“智慧政务”中的文本挖掘应用

摘要

随信息时代的发展，各类网络政务文本处理问题出现，分类时间长，处理繁琐等。本文针对文本挖掘应用的问题，提出了解决方法。

首先我们利用 TF-IDF 对群众留言进行关键词抽取，过滤掉无关词汇，按照一级标签拆分，然后利用 Python 第三方工具 jieba 工具切分留言，按照不同标签提取关键词，并计算其权重，进行分类。

对于热点问题挖掘，我们将一个问题的重要性指标拆分成三个方面，留言数量、留言时间和留言的社会反馈。其中，留言数量依据词袋模型和层次聚类模型进行聚类后得出留言较高的话题。而后根据牛顿冷却定律，利用时间戳和点赞反对数计算话题热度，从而挖掘得出热点。

针对答复意见评价，从回复时间比和回复篇幅两个方面出发。利用一级标签分类，进行不同类的比较。

关键词：TF-IDF 余弦相似度 层次聚类

绪论

问题重述

（一）在处理网络问政平台的群众留言时，按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。并使用 F-Score 对分类方法进行评价。

（二）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。请根据附件 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.

（三）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。

背景介绍

随着我国网络技术的飞速发展，网络平台上的各类社情民意留言文本剧增，从各种线上平台了解民意、解决民众困难成为我国与人民互动的重要部分。然而大量的文本数据，处理起来往往繁琐，留言杂乱很多时候没有办法移交相关部门处理。这就需要我们与时俱进，利用自然语言处理和文本挖掘的方法，建立有效的系统，针对热点问题进行及时的分类和可靠的回复。

一、 数据说明

1、 数据来源及时段

数据来源：“泰迪杯”数据挖掘挑战赛平台

数据起止年度：

附件二：2011 年至 2020 年

附件三：2017 年至 2020 年

附件四：2011 年至 2019 年

2、 数据量

附件二：9210 条留言

附件三：4326 条留言

附件四：2816 条留言

二、 群众留言分类

TF-IDF 是一种多用于信息处理和信息挖掘的加权技术，常用于过滤一些文章中无关紧要的词汇后，抽取影响整个文本的重要字词（即抽取文章的关键字词）。其分为 TF 和 IDF 两个部分。TF 表示的是某个关键词在整篇文章中出现的频率，而 IDF 表示的是整个语料库中所有的文章中出现了这个词的频率。

TF 的计算方法为：

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

IDF 的计算方法为：

$$\mathrm{id}f_i = \lg \frac{|D|}{|\{j: t_i \in \mathrm{d}_j\}|}$$

TF-IDF 计算方法为：

$$tf\mathrm{id}f_{i,j} = tf_{i,j} \times \mathrm{id}f_i$$

在此处我们将留言按一级标签拆分，每类留言利用 Python 的第三方工具 jieba 工具对留言内容进行切分，形成每类留言的词频矩阵。通过词频矩阵创建词袋。计算每个词的 TF-IDF 值后提取出每类文章关键词。利用 Python 的第三方工具 sklearn 工具的算法如下：

```
//Algorithm: createTFIDFVec()
//INPUT: dataSet 存放留言内容列表； stopwords 存放停用词列表
//OUTPUT: word 存放词汇列表； weight 存放权重列表
bufferList = []
buffer = “”
for data in dataset:
    buffer_cut = jieba.cut(data)
    for word in buffer_cut:
        if word not in stopwords:
            buffer = buffer + word + “ “
    bufferList.append(buffer_temp)
vectorizer = CountVectorizer()
transformer = TfidfTransformer()
tfidf = transformer.fit_transform(vectorizer.fit_transform(bufferList))
word = vectorizer.get_feature_names()
weight = tfidf.toarray()
return [word, weight]
```

通过分析数据，我们可知有数据的一级标签有七类，分别是城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生。最多的是城乡建设类，有 2009 篇；最少的是交通运输类，有 613 篇。通过此模型获得每类文章关键词如下（此取每类文章权重前 50 个）。将关键词回投进文章中的 F-SCORE 结果为 0.5162。

表 1：城乡建设类关键词及权重表

词语	权重
业主	0.321648
小区	0.291319
开发商	0.219782
政府	0.19191
领导	0.172871
部门	0.155011
房屋	0.152484
建设	0.140352
规划	0.133781
相关	0.13277
公司	0.130411
居民	0.130411
a 市	0.130243
公积金	0.116905
项目	0.105643
住房	0.099696
情况	0.097387

物业	0.096713
解决	0.095871
老百姓	0.095028
西地省	0.08812
工程	0.086098
希望	0.081886
生活	0.080033
房子	0.079864
城市	0.07919
影响	0.075147
单位	0.073293
施工	0.072767
工作	0.071103
时间	0.070429
管理	0.070429
办理	0.070092
建筑	0.070092
尊敬	0.06588
社区	0.063015
征收	0.062004
质量	0.061667
国家	0.059814
改造	0.058825
招标	0.057287
有限公司	0.057287
合同	0.055939
发展	0.054759
书记	0.054254
请问	0.054085
政策	0.053074
住户	0.052331
公园	0.052232
开发	0.051221

表 2：环境保护类关键词及权重表

词语	权重
污染	0.286399
村民	0.263161
居民	0.256824
排放	0.204901
环保局	0.184394
生产	0.160553
生活	0.157836
领导	0.157836

影响	0.151499
环境	0.14999
部门	0.144256
环保	0.131279
政府	0.130675
相关	0.113473
环评	0.10986
企业	0.108946
噪音	0.107758
公司	0.103816
周边	0.09778
污水	0.089969
采石场	0.089421
小区	0.089028
老百姓	0.086614
希望	0.085105
电磁辐射	0.084979
基站	0.084822
情况	0.084803
开采	0.082778
环保部门	0.080298
解决	0.079974
投诉	0.07756
西地省	0.074542
环境保护	0.074087
群众	0.071524
建设	0.06911
国家	0.066696
健康	0.066092
废水	0.06516
空气	0.064655
养猪场	0.063872
百姓	0.061565
排污	0.060778
水源	0.059447
导致	0.057642
破坏	0.057038
尊敬	0.056133
粉尘	0.055898
气味	0.055571
砖厂	0.055185
危害	0.053719

表 3：交通运输类关键词及权重表

词语	权重
出租车	0.480415
快递	0.23385
公司	0.228406
的士	0.210941
司机	0.182189
政府	0.155396
领导	0.146689
部门	0.132623
希望	0.127264
a 市	0.122576
收费	0.120721
车辆	0.117887
邮政	0.111263
公交	0.107813
交通	0.10516
客运	0.096493
城市	0.093774
滴滴	0.090221
老百姓	0.084396
相关	0.081717
打表	0.080934
出行	0.077028
管理	0.075019
道路	0.06966
路面	0.068138
驾驶员	0.067275
解决	0.066981
费用	0.066981
公路	0.066055
乘客	0.06597
取件	0.065533
出租汽车	0.063935
车主	0.063777
发展	0.063632
服务	0.062292
西地省	0.062292
物流	0.061499
情况	0.059613
营运	0.059222
时间	0.057604
10	0.056264
尊敬	0.055594
线路	0.054666

工作	0.054255
收取	0.054255
运输	0.053585
市民	0.052915
建设	0.052915
承包	0.051575
网约车	0.051148

表 4：教育文体类关键词及权重表

词语	权重
学校	0.465394
学生	0.306784
教师	0.25148
老师	0.244844
教育局	0.23128
教育	0.221219
孩子	0.195784
领导	0.167479
家长	0.155218
小学	0.127435
工作	0.111001
补课	0.10424
中学	0.095218
幼儿园	0.085174
招生	0.084276
文化	0.082696
西地省	0.079826
国家	0.076957
希望	0.073566
校长	0.072861
发展	0.068479
政府	0.067957
情况	0.061435
a 市	0.061174
小孩	0.058826
相关	0.058305
录取	0.058083
政策	0.05687
培训	0.056739
学习	0.055565
部门	0.054652
时间	0.053218
公办	0.052153
收费	0.051896

代课	0.051157
学籍	0.049607
解决	0.049435
社会	0.049174
一中	0.048708
尊敬	0.048261
考生	0.046018
工资	0.045522
班主任	0.045181
机构	0.044348
办学	0.043142
教学	0.04295
建设	0.042913
局长	0.042522
成绩	0.04187
c 市	0.041609

表 5：劳动和社会保障类关键词及权重表

词语	权重
工作	0.268141
公司	0.222587
职工	0.205056
社保	0.199062
领导	0.196424
工资	0.192971
单位	0.178362
人员	0.168933
退休	0.163886
西地省	0.162956
劳动	0.156449
企业	0.133738
政府	0.129488
政策	0.126168
国家	0.121254
劳动合同	0.115606
员工	0.10545
养老保险	0.100543
医保	0.099057
劳动者	0.098163
用人单位	0.092644
办理	0.092036
解决	0.08938
工伤	0.087658
文件	0.087654

相关	0.084732
待遇	0.083852
情况	0.081412
a 市	0.079951
部门	0.073842
医院	0.073576
享受	0.072912
新农	0.071307
社保局	0.071282
缴纳	0.070787
养老金	0.070478
关系	0.069459
12	0.068264
报销	0.068196
社局	0.067549
10	0.067201
尊敬	0.064545
时间	0.063748
社会保险	0.062669
参保	0.062611
厅长	0.060959
社会	0.060959
事业单位	0.060561
生活	0.060295
公务员	0.059631

表 6：商贸旅游类关键词及权重表

词语	权重
电梯	0.326211
公司	0.219405
传销	0.209773
部门	0.180162
业主	0.160243
相关	0.155486
西地省	0.148648
屠宰场	0.131982
领导	0.130513
政府	0.121891
a 市	0.121297
旅游	0.118919
小区	0.118027
收费	0.115926
景区	0.115252
希望	0.114162

市场	0.104946
价格	0.103162
开发商	0.101448
游客	0.091495
国家	0.09127
老百姓	0.090973
生猪	0.089512
发展	0.086216
情况	0.085324
人员	0.084135
屠宰	0.083623
消费者	0.08189
产品	0.081756
物业	0.080567
投诉	0.077892
检疫	0.076512
定点	0.075169
屠商	0.073072
10	0.072243
工作人员	0.068081
费用	0.066594
收取	0.065702
垄断	0.063621
猪肉	0.063549
销售	0.060946
电话	0.060351
购买	0.059162
维修	0.058637
居民	0.057378
管理	0.056189
经营	0.055297
企业	0.055
b 市	0.05173
质量	0.05173

表 7：卫生计生类关键词及权重表

词语	权重
医院	0.510585
医生	0.365425
生育	0.15561
患者	0.152887
领导	0.137454
西地省	0.136881
手术	0.130904

政策	0.124854
情况	0.113113
国家	0.103091
治疗	0.101276
病人	0.099081
医疗	0.09803
小孩	0.095072
检查	0.090491
护士	0.089213
独生子女	0.086296
计划生育	0.08377
工作	0.083332
孩子	0.081041
家属	0.076172
卫生室	0.07601
计生	0.073698
住院	0.07336
家庭	0.072736
希望	0.070445
户口	0.067111
政府	0.066723
药品	0.066374
再婚	0.06594
部门	0.065577
卫生院	0.064596
医师	0.064271
院方	0.06353
病历	0.061829
卫生局	0.060843
主任	0.060423
相关	0.05985
办理	0.058418
请问	0.058132
准生证	0.057858
医疗事故	0.057698
计生办	0.056728
社会	0.056413
尊敬	0.056127
10	0.054982
证明	0.054982
二胎	0.054304
病情	0.054303
老百姓	0.052691

三、 热点问题挖掘

某一个时段内群众集中留言的某一话题就可称为热点问题。在此我们对事件的重要性指标按重要程度分成留言数量、留言时间、和留言的社会反馈。

（一） 留言数量

热点问题的挖掘难点首先在于将以自然语言组成的留言归类。不同于规范化语言，自然语言的留言存在很多非结构化、非书面化、非中立化的表达。因此普通的匹配无法将事件很好的分类。在此我们引入词袋模型和层次聚类模型。

1、 构建词袋模型

要对中文文本进行聚类，首先要对文本进行预处理。预处理即对文本进行切分，去除停用词和无意义的口语化表达。例如 PASS(找个例子) 。尔后将每一个文本进一步转化为词向量，构建词袋模型。算法如下：

```
// Algorithm: createBagofWordVec()
// INPUT: dataSet 存放留言列表；stopwords 存放停用词
// OUTPUT: vocabVec 存放词袋模型
vocabSet = set()
for data in dataSet:
    data_list = buffer_cut(data, stopwords)
    for data_temp in data_list:
        vocabSet.add(data_temp)
vocabSet = list(vocabSet)                    //创建词列表
vocabVec = []
for data in dataSet:
    row = []
    for vo in vocabSet:
        if vo in data:
            row.append(1)
        else:
            row.append(0)
    vocabVec.append(row)
return vocabVec                              //返回词袋模型
```

2、 计算余弦相似度

比起普通的欧氏距离计算文本相似度，余弦相似度更符合我们的要求。在获得词袋模型后，通过计算每个文本的词向量的夹角余弦值，可以评估他们的相似度。余弦相似度的计算方式如下：

$$\cos\theta = \frac{\Sigma_1^n (A_i \times B_i)}{\sqrt{\Sigma_1^n A_i^2} \times \sqrt{\Sigma_1^n B_i^2}}$$

余弦值的范围会在[-1，1]之间，当余弦值越接近 1，则说明两个文本的词向量方向越近，文本越可能在描述同一件事；反之，则说明两个文本越不相关。算法如下：

```
//Algorithm: cosine_similarity()
//INPUT: data1 存放第一个文本的词向量；data2 存放第二个文本的词向量
//OUTPUT: sim 存放余弦相似性
a = 0
b = 0
c = 0
for i in range(len(data1)):
    a = a + pow(data1[ i ], 2)
for j in range(len(data2)):
    b = b + pow(data2[ j ], 2)
for k in range(len(data1)):                    //data1 和 data2 长度相等
    c = c + data1[ i ] * data2[ j ]
return c / (math.sqrt(a) * math.sqrt(b))
```

3、 层次聚类模型

层次聚类模型的基本思想是，先将所有的点转化为孤立结点作为初始状态，通过某种相似度计算方式来计算结点的相似性，并由相似度由高到低排列，连接最相似的结点。反复迭代这一过程，逐步重新连接所有结点。层次聚类有“自顶向下”和“自底向上”两种策略。

在此我们采用“自底向上”的聚类策略。首先给每个分配事件号。然后每个事件号作为一个类，直接以余弦相似度作为每个类的距离。寻找到距离最近的两个类，即余弦相似度最大的两个类进行合并，并对每个类重新编号。然后每个类互相计算所有类内向量的距离，形成最短距离矩阵。最短距离计算方式如下：

$$d_{min}(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} dist(x_i, x_j)$$

最短距离的计算算法如下：

```
//Algorithm: findMinDist()
//INPUT: vector 存放最小距离矩阵
//OUTPUT: mini 存放最小值的 i；minj 存放最小值的 j；minDist 存放最小值
minDist = 10000
for i in range(len(vector)):
    for j in range(len(vector)):
        if (i != j) & (vector[ i ][ j ] < minDist):
            minDist = vector[ i ][ j ]
            mini = i
            minj = j
return [mini, minj, minDist]
```

在最短距离矩阵中寻找最短距离及相应类号，对该两类进行合并。以此重复直至形成较好的分类。

```
//Algorithm: hierarchicalModel()
//INPUT: clusterList 存放分类；vocabVec 存放词袋模型
//OUTPUT: clusterList 存放分类
distVector = []
for i in range(len(clusterList)):
    row = []
    for j in range(len(clusterList)):
        minDist = distanceMin(clusterList[ i ], clusterList[ j ])
        row.append(minDist)
    distVector.append(row)
mini, minj, minDist = findMinDist(distVector)
for i in range(len(clusterList[ minj ])):
    clusterList[ mini ].append(clusterList[ minj ][ i ])
del clusterList[ minj ]
return clusterList
```

（二）留言时间与留言社会反馈

将事件聚类后，在此我们采用牛顿冷却定律计算热度指标。牛顿冷却定律是一个经验性的关系，由英国物理学家艾萨克·牛顿提出。这个定律是指所损失的热速率是与物体和其周围环境的温度具有线性关系。当物体表面与周围存在温差的时候，单位时间与单位面积失去的热量和温度差成正比，这个比例系数就是热传递系数。

由于留言同样具有时效性，和热量散失规律类似。我们采用牛顿冷却定律，即假定一个事件有一个热度，计算公式为：

当前热度分值 = 上一期得分 * exp（ − (冷却系数) * 间隔的小时数）

另外，社会反馈效果可以落实为留言的点赞数和反对数，点赞数高的事件同样具有热度，即把点赞数亦归结为热度指标之一。

四、答复意见的评价

针对相关部门对留言的答复意见，我们对答复意见的质量通过回复速度、回复篇幅比对答复进行评价。

（一） 回复速度

对于群众的留言，及时交办相关部门，对相关情况进行核实，沟通协商处理方案，并尽快回复公众，才能争取得到公众认可。通过分析数据，我们将有回复的留言分为五类，即城乡建设、环境保护、交通运输、教育文体、劳动和社会保障。从结果可看出，各类留言的回复平均用时在 20 天左右，可以较好的、及时的回复给群众。个别可能需要长时间解决的事件在事后也都得到了回复。

图 1：各类回复平均用时柱状图

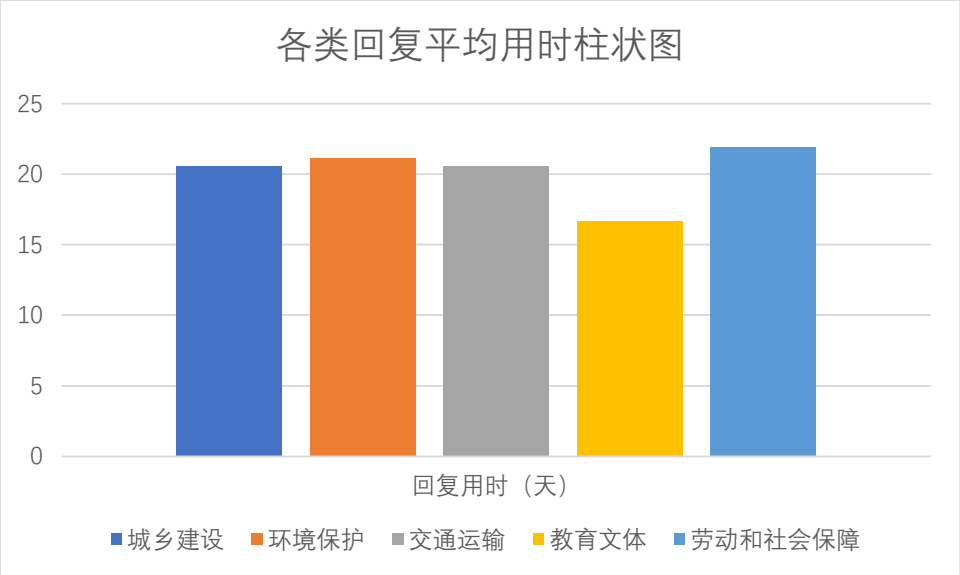


图 2：城乡建设类留言回复用时饼图

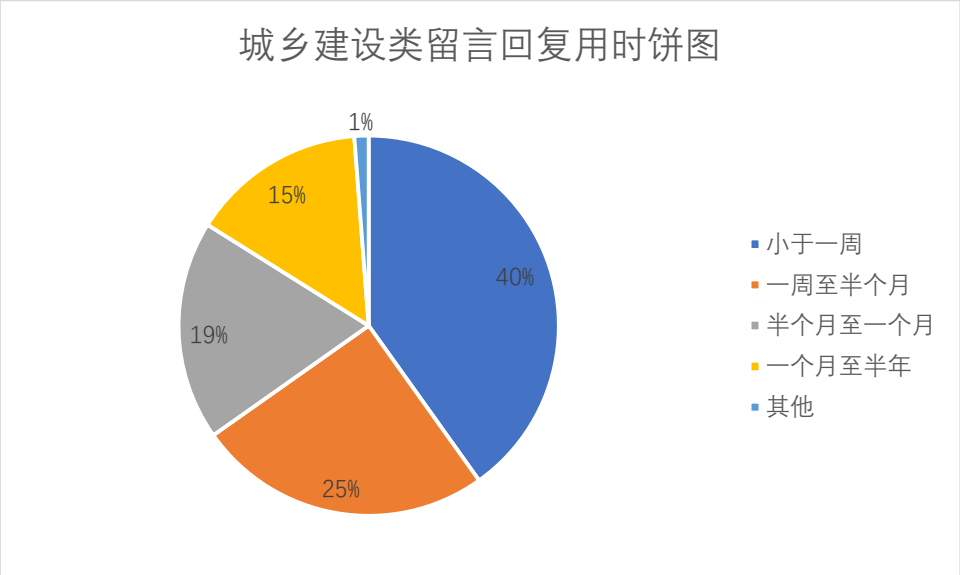


图 3：环境保护类留言回复用时饼图

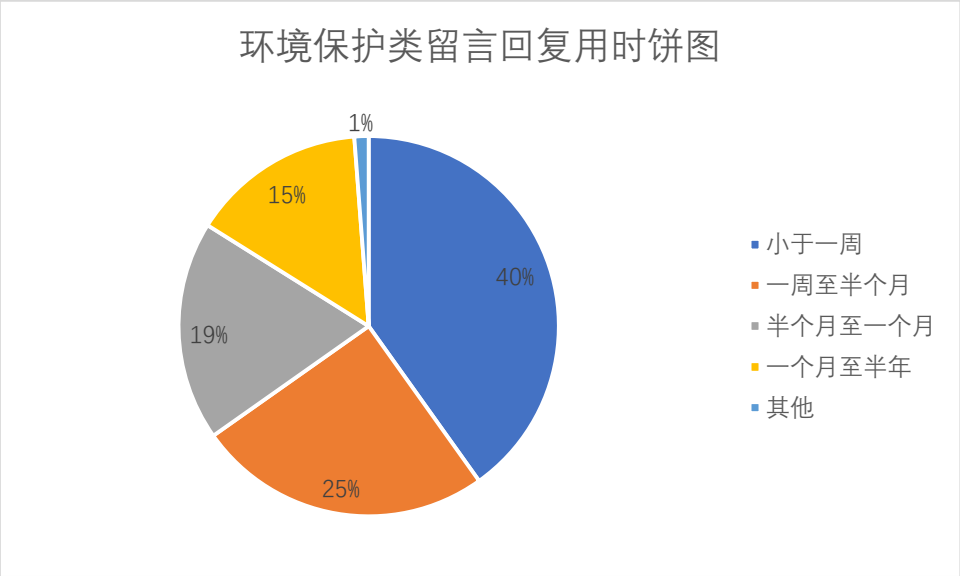


图 4：交通运输类留言回复用时饼图

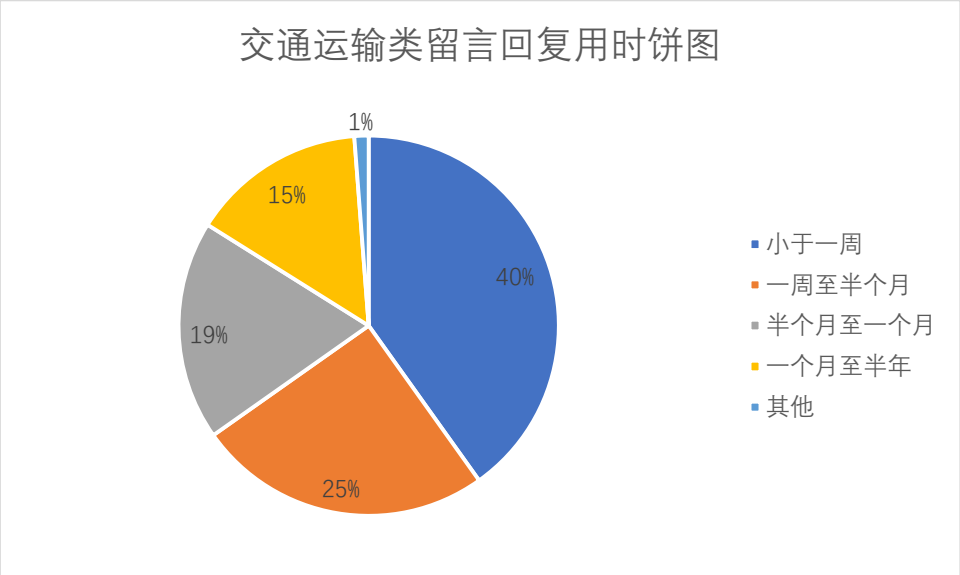


图 5：教育文体类留言回复用时饼图

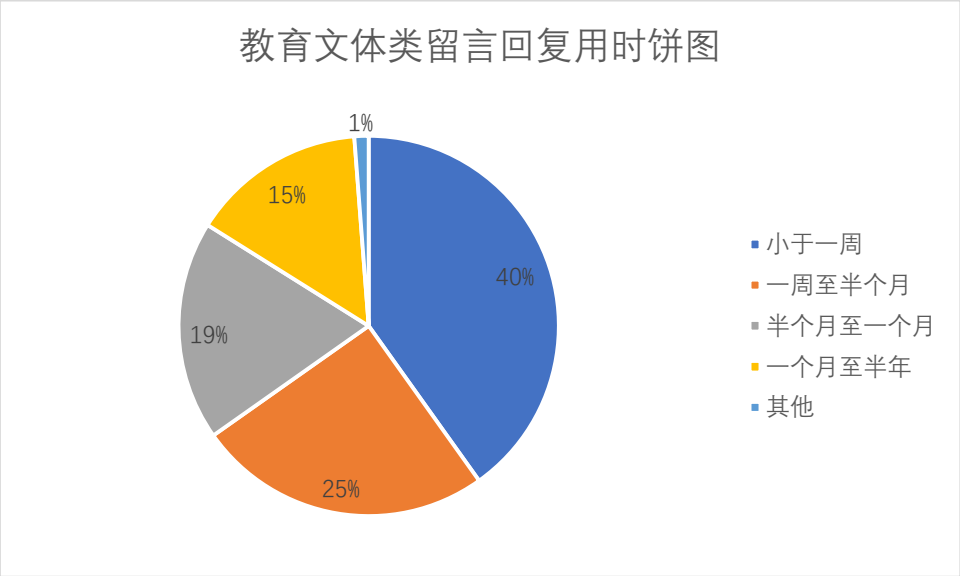
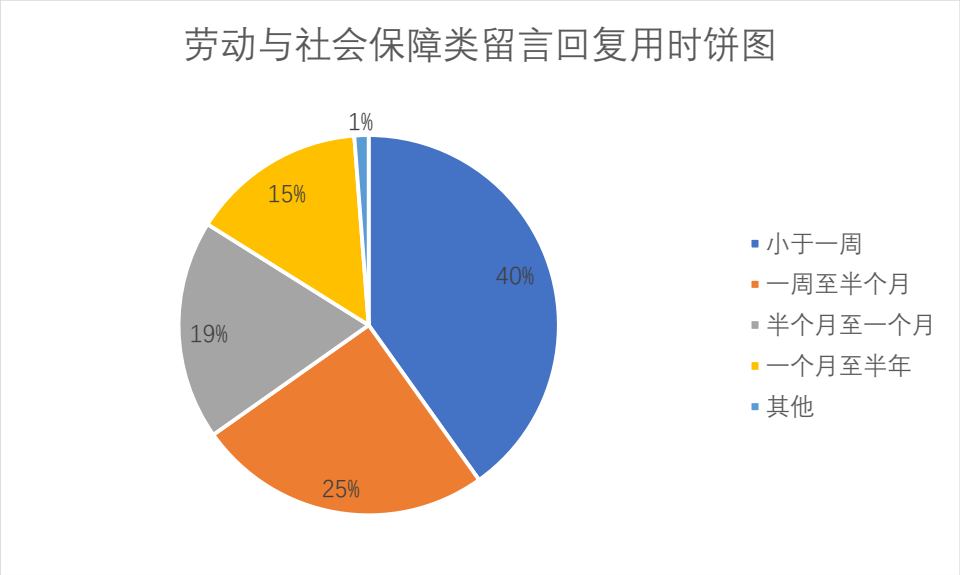


图 6：劳动与社会保障留言回复用时饼图



（二）留言篇幅比

在分析样本中，有的留言比较简单，有的留言比较复杂，有的留言则非常复杂，回复内容也是如此，这可以通过留言字数和回复字数做直观判断。通过分析得知，各类的绝大部分回复篇幅都在留言篇幅的 1.5 倍左右，可以得知回复都具有比较好的完整性。

五、参考文献

[1]蔡梨萍.基于 MATLAB 的柴油机高压喷油过程的模拟计算[D].湖北省：华中科技大学，2005.