摘要

随着时代的发展和网络的普及。中央明确要求，各单位要坚持开门搞活动，要深入群众听取意见和建议，问政于民、问需于民、问计于民。面对众多群众的留言数据，由人工对数据进行一一归类、分析、答复等繁琐的过程，这不仅对人工技术方面有很高的要求，而且间接性的给相应部门人员带来沉重的压力，加大政府部门的管理难度。人工的能力是有限的，更避免不了错误，因此基于 Python 语言，可以做到快速高效且标准的解决这些问题，能给群众带来前所未有的便利。

针对问题 1，首先利用 pandas 库读取附件 2 内容，通过留言编号知道留言用户，同一个留言用户可以多次留言。使用 jieba 库对原始留言详情进行分词，并进行数据清洗和停用词过滤，再根据一定规则选择特征词之后，结合朴素贝叶斯方法训练样本分类，与其他学习方法相比，取得较高的结果。

针对问题 2，根据热点问题的挖掘，先对附件 3 的内容中无意义的词停用处理，后采用函数处理对文件内容进行理解，在对已经被分割识别的词进行停用处理，过滤出有用的信息提高机器对语言文字的处理效率和对问题的识别正确率。

针对问题 3，对答复意见的评价先利用 Python 对数据预处理，使用 chineseStopWords.txt 停用词文本，去掉停用词和剔除重复无效的数据，可以提高信息的有效性。基于 gensim 对答复意见给予评价：将数据进行遍历后分词，生成词典，通过 doc2dow 稀疏向量生成语料库，再利用 TF-IDF 模型算法，计算出 TF 值可以得到稀疏矩阵的相似度，最后采用了两重 for 循环让留言详情和答复意见可以循环进行两者之间的相关性的比较，即两者的相似度。根据答复意见评价标准。

关键词： Python 语言　文本挖掘应用　朴素贝叶斯　数据分析分类

Abstract

With the development of the times and the popularization of the network. The Central Committee clearly requires that all units open their doors to carry out activities, listen to the opinions and suggestions of the masses in depth, and ask the government for the people, ask the needs for the people, and ask the plan for the people. In the face of many people's message data, manual data classification, analysis, response and other tedious process, which not only has a high demand for artificial technology, but also indirectly brings heavy pressure to the corresponding department personnel, and increases the management difficulty of government departments. The human ability is limited, and mistakes can not be avoided. Therefore, based on Python language, these problems can be solved quickly, efficiently and standard, which can bring unprecedented convenience to the masses.

For problem 1, firstly, we use pandas library to read the content of attachment 2, and know the message user through the message number. The same message user can leave multiple messages. The original message details are segmented by using the jeeba database, the data is cleaned and the stop words are filtered, and then the feature words are selected according to certain rules, and the sample classification is trained by combining the naive Bayes method. Compared with other learning methods, the result is higher.

For problem 2, according to the mining of hot issues, the meaningless words in the content of Annex 3 are stopped first, and then the function processing is used to understand the content of the file. After the word that has been segmented and identified is stopped, useful information is filtered out to improve the processing efficiency of the machine for language words and the recognition accuracy of the problem.

For question 3, the evaluation of the response first uses Python to preprocess the data, uses chinesestopwords.txt to stop the word text, removes the stop word and removes the duplicate invalid data, which can improve the effectiveness of the information. Based on gensim, we evaluate the response: after traversing the data, we segment words, generate dictionaries, generate idiom database through doc2dow sparse vector, and then use TF-IDF model algorithm to calculate TF value to get the similarity of sparse matrix. Finally, we use the double for loop to make the correlation comparison between message details and response comments circularly, that is, the similarity between the two Degree.According to the evaluation criteria of the replies.

Key words: Python language; text mining application; naive Bayes; data analysis and classification

目录

1 挖掘目标

1.1 挖掘背景

1.2 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。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的

答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

2 问题分析

2.1 问题 1 群众留言分类

2.1.1 中文文本预处理

【中文文本分词】

（1）用 pandas 读取附件 2 内容，将留言详情这一列表转为中文文本分词

（2)利用 python 语言中的 jieba 库将附件 2 中的留言详情这一列数据进行分词

【数据清洗】

(3)发现并纠正文件中出现错误的数据，保证处理的文件内容和格式的准确性，完整性和一致性。

【去掉停用词】

（4)为提高文本分类的准确性，词汇之间不具有特殊性，建立一个停用词词典，在分词后的每个词与停用词字典中的条目进行匹配，如果匹配成功，该词将为删除，过滤到停用词词典中并找出有价值词后输出列表，转成字符串。

```
#读入停用词库，事先分析text，将没有正确分词的词加入词库stoplist.txt
stop_word=pd.read_table('stoplist.txt',sep='aaaa',encoding='utf-8')
stop_word=stop_word.iloc[:,0].tolist()
#停用词过滤，找出有价值词
```

2.1.2 文本特征选择

特征文本是为了提高文本分类的效率，减少计算复杂度。
常用方法有：文档频率、信息增益、互信息、统计量和期望交叉熵等
特征预处理：去掉停用词，可进行同义词合并，还可以去掉通用词

2.1.3 文本分类

（1）文本分类用一个已标记类别的文本数据集来训练分类器，这个文本数据集称为训练集。
（2）常用的文本分类算法有：KNN 算法、SVM 算法、贝叶斯算法等。
（3）贝叶斯算法能运用到大型数据库中，且方法简单，分类准确率高、速度快。
贝叶斯法则是关于随机事件 A 和 B 的条件概率和边缘概率的。

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

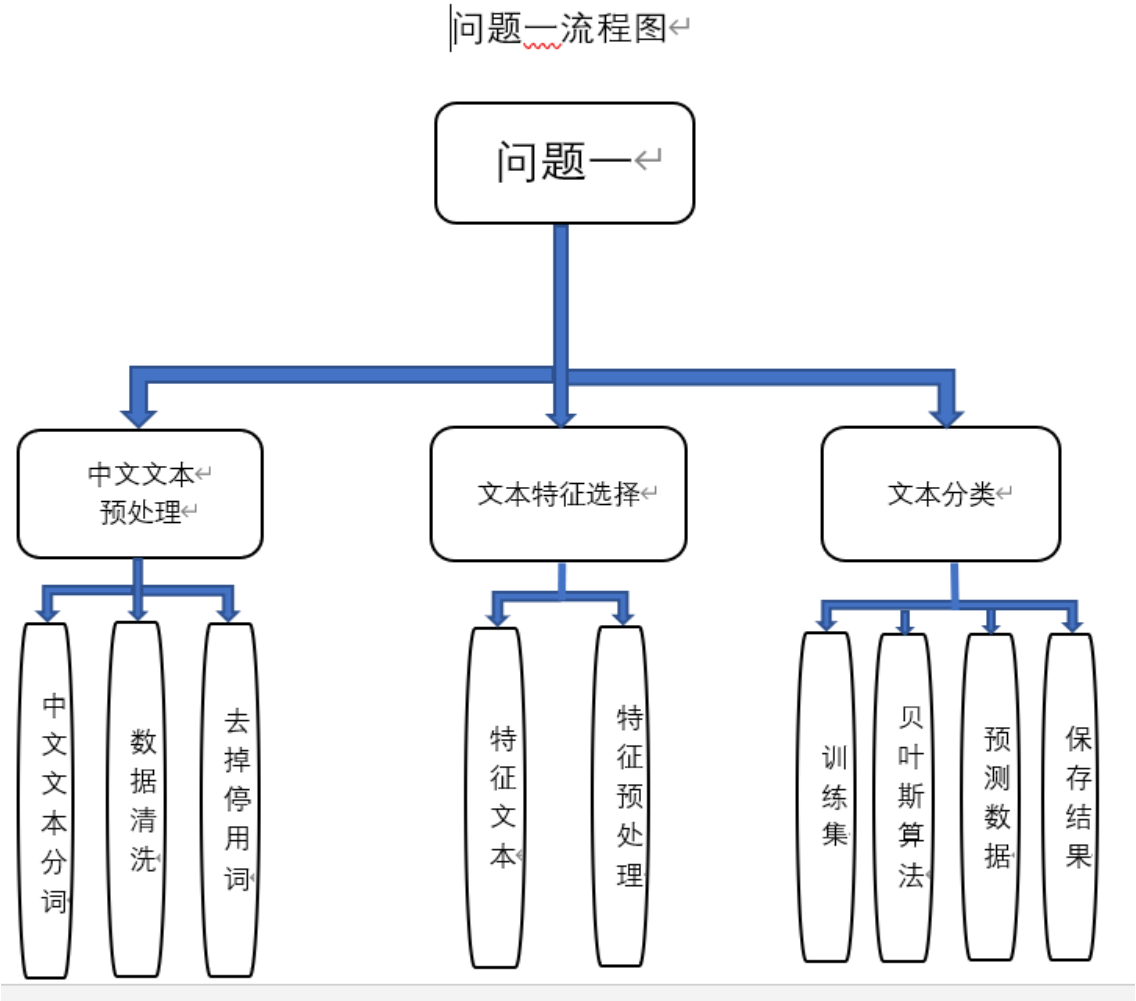
贝叶斯定理

P(A|B,C)=P(B|A)*P(A)*P(C|A,B)/(P(B)*P(C|B))

从理论上讲，与其他所有的分类算法相比，贝叶斯分类具有最小的出错率，在其他条件独立的假定成立的前提下，它是最佳的分类算法。
(4)样本训练模块完成对样本集的训练，模块的输入是训练样本集，样本的输出是训练结果，可作为分类计算模块的输入数据。在训练模块中设计了 TrainedResult 类作为保存训练结果的数据结果。同时，在训练过程中需要对文本切分，形成特征向量，为此设计了 DocmentSeperator 类和 StopWordsHandler 类。每一类的功能有：
TrainedResult 类：训练结果类，用于保存训练的结果。
Trainer 类;样本训练类，完成对样本集的分析，统计样本类别、样本数量。分析特征向量，统计特征词和特征词出现的次数，形成寻来你结果。
DocmentSeperator 类：文本切分类，完成文本切分，形成词语串。
StopWordsHandler 类：停用词处理类，剔除对分类不起作用的停用词。
预测数据，提高文本分类的计算，减少不必要的失误。

```
from sklearn.model_selection import train_test_split
train_data,test_data,train_label,test_label = train_test_split(
    data['留言详情'],data['一级标签'],test_size=0.2)
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
train_cv = cv.fit_transform(train_data)
train_cv = train_cv.toarray() #MemoryError
#  贝叶斯算法
from sklearn.naive_bayes import MultinomialNB
model_nb = MultinomialNB()
model_nb.fit(train_cv,train_label)
```

(5)本文采用文本分类普遍接受的评估指标来评价文本分类的性能，即准确性(Precision)、查全率（Real）和 FI 测试值。
(6)保存结果，保存模型



2.2 问题二热点问题挖掘

·热点问题的挖掘

·某一时间内群众所反映的问题

先对附件 3 的内容对无实质意义的词进行停用处理，对长句进行有效的分离，将文件内容用函数处理为机器易于理解的形式。再对已经被分割识别的词进行一些停用处理，调用文件 chineseStopWords.txt 将没有太大的实际意义的词组和所有的标点符号还有空格等进行停用。过滤出有用的信息，提高机器对语言文字的处理效率，提高机器对问题的识别正确率。

具体的方法是 def split_word(document):

"""

分词，去除停用词

```
        """
        path = '/文件保存路径 /'
        file = path + 'chineseStopWords.txt'
```

·文本的相似度算法

文本中有一些词是相似但是不完全相同，我们就要计算文本的相似度。利用相似度的矩阵的方式进行相似度对比。要载入一些有关的数据

```
def calculate_similar_matrix(self):
        """
        计算相似度矩阵及一些必要数据
        """
        words = [self.split_word(document) for document in self.documents]

        self.dictionary = corpora.Dictionary(words)
        corpus = [self.dictionary.doc2bow(word) for word in words]
        self.tfidf = models.TfidfModel(corpus)
        corpus_tfidf = self.tfidf[corpus]
        self.similar_matrix = similarities.MatrixSimilarity(corpus_tfidf)
def get_similar(self, document):
        """
        计算要比较的文档与语料库中每篇文档的相似度
        """
        words = self.split_word(document)
        corpus = self.dictionary.doc2bow(words)
        corpus_tfidf = self.tfidf[corpus]
        return self.similar_matrix[corpus_tfidf]
```

使用 aa=[]将计算结果形成一个新的文件夹，

·对留言地的归类，定义热度的指标，评价结果

定义热度指标，将留言详情类似的数目多少，每条留言的点赞数的多少为定义热度的指标。留言数量越多说明民众反应越多热点指数就越高，点赞数越多关注度就越高，问题就学热点。

将附件 3 的每行的点赞数用 excel 降序排序。

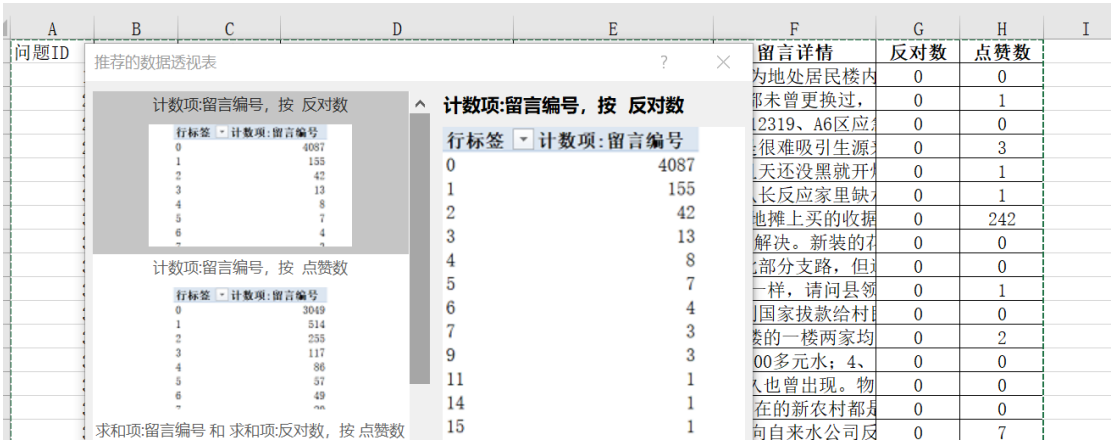
0	0	
5	1762	
0	0	
0	0	
0	0	
0	1	
0	2	
0	2	
0	1	
1	0	
0	0	
0	0	
1	1	
0	0	

- **热度评价**

对以上形成的新文件进行相似度评价后归类相似度大于等于 0.1 的数据标记成同一个标签并将具体的数据进行统计，相似度小于 0.1 标上不相同的标签。相同的标签数总数越多越排前，问题的关注度越高，说明这个问题越热点。具体的操作步骤为

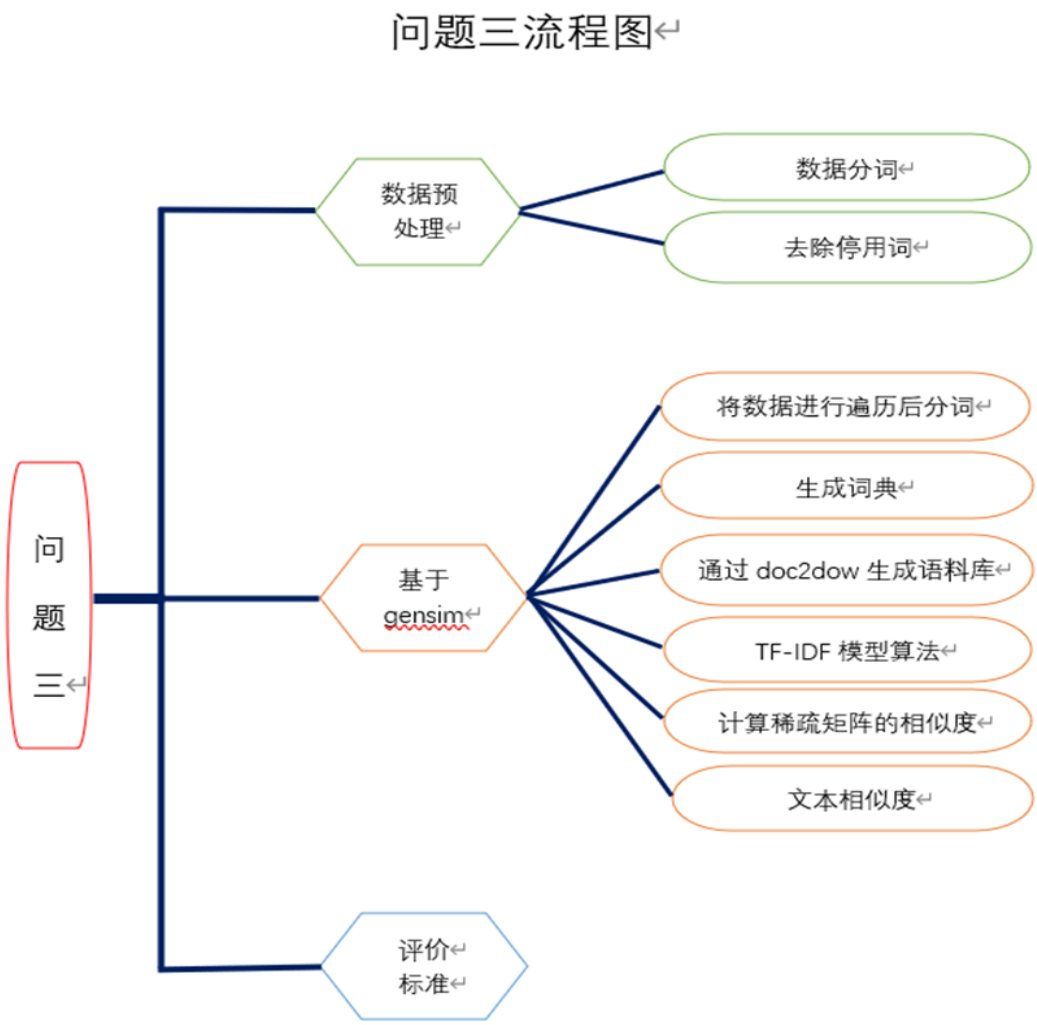
```
19
20
21 ab=[]
22 for i in range(len(aa)):
23     ab.append(aa[i]>0.1)
24
```

我们利用 Excel 对文本进行了筛选得到了热点问题留言明细表及热点问题表



2.3 问题三答复意见的评价

1、 流程图



2、 数据预处理

2.1 数据分词及去停用词

在题目所给出的附件四中，其数据参差不齐，我们在此利用 Python 对其数据进行分词并去停用词，把一些重复无效的数据剔除掉，大大提高了信息的有效性， 我们在这里使用了 chineseStopWords.txt 停用词文本，大幅度提高了其无效信息的存在。

3、 基于 gensim 对答复意见给予评价

a) 将数据进行遍历后分词

通过载入题目所给的附件四中的留言详情和答复意见将其分成分词

列表。

b) 生成词典

通过题目所提供的数据生成词典，并且得到了词典特征数。

c) 通过 doc2dow 稀疏向量生成语料库

存贮于词袋模型中的数据，我们可以将其向量化，减少运算时的复杂程度，一定程度上提高了其运算的效率，生成了 corpus_tfidf 语料库。

d) 通过 TF-IDF 模型算法，计算出 TF 值

TFIDF 的主要思想是：如果一个词或短语经常出现在一篇文章中，而很少出现在其他文章中，则认为该词或短语具有良好的分类能力，适合于分类。我们在给定的文件里，可以把 TF 理解为我们所要求的词语在文档数据中所出现的频率,这个数值是数据归一化后处理的结果,就是对向量的长度进行缩放处理，其全部的元素合计值均等于一，以免其偏向长文本或者是短的文本，在给定的文件中词语 Ti 我们可以通过下式来体现其重要性：

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

IDF 又称为逆文档频率，一定程度上可以将其作为体现一个词语是否具有普遍重要性的标准，但不能仅仅依赖于每个词语出现的频率，有些无效、重复的词语的出现，会对我们的分析造成极大的误差，因此识别出题目所给出的数据中独一无的的特征量是非常重要的，对于数据中一特定词语 IDF 求解方式是，总文件数量÷包含此词语的文件数量，然后把两者相除所得值取对数即为 IDF，具体公式为：

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

则 TF-IDF 的值即为其乘积：

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

e) 计算稀疏矩阵的相似度

```
corpus_tfidf = self.tfidf[corpus]
self.similar_matrix = similarities.MatrixSimilarity(corpus_tfidf)
```

利用以上代码可以将由稀疏向量组成的稀疏矩阵的中数据相似度算出，

f) 文本相似度

在本题中，我们将 Excel 中留言详情和答复意见的数据读成列表，方便我们进行运算，我们采用了两重 for 循环。FOR——ENDFOE 两重循环结构的格式:

语句使用注意事项:

1、内循环为外循环的循环体。

2、内、外循环所使用的变量一般不使用同一变量,以防止出现死循环。

3、内、外循环语句不得交叉使用。

4、当外部循环变量取一次值,内部循环从初值到终值完全执行一次。

正是两重 for 循环让留言详情和答复意见可以循环进行两者之间的相关性的比较，即两者的相似度。

4、 答复意见评价的标准

4.1 通过以上文本的叙述，我们得到了关于留言详情和答复意见的相似度的数据，其为题目所提到的相关性角度，我们根据两者一一对应的相似度数据高低，依次进行排序，排序越靠前的留言详情就说明其答复意见越好，越能解决人们生活中所发生的问题。

4 总结

对于本次泰迪杯竞赛的 C 题，切合了实际，我们在平常政务中文本挖掘众多，我们在基于这样的情况下，分别对本题的一、二、三题做出了回答，让文本挖掘得到了进一步的提升，我们在本次竞赛中收获了团结、不懈，希望我们可以变得更加优秀。

5 未来展望

在未来大时代的发展下，我相信利用网络进行文本分类亦或是热点问题挖掘将

会是一个大的、好的趋势，我相信网络时代将造就更好的世界。

6 参考文献

[1]. 祁小军，兰海翔，卢涵宇，丁蕾锭 ，薛安琪. 贝叶斯、KNN 和 SVM 算法在新闻文本分类中的对比研究[M]. 贵州贵阳: 贵州大学大数据与信息工程学院 550025; 贵州贵阳:贵州力创科技发展有限公司， 550018.
[2].王艺颖. 朴素贝叶斯方法在中文文本分类中的应用[J].中国高新技术，2019:07-057-04。
[3]. 孟天乐. 朴素贝叶斯在文本分类上的应用[M]. 天津市:天津市海河中学，300202.
[4] 熊志斌，刘 冬.
[5]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.