

基于自然语言处理的文本挖掘模型

摘要

近年来随着信息的发展，许多网络问政平台逐步成为了政府了解民生的渠道，各类的民意的文本数量不断增多，给留言分类和热点整理的相关部门带来了极大的挑战。但是，随着大数据，云计算的发展，建立基于自然语言处理的智慧政务系统提供了新的治理创新发展趋势，对相关部门提高行政水平有极大的推动作用。

本队伍做 C 题的过程中，第一步运用 jieba 分词，正则表达式对数据进行预处理转换成 fasttext 格式，再通过 fasttext 对文本数据进行处理得出有监督模型，再通过测试集对模型进行检验。

第二题通过使用 TankRank 进行关键词权重的提取，利用 excel 对点赞数与访问量的数量和 K-means 聚类获得相同类型最多的二十类，再进行对比分析选择热度前五的问题。

第三题通过建立层次分析法对答复建议的质量进行评价，对答复的相关性、完整性和可解释性等因素建立了关键词、文本相似性、语法语义这三个主要指标。利用 matlab 对此模型进行训练得出综合排序的一致性是可以接受的，且在权值上相关性大于完整性大于可解释性。

关键词： FastText ,TankRank ,K-means ,matlab,文本处理

目录

一、群众留言分类.....	3
1.1 模型的建立.....	3
二、热点问题挖掘.....	6
2.1 模型的建立	6
2.1.1 K-means 聚类.....	6
2.2.1 分词处理	6
三、答复意见的评价.....	10
3.1 建立层次结构模型	10
3.2 举例判断答复的质量.....	15
四、参考文献	19

一、群众留言分类

1.1 模型的建立

1.1.1 FastText 模型

FastText 模型将整篇文档的词或 N-gram 向量叠加平均得到文档向量，然后使用文档向量分类。适合海量数据和高速训练，能将训练时间由几小时缩短到几分钟。

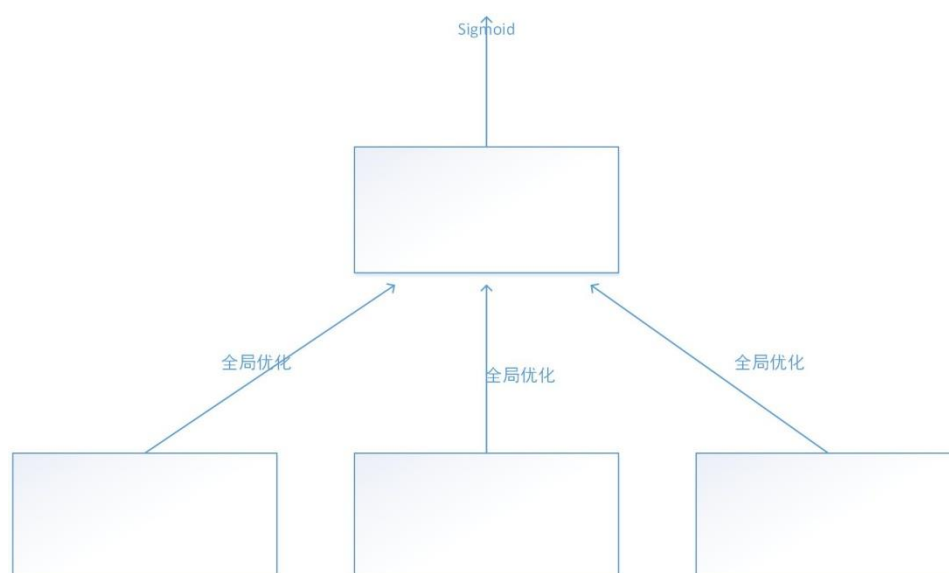


图 1FastText 思想

模型架构：词嵌入后，隐藏层只是一个简单的全局平均池化层，然后连接经过池化的问题和回答向量，最后的全连接层作为分类器使用。注意这里的输入可以是单词，也可以是 N-gram 组合。

在本模型中，我们主要采用将自然语言处理的问题要转化为机器学习的方式进行学习。分词采用 python 自然语言处理工具 jieba，开发者可以自定义词典，以便包含 jieba 词库里没有的词，虽然 jieba 有新词识别能力，但是自行添加新词可以保证更高准确率。

```

jieba.load_userdict(file_name) # 载入自定义词典
jieba.add_word(word, freq=None, tag=None) # 在程序中动态修改词典
jieba.del_word(word)
jieba.suggest_freq(segment, tune=True) # 调节单个词语的词频, 使其能/不能被分词开

```

图 2. jieba 新词

在自行添加字典以后对文本进行分词, 先将 excel 中的内容利用 pandas 输出到文档中进行处理。

```

import pandas as pd
df = pd.read_excel(r'C:/Users/pro/Desktop/C题全部数据/测试集.xlsx', sheet_name='Sheet1', header=None, usecols = [4,5]) # 使用pandas模块读取数据
print('开始写入txt文件...')
df.to_csv(r'C:/Users/pro/Desktop/C题全部数据/news_fasttext_test', header=None, sep=' ', index=False) # 写入, 逗号分隔
print('文件写入成功!')

```

图 3. 提取内容

再对文本内容进行去除停用词和正则表达式进行优化。

```

import re
from types import MethodType, FunctionType

import jieba

def clean_txt(raw):
    fil = re.compile(r"^[^0-9a-zA-Z\u4e00-\u9fa5]+")
    return fil.sub(' ', raw)

def seg(sentence, sw, apply=None):
    if isinstance(apply, FunctionType) or isinstance(apply, MethodType):
        sentence = apply(sentence)
    return ' '.join([i for i in jieba.cut(sentence) if i.strip() and i not in sw])

def stop_words():
    with open(r'C:/Users/pro/Desktop/shilishuju/stop_words.txt', 'r', encoding='utf-8') as swf:
        return [line.strip() for line in swf]

# 对某个sentence进行处理:
content = (r'C:/Users/pro/Desktop/shilishuju/file6.txt')
res = seg(content.lower().replace('\n', ' '), stop_words(), apply= clean_txt)

```

图 4. 去除停用词与正则表达式

利用 jieba 对文档进行分词

```

import jieba

fR = open(r'C:/Users/pro/Desktop/C题全部数据/news_fasttext_train', 'r', encoding='UTF-8')

sent = fR.read()
sent_list = jieba.cut(sent)

fW = open(r'C:/Users/pro/Desktop/C题全部数据/news_fasttext_train', 'w', encoding='UTF-8')
fW.write(' '.join(sent_list))

fR.close()
fW.close()
jieba.load_userdict(file_name) # 载入自定义词典
jieba.add_word(word, freq=None, tag=None) # 在程序中动态修改词典
jieba.del_word(word)
jieba.suggest_freq(segment, tune=True) # 调节单个词语的词频, 使其能/不能被分词开

```

图 5. jieba 分词

二、热点问题挖掘

2.1 模型的建立

2.1.1 K-means 聚类

k 均值聚类算法 (k-means clustering algorithm) 是一种迭代求解的聚类分析算法, 其步骤是, 预将数据分为 K 组, 则随机选取 K 个对象作为初始的聚类中心, 然后计算每个对象与各个种子聚类中心之间的距离, 把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本, 聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有 (或最小数目) 对象被重新分配给不同的聚类, 没有 (或最小数目) 聚类中心再发生变化, 误差平方和局部最小。

2.2.1 分词处理

通过 jieba 对附件 3 中的留言内容与主题进行分词

座落在 A 市 A3 区联丰路米兰春天 G2 栋 320, 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百万, 因为地处居民楼内部, 而且有蛮长的时间了, 请税务局和工商局查一下, 看看这个一米阳光有没有正常纳税! 如果没有, 应该会怎么操作!

图 9. 留言内容分词

再通过 TextRank 进行关键词的提取, TextRank 算法是基于 PageRank, 用于为文本生成关键字和摘要。

TextRank 一般模型可以表示为一个有向有权图 $G=(V, E)$ ，由点集合 V 和边集合 E 组成，图中任两点 V_i, V_j 之间边的权重为 w_{ji} 。点 V_i 的得分定义如下：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in \text{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \text{Out}(V_j)} w_{jk}} WS(V_j)$$

其中， d 为阻尼系数，取值范围为 0 到 1，代表从图中某一特定点指向其他任意点的概率，一般取值为 0.85。使用 TextRank 算法计算图中各点的得分时，需要给图中的点指定任意的初值，并递归计算直到收敛，即图中任意一点的误差率小于给定的极限值时就可以达到收敛，一般该极限值取 0.0001。

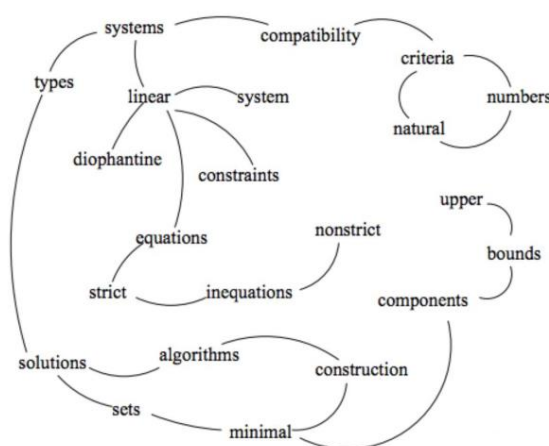


图 10. tankrank 原理流程

首先从 jieba 分词的分析工具箱里导入所有关键词提取功能，使用 TF-idf 方式提取关键词和权重。

```
from jieba.analyse import *
with open(r'C:/Users/pro/Desktop/C题全部数据/file3.txt', 'rb') as f:
    data = f.read()
for keyword, weight in textrank(data, withWeight=True):
    print('%s %s' % (keyword, weight))
```

图 11. 权重计算

进而得出了留言内容的关键词权重

```
业主 0.08165717820215436
小区 0.07330801165325948
开发商 0.031239912251116175
2019 0.0297129872290514
领导 0.028313662957078618
西地省 0.026775005821081995
2018 0.025200005478665407
部门 0.023629250483410285
物业 0.02359706648505745
居民 0.023480915382605643
A3 0.02307981271002769
相关 0.02205306318374855
A7 0.02086875453701979
社区 0.02080926523688746
政府 0.019645582475108603
投诉 0.018541651878225422
希望 0.017843618487899678
解决 0.016600066556582846
街道 0.016375829321252863
A1 0.015840868828536067
施工 0.0152813328539506
学校 0.015243694521503027
车位 0.01500403879824017
A4 0.014720195507970418
交房 0.014385937052787328
房屋 0.014348154074986534
A5 0.014296156954242875
A2 0.0141144261455025
公司 0.013891492617635047
规划 0.013745497476485
```

图 12. tankrank 权重

同时使用 jieba 进行关键词的提取得到以下结果

```
In [33]: runfile('C:/Users/10634/Desktop/
untitled2.py', wdir='C:/Users/10634/Desktop')
小区 1.0
问题 0.4573691288696904
街道 0.3950623423094337
西地省 0.38226258313369005
业主 0.3028529626949264
社区 0.278609216211631
扰民 0.2179919968586369
有限公司 0.21488703178780014
领导 0.21332975444886662
居民 0.20646079589194113
国际 0.18996253334465
建设 0.1835764683222522
违规 0.15914299264907908
噪音 0.15762564447300959
物业 0.1546478966279439
```

图 13.jieba 权重

再利用 excel 的相关技术获得了点赞数的排名并通过降序进行排序

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	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	回复数	回复内容				
2	208636	A00077171	汇金路五矿万境K9县存在一	2019/8/19 11:34:04	狗咬人，请问有人对	0	2097	2097					
3	223297	A00087522	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	5	1762	1767					
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	821					
5	217032	A00056543	A市58车贷特大集资诈骗案保	2019/2/25 9:58:37	股东、苏纳弟弟苏	0	790	790					
6	194343	A000106161	A市58车贷案警官应跟进关注	2019/3/1 22:12:30	等候并没有跟进市领	0	733	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	669					
8	193091	A00097965	省绿物业发新城强行断业主	2019/6/19 23:28:27	提供地摊上买的收据	0	242	242					
9	284571	A00074795	省尽快外迁京港澳高速城区	2019/1/10 15:01:26	、长河高速出口，及	0	80	80					
10	200667	A00079480	是把和包支付作为任务而不让	2019/1/16 17:01:25	层工作者也不理解，	0	78	78					
11	262052	A00072424	月亮路沿线架设110kv高压线	2019/3/26 14:33:47	上电力线路，应采	0	78	78					
12	226723	A00040222	一大道全线快速化改造何时	2019/9/15 15:31:19	改造，打通机场北通	0	66	66					
13	203187	A00024716	咨询A9市高铁站选址的问题	2019/8/1 13:48:57	城区东进的步伐，为	53	10	63					
14	281898	A00096623	云计算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	60					
15	272089	A00061602	A6区月亮路110kv高压线的	2019/4/9 17:10:01	地省体操学校、西	2	55	57					
16	288398	A00053962	出止星沙滨湖路以南，特立	2019/2/11 14:09:40	间有大量的闲置土	5	40	45					
17	239595	A00057814	回东六路恒大九五工厂地块，	2019/11/8 15:48:07	地区，这里潜力巨大	0	44	44					
18	279062	A00027836	加大A7县东六线柳梨段拆迁	2019/1/17 19:25:45	有土地（正钢机械厂	1	42	43					
19	267630	A000106648	铁3号线松雅湖站点附近地	2019/5/22 23:37:38	东四路和东四路西	0	42	42					
20	209742	A00012969	K郝家坪小学什么时候能改扩	2019/3/24 21:07:12	小学，淡湾路小学	0	41	41					
21	239670	A00080329	六线以西泉塘昌和商业中心	2019/1/11 15:46:04	置业有限公司厂房	0	41	41					
22	257376	A909155	关于加快修建A市南横线的建	2019/5/10 18:01:52	通落后的现状，拉	0	39	39					
23	244178	A00057874	4号线北延线“同心路”设	2019/1/30 23:59:12	站路设在雷峰大道	0	38	38					
24	205217	A00040562	实发放原A市七中01年后退休	2019/10/29 12:42:17	K下放教师待遇有	0	33	33					
25	193286	A000103197	KA7县松雅西地省站西北方	2019/4/17 11:13:12	K北侧穿越东四路	0	32	32					
26	226996	A00022217	A市汽车站何时能建好？	2019/3/20 9:20:46	群众的出行和生活	0	32	32					
27	243808	A00053304	义将地铁7号线南延至A市生态	2019/3/6 14:20:16	车赶到“尚双塘”	0	31	31					
28	283631	A00042509	步联络线噪音污染严重超过国	2019/11/7 10:13:18	不开窗透气，可是只	2	29	31					
29	253369	A00074795	澳高速《长建高速》什么时	2019/11/18 15:35:11	0万居民苦不堪言，	0	29	29					
30	256358	A00080329	六线以西泉塘昌和商业中心	2019/1/2 20:27:07	租上原厂和闲置的	0	29	29					
31	280425	A00032687	K泉塘小塘路路灯太暗，建议	2019/11/8 16:38:44	塘路路灯太暗，只	0	29	29					

图 14. 点赞数排名

再通过 k-means 聚类对数据进行 20 类聚类来获得各类问题的热度。

```

# k-means 聚类算法
def kMeans(dataSet, k, distMeans = distEclud, createCent = randCent):
    m = shape(dataSet)[0]
    clusterAssment = mat(zeros((m,2))) # 用于存放该样本属于哪类及质心距离
    # clusterAssment第一列存放该数据所属的中心点，第二列是该数据到中心点的距离
    centroids = createCent(dataSet, k)
    clusterChanged = True # 用来判断聚类是否已经收敛
    while clusterChanged:
        clusterChanged = False;
        for i in range(m): # 把每一个数据点划分到离它最近的中心点
            minDist = inf; minIndex = -1;
            for j in range(k):
                distJI = distMeans(centroids[j,:], dataSet[i,:])
                if distJI < minDist:
                    minDist = distJI; minIndex = j # 如果第i个数据点到第j个中心点更近，则将i归属为j
            if clusterAssment[i,0] != minIndex: clusterChanged = True; # 如果分配发生变化，则需要继续迭代
            clusterAssment[i,:] = minIndex,minDist**2 # 并将第i个数据点的分配情况存入字典
        print (centroids)
        for cent in range(k): # 重新计算中心点
            ptsInClust = dataSet[nonzero(clusterAssment[:,0] == cent)[0]] # 去第一列等于cent的所有列
            centroids[cent,:] = mean(ptsInClust, axis = 0) # 算出这些数据的中心点
    return centroids, clusterAssment

# -----测试-----
# 用测试数据及测试kmeans算法
datMat = mat(loadDataSet(r'C:/Users/pro/Desktop/C题全部数据/新建文本文档.txt'))
myCentroids, clustAssing = kMeans(datMat,10)
print (myCentroids)
print (clustAssing)

```

图 15. k-means 聚类

Start Kmeans:	出现次数
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,	368
n_clusters=20, n_init=10, n_jobs=None, precompute_distances='auto',	420
random_state=None, tol=0.0001, verbose=0)	83
[[-2.16840434e-19 -2.03287907e-20 4.33680869e-19 ... 2.03287907e-20	236
-3.38813179e-21 -3.38813179e-21]	88
[8.23993651e-18 2.84603070e-19 4.98816253e-04 ... -1.28749008e-19	172
2.03287907e-20 2.11758237e-20]	98
[1.00032504e-03 4.70469451e-04 4.41927268e-03 ... -1.35525272e-19	81
4.91279109e-20 4.06575815e-20]	172
...	87
[2.59493494e-04 5.08219768e-19 8.79176816e-04 ... -1.42301535e-19	158
-2.87991202e-19 -8.72443936e-20]	238
[1.75599108e-03 5.42101086e-20 -6.50521303e-19 ... -6.09863722e-20	457
1.69406589e-20 6.77626358e-21]	147
[0.00000000e+00 6.77626358e-21 -3.90312782e-18 ... -1.15196481e-19	51
-4.23516474e-20 -2.03287907e-20]]	62
[17 17 2 ... 19 19 19]	1138
	71
	199
	0

图 16. 各类别出现次数

综合聚类、点赞数、关键词的权重建立评价指标，通过选取共同出现的五个问题得出热点问题。热点问题挖掘

三、答复意见的评价

3.1 建立层次结构模型

3.1.1 确立评价指标

建立层次分析法(AHP)模型的首要问题，是确定指标体系。指标体系应能够反映留言答复意见的有效性和质量。由于留言答复的复杂性，衡量答复意见的质量的指标体系是由若干相互联系、相互补充，具有层次性和结构性的指标构成的有机系列。

答复意见会根据留言详情和留言主题发生变化，对系统进行综合分析，可以看出决定和影响答复意见质量的因素主要包括答复的相关性(B_1)、完整性(B_2)、可解释性(B_3)这三个方面。

答复的相关性因素是指答复意见与留言详情的关联度，主要指标

包括文本相似性(c_1)和文本提取的关键词(c_2)。

答复的完整性因素是指答复意见是否完整，是否具有意义，影响因素的主要指标有文本相似性、关键词和语法语义(c_3)。

答复的可解释性因素是指答复意见是否可以解决留言详情提出的问题，主要指标有文本相似性、关键词和语法语义。

3.1.2 建立层次结构模型

根据层次分析法的基本步骤，建立如图 1 的答复意见的质量评价层次结构模型。

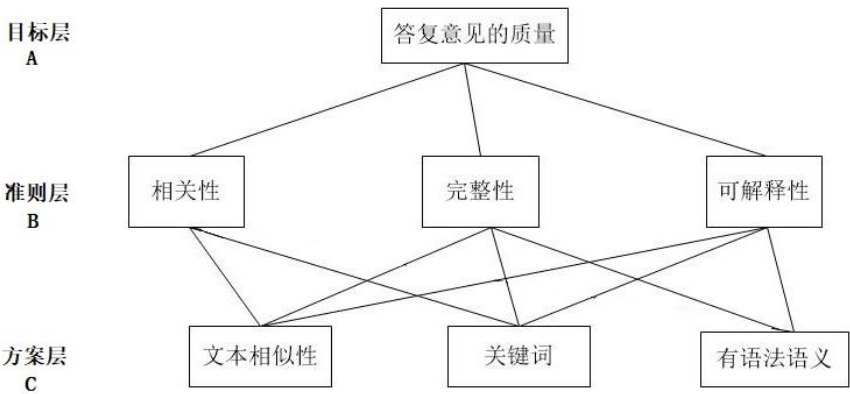


图 17.答复意见的质量评价层次结构模型

3.1.3 构造判断矩阵并一致性检验

先构建准则层的判断矩阵，再构建方案层的判断矩阵。设某一层有 n 个因素， $X=\{x_1,x_2,\dots,x_n\}$ 。要比较该层的每一个因素对上一层的某个因素的影响程度，确定在该层中相对于某一准则所占的比重。可以采取对因子进行两两比较建立成对比较矩阵的办法。最后构建层次总排序。

准则层的判断矩阵如表 1 所示。

表 1 准则层的判断矩阵

层 A \ 层 B	A_1	A_2	...	A_m	B 层总排序权值
B_1	b_{11}	b_{12}	...	b_{1m}	$\sum_{j=1}^m b_{1j}a_j$
B_2	b_{21}	b_{22}	...	b_{2m}	$\sum_{j=1}^m b_{2j}a_j$
\vdots	\vdots
B_n	b_{n1}	b_{n2}	...	b_{nm}	$\sum_{j=1}^m b_{nj}a_j$

利用 1~9 标度法进行成对比较，同时参考意见，确定各因素之间的相对重要性并赋以相应的分值，构造出各层次中的所有判断矩阵，并计算权向量和一致性检验。

表 2 1~9 标度的意义

标度	含义
1	表示两个因素相比，具有相同重要性
3	表示两个因素相比，前者比后者稍重要
5	表示两个因素相比，前者比后者明显重要
7	表示两个因素相比，前者比后者强烈重要
9	表示两个因素相比，前者比后者极端重要
2, 4, 6, 8	表示上述相邻判断的中间值

1, 1/2, ...1/9	与上述说明相反
----------------	---------

A-B 之间构成的判断矩阵：

表 3 A-B 判断矩阵

A	B ₁	B ₂	B ₃
B ₁	1	2	3
B ₂	1/2	1	2/3
B ₃	1/3	3/2	1

利用 matlab 软件求出其特征值：

$$W = (w_1, w_2, w_3) = (0.5499, 0.2098, 0.2402)$$

$$CR = \frac{CI}{RI} = 0.0634 < 0.1,$$

因此，该判断矩阵的一致性是可以接受的。

B-C 之间构成的各个判断矩阵形式如下：

表 4B-C 判断矩阵

B ₁	C ₁	C ₂	C ₃
C ₁	1	2	3/2
C ₂	1/2	1	2
C ₃	2/3	1/2	1

该矩阵的特征值为：

$$W = (w_1, w_2, w_3) = (0.4600, 0.3189, 0.2211)$$

$$CR = \frac{CI}{RI} = 0.0930 < 0.1,$$

因此，该判断矩阵的一致性是可以接受的。

表 5B-C 判断矩阵

B ₂	C ₁	C ₂	C ₃
----------------	----------------	----------------	----------------

C_1	1	2/3	1/3
C_2	3/2	1	1/2
C_3	3	2	1

该矩阵的特征值为：

$$W = (w_1, w_2, w_3) = (0.1818, 0.2727, 0.5455)$$

$$CR = \frac{CI}{RI} = 0 < 0.1$$

因此，该判断矩阵的一致性是可以接受的。

表 5B-C 判断矩阵

B_3	C_1	C_2	C_3
C_1	1	1/2	1/3
C_2	2	1	3/2
C_3	3	2/3	1

该矩阵的特征值为：

$$W = (w_1, w_2, w_3) = (0.1692, 0.4434, 0.3874)$$

$$CR = \frac{CI}{RI} = 0.0634 < 0.1$$

因此，该判断矩阵的一致性是可以接受的。

可以看出，所有单排序的 $CR < 0.1$ ，认为每个判断矩阵的一致性都是可以接受的。

3.1.4 层次总排序及一致性检验

上面得到的是一组元素对其上一层中某元素的权重向量。要最终得到各元素特别是最低层中各方案对于目标的排序权重，需要进行总排序。总排序是指同一层次所有因素对于目标层（最上层）的相对重

要性的排序权重（以上第一个矩阵的排序也为一个总排序）。总排序权重要自上而下地将单准则下的权重进行合成。对总排序结果进行一致性检验。计算综合检验指标：

$$CR = \frac{\sum CI}{\sum RI} = 0.0664 < 0.1$$

可以认为，综合排序的一致性是可以接受的。因此，用层次分析法评价答复意见的质量的因素，确定各种因素之间的相对重要性程度是可行的。

3.2 举例判断答复的质量

利用 TextRank 算法对附件 4 的留言详情和留言答复进行关键词和摘要提取。

```
import codecs
from textrank4zh import TextRank4Keyword, TextRank4Sentence

# 读取文件
text = codecs.open('答复.txt', 'r', 'utf-8').read()

# 关键词和关键短语
tr4w = TextRank4Keyword()
tr4w.analyze(text)

print('关键词：')
for item in tr4w.get_keywords(num=5, word_min_len=2): # 提取5个关键词，关键词最少为2个字
    print(item.word, item.weight)
print()

print('关键短语：')
for phrase in tr4w.get_keyphrases(keywords_num=20, min_occur_num=2): # 从20个关键词中选出出现次数至
    print(phrase)
print()

# 摘要
tr4s = TextRank4Sentence()
tr4s.analyze(text)
print('摘要：')
for item in tr4s.get_key_sentences(num=3):
    print(item.index, item.weight, item.sentence) # index是语句在文本中位置，weight是权重
```

图 18.关键词和摘要提取代码

以留言编号 2549，主题“A2 区景蓉华苑物业管理有问题”为例，

提取主题关键词为：

问题 1.0
物业管理 0.9961264494011037

图 19.主题关键词提取

留言详情的关键词与摘要提取如下：

关键词：
小区 0.04672924422571345
物业公司 0.03705469608000855
业主 0.0354470732249849
投票 0.0249795584959688
业委会 0.021668258148228967

关键短语：
A2区

摘要：
0 1.0 2019年4月以来，位于A市A2区桂花坪街道的A2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言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了（原水电在小区买，水4.23—吨，电0.64—度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交20万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪

图 20.留言详情关键词和摘要提取

答复意见的关键词与摘要提取如下：

关键词：
业委会 0.03545612130777799
业主大会 0.02729053437675484
问题 0.02511371978154965
反映 0.023670892108504288
停车 0.023131943772717902

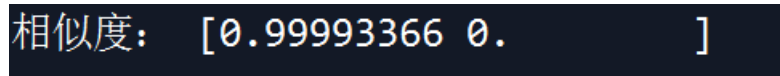
关键短语：

摘要：
1 0.24441859545999328 现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于2019年4月10日至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真梳理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格
2 0.19418168577523853 针对来信所反映的“物业公司去留问题”，5月5日下午，辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议
0 0.17320501287167364 现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉华苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2区景蓉华苑物业管理有问题”的情况已收悉

图 21.答复意见关键词和摘要提取

留言详情和答复意见的关键词都有业委会，且关键词中都与主题关键词有相关性，可知回复相关性、完整性高。再对留言详情和答复意见进行相似度分析，先对文档进行分词，采用默认的精准确模式分词，整理分词后的文档，整理成如图所示：

向量进一步处理，得到新语料库，通过 tf-idf 模型处理新语料库，得到 tf-idf 值并通过 token2id 得到特征数(字典里面的键的个数)。计算稀疏矩阵相似度，建立一个索引，根据索引得到最终的相似度。



```
相似度: [0.99993366 0.]
```

图 25.相似度

可进一步验证答复的相关性高，可解释性高。综合得知该答复意见质量高。

四、参考文献

- [1]郭金玉,张忠彬,孙庆云.层次分析法的研究与应用[J].中国安全科学学报,2008(5):148-153.
- [2]荆全忠,姜秀慧,杨鉴淞,等.基于层次分析法(AHP)的煤矿安全生产能力指标体系研究[J].中国安全科学学报,2006,16(9):74-79. DOI:10.3969/j.issn.1003-3033.2006.09.013.
- [3]邹俊杰.受限域问答系统文本检索研究[D].昆明理工大学,2011.
- [4]《机器学习》,周志华,清华大学出版社.
- [5]《A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.