智慧政务中的文本挖掘与分析

摘　要

近年来，随着网络科技的发展，线上问政成为政府了解民意、汇聚民智、凝聚民气的重要渠道。为解决靠人工进行留言划分和热点整理的不便，基于自然语言处理技术的智慧政务系统应运而生。因此，运用自然语言和文本挖掘的方法对留言和热点问题的处理具有重大意义。

对于问题 1，首先查看一级标签和留言详情是否含有空缺值，并统计一下一级标签各类所含的数据数量，接下来我们用图形化的方式再查看一下各个类别的分布。我们将一级标签转换成了 Id(1 到 7),由于我们的留言详情都是中文,所以要对中文进行一些预处理工作,这包括删除文本中的标点符号,特殊符号,还要删除一些无意义的常用词(stopword),因为这些词和符号对系统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,所有在使用这些文本数据之前必须要将它们清理干净。利用 jieba 中文分词工具对留言详情进行分词，并通过 TF-IDF 算法得到每个留言详情的权重向量，接下来用测试数据对不同的模型 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)进行准确率评估，最后建立最适合的线性支持向量机模型进行分类，并使用 F-Score 对分类方法进行评价。

对于问题 2，将留言主题按地区分类，利用 re 模块实现正则表达式匹配，结合 for 语句循环首先筛选出留言主题中只含 A 市的留言主题行号，并将含区，县的留言主题按字典形式将行号和区县内容结合起来,通过字典得到相同区县名称的行号,在 Dataframe 中建立列标题为‘问题 ID’的一列，通过前面筛选的行号进行归类，输入到问题 ID，最后输出 Excel‘热点问题明细表’，然后通过利用 Excel 将点赞数和反对数相加得到一列总票数，将总票数定为热度评价指标并从高到低进行排序，选出热度排行前五的留言主题并列出相应的留言信息输出到 Excel，命名为‘热点问题表’。

对于问题 3，针对相关部门对留言的答复意见，我们采用 LDA 主题模型来评价相关性的。一条留言，首先是以一定的概率选择某个主题的，然后再在这个主题下以一定的概率选出某一个词，这样就生成这条的第一个词，不断重复这个过程，就构成整条留言。首先，选择一个要求的参数，其次选择一个主题第 m 个留言被赋予的第 n 个词被赋予的

主题，然后生成第 m 条留言的第 n 个词，最后推导出主题分布。针对可解释性，从大规模的回复中提取特征和情感，可以从评论回复中提取出“特征-观点-情感”。

关键词：中文分词，去掉停用词，线性支持向量机，TF-IDF 算法，f-score 评价，热点问题挖掘

Text mining and analysis in smart government

Abstract

In recent years, with the development of network technology, online political inquiry has become an important channel for the government to understand public opinion, gather people's wisdom and gather people's morale. In order to solve the inconvenience of manual message division and hot spot sorting, intelligent government system based on natural language processing technology came into being. Therefore, it is of great significance to use natural language and text mining methods to deal with messages and hot issues.

For question 1, first check whether the first level tags and message details contain vacancy values, and count the data quantity of each type of first level tags. Next, we can check the distribution of each category in a graphical way. We have converted the first level label to ID (1 to 7). Since our message details are all in Chinese, we need to do some preprocessing work for Chinese, including deleting punctuation marks, special symbols and meaningless common words (stopword), because these words and symbols are not helpful for system analysis and prediction of text content, On the contrary, it will increase the complexity of calculation and the overhead of the system. All these text data must be cleaned up before use. The Chinese word segmentation tool of Jieba is used to segment the message details, and the weight vector of each message details is obtained by TF-IDF algorithm. Next, the test data are used to analyze the different models: logistic regression, polynomial naive Bayes, linear support vector machine, random. Finally, the most suitable linear support vector machine model is established for classification, and F-score is used to evaluate the classification method.

For question 2, classify the message subject by region, use re module to achieve regular expression matching, combine for statement cycle to first filter out the message subject line number which only contains city a in the message subject, and combine the message subject which contains district and county with the content of district and county in the form of dictionary, get the line number of the same district and county name through the dictionary, and establish the column title in dataframe as‘ The column of "question Id" will be classified by the row number screened in the front, input to the question Id, and finally output the excel "list of hot issues". Then, the total number of votes will be obtained by adding the number of likes and objections in Excel. The total number of votes will be set as the heat evaluation index and sorted from high to low. The top five message topics in the heat ranking will be

selected and the corresponding message information will be listed Go to excel and name it "hot issues table"

For question 3, we use the LDA theme model to evaluate the relevance of the comments from relevant departments. For a message, first select a topic with a certain probability, and then select a word with a certain probability under the topic, so as to generate the first word of the message, repeat the process continuously, and form the whole message. First, select a required parameter, then select a topic to which the nth word of the m message is assigned, and then generate the nth word of the m message, and finally derive the topic distribution.

Keywords: Chinese word segmentation, elimination of stop words, linear support vector machine, TF IDF algorithm, F-score evaluation, hot issues mining, LDA model.For interpretability, features and emotions are extracted from large-scale responses, and "feature-view-emotion" can be extracted from comment responses.

Keywords: Chinese word segmentation, removal of stop words, linear support vector machine, TF-IDF algorithm, f-score evaluation, hot spot problem mining

目 录

1. 挖掘目标

本次建模目标是利用数据挖掘技术,利用 jieba 中文分词工具对留言详情进行分词，线性支持向量机算法，re 模块的正则匹配，达到以下三个目标：

1）利用文本分词和中文文本多分类方法对非结构化的数据进行文本挖掘，解决大部分电子政务系统还是依靠人工根据经验处理所存在的工作量大、效率低，且差错率高等问题，并进行群众留言分类。

2）根据提供的留言数据及时发现热点问题，并将某一时段内反映特定地点的留言进行归类，定义合理的热度评价指标，并给出评价结果。

3)针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实行。

2. 分析方法与过程

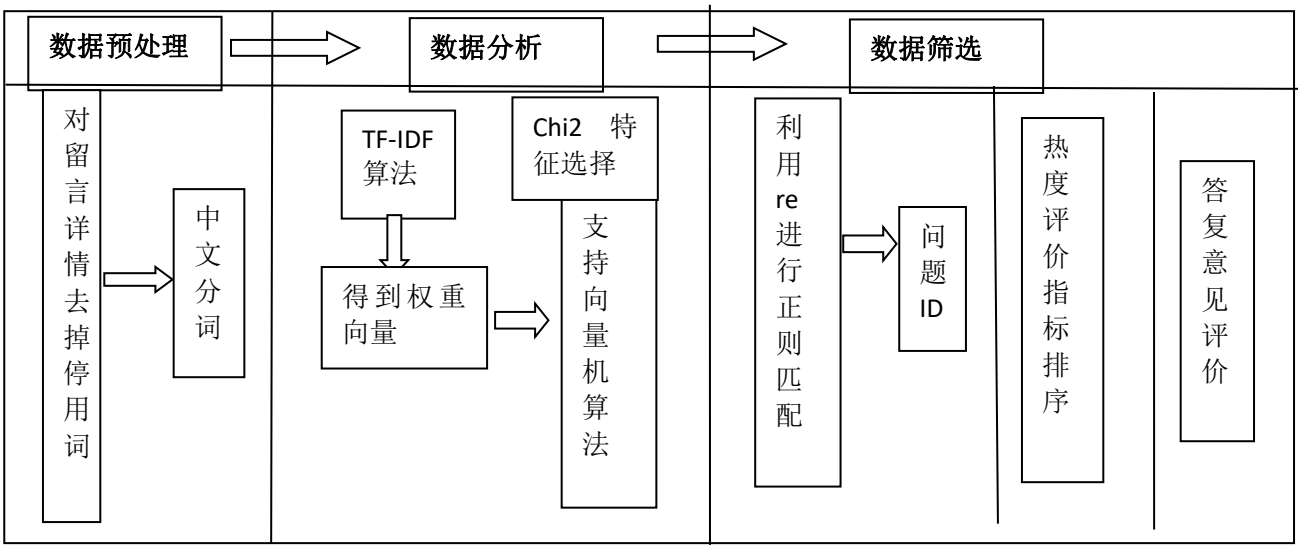

图 1：总体流程图

本用例主要包括如下过程：

步骤一：数据预处理，在原始的数据基础上，对留言详情去掉停用词，这包括删除文本中的标点符号，特殊符号，包括一些无意义的常用词，在此基础上进行中文分词。

步骤二：数据分析，利用 jieba 中文分词工具对留言详情进行分词，并通过 TF-IDF 算法得到每个留言详情的权重向量，以供挖掘分析使用。接下来用测试数据对不同的模型进行准确率评估，最后建立最适合的线性支持向量机模型进行分类，并使用 F-Score 对分类方法进行评价。

步骤三：数据筛选，利用 re 模块实现正则表达式匹配，并列出相应的留言信息输出

到 Excel。

步骤四：针对相关部门对留言的答复意见，主要从答复的相关性的角度，建立 LDA 模型对答复意见的质量给出一套评价方案。

2.1 数据预处理

2.1.1

（1）留言详情和一级标签无缺失值

（2）各类别总数

	一级标签	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 2：一级标签

（3）各类别分布图

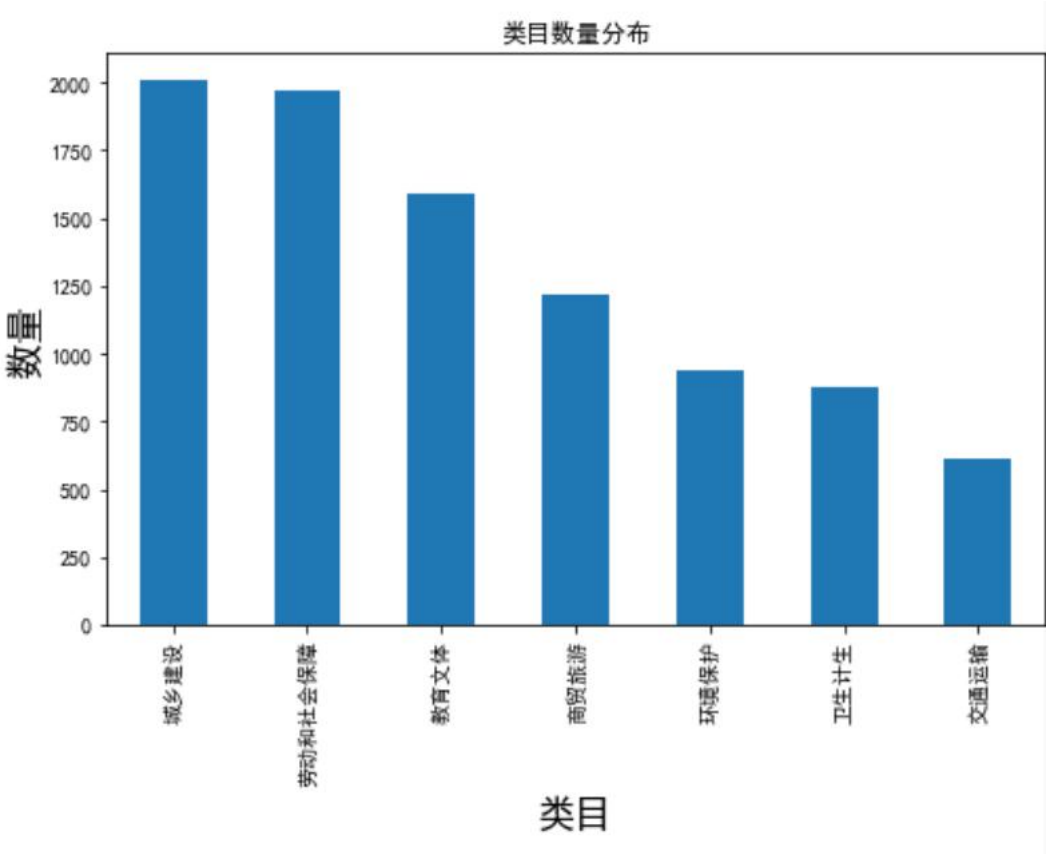

图 3：各类别分布图形

（4）在对留言详情去除停用词，中文停用词包含了很多日常使用频率很高的常用词,如 吧，吗，呢，啥等一些感叹词等,这些高频常用词无法反应出文本的主要意思,所以要被过滤掉。

2.1.2 对留言详情进行中文分词

在对留言详情进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件留言表中，以中文文本的方式给出了数据。为了便于转换，先要对这些留言主题进行中文分词。我们选择采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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）。在分词的同时，采用了 TF-IDF 算法，将留言详情转换成数字进行处理。

2.1.3 TF-IDF 算法

在对留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把留言信息转换为权重向量。TF-IDF（term frequency–inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF 意思是词频(Term Frequency)，IDF 意思是逆文本频率指数(Inverse Document Frequency)。TF-IDF 是在单词计数的基础上，降低了常用高频词的权重,增加罕见词的权重。因为罕见词更能表达文章的主题思想,比如在一篇文章中出现了“中国”和“卷积神经网络”两个词,那么后者将更能体现文章的主题思想,而前者是常见的高频词,它不能表达文章的主题思想。所以“卷积神经网络”的 TF-IDF 值要高于“中国”的 TF-IDF 值。这里我们会使用 sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取文本的 TF-IDF 的特征值。这里我们使用了参数 ngram_range=(1,2),这表示我们除了抽取评论中的每个词语外,还要抽取每个词相邻的词并组成一个“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准确度。

TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）

词频（TF）=某个词在文本中出现的次数

考虑到文章有长短之分，为了便于不用文章的比较，进行“词频”标准化，除以文

本的总词数或者除以该文本中出现次数最多的词的出现次数即：

词频（TF）=某个词在文本中的出现次数/文本的总词数

或词频（TF）=某个词在文本中的出现次数/该文本出现次数最多的词的出现次数

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

逆文档频率（IDF）=log（语料库的文本总数/（包含该词的文本数+1））

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）*逆文档频率（IDF）

实际分析得到 TF-IDF 值与一个词在留言信息表中文本出现的次数成正比，某个词的文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的留言信息表中文本的关键词。

2.1.4 生成 TF-IDF 向量

生成各个留言主题的 TF-IDF 权重向量，计算公式如下：

TF - IDF = 词频（TF）*逆文档频率（IDF）

2.1.5 特征选择

我们使用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。在这里我们使用 sklearn 中的 chi2 方法。

主要公式：

$$\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}\sim\chi^2$$

通过卡方检验得到每类关联度最大的词语和词语对

```
# '城乡建设':
  . Most correlated unigrams:
       . 小区
       . 业主
  . Most correlated bigrams:
       . 住房 公积金
       . 公积金 贷款
# '环境保护':
  . Most correlated unigrams:
       . 环保局
       . 污染
  . Most correlated bigrams:
       . 周边 居民
       . 环保局 领导
# '交通运输':
  . Most correlated unigrams:
       . 快递
       . 出租车
  . Most correlated bigrams:
       . 的士 司机
       . 出租车 司机
```

图 4.1：每类关联度最大的词语和词语对

```
# '教育文体':
  . Most correlated unigrams:
       . 学生
       . 学校
  . Most correlated bigrams:
       . 教育局 领导
       . 培训 机构
# '劳动和社会保障':
  . Most correlated unigrams:
       . 退休
       . 社保
  . Most correlated bigrams:
       . 劳动 关系
       . 退休 人员
# '商贸旅游':
  . Most correlated unigrams:
       . 传销
       . 电梯
  . Most correlated bigrams:
       . 小区 电梯
       . 传销 组织
# '卫生计生':
  . Most correlated unigrams:
       . 医生
       . 医院
  . Most correlated bigrams:
       . 社会 抚养费
       . 乡村 医生
```

图 4.2 每类关联度最大的词语和词语对

2.1.6 模型选择

首先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)

利用编程画出四种模型的箱体图

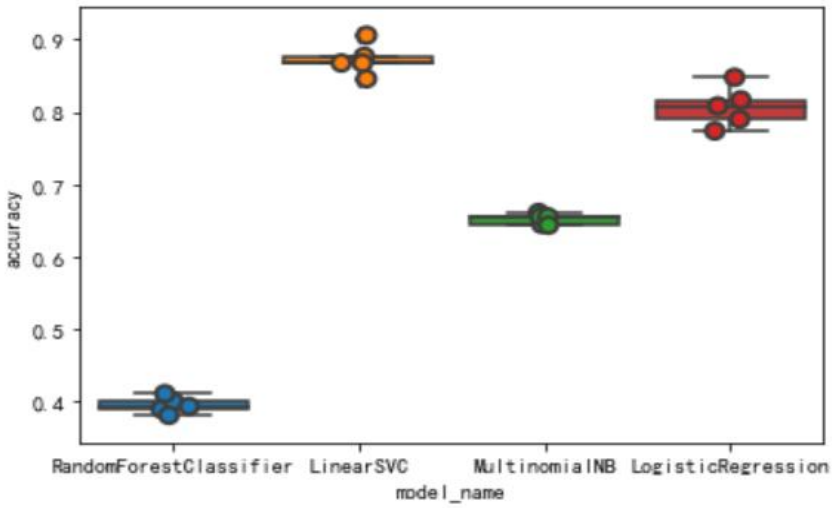

图 5：四种模型的箱体图

可以箱体图上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据),因为文本数据有太多的特征值,使得集成分类器难以应付,另外三个分类器的平均准确率都在 80%以上。其中线性支持向量机的准确率最高。

```
]: model_name
   LinearSVC                0.872863
   LogisticRegression       0.807385
   MultinomialNB            0.652005
   RandomForestClassifier   0.394889
   Name: accuracy, dtype: float64
```

图 6：四种模型的准确率

2.1.7 利用线性支持向量机进行一级标签分类

SVM 的算法过程总结：

输入 m 个样本(x1,y1),(x2,y2),...,(xm,ym),(x1,y1),(x2,y2),...,(xm,ym),其中 x 为 n 维特征向量。y 为二元输出，值为 1，或者-1.

输出是分离超平面的参数 W*和 b*和分类决策函数。

算法过程如下：

1）选择适当的核函数 K(x,z)和一个惩罚系数 C>0C>0，构造约束优化问题

$$\underbrace{min}_{\alpha}\frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{m}\alpha_i$$

$$s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0$$

$$0\le\alpha_i\le C$$

2）用 SMO 算法求出上式最小值时对应的α向量的值α^*向量.

3）得到 $w^*=\sum_{i=1}^{m}\alpha_i^* y_i\phi(x_i)$，此处可以不直接显式的计算$w^*$。

4）找出所有的S个支持向量,即满足$0<\alpha_s<C$对应的样本(x_s,y_s)，通过 $y_s(\sum_{i=1}^{m}\alpha_i y_i K(x_i,x_s)+b)=1$，计算出每个支持向量$(x_s,y_s)$对应的$b_s^*$,计算出这些 $b_s^*=y_s-\sum_{i=1}^{m}\alpha_i y_i K(x_i,x_s)$. 所有的 b_s^* 对应的平均值即为最终的 $b^*=\frac{1}{S}\sum_{i=1}^{S}b_s^*$

这样最终的分类超平面为：$\sum_{i=1}^{m}\alpha_i^* y_i K(x,x_i)+b^*=0$，最终的分类决策函数为：$f(x)=sign(\sum_{i=1}^{m}\alpha_i^* y_i K(x,x_i)+b^*)$

2.1.8 F—score 模型评估

下面我们就准确率最高的 LinearSVC 模型，查看其**混淆矩阵，**

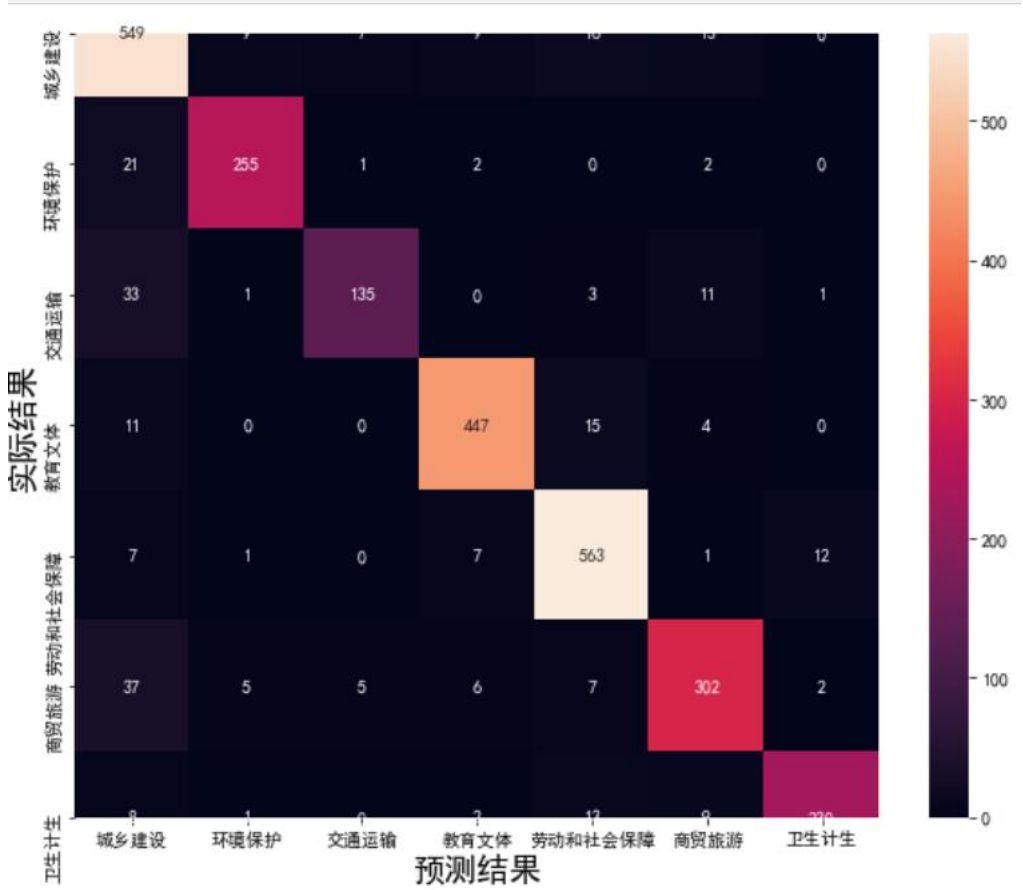

图 7：linearSVC 模型的混淆矩阵

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. 。

多分类模型一般不使用准确率(accuracy)来评估模型的质量,因为 accuracy 不能反应出每一个分类的准确性,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 分数来评估模型。

F1 分数（F1 Score），是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率。F1 分数可以看作是模型精确率和召回率的一种加权平均，它的最大值是 1，最小值是 0。

精准度 / 查准率(precision)：指被分类器判定正例中的正样本的比重

$$精确率\ P = \frac{TP\,真阳性}{TP\,真阳性 + FP\,假阳性}$$

召回率 / 查全率 (recall)：指的是被预测为正例的占总的正例的比重

$$召回率\ R = \frac{TP\,真阳性}{TP\,真阳性 + FN\,假阴性}$$

数学定义：F1 分数（F1-Score），又称为平衡 F 分数（BalancedScore），它被定义为精确率和召回率的调和平均数。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precison + recall}$$

score 模型的具体原理如下：

一般情况下，为了综合权衡召回率和精确率，就引入了一个新的指标 F-score。这是综合考虑 Precision 和 Recall 的调和值。

其中 Pi 为第 i 类的查准率， Ri 为第 i 类的查全率。

Precision: 0.900
Recall: 0.898
F1: 0.898

	precision	recall	f1-score	support
城乡建设	0.82	0.91	0.87	603
环境保护	0.94	0.91	0.92	281
交通运输	0.91	0.73	0.81	184
教育文体	0.95	0.94	0.94	477
劳动和社会保障	0.91	0.95	0.93	591
商贸旅游	0.88	0.83	0.86	364
卫生计生	0.94	0.87	0.91	263
accuracy			0.90	2763
macro avg	0.91	0.88	0.89	2763
weighted avg	0.90	0.90	0.90	2763

图 8：score 模型测评结果

由上图我们得到测试集准确率为 0.9，召回率为 0.898，F1—score 为 0.98

还显示了各分类的准确率，召回率，F1—score，和分对的样本数量，以及宏平均值：macro average，所有标签结果的平均值

加权平均值：weighted average，所有标签结果的加权平均值

从以上 F1 分数上看,"教育文体"类的 F1 分数最大，“交通运输”类 F1 分数最差，究其原因可能是因为“交通运输”分类的训练数据最少只有 613 条,使得模型学习的不够充分,导致预测失误较多。

2.2 问题 2 分析方法与过程

2.2.1 数据筛选

（1）根据附件 3 将留言主题按地区分类，利用 re 模块实现正则表达式匹配，结合 for 语句循环首先筛选出留言主题中只含 A 市的留言主题行号

（2）在 for 循环中，将含区，县的留言主题按字典形式将行号和区县内容结合起来，并放入空字典中，通过字典得到相同区县名称的行号（同 values 的 key 值）

（3）在 Dataframe 中建立列标题为‘问题 ID’的一列，并将 A 市的问题 ID 赋值为 1，并且把不含任何市，区，县的留言主题的行的问题 ID 也当做 A 市来分类，赋值为 1，其余各不同区县的问题 ID 从 2 开始赋值依次类推，最后输出 Excel ‘热点问题明细表’

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	239811	A0006491	A市通商食	2019/3/15	本人于201	0	0
1	246035	A0005830	A市丁字湾	2019/10/3	丁字湾中国	0	0
1	246093	A00057	西地省师范	2019/11/1	11月10日凌	0	0
1	246325	A0008393	A市柠檬丽	2019/6/24	A市柠檬丽	0	0
1	246329	A0008352	A市南站候	2019/7/2 9	A市南站作	0	0
1	246362	A909114	A市魅力之	2019/08/2(	2019年5月	0	0
1	246366	A0002226	A市麻园湾	2019/4/15	本人是A市	0	0
1	246384	A0005258	建议对A市	2019/5/29	目前三一大	0	0
1	246407	A0009959	举报广铁集	########	我要举报广	0	0
1	246019	A0009369	A市365公交	2019/1/18	365公交车	0	0
1	246428	A0002624	A市愿景山	2019/11/1:	我们的房子	0	0
1	246528	A0006404	咨询A市军	2019/12/1(	今年我三期	0	0
1	246637	A0001039	A市明发国	2019/4/17	本人去年8	0	0
1	246662	A0009310	反映北京师	2019/8/19	尊敬的县领	0	0
1	246708	A0001039	A市海德公	2019/1/3 1	星沙海德公	0	0
1	246771	A0004912	投诉A市先	2019/8/23	我家住在A	0	0
1	246785	A0001592	不需要车位	########	我们是还没	0	0
1	246891	A0001128	A市一师三	2019/10/1	您好！希望	0	0
1	247052	A0008633	A市火车南	2019/6/30	6月30日上	0	0
1	246518	A0002352	A市西湖公	2019/4/14	西湖文化公	0	0
1	247095	A0009688	A市润和山	2019/12/2:	尊敬的西地	0	0
1	246011	A0004538	建议A市1	2019/7/9 1	A市碧沙湖	0	0
1	245780	A0001067	反映A市丁	2019/9/2 2	1:丁字湾属	0	1
1	244951	A0002603	反映A市人	2019/7/7 1	A市人才租	0	0

Sheet1

表 1 热点问题明细表 1

（4）根据附件 4，通过利用 Excel 将点赞数和反对数相加得到一列总票数，然后删掉'留言编号','留言用户','反对数','点赞数'为列名的列，修改列名称，将'留言主题'改为'地点/人群', '留言时间'改为'时间范围', '留言详情'改为'问题描述','总票数'改为'热度指数'，将总票数定为热度评价指标并从高

到低进行排序，选出热度排行前五的留言主题并列出相应的留言信息输出到 Excel，命名为‘热点问题表’

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	11	2097	2019/8/19	A市A5区汇	我是A市A5区汇金路五矿万境
2	1	1767	2019/4/11	反映A市金	书记先生：您好！我是梅溪湖
3	6	821	2019/2/21	请书记关注	尊敬的胡书记：您好!A4区p2
4	1	790	2019/2/25	严惩A市58	胡市长：您好！西地省展星扌
5	1	733	2019/3/1 2	承办A市58	胡书记：您好！58车贷案发，

表 2：热点问题明细表 2

2.3 问题 3 问题分析方法与过程

2.3.1 问题分析

针对相关部门对留言的答复意见，我们采用 LDA 主题模型来评价相关性，从而构建答复质量测评的方案。针对评论回复的解释性可以分为两个层面：一个是特征层面，一个是句子层面。从大规模的回复中提取特征和情感，可以从评论回复中提取出“特征-观点-情感”。针对评论回复的解释性可以分为两个层面：一个是特征层面，一个是句子层面。基于句子层面：解释是完整的，语义是连贯的。句子可以通过模板生成（template）,也可以利用自然语言处理技术生成。（LSTM,GRU）

2.3.2 LDA 模型

一篇留言的构造过程，一条留言，首先是以一定的概率选择某个主题的，然后再在这个主题下以一定的概率选出某一个词，这样就生成这条的第一个词，不断重复这个过程，就构成整条留言。首先，选择一个要求的参数，其次选择一个主题第 m 个留言被赋予的第 n 个词被赋予的主题，然后生成第 m 条留言的第 n 个词，最后推导出主题分布。

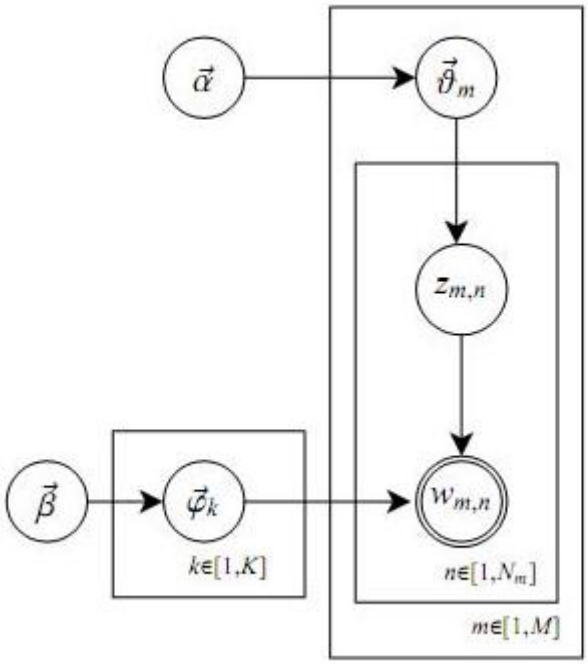

图 9：形式化 LDA

符号	解释
M	文章的数量
K	主题的个数
V	词袋的长度
Nm	第m篇文章中单词的总数
$\vec{\alpha}$	是每篇文章的主题分布的先验分布Dirichlet分布的参数（也被称为超参数）通常是手动设定的
$\vec{\beta}$	是每个主题的词分布的先验分布Dirichlet分布的参数（也被称为超参数）通常是手动设定的
$\vec{\vartheta}_m$	θ是一个M*K的矩阵,$\vec{\vartheta}_m$表示第m篇文章的主题分布，$\vec{\vartheta}_m \sim Dir(\vec{\alpha})$是我们要求的参数
$\vec{\varphi}_k$	φ是一个K*V的矩阵,$\vec{\varphi}_k$表示第k个主题的词分布，$\vec{\varphi}_k \sim Dir(\vec{\beta})$是我们要求的参数
$z_{m,n}$	第m篇文章第n个词被赋予的主题，隐变量
$w_{m,n}$	第m篇文章第n个词，这个是可以被我们观测到的

表 3：符号解释表

选择一个ϑ⃗ m~Dir(α⃗)ϑ→m~Dir(α→)

对每个准备生成的单词 Wm,n

(a)选择一个主题 Zm,n ~Multinomial(ϑ⃗ m)

(b)生成一个单词 Wm,n 从 $P(W_{m,n}|Z_{m,n},\vec{\beta})$

根据上节的讨论，可以推导出主题分布：

$$p(\vec{\mathbf{z}}|\vec{\alpha}) = \prod_{m=1}^{M} p(\vec{z}_m|\vec{\alpha}) = \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

词分布有类似的形式:

$$p(\vec{\mathbf{w}}, \vec{\mathbf{z}}|\vec{\alpha}, \vec{\beta}) = p(\vec{\mathbf{w}}|\vec{\mathbf{z}}, \vec{\beta})p(\vec{\mathbf{z}}|\vec{\alpha}) = \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

最后，可以写出资料库的联合概率。

这里参数转化为每个主题的单词计数以及词分布的单词计数。概率将有这些计数除它们对应的总数所得。

3. 参考文献

【1】王千，王成，冯振元、叶金凤，k-meas 聚类算法研究综述.2012

[2]基于支持向量机的计算机键盘用户身份验真【J】.刘学军，陈松灿，彭宏京。