“智慧政务”中的文本挖掘应用

摘要：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，人工处理方法已经显得远远不够。此时智慧政府运用而生，通过运用互联网、大数据等现代信息技术，加快推进部门间信息共享和业务协同，简化群众办事环节、提升政府行政效能、畅通政务服务渠道，解决群众“办证多、办事难“等问题。

对于问题一：首先是初步数据分析，有了大概了解后开始进行文本预处理，即删除数据中的标点符号，特殊符号。接着进行文本留言分词时，即把连续的字序列按照一定的规范重新合成词序列的过程，完成了文本留言分词后，将运用机器学习的方法，计算处理后的文本数据（cut_review）的 TF-IDF 特征值，把留言例的关键字进行提取分类。最后用 F-score 对分类方法评估。

对于问题二：由于要提取留言内容中的热力问题，可以想到使用留言内容中使用频率高的具有代表性的词，词组，短句作为参照，将出现频率高的词组等抽取出来，用来反映留言内容的热度，以便建立热度评价指标。运用到了 TextRank 算法。同时此问题运用了简单的工具 excel 将点赞多的问题进行排序筛选。结合两者的结论出最终结果。

对于问题三：对于此问题采用较为简单的 excel 进行数据筛选，首先大致分析数据，将一看就是无效留言列举出来。然后利用筛选功能将其无效的答复筛选除去。还进行时间的分析，通过对时间进行了纵向对比。结合回复的内容好坏和回复时间的速度，找到想要的答案。

关键词：智慧政府 深度挖掘 留言答复 自然语言处理 TextRank TF-IDF 特征值

目录

一. 引言

1.研究背景

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。 智慧政务即通过“互联网+政务服务”构建智慧型政府，利用云计算、移动物联网、人工智能、数据挖掘、知识管理等技术，提高政府在办公、监管、服务、决策的智能水平，形成高效、敏捷、公开、便民的新型政府，实现由“电子政务”向“智慧政务”的转变。运用互联网、大数据等现代信息技术，加快推进部门间信息共享和业务协同，简化群众办事环节、提升政府行政效能、畅通政务服务渠道，解决群众“办证多、办事难“等问题。此论文则是在此基础上进行的研究

2.研究意义

通过数据挖掘，可以处理文本量更大的数据，同时减少了人工的成本，使得政府的效率加大，也更智能化人性化，数据挖掘本质上像是机器学习和人工智能的基础，它的主要目的是从各种各样的数据来源中，提取出超集的信息，然后将这些信息合并让你发现你从来没有想到过的模式和内在关系。这就意味着，数据挖掘不是一种用来证明假说的方法，而是用来构建各种各样的假说的方法。我们也根据研究这些问题得到方法拓展了我们的知识面。同时通过数据挖掘来丰富我们的知识面，使技能得到提高。

二．问题的解决及实现

2.1 总体流程

本论文的分析流程大致可分为以下四步：

第一步：获取分析所用的原始数据（所给数据）；

第二步：对获取的数据进行基本的处理操作，包括数据预处理，中文分词，停用词过滤等操作；

第三步：文本评论数据经过处理后，运用多种手段对评论数据进行多方面的分析；

第四步：从对应结果的分析中获取文本评论数据中有价值的内容。

2.2 问题求解

2.2.1 问题一过程及结果

2.2.1.1 数据分析

所给数据（附件 2）中包含了七个一级类别，这些数据都是来自于网络问政平台的留言，为了让留言信息与相关部门相匹配，我们要把不同的留言数据分到不同的类别中去，使留言得到较好的分类使政府好处理。且每条数据只能对应七个类中的一个类。首先我们统计了一下各个类别的数据量，各个类别的数据量分别为：城乡建设 2009 条，劳动和社会保障 1969 条，教育文体 1589 条，商贸旅游 1215 条，环境保护 938 条， 卫生计生 877 条， 交通运输 613 条。

如下图 1.1 所示：

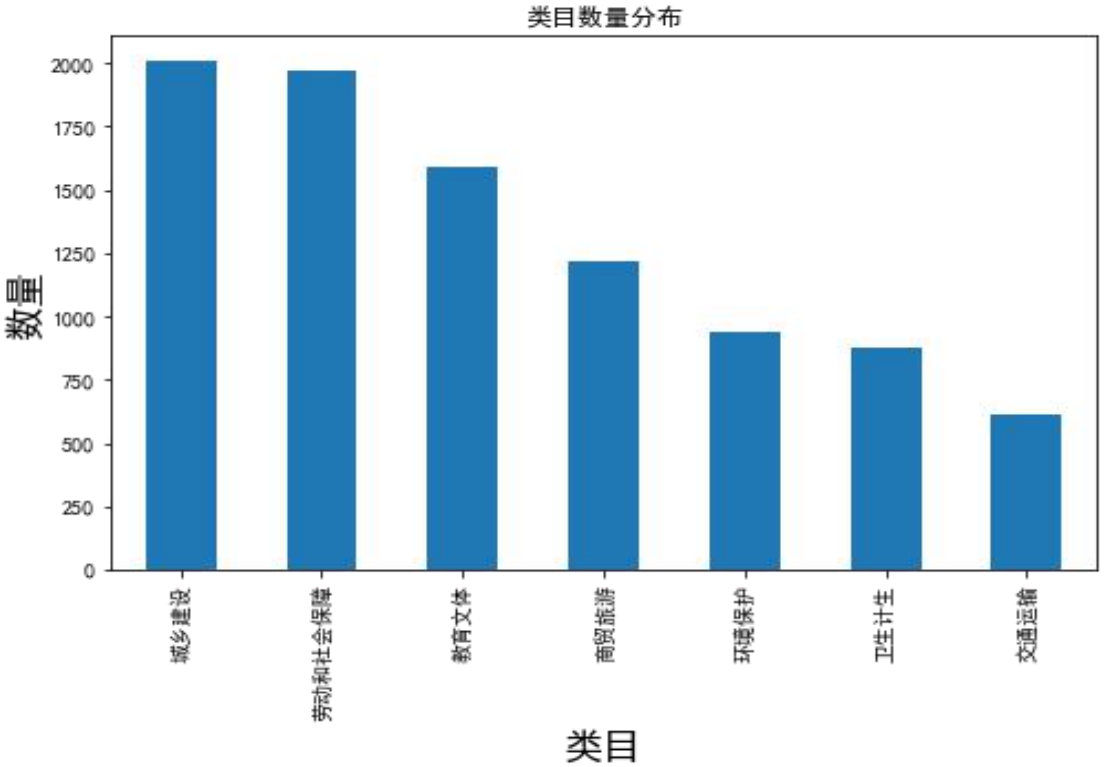

图 1-1

2.2.1.2 文本预处理

因为发现的留言内容都是文字，所以要对这些文字进行一些预处理工作，首先需要删除数据中的标点符号，特殊符号。结果如图 2.1 所示：

删除符号前后的实例:

	cat	review	cat_id	\
0	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	0	
1	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	0	
2	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	0	
3	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	
4	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	
5	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t我在2015年购买了盛世耀凯小区17栋3楼，...	0	
6	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t由于西地省地区常年阴冷潮湿的气候，加之近年气...	0	
7	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：您好！家住A市A3区桐梓坡西路...	0	
8	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t我们是梅家田社区辖区内的小区居民，我们每年都...	0	
9	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的A市政府领导：您们好！我是A市A3区魏...	0	

	clean_review
0	A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工...
1	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...
2	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...
3	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...
4	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...
5	我在2015年购买了盛世耀凯小区17栋3楼4楼两层共计2千平方一直以来我们按时足额缴纳物业费...
6	由于西地省地区常年阴冷潮湿的气候加之近年气候逐渐更加恶劣地处月亮岛片区近年规划有楚江集中供暖...
7	尊敬的胡书记您好家住A市A3区桐梓坡西路可可小城的居民长期以来经常停水小区业主业委会多次找物...
8	我们是梅家田社区辖区内的小区居民我们每年都依法依规向小区物业公司交纳了城市垃圾处理费我也认为...
9	尊敬的A市政府领导您们好我是A市A3区魏家坡巷的业主多年以来我们小区的脏乱差多次向社区反映都...

图 2-1

2.2.1.3 文本留言分词

因为在中文文本中，只有字，句和段落能够通过明显的分界符进行简单的划分，所以对于“词”和“词组”来说，他们的边界模糊，没有一个正式的划分原则。因此，进行中文文本挖掘时，应对文本分词，就是把连续的字序列按照一定的规范重新合成词序列的过程。

分词的结果对于后续的文本挖掘算法有着不容忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。

该库共提供了三种分析模式，本文选取了其中的精确模式来进行分词，在分词的过程中顺便去除了停用词。因为这些无意义的常用词（停用词）对分析预测文本的内容没有任何的帮助，反而会增加计算的复杂度和增加系统开销，所以在使用这些文本数据之前必须要把它们清理干净。

结果示例如图 2.2 如下：

分词and去除停用词后的实例：

	cat	review	cat_id	\
0	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	0	
1	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	0	
2	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小...	0	
3	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	
4	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\tA1区A2区华庭小区高层为二次供水，楼顶水箱...	0	
5	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t我在2015年购买了盛世耀凯小区17栋3楼，...	0	
6	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t由于西地省地区常年阴冷潮湿的气候，加之近年气...	0	
7	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的胡书记：您好！家住A市A3区桐梓坡西路...	0	
8	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t我们是梅家田社区辖区内的小区居民，我们每年都...	0	
9	城乡建设	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的A市政府领导：您们好！我是A市A3区魏...	0	

	clean_review \
0	A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工...
1	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...
2	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...
3	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...
4	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...
5	我在2015年购买了盛世耀凯小区17栋3楼4楼两层共计2千平方一直以来我们按时足额缴纳物业费...
6	由于西地省地区常年阴冷潮湿的气候加之近年气候逐渐更加恶劣地处月亮岛片区近年规划有楚江集中供暖...
7	尊敬的胡书记您好家住A市A3区桐梓坡西路可可小城的居民长期以来经常停水小区业主业委会多次找物...
8	我们是梅家田社区辖区内的小区居民我们每年都依法依规向小区物业公司交纳了城市垃圾处理费我也认为...
9	尊敬的A市政府领导您们好我是A市A3区魏家坡巷的业主多年以来我们小区的脏乱差多次向社区反映都...

	cut_review
0	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
1	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
2	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公...
3	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉...
4	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉...
5	2015 年 购买 盛世 耀凯 小区 17 栋 楼 楼 两层 共计 平方 足额 缴纳 物业费...
6	西地省 地区 常年 阴冷 潮湿 气候 近年 气候 恶劣 地处 月亮 岛 片区 近年 规划 楚...
7	尊敬 胡书记 您好 家住 市 A3 区 桐梓 坡 西路 可可 小城 居民 停水 小区业主 业...
8	梅家田 社区 辖区 小区 居民 依法 依规 小区 物业公司 交纳 城市 垃圾处理 费 环卫局...
9	尊敬 市政府 领导 您们好 市 A3 区 魏家坡巷 业主 多年 小区 脏乱差 社区 得不到 ...

图 2-2

2.2.1.4 找出主题词

运用机器学习的方法，计算处理后的文本数据（cut_review）的 TF-IDF 特征值，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 意思是词频（Term Frequency），IDF 意思是逆文本频率指数（Inverse Document Frequency）。TF-IDF 是在单纯计

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“新型冠状病毒”两个词，那么后者更能体现文章的中心思想，而前者是常见的高频词，它不能表达文章的主题思想。所有“新型冠状病毒”的 TF-IDF 值要高于“中国”的 TF-IDF 值。本文使用 TfidfVectorizer 方法来抽取文本的 TF-IDF 的特征值。除了抽取留言中的每个词语外，我们还抽取了每个词相邻的词组成“词语对”，拓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。

```
# 计算cut_review的TF-IDF特征值

from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(norm='l2', ngram_range=(1, 2))
features = tfidf.fit_transform(df.cut_review)
labels = df.cat_id
print(features.shape)
print('------------------------------')
print(features)
```

```
(9210, 695514)
------------------------------
  (0, 409709)   0.15622076750032005
  (0, 406948)   0.15622076750032005
  (0, 409766)   0.15622076750032005
  (0, 617318)   0.15622076750032005
  (0, 274459)   0.14104922086949562
  (0, 637788)   0.15622076750032005
  (0, 99530)    0.14950725116664162
  (0, 634622)   0.15622076750032005
  (0, 435693)   0.15622076750032005
  (0, 49251)    0.15622076750032005
  (0, 232893)   0.15622076750032005
  (0, 413457)   0.14950725116664162
  (0, 681671)   0.13547805873688806
  (0, 275701)   0.15622076750032005
  (0, 493309)   0.15622076750032005
  (0, 677215)   0.15622076750032005
  (0, 325818)   0.14104922086949562
  (0, 600985)   0.14474393552322437
```

图 2-3

由上图 2.3 可看到我们 feature 的维度是（9210，695514），9210 表示我们总共有 9210 条评价数据，695514 表示我们的特征向量，这包括所有留言中的所有词语数和词语对（相邻两个词语的组合）的总数。

接下来我们使用卡方检验的方法找出了每个分类中关联度最大的两个词语和词语对。卡方实验是一种统计学的工具，用来检验数据的拟合度和关联度。

得到的结果如下所示：

```
# '交通运输':
  . Most correlated unigrams:
        . 快递
        . 出租车
  . Most correlated bigrams:
        . 的士 司机
        . 出租车 司机
# '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
# '卫生计生':
  . Most correlated unigrams:
        . 医生
        . 医院
  . Most correlated bigrams:
        . 社会 抚养费
        . 乡村 医生
# '商贸旅游':
  . Most correlated unigrams:
        . 传销
        . 电梯
  . Most correlated bigrams:
        . 小区 电梯
        . 传销 组织
# '城乡建设':
  . Most correlated unigrams:
        . 小区
        . 业主
  . Most correlated bigrams:
        . 住房 公积金
        . 公积金 贷款
# '教育文体':
  . Most correlated unigrams:
        . 学生
        . 学校
  . Most correlated bigrams:
        . 教育局 领导
        . 培训 机构
# '环境保护':
  . Most correlated unigrams:
        . 环保局
        . 污染
  . Most correlated bigrams:
        . 周边 居民
        . 环保局 领导
```

图 2-4

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。这些词和词语对能很好的反映除分类的主题。

2.2.1.5 训练分类器

这里我们使用的是朴素贝叶斯分类器 MultinomialNB，我们首先将留言内容（review）转换为词频向量，然后将词频向量转化为 TF-IDF 向量，还有另种简化的方式是直接用 TfidVectorizer 来生成 TF-IDF 向量（如前面生成 features 的过程），在这里，我们采用一般的方式将生成 TF-IDF 向量分成两个步骤：1.生成词频向量，2.生成 TF-IDF 向量。最后开始训练我们的朴素贝叶斯（MulitinomiaNB）分类器。当模型训练完成之后，我们让它预测分类一些我们自定义的留言内容，编写一个预测函数，写入自定义的留言内容，开始运行。

运行结果如图 2-5 所示：

```
# 编写一个预测函数myPredict
def myPredict(sec):
    format_sec=" ".join([w for w in list(jb.cut(remove_punctuation(sec))) if w not in stopwords])
    pred_cat_id=clf.predict(count_vect.transform([format_sec]))
    print(id_to_cat[pred_cat_id[0]])
myPredict('XX县XXX酒店外墙装修施工实在太吵！！！')
myPredict('A市高铁站的出租车太乱了，车辆管理需规范！！')
myPredict('教育局也不管管，教学资源太不公平！')
```

```
城乡建设
交通运输
教育文体
```

图 2-5

2.2.1.6 模型的选择

接下来我们尝试使用了不同的机器学习模型，并评估他们的准确率，我们使用了逻辑回归（Logistic Regression），多项式朴素贝叶斯（（Mulitnomial）Naive Bayes），线性支持向量机（Linear Support Vector Machine），随机森林（Random Forest）这四种模型。

运行结果如下图所示：

	model_name	fold_idx	accuracy
0	RandomForestClassifier	0	0.389371
1	RandomForestClassifier	1	0.400217
2	RandomForestClassifier	2	0.410743
3	RandomForestClassifier	3	0.393265
4	RandomForestClassifier	4	0.380849
5	LinearSVC	0	0.845987
6	LinearSVC	1	0.876898
7	LinearSVC	2	0.867607
8	LinearSVC	3	0.906029
9	LinearSVC	4	0.867791
10	MultinomialNB	0	0.644794
11	MultinomialNB	1	0.661063
12	MultinomialNB	2	0.655453
13	MultinomialNB	3	0.655079
14	MultinomialNB	4	0.643634
15	LogisticRegression	0	0.773861
16	LogisticRegression	1	0.816703
17	LogisticRegression	2	0.808464
18	LogisticRegression	3	0.847909
19	LogisticRegression	4	0.789989

图 2-6

可以看出随机森林分类器（Random Forest）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），线性支持向量机的准确率（Linear Support Vector Machine）最高。

2.2.1.7 模型的评估

我们选择针对准确率最高的 linearSVC（线性支持向量机）模型，查看混淆矩阵，并且显示预测标签和实际标签之间的差异。

混淆矩阵如图所示：

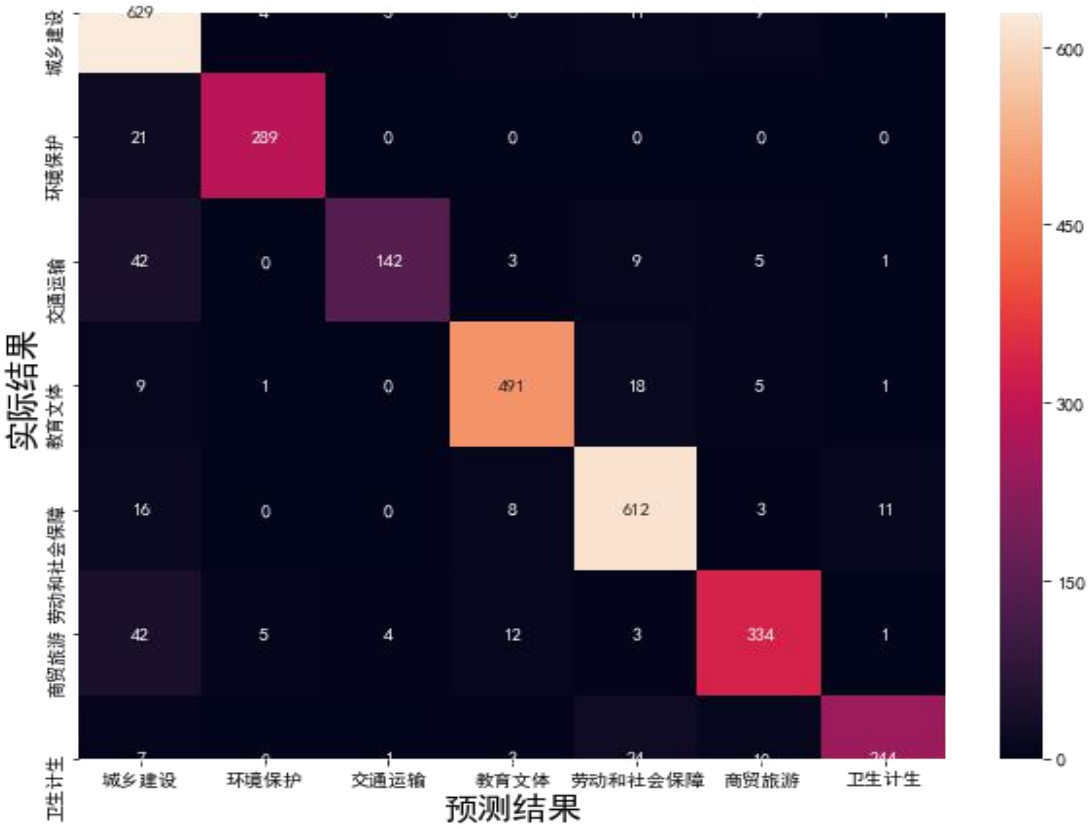

图 2-7

混淆矩阵的主对角线表示预测正确的结果的数量，除主对角线外其余都是预测错误的数量，从上面的混淆矩阵可以看出，环境保护，交通运输，商贸旅游的预测最准确，没有一例预测错误。城乡建设，劳动和社会保障预测的错误数量较多。

但是多分类模型一般不使用准确率（accuracy）来评估模型的质量，因为准确率(accuracy)不能反映除每一个分类的准确性，因为当训练数据不平衡时（有的类数据很多，有的类数据很少）时，准确率（accuracy）就不能反映除模型的实际预测精度，于是我们可以借助 F1 分数，ROC 等指标来评估模型。本文选用 F1 分数来作为评估指标。--**【思想参考文献 CSDN 博文：使用 Python 和 sklearn 的中文文本分类实战开发】**

各个类的 F1 分数如下图 2-8 所示：

	precision	recall	f1-score	support
城乡建设	0.82	0.95	0.88	663
环境保护	0.97	0.93	0.95	310
交通运输	0.95	0.70	0.81	202
教育文体	0.94	0.94	0.94	525
劳动和社会保障	0.90	0.94	0.92	650
商贸旅游	0.91	0.83	0.87	401
卫生计生	0.94	0.84	0.89	289
accuracy			0.90	3040
macro avg	0.92	0.88	0.89	3040
weighted avg	0.91	0.90	0.90	3040

2.2.2 问题二过程及结果

2.2.2.1 问题分析

问题二要求我们根据附件三所给信息找出某一时间段内群众集中反映的某些热点问题，自己定义合理的热度评价指标，给出评价结果。从而及时发现热点问题，帮助有关部门进行有针对性的处理，提高解决问题的服务效率。最后将排名前五的热点问题按照所给形式绘制表 1，另按格式绘制表 2 给出相应热点问题对应的留言信息。于是我们将问题分为两步解决，第一步是定义热度评价指标，第二步则是绘图。

2.2.2.2 定义热度评价指标

由于想提取留言内容中的热力问题，于是想到使用留言内容中使用频率高的具有代表性的词，词组，短句作为参照，将出现频率高的词组等抽取出来，用来反映留言内容的热度，以便建立热度评价指标。

我们了解到 TextRank 算法是一种基于图的用于关键词抽取和文档摘要的排序算法。它可以利用一篇文档内部的词语间档内部的词语间的共现信息(语义)便可以抽取关键词，它能够从一个给定的文本中抽取出该文本的关键词、关键词组，并使用抽取式的自动文摘方法抽取出该文本的关键句。

TextRank 算法是由 PageRank 算法改进而来的，二者的思想有相同之处，区别在于：PageRank 算法根据网页之间的链接关系构造网络，而 TextRank 算法根据词之间的共现关系构造网络；PageRank 算法构造的网络中的边是有向无权边，

而 TextRank 算法构造的网络中的边是无向有权边。TextRank 算法的核心公式如下，其中 ω_{ji} 用于表示两个节点之间的边连接具有不同的重要程度：

$$\mathrm{WS}(V_i) = (1-\mathrm{d}) + \mathrm{d} \times \sum_{V_j \epsilon In(V_i)} \frac{\omega_{ji}}{\sum_{V_k \epsilon Out(V_j)} \omega_{jk}} WS(V_j)$$

TextRank 算法提取关键词，关键词组和关键句的具体步骤如下：

1）将给定的文本按照整句进行分割，即；$\mathrm{T} = [S_1, S_2, \ldots, S_m]$；

2）对于每个句子 $S_i \epsilon T$，对其进行分词和词性标注，然后剔除停用词，只保留指定词性的词，如名词、动词、形容词等，即 $S_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,n}]$，其中 $t_{i,j}$ 为句子 i 中保留下的词；

3）构建词图 $\mathrm{G} = (\mathrm{V}, \mathrm{E})$，其中 V 为节点集合，由以上步骤生成的词组成，然后采用共现关系构造任意两个节点之间的边：两个节点之间存在边仅当它们对应的词在长度为 K 的窗口中共现，K 表示窗口大小，即最多共现 K 个单词，一般 K 取 2；

4）根据上面的公式，迭代计算各节点的权重，直至收敛；

5）对节点的权重进行倒序排序，从中得到最重要的 t 个单词，作为 top-t 关键词；

6）对于得到的 top-t 关键词，在原始文本中进行标记，若它们之间形成了相邻词组，则作为关键词组提取出来。

从给定文本中提取关键句时，将文本中的每个句子分别看作一个节点，如果两个句子有相似性，则认为这两个句子对应的节点之间存在一条无向有权边，衡量句子之间相似性的公式如下：

$$\mathrm{Similarity}(S_i, S_j) = \frac{|w_k | w_k \epsilon S_i \cap w_k \epsilon S_j|}{\log(|S_i|) + \log(|S_j|)}$$

S_i、S_j：两个句子

w_k：句子中的词

分子部分的意思是同时出现在两个句子中的同一个词的数量，分母是对句子中词的个数求对数后求和，这样设计可以遏制较长的句子在相似度计算上的优势。

根据以上相似度计算公式循环计算任意两个节点之间的相似度，设置阈值去掉两个节点之间相似度较低的边连接，构建出节点连接图，然后迭代计算每个节点的 TextRank 值，排序后选出 TextRank 值最高的几个节点对应的句子作为关键句。

--【该算法文献说明取自绽放的四叶草博客园】

于是用以上方法使用 python，运用 textrank4zh 模块，将附件 3 的留言详情中的关键词，关键词组，关键句提取出一部分。用来反映热度问题。根据编程，我们可以将关键句子提取出来，数据如下

2.2.2.3 问题解决

我们通过代码得出以下几个图，截图如下

```
关键句：
349 0.003506375098361072 ', '请A市加快严查学校违规上课时间', '万科魅力之城小区底居门店深夜经营，各
69 0.003500562737362312 ', '投诉A市盛世耀凯小区物业无故停水', 'A市X115等待时间太久了', '反映A4区月
572 0.003461733845832113 ', 'A4区御景龙城小区20多年都没办房产证', '请加快A市王陵国家考古公园建设',
609 0.003394345912758923 ', 'A市万科魅力之城商铺无排烟管道，小区内到处油烟味', '反映A市文景领秀小区
185 0.003375785704973595 ', 'A市梅溪正荣府自2018年底至今长期停工面临烂尾风险', 'A6区润和紫都小区停
292 0.0033355799368355358 ', 'A3区梅溪湖环湖路沿文化艺术中心路段乱停乱靠现象严重', 'A7县楚龙街道山
138 0.0033073033099136713 ', '请解决A市星沙城建设第一批拆迁户的生产安置预留用地及养老保险问题', '反
```

图 2-9

```
关键词为：
小区 0.009983698960208084
问题 0.0072943151511478425
a7 0.006628824807269341
a3 0.005830321389454624
西地省 0.005208134317043861
街道 0.004181630221405598
a2 0.0038579035065397825
社区 0.0037514287439613095
a4 0.00372858878877259
投诉 0.0031454214190545942
业主 0.003117200488322441
```

图 2-10

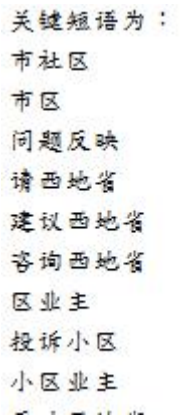

图 2-11

同时用 excel 对点赞排行进行整理，图如下：
热点问题留言明细表：

	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	208636	A00077171	区汇金路五矿万境K9县存在一系	2019/8/19 11:34:04	狗咬人，请问有人对	0	2097
3	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	纳入配套入学，A3区	5	1762
4	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	案情消息总是失望，	0	821
5	217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	小股东、苏纳弟弟苏	0	790
6	194343	A000106161	A市58车贷案警官应跟进关注	2019/3/1 22:12:30	经侦并没有跟进市领	0	733
7	263672	A00041448	小区距长赣高铁最近只有30米	2019/9/5 13:06:55	复到我如下问题：1、	0	669
8	193091	A00097965	富绿物业丽发新城强行断业主	2019/6/19 23:28:27	只提供地摊上买的收据	0	242
9	284571	A00074795	他省尽快外迁京港澳高速城区	2019/1/10 15:01:26	4、长浏高速出口，进	0	80
10	200667	A00079480	要把和包支付作为任务而不让	2019/1/16 17:01:25	基层工作者也不理解，	0	78
11	262052	A00072424	月亮岛路沿线架设110kv高压线	2019/3/26 14:33:47	第14号令《建设项目环	0	78
12	226723	A00040222	一大道全线快速化改造何时启	2019/9/15 15:31:19	改造，打通机场北通	0	66
13	272089	A00061602	A6区月亮岛路110kv高压线的	2019/4/9 17:10:01	西地省体操学校、西地	2	55
14	281898	A00096623	房云时代多栋房子现裂缝，质	2019/2/25 15:17:38	检站处理，陷入了与	5	55
15	239595	A00057814	回东六路恒天九五工厂地块，	2019/11/8 15:48:07	地区，这里潜力巨大，	0	44
16	267630	A000100648	铁3号线松雅湖站点附近地下	2019/5/22 23:37:38	东四路和东四路西侧	0	42
17	279062	A00027836	议加大A7县东六线榔梨段拆迁力	2019/1/17 19:25:45	有土地（正钢机械厂	1	42
18	209742	A00012969	区郝家坪小学什么时候能改扩	2019/3/24 21:07:12	民小学、潇湾路小学、	0	41
19	239670	A00080329	东六线以西泉塘昌和商业中心	2019/1/11 15:46:04	源置业有限公司厂房，	0	41
20	288398	A00053962	快出让星沙滨湖路以南，特立	2019/2/11 14:09:40	之间有大量的闲置土地	5	40
21	257376	A909155	关于加快修建A市南横线的建议	2019/5/10 18:01:52	交通落后的现状，拉直	0	39
22	244178	A00057874	四号线北延线“同心路站”设	2019/1/30 23:59:12	心路站设在雷峰大道	0	38
23	205217	A00040562	落实发放原A市七中01年后退休	2019/10/29 12:42:17	区下放教师待遇有关问	0	33
24	193286	A000103197	线A7县松雅西地省站西北方向	2019/4/17 11:13:12	路北侧穿越东四路和	0	32
25	226996	A00022217	A市汽车南站何时能建好？	2019/3/20 9:20:46	群众的出行和生活带	0	32
26	243808	A00053304	议将地铁7号线南延至A市生态	2019/3/6 14:20:16	东村到“尚双塘”	0	31

图 2-12

根据以上资料，综合分析得出结果
以下是热点问题表截图：

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	0.35	2019/1/14至2(	A市A4区有关车贷案件人员及	A市58车贷案警方抓紧时间(
2	2	0.24	2019/8/19	A市A5区汇金路五矿万境K9层	A市A5区汇金路五矿万境K9(
3	3	0.21	2019/4/11	A市梅溪湖金毛湾有关人员	A市梅溪湖金毛湾楼盘学区划
4	4	0.14	2019/8/23至2(	A市A4区绿地海外滩小区业主	A市A4区绿地海外滩小区距(
5	5	0.06	2019/6/19至2(	A市A2区暮云街道丽发新城物	A市A2区暮云街道丽发新城(

图 2-13

2.2.3
问题三：答复质量评价

2.2.3.1
问题分析

针对问题三，将要对答复意见的质量进行评价和计算。由此我们要建立对答复质量的指标，答复的质量指标分为：相关性，完整性，可解释性。这三者皆有则称这是一个优秀的答复，由于答复数量大，由此可以将无相关性，无完整性，无可解释性的答复从中剔除，评出优秀的答复。优秀的答复/总答复，则得出我们对此个答复系统的优秀率。

分析过程如下：

首先来粗略的看一下答复，发现有一下几个问题：

（1）无相关性：

例一：答复为日期

问：请问 B9 市带小孩打疫苗要带什么证件

答 ：2018 年 12 月 12 日

例二：

问：G2 区的政务中心预计何时搬至新址？

答：您好，你所反映的问题已转交相关单位调查处置。

例三：

问：我想在 C4 市办个甲醇和危险化工用品的仓储，请问各位领导需要离市区多远，远离多少人口才能符合条件，望各领导回复和指点！

答：好！您的咨询已收悉。以下是危险化学品经营许可证（带储存）的申办条件，如有疑问，请致电 0731-0000-00000000。申请条件：（一）依法登记注册为企业；

（二）经营和储存场所、设施、建筑物符合《建筑设计防火规范》（gb50016）、.....以上这些都是答非所问，与内容毫不相关。属于劣质回答

（2）不可解释性：

例一：

问：为何 EMS 快递身份证收费这么贵？

答：A00085038：您在《问政西地省感谢您的理解与支持！2018 年 8 月 6 日

例二：

问：B2 区江南世家水改后，水压不正常

答：网友：您好！您反映的问题已收悉，因小区水改不属于 B2 区管辖权限范围内，建议您向 B 市自来水公司反馈，联系电话:(0731)0000-00000000。感谢您对我区工作的支持和理解，祝您生活和工作愉快。

以上情况的回复并没有直接回复题主的问题，这种答复无法解决当事者问题，有推脱的责任，所以也列为不合理答复

（3）不完整性

例如：

问：咨询 I4 县农村危房改造资金

答：尊敬的网友：　　您好！您所反映的的事情已收悉，现回复如下：》.....

以上问题的答复并没有给出完整的说明，可能是工作人员疏漏，没有将答复补充完整，所以为不合理答复。

经过一轮分析，我们直接在 excel 中进行对数据的筛选:

已知总数据有 2817 条，我们先处理好处理的特征明显的数据，对于不相关性和不完整性的数据我们用筛选这一命令进行处理

1. 首先我们用筛选将答复只有日期的答复从表中去除得到条回复

2. 我们发现在某些答复中，将问题转移给他人的答复某些句的重复使用率过高，而且特征明显，如：你所反映的问题已转交相关单位调查处置；您所反映的问题，已转交相关部门调查处置。由此我们将此类数据进行移除，剩下 1862 条。

3. 答复不清楚，类似于第二条性质但是语句形势并不完全一样的答复，我们也将去除，我们找到了这种类型常见的高频词如：转交。除去剩下 1653 条。

4. 对于答复不完整，即冒号后面没有句子，这类不完整答复进行移除，得到的结果剩下 1522 条。

对于那些特征不这么明显，需要通过语意理解的答复，即答复不能说明问题，较难处理，不是通过软件筛选就能得出的，由于样本较少，我们采取调查模式，即理解对话，如答复不符合问题十分敷衍的情况进行去除，在剩下的答复中，我们找到了 38 条

2.2.3.2

结果给出

最终在2817条答复中有1484条是较为准确具体的，由此可以计算出答复的质量，内容为52.7%的优秀率.

二．开始对答复所用时间进行计算

我们还将政府答复的效率进行了时间计算对比，部分截图如下以一个星期内（30天）答复为效率为标准，一星期内答复的有条，对比总的 2800 多条，显示得出有 461 条是一个月（30 天）以上的。政府效率有待加强，部分结果排序如下

图 2-14

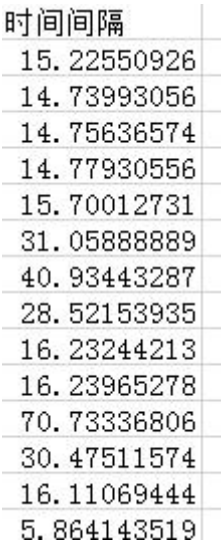

图 2-15

参考资料：

https://www.nature.com/articles/d41586-019-03013-5;

https://blog.csdn.net/asialee_bird/article/details/81486700;

https://github.com/zhbbupt/TF_IDF;

https://blog.csdn.net/yas12345678/article/details/52188287;

https://www.cnblogs.com/clover-siyecao/p/5726480.html;

https://blog.csdn.net/holysll/article/details/89396976?depth_1-utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task