“智慧政务”中的文本挖掘应用

摘要：近年来，随着大数据、云计算、人工智能等互联网技术的发展和盛行，群众与政府的互动与交流愈发网络化。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群众并与其进行互动的基本途径, 这些网络问政平台的兴起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，巨大的信息量给政府解决问题增加了一定的困难。为了提高政府的管理水平和施政效率，创建基于自然语言处理技术和文本挖掘方法的“智慧政务”系统已经被认为是中国特色社会主义行政治理的创新和经济发展的新常态和新趋势。

针对问题 1, 首先在 Python 语言环境中对附件 2 的数据进行数据清洗、分词、去停用词和绘制词云图等一系列文本预处理操作，计算文本的 TF-IDF 特征值，将预处理过后的数据转换成词频 TF-IDF 向量。通过对 4 种机器学习模型 (Logistic Regression、(Multinomial) Naive Bayes、Linear Support Vector Machine 和 Random Forest）的对比，最终选择了效果最好的 LinearSVC 模型，该模型的 F-Score 值达到了 0.90，模型整体效果很好。

针对问题 2, 首先对附件 3 中的数据进行文本预处理操作, 绘制高频词云图。然后建立 LDA 模型，根据困惑度越小模型越好的原理，确定最优主题数为 12，然后根据主题对群众留言进行分类，并根据热度排行 Reddit 算法建立热度评价指标，给出热度评价结果。结果显示，噪音扰民和强制学生去定点企业实习是热度指数最高的两个主题。

针对问题 3，与上述方法相同，先对附件 4 中的文本数据进行预处理。从留言详情和答复建议的词云图和前 100 个高频词可以看出, 政府对留言详情中所涉及到的问题基本都进行了相应的调查，答复意见都是围绕留言主题而回复的，基本解决了留言中所提到的问题。说明其完整性和可解释性都较好。通过对留言详情和答复意见的相似度的计算，相似度值都大于 0，说明答复意见和群众留言之间存在相关性。

关键词：网络问政；文本挖掘；群众留言；Python

1 文本挖掘目标

1.1 文本挖掘背景

近年来，随着互联网技术的发展和盛行，大数据时代已悄然而至，群众参与政府工作的途径也越来越多，于是群众留言信息成为网络问政的产物。互联网的普及加快了网络问政的出现，它是信息技术快速发展和群众积极性不断提高的产物，它不但解决了群众问政的的时空障碍，而且可以让群众随时了解政治动态并直接与政府沟通。2008 年，胡锦涛书记强调网络是了解民意的重要途径，从此，网络问政开始兴起。2009—2011 年，温家宝总理曾三次通过网络与群众交流，促进了网络问政的发展。此后，各地政府借助网络平台广泛了解群众的参政意见，从而更好地为群众解决问题。与此同时，网络问政也成为了学者的聚焦点。李传君、李怀阳学者[1]通过分析政府回应网络问政存在的问题，提出了构建良性的政府回应机制的相关建议；孟天广、赵娟学者[2]探讨了网络问政回应制度在我国的应用、扩散发展态势、制度管理体系设计和其运行管理模式,考察了政府在多元化的制度管理模式下的网络问政回应管理绩效,为进一步建设现代化的中国式具有回应性的政府提供了理论和实践参考。沙勇忠[3]等学者探究政民互动行为对网络问政效果的直接影响，运用文本挖掘和机器学习等方法，利用数据探索作为推论——统计分析检验的"数据驱动"研究理论模式，识别和分析描绘了网络问政中社会公众与其他政府以及社会组织的网络问政主体互动行为及相关话题的结构,进而讨论影响网络问政效果的因素；于君博等[4]学者针对网络问政中存在政府回应的"选择性"和"有条件性"这一问题，选取 K 市领导信箱为研究对象，对其公众诉求与政府回应情况进行了定性与定量相结合的分析。提出了探索合作治理型的回应模式，通过对制度的规范化和标准化，从而改进网络问政中的问题。

现今，随着各种网络问政平台的出现，群众留言也越来越多，而如何有效的对这些留言信息进行分类并交由相关部门处理成为提高政府工作效率的一大难题。因此，建立群众留言分类模型可以尽快将留言分派至相应的职能部门处理。通过对群众留言问题的分析，可以及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率，对“智慧政务”的建立具有重要的意义。

1.2 文本挖掘目标

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型（参考附件 1 提供的内容分类三级标签体系），并用 F-Score 方法对分类方法进行评价。

根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类, 定义合理的热度评价指标，并给出评价结果。

针对附件 4 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，并尝试实现。

2 问题分析

2.1 问题一的分析

对附件 2 的数据进行分析，进行数据清洗、分词和去停用词等文本预处理操作，然后根据分词后的结果画出高频词的词云图。计算文本的 TF-IDF 特征值, 将文本数据用向量表示。用得到的词频 TF-IDF 向量训练分类器，通过比较不同分类模型的结果，选择一个最为合适的多分类模型，用 F-Score 对该分类模型进行评价，从而得出该分类模型的效果。

2.2 问题二的分析

针对问题 2，根据附件 3，首先对文本进行去重、去停用词和分词等一系列预处理操作。在 LDA 模型中，困惑度越小，模型表现越优。通过对比不同主题下困惑度的变化，确定最优主题数，训练 LDA 模型，根据确定出来的主题将群众留言进行归类，定义合理的热度评价指标，并给出评价结果。

2.3 问题三的分析

针对问题 3，根据附件 4，首先对其进行预处理操作。分别对留言详情和答复意见进行 TF-IDF 向量化处理，然后计算其相似度，通过对相似度的分析，结合群众留言高频词和答复意见高频词的分布，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评价方案。

3 模型假设与符号说明

3.1 模型假设

(1) 假设所有的数据都是真实可靠的；

(2) 假设政府对群众留言的答复率受留言主题分布的影响，热点问题的答复率未必高；

(3) 假设持有鲜明态度的比持有中性态度的群众留言更易得到政府的答复率；

(4) 假设网络问政平台得分越高的得到政府的答复率越高。

3.2 符号说明

表 1: 符号说明

F_1	F 分数	n	类别总数
R_i	第 i 类的查全率（召回率）	P_i	第 i 类的查准率（精度）
N	某关键词在文件中出现的次数	M	在文件中所有关键词的次数之和
D	总文件数目	D_w	包含某关键词的文件的数目

注: 本表未涉及的符号在文中有具体介绍。

4 数据预处理

4.1 数据清洗

由于给的都是文本数据，可能会包含一些重复的群众留言，所以要先对数据进行清洗。Python 中使用 duplicated()方法检查留言详情中的重复对象，在重复的留言中，保留一个即可，其余的用 drop_duplicates()删除。由于留言中还可能包含一些对文本分析无用的数字，使用 lambda()函数将数字替换成空，从而完成对数据的初步处理即数据清理。

4.2 分词

在文本分析中，为了使计算机快速地识别出各语句的重点，通常以词为基本单元，用计算机自动对中文文本进行词语划分，使词之间有空格。中文分词的目的就是将一个连贯的句子按照一定的分词标准将其分成一个个具有独立含义的词[5]。分词的好坏直接影响到后期模型的准确率，它是文本挖掘的基础。只有经过分词处理，才能把原始的文本数据进行向量化处理。本文利用 Python 语言环境中的 jieba 工具包对文本进行分词处理。

4.3 去除停用词

用 jieba 工具包对留言详情进行分词后，并不是所有词都是有意义的，类似的、了、非常、十分、谢谢、尊敬、啊、嗯等一些无实际意义的连词、虚词或者介词也会保存在其中，这时，就需要剔除这些无意义的词。在本文中，首先扩展了通用的停用词表，然后又自定义了一些新的停用词，将分词处理过后所得到的词与停用词表进行匹配，若匹配成功，则删除该词，

反之保留。

4.4 词云绘制

词云图是文本结果展示的有利工具，通过词云图的展示可以让留言详情文本数据分词和去停用词后的高频词在视觉上给人强烈的突出效果，使得读者一眼就可获取到主要信息。

5 问题求解

5.1 群众留言分类

（1）附件 2 文本预处理

附件 2 中共有包含 7 类一级标签的 9210 条数据，在 Python 语言环境中，对这些数据重新按标签进行排序，代码如图 5.1-1 所示，重新排序后的效果如图 5.1-2 所示：

```
#分别用a—g表示各标签
a = data1[data1['label'] == '劳动和社会保障']
b = data1[data1['label'] == '城乡建设']
c = data1[data1['label'] == '教育文体']
d = data1[data1['label'] == '卫生计生']
e = data1[data1['label'] == '交通运输']
f = data1[data1['label'] == '商贸旅游']
g = data1[data1['label'] == '环境保护']
# 将a—g拼接起来 因为列数相同，所以用纵向拼接（axis=0）
data1_new = pd.concat([a, b, c, d, e, f, g], axis=0)
```

图 5.1-1：数据按标签顺序重新排序代码

Index	num	user	topic	time	message	label
5149	6	A00043564	为什么处理社保…	2020/1/7 21:…	…	劳动和社会保障
5150	250	U0002976	240元/年的人…	2012/5/3 11:…		劳动和社会保障
5151	251	U0002976	关于落实精简退…	2016/2/2 15:…		劳动和社会保障
5152	260	U0004362	希望西地省人事…	2013/8/2 9:4…		劳动和社会保障
5153	265	U000872	省、市社保局要…	2010/12/22 9…		劳动和社会保障
5154	284	U000388	退休工人加工资	2011/1/9 20:…		劳动和社会保障
5155	456	U000678	诉求解决血防战…	2011/2/26 17…		劳动和社会保障
5156	466	A00081652	咨询A市楚税社…	2019/11/17 1…	…	劳动和社会保障
5157	509	A00079552	咨询小孩城乡居…	2019/11/12 1…	…	劳动和社会保障
5158	555	A00097615	公司不交公积金…	2019/11/7 0:…	…	劳动和社会保障

图 5.1-2：按标签重新排序后的部分文本数据

使用 duplicated()方法检查留言详情中的重复对象，结果显示有 158 条重复对象，删除

重复对象（保留第一个）后，还剩 9052 条数据。文本去重代码如图 5.1-3 所示，删除重复对象后各类别的分布情况如图 5.1-4 所示。

```
#使用duplicated()方法检查留言详情中的重复对象
print(data1_new['message'].duplicated().sum())
#删除留言详情中重复对象所在的行，保留第一个
data1_dup=data1_new.drop_duplicates(subset='message', keep='first',inplace=False)
print(data1_dup.shape)
```

图 5.1-3：文本去重代码

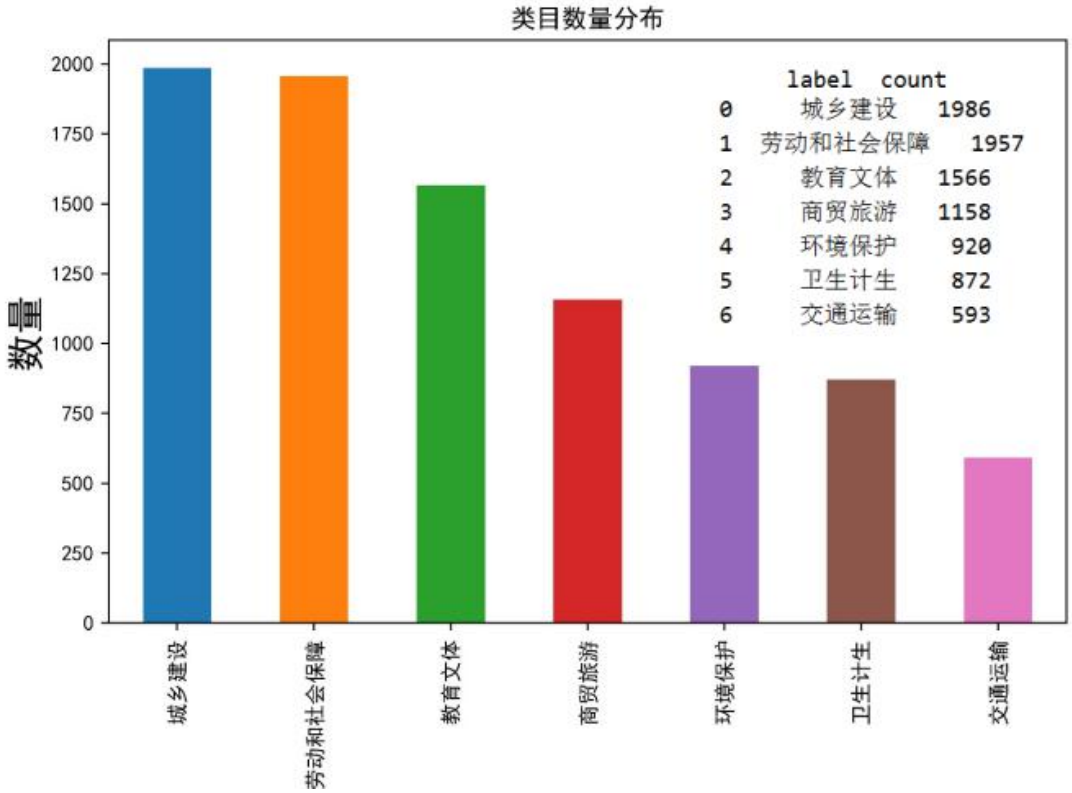


图 5.1-4：去重后各类别情况

为了方便后续分类模型的训练，我们先将 label 类转换成 Id(0-6)，Id 与 label 的对应关系如图 5.1-5 所示：

```
{0: '劳动和社会保障',
 1: '城乡建设',
 2: '教育文体',
 3: '卫生计生',
 4: '交通运输',
 5: '商贸旅游',
 6: '环境保护'}
```

图 5.1-5：Id 与 label 的对应关系

接着对留言详情中的文本内容进行处理，先将其中数字替换成空，然后用 jieba 工具包对文本数据进行分词处理，再对经过分词处理后的数据进行去除停用词操作，本文采取了通

用停用词表和自定义停用词表。通过词云图来展示对留言详情进行这一系列操作后的效果，让读者一眼便能看出主要内容。如图 5.1-6—图 5.1-12 所示。

图 5.1-6：劳动和社会保障类群众留言词云图

图 5.1-7：城乡建设类群众留言词云图

图 5.1-8：教育文体类群众留言词云图

图 5.1-9：卫生计生类群众留言词云图

图 5.1-10：交通运输类群众留言词云图

图 5.1-11：商贸旅游类群众留言词云图

图 5.1-12：环境保护类群众留言词云图

（2）计算 TF-IDF 特征值

TF-IDF 是一种常用的加权技术，主要用于信息检索与数据挖掘。TF 表示词频，IDF 表示逆文本频率指数。其计算公式如下：

$$\mathrm{TF} = \mathrm{N}/\mathrm{M}\ ,$$

$$\mathrm{IDF} = \log(\frac{\mathrm{D}}{\mathrm{D_w}})\ ,$$

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF}.$$

TF-IDF 是在单词计数的基础上，降低常用高频词的权重,增加罕见词的权重。本文使用 sklearn.feature_extraction.text. 的 CountVectorizer 和 TfidfTransformer 方法来抽取文本的 TF-IDF 的特征值。为了扩展特征集的数量，提高分类文本的准确率，本文采用了参数 ngram_range=(1,2)，除了抽取留言中的单个词语外，还抽取了词语对。文本进行预处理后的 TF-IDF 特征值的结果如图 5.1-13 所示。

```
(9052, 644863)
------------------------------
  (0, 603917)   0.11436081069064982
  (0, 21146)    0.04678065709075157
  (0, 433270)   0.04678065709075157
  (0, 5517)     0.03881340686497496
  (0, 493860)   0.04905676382861154
  (0, 121863)   0.05395908149361711
  (0, 393448)   0.06149457271612354
  (0, 388084)   0.06484964189800309
  (0, 269683)   0.03595601665264419
  (0, 563539)   0.050076535851737915
```

图 5.1-13：TF-IDF 部分特征值

从图 5.1-13 中可以看出，TF-IDF 特征值的维度是(9052, 644863)，9052 表示共有 9052 条样本数据，644863 表示特征数量，即样本中的所有词语数+词语对(相邻两个单词的组合)的总数。

将处理后的文本数据用向量形式表示并划分测试集和训练集。代码如图 5.1-14 所示。

```
#划分训练集和测试集
data1_tr,data1_te,labels_tr,labels_te = train_test_split(adata1,labels,test_size=0.2)
#创建词袋数据结构
Vectorizer = CountVectorizer()

#转化词频向量
data1_tr=Vectorizer.fit_transform(data1_tr)
print(data1_tr)
transformer = TfidfTransformer()
#训练集的样本数据的TF_IDF权值矩阵
X_tr=transformer.fit_transform(data1_tr)
X_tr.shape
print(X_tr)
#测试集
data1_te = CountVectorizer(vocabulary=Vectorizer.vocabulary_).fit_transform(data1_te)
X_te=TfidfTransformer().fit_transform(data1_te)
print(X_te)
```

图 5.1-14：对训练集和测试集向量化的代码

（3）模型的选择

首先将 message(留言详情)转换成词频向量，然后将词频向量转换成 TF-IDF 向量，最后开始训练分类器。本文尝试了 4 种不同的机器学习模型，分别为：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)和 Random Forest(随机森林)。通过比较这四种模型的准确率，从而选择最为合适的一种模型。这四种模型的准确率可视化图如图 5.1-15 所示。从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(由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。故我们选择支持向量机模型。图 5.1-16 为 4 种模型的具体准确率。

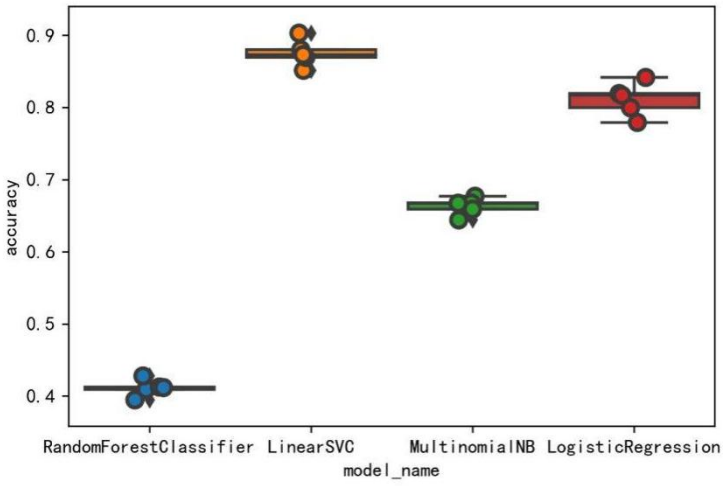


图 5.1-15：4 种机器学习模型的准确率

```
model_name
LinearSVC                0.875623
LogisticRegression       0.811550
MultinomialNB            0.663284
RandomForestClassifier   0.411516
```

图 5.1-16：4 种模型的具体准确率

（4）模型的评估

针对准确率最高的 LinearSVC 模型，查看其混淆矩阵，并显示预测标签和实际标签之间的差异。混淆矩阵如图 5.1-17 所示。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.从图 5.1-17 的混淆矩阵可以看出"教育文体"类预测最准确,只有 5 例预测错误。“交通运输”预测的错误数量最多，有 89 例。

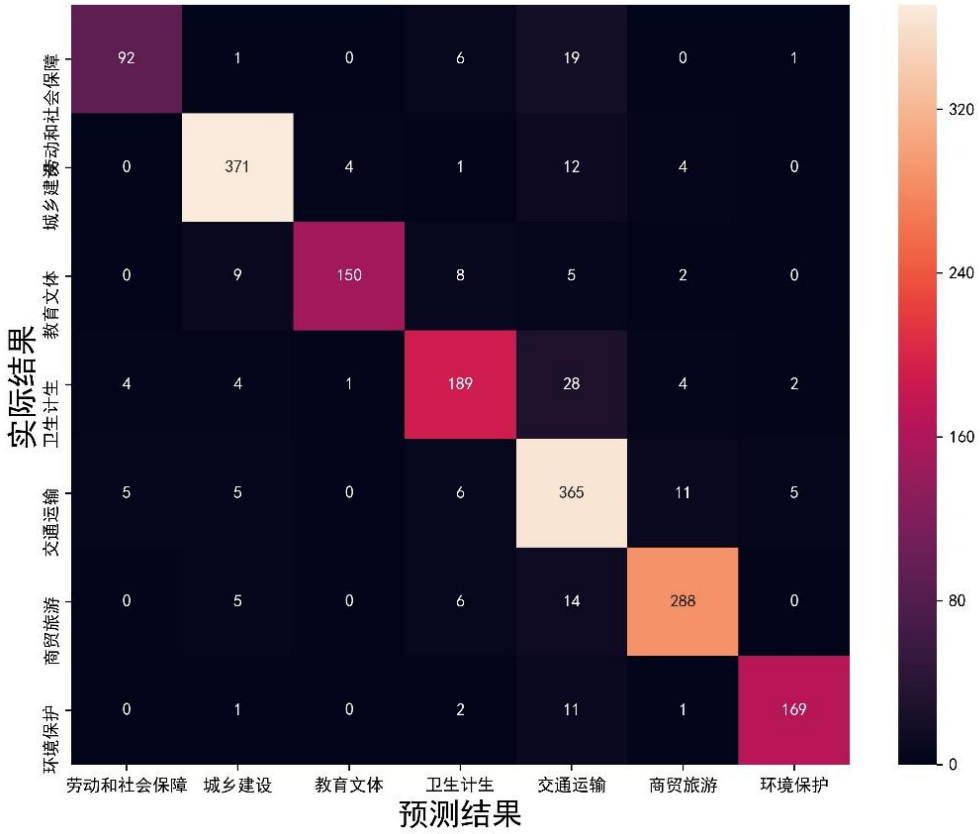


图 5.1-17: LinearSVC 模型混淆矩阵

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。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,这时候就需要借助于 F-Score、ROC 等指标来评估模型。

F-Score 的表达式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

图 5.1-18 为各个类的 F-Score。从图 5.1-20 中 F1 分数来看，"城乡建设"和‘环境保护’类的 F1 分数最高，达到 94%，“劳动和社会保障”和‘卫生计生’类 F1 分数最差，只有 84%，究其原因可能是“劳动和社会保障”和‘卫生计生’分类的训练数据较少,使得模型学习的不够充分,导致预测失误较多。从该图中还能看出 LinearSVC 模型的精度为 0.90，召回率为 0.90，F-Score 值为 0.90，整体效果很好。

accuracy 0.8967421314191054

	precision	recall	f1-score	support
劳动和社会保障	0.91	0.77	0.84	119
城乡建设	0.94	0.95	0.94	392
教育文体	0.97	0.86	0.91	174
卫生计生	0.87	0.81	0.84	232
交通运输	0.80	0.92	0.86	397
商贸旅游	0.93	0.92	0.92	313
环境保护	0.95	0.92	0.94	184
avg / total	0.90	0.90	0.90	1811

图 5.1-18: 各个类的 F1 分数

5.2 热点问题挖掘

（1）附件 3 文本预处理

附件 3 中共有 4326 条数据，用 duplicated()方法检查出留言详情中有 101 条重复对象，删除重复对象（保留第一个）后，还剩 4225 条数据。因为留言详情重复很可能是同一个人写的，所以要对其进行删除。本题为挖掘热点问题，故对留言主题进行接下来的一系列操作。再对去重后的数据进行分词、去停用词处理，本文统计了群众留言详情中前 100 个高频词，如图 5.2-1 所示。并画出了附件 3 中群众留言详情词云图，见图 5.2-2。

```
前100条高频词:
[('业主', 4028), ('部门', 1782), ('开发商', 1729), ('政府', 1662), ('居民', 1535), ('物业', 1376),
('\xa0', 1315), ('情况', 1230), ('影响', 1184), ('社区', 1075), ('学校', 1041), ('建设', 987), ('生
活', 971), ('西地省', 912), ('街道', 905), ('投诉', 893), ('规划', 882), ('时间', 862), ('房屋',
825), ('施工', 818), ('合同', 714), ('回复', 675), ('人员', 647), ('导致', 640), ('管理', 624), ('项
目', 618), ('幼儿园', 614), ('环境', 600), ('小学', 597), ('老百姓', 594), ('有限公司', 594), ('城市',
582), ('办理', 582), ('装修', 580), ('孩子', 576), ('车位', 574), ('交房', 572), ('车辆', 570), ('周
边', 565), ('学生', 563), ('政策', 559), ('房子', 558), ('您好', 549), ('道路', 548), ('国家', 548),
('购买', 535), ('单位', 528), ('违法', 526), ('噪音', 477), ('医院', 467), ('出行', 463), ('恳请',
462), ('工作人员', 455), ('村民', 450), ('通知', 448), ('找', 440), ('晚上', 437), ('文件', 436),
('销售', 436), ('电话', 431), ('质量', 416), ('中心', 415), ('楼盘', 412), ('整改', 411), ('购房',
405), ('大道', 405), ('小孩', 405), ('标准', 401), ('公示', 400), ('市民', 400), ('发现', 399), ('经
营', 396), ('收费', 396), ('教育', 394), ('发生', 393), ('地铁', 393), ('国际', 388), ('提供', 388),
('改造', 388), ('企业', 387), ('地方', 385), ('电梯', 385), ('平台', 383), ('物业公司', 383), ('工程',
381), ('只能', 380), ('违规', 380), ('拆迁', 379), ('走', 377), ('市政府', 374), ('申请', 372), ('居
住', 364), ('安全隐患', 363), ('费用', 363), ('诉求', 362), ('群众', 357), ('钱', 354), ('告知',
352), ('车', 351), ('安置', 347)]
```

图 5.2-1: 前 100 个群众留言详情高频词

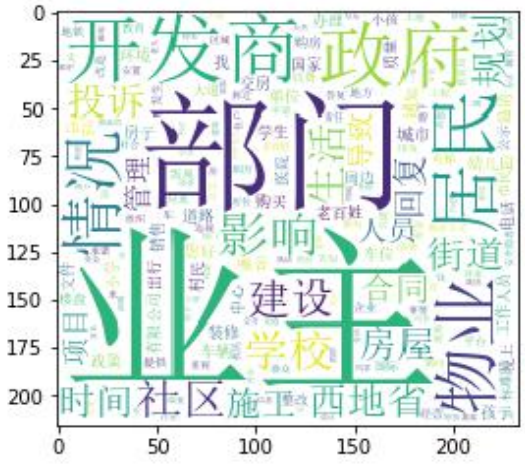


图 5.2-2: 附件 3 中群众留言详情词云图

（2）LDA 模型

主题模型（LDA）算法是文本处理与数据挖掘中一个非常重要的方法，以概率分布的形式给出文档集中每篇文档的主题，并从文本语义中提取有效的主题信息。对文字隐含主题进行建模，克服了传统信息检索中文档相似度计算方法的缺点。在进行 LDA 建模时，需要先确定主题数量 K 的值。主题数量 K 的值直接影响到最终结果的好坏。对于一个未知分布，Perplexity(困惑度)越低，则模型效果越好。从图 5.2-3 中可以看出，当主题数为 12 时，困惑度达到最低，故我们可以确定最优主题数为 12.

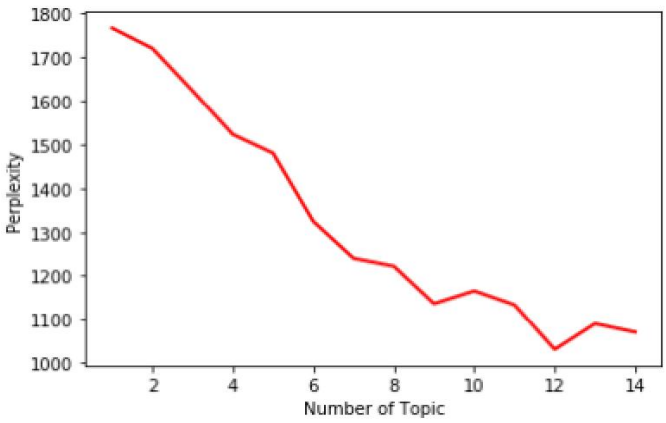


图 5.2-3；困惑度随主题数量的变化图

根据确定的最优主题数训练 LDA 模型，将某一时段内反映特定地点或特定人群问题的留言进行归类，然后建立热度评价指标。本文采用了热度排行 Reddit 算法。Reddit 算法最终得分公式：

$$\text{Score} = \log_{10}^{z} + \frac{\text{yt}}{45000}$$

其中，log_{10}^{z}表示赞成票超过反对票的数量越多，得分越高。$\frac{yt}{45000}$表示 t 越大，得分越高。发表时间对排名有很大影响，该算法使得新的话题比旧的话题排名靠前，话题的得分不会因为时间的流失而减少，但是新的话题会比旧的话题得分高。这与 Hacker New 的算法不同 （随着时间的发展降低话题的得分）。经过计算，得到的热点问题如表 2 所示。部分热点问题明细见表 3（其余见附件热点问题明细表）。

表 2：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	1	95.3	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	87.6	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
3	3	80.1	2019/11/13 至 2020/1/26	A 市丽发新城小区	小区附近搅拌站噪音扰民
4	4	71.3	2019/05/05 至 2019/09/06	A 市五矿万境 K9 县	房屋质量问题
5	5	63	2019/01/11 至 2019/07/08	A 市 58 车贷诈骗集团	民众询问车贷案进度

表 3: 热点问题留言明细表（部分结果）

问题 ID	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
1	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/8/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道...	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/8/26 8:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层...	1	0
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/8/26 1:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不飘...	0	
1	360107	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	2019/7/21 10:29:36	局长： 你好，A5 区劳动东路魅力之城小区的一...	3	0
1	360108	A0283523	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理!	2019/8/1 16:20:02	局长： 你好，A5 区劳动东路魅力之城小区的一楼...	6	0
1	360109	A0080252	万科魅力之城小区底层门店深夜经	2019/9/4 21:00:18	您好：我是万科魅力之...	0	0

			营，各种噪音扰民				
2	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 A 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...	0	0
2	360111	A1204455	A 市经济学院组织学生外出打工合理吗?	2019/11/5 10:31:38	一名中职院校的学生,学校组织我们学...	1	0
2	360112	A220235	A 市经济学院强制学生实习	2019/4/28 17:32:51	各位领导干部大家好，我是资 ...	0	0
2	360113	A3352352	A 市经济学院强制学生外出实习	2018/5/17 8:32:04	A 市经济学院强制 16 届电子商务跟企业物流专业实习...	3	0
2	360114	A0182491	A 市经济学院体育学院变相强制实习	2017/6/8 17:31:20	书记您好，我是来自西地省经济学院体育..	9	0
3	208714	A00042015	A2 区丽发新城附近修建搅拌站，污染环境，影响生活	2020-01-02 00:00:00	尊敬的领导：....	0	4

5.3 答复意见的评价

（1）附件 4 文本预处理

首先对附件 4 进行文本预处理, 对群众留言详情和政府答复意见进行去重、分词等操作。用 duplicated()方法检查出留言详情中有 33 条重复对象，删除重复对象（保留第一个）后，还剩 2783 条数据。本文统计了群众留言详情和答复意见中前 100 个高频词，如图 5.3-1 和图 5.3-2 所示。并画出了附件 4 中群众留言详情和答复意见词云图，见图 5.3-3 和图 5.3-4。

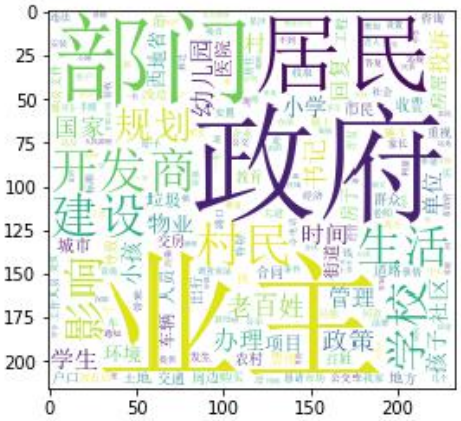
从图 5.3-1—图 5.3-4 中，我们可以看出，群众留言的高频词为“开发商”、“建设”、“投诉”等词，从这些关键词中，不难发现大多数群众留言都是围绕如：开发商违规建设、物业管理、户口等问题展开的。从答复意见中，高频词为“收悉”、“核实”、“调查”等，从这些词中，我们可以分析出政府答复意见都是经过调查研究后才给的回复，而且基本上都进行了回复，从而可以说明政府答复意见可解释性和完整性都较好。

群众留言详情前100条高频词:
[('业主', 1559), ('政府', 1214), ('部门', 1055), ('居民', 811), ('开发商', 700), ('建设', 698), ('村民', 678), ('生活', 636), ('学校', 597), ('影响', 546), ('规划', 536), ('老百姓', 517), ('幼儿园', 512), ('办理', 511), ('国家', 489), ('政策', 487), ('书记', 467), ('学生', 434), ('时间', 430), ('物业', 397), ('社区', 397), ('管理', 397), ('小学', 389), ('回复', 388), ('小孩', 386), ('投诉', 385), ('\r\n', 380), ('单位', 365), ('村', 362), ('孩子', 361), ('项目', 361), ('医院', 352), ('环境', 347), ('西地省', 343), ('城市', 341), ('街道', 333), ('房屋', 332), ('垃圾', 330), ('车辆', 329), ('人员', 327), ('房子', 323), ('收费', 314), ('周边', 313), ('费用', 310), ('导致', 301), ('道路', 300), ('户口', 292), ('群众', 290), ('地方', 284), ('合同', 282), ('出行', 275), ('前', 273), ('交通', 268), ('市民', 259), ('土地', 259), ('教育', 255), ('百姓', 255), ('交房', 250), ('农村', 248), ('工程', 247), ('施工', 243), ('车', 241), ('咨询', 240), ('购买', 239), ('找', 239), ('重视', 239), ('改造', 237), ('文件', 237), ('公交车', 231), ('安置', 229), ('教育局', 227), ('恳请', 226), ('社会', 225), ('违法', 225), ('只能', 225), ('开发', 225), ('收取', 222), ('大道', 221), ('企业', 219), ('居住', 218), ('买', 216), ('家长', 215), ('晚上', 214), ('经济', 211), ('工作人员', 210), ('钱', 210), ('路口', 209), ('走', 209), ('发生', 208), ('事情', 207), ('我家', 206), ('市场', 205), ('带来', 204), ('老师', 204), ('站', 202), ('同意', 201), ('标准', 200), ('特别', 199), ('购房', 197), ('申请', 197)]

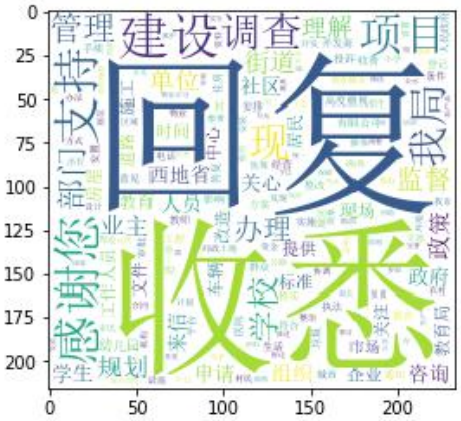
5.3-1: 附件 4 中群众留言详情高频词

答复意见前100条高频词:
[('回复', 2030), ('收悉', 1824), ('建设', 1578), ('感谢您', 1548), ('支持', 1437), ('项目', 1196), ('我局', 989), ('现', 949), ('调查', 934), ('管理', 920), ('学校', 911), ('部门', 878), ('办理', 865), ('监督', 859), ('理解', 784), ('单位', 769), ('街道', 762), ('规划', 758), ('业主', 725), ('政策', 683), ('西地省', 676), ('关心', 646), ('组织', 626), ('人员', 610), ('来信', 570), ('社区', 560), ('咨询', 558), ('学生', 547), ('申请', 530), ('政府', 509), ('现场', 496), ('教育', 494), ('标准', 493), ('道路', 467), ('时间', 454), ('居民', 447), ('教育局', 447), ('提供', 440), ('企业', 436), ('施工', 431), ('房屋', 428), ('中心', 420), ('关注', 415), ('工作人员', 408), ('改造', 405), ('文件', 404), ('车辆', 403), ('市场', 400), ('幼儿园', 399), ('高度重视', 388), ('投诉', 388), ('方案', 384), ('城市', 383), ('有限公司', 383), ('意见', 381), ('住房', 375), ('安排', 375), ('实施', 372), ('村', 372), ('生活', 370), ('经营', 368), ('登记', 361), ('核实', 359), ('工程', 357), ('收费', 354), ('条件', 354), ('整改', 349), ('人民政府', 347), ('群众', 347), ('手续', 346), ('符合', 345), ('进一步', 341), ('通知', 339), ('提出', 337), ('教师', 333), ('开发商', 328), ('答复', 326), ('影响', 325), ('电话', 324), ('医院', 321), ('垃圾', 311), ('执法', 310), ('审批', 309), ('安置', 309), ('城区', 307), ('平台', 304), ('市民', 303), ('村民', 302), ('费用', 301), ('小学', 290), ('设施', 289), ('确保', 288), ('物业', 283), ('环境', 282), ('我市', 281), ('资金', 280), ('户', 279), ('发现', 275), ('你好', 273), ('负责', 273)]

5.3-1: 附件 4 中答复意见高频词



5.3-3: 附件 4 中群众留言词云图



5.3-4: 附件 4 中答复意见词云图

（2）文本相似度分析

分析附件 4 中相关部门对留言的答复意见，本文采用计算群众留言详情和答复意见之间的相似性来判断答复的相关性。计算文本相似度的步骤：根据(1)中预处理后的数据，建立群众留言详情的语料库词典，将答复意见通过 doc2bow 转化为词袋模型，对该模型进行进一步的处理，获得新的语料库，将其通过 tfidfmodel 处理，得到 tfidf。通过 token2id 得到特征数，通过稀疏矩阵相似度的计算，从而建立索引，最终得到相似度结果。该部分的代码如图 5.3-5 所示。部分结果见图 5.3-6。

```
###文本相似度分析
from gensim import corpora,models,similarities
#用dictionary方法获取词袋（bag-of-words）
dictionary = corpora.Dictionary(message_after_stop)
#词袋中用数字对所有词进行了编号
dictionary.keys()
#编号与词之间的对应关系
dictionary.token2id
#使用doc2bow制作留言详情语料库
corpus = [dictionary.doc2bow(doc) for doc in message_after_stop]
sim={}
tfidf = models.TfidfModel(corpus)
##对每个目标文档，分析测试文档的相似度
index = similarities.SparseMatrixSimilarity(tfidf[corpus], num_features=len(dictionary.keys()))
a=range(len(reply_after_stop))
for i in a:
    reply_vec= dictionary.doc2bow(reply_after_stop[i])
    tfidf[reply_vec]
    sim[i] = index[tfidf[reply_vec]]
```

图 5.3-5：留言详情和答复意见相似度计算代码

Key	Type	Size	Value
0	float32	(2783,)	[0.23382655 0. 0. ... 0.00942786 0. 0. ...
1	float32	(2783,)	[0.0141529 0.03805237 0. ... 0. 0. 0.02657234 ...
2	float32	(2783,)	[0.00853988 0.01141298 0.33576047 ... 0. 0. 0.00418473 ...
3	float32	(2783,)	[0. 0. 0. ... 0.01413765 0. 0. ...
4	float32	(2783,)	[0.00426717 0.00545431 0. ... 0.02011033 0. 0. ...
5	float32	(2783,)	[0.02261733 0. 0. ... 0. 0. 0. ...
6	float32	(2783,)	[0.01786833 0. 0.05064777 ... 0. 0.0087748 0. ...
7	float32	(2783,)	[0.01511065 0. 0.1037294 ... 0. 0. 0.01251079 ...
8	float32	(2783,)	[0.03297013 0. 0. ... 0. 0.00426323 0.02469809 ...
9	float32	(2783,)	[0.00733438 0. 0. ... 0. 0. 0.02020578 ...
10	float32	(2783,)	[0.02617521 0. 0.01083093 ... 0. 0.06284463 0.00831565 ...
11	float32	(2783,)	[0.00138182 0. 0. ... 0. 0.00901062 0.01008033 ...
12	float32	(2783,)	[0. 0.00852714 0. ... 0. 0. 0.01117648 ...
13	float32	(2783,)	[0. 0. 0. ... 0. 0. 0.]
14	float32	(2783,)	[0.01118329 0.00311836 0.00548173 ... 0. 0. 0. ...
15	float32	(2783,)	[0.03413525 0.01688026 0.01215982 ... 0. 0.01230503 0.01683548 ...
16	float32	(2783,)	[0. 0. 0. ... 0. 0. 0.]

图 5.3-6：留言详情与答复意见相似度部分结果

图 5.3-6 中，索引(Key)对应的是群众留言详情，Value 对应的是真服答复意见和留言的

相似度情况。主对角线上的元素表示每一条留言和其对应答复意见之间的相似度。从结果中，我们可以看出，政府答复意见和对应留言情况的相似度都大于 0(主对角线元素全都大于 0)，说明政府的答复意见和群众留言详情之间有相关性。从结果中，我们还能看出相似度系数值都不高，究其原因，在实际生活中，答复意见都是根据留言内容而定的，一般来说，这两者之间虽说的是同一个问题，但一个是问，一个是答，两者之间的联系仅仅是主题相同而已，其他内容都是不尽相同的，故就造成了相似度值不高这种现象。本问仅以此来证明政府的答复意见和留言内容是相关的。

6 总结

本文的主要目的是利用文本挖掘和机器学习技术建立对群众留言的多分类模型，并对该模型进行评价。针对问题 1，通过对 4 种机器学习模型的对比分析，本文选择了 LinearSVC 分类模型，其 F1 值达到了 0.90，分类效果很好。针对问题 2，通过进行 LDA 建模，首先确定了最优主题数为 12，然后根据确定的主题数对留言详情进行划分，通过 Reddit 算法建立热度评价指标，结果显示噪音扰民和强制学生去定点企业实习是热度指数最高的两个问题。政府应多花费一些精力去解决这些热点问题。针对问题 3，通过计算留言详情和政府答复意见的相似度，发现相似度值都大于 0，从而说明了政府的答复意见和群众的留言都是相关的。通过查看词云图和高频词，可以明显看出，答复意见中的前几个高频词中有“回复”、“收悉”、“调查”等，说明政府对留言详情中所涉及到的问题基本都进行了相应的调查并给予了答复，这些高频词表明了政府对所搜集到的留言都进行了答复，并进行了相应的调查去核实，从这些高频词反映了政府答复意见的完整性和可解释性都较好。

参考文献:

[1]李传军,李怀阳.公民网络问政与政府回应机制的建构[J].电子政务,2017,No.169,77-84.

[2]孟天广,赵娟.网络驱动的回应性政府:网络问政的制度扩散及运行模式[J].上海行政学院学报,2018,v.19;No.100,37-45.

[3]沙勇忠,王峥嵘,詹建.政民互动行为如何影响网络问政效果?——基于“问政泸州”的大数据探索与推论[J].公共管理学报,2019,v.16;No.62,20-32+174.

[4]于君博,李慧龙,于书鳗.“网络问政”中的回应性——对 K 市领导信箱的一个探索性研究[J].长白学刊,2018,No.200,71-80.

[5]李少温.基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[C].华中科技大学,2019.