基于自然语言处理及文本挖掘的“智慧政务”系统

绪论

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。

在本次数据挖掘的过程中，本小组主要使用 pycharm 进行自然语言的处理，并利用 jieba 库完成中文分词和文本挖掘，进一步对留言信息进行处理。

在数据挖掘的过程中，本小组首先运用 pycharm 将 csv 格式的数据导入，并运用 jieba 库中文分词筛选出各条留言的关键词，并根据关键词及分类标签，将所有留言进行分类。之后针对留言热点和政府回复，分别运用中文分词的方法提取各条信息的关键词，并完成热点留言的提前和留言答复的评价。

最后，我们将数据挖掘模型的结果输出，形成了具有三级分类的留言数据，并筛选归纳出热点留言的时间范围、地点人群、留言内容，同时从相关性、完整性、可解释性等角度完成了对留言答复的评价工作/

关键词：留言分类、热点提前、答复评价、自然语言处理、文本挖掘、中文分词

目 录

一．总体目标

二．总体分析流程

三．数据挖掘过程

3.1 问题一

3.1.1 问题重述

3.1.2 解题思路

3.1.3 解题过程

3.2 问题二

3.1.1 问题重述

3.1.2 解题思路

3.1.3 解题过程

3.3 问题三

3.1.1 问题重述

3.1.2 解题思路

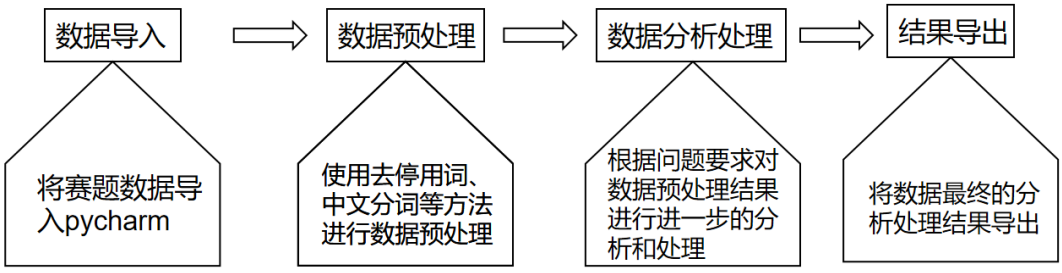
3.1.3 解题过程

四．结论

一、总体目标

本次数据挖掘的总体目标是主要通过去停用词、中文分词等自然语言处理的方法，对数量巨大的留言进行处理，提取留言中的有效性信息去除无关信息，根据各项问题的目标要求，对有效信息进行处理，提取解决问题所需要的有效信息。进而达到在实际生活中，可以通过计算机对大量留言信息进行自动处理，并得到人们所需要的处理结果，从而降低了人工成本、提高了工作效率。

二、总体分析流程



本小组数据挖掘的总体流程大致如下：首先将赛题数据导入 pycharm，对留言进行第一步去停用词和中文分词的预处理，之后根据各小问的目标要求，对预处理的数据结果进行进一步细化的分析和处理，最后根据题目要求按规定格式将留言数据的最终处理结果导出

三、数据挖掘过程

3.1 问题一

3.1.1 问题重述

根据赛题所给数据，参考附件所给分类标签体系，建立关于留言内容的一级分类标签模型。

3.1.2 解题思路

通过 gensim 库使用 word2vec 模型，引入网络资料中已建立好的 word2vec 模型，提取模型中的词向量，对附件 2 中的留言详情进行逐行读取并提取关键词，将关键词与附件 1 的三级分类进行相似度分析，进行归类，判断该留言属于哪个一级分类。

3.1.3 解题过程

1.数据导入

首先将 xlsx 格式的数据文件改为 csv 格式 ，使用 pycharm 将 csv 格式的数据文件导入 pandas 库，并将一级分类去重，语句如下

```
data = pd.read_csv('fujian2.csv', encoding='utf-8')
data1 = data['留言详情']
fen1_1 = fen1['一级分类'].drop_duplicates()#去重
```

2. 建立空表格

人工建立名称为“热点表”的空表格

3.判断词语权重

通过 word2vec 模型训练得到的词向量 improve，通过

```
#此函数计算某词对于模型中各个词的转移概率 p(wk|wi)
def predict_proba(oword, iword):
        #获取输入词的词向量
        iword_vec = model[iword]
        #获取保存权重的词的词库
        oword = model.wv.vocab[oword]
        oword_l = model.trainables.syn1[oword.point].T
        dot = np.dot(iword_vec, oword_l)
        lprob = -sum(np.logaddexp(0, -dot) + oword.code*dot)
        return lprob

#各个词对于某词 wi 转移概率的乘积即为 p(content|wi)，
#如果 p(content|wi)越大就说明在出现 wi 这个词的条件下，此内容概率越大，
#那么把所有词的 p(content|wi)按照大小降序排列，越靠前的词就越重要，越
应该看成是本文的关键词。
from collections import Counter
def keywords(s):
        #抽出 s 中和与训练的 model 重叠的词
        s = [w for w in s if w in model]
        ws = {w:sum([predict_proba(u, w) for u in s]) for w in
  s}
        return Counter(ws).most_common()
```

语句判断留言中每个词语的权重

3. 筛选留言中权重最高的词语

通过下述语句取出各条留言中权重最高的词语，并将词语写入空列表

```
list=[]
data = pd.read_csv('fujian2.csv', encoding='utf-8')
data1 = data['留言详情']
counter = 0
for index, line in enumerate(data1):
```

```
        counter  +=  1
for  i  in  range(0,  counter):
        x  =  pd.Series(keywords(jieba.cut(data1[i])))
        list.append(x[0:1])
```

4. 判断一级分类与最高权重词语的相似度

通过下述语句，将与一级分类相似度最高的词语所在留言

```
counter1  =  0
for  index,  line  in  enumerate(fen1_1):
        counter1  +=  1
for  i  in  range(0,counter):
        list_1=[]
        for  j  in  range(0,counter1):
                sim1  =  model.similarity(list[i][0],fen1_1[j])
                list_1.append(sim1)
        x=max(list_1)
        for  m  in  range(0,counter1):
                list_2=[]
                if  x  ==  list_1[m]:
                        y  =  m
        list_2.append(fen1_1[y])
```

5. 生成一级分类表

通过下述语句，生成一级分类表

```
sh  =xlrd.open_workbook(r'output.xls',formatting_info=True)
xl  =  copy(sh)
shtc  =  xl.get_sheet(0)
for  n  in  range(0,counter):
        shtc.write(n,0,list[n][0])
        shtc.write(n,1,list_2[n])
xl.save(r'一级分类.xls')
```

3.2 问题二

3.2.1 问题重述

根据赛题所给数据，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。最后按照赛题所要求格式，分别输出“热点问题表”和“热点问题留言明细表”。

3.2.2 解题思路

针对问题二的热度评价指标，本小组从留言的关键词词频入手，含有相同关键词的留言为反映同一问题的留言，关键词频率越高该留言反映的问题热度越高。因此我们运用中文分词和统计词频的方法，先逐条将留言进行分词得到留言

的关键词，之后通过统计关键词词频，使用点赞数减去反对数加上同类型留言的个数，建立起热度评价体系。

3.2.3 解题过程

1.数据导入

首先将 xlsx 格式的数据文件改为 csv 格式 ，使用 pycharm 将 csv 格式的数据文件导入 pandas 库，通过

```
data_2=pd.read_csv('fujian.csv',encoding='gbk')
data5 = data_2['留言主题']
data_liuyan=data_2['留言时间']
data_xiangqing=data_2['留言详情']
data_bianhao=data_2['留言编号']
data_dianzai=data_2['点赞数']
data_fandui=data_2['反对数']
data_yonghu=data_2['留言用户']
```

语句将赛题数据导入

2. 建立空表格

人工建立名称为“热点表”的空表格

3.判断词语权重

通过问题一 word2vec 模型训练得到的词向量 improve，通过

```
#此函数计算某词对于模型中各个词的转移概率 p(wk|wi)
def predict_proba(oword, iword):
    #获取输入词的词向量
    iword_vec = model[iword]
    #获取保存权重的词的词库
    oword = model.wv.vocab[oword]
    oword_l = model.trainables.syn1[oword.point].T
    dot = np.dot(iword_vec, oword_l)
    lprob = -sum(np.logaddexp(0, -dot) + oword.code*dot)
    return lprob

#各个词对于某词 wi 转移概率的乘积即为 p(content|wi)，
#如果 p(content|wi)越大就说明在出现 wi 这个词的条件下，此内容概率越大，
#那么把所有词的 p(content|wi)按照大小降序排列，越靠前的词就越重要，越应该看成是本文的关键词。

from collections import Counter
def keywords(s):
    #抽出 s 中和与训练的 model 重叠的词
    s = [w for w in s if w in model]
    ws = {w:sum([predict_proba(u, w) for u in s]) for w in
```

```
  s}
        return  Counter(ws).most_common()
```
语句判断留言中每个词语的权重

4.筛选留言中权重前三的词语

通过
```
guanjian=[]
count4=0
for  index,  line  in  enumerate(data5):
        count4+=1
for  j  in  range(0,count4):
        x  =  pd.Series(keywords(jieba.cut(data5[j])))
        guanjian.append(x[0:3])
```
语句取出每个句子中权重前三的词语

5.生成热点问题表

通过以下语句，依次将各句子中取出的权重前三的关键词进行查重，若某两条留言的三个关键词中有一个词语相同，即将两条留言归为描述同一问题的留言类型。每次找到同类型的留言后，即记录下同类型留言的序号，并在之后的查重和归类中剔除已经完成分类的留言。之后归纳同类型留言描述问题的主题，并将同类型留言的留言时间放入同一列表中，计算最早留言时间和最晚留言时间，并将所有类型留言的最早留言时间和最晚留言时间写入列表。通过“点赞数-反对数+同类型留言条数”计算问题的热度。最后将以上结果写入“热点问题表”并保存。
```
hot=[]
x=[]
y=[]
sh  =xlrd.open_workbook(r'热点表.xls',formatting_info=True)
xl  =  copy(sh)
shtc  =  xl.get_sheet(0)
time=[]
sum  =  0
out=[]
row=0
for  j  in  range(0,count4):
        if  j  not  in  out:
                row=row+1
        if  j<count4:
                k=0
                sum  =  0
                for  i  in  range(1,count4-j):
                        if  j  in  out:
                                continue
                        elif  guanjian[j][0][0]  ==  guanjian[j+i-1][0][0]  or  g
```

```
uanjian[j][0][0] == guanjian[j+i-1][2][0] or guanjian[j][0][0] =
= guanjian[j+i-1][1][0] or guanjian[j][1][0] == guanjian[j+i-1][
0][0] or guanjian[j][0][0] == guanjian[j+i-1][2][0] or guanjian
[j][1][0] == guanjian[j+i-1][1][0]or guanjian[j][2][0] == guanji
an[j+i-1][0][0] or guanjian[j][2][0] == guanjian[j+i-1][1][0] or
  guanjian[j][2][0] == guanjian[j+i-1][2][0]:
                                time.append(data_liuyan[j])
                               sum = sum+data_dianzai[j]-data_fandui[j]
                                shtc.write(row, 5, data5[j])
                                k=k+1
                                out.append(j+i-1)
                        else:
                                time.append(data_liuyan[j])
                             sum = sum + data_dianzai[j] - data_fandui[j]
                                shtc.write(row, 5, data5[j])
                                k = k + 1
                sum = sum + k
                x.append(max(time))
                y.append(min(time))
                hot.append(int(sum))
for j in range(0,row):
        shtc.write(j+1,0,j+1)
        shtc.write(j+1, 1, j+1)
        shtc.write(j+1,2,hot[j])
        shtc.write(j+1, 3, x[j])
        shtc.write(j+1, 4, y[j])
shtc.write(0,0,'热度排名')
shtc.write(0,1,'问题 ID')
shtc.write(0,2,'热度指数')
shtc.write(0,3,'时间范围')
shtc.write(0,5,'问题详情')
xl.save(r'热点问题表.xls')
```

6.生成热点问题明细表

通过下述语句，按步骤 5 中的方法依次将“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”写入“热点问题明细表”。

```
sh =xlrd.open_workbook(r'热点表.xls',formatting_info=True)
xl = copy(sh)
shtc = xl.get_sheet(0)
row=0
for j in range(0,count4):
        if j not in out:
                row=row+1
```

```
        if  j<count4:
                for  i  in  range(1,count4-j):
                  if  j  in  out:
                                continue
                  elif  guanjian[j][0][0]  ==  guanjian[j  +  i  -  1][0][0]
or  guanjian[j][0][0]  ==  guanjian[j  +  i  -  1][2][0]  or  guanjia
n[j][0][0]  ==  guanjian[j  +  i  -  1][1][0]  or  guanjian[j][1][0]
  ==  guanjian[j  +  i  -  1][0][0]  or  guanjian[j][0][0]  ==  guanj
ian[j  +  i  -  1][2][0]  or  guanjian[j][1][0]  ==  guanjian[j  +  i
  -  1][1][0]  or  guanjian[j][2][0]  ==  guanjian[j  +  i  -  1][0][
0]  or  guanjian[j][2][0]  ==  guanjian[j  +  i  -  1][1][0]  or  gua
njian[j][2][0]  ==  guanjian[j  +  i  -  1][2][0]:
                                shtc.write(row,  5,  data_xiangqing[j])
                                shtc.write(row,  1,  int(data_bianhao[j]))
                                shtc.write(row,  3,  data5[j])
                                shtc.write(row,  2,  data_yonghu[j])
                                shtc.write(row,  4,  data_liuyan[j])
                                shtc.write(row,  6,  int(data_dianzai[j]))
                                shtc.write(row,  7,  int(data_fandui[j]))
                                out.append(j+i)
                  else:
                          shtc.write(row,  5,  data_xiangqing[j])
                          shtc.write(row,  1,  int(data_bianhao[j]))
                          shtc.write(row,  3,  data5[j])
                          shtc.write(row,  2,  data_yonghu[j])
                          shtc.write(row,  4,  data_liuyan[j])
                          shtc.write(row,  6,  int(data_dianzai[j]))
                          shtc.write(row,  7,  int(data_fandui[j]))

for  j  in  range(0,row):
        shtc.write(j+1,0,j+1)
shtc.write(0,0,'问题 ID')
shtc.write(0,1,'留言编号')
shtc.write(0,2,'留言用户')
shtc.write(0,3,'留言主题')
shtc.write(0,4,'留言时间')
shtc.write(0,5,'留言详情')
shtc.write(0,6,'点赞数')
shtc.write(0,7,'反对数')
xl.save(r'热点问题留言明细表.xls')
```

7.结果分析

根据上述步骤得到如下“热点问题表”和“热点问题明细表”

热度排名	问题ID	热度指数	时间范围		问题详情
1	1	290	2017-06-0	2017-06-0	A市经济学院体育学院变相强制实习
2	2	28	2018-11-1	2017-06-0	在A市人才app上申请购房补贴为什么通不过
3	3	27	2019-09-0	2017-06-0	希望西地省把抗癌药品纳入医保范围
4	4	52	2019-09-2	2017-06-0	A5区劳动东路魅力之城小区临街门面烧烤夜宵摊
5	5	25	2019-09-2	2017-06-0	请给K3县乡村医生发卫生室执业许可证
6	6	96	2019-09-2	2017-06-0	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
7	7	0	2019-09-2	2017-06-0	A市能否设立南塘城轨公交站?
8	8	22	2019-10-3	2017-06-0	请求A市地铁2#线在梅溪湖CBD处增设一个站
9	9	84	2019-10-3	2017-06-0	请问A市什么时候能普及5G网络?
10	10	20	2019-10-3	2017-06-0	A市经济学院寒假过年期间组织学生去工厂工作
11	11	0	2019-10-3	2017-06-0	L市物业服务收费标准应考虑居民的经济承受能力
12	12	0	2019-10-3	2017-06-0	A市江山帝景新房有严重安全隐患
13	13	17	2019-11-2	2017-06-0	12123上申请驾驶证期满换证，一个星期了都无人受理
14	14	16	2019-11-2	2017-06-0	能否分层单独补交超面积地款?
15	15	15	2019-11-2	2017-06-0	A市魅力之城商铺无排烟管道，小区内到处油烟味
16	16	0	2019-11-2	2017-06-0	J4县供销合作社在岗失业职工追缴社保
17	17	13	2019-11-2	2017-06-0	A市能不能提高医疗门诊报销范畴
18	18	0	2019-11-2	2017-06-0	对A市参保记录的几点疑问
19	19	11	2019-11-2	2017-06-08	17:31:20

	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
2	1	360114	A0182491	A市经济学	2017-06-0	书记您	9	0
3	2	289408	A0012413	在A市人才	2018-11-1	我叫朱琦梦	0	0
4	3	336608	A0005623	希望西地省	2019-09-0	让癌症病人	0	0
5	4	360103	A0012425	A5区劳动东	2019-09-2	A5区劳动东	1	0
6	5	323149	A1241141	请给K3县乡	2019-06-2	K3县的乡村	0	0
7	6	360107	A0283523	A5区劳动东	2019-07-2	局长：	3	0
8	7	343985	A108051	A市能否设	2019-10-3	A2区南托街	0	0
9	8	286572	A23525	请求A市地	2018-10-2	领导好！今	3	0
10	9	316619	A235259	请问A市什	2019-05-1	A市A2区之	0	0
11	10	360110	A110021	A市经济学	2019-11-2	关于西地省	0	0
12	11	323034	A012414	L市物业服	2019-06-1	L市发展和	0	0
13	12	319659	A023956	A市江山帝	2019-05-3	我是江山帝	0	0
14	13	313964	A108906	12123上申	2019-04-2	说是推出了	0	0
15	14	337458	A078325	能否分层单	2019-09-1	尊敬的蒋局	0	0
16	15	360104	A012417	A市魅力之	2019-08-1	A市魅力之	0	0
17	16	353426	A0098773	J4县供销合	2020-01-0	关于J4县供	2	0
18	17	321736	A9992521	A市能不能	2019-06-1	尊敬的领导	1	0
19	18	351074	A0012414	对A市参保	2019-12-1	尊敬的A市	0	0
20	19							

3.3 问题三

3.3.1 问题重述

针对赛题数据中所给出的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的质量给出一套评价方案。

3.3.2 解题思路

针对问题三的相关性、完整性和可解释性，本小组采取打分制对留言答复进行评价，从答复的字数、答复关键词筛选等角度入手，以答复字数作为答复完整性的评价标准，以答复关键词作为答复相关性和可解释性的评价标准，对所有留言答复按照既定的评价标准进行打分，最终根据留言答复的分数给出一套评价方案。

3.3.3 解题过程

1. 数据导入

首先将 xlsx 格式的数据文件改为 csv 格式

使用 pycharm 导入 pandas 库，通过

“data_1 = pd.read_csv('text4.csv',encoding='utf-8')

data_huifu = data_1['答复意见']”

语句将赛题数据导入

2. 答复完整性评价

本小组逐条统计各条留言的字数，以 250 字作为字数评价的标准，筛选出符合标准的留言共计 X 条。通过“

```
count=0
score1=[]
score=0
count2=0
for index, line in enumerate(data_huifu):
 count+=1
for i in range(0,count):
    count2 = 0
    for s in data_huifu[i]:
        if 0x4e00 <= ord(s) <= 0x9fa5:
            count2 += 1
        elif 'a'<=s<='z' or 'A'<=s<='Z':
            count2 += 1
    if count2 > 250:
       score=30
    else:
        score=0
    score1.append(score)
```

”语句我们将所有符合 250 字标准的答复加 30 分并保存所有答复的分数数据

3. 答复相关性和可解释性评价

我们根据政府留言答复的日常标准格式和用词，归纳出如下“ ’感谢’,’您好’,’你好’,’支持’,’关注’,’关心’,’监督’,’反映’”词语作为答复相关性和可解释性的评价标准。使用 pycharm 导入 jieba 库，之后我们通过 jieba 分词逐条对留言答复进行中文分词，并通过

```
“data9 = data_huifu.apply(lambda x: list(jieba.cut(x)))
for j in range(0,count):
    for i in a:
        if i in data9[j]:
            score1[j]=score1[j]+6”
```

语句判断各条答复的分词结果中，是否包含评价标准里的关键词。之后我们根据既定的分数，按每项关键词 6 分为标准对各条答复进行相关性和可解释性的打分，同时将打分结果进行保存。

4.**创建完整评价方案**

根据步骤 2、3 保存的答复分数，我们按照分数高低给出了一套留言答复的评价方案，使用 pycharm 导入 xlrd、xlutils 库，并将该方案通过

“xls =xlrd.open_workbook(r'text4.xls',formatting_info=True)
xlsc = copy(xls)
shtc = xlsc.get_sheet(0)
shtc.write(0,7,'评论得分')
for i in range(0,count):
    shtc.write(i+1,7,score1[i])
xlsc.save(r'text4-得分.xls')”

格式的形式保存为 text-4 得分表。

5.结果分析

本小组根据步骤 4 输出的评价方案，得到如下留言评价表格。由表格可以看出，通过既定的留言回复评价方案，我们对所有留言回复进行了评分，评分结果从 12—60 分不等，完整性、相关性和可解释性越高的留言回复得分越高。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	评论得分
2549	A00045581	区景蓉华苑物业管理有问题	2019/4/25 9:32:09	业公司却以交20万保证金，不	收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	54
2554	A00023583	潇楚南路洋湖段怎么还没修	2019/4/24 16:03:40	面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	60
2555	A00031618	决提高A市民营幼儿园老师	2019/4/24 15:40:04	同时更是加大了教师的工作	办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	54
2557	A000110735	公寓能享受人才新政购房补	2019/4/24 15:07:30	落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	12
2574	A0009233	于A市公交站点名称变更的	2019/4/23 17:03:19	“马坡岭小学”，原“马坡岭	留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	24
2759	A00077538	A3区含浦镇马路卫生很差	4/8/19 8:37	每把泥巴冲到右边，越是上下	于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	30
2849	A000100804	教师村小区盼望早日安装	2019/3/29 11:53:23	为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	30
3681	UU00812	区东澜湾社区居民的集体民	2018/12/31 22:21:59	好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	48
3683	UU008792	麓阳光住宅楼无故停工以及	2018/12/31 9:55:00	也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	48
3684	UU008687	和顺路洋湖壹号小区路段名	2018/12/31 9:45:59	立交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	24
3685	UU0082204	A2区大托街道大托新村违建	2018/12/30 22:30:30	规划局审批通过《温室养殖用	列支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	48
3692	UU008829	鄱阳村D区安置房人防工程能	2018/12/29 23:27:51	区安置房地下室近两万平方米	续，按长人防发[2014]7号文件要求，鄱阳村	2019/1/29 10:52:01	48
3700	UU00877	区段请求修建一座人行天桥	2018/12/29 11:55:34	峰，大量从小区开车出去的业	分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	18
3704	UU0081480	报A市芒果金融平台涉嫌诈	2018/12/28 17:18:45	省省相关政府部门的大力支持	映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	24
3713	UU0081227	建议增开A市261路公交车	2018/12/28 7:53:25	小时以上！天寒地冻，其他公	正常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	18
3720	UU008444	路与披塘路交叉路口通行安	2018/12/27 15:18:07	址：https://baidu.com/。但	的“披塘路路口两端各拆除20米中间花坛，	2019/3/6 10:26:14	48
3727	UU0081194	B区桐梓坡路益丰大药房以北	2018/12/27 1:55:21	便以各种理由拒绝退货，并获	根据您提供的信息进行投诉信息的登记分送	2019/1/3 14:02:47	24
3733	UU008706	义在A市梅溪湖开办一个图书	2018/12/26 16:51:40	称。建议在艺术中心先期借	营业。梅溪湖二期金菊路与雪松路东南角规	2019/1/14 14:32:40	18
3747	UU008201	是A3区中海国际社区一期旁边	2018/12/25 19:35:12	上很早就施工，严重影响居民	查，施工单位由于需要夜间连续作业，已办	2019/1/8 16:19:16	18
3755	UU0081681	保卡、医保卡、居民健康卡	2018/12/25 16:23:27	希望可以尽快合一。让社保卡	以上不同机构，需三方或三方以上不同机构	2019/1/4 15:48:23	18
3756	UU0081681	楚一卡通尽快支持手机nfc刷	2018/12/25 16:19:49	华为、苹果等手机都无法开通	具体上线时间请关注潇洒支付公司官网http	2019/1/4 15:49:46	24
3760	UU0081500	区对泉水村塘下组土地征收	2018/12/25 14:40:13	有权国家行政机关进行了申请	桥北组签订了土地补偿协议，并按协议达成	2019/1/8 16:18:00	48
3762	UU0081057	警大队纠正电子交通警察的	2018/12/25 13:56:31	直行车辆和行人通行，此路口	实施条例》第三十八条第一款第三项“红灯	2019/1/16 15:22:16	18
3777	UU008162	3号线北段在楚江北路上设站	2018/12/23 21:47:34	，事故频发。如果8路线设立	9年1月15日您好，非常感谢您对于A市轨道	2019/1/29 10:50:31	24
3788	UU0081604	市商业住房贷款转公积金贷	2018/12/21 11:01:00	金，是否能在A市办理商业住	理中心不支持非本中心的缴存人以及异地商	2019/1/3 14:00:47	18
3791	UU008694	段（劳动东路-机场高架）段	2018/12/20 17:28:09	在到A市国际会展中心非常不	2公里，已完成约800米路基，其余路段因洪	2019/1/4 15:47:36	18
3797	UU008765	B区西湖街道茶场村公路规	2018/12/20 11:16:07	政府修A3区山景区西大门，拆	，因政府投资计划调整，该项目已暂停。至	2019/1/3 13:59:33	24
3838	UU0082119	B区靳江洋湖集体资产的有	2018/12/15 15:17:53	是一个多亿好远，这笔大资金	二是村级举办的西地省洋兴置业公司。土地	2019/1/4 15:44:31	54

四、结论

本小组通过自然语言处理及文本挖掘的方法，将政务系统中数量巨大的留言进行了分类、归纳和评分等操作，顺利完成了赛题要求的群众留言分类、热点问题挖掘和答复意见评价的三项任务，我们将每个问题的结论和输出结果以表格文件的形式输出。

通过本次解题本小组认识到，数据挖掘在当今社会具有极为广泛的应用前景。例如在当下的政务系统中，拥有大量的数据需要处理和分析，通过计算机对这类数据进行自动化的处理分析能够极大的降低人工成本，同时可以规避一些人工操作存在的不足和缺陷，未来数据挖掘和大数据的应用一定会给人类的生活带来巨大的改变。