

“智慧政务”中的文本挖掘应用

摘 要

近年来，随着科技进步，互联网已逐渐成为群众发表意见、提出诉求的重要渠道，网络民意已成为人民群众诉求的重要组成部分。由于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。因此运用文本分析和数据挖掘技术对群众留言分类和热点整理的研究具有重大的意义。

对于群众留言分类问题：通过对数据可视化，我们发现该数据属于不平衡文本的多分类问题。我们首先从特征值的选择和分类器的选择两部分出发，特征值采取的 TF-IDF 的特征，通过 F-Score 作为评估指标，比较了逻辑回归、朴素贝叶斯、线性支持向量机、决策树等分类器；并且我们还使用了 LSTM (Long Short-Term Memory) 长短期记忆模型来进行分类，最后得出线性支持向量机具有最高的分类指标，并且通过绘制混淆矩阵热力图进行了结果展示。

对于热点问题挖掘问题：我们比较了 LDA 和 LDA Mallet 这两个模型，并计算出来二者的困惑度和一致性分数，根据结果选择用 LDA Mallet 模型。计算不同主题数的一致性得分，得出最佳主题数为 20 个。找出 20 个主题中所包含的关键词，并用 pyLDAvis 软件包对其进行可视化。将点赞和反对都当作文档加一处理，并生成热度图，计算出本文所需要的热度指数对其进行排序。当热度指数求出后，找出排名前五的主题及其对应的留言，就可以导出该问题所需要的文档。

对于答复意见的评价问题：我们建立基本特征工程来计算留言详情和答复意见中的词数，并且计算两个问题的常用单词数。使用 CountVectorizer 函数来构建词袋模型。然后生成 TF-IDF 向量，再使用 SVD 进行降维处理。最后使用 fuzzywuzzy 来匹配模糊字符串，fuzzywuzz 使用不同比率来描述留言详情和答复评价之间的相似程度。

关键词：数据挖掘；TF-IDF；线性支持向量机；LDA Mallet 模型

Abstract

In recent years, with the progress of science and technology, the Internet has gradually become an important channel for the people to express their opinions and make appeals, and online public opinion has become an important part of the people's appeals.

For the problem of the crowd message classification: we found that the data was divided into the multi-fractal problem of the uneven text. We first chose two points of the choice and classifier of the eigenvalues to select the characteristics of the TF-IDF, which was used by the characteristic value, and compared the logical regression, the simple beeses, the linear support vector machine, the decision tree equidivider. And we also use the LSTM(Long Short-Term Memory) long term memory model to classify, and finally obtained the highest classification index for linear support vector devices, and the results of the completion of the matrix heat diagram were drawn by drawing the confusion matrix.

For hot-spot problem mining: we compared two models, LDA and LDA Mallet, and calculated their confusions and consistency scores. According to the results, we chose the LDA Mallet model. The consistency score of the number of different topics was calculated, and the best topic number was 20. Identify the keywords contained in the 20 topics and visualize them using the pyLDavis software package. Treat both thumb up and opposition as documents plus one, and generate a heat map to calculate the heat index needed for this article to sort them. After the heat index is calculated, find the top five topics and their corresponding comments, and then export the documents needed for the problem.

Evaluation questions for response comments: we built the basic features project to calculate the message details and the number of words in the response comments, and to calculate the number of commonly used words for both questions. The CountVectorizer function is used to construct the word bag model. Then generate tf-idf vector, and use SVD for dimension reduction. Finally, fuzzywuzz is used to match fuzzy strings, and fuzzywuzz used different rates to describe the degree of similarity between message details and response ratings.

Keywords: Data mining; TF - IDF; Linear support vector machine;

目 录

第 1 章 问题分析	4
1.1 群众留言分类问题分析.....	4
1.2 热点问题挖掘问题分析.....	4
1.3 答复意见的评价问题分析	4
第 2 章 问题求解	5
2.1 群众留言分类问题求解.....	5
2.1.1 整体流程.....	5
2.1.2 具体步骤.....	6
2.1.3 结果分析.....	14
2.2 热点问题挖掘问题求解.....	15
2.2.1 整体流程.....	15
2.2.2 具体步骤.....	16
2.2.3 结果分析.....	21
2.3 答复意见的评价问题求解	22
2.3.1 整体流程.....	22
2.3.2 具体步骤.....	23
2.3.3 结果分析.....	27
第 3 章 结论	28
参考文献	29

第 1 章 问题分析

1.1 群众留言分类问题分析

工作人员在处理网络问政平台的群众留言时，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到相应的部门来处理。目前，大部分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。

群众留言分类研究是一个多分类问题。多分类是指将文本分成若干个类中的某一个类。我们尝试使用了不同的机器学习模型：逻辑回归、多项式朴素贝叶斯、线性支持向量机、随机森林。并借助于 F1 分数、ROC 等指标来评估模型的准确率。

1.2 热点问题挖掘问题分析

在某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。热点问题挖掘是对一定的网络数据源来进行分析，综合使用统计、聚类等方法识别出被群众广泛讨论的热点。热点问题挖掘在一定程度上可以帮助政府迅速了解当前社会群众关注的热点，也能帮助政府的相关部门能针对性的处理问题，提高政府的服务效率。

在本文的热点问题挖掘研究中，留言中的特征词是整个模型中唯一的可观察变量，LDA 模型在已知主题数目的情况下，通过调节特征词语在主题上的概率分布完成每篇留言的生成过程。

1.3 答复意见的评价问题分析

如今很多政府部门设立的网络问政平台存在着万能回复，比如：“网友，你好！你反映的问题已在办理中，请等待回复。”这种互动没有真正地解决群众的需求，降低了政府的公信力。同时也让网站成为了“僵尸网站”，浪费了社会资源。对答复意见的评价可以督促政府相关部门按时并保证高质量得进行回复，争取能为群众解决实际问题。

答复意见的评价问题，本文研究了答复意见与留言详情的相观性。通过建立基本特征工程来丰富我们的词向量表达。Fuzzywuzzy 的不同比率来描述两个字符串之间的相似程度，用问题详情和答复意见构建 LDA 模型，计算两者之间的文本相似度，这样就能比较出留言详情和答复意见的相关程度。

第 2 章 问题求解

2.1 群众留言分类问题求解

2.1.1 整体流程

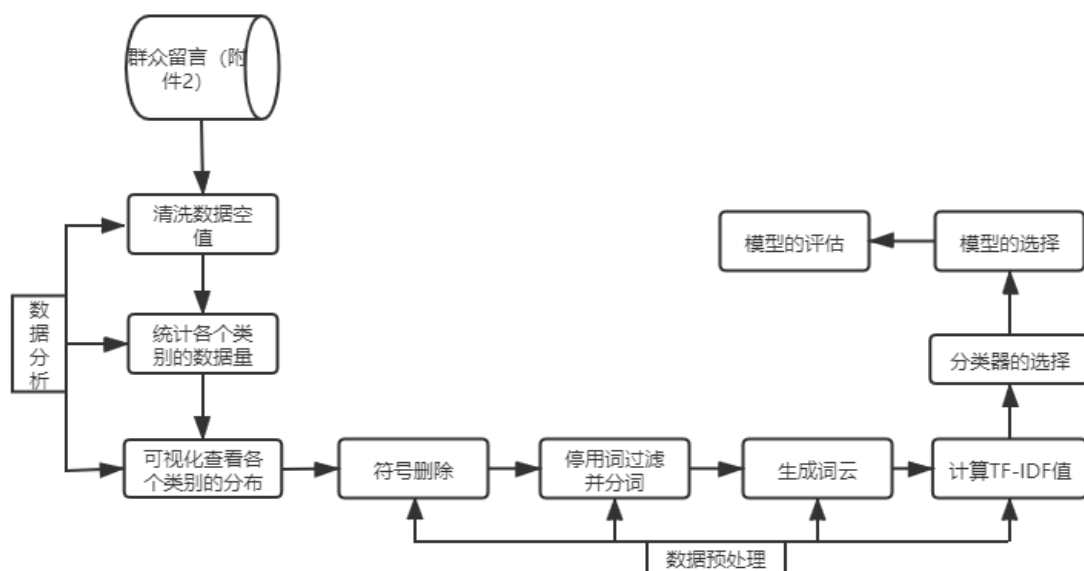


图 2-1 整体流程图

该问题主要包括以下几个步骤：

- 步骤一：在数据挖掘挑战赛的官网上可以下载本问题的全部数据，即可得到该问题实验所需的数据。
- 步骤二：数据分析，首先清除数据中的空值，然后统计各个类别的数据量，最后以图形化的方式再查看一下各个类别的分布。
- 步骤三：数据预处理，第一步删除除了汉字以外的所有符号；第二步进行停用词过滤并进行分词，去掉留言中大部分日常生活中使用频率很高的常用词，并将一句评论分成多个词语进行分析；第三步生成前 100 个高频词的词云；第四步，对词进行向量化，并计算 TF-IDF 值。
- 步骤四：分类器的选择，对于该问题，本文选取线性支持向量机分类器，使用已有的词向量来训练该分类器。
- 步骤五：模型的选择，本文尝试不同的机器学习模型，并评估每个模型的准确率，以此来选择合适的模型。
- 步骤六：模型的评估，根据平均准确率最高来选取模型之后，再查看混淆矩阵，并显示预测标签和实际标签的差异。由于多分类模型一般不使用准确率

来评估模型的质量,因此本文通过查看各个类的 F1 分数来对模型进行评估。

2.1.2 具体步骤

步骤一：获取实验的数据

打开数据挖掘挑战赛的官网,找到如下图 2-2 所示内容,点击下载 C 题全部数据,即可得到该问题实验所需的全部数据。

[点击下载C题赛题](#)
[点击下载C题示例数据 \(提取码: nuhu\)](#)
[点击下载C题全部数据 \(提取码: v2ot\)](#)

图 2-2 全部数据下载地址

步骤二：数据分析

(1) 清除数据中的空值

对于获取到的数据,每行代表了一个留言文本但是其中难免会出现一些空行。所以本文首先进行了“去空”的预处理操作。

在导入留言文本时,同时输出了多少空值,并且进行“去空”处理,“去空”结果如图 2-3 所示:

在 留言详情 列中总共有 0 个空值.
在 一级标签 列中总共有 0 个空值.

图 2-3 “去空”结果

(2) 统计各个类别的数据量

由于留言文本的数目比较多,将文本分类的同时,统计出各个类别的数量,统计结果如图 2-4 所示:

	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 2-4 统计各个类别数据量

(3) 可视化各个类别的数据量

各个类别的数据量非常不一致,其中城乡建设、劳动和社会保障、教育文体、商贸旅游的数据量已经超过 1000 条,而环境保护、卫生计生、交通运输的数据量则在 1000 条以下,分布不是很均匀,本文用图形化的方式来更直观的查看一下各个类别的分布,结果如图 2-5 所示:

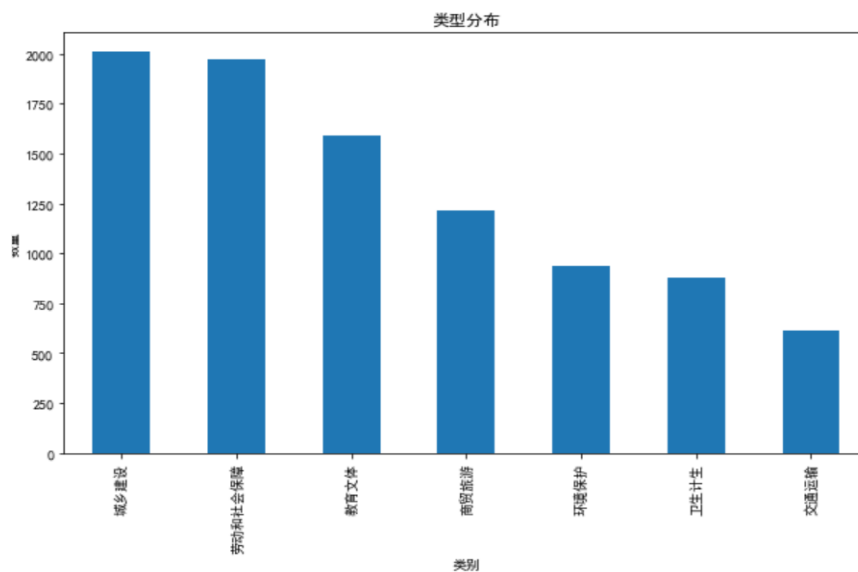


图 2-5 可视化各个类别的数据量

步骤三：数据预处理

(1) 符号删除

留言文本中有许多标点符号,这些标点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,所以必须将标点符号给清

除，符号删除的结果如图 2-6 所示：

		留言详情	一级标签
1628	西地省住房和城乡建设厅现根据中华人民共和国建筑法建筑工程质量管理条例向贵单位实名举报县泰鸿置...	城乡建设	
9017	西地省市人民医院一个小小的介入微创手术要了患者的命年仅岁的徐望花乐观坚强善良能吃能喝能跳舞就...	卫生计生	
7942	市电力公司对面对中粮商行挂羊头卖狗肉利用工商局的保护经常不断欺骗老年人卖保健品难道就没有人管难...	商贸旅游	
1752	年全款购得房产至今无法办理房屋所有权证我们是年月在西地省市市生活湾置业有限公司交付了万元的购...	城乡建设	
7703	自今年月以来县屠宰场猪贩多人组织到一起经多次开会商议分工明确确定负责人统一采购定价收款获取高...	商贸旅游	
706	县北门市场西侧正在建造楼房该楼离老烟草局的住房最近点不到五米将严重影响老烟草局住房的采光与通...	城乡建设	
4534	尊敬的市委书记您好我是星沙松雅湖小学原名华润小学一名学生的家长请求政府能够协调县教育局积极妥...	教育文体	
2814	请关注一下市市川山坪等乡镇的麻石开采切割和再加工生产近年来麻石生产使得川山坪等乡镇污水横流空...	环境保护	
4876	市楚一培训学校实质是在职老师贺勇强主办的高额收费培训班学校市第一中学在职教师贺勇强借助一中教...	教育文体	
8192	局长您好我是美林银谷的业主美林银谷小区的电梯号称是日立电梯存在安全隐患我怀疑这个是伪劣产品理...	商贸旅游	

图 2-6 符号删除的结果

(2) 停用词过滤并分词

中文分词(Chinese Word Segmentation)，也可称为中文切词，指的是通过某种特定的规则，将中文文本切分成一个一个单独的词，之后进行分析。

留言中包含很多日常使用频率很高的常用词，需要将它们过滤掉，否则将会影响下文的分析的正确率。在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言之前会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。

本文将停用词过滤和中文分词一起处理，处理后的结果如图 2-7 所示：

	留言详情	一级标签	label_id	cut_notes
8304	县御景中央电梯经常掉层为何没人管这两天又发生了但是发生以后物业相关部门并没及时检修处理屢屢发...	商贸旅游	5	县御景中央电梯经常掉层没人管两天发生发生以后物业相关部门没及时...
5851	县领导你好我叫李三年月我到城北中学打工现更名深柳中学既做白案也炒菜年聂校长表扬食堂员工对老师...	劳动和社会保障	4	县领导你好李三年月城北中学打工现更名深柳中学做白案炒菜年聂...
5855	胡厅长您好西地省人事考试网是贵部门的一个服务窗口以往都在考后及时提供各项专业资格考试的成续查...	劳动和社会保障	4	胡厅长您好西地省人事考试网贵部门服务窗口以往都考后及时提供各项...
872	尊敬的市领导你们好之前是在县住房公积金贷款可是县住房公积金管理部说没钱放贷但是购房期限已到...	城乡建设	0	尊敬市领导你们好之前想县住房公积金贷款县住房公积金管理部说没...
4080	杨书记你好我们是县当年被辞退的民办代课教师及幼师早在年份时都已经填表参加调查摸底至今将近三...	教育文体	3	杨书记你好县当年辞退民办代课教师幼师早年份时都已经填表参...
3243	年月日点分左右本人从城南加油站乘坐一辆出租车到城北汽车站司机问到哪里我说到北站然后司机张嘴说...	交通运输	2	年月日点分左右本人城南加油站乘坐一辆出租车城北汽车站司机问说北站...
8785	省生计委你们是不是只管发文不管事原来省计生委规定独生子女补贴从退休之日起而你们确定从元年元月...	卫生计生	6	省生计委是不是只管发文事原来省计生委规定独生子女补贴退休之日起确...
8541	我们夫妇雷珍秀胡湘电话是县麻林乡上林村人我老婆于年月号在县真山医院妇产科住院待产当时是妇产科...	卫生计生	6	夫妇雷珍秀胡湘电话县麻林乡上林村人老婆年月号县真山医院妇产科...
1955	本人居住在双安社区居委会组非商品房基本上每天晚上点开始到晚上点左右不要想着去洗澡肯定水压不够...	城乡建设	0	本人居住双安社区居委会组非商品房基本上每天晚上点晚上点左右不要想...
1776	在一个工程上建造师是工程质量与安全生产的第一责任人但是对于实行项目负责人的项目确不是人财物的...	城乡建设	0	工程上建造师工程质量安全生产第一责任人实行项目负责人项目确不是人...

图 2-7 停用词过滤并分词后的结果

(3) 生成词云

在 data0 中每个词语中间都是用空格隔开，本文在每个分类中罗列了前 100 个高频词，并画出这些高频词的词云，以此更加直观的看出每个分类的主题词，

生成词云的结果如图 2-8 所示：



图 2-8 生成词云的结果

(4) 计算 TF-IDF 值

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 的意思是词频(Term Frequency)，IDF 的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。

TF-IDF 算法的具体原理如下：

第一步，计算词频，即权重。

$$\text{词频 (TF)} = \frac{\text{某个词在文本中的出现次数}}{\text{文本的总次数}}$$

第二步，计算 IDF 权重，即逆文档频率，需要建立一个语料库，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。

$$\text{逆文档频率 (IDF)} = \log \left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数}+1} \right)$$

第三步，计算 TF-IDF 值

$$\text{TF-IDF} = \text{词频 (TF)} \times \text{逆文档频率 (IDF)}$$

实际分析得出 TF-IDF 值与一个词在留言文本出现的次数成正比，某个词文本的重要性越高，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文本的关键词。

计算 TF-IDF 值的结果如图 2-9 所示：

(9210, 742998)		:	:
(0, 421489)	0.14823785951918117	(9209, 503495)	0.055056221278478264
(0, 418847)	0.14823785951918117	(9209, 214022)	0.08568539756691501
(0, 295005)	0.1373474956221886	(9209, 244911)	0.10292752398514576
(0, 421535)	0.14823785951918117	(9209, 24778)	0.04340122170537562
(0, 652913)	0.14823785951918117	(9209, 559197)	0.104030504239541
(0, 339121)	0.10793872440974009	(9209, 730915)	0.060602455389819024
(0, 726164)	0.14823785951918117	(9209, 199934)	0.03722597599610743
(0, 273045)	0.141867405020233	(9209, 98241)	0.040346690708675066
(0, 673542)	0.14823785951918117	(9209, 340602)	0.08492285056236702
(0, 77767)	0.141867405020233	(9209, 275078)	0.046648259778600064
(0, 670372)	0.14823785951918117	(9209, 646951)	0.044586970420188096
(0, 451751)	0.14823785951918117	(9209, 395419)	0.045753337167274366
(0, 17248)	0.14823785951918117	(9209, 270826)	0.038072536391861166
(0, 293329)	0.14823785951918117	(9209, 290761)	0.11556979290119647
(0, 481747)	0.14823785951918117	(9209, 724944)	0.07915927906619802
(0, 226399)	0.14823785951918117	(9209, 324112)	0.05440361106298604
(0, 425814)	0.14823785951918117	(9209, 309896)	0.029348743471284886
(0, 728905)	0.13097704112324046	(9209, 158435)	0.07066539741058149
(0, 274262)	0.14823785951918117	(9209, 546527)	0.027634311574133326
(0, 514567)	0.14823785951918117	(9209, 593413)	0.08626299831436156
(0, 723278)	0.14823785951918117	(9209, 724355)	0.10445710519479576
(0, 330880)	0.1338415815201997	(9209, 298469)	0.03969629988478779
(0, 633026)	0.1373474956221886	(9209, 600737)	0.09687222334041846
(0, 670963)	0.14823785951918117	(9209, 590507)	0.041194542340439164
(0, 164381)	0.141867405020233	(9209, 730609)	0.022325941042395995

图 2-9 TF-IDF 值

图中结果显示 features 的维度是(9210, 747512), 总共有 9210 条评价数据，所有的词语数+词语对总数为 747512。

步骤四：分类器的选择

为了训练监督学习的分类器，我们首先将“留言详情”转变为包含数字的词向量。例如我们前面已经转换好的 TF-IDF 的 features。

当我们有了词向量以后我们就可以开始训练我们的分类器。分类器训练完成

后，就可以对没有见过的留言详情进行预测。经过后面的实验发现，线性支持向量机分类器的效果比较好，之后使用已有的词向量来训练该分类器。

步骤五：模型的选择

本文使用如下四个机器学习模型：逻辑回归、多项式朴素贝叶斯、线性支持向量机和随机森林，并评估每个模型的准确率。为了更好的区分每个模型的准确率，本文通过箱体图来展示，四个模型的箱体图如图 2-10 所示：

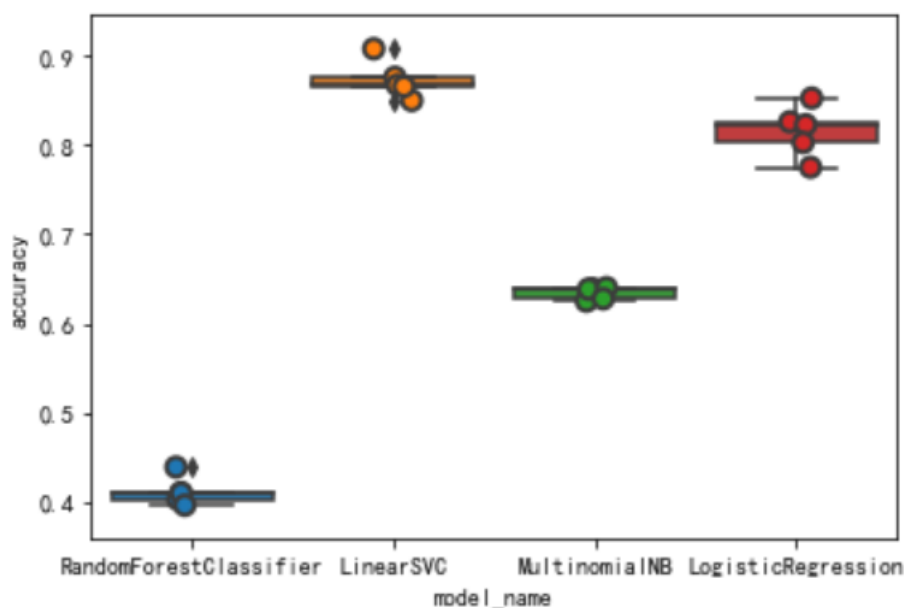


图 2-10 箱体图

从上图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 60%以上。其中线性支持向量机的准确率最高。

步骤六：模型的评估

根据模型的选择之后，我们发现 LinearSVC 平均准确率最高，之后再查看其混淆矩阵，并显示预测标签和实际标签的差异，混淆矩阵如图 2-11 所示：

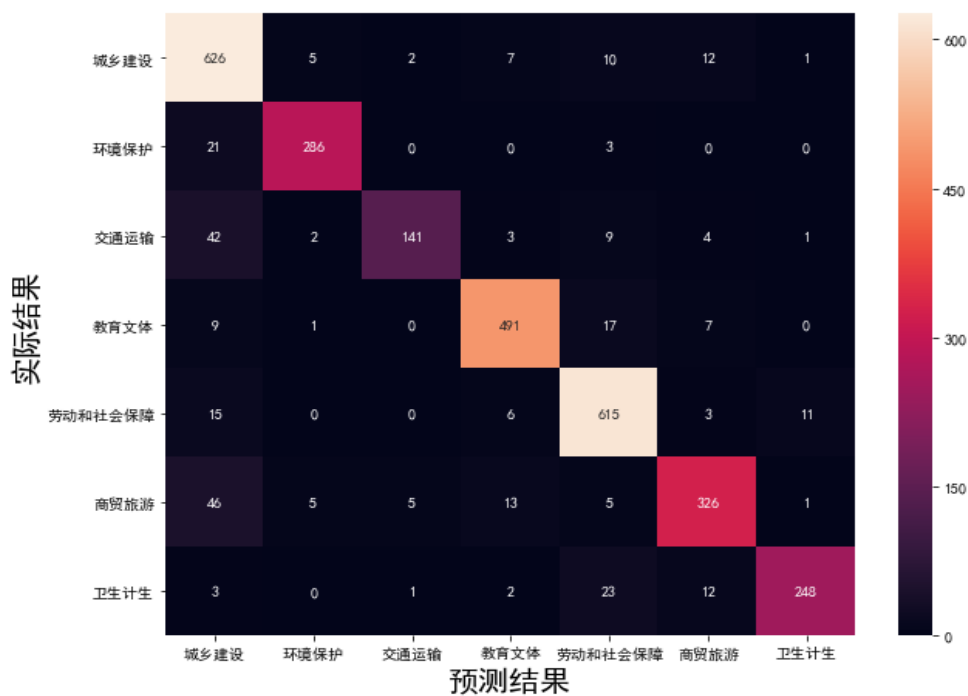


图 2-11 LinearSVC 模型的混淆矩阵

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。从上面的混淆矩阵中可以看出“交通运输”类预测最准确, “城乡建设”预测的错误数量较多。由于多分类模型一般不使用准确率 (accuracy) 来评估模型的质量, 因此本文通过查看各个类的 F1 分数来对模型进行评估。LinearSVC 模型的 F1 分数如图 2-12 所示:

accuracy 0.8990131578947368				
	precision	recall	f1-score	support
城乡建设	0.82	0.94	0.88	663
环境保护	0.96	0.92	0.94	310
交通运输	0.95	0.70	0.80	202
教育文体	0.94	0.94	0.94	525
劳动和社会保障	0.90	0.95	0.92	650
商贸旅游	0.90	0.81	0.85	401
卫生计生	0.95	0.86	0.90	289
accuracy			0.90	3040
macro avg	0.92	0.87	0.89	3040
weighted avg	0.90	0.90	0.90	3040

图 2-12 LinearSVC 模型的 F1 分数

此外, LSTM 也可以用来做中文文本多分类。LSTM 它是一种时间循环神经网络, 适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。所有的 RNN

都具有一种重复神经网络模块的链式形式。在标准 RNN 中，这个重复的结构模块只有一个非常简单的结构，例如一个 tanh 层。LSTM 同样是这样的结构，但是重复的模块拥有一个不同的结构。不同于单一神经网络层，这里是有四个，以一种非常特殊的方式进行交互。

LSTM 通过三个门控制信息的输入和输出，分别为“输入门”，“遗忘门”和“输出门”，其中“输入门”和“遗忘门”是至关重要的。“遗忘门”的作用是让循环神经网络“忘记”之前没有用的信息。“输入门”是在循环神经网络“忘记”了部分之前的状态后，它还需要从当前的输入补充最新的记忆。“输出门”则会根据最新的状态、上一刻的输出和当前的输入来决定该时刻的输出。

由于 LSTM 的分类效果比较好，本文对其进行了评估，并做出了其混淆矩阵和求它的 F1 分数，其混淆矩阵和求的 F1 分数如图 2-13 和图 2-14 所示：

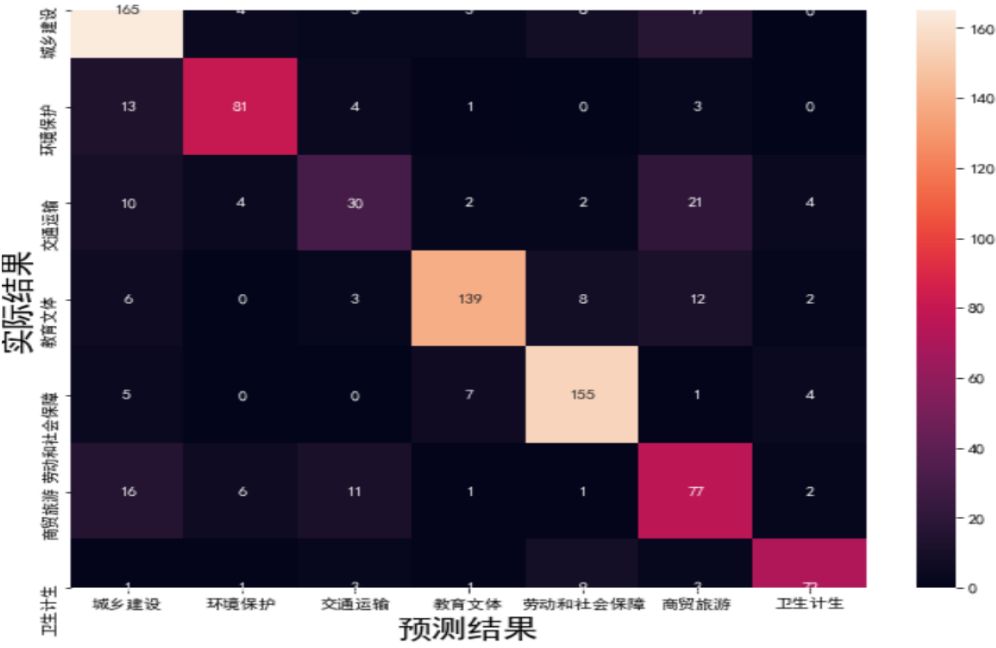


图 2-13 LSTM 的混淆矩阵

```

accuracy 0.7806731813246471
          precision    recall  f1-score   support

   城乡建设      0.76      0.82      0.79      200
   环境保护      0.84      0.79      0.82      102
   交通运输      0.56      0.41      0.47       73
   教育文体      0.90      0.82      0.86      170
   劳动和社会保障      0.85      0.90      0.87      172
   商贸旅游      0.57      0.68      0.62      114
   卫生计生      0.86      0.80      0.83       90

 accuracy              0.78      921
 macro avg      0.76      0.75      0.75      921
 weighted avg   0.78      0.78      0.78      921

```

图 2-14 LSTM 的 F1 分数的程序段

通过二者的 F1 分数的比较可以发现，LinearSVC 模型各个类的 F1 分数都大于 LSTM 模型各个类的 F1 分数，所以 LinearSVC 模型的效果比 LSTM 模型的效果要好一些。

2.1.3 结果分析

根据分析结果对群众留言进行合理分类，进而使相应的职能部门更好的处理群众留言。

通过对群众留言文本的分类可以发现：

1、群众的留言数量确实很多，我们可以通过七个主题来简单概括其思想：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、卫生计生。

2、尽管用于分类的模型比较多，但每个分类器的效果是不一样的。其中，随机森林不适合处理文本数据，而在剩余的模型当中，LinearSVC 模型和 LSTM 模型的效果是比较好的，在本文的实验中，LinearSVC 模型的效果更胜一筹。

3、通过求 F1 分数可以看出，F1 分数的高低可能与分类的训练数据量的多少有一定的关系，本文是分类的训练数据量越多，F1 分数就相对的高一些。

2.2 热点问题挖掘问题求解

2.2.1 整体流程

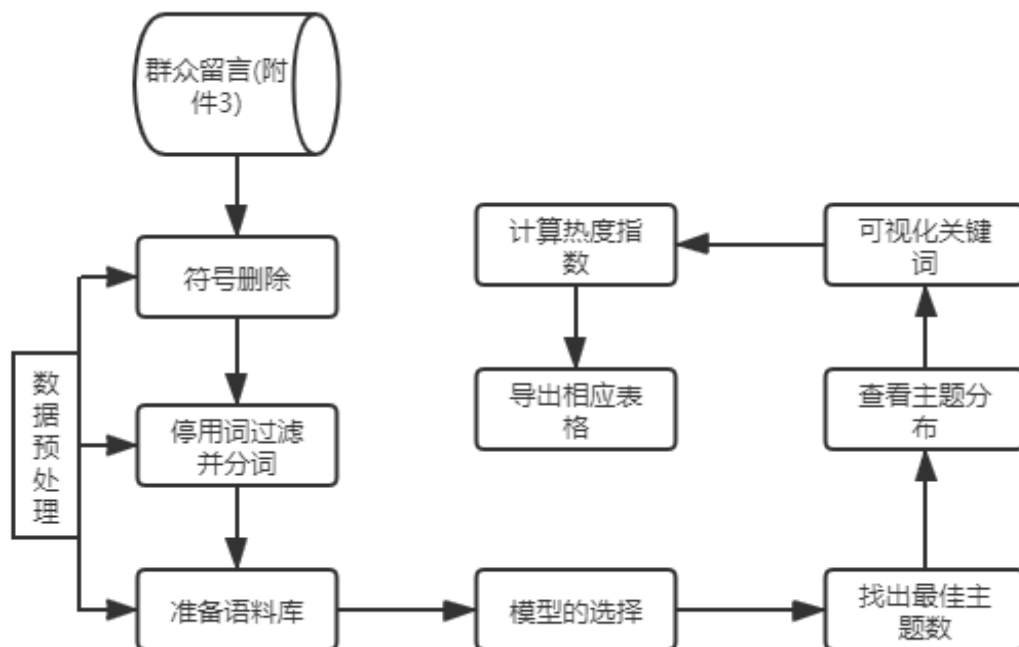


图 2-15 整体流程图

该问题主要包括以下几个步骤：

- 步骤一：在数据挖掘挑战赛的官网上下载本问题的全部数据，即可得到该问题实验所需的数据。
- 步骤二：数据预处理，首先删除汉字以外所有的符号；接下来进行通用词过滤并进行分词，去掉留言中很多日常使用频率很高的常用词，并将一句评论分成多个词语进行分析；最后准备语料库。
- 步骤三：模型的选择，在对模型进行比价之后，选择最佳的模型。
- 步骤四：找出最佳主题数，通过选择不同的主题数，得到不同的一致性得分，从而找到最佳的主题数。
- 步骤五：查看主题分布，在所有的留言当中对其主题进行分配，之后就可以查看所有留言中主题的分布。
- 步骤六：可视化关键词，对每个主题的关键词进行可视化。
- 步骤七：计算热度指数，由于每个主题有许多文档，并且还点赞数和反对数，此时本文将点赞和反对都当作文档加一处理，从而计算出热度指数。

- 步骤八：导出相应文档，当热度指数求出后，找出排名前五的热度，就可以导出该问题所需要的文档。

2.2.2 具体步骤

步骤一：获取实验的数据

打开数据挖掘挑战赛的官网，找到如图 2-2 所示内容，点击下载 C 题全部数据，即可得到该问题实验所需的全部数据。

步骤二：数据预处理

(1) 符号删除

留言文本中有许多标点符号，并且这些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，所以必须将标点符号给清除，符号删除的结果如图 2-16 所示：

	留言详情	留言时间	点赞数	反对数
775	市河汉桥东岸附近步行道安装了车辆减速器对普通骑车人造成危险见到小孩溜冰没注意都会头朝下摔倒希望拆除	2019/4/7 17:55:36	1	0
1512	近日刊登了关于栋社区用房问题的回复如下关于社区办公用房的问题经过实地走访调查了解一是为了社区...	2019/2/13 3:57:31	4	1
675	市区黄谷路号山水嘉园栋三单元房被改建成户分别出租给不同人员存在严重的消防治安等隐患投诉给物业...	2019/1/8 10:08:32	0	0
2701	我使用了半个月左右市地铁本以为是很方便的出门不用带卡一步手机全搞定结果这半个月里扫码闸门不开...	2019/7/19 8:23:40	6	0
711	请问县楚龙街道山水湾小区和保利香樟国际小区之间的龙塘路苑景路何时拉通到星沙联络线每天经过该断...	2019/8/21 14:51:44	3	0
641	目前省会在交通强国文明创建和都市公交等常态工作中我市常规公交无论是新能源公交普及智能公交技...	2019/12/16 11:07:20	0	0
3370	我们是西地省市江干锦园的团购户通过西地省司法警官职业学院以下简称司法警院向西地省江干投资有限...	2019/4/25 0:18:28	4	0
3953	关于市人才新政购房补贴的高级技师申报要求申请人身份证所在单位出具的工作证明及单位营业执照复印...	2019/7/29 10:42:38	1	0
2973	本人年在市万科金域蓝湾买了一套房本想开开心心入住没想到刚交房却发现一楼底商一个大大的烟道对着自...	2019/3/5 14:24:33	0	0
188	麓天小区三期栋旁边挖自然绿化山体大建违章视规则为无物影响极其恶劣开挖山地占用公共绿地私自开挖...	2019/10/12 8:39:18	0	0

图 2-16 符号删除的结果

(2) 停用词过滤并分词

本文将停用词过滤和中文分词一起处理，处理后的结果如图 2-17 所示：

	留言详情	留言时间	点赞数	反对数	cut_notes
2097	又到开学的时候了小朋友们都来到了幼儿园娱乐的新学期开始了可是市河幼儿园的家长接到通知又要交门...	2019/9/6 21:56:22	0	0	开学 小朋友 都 来到 幼儿园 娱乐 新 学期开始 市 河 幼儿园 家长 接到 通知 交 门禁...
3751	尊敬的领导做为泉塘本地居民很欣慰在党和政府的英明决策和领导下泉塘这些年发生了翻天覆地的变化高...	2019/4/24 17:45:08	10	0	尊敬 领导 做 泉塘 本地 居民 很 欣慰 党和政府 英 明 决策 领导 下 泉塘 年 发生 翻...
1108	本人购买的金域蓝湾三期的房子年月日收房至今出现多处质量问题包括漏水导致房屋发霉地板被泡开关没...	2019/5/10 17:21:25	0	0	本人 购买 金域 蓝湾 三期 房子 年月日 收房 至今 出现 多处 质量 问题 包括 漏水 ...
3154	我们是县星沙镇杉仙岭社区六区的住户几乎每天都被栋的钻石皇家会所吵得难以入睡之前多次拨打过投诉...	2019/5/28 16:31:45	0	0	县星 沙镇 杉仙岭 社区 六区 住户 几乎 每天 都 栋 钻石 皇家 会 吵 难以 入睡 之...
3236	富绿新村提质改造工程本是一项民心工程是件大好事可是该工程在运作中项目与公示有较大缩水特别是路...	2019/3/23 15:38:35	3	0	富绿 新村 提质 改造 工程 本 是一 项 民心 工程 件 大 好事 工程 运作 中 项目 公示 ...
3699	我使用的市银行网上服务已经因为他们的控件升级而无法使用对我的工作影响极大请敦促该行马上改变他...	2019/8/28 9:46:57	0	0	使用 市 银行 网上 服务 已经 控件 升级 无法 使用 工作 影响 极大 请 敦促 该行 ...
2333	请问为何松雅湖西北部工程烂尾无人处理建筑材料随意堆积几年之久脚手架锈蚀桥面无护栏多次看到小朋...	2019/5/24 23:33:56	9	1	请问 松雅湖 西北部 工程 烂尾 无人 处理 建筑材 料 随意 堆积 几年 久 脚手架 锈蚀 ...
1514	市房屋交易管理中心关于执行政府发文号文件操作细则的通知中规定市人才购房及购房补贴实施办法试行...	2019/1/16 11:58:48	0	0	市 房屋 交易 管理 中心 执行 政府 发文 号 文件 操 作 细则 通知 中 规定 市 人才 购...
1717	辉旭地产无施工许可强制开工百姓生命财产安全受重大威胁辉旭地产公司位于区兴联路的湖山赋项目学校...	2019/2/28 11:09:49	1	0	辉旭 地产 无 施工 许可 强制 开工 百姓 生命 财产 安全 受 重大 威胁 辉旭 地产 公...
1248	涉外和第一师范中间这条路枫林三路凌晨一点拖土车往步步高新天地方向空车从步步高新天地返回几十个...	2019/9/9 15:21:17	0	0	涉外 第一 师范 中间 这条 路 枫林 三路 凌晨 一点 拖土车 步步高 新天地 方向 空车...

图 2-17 停用词过滤并分词后的结果

(3) 准备语料库

准备语料库，主要是创建一个文档术语矩阵和术语词典。

步骤三：模型的选择

本文比较了 LDA 和 LDA Mallet 这两个模型，并计算出来二者的困惑度和一致性分数，计算结果图 2-18 和图 2-19 所示：

```
Perplexity: -9.656275107009009
Coherence Score: 0.437291180078506
```

图 2-18 LDA 模型的困惑度和一致性分数

```
Perplexity: -9.656245336453674
Coherence Score: 0.5068833724772478
```

图 2-19 LDA Mallet 模型的困惑度和一致性分数

通过比较发现，LDA Mallet 模型比 LDA 模型的一致性分数高，故本文选择用 LDA Mallet 模型。

步骤四：找出最佳主题数

本文通过选择不同的主题数，来计算其一致性得分，并将其结果进行可视化，从图中找出最佳的主题数，结果如图 2-20 所示：

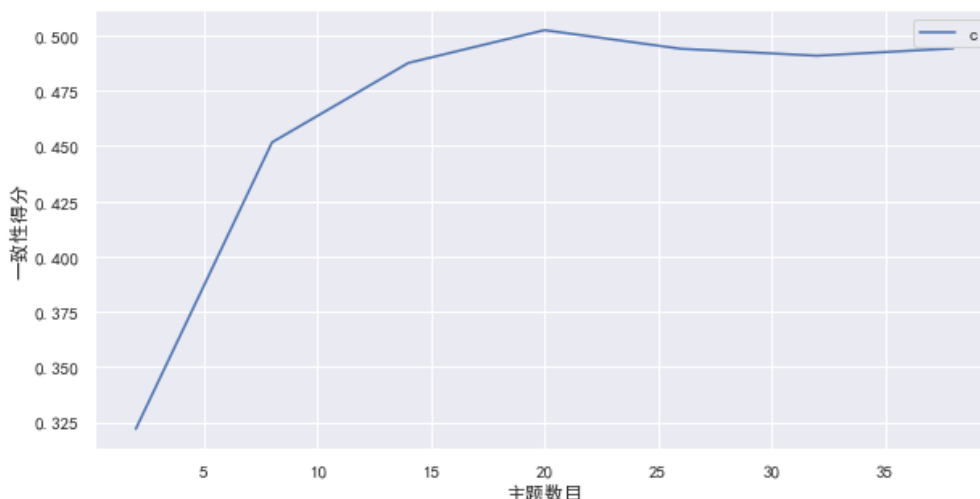


图 2-20 不同主题数的一致性得分

从图中可以发现，当主题数为 20 个时，一致性得分最高，故本文选取最佳主题数为 20 个。

步骤五：查看主题分布

首先，在所有的留言当中对其主题进行分配，其类似 TF-IDF，创建主题权重矩阵，以留言为行，主题为列，生成前 10 个文档的主题分布如图 2-21 所示：

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0.302382	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.055542	0.114512	0.0	0.000000	0.000000	0.028926	0.000000	0.478090	0.000000	0.000000	0.0	0.000000	0.000000
1	0.056861	0.033154	0.080503	0.000000	0.000000	0.000000	0.000000	0.000000	0.043916	0.0	0.396070	0.000000	0.068874	0.143250	0.000000	0.079725	0.000000	0.0	0.088563	0.000000
2	0.572864	0.000000	0.000000	0.000000	0.193354	0.146767	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.057759	0.021000	0.0	0.000000	0.000000
3	0.082608	0.000000	0.000000	0.000000	0.000000	0.000000	0.066585	0.000000	0.197644	0.0	0.000000	0.000000	0.000000	0.000000	0.110059	0.255192	0.130094	0.0	0.119256	0.032668
4	0.423379	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.071218	0.119386	0.000000	0.000000	0.040871	0.069710	0.221767	0.0	0.043935	0.000000
5	0.143519	0.000000	0.000000	0.000000	0.000000	0.263304	0.000000	0.000000	0.065359	0.0	0.000000	0.000000	0.000000	0.000000	0.241234	0.060307	0.000000	0.0	0.000000	0.212588
6	0.069726	0.000000	0.000000	0.000000	0.313888	0.309030	0.088886	0.000000	0.000000	0.0	0.083211	0.000000	0.000000	0.000000	0.091063	0.000000	0.000000	0.0	0.027053	0.000000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.113277	0.0	0.494309	0.000000	0.000000	0.234624	0.000000	0.052467	0.096260	0.0	0.000000	0.000000
8	0.101345	0.642712	0.031844	0.07703	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.027660	0.104775	0.000000	0.0	0.000000	0.000000
9	0.000000	0.050893	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.093607	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.151995	0.684403	0.0	0.000000	0.000000

图 2-21 前 10 个文档的主题分布

其次，本文进而进行查看所有留言文档的主题分布，并对其进行可视化处理，所生成的结果如图 2-22 所示：

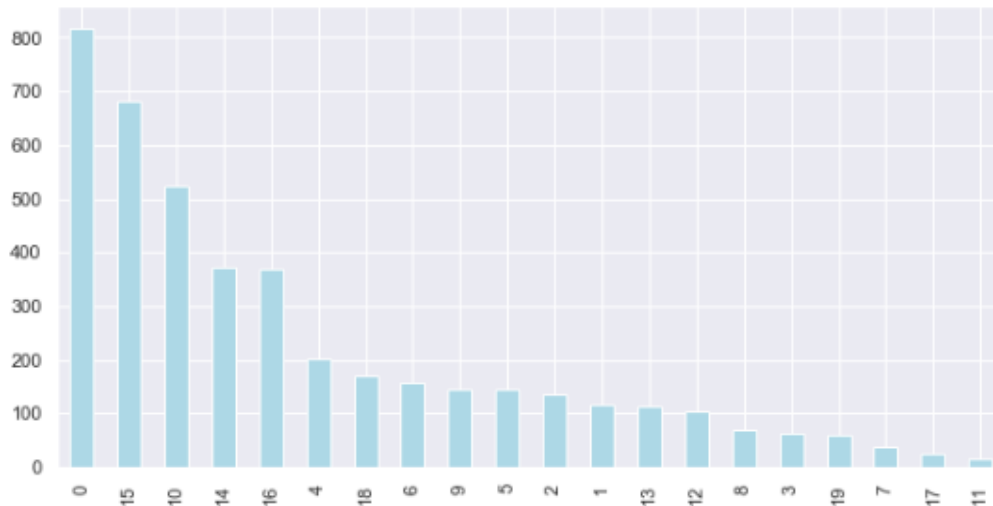


图 2-22 所有留言文档的主题分布

步骤六：可视化关键词

对已经确定的 20 个主题，找出其每个主题中所包含的关键词，并用 pyLDAvis 软件包对其进行可视化，结果如图 2-23 所示：

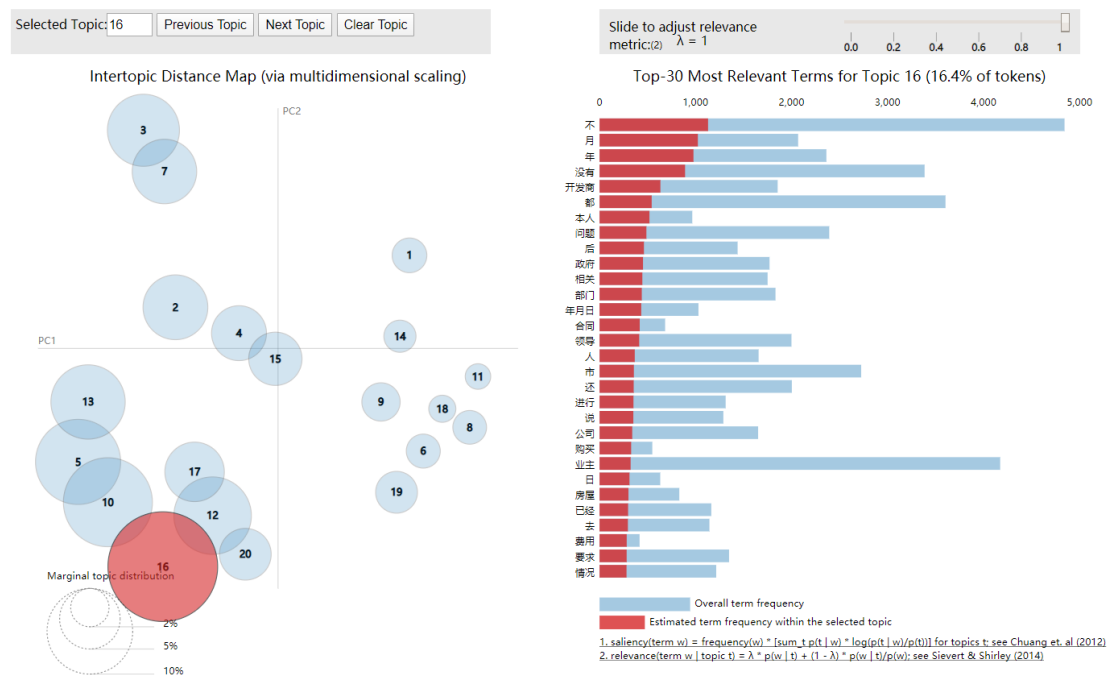


图 2-23 关键词的可视化结果

左侧图上的每个气泡代表一个主题。气泡越大，该话题越普遍。一个好的主题模型将有相当大的，不重叠的气泡分散在整个图表中，而不是聚集在一个象限中。一个主题太多的模型通常会在图表的一个区域中聚集许多重叠的小气泡。从图中可以明显的看出，16 这个主题的起泡最大，故该主题里面的话题是最为普

遍的。

步骤七：计算热度指数

由于每个主题有许多文档，并且还有点赞数和反对数，此时本文将点赞和反对都当作文档加一处理，并生成热度图，如图 2-24 所示：

热度图

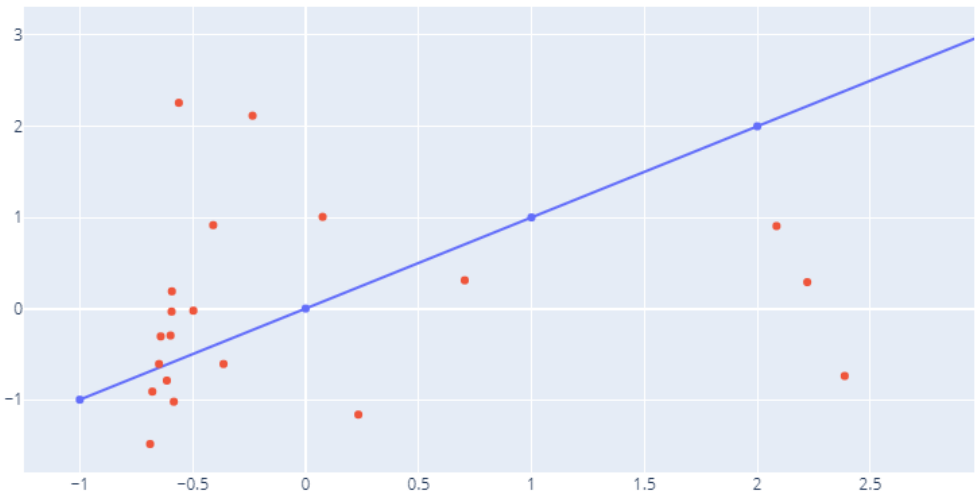


图 2-24 热度图

通过计算图中的点到原点的距离，就可以计算出本文所需要的热度指数，并对其进行排序，结果如图 2-25 所示：

	点赞反对之和	topic_nums	heat_index
18	2.367021	-0.799288	2.498329
2	-0.577047	2.405826	2.474062
10	2.069704	0.186135	2.078057
12	1.977346	0.624101	2.073499
8	1.380182	1.390541	1.959211
1	-0.241775	1.838461	1.854291
9	-0.463181	-1.456236	1.528123
14	-0.671936	-1.217346	1.390478
4	-0.688383	-1.038178	1.245666
7	-0.407514	1.101882	1.174824
0	-0.665610	-0.878918	1.102512
17	-0.652958	-0.460859	0.799216
15	-0.632715	-0.361322	0.728617
16	-0.644102	0.305581	0.712914
13	-0.370824	-0.540490	0.655469
3	-0.050734	-0.590258	0.592435
11	-0.571987	-0.072663	0.576584
6	-0.420165	-0.202061	0.466227
19	-0.403718	-0.022894	0.404367
5	-0.331603	-0.212015	0.393587

图 2-25 热度指数排名

从上图可以看出，排在前五的主题为 18, 2, 10, 12, 8。

步骤八：导出相应文档

当热度指数求出后，找出排名前五的主题及其对应的留言，就可以导出该问题所需要的文档，本文所找的结果如图 2-26 所示：

	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数	Dominant_Topic
0	188006	A000102948	A3区一米阳光婚纱摄影是否合法纳税了?	2019/2/28 11:25:05	座落在市区联丰路米兰春天栋一家名叫一米阳光婚纱摄影的影楼据说年单这一个工作室营业额就上百...	0	0	5.0
1	188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	市区道路命名规划已经初步成果公示文件什么时候能转化成为正式的成果希望能加快完成的命名规范给道...	1	0	8.0
2	188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	本人系春华镇金鼎村七里组村民不知是否有相关水泥路到户政策和自来水到户政策如政府主导投资村民部...	1	0	16.0
3	188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	靠近黄兴路步行街城南路街道大古道巷一步两搭桥小区停车场东面围墙外第一单元一住户卫生间粪便长年...	1	0	16.0
4	188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	市区中海国际社区三期四期中间即蓝天璞和洲幼儿园旁边那块空地一直处于三不管状态物业不管城管不管...	0	0	2.0
...
4321	360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	关于西地省市经济学院寒假过年期间组织学生去工厂工作过年本该是家人团聚的时光很多家长一年回来一...	0	0	12.0
4322	360111	A1204455	A市经济学院组织学生外出打工合理吗?	2019-11-05 10:31:38	一名中职院校的学生学校组织我们学生在外边打工在外省做流水线工作还要倒白夜班本来都在学校好好上...	0	1	12.0
4323	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	各位领导干部大家好我是市经济学院的一名学生临近毕业学校开始组织学生参加实习当然学生是必须实习...	0	0	12.0
4324	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	市经济学院强制组织电子商务跟企业物流专业实习其中我们企业物流专业实习个月去江苏暑假去过回来不...	0	3	12.0
4325	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	书记您好我是来自西地省经济学院体育学院的一名即将大四的学生系里要求我们在实习前分别去指定的不...	0	9	11.0

图 2-26 主题及其对应的留言

2.2.3 结果分析

1. 在考虑特定时间，地点/人物分析时，我们将留言的最大间隔时间段，点赞数和反对数取出来，计算这些变量对生成主题的影响。
2. 在多次测试后，发现时间间隔因素对最大主题数影响很小，这里考虑将时间间隔因素忽略掉。
3. 分析留言的点赞数和反对数时，发现点赞和反对均代表了对留言的不同看法，代表了问题的热度，故将点赞数和反对数加和处理。
4. 将点赞反对之和，每个主题文档数做均值方差归一化处理，考虑到点赞反

对也是留言的另一种转述形式，这里将两个因素各给予 50%的权重，通过计算归一化后的每一行主题参数，得出最后的热点指数。

2.3 答复意见的评价问题求解

2.3.1 整体流程

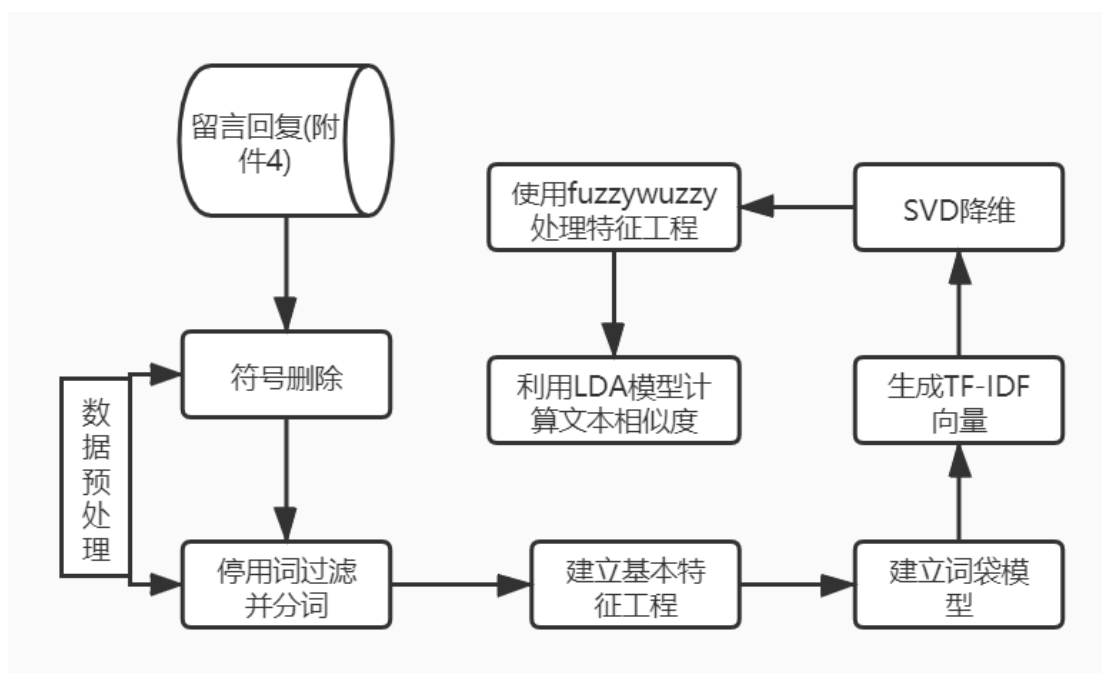


图 2-27 整体流程图

- 步骤一：在数据挖掘挑战赛的官网上可以下载本问题的全部数据，即可得到该问题实验所需的数据。
- 步骤二：数据预处理，首先删除留言详情和答复意见中的空格和标点符号；再使用停用词过滤并进行分词，去掉留言中大部分日常生活中使用频率很高的常用词，并分成多个词语进行分析。
- 步骤三：建立基本特征工程来计算留言详情和答复意见的词数，并且计算两个问题的常用单词数。
- 步骤四：使用 CountVectorizer 函数来构建 BOW (词袋) 模型。
- 步骤五：词向量化，生成 TF-IDF 向量。
- 步骤六：使用 SVD 降维。使用 SVD 后将 TF-IDF 的维数降低到 180 维。
- 步骤七：使用 fuzzywuzzy 来处理特征工程。
- 步骤八：利用 LDA 模型计算文本相似度。

2.3.2 具体步骤

步骤一：获取实验的数据

打开数据挖掘挑战赛的官网，找到如图 2-2 所示内容，点击下载 C 题全部数据，即可得到该问题实验所需的全部数据。

步骤二：数据预处理

(1) 空格与符号删除

留言文本中有许多标点符号和空格，且这些符号和空格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，所以必须将标点符号和空格给清除，清楚的结果如图 2-28 所示：

	留言详情	留言时间	答复意见	答复时间
0	年月以来位于市区桂花坪街道的区公安分局宿舍区景蓉华苑出现了一番乱象该小区的物业公司美顺物业扬...	2019/4/25 9:32:09	现将网友在平台问政西地省栏目向胡华衡书记留言反映区景蓉华苑物业管理有问题的调查核实情况向该网...	2019/5/10 14:56:53
1	潇楚南路从年开始修到现在都快一年了路挖得稀烂用围栏围起一直不怎么动工有时候今天来台挖机挖两几...	2019/4/24 16:03:40	网友您好针对您反映区潇楚南路洋湖段怎么还没修好的问题区洋湖街道高度重视立即组织精干力量调查处...	2019/5/9 9:49:10
2	地处省会市民营幼儿园众多小孩是祖国的未来但民营幼儿园教师一直都是超负荷工作且收入又是所有行业...	2019/4/24 15:40:04	市民同志你好您反映的请加快提高民营幼儿园教师的待遇的来信已收悉现回复如下为了改善和提高民办幼...	2019/5/9 9:49:14
3	尊敬的书记您好我研究生毕业后根据人才新政落户市想买套公寓请问购买公寓能否享受研究生万元的购房...	2019/4/24 15:07:30	网友您好您在平台问政西地省上的留言已收悉市住建局及时将您反映的问题交由市房屋交易管理中心办理...	2019/5/9 9:49:42
4	建议将白竹坡路口更名为马坡岭小学原马坡岭小学取消保留马坡岭	2019/4/23 17:03:19	网友您好您的留言已收悉现将具体内容答复如下关于来信人建议白竹坡路口更名为马坡岭小学原马坡岭小...	2019/5/9 9:51:30
...
2811	我们是市汽车北站进站口的周围居民在这里的马良社区马园口组有一栋六层近一万平方米的新大楼月份该...	2018/12/12 15:20:46	您的留言已收悉关于您反映的问题已转区委区政府调查处理	2019/1/8 16:54:53
2812	强烈反对市路公交车改线路获悉从月日起路公交车要更改线路滨江路将被弃线滨江路乘车群众表示强烈反...	2018/6/12 8:51:03	您的留言已收悉关于您反映的问题已转市交通运输局调查处理	2018/7/4 16:55:53
2813	县文盛小学引入特色班每个学生必须参加然后特色班老师就要学生购买相关的物品一星期一节课很多家长...	2018/10/11 20:02:52	您好获悉关于对县文盛小学特色班的质疑的网帖后我局领导高度重视并责成教育局基教股调查处理现将调...	2018/10/24 9:22:07
2814	贺厅长您好自燃油税费改革以来到市财政局咨询得不到结果现特向您咨询中央财政转移支付的公路养路费...	2012/9/4 23:14:44	西地省平台问政西地省栏目组网民在贵栏目留言咨询中央转移支付我省燃油税费资金情况现函复如下关于...	2013/1/6 15:41:02
2815	县朱良桥乡可以说是县最破烂的乡了集镇建设相关基础设施建设差是该乡的显著特征况且没有一条好路通...	2011/10/3 21:52:37	您好您的留言我厅领导高度重视要求相关部门进行了认真的调查和研究现就有关问题回复如下我厅对县城...	2012/2/28 10:19:55

图 2-28 空格与符号删除的结果

(2) 停用词过滤并分词

本文将停用词过滤和中文分词一起处理，处理后的结果如图 2-29 所示：

	留言详情	留言时间	答复意见	答复时间	cut_notes	cut_response
2231	尊敬的县长您好我叫孙琛身份证号于年月日在县城锦豪雅景园小区购买一套商品房现在有关费用问题发生...	2013/6/3 13:05:49	已收悉	2013/7/5 16:47:46	尊敬 县长 您好 孙琛 身份证号 年月日 县 城 锦豪雅 景园 小区 购买 一套 商品房 现 在...	已 收悉
2797	质疑就县教育部门住房公积金标准为什么不统一的回复什么叫各单位可视财力状况住房公积金来源于学生...	2019/3/11 22:37:50	您好您在年月日问政西地省上发帖反映的问题我局高度重视迅速调查了解情况现回复如下根据国务院住房...	2019/3/18 15:41:16	质疑 县 教育部门 住 房 公 积 金 标 准 不 统 一 回 复 单 位 可 视 财 力 状 况 住 房 公 积 金...	您好 年月日 问政 西 地 省 上 发 帖 反 映 问 题 我 局 高 度 重 视 迅 速 调 查 了 解 情 况 现...
2555	本人符东晓岁在县四中读高一因家庭不幸父母在我两岁时因感情不合离婚当时哥哥符州权由妈赵秀卵抚养...	2017/7/12 16:15:53	您好您于月日在平台问政西地省中反映县火场乡低保评选问题咨询的网帖我单位收悉后安排专人进行了调...	2017/7/13 16:32:13	本人 符东晓岁 县 四 中 读 高 一 家 庭 不 幸 父 母 两 岁 时 因 感 情 不 合 离 婚 当 时 哥...	您好 月 日 平 台 问 政 西 地 省 中 反 映 县 火 场 乡 低 保 评 选 问 题 咨 询 网 站 单 位...
2407	我爸爸年有带了两学期课前年有登记代课老师拿补贴我们没有人通知我爸爸他不知道去年是听别人说才知...	2019/1/9 17:11:10	网友您反映的问题已转县调查核处年月日您好由于您未提供当事人姓名等信息无法核查您可到县教育局人...	2019/2/13 11:25:09	爸爸 年 带 两 学 期 课 前 年 登 记 代 课 老 师 补 贴 没 有 人 通 知 爸爸 不 知 道 去 年...	网友 反映 问题 已 转 县 调 查 核 处 年 月 日 您 好 未 提 供 当 事 人 姓 名 信 息 无 法 核...
1655	长期一来江口大桥迟迟未动工县江口镇的留守儿童每年夏天都有到东水河畔去洗澡安全隐患很重每年都...	2019/8/20 16:54:53	广大网友近日有网友在西地省平台留言希望县江口镇政府前的池塘改建成游泳馆我镇党委政府高度重视当...	2019/9/18 11:07:43	长期 江 口 大 桥 迟 迟 未 动 工 县 江 口 镇 留 守 儿 童 每 年 夏 天 都 来 水 河 畔 去 ...	广大 网友 近 日 网 友 西 地 省 平 台 留 言 希 望 县 江 口 镇 政 府 前 池 塘 改 建 成 游 泳...
2472	位于县人民西路月亮湾对面佳惠超市大汉店门口贴了个月的装修公告不见有任何动静一堆破铁架子天天...	2018/3/26 9:50:33	您好您反映的问题已转城管大队核查办复感谢您的留言祝您事事顺心年月日	2018/3/26 15:57:52	位于 县 人 民 西 路 月 亮 湾 对 面 佳 惠 超 市 大 汉 店 门 口 贴 月 装 修 公 告 不 见 动...	您好 反映 问题 已 转 城 管 大 队 核 查 办 复 感 谢 您 留 言 祝 您 事 事 顺 心 年 月 日
2013	柏林工业园勐建环保公司自称有环保局领导关系在未验收合格私自违法生产达一年多经营排出恶臭黄烟请...	2019/11/7 8:02:45	网友您好收悉您的留言后我局高度重视立即调查核实现回复如下西地省勐建环保资源科技发展有限公司是...	2019/11/19 15:48:44	柏 林 工 业 园 勐 建 环 保 公 司 自 称 环 保 局 领 导 关 系 未 验 收 合 格 私 自 违 法 生 产 ...	网 友 您 好 收 悉 留 言 后 我 局 高 度 重 视 立 即 调 查 核 实 现 回 复 如 下 西 地 省 勐 建 环...
240	我的户口不是在市现在在市工作想把档案转到市人社局需要如何做市人社局上班时不需不需要调档函	2018/3/22 9:55:23	网友您好您的留言已收悉现将有关情况回复如下根据中共中央组织部人力资源社会保障部等五部门关于进...	2018/3/29 14:41:06	户 口 不 是 市 现 在 市 工 作 想 档 案 转 到 市 人 社 局 需 要 做 市 人 社 局 上 班 时 间 ...	网 友 您 好 留 言 已 收 悉 现 将 情 况 回 复 如 下 中 共 中 央 组 织 部 人 力 资 源 社 会 保 障 部 五 ...
64	九点四十几的时候在西地省中医药大学站公交车区公交线路车牌号为的公交车没给老人上车七十多岁的奶...	2018/11/9 10:27:57	网友您好您的留言已收悉现将有关情况回复如下经查看车载视频驾驶员驾驶区自编号为楚的车辆于月日时...	2018/11/29 14:58:43	九 点 四 十 几 西 地 省 中 医 药 大 学 站 公 交 车 区 公 交 路 线 车 牌 号 公 交 车 没 给 老 人 上 车...	网 友 您 好 留 言 已 收 悉 现 将 情 况 回 复 如 下 查 看 车 载 视 频 驾 驶 员 驾 驶 区 自 编...
1383	市急救中心是全市人民共享的急救中心为什么市中医院独家长期占有其它医院却不街边明显违反了国家...	2018/5/23 22:03:21	市卫生和计划生育局回复收到网友反映的问题后我局高度重视因事关全市医疗急救体系建设我局正组织相...	2018/5/29 8:31:07	市 急 救 中 心 全 市 人 民 共 享 急 救 中 心 市 中 医 院 独 家 长 期 占 有 医 院 却 不 准 沾 边...	市 卫 生 计 划 生 育 局 回 复 收 到 网 友 反 映 问 题 后 我 局 高 度 重 视 事 关 全 市 医 疗 ...

图 2-29 停用词过滤并分词后的结果

步骤三：建立基本特征工程

基本特征工程主要包括以下功能：

1. 计算留言详情和答复意见的词数；
2. 计算留言详情和答复意见的常用单词数。

计算的部分结果如图 2-30 所示：

留言详情	留言时间	答复意见	答复时间	cut_notes	cut_response	len_word_notes	len_word_response	common_words
0	年月以来位于市区桂花坪街道的区公安分局宿舍区晨暮华庭出现了一	现将网友在平台问政西地省栏目向胡华街书记留言反映区晨暮华庭物业管理有问题的调查核实情况向该网...	2019/4/25 9:32:09	2019/5/10 14:56:53	年月以来位于市区桂花坪街道区公安分局宿舍区晨暮华庭出现乱象小区...	381	490	72
1	潇湘南路从年开始修到现在都快一年了路挖得稀烂用围栏围起一直不怎么动工有时候今天来台挖机挖两几...	网友您好针对您反映区潇湘南路洋湖段怎么还没修好的问题区洋湖街道高度重视立即组织精干力量调查处...	2019/4/24 16:03:40	2019/5/9 9:49:10	潇湘南路现在都快一年路挖得稀烂围栏围起怎么动工有时候今天...	169	335	28
2	地处省会市民喜幼儿园众多小孩是祖国的未来但民喜幼儿园教师一直是超负荷工作且收入又是所有行...	市民同志您好您反映的请加快提高民喜幼儿园教师的待遇的来信已收悉现回复如下为了改善和提高民办幼...	2019/4/24 15:40:04	2019/5/9 9:49:14	地处省会市民喜幼儿园众多小孩祖国的未来民喜幼儿园教师都超负荷工作收...	189	402	39

图 2-30 特征工程的部分结果

步骤四：建立词袋模型

本文使用的 sklearn 中的 CountVectorizer 构建 BOW(词袋)模型，该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。BoW 使用一组无序的单词(words)来表达一段文字或一个文档。

步骤五：生成 TF-IDF 向量

与问题二中求解 TF-IDF 值过程一样，这里就不再做详细说明。

步骤六：使用 SVD 降维

在机器学习中经常会碰到一些高维的数据集，而在高维数据情形下会出现数据样本稀疏，距离计算等困难，且在高维特征中容易出现特征之间的线性相关，这也就意味着有的特征是冗余存在的，于是解决这些问题，就要通过降维来实现。

本文使用 SVD 后将 TF-IDF 的维数降低到 180 维，降维后的结果如下图 2-31 所示：

```
[[ 0.38434705 -0.08378351 -0.08971599 ... -0.01974814 -0.00786557
  0.00777293]
 [ 0.28775227 -0.07590031 -0.01122755 ... -0.02493102 -0.0530998
  0.01888859]
 [ 0.17066279 -0.04217538 -0.10171224 ... -0.02092735 -0.0760269
  0.04773524]
 ...
 [ 0.15346023 -0.02014798 -0.14560392 ... 0.01936855 0.04311808
 -0.05714308]
 [ 0.11795183 -0.03100117 -0.0582893 ... -0.013608 0.01395568
 0.10971404]
 [ 0.28814505 -0.08108234 -0.045629 ... 0.04312402 0.04325601
 -0.02337485]]
```

图 2-31 SVD 降维后的结果

步骤七：使用 fuzzywuzzy 来处理特征工程

Fuzzywuzzy 是一个简单易用的模糊字符串匹配工具包。它依据 Levenshtein Distance 算法计算两个序列之间的差异。

Levenshtein Distance 算法，又叫 Edit Distance 算法，是指两个字符串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。

故本文使用来 fuzzywuzzy 匹配模糊字符串，fuzzywuzzy 模块使用不同比率来描述两个字符串之间的相似程度。其中数值越大，相似程度越高。处理后的结果如图 2-32 所示：

	fuzz_ratio	fuzz_partial_ratio	fuzz_partial_token_set_ratio	fuzz_partial_token_sort_ratio	fuzz_token_set_ratio	fuzz_token_sort_ratio
0	9	14	100	26	38	25
1	6	20	100	24	10	5
2	3	23	100	35	29	11
3	15	45	100	36	60	18
4	38	100	100	50	100	38
...
2811	9	31	38	38	11	9
2812	3	39	100	39	13	3
2813	7	27	100	43	60	11
2814	10	41	100	37	58	22
2815	9	20	100	35	30	18

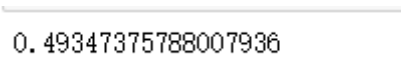
2816 rows x 6 columns

图 2-32 使用 fuzzywuzzy 处理后的结果

步骤八：利用 LDA 计算文本相似度

用问题详情和答复意见构建 LDA 模型，用 gensim.matutils.cossim 方法计

算两者之间的文本相似度，计算的相似度如图 2-33 所示：



0.49347375788007936

图 2-33 计算相似度

从图中可以看出有的留言详情和答复意见文本相似度很高，有的则较低。

2.3.3 结果分析

由于网络问政刚刚兴起,还存在着政府工作人员对网络问政认识不足、回应信息不完备、对待回应的认识分歧、缺乏系统保障和群众参与意识有待提高等问题。良性的政府回应机制是群众网络问政健康发展的前提。

政府应积极回应公民的利益诉求;推进政务公开,保障民众的知情权;以合作理念重塑政府与民众的关系,达成民意共识;拓展网络沟通的渠道,切实保障民众的表达权;加强网络群体类型化分析,采取不同的应对措施;加强网络信息管理,通过组织引导与部门协同,确保及时回应。

第3章 结论

随着网络在中国的日益普及,网络问政平台在政治、经济和社会生活中扮演着日益重要的角色,成为公民行使知情权、参与权、表达权和监督权的重要渠道。因此,对群众问政留言记录进行分类、热点问题挖掘和答复意见的评价进行研究,有助于了解群众需求和政府可以更好的服务群众。

由分析结果可以得出,群众留言可以分为城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生七个主题。在本实验中,对比了四个机器学习模型:逻辑回归、多项式朴素贝叶斯、线性支持向量机、随机森林。选择了使用分类效果更好的 LinearSVC 模型。从混淆矩阵中可以得出“交通运输”类预测最准确,“城乡建设”预测的错误数量较多。

在对热点问题挖掘时,发现时间间隔因素对最大主题数影响很小,因此将时间间隔因素忽略。点赞数和反对数都代表了对留言的不同看法,同时代表了问题的热度,将点赞数和反对数加和处理来分析热点问题。

答复意见的评价需要在答复的相关性、完整性、可解释性等诸多角度进行考虑。本文在答复意见与留言详情内容相关性方面进行了研究。使用 fuzzyzuzzy 的不同比率来描述字符串之间的相似程度,数值越大,相似程度越高。用问题详情和答复意见构建 LDA 模型,计算两者之间的文本相似度,这样就能比较出留言详情和答复意见的相关程度。针对相关部门对留言的答复存在若干问题与不足,需要分析其内在原因,在此基础上,进一步提升网络问政主体的素养,疏通网络问政主体的渠道,增强网络问政的实际效果。

参考文献

- [01] 罗辉停. 基于 CRP 模型的评论热点挖掘研究修正版[J]. 技术与创新管理, 2012:166-169
- [02] 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
- [03] 余传明.基于 LDA 模型的评论热点挖掘 原理与实现[J]. 情报理论与实践 2010:103-106
- [04] 曾依灵, 许洪波. 网络热点信息发现研究[J]. 通信学报, 2007 ,28(12):141-145
- [05] 丁晟春, 蔡骅.在线评论信息挖掘研究[M].北京:科学出版社, 2014:32-34
- [06] Topic Modeling with Gensim (Python)
<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#17howtofindtheoptimalnumberoftopicsforlda>