

“智慧政务”中的文本挖掘应用

摘要:

在智慧城市的规划建设中,“智慧政务”无疑是其中的一个重点领域,而电子政务是智慧政务的最直观体现。因此,挖掘互联网公开来源的群众问政留言记录及其相关答复,对于政府工作效率和针对性的提高具有重要的意义。本文通过比较不同的机器学习方法和优化模型,分步完成了群众留言分类、热点问题挖掘、和答复意见评价三个任务。

第一个阶段,我们先对文本数据进行去噪处理,除去重复留言和空留言。除此之外,将停用词以及无意义的符号去除,利用 python 中的 **jieba** 库使文本切割为词语。通过使用**词频-逆文件频率(TF-IDF)模型**计算词语特征值,接着利用**卡方检验**找出每个分类中关联度最大的两个词语和两个词语对。使用**朴素贝叶斯分类器**将留言进行分类。最后,通过**f1-score 模型**评估方法,得到最终结果准确率为 85%。

第二个阶段,同样地对文本数据先预处理,再构建**词袋模型(BOW)**从文本中提取特征,利用**TF-IDF**进行权值转换。接着我们需要度量文本相似度。为了方便聚类分析,我们先将余弦值做归一化处理,再做**层次聚类**。最后依据每个热点问题的留言数量分别计算出每个问题的热度值。

第三个阶段,为确定影响答复意见质量因素的比例,我们决定构造判断矩阵,使用的是萨蒂提出的**1-9 标度法**。该方法得到的结果通过了一致性检验,于是我们对上述矩阵做**归一化处理**,最终获得答复的**相关性、完整性、可解释性和礼貌性**的权重向量为(0.471, 0.179, 0.253, 0.096)。相关性,采用余弦相似度算法;完整性,以回复的篇幅和格式的规范性为依据;可解释性,以相似词语的关联度和相似性为依据;礼貌性,构建礼貌用词的字典,计算回复中以上词语的使用程度。以 100 分为满分给每个回复打分以衡量回复意见的质量。

关键词: 机器分类与评价 自然语言处理 TF-IDF 模型 关联度 聚类

Abstract:

In the planning and construction of smart city, "smart government affairs" is undoubtedly one of the key areas, and e-government is the most intuitive embodiment of smart government affairs. Therefore, it is of great significance for the government to dig into the records of people's political messages from open sources on the Internet and to improve the government's work efficiency and pertinence. In this paper, by comparing different machine learning methods and optimization models, three tasks are completed step by step, namely, classification of public comments, mining of hot issues, and evaluation of responses to comments.

In the first stage, we first de-noising the text data to remove repeated messages and empty messages. In addition, the stop words and meaningless symbols are removed, and the text is cut into words using the **jieba** library in python. The word eigenvalues were calculated using the word frequency-inverse file frequency (**tf-idf**) **model**, and then the chi-square test was used to find the two words and the two word pairs with the highest correlation degree in each classification. The **naive bayes classifier** is used to classify the messages. Finally, by **f1-score model evaluation method**, the accuracy of the final result is 85%.

In the second stage, the text data is similarly preprocessed, and then the word **bag model (BOW)** is constructed to extract features from the text, and tf-idf is used for weight conversion. Then we need to measure the textual similarity. In order to facilitate cluster analysis, we first normalize the cosine value and then do hierarchical clustering. Finally, the heat value of each hot issue is calculated according to the number of comments.

In the third stage, in order to determine the proportion of factors affecting the quality of replies, we decided to construct a judgment matrix, using the **1-9 scale method** proposed by satie. The results obtained by this method passed the consistency test, so we normalized the above matrix and finally obtained the weight vectors of relevance, completeness, interpretability and politeness of the replies (0.471, 0.179, 0.253, 0.096). Correlation, using cosine similarity algorithm; Completeness, based on the normative length and format of the reply; Interpretability is based on the correlation degree and similarity of similar words. Courtesy, build a dictionary of polite words, and calculate the usage of the above words in the reply. Each response was scored on a scale of 100 to measure the quality of the feedback.

Key words: machine classification and evaluation natural language processing tf-idf model correlation degree clustering

目录

引言.....	6
1 群众留言分类.....	7
1.1 数据分析.....	7
1.2 数据预处理.....	8
1.3 计算 TF-IDF 的特征值 ^[1]	10
1.3.1 TF-IDF 的原理.....	10
1.3.2 TF-IDF 的具体实现.....	11
1.4 朴素贝叶斯分类器 ^[3]	13
1.4.1 朴素贝叶斯分类器的原理.....	13
1.4.2 朴素贝叶斯分类器的具体实现.....	14
1.5 模型选择.....	14
1.6 模型评估.....	15
2 热点问题挖掘.....	16
2.1 了解数据与建模流程.....	16
2.1.1 数据分析.....	17
2.1.2 建模流程.....	17
2.2 “留言主题”数据预处理.....	18
2.2.1 中文分词与去除停用词.....	18
2.2.2 构建词袋模型.....	18
2.2.3 TF-IDF 权值转换.....	19
2.3 “留言主题”建模与应用.....	19

2.3.1 余弦相似度计算.....	19
2.3.2 层次聚类.....	20
2.4 “留言主题”的热点问题.....	21
2.4.1 热度排序.....	21
2.4.2 热度问题.....	22
3 答复意见的评价.....	23
3.1 确定影响答复意见质量因素的比例.....	23
3.1.1 构造判断矩阵知识.....	23
3.1.2 构造判断矩阵.....	24
3.1.3 一致性检验.....	25
3.1.4 归一化处理.....	25
3.2 礼貌性的量化.....	26
3.2.1 数据预处理.....	26
3.2.2 词频统计.....	27
3.2.3 根据词频绘制词云.....	27
3.2.4 得到礼貌性的词典.....	28
3.2.5 得到礼貌性词语的个数.....	28
3.2.6 得到礼貌性的评分.....	28
3.3 相关性的量化.....	29
3.3.1 对留言详情和留言回复进行相似度分析.....	29
3.3.2 对相似度的量化.....	29
3.4 可解释性的量化.....	30

3.4.1 量化分析思路.....	30
3.4.2 将可解释性量化.....	30
3.5 完整性的量化.....	30
3.5.1 量化分析思路.....	30
3.5.2 将完整性量化.....	30
3.6 整体评分结果的量化.....	31
4 总结与展望.....	32
参考文献.....	33

引言

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。在一系列的文本数据中对数据进行分类，并且找出亟待解决的热点问题，对政府的答复意见进行评价就成为了主要的问题。

自然语言处理技术为上述问题提供了解决方案。近年来，自然语言处理技术作为人工智能的一个重要领域得到了飞速发展，构建基于自然语言处理技术的智能分类模型，以及对文本聚类模型从而挖掘出热点问题，并将答复意见进行评价。

基于对智慧政务系统的理解和认识，本文将立足于以上背景和问题，构建基于自然语言处理技术及文本挖掘的分类和聚类模型，完成基于特定文本的分类模型，并且完成基于层次聚类模型挖掘出热点问题，并构建热度评价指标。在完成对题目所给问题集的数据分析以及预处理工作后，该模型与其他主流方案相比，在准确率以及泛化能力上都表现出优越的效果，有利于推动智慧政务系统的建设。

1 群众留言分类

1.1 数据分析

附件 2 所给的数据中，共包含了 7 个不同的一级标签(城乡建设, 劳动和社会保障, 教育问题, 商贸旅游, 环境保护, 卫生计生, 交通运输)，9000 余条群众留言信息。

	label	count
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

图 1.1.1

具体统计情况如右图：

我们操作的主要对象有两个字段，其中 theme 字段表示群众留言主题，label 字段表示相应的一级标签。这些数据内容基本上都由中文表示。下图为随机抽取的 10 个样本示例：

	theme	label
1068	K市龙腾大厦开发商隐瞒房产性质，向购房者强制征收不合理税率	城乡建设
5917	咨询产假工资发放问题	劳动和社会保障
4238	A8县高中老师的农村教师补助为何迟迟不到位？	教育文体
4394	请A7县教育局更换县属中学不符合国家标准的体育器械	教育文体
5461	C4市谭市镇兽医何时能领着退休金？	劳动和社会保障
5872	关于公务员身份认定及工资改革的困惑	劳动和社会保障
9181	M1区妇幼保健院结核门诊不开门	卫生计生
1672	A6区城乡规划局随意批准变更规划	城乡建设
8733	L市中医院私自采购药品	卫生计生
1765	廉租房质量大廉租了	城乡建设

图 1.1.2

将留言数据中的空值和重复的部分进行清洗，得到留言主题的总体情况。

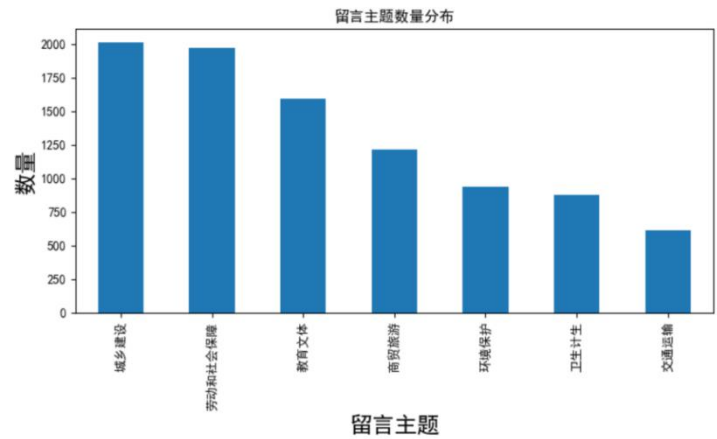


图 1.1.3

由上图，发现不同的标签下主题数量的分布不够均匀，对后面分类的模型构建会产生一定的影响，在接下来的工作中需针对此情况进行一定的调整。

1.2 数据预处理

首先，将 label 类转换为 id，便于后续的模式建立与训练。即按顺序将“城乡建设”到“卫生计生”7 个标签标记为 0, 1, 2, 3, 4, 5, 6。

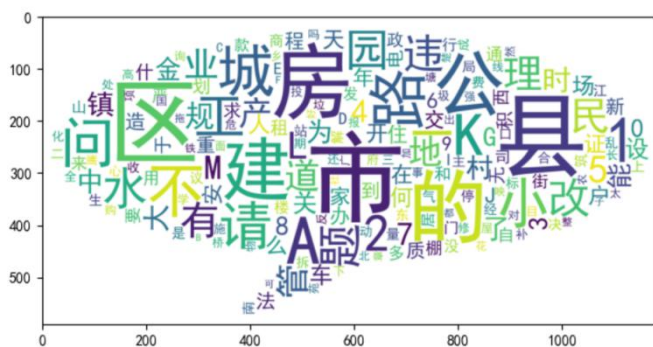
其次，因为我们的数据均为中文文本，其特点是词与词之间没有明显的界限，从文本中提取所需的关键词时，需要先删除文本中的标点符号，特殊符号，以及一些无意义的常用词 (stopword)。这些词和符号对分析预测文本的内容没有任何帮助，反而会增加计算的复杂度和增加系统开销，甚至影响预测结果的准确度，因此在使用文本数据之前必须把他们清理干净。处理后的词语示例如下：

3855	C4 市 东山 学校 学生 参加 高考 600 元 补课费
1876	H4 县 澧源 镇 八斗 溪 违章建筑 猖獗
3402	G5 县 汽车 东站 黑 出租车 扰民
3664	全民 学 英语 浪费 太 教育资源
9170	西地省 公共卫生 工作 服务 考核 为难
1850	G5 县 澧 浦路 嘉 小区 不动产 证 尚未 发放
3086	D 市 出租车 违法 转包 新 政策
8267	西地省 旅游 管理 重视
8660	请问 2014 年 主治医师 考试 西 省内 乡镇 基层 分数线 出台
7388	D8 县 景区 华夏 祖庙 诱导 游客 添 香油 敛财

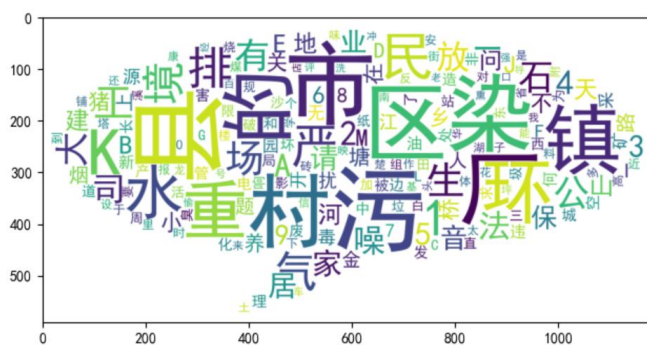
图 1.2.1

经过分词之后我们生成了所有群众留言主题的单个有意义的词语，词语之间由空格隔开，接下来我们尝试在这些词语的基础上生成每个一级标签的词云。具体做法时在每个标签中罗列前 100 个高频词，再画出他们的词云。

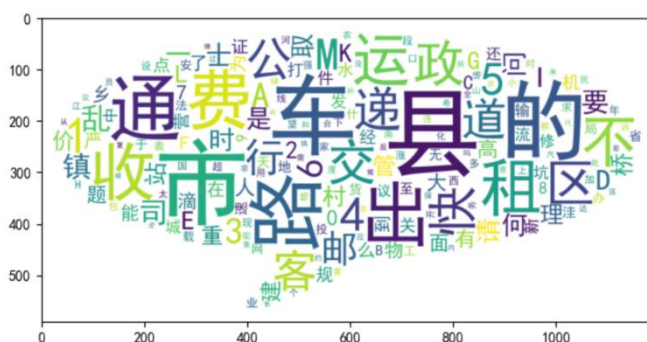
城乡建设



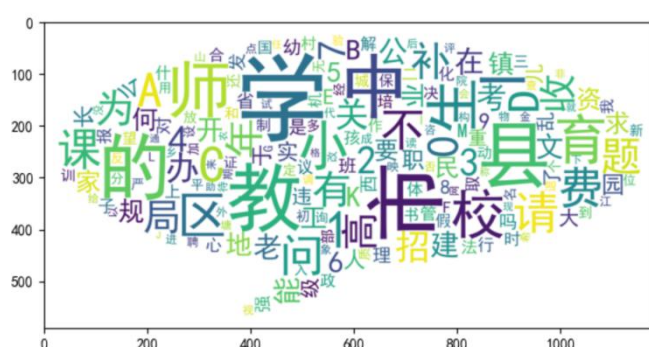
环境保护



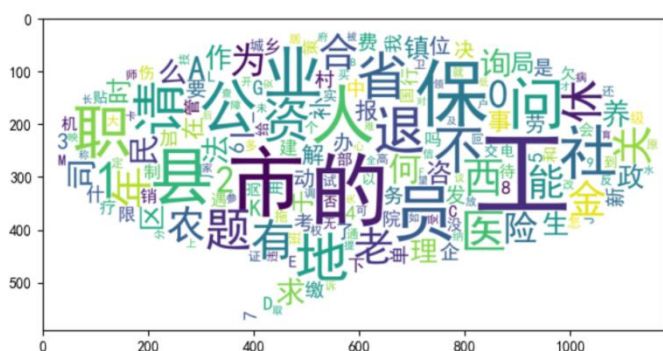
交通运输



教育文体



劳动和社会保障



商贸旅游

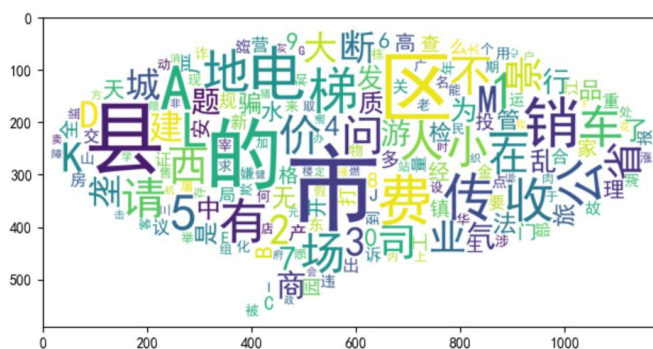
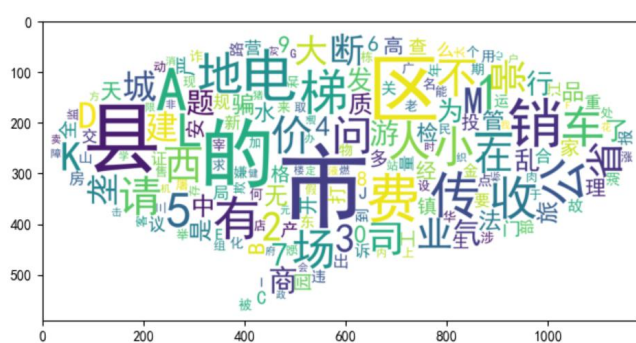


图 1.2.2

1.3 计算 TF-IDF 的特征值^[1]

1.3.1 TF-IDF 的原理

TF-IDF (term frequency - inverse document frequency, 词频-逆向文件频率) 是一种用于信息检索 (information retrieval) 与文本挖掘 (text mining) 的常用加权技术。

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。字词的重要性随着它在文件中出现的次数成正比增加, 但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF 的主要思想是: 如果某个单词在一篇文章中出现的频率 TF 高, 并且在其他文章中很少出现, 则认为此词或者短语具有很好的类别区分能力, 适合用来分类。

① TF 是词频(Term Frequency)

词频(TF)表示词条(关键词)在文本出现的频率。这个数字通常会被归一化, 以防止它偏向长的文件。

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \text{ 即 } TF_w = \frac{\text{在某一类中词条 } w \text{ 出现的次数}}{\text{该类中所有的词条数目}}$$

其中 $n_{i,j}$ 是该词在文件中出现的次数, 分母则是文件中所有词汇出现的次数总和。

② IDF 是逆向文件频率(Inverse Document Frequency)

逆向文件频率(IDF): 某一特定词语的 IDF, 可以由总文件数目除以包含该词语的文件的数目, 再将得到的商取对数得到。

如果包含词条的文档越少, IDF 越大, 则说明词条具有很好的类别区分能力。

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

其中, $|D|$ 是语料库的文件总数, $|\{j: t_i \in d_j\}|$ 表示包含词语 t_i 的文件数目。

③ TF-IDF 实际上是：TF×IDF

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF-IDF = TF * IDF$$

1.3.2 TF-IDF 的具体实现

这里我们会使用 `sklearn.feature_extraction.text`.

`TfidfVectorizer` 方法来抽取留言主题中的 TF-IDF 的特征值。^[2] 这里我们使用了参数 `gram_range=(1, 2)`，这表示我们除了抽取留言中的每个词语外，还要抽取每个词语并组成一个“词语对”，如：词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2，词 3)，(词 3，词 4)。这样就扩展了我们特征集的数量，有了丰富的特征集才有可能提高我们分类文本的准确度。参数 `norm='l2'`，是一种数据标准划处理的方式，可以将数据限制在一定的范围内，比如说 $(-1, 1)$ 。

```
(9210, 53139)
-----
(0, 32724) 0.38992145246512155
(0, 51564) 0.38992145246512155
(0, 27245) 0.38992145246512155
(0, 46243) 0.38992145246512155
(0, 22157) 0.2568946512414059
(0, 32717) 0.2682363300472749
(0, 51547) 0.26611652697072047
(0, 27222) 0.26217763369454206
(0, 46242) 0.3381483927207964
(1, 20628) 0.33061857028391556
(1, 38233) 0.33061857028391556
(1, 8712) 0.33061857028391556
(1, 20746) 0.33061857028391556
(1, 19425) 0.33061857028391556
```

我们看到我们的 features 的维度是 (9210, 53139)，这里的 9210 表示我们总共有 9210 条留言主题数据，53139 表示我们的特征数量这包括全部留言主题中的所有词语数+词语对(相邻两个单词的组合)的总数。

图 1.3.1

下面我们通过卡方检验的方法来找出每个分类中关联度最大的两个词语和两个词语对。卡方检验是一种统计学的工具,用来检验数据的拟合度和关联度。卡方检验的计算公式为:

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

其中, A 为实际值, T 为理论值。

χ^2 用于衡量实际值与理论值的差异程度(也就是卡方检验的核心思想), 包含了以下两个信息:

1. 实际值与理论值偏差的绝对大小(由于平方的存在, 差异是被放大的)
2. 差异程度与理论值的相对大小。

<pre># '交通运输': . Most correlated unigrams: . 的士 . 出租车 . Most correlated bigrams: . 出租车 乱收费 . 滴滴 出行</pre>	<pre># '劳动和社会保障': . Most correlated unigrams: . 职工 . 社保 . Most correlated bigrams: . 退休 工资 . 退休 人员</pre>	
<pre># '城乡建设': . Most correlated unigrams: . 公积金 . 房产证 . Most correlated bigrams: . 棚户区 改造 . 住房 公积金</pre>	<pre># '教育文体': . Most correlated unigrams: . 补课 . 教师 . Most correlated bigrams: . 教师 招聘 . 培训 机构</pre>	<pre># '环境保护': . Most correlated unigrams: . 排放 . 污染 . Most correlated bigrams: . 噪音 扰民 . 噪音 污染</pre>
<pre># '卫生计生': . Most correlated unigrams: . 独生子女 . 医院 . Most correlated bigrams: . 市中心 医院 . 再婚 家庭</pre>	<pre># '商贸旅游': . Most correlated unigrams: . 电梯 . 传销 . Most correlated bigrams: . 传销 组织 . 小区 电梯</pre>	

图 1.3.2

我们可以看到经过卡方(chi2)检验后, 找出了每个分类中关联度最强的两个词和两个词语对。这些词语和词语对能很好的反应出分类的标签。

1.4 朴素贝叶斯分类器^[3]

1.4.1 朴素贝叶斯分类器的原理

首先，用数学公式来表述贝叶斯定理：

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(x,c)}{P(x)}$$

c 表示的是随机事件发生的一种情况。 x 表示的就是证据 (evidence) \状况(condition)，泛指与随机事件相关的因素。

- $P(c|x)$ ：在 x 的条件下，随机事件出现 c 情况的概率。（后验概率）
- $P(c)$ ：（不考虑相关因素）随机事件出现 c 情况的概率。（先验概率）
- $P(x|c)$ ：在已知事件出现 c 情况的条件下，条件 x 出现的概率。（后验概率）
- $P(x)$ ： x 出现的概率。（先验概率）

那么，落实到实际的问题当中，我们想获得的核心结果其实也就是 $P(c|x)$ ，即我们想知道，在考虑了一些现有的因素后，随机事件会以多大概率出现各种情况，通过参考这个结果，我们针对性地作出决策。

而从计算上来说，我们需要同时知道 $P(c)$, $P(x|c)$ 和 $P(x)$ 才能算出目标值 $P(c|x)$ ，而 $P(x)$ 由于是 c 无关，而且作为共同的分母，在我们计算 c 的各种取值的可能性时并不会对各结果的相对大小产生影响，因此可以忽略。略掉了 $P(x)$ 后，最后难点也就落在了计算 $P(x|c)$ 与 $P(c)$ 上，而这两个概率分布是必须要通过我们手上有的数据集来进行估计的。

朴素贝叶斯采用属性条件独立性假设，用公式表达为：

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)\prod_{i=1}^d P(x_i|c)}{P(x)}$$

分类准则用公式表达如下：

$$h_{nb}(x) = \arg \max_{c \in y} P(c) \prod_{i=1}^d P(x_i|c)$$

h_{nb} 代表一个由 naive bayesian(nb) 算法训练出来的 hypothesis（假设），它的值就是贝叶斯分类器对于给定 x 的因素下，最可能出现的

情况 c 。 y 是 c 的取值集合。

在这里，数据都为离散值属性，因此我们只需计算每个属性取值占有所有样本的数量比例：

$$P(x_i | c) = \frac{|D_{c,x_i}|}{D_c}$$

其中， D_c 表示训练集 D 中第 c 类样本组成的集合， $|D_c|$ 表示集合的元数量。 D_{c,x_i} 表示 D_c 中在第 i 个属性值上取值为 x_i 的样本组成的集合。

1.4.2 朴素贝叶斯分类器的具体实现

朴素贝叶斯分类器最适合用于基于词频的高维数据分类器，最典型的应用如垃圾邮件分类器等，准确率可以高达 95% 以上。这里我们使用的是 sklearn 的朴素贝叶斯分类器 MultinomialNB，我们首先将 review 转换成词频向量，然后将词频向量再转换成 TF-IDF 向量。最后我们开始训练我们的 MultinomialNB 分类器。

完成对模型的训练以后，我们编写了一个预测函数 myPredict 查看模型的分类效果：

例如，

```
myPredict('C市昭山示范区违反劳动法，强迫员工加班却不发工资')
```

得到预测结果为： 劳动和社会保障

1.5 模型选择

首先，用数学公式来表述我们尝试不同的机器学习模型，并评估它们的准确率，我们将使用如下四种模型：

- Logistic Regression (逻辑回归)
- (Multinomial) Naive Bayes (多项式朴素贝叶斯)
- Linear Support Vector Machine (线性支持向量机)
- Random Forest (随机森林)

分别进行训练后，得到四种不同的机器学习模型准确率的数值及其箱型图：

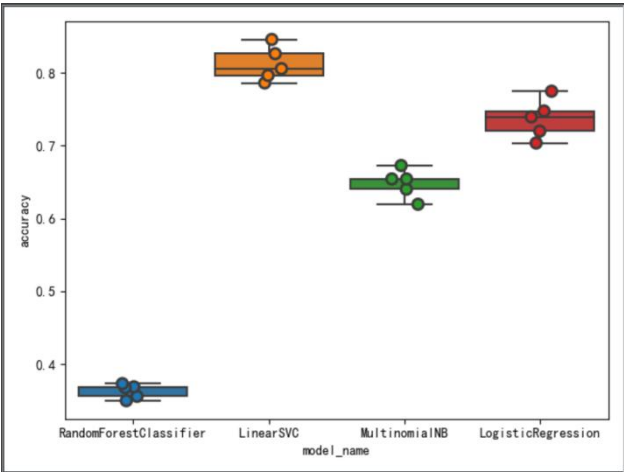


图 1.5.1

```
model_name
LinearSVC      0.812052
LogisticRegression  0.737025
MultinomialNB  0.648100
RandomForestClassifier  0.362866
Name: accuracy, dtype: float64
```

图 1.5.2

由此得到结论：线性支持向量机的平均准确率达到到了 81.2%，其次是逻辑回归和多项式朴素贝叶斯，随机森林的准确率最低，只有 36.3%。

1.6 模型评估

针对准确率最高的 Linear Support Vector Machine 模型，它的混淆矩阵如图：

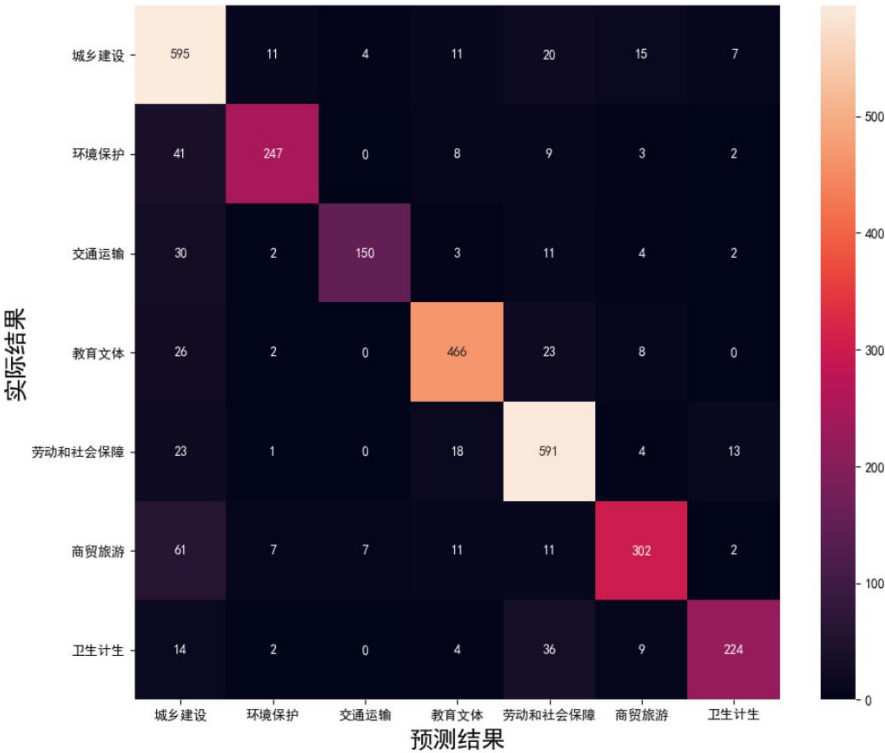
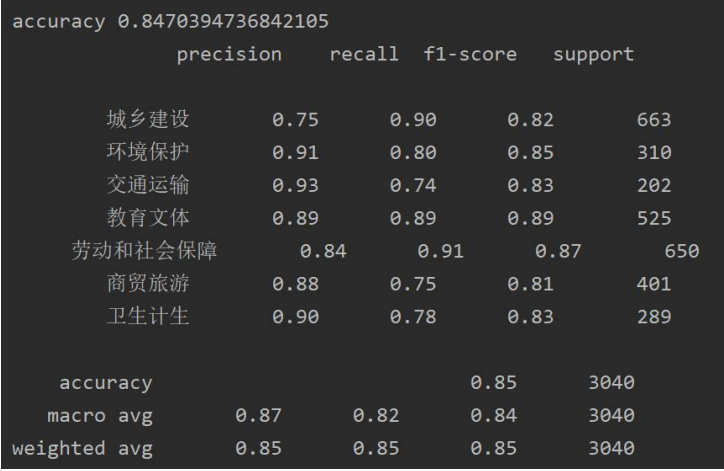


图 1.6.1

混淆矩阵的主对角线表示预测正确的数量,除主对角线外其余都是预测错误的数量。由此可发现城乡建设和商贸旅游的留言主题被分类错误的数量较多。

由于多分类模型一般不使用准确率 (accuracy) 来评估模型的质量,因为当训练数据不平衡(有的类数据很多,有的类数据很少)时, accuracy 不能反映出模型的实际预测精度,这时候我们就需要借助于 F1 分数、ROC 等指标来评估模型。



	precision	recall	f1-score	support
城乡建设	0.75	0.90	0.82	663
环境保护	0.91	0.80	0.85	310
交通运输	0.93	0.74	0.83	202
教育文体	0.89	0.89	0.89	525
劳动和社会保障	0.84	0.91	0.87	650
商贸旅游	0.88	0.75	0.81	401
卫生计生	0.90	0.78	0.83	289
accuracy			0.85	3040
macro avg	0.87	0.82	0.84	3040
weighted avg	0.85	0.85	0.85	3040

图 1.6.2

从 f1-score 看来,总体预测的准确率为 0.85,分类的结果比较理想。

2 热点问题挖掘

2.1 了解数据与建模流程

任务描述

本次建模针对群众问政留言记录数据,在对文本进行基本的中文分词、停用词过滤后,通过建立词袋模型、文本相似度计算、权值转换、层次聚类等自然语言处理和文本挖掘方法,实现对文本相似留言数据的归类以及热点问题的挖掘。

2.1.1 数据分析

某一时段内群众集中反映的某一问题可称为热点问题，在附件 3 所给的数据中，有 4327 行 6 列，共有 4326 条留言数据，留言时间跨度为从 2017-06-08 17:31:20 至 2020/1/8 9:32:33。

附件 3 数据中含有“留言主题”与“留言详情”，留言主题是留言详情的概括，故我们采用“留言主题”的文本来进行相似留言归类。

2.1.2 建模流程

- 1) 分词
- 2) 构建词袋模型
- 3) 权值转换
- 4) 计算余弦相似度
- 5) 层次聚类
- 6) 进行热度排名

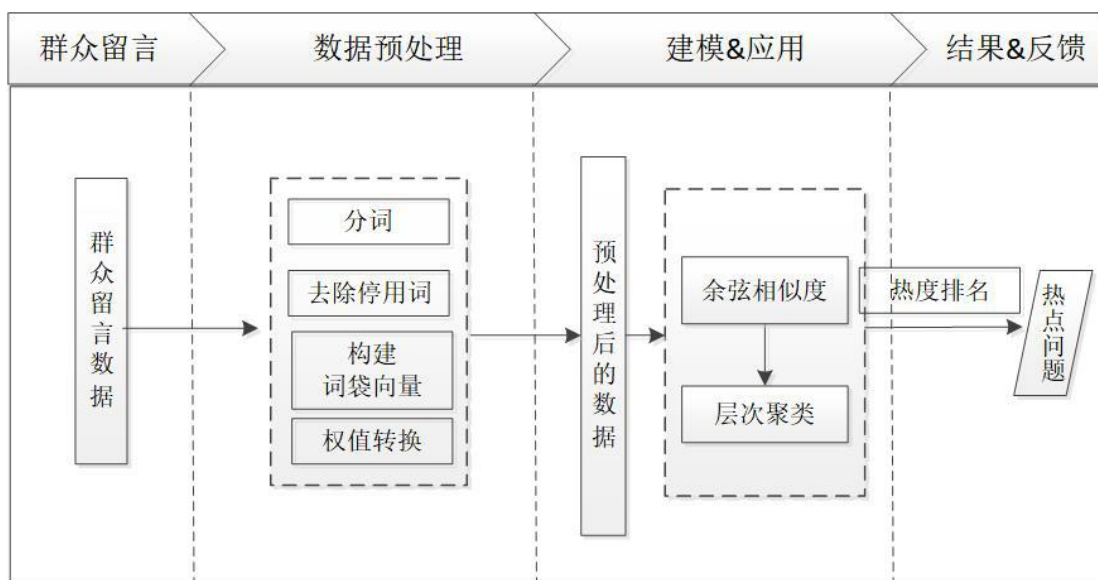


图 2.1.1

2.2 “留言主题”数据预处理

将附件 3 按留言先后时间进行了一次排序，然后将附件 3 中的留言主题的内容提取出来，存放在“data.txt”文件内。

2.2.1 中文分词与去除停用词

分词是文本信息处理的基础环节，是将一个单词序列切分成一个一个单词的过程。准确的分词可以极大的提高计算机对文本信息的是被和理解能力。分词最常用的工作包是 jieba 分词包，jieba 分词是 python 写成的一个分词开源库，专门用于中文分词。

针对“留言主题”的文本内容的分词，我们采用 jieba 分词并借助停用词表“stoplist.txt”去除停用词，减少通用词的干扰，结果保存在“cutWords.txt”文件中，方便后续调用。下图为部分结果输出。



4	人才	APP	申请	购房	补贴	通不过
5	实行	独生子女	护理	假		
6	12123	申请	驾驶证	期满	换证	星期 无人 受理
7	经济	学院	强制	学生	实习	
8	请问	普及	5G	网络		

图 2.2.1

2.2.2 构建词袋模型

词袋模型（BOW），是一种使用机器学习算法从文本中提取特征的方法。该方法非常简单和灵活，可以用于从文档中提取各种功能的各种方法。词袋（Bag-of-words）是描述文档中单词出现的文本的一种表示形式。它涉及两件方面：一是已知词汇的集合；二是测试已知单词的存在。[1]

BOW 对词的统计有不同的评分方法，例用二进制来表示单词的存在或不存在，计算，频率等。

我们将“留言主题”的文本切分成单词后，需要进一步转换成向量，方便后续的计算。在这里，我们采用频率的评分方法对词进行统计。

为方便表述，假设留言主题中出现的所有词语的集合标记为 $W = (w_1, w_2, \dots, w_M)$ 。通过 TF-IDF 算法，得到留言主题 A 中每个词条 TF-IDF 值的向量，记做 $t = (t_1, t_2, \dots, t_M)$ ，其中 t_1 表示 w_1 在留言主题 A 中的 TF-IDF 值。

于是可以将要比较留言主题 A1, A2 表示为 TF-IDF 值的向量：

$$A_1 = (t_{11}, t_{12}, \dots, t_{1M})$$

$$A_2 = (t_{21}, t_{22}, \dots, t_{2M})$$

则留言主题 A_1, A_2 之间的相似度为：

$$\cos \theta = \frac{\sum_1^n A_1 \times A_2}{\sqrt{\sum_1^n A_1^2} \times \sqrt{\sum_1^n A_2^2}}$$

下图为计算余弦相似度代码。

```

133 # 计算余弦距离
134 def gen_sim(A, B):
135     num = float(dot(mat(A), mat(B).T))
136     denum = linalg.norm(A) * linalg.norm(B)
137     if denum == 0:
138         denum = 1
139     cosn = num / denum
140     sim = 0.5 + 0.5 * cosn # 余弦值为 [-1,1], 归一化为 [0,1], 值越大相似度越大
141     sim = 1 - sim # 将其转化为值越小距离越近
142     return sim

```

图 2.3.1

2.3.2 层次聚类

基于前面的准备，我们采取文本聚类的方法对相似留言进行归类，由于留言归类的分类数未知，且基于 DBI (DB 指数) 的值越小，同时 DI (Dumn 指数) 的值越大，聚类的效果越好的性能度量。我们采取的是层次聚类的方法将相似的留言进行归类，该算法可以通过学习得到距离阈值作为聚类结束的条件，从而解决分类数未知的问题。

在进行聚类前，将附件 3 中的留言主题的内容提取出来，存放在

“data.txt”文件内。经过数据预处理后，设定距离阈值为 0.3 开始进行聚类，将聚类的类别按照主题序号给出，保存在“data1.txt”文件中。部分结果如下图所示：

1	0.0
2	0.0
3	1.0
4	2.0
5	3.0
6	4.0
7	0.0
8	5.0
9	6.0
10	7.0

图 2.3.2 聚类类别

4326 条留言分为了 3207 个类别（由于群众语义表达的不同，部分相似留言没有聚类在一起，忽略部分不相似留言聚在一起的问题，后面再根据数量最多前 5 个类型用过筛选地址归类）
将所得结果放入附件三，根据类别进行排序，下图部分结果展示：

	A	B	C	D	E	F	G	H
1	类别	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
2	0	360114	A0182491	A市经济学院体育学院	2017-06-08 17:31:20	书记您好，我是来自西地省经	9	0
3	0	360113	A3352352	A市经济学院强制学生	2018-05-17 08:32:04	A市经济学院强制16届电子商务跟	3	0
4	0	360112	A220235	A市经济学院强制学生	2019-04-28 17:32:51	各位领导干部大家好，我是A市经	0	0
5	0	233759	A909118	A市涉外经济学院强制	2019/04/28 17:32:51	各位领导干部大家好，我是A市涉	0	0
6	0	242062	A00028889	西地省涉外经济学院	2019/11/27 23:14:33	请制止和修改西地省涉外经济学院	0	0
7	0	255719	A00060810	A市商贸旅游职业技术	2019/4/28 17:32:51	各位领导干部大家好，我是A市商	0	1
8	1	286572	A23525	请求A市地铁2#线在梅	2018-10-27 15:13:26	领导好！今天游了A3区山，在A3区	3	0
9	2	289408	A0012413	在A市人才app上申请	2018-11-15 16:07:12	我叫朱琦梦，是2017年12月落户，	0	0
10	2	224042	A00014225	咨询A市人才购房及购	2019/1/16 11:58:48	《A市房屋交易管理中心关于执行	0	0
11	2	282248	A000106091	咨询A市人才购房补贴	2019/1/30 15:42:19	您好！政府2017年出了”A市域内	0	0
12	2	268023	A00030760	反映A市人才补贴问题	2019/10/29 19:32:19	关于外来人才进行补贴的政策，这	0	0
13	2	265577	A0004787	咨询A市人才购房补贴	2019/12/2 11:57:49	您好！本人于明年硕士毕业，目前	0	1
14	2	225657	A00051791	关于A市人才购房补贴	2019/2/25 14:43:15	根据《A市人才购房及购房补	2	6
15	2	270015	A00020115	为何我的A市人才购房	2019/5/29 16:51:26	我于2014年结婚，2017年户口迁入	0	0

图 2.3.3 根据类别排序

2.4 “留言主题”的热点问题

2.4.1 热度排序

通过观察附件 3 的点赞数和反对数这两列，发现大部分都较小，只有部分留言的点赞数和反对数较大，相比之下，评论数更能反映问题热度的大小。故我们选择以相似问题的评论数大小作为热度排序的度量，评论数越大，热度越高。

2.4.2 热度问题

通过对同一类别的数量统计，如图 2.4.1 所示，可知前 5 个热点问题为第 13、41、2、66、11、186 类别所反映的问题，即 A 市伊景园滨河苑捆绑销售车位问题，A 市丽发新城小区附近的搅拌站噪音严重扰民，A 市人才补贴问题，反映请 A 市加快国家中心城市建设，A5 区劳动东路魅力之城小区油烟扰民问题，A3 区西湖街道茶场村五组拆迁问题，部分数据如图 2.4.2 所示。

	label	count
0	13	28
1	41	15
2	2	12
3	66	11
4	11	10
5	186	10

图 2.4.1

11	360107	A0283523	A5区劳动东路魅力之城	2019-07-21 10:29:36	局长： 你好，A5区劳动东路魅	3	0
11	360101	A324156	A5区劳动东路魅力之城	2019-07-28 12:49:18	尊敬的政府：A5区劳动东路魅力之	4	0
11	360108	A0283523	A5区劳动东路魅力之城	2019-08-01 16:20:02	局长： 你好，A5区劳动东路魅	6	0
11	360102	A1234140	A5区劳动东路魅力之城	2019-09-10 06:13:27	A5区劳动东路魅力之城小区，底层	0	0
11	360103	A0012425	A5区劳动东路魅力之城	2019-09-25 00:31:33	A5区劳动东路魅力之城小区临街夜	1	0
11	284147	A909113	A5区劳动东路魅力之城	2019/07/21 10:29:36	局长： 你好，A5区劳动东路魅	3	
11	236798	A00039089	A5区劳动东路魅力之城	2019/07/28 12:49:18	尊敬的政府：A5区劳动东路魅力之	0	4
11	272122	A909113	A5区劳动东路魅力之城	2019/08/01 16:20:02	局长： 你好，A5区劳动东路魅	0	6
11	268914	A0006238	A5区劳动东路魅力之城	2019/09/10 06:13:27	A5区劳动东路魅力之城小区，底层	0	0
11	246598	A00054842	A5区劳动东路魅力之城	2019/09/25 00:31:33	A5区劳动东路魅力之城小区临街夜	0	1
13	224767	A909176	伊景园滨河苑车位捆绑	2019-07-30 14:20:08	伊景园滨河苑车位捆绑销售！广铁	0	0
13	285897	A909191	武广新城伊景园滨河苑	2019-08-01 20:06:52	我们是广铁集团铁路职工，武广新	4	0
13	205982	A909168	坚决反对伊景园滨河苑	2019-08-03 10:03:10	我坚决反对伊景园滨河苑捆绑销售	0	2
13	276016	A909181	车位属于业主所有，不	2019-08-06 00:00:00	尊敬的胡书记，您好！我叫陈玉春	0	2
13	271517	A909238	开发商联合广铁集团推	2019-08-11 12:02:27	你好，本人购买伊景园滨河苑楼盘	0	0
13	205277	A909234	伊景园滨河苑捆绑车位	2019-08-14 09:28:31	广铁集团强制要求职工购买伊景园	0	1
13	195511	A909237	车位捆绑违规销售	2019-08-16 14:20:26	对于伊景园滨河苑商品房，A市广	4	0
13	230554	A909174	投诉A市伊景园滨河苑	2019-08-19 10:22:44	投诉A市伊景园滨河苑开发商强制	0	0
13	234633	A909194	无视消费者权益的A市	2019-08-20 12:34:20	伊景园滨河苑项目商品房，广铁集	0	0

图 2.4.2 部分前五热点问题

由于语义原因，存在着相似问题却没有划分到同一类的问题，我们在前 5 个热点问题的基础上，按地点或事件完善相似问题评论数。

通过以地点或事件再次筛选，发现对 A 市伊景园滨河苑捆绑销售车位问题，共有 51 条留言；对 A 市丽发新城小区附近的搅拌站噪音严重扰民，共有 46 条留言；对 A 市人才补贴问题，共有 22 条留言，对反映请 A 市加快国家中心城市建设，共有 13 条留言；对 A5 区劳动

东路魅力之城小区油烟扰民问题，共有 21 条留言；对 A3 区西湖街道茶场村五组拆迁问题，共有 10 条留言。

故排名前 5 的热点问题为：A 市伊景园滨河苑捆绑销售车位问题，A 市丽发新城小区附近的搅拌站噪音严重扰民，A 市人才补贴问题，A5 区劳动东路魅力之城小区油烟扰民问题，反映请 A 市加快国家中心城市建设问题。

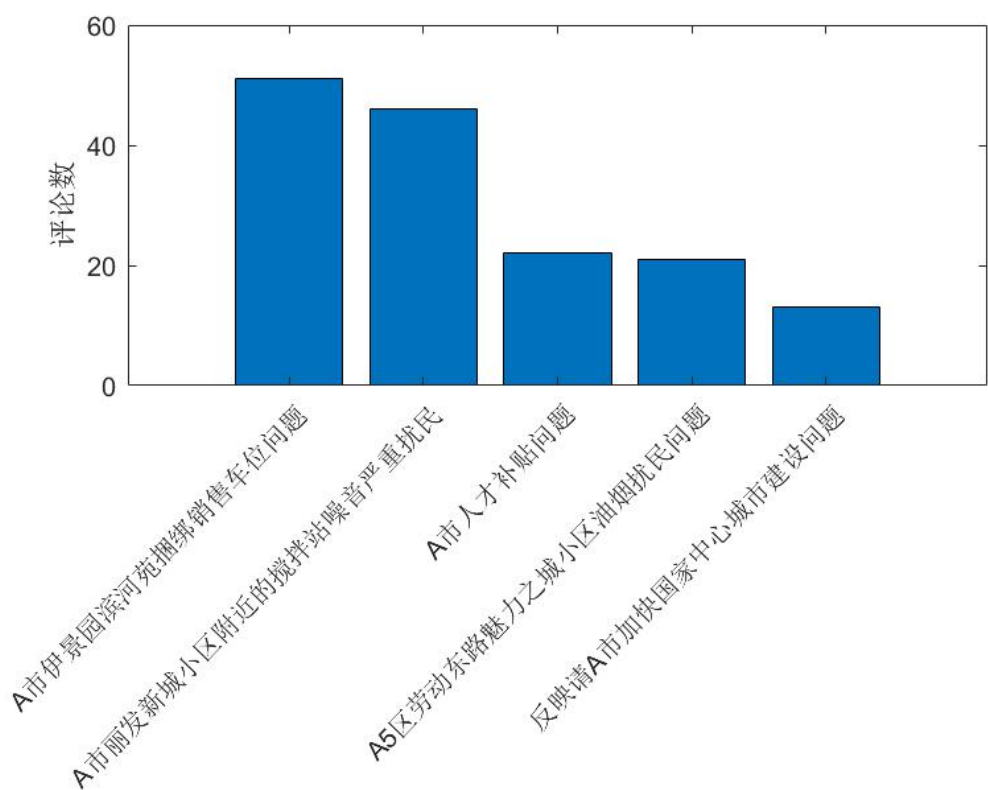


图 2.4.3

3 答复意见的评价

3.1 确定影响答复意见质量因素的比例

3.1.1 构造判断矩阵知识

在评价指标的权重值设置中，我们通过层次分析法对比分析指标之间的两两重要性，得到各评价指标权重比重，构建两两对比比较矩阵；在两两比较矩阵进行构建的过程中，使用萨蒂提出的 1-9 标度法进行确定^[7]。

数值	数值代表的含义
1	若 x_i 等价于 x_j
3	若 x_i 比 x_j 稍微重要
5	若 x_i 比 x_j 明显重要
7	若 x_i 是强烈重要的
9	若 x_i 是极端重要的
2, 4, 6, 8	小于和大于两个基数之间的模糊表达
$1/2, 1/3, \dots, 1/9$	对应于以上等级的 x_i 和 x_j 之间的关系

表 3.1.1

在对比设置指标间的权重比值时，最终形成的评价指标的比较矩阵如下：

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}$$

其中 A 为比较矩阵， a_{ij} 是要素 i 与要素 j 重要性比较结果，且有如下关系： $a_{ij} = 1/a_{ji}$

3.1.2 构造判断矩阵

字母	代表的含义
X1	相关性
X2	完整性
X3	可解释性
X4	礼貌性

表 3.1.2

将四个元素 X1、X2、X3、X4 两两比较，得成对比较矩阵。

得到对比矩阵如下：

	X1	X2	X3	X4
X1	1	3	2	4
X2	1/3	1	1/2	3
X3	1/2	2	1	2
X4	1/4	1/3	1/2	1

表 3.1.3

则得出判断矩阵为：

$$A = \begin{bmatrix} 1 & 3 & 2 & 4 \\ \frac{1}{3} & 1 & \frac{1}{2} & 3 \\ \frac{1}{2} & 2 & 1 & 2 \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{2} & 1 \end{bmatrix}$$

3.1.3 一致性检验

求解 M-C 的特征值，易解得 $\lambda = 4.124$

有公式 $CI = (\lambda - n) / (n - 1)$

于是根据 $CR = CI / RI$, 计算得到 $CR = 0.0413 < 0.1$, 通过了一致性检验。

n	2	3	4	5	6	7	8	9	10	11
RI	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49	1.51

表 3.1.4

3.1.4 归一化处理

对上述矩阵 A 中的每一行元素做乘积运算： $M_i = \prod_{j=1}^n a_{ij}$ ($i=1, 2, \dots, n; j=1, 2, \dots, n$)。然后计算 M_i 的 n 次方根值，计算公式为：

$$W_i = \sqrt[n]{M_i}$$

接着再对 W_i 做归一化处理，计算方法为： $W = W_i / \sum W_i$

经过系列处理计算后，即得到评价指标集的权重向量 $W = \{w_1, w_2, \dots, w_n\}$

A 矩阵经过归一化处理可得权向量为 (0.471, 0.179, 0.253, 0.096)

	权重
相关性	0.471
完整性	0.179
可解释性	0.253
礼貌性	0.096

表 3.1.5

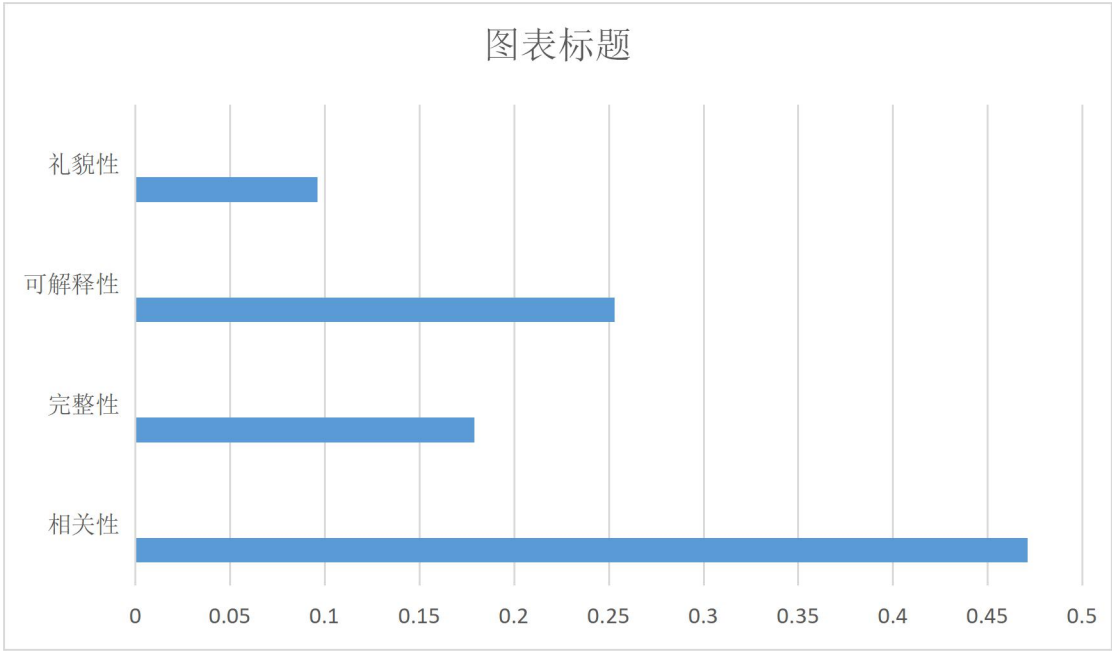


图 3.1.1

3.2 礼貌性的量化

3.2.1 数据预处理

取出附件四中的答复意见，利用 Python 中的 jieba 库对答复意见进行分词，分词结果如下：

'激烈', ' ', ' ', '造成', '有线', '网络', '公司', '用户', '减少', ' ', ' ', '收入', '下滑', ' ', ' ', '职工', '生活', '难以',

图 3.2.1

发现分词结果中存在对数据处理没有用处的分词结果。

接着对分词后的结果进行数据清洗，去除掉停用词，发现结果中存在很多电话号码，将数字加入停用词表，得到结果如下：

'吃喝玩乐', '违规', '违纪行为', '特此', '回复', '年', '31', '网友', '您好', '收悉', '并交', '单位', '调查',

图 3.2.2

得到的结果都是有实际意义的词语，我们得到了数据预处理的结果。

3.2.2 词频统计

对分词后的结果进行统计,统计每一个词语的词频,得到的结果如下:

您好	2395
市	2266
网友	2162
县	2100
回复	2054
收悉	1840
相关	1798
建设	1660
感谢您	1564
支持	1452
项目	1233
留言	1231
区	1151
我局	1001

图 3.2.3

3.2.3 根据词频绘制词云

根据得到的词频绘制答复的词云，得到的结果如下：



图 3.2.4

3. 2. 4 得到礼貌性的词典

根据词云得到出现较多的关于礼貌性的词语，统计结果如下：

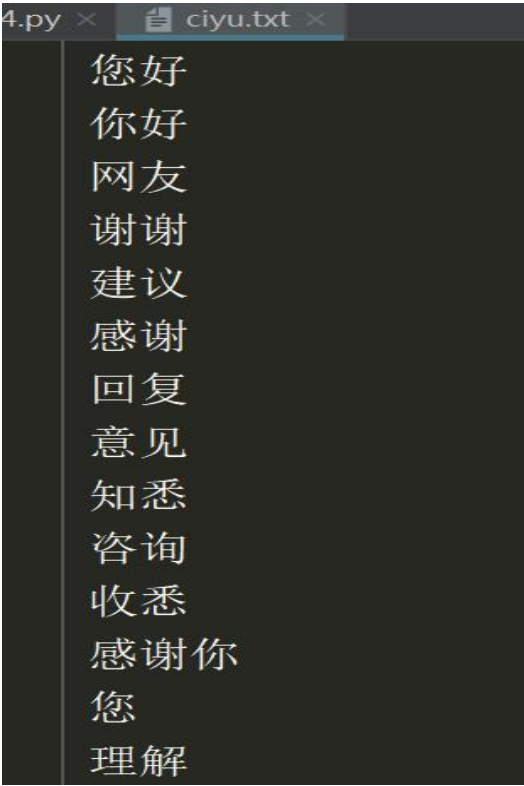


图 3.2.5

3. 2. 5 得到礼貌性词语的个数

得到每一条答复中词典中的词语的个数，统计结果如下：



图 3.2.6

根据答复中礼貌性词语的个数对礼貌性进行打分，具体评分结果为：礼貌性词语个数大于 4，结果为 100 分，礼貌性词语个数小于 4，结果为 30 乘以礼貌性词语的个数。

3. 2. 6 得到礼貌性的评分

根据评分规则，得到每一条答复关于礼貌性的评分，具体评分结果如下：



图 3.2.7

3.3 相关性的量化

3.3.1 对留言详情和留言回复进行相似度分析

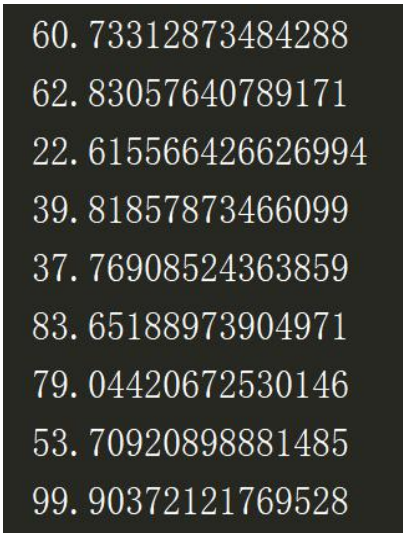
采用余弦相似度算法，思路如下：^[8]

- (1) 利用Python中的jieba分词，分别得到两个列表listA和listB
- (2) 列出所有词，将listA和listB放在一个set中，得到集合set，将set转为dict，key为set中的词，value为set中词出现的位置。
- (3) 将listA和listB进行编码，将每个字转换为出现在set中的位置
- (4) 对listAcode和listBcode进行oneHot编码，就是计算每个分词出现的次数。
- (5) 得出两个句子的词频向量之后，就变成了计算两个向量之间夹角的余弦值，值越大相似度越高。

3.3.2 对相似度的量化

根据附录4中的留言详情和留言答复，确定每一条留言详情和留言答复的相似度，对相似度进行量化。

经过多次反复试验，确定比例系数为170，所以最终相似度的量化结果=余弦相似度值×200，运行的结果如下：



```
60.73312873484288
62.83057640789171
22.615566426626994
39.81857873466099
37.76908524363859
83.65188973904971
79.04420672530146
53.70920898881485
99.90372121769528
```

图 3.3.1

3.4 可解释性的量化

3.4.1 量化分析思路

- (1) 将留言详情和留言回复分别采用 jieba 分词
- (2) 将 jieba 分词后的结果放在列表中
- (3) 统计两个列表中相同词语出现的个数
- (4) 根据个数进行量化

3.4.2 将可解释性量化

根据两个列表中相同词语出现的个数进行打分，具体评分结果为：相同词语个数大于等于 10，结果为 100 分，相同词语个数小于 10，结果为 10 乘以相同词语的个数。

最终得到可解释性的评分结果如下：

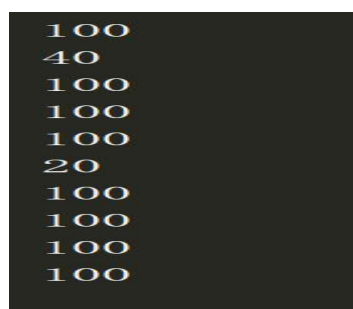


图 3.4.1

3.5 完整性的量化

3.5.1 量化分析思路

- (1) 将留言回复采用 jieba 分词进行分词
- (2) 将 jieba 分词后的结果放在列表中
- (3) 统计列表中词语的个数
- (4) 检测留言回复中是否存在完整的开头敬词和结尾的留言回复时间。

3.5.2 将完整性量化

根据列表中词语的个数进行打分以及格式的规范性进行打分，具体评分结果为：词语个数得分 $\times 70\%$ +格式的规范性得分 $\times 30\%$

词语分数评分规则：词语个数大于等于 70，结果为 100 分，词语个数小于 70，结果为 1.5 乘以相同词语的个数。

格式的规范性评分规则：具有完整的开头或则结尾得分 50，完整的开头和结尾得分 100，否则为 0 分。

将完整性量化的结果如下：

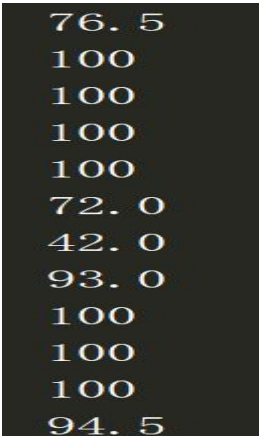


图 3.5.1

3.6 整体评分结果的量化

答复的评价结果为 = 相关性×0.471+完整性×0.179+可解释性×0.253+礼貌性×0.096

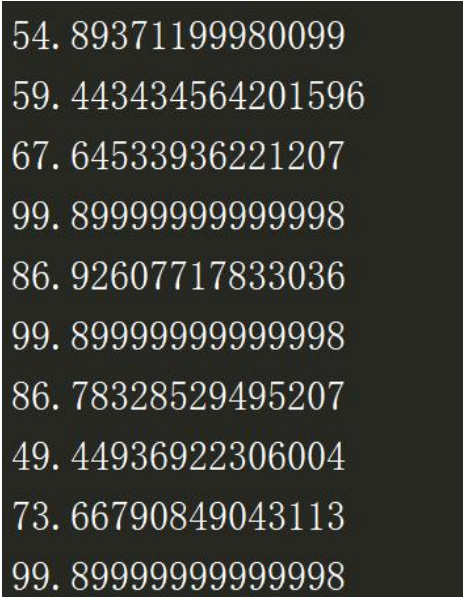


图 3.6.1

4 总结与展望

综上所述，我们提出了基于 TF-IDF 构建的文本分类模型，以及基于层次聚类法的热点问题分析模型，它可以判别两段文本的匹配程度，我们的模型在线下测试集上取得了良好而鲁棒的特性。总体来看我们的贡献集中在文本挖掘与深度学习的融合上。在接下来的工作中，我们将主要关注在如何更好的解决进行分类，进一步提高模型的准确率，热点分析更加的真实可靠，反应人民群众的迫切需求，对答复意见合理评价，则反应了政府的工作水平。

致谢：十分感谢泰迪杯官方给出关于智慧政务系统的构建赛题，我们都知道，在当今互联网快速发展的时代，人工智能、大数据的发展能够更好的服务我们的生活，我们希望我们的初步尝试能够让民众能够更容易的反映生活中的问题，政府的效率更加提高，更好的解决人民遇到的亟待解决的问题。希望我们的工作可以成为人工智能黄金时代的一朵小小的浪花，为人类的事业添砖加瓦。

参考文献

- [1] TF-IDF 算法介绍及实现
<https://blog.csdn.net/asialeebird/article/details/81486700>
- [2] 使用 python 和 sklearn 的中文文本多分类实战开发
https://blog.csdn.net/weixin_42608414/article/details/88046380?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522158826152119725222411382%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=158826152119725222411382&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~first_rank_v2~rank_v25-2
- [3] 朴素贝叶斯分类器 (Naive Bayesian Classifier)
https://blog.csdn.net/qq_32690999/article/details/78737393?ops_request_misc=&request_id=&biz_id=102&utm_term=%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%88%86%E7%B1%BB%E5%99%A8&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduweb~default-3-78737393
- [4] 词袋模型的通俗介绍
<https://www.jianshu.com/p/97862a7e3740>
- [5] TF-IDF 算法介绍及实现
<https://blog.csdn.net/asialeebird/article/details/81486700>
- [6] 机器学习笔记 (3) ——使用聚类分析算法对文本分类 (分类数 k 未知)
https://blog.csdn.net/leaf_zizi/article/details/82684921
- [7] 基于层次分析的模糊综合评价方法-道客巴巴
<https://www.doc88.com/p-3867896433901.html>
- [8] 使用余弦相似度算法计算文本相似度
<https://www.cnblogs.com/airnew/p/9563703.html>