂石料噪音污水影响丽发新城小	2019/12/25 12:15	这些严重扰民的噪音	0	0
1	217700	A909239	丽发新城小区旁的搅拌站严重影响生活	2019/12/21 2:33	为搬迁到丽发新城小区	0	1
1	219174	A00081998	A2区丽发新城小区内垃圾站散发严重臭味	2019/7/3 23:27	候吃饭睡觉都能闻到	0	3
1	222831	A909228	噪音、灰尘污染的A2区丽发新城附近环保部门不	2019/12/22 10:23	能修改产生大量灰尘	0	0
1	225217	A909223	A2区丽发新城附近修建搅拌厂严重影响睡眠	2019/11/15 9:17	一搅拌站,每天尘土飞	0	0
1	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/12/2 11:20	距离上次投诉已经过	0	0
1	233158	A909242	丽发新城小区旁建搅拌厂严重扰民！	2019/12/5 8:46	在家我还能忍忍，但	0	0

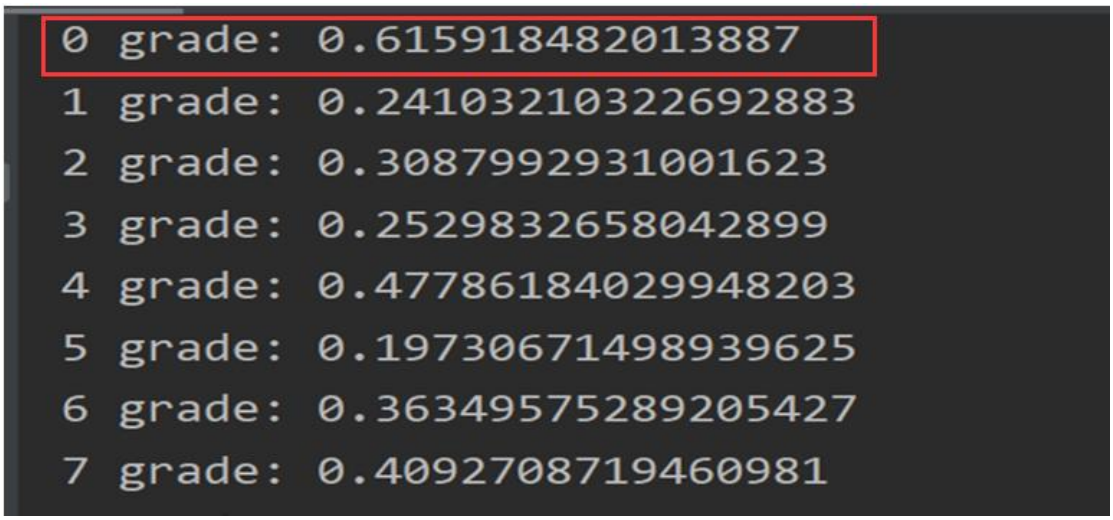
由于篇幅原因，只截取热度最高的问题留言明细表如上。

5.3 答复意见评价

评价方案为：

$$答复意见质量 = 0.3 * 关键词覆盖率 + 0.05 * 时间间隔 + 0.5 * 文本相似度 + 0.15 * 文本情感$$

答复意见质量值如下：



从前八条的答复意见中，可以看出第一条的答复意见质量较高，评价值达到了 0.616，而第六条的答复意见质量仅为 0.197。

6. 参考文献

[1] 薛为民,陆玉昌.文本挖掘技术研究[J].北京联合大学学报(自然科学版),2005(04):59-63.

[2] 韩明宇. Jieba 分词简介. https://blog.csdn.net/qq_37098526/article/details/88877798，2019.03.28

[3] 使用 scikit-learn 计算词语权重. https://blog.csdn.net/levy_cui/article/details/77962768，2017.09.13

[4] 王继成,潘金贵,张福炎.Web 文本挖掘技术研究[J].计算机研究与发展,2000(05):513-520.

[5] 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究[J]. 中文信息学报, 2007(05):73-79.

[6] 时志芳. 移动投诉信息中热点问题的自动发现与分析[D]. 北京邮电大学, 2013.

[7] 吴柳, 程恺, 胡琪. 基于文本挖掘的论坛热点问题时变分析[J]. 软件, 2017, 38(04):47-51.