“智慧政务”的文本挖掘应用

摘要

近年来，随着微信，微博，市长信箱等网络问政平台不断成为政府了解民意，汇聚民智，凝聚民气的重要渠道，各类数据的文本数量急剧攀升，传统的人工分类处理数据方法已经出现效率低下、工作繁冗等问题。因此运用文本分析以及数据挖掘等方法对处理群众留言提高政府行事效率有着重大的意义。

对于问题一，操作者可以首先通过 python 等工具将每条留言出现频率最多的十五个词选出来成为一个集合令其代替该条留言，然后将每个相似度最高的词语集合放到个更大集合，最后通过简易人工可轻易找到每条留言所对应的一级标签。

对于问题二，先通过地区或人群留言分成若干个数据组，再通过计算可以得出每个数据组的热度评价指标，再依次排列并最后将明细排列出来即可。

对于问题三，先列出所能评价答复的各个角度，然后再赋予各个角度相应的权重，再从各个角度进行一系列的计算评测得出相应的分数，最后通过加权平均即可得出该条回复的分数，并在最后用各种公式验证分数的可靠性。

关键词:TF-IDF, F-score，相关系数，显著性水平

The text mining application of "intelligent government affairs"

ABSTRACT

In recent years, as online platforms such as WeChat, Weibo and mayor's mailbox have become an important channel for the government to understand public opinions, pool public wisdom and gather public opinion, the number of texts of various types of data has risen sharply, the traditional method of data processing by manual classification has already appeared the problems of low efficiency and heavy work. Therefore, using text analysis and data mining and other methods to deal with the message of the masses to improve the efficiency of government action is of great significance.

For Question One, the operator can first select the 15 words that appear most frequently in each message using tools such as python to form a collection to replace the message, the most similar set of words is then placed into a larger set, and the first-level label for each message can be easily found by hand.

For the second question, first through the region or the crowd message into a number of data groups, and then through the calculation can be derived from each data group of heat evaluation indicators, and then arranged in turn and finally detailed arrangement can be.

For Question Three, first list all the angles that can be evaluated and then give each angle the corresponding weight, and then carry out a series of calculation and evaluation from each angle to get the corresponding score, finally, the score of this article can be obtained by weighted average, and the reliability of the score can be verified by various formulas

Keywords: TF-IDF; F-score; significance level; correlation coefficient

目录

1,挖掘目标

本次建模目标是利用已得到的群众问政留言记录，通过若干算法和若干数学工具来实现如下若干问题。

（1）、通过若干算法实现将若干个留言信息分别归类到若干的一级标签，并在最后通过 F-score 模型对分类方法进行评价。

（2）、通过分类将某一人群或者某一地区的人提出来的问题归为一组，然后按照热度高低的顺序依次排列出前五，并在最后将热度前五组的数据的明细细分出来。

（3）、针对附件 4 相关部门的答复意见，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.问题假设

（1）、假设所有数据真实可靠。

（2）、假设一个地区或者一个人群提出的问题只针对一个方面。

3.基本定义和说明

（1）、权重：指某一因素或指标相对于某一事物的重要程度，其不同于一般的比重，体现的不仅仅是某一因素或指标所占的百分比，强调的是因素或指标的相对重要程度，倾向于贡献度或重要性。

（2）、TF-IDF 算法：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 是词频(Term Frequency)，IDF 是逆文本频率指数(Inverse Document Frequency)。

（3）、精确率：精确率是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。

（4）、召回率：召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。

（5）、F-score：F-score 是一种统计量，F-score 又称为 F-Measure，F-Measure 是 Precision 和 Recall 加权调和平均，是 IR（信息检索）领域的常用

的一个评价标准，常用于评价分类模型的好坏。

（5)、相关系数：相关系数是研究变量之间线性相关程度的量。

（6)、显著性水平：显著性水平是估计总体参数落在某一区间内，可能犯错误的概率，用α表示。

4.符号说明

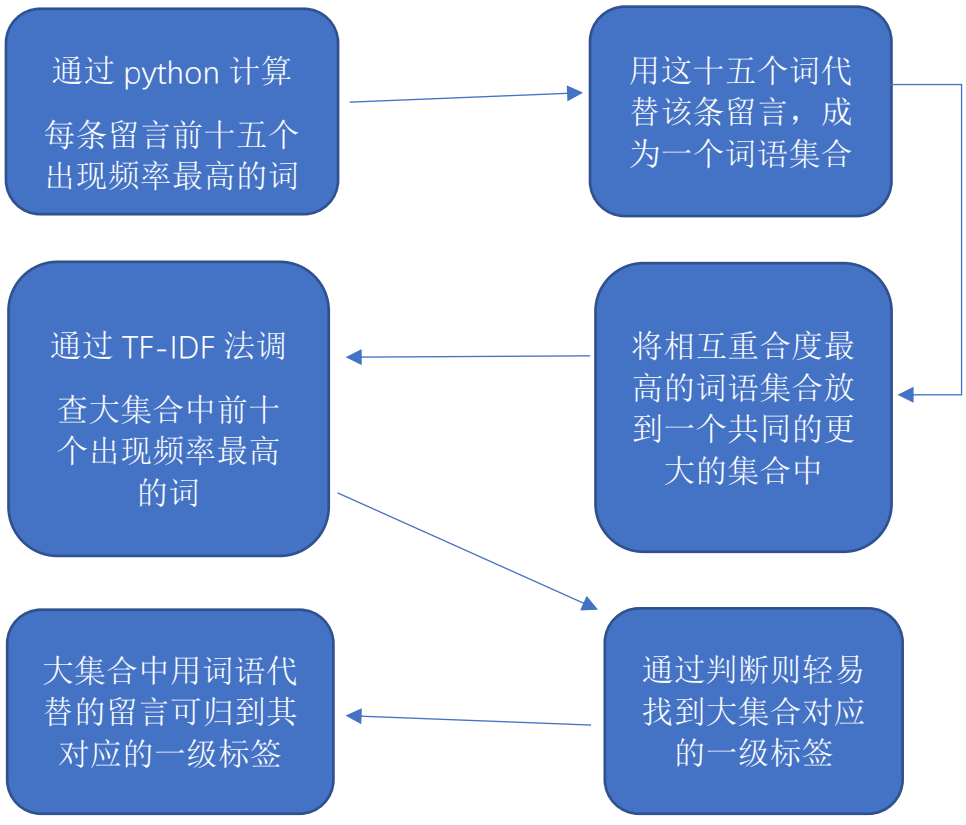
符号	符号意义
N	词语频数
n	文本总字数
P	精确率
R	召回率
β	F-score 公式中权重
θ	分类后的数据组
a	某组数据组中含有的留言数量
b	某条留言中的
c	某条留言中的点赞数
d	某条留言中点赞数与反对数之差
W	某组数据中的每条留言赞成数与反对数之差的和(热度评价指标)
e	相关性的权重
o	完整性的权重
g	可解释性的权重
v	滞后度的权重
A	机器对答复意见相关性的打分
B	机器对答复意见的完整性的打分
C	机器对答复意见的可解释性打的分
D	答复意见的加权平均分
F	第三题中从总体中抽取的样本中的答

	复意见数量
h	第三题中可接受的期望误差程度
l	第三题中可接受的方差误差程度
m	第三题中可接受的相关系数最小值
Z	第三题中检验统计量
t	第三题检验统计量
f	第三题自由度
α	第三题显著性水平
H0	第三题中原假设
H1	第三题中备选假设
r	点赞数的权重
s	反对数的权重
u	留言条数的权重

5.分析方法过程

5.1 问题一的分析方法和过程

5.1.1.流程图



5.1.2 TF-IDF 方法

如今有若干条留言，操作者可以将留言一条一条对应归到其应有的一级标签；也可以先将相同类型的留言聚在一起，最后一同归到其应有的标签。显然是后者效率更为高。要想将相同类型的归到一起，那他们一定拥有相同的特征，在这里我们选用词频作为它们的特征，相同类型的留言一般会拥有相同的频率较高的词，比如根据经验可推断出要归位到公安部门的留言出现频率最多的词语无非是“打架”、“抢劫”等跟犯罪含义相关的词汇。

若要把留言中的关键词较为准确地从留言中将出现频率较多的词筛选出来，操作者要用到 TF-IDF 算法。TF-IDF 算法的具体原理如下：

第一步计算词频：

词语频数（N_i）=某个词在文本中出现的次数，可表示为：

$$\sum_{i=1}^{n} N_i$$

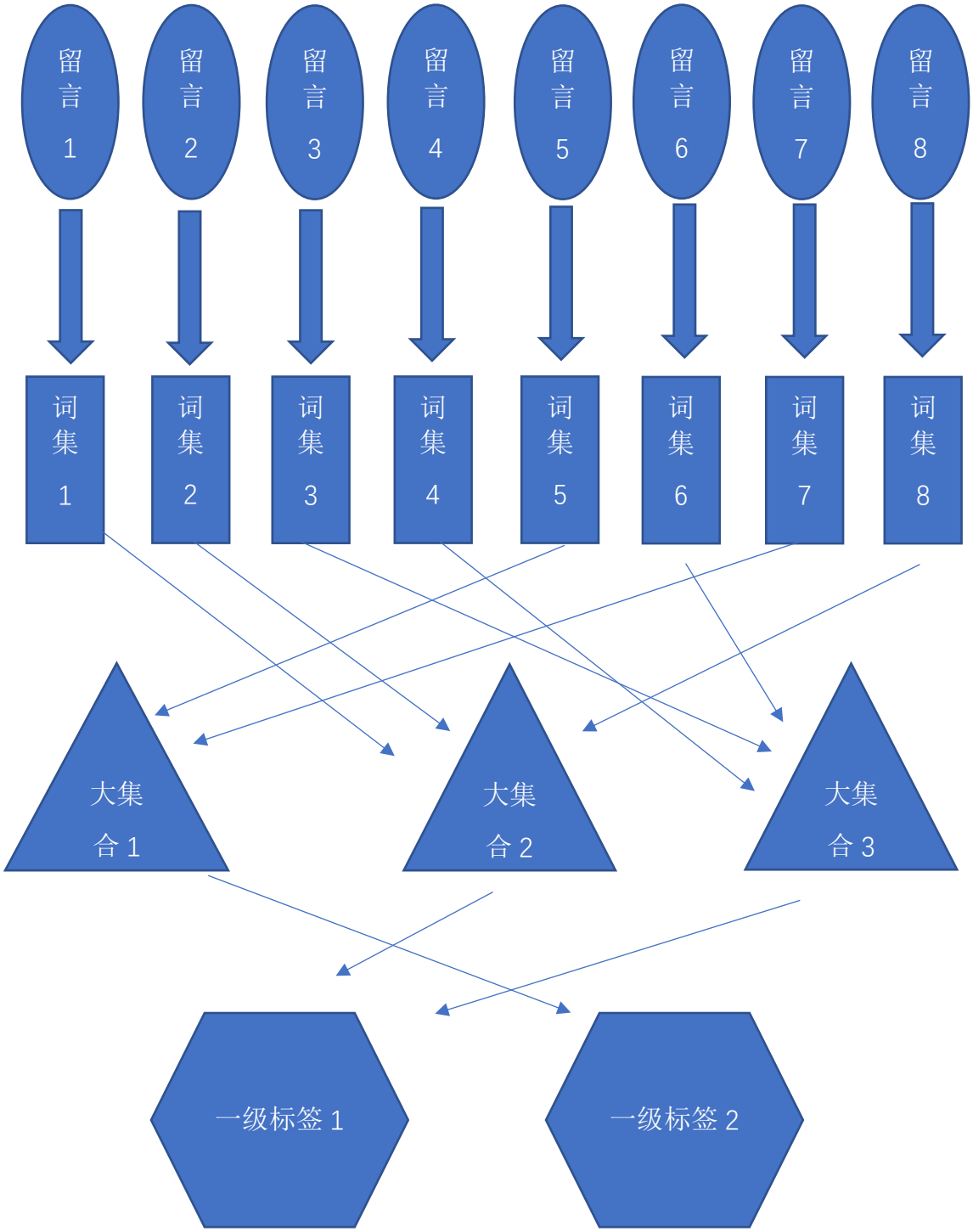
考虑到每个留言的长度不一致，故操作者可以将词频标准化，设第 p 条文本总词数为 nq：

词频（TF）=（某个词在文本中的出现次数）/（文本的总词数），可表示为：

$$TF = \frac{\sum_{i=1}^{n} N_i}{nq}$$

5.1.3 操作过程

计算出每个留言每个关键词的词频以后，操作者可以从每个留言中选出前十五个词频最大的词作为一个词语集合代替该条留言。总共有十四个一级标签，有若干条用词频代替的留言的词语集合，此时我们可以先通过归集重合度最高的词语集合从而得到十四个集合内留言词频重合率较高的集合。再通过 TF-IDF 法计算每个集合中重合率最高的前十个词。最后通过人工判断则可将这些由词语集合代替的留言的留言集合全部归集到其对应的一级标签。具体流程如下图所示：



5.1.4 对上述方法进行评估

要想得知操作者上述方法是否得当我们要运用 F-score 公式如下：

$$F-score=\frac{1}{n}\sum_{i=1}^{n}*\ （1+\beta）^{2}*\frac{P_i*R_i}{R_i+P_i*\beta^2}$$

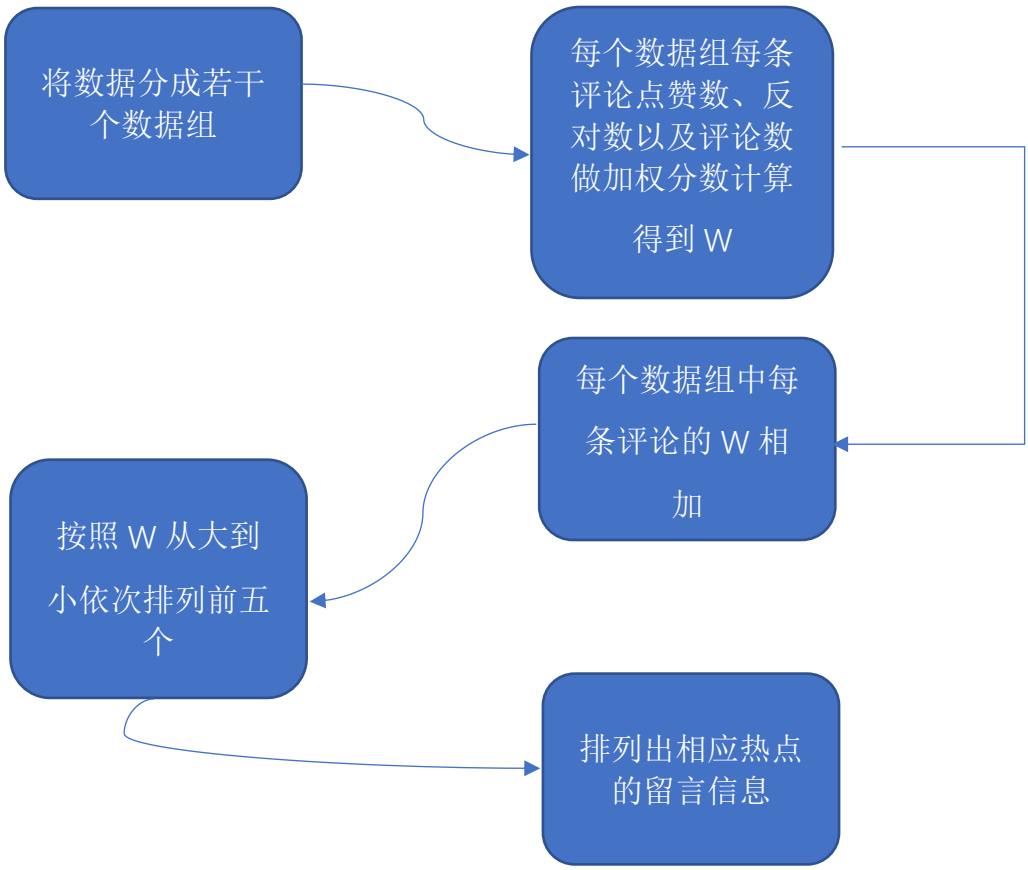
其中 p 为精确率（Precision），R 为召回率（Recall）虽然从计算公式来看，这二者并没有什么联系，但是在大规模数据集合中这二者往往不能相互兼得而是相互制约的，一般情况下 P 越高则 R 越低，R 越高则 P 越低。β 在该公式中表示权重，当 β >1 时操作者认为召回率更加重要，当 β <1 时操作者认为精确率更高，当操作者认为精准率和召回率同样重要时 β =1，所以有时候常常需要做出取舍。在本题中操作者认为 P 和 R 都十分重要于是操作者给 β 赋值为 1 故上述公变为：

$$F-score=\frac{1}{n}*\sum_{i=1}^{n}\frac{2*Pi*Ri}{Pi+Ri}$$

故上述方法越好则 F-score 数值越高，方法越差则 F-score 数值越低。

5.2 问题二的分析方法与过程

5.2.1 流程图



5.2.2 数据筛选

（1）、附件三中的留言可以先通过以某一时段某一人群或者某一地区的标准将留言分成若干部分，从而得到了若干组数据。

5.2.3 数据排序

（1）、得到了若干组数据后，设一共有 k 组数据，每组数据可表示为 θ_j（0≤j≤k）。每组数据包含了若干原附件三中的留言，附件三中每条留言有若干反对数和点赞数。假设第 θ_j 组数据中包含了 a 条留言；设第 i(0≤i≤a)条留言中有 b_i 条反对数，c_i 个点赞数；设 di 为点赞数与反对数的加权和，设点赞数的权重为 r，反对数的权重为 s，加权分数表示为：

$$di = rc_i + sb_i$$

每组数据中留言数的权重为 u，故第 θ_j 组数据中的加权总分数为 W_j；

$$w_j = ua + \sum_{i=0}^{n} d_i$$

热度评价指标同时也可以用 W 来表示，该公式可以根据认为的需要修改权重的大小，若操作者觉得一组数据组中的评论数或者点赞数或者反对数相较于其它二者更能反应该问题的热度，则可将该权重设置高于其它二者权重；若操作者觉得一组数据组中的评论数或者点赞数或者反对数对该问题的热点起到了副作用，也可将该权重设定为负值。

5.2.4 数据组排列

（1）、在得到 k 个 W 后，通过相互比较可以得出 W 值排名前五的数据组，热度从高到低排列问题 ID 依次为 1,2,3,4,5。

（2）、在具体研究该问题时，操作者将 s 设为 0 即不考虑反对数对问题热度的影响，将每个数据组中一条留言数视为五个点赞数：s=0、u=5、r=1；通过对附件三的筛选操作者可以得到前五大热点问题分别是“车贷诈骗案”、“高铁造成

的噪音问题”、“该路段存在公共交通、基础建设、校园等问题”、“高压杆线问题”及“小区夜宵油烟噪音扰民，还存在房屋质量、购房政策、入学困难等问题”。通过计算他们的热度指数分别为1587、726、225、208及165。将他们依次排列得出如下表格：

热度排名	问题ID	热度指数	时间范围	地点/人群
1	1	1587	2019/1/11至2019/7/8	A市58车贷案
2	2	726	2019/1/30至2019/9/6	A市绿地海外滩小区
3	3	225	2019/1/14至2019/11/21	A7县松雅湖
4	4	208	2019/3/26至2019/4/12	A6区月亮岛路
5	5	165	2019/1/8至2019/12/4	A市魅力之城小区

问题ID	问题描述	留言数	点赞数
1	车贷诈骗案	6	1557
2	高铁造成的噪音问题	7	691
3	该路段存在公共交通、基础建设、校园等问题	22	115
4	高压杆线问题	7	173
5	小区夜宵油烟噪音扰民，还存在房屋质量、购房政策、入学困难等问题	29	20

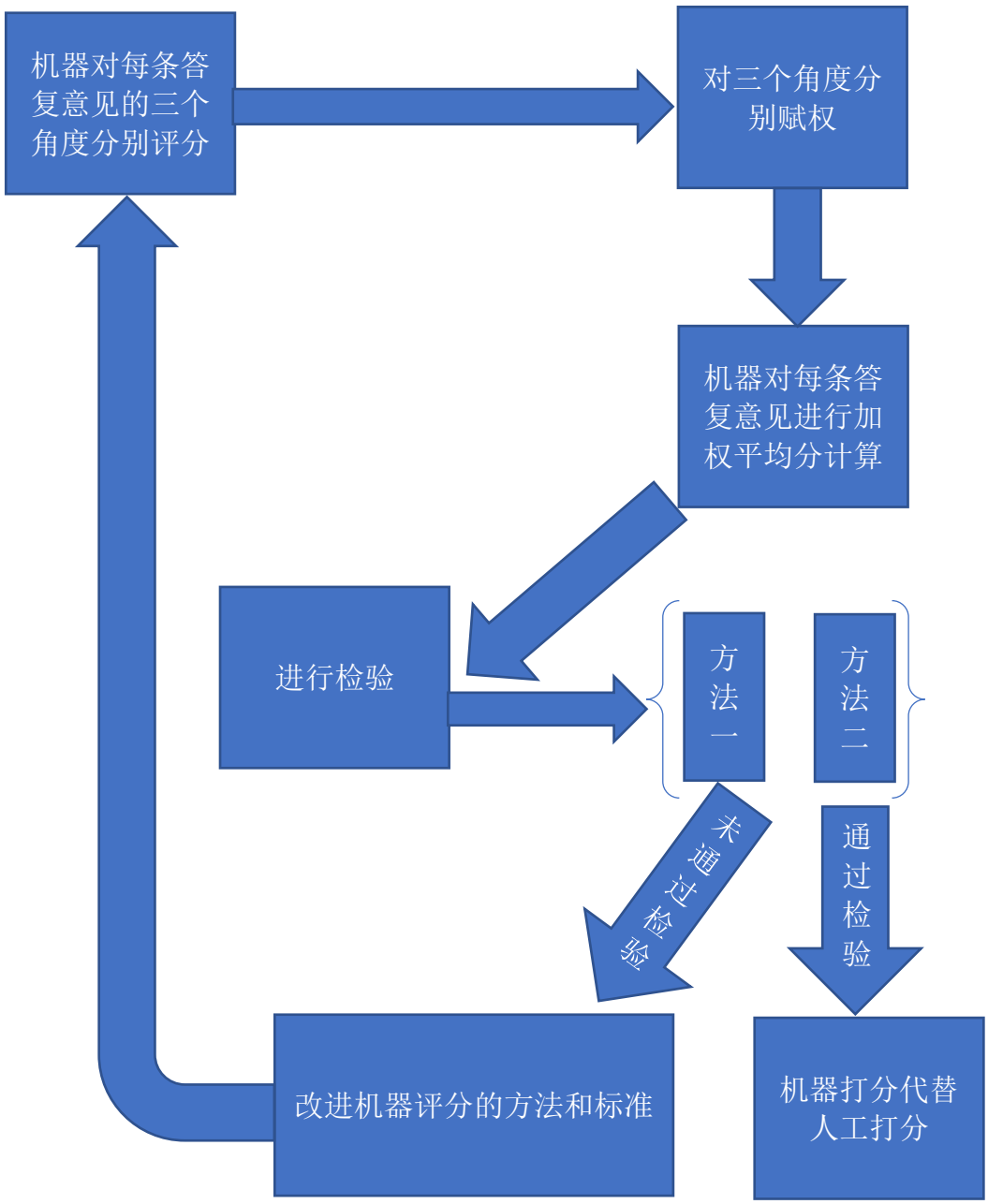
将上述表格进一步汇编制作成为文件“热点问题表.xls”。

5.2.5 列出热点问题明细

（1）、得到 W 值排名前五的数据组后，将每个数据组的每条留言明细按照问题 ID 升序排列出来即可得到“热点问题留言明细表.xls”。

5.3 问题三的分析方法和过程

5.3.1 流程图



5.3.2 答复意见打分

对于相关性的评价，依然采用 5.1.2 中所述 TF-IDF 算法同时对答复意见进行关键词筛选，对留言与回复意见的关键词的重合数进行判别后进行评分，相关性应当与关键词重合数成正相关。

答复意见的可解释性，实际上应当受答复本身的逻辑性、是否有相应法律法规做为依据等因素影响。机器打分很难对语言的逻辑性进行判别，故对答复中的‘依法’、‘依规’、‘规定’、‘法律条文’‘政府发文’等相关字词提取，依此类具有‘公信力’特征的词的出现频数进行评分。可解释性与该类词的出现频数成正相关。

答复的完整性的评分可综合相关性和可解释性两者的评价，容易理解若答复与问题愈相关，可解释性越高，其留言意见则愈完整。故完整性应当与相关性和可解释性成正相关。

由于留言答复的滞后性，不及时性，留言中所描述问题的实际情况可能随着时间的延长逐渐熵增；即答复越迟，留言所描述问题的具体情况可能已经改变，甚至有可能出现相关法律规定的更改修订等不可抗力的影响。故有理由认为问题留言时间和答复时间的时间间隔为评分的一个重要指标，为此对该情况引入新的评价指标‘滞后度’，该指标应当与评分成负相关。

5.3.3 答复意见加权平均分计算

假设操作者从四个角度（相关性，完整性，可解释性，滞后度）进行了打分。由 5.3.2 得出相关性的分数为 A，完整性的打分为 B，可解释性的分数为 C，滞后度分数为 J。接下来操作者需要对相关性完整性可解释性分别进行加权，设他们的权重分别为 e，o，g，v。由此操作者可以得到第 i 条回复的的加权平均分为 Di，公式可表达：

$$Di = \frac{e * A_i + o * B_i + g * C_i + v * J_i}{100 * e + 100 * o + 100 * g + 100 * J}$$

5.3.4 可行性检验（方法 1）

操作者经用机器打分得出了每条答复意见的分数，但为了检验机器检验是否得当，我们需对其进行复查。

操作者先在已全受过机器打分的答复意见中抽取 F 条答复意见的分数作为样本 X，第 i 条答复意见分数为 x_i。

计算可得出样本 X 的机器打分数学期望：

$$E(X) = \frac{\sum_1^F x_i}{F}$$

通过计算可得出样本 X 的方差为:

$$D(X) = \frac{1}{F-1} * \sum_1^F [x_i - E(x)] * [x_i - E(X)]$$

求出期望方差后，操作者将对这 F 条答复意见进行人工打分，将得到的分数作为样本 Y，第 i 条答复意见分数为 y_i。

通过计算可得出样本 Y 的数学期望：

$$E(Y) = \frac{\sum_1^F y_i}{F}$$

通过计算可以得出样本 Y 的方差为：

$$D(Y) = \frac{1}{F-1} * \sum_1^F [y_i - E(Y)] * [y_i - E(Y)]$$

要判断二者的评分是否较为吻合，操作者可以从他们的期望方差相关系数入手。

假设操作者能接受的期望误差程度为 h，当

$$h \leq |E(X) - E(Y)|$$

操作者视为不满足需求，当

$$h > |E(X) - E(Y)| \qquad (1)$$

操作者视为满足条件；

同理，假设操作者能接受的方差误差程度为 l，当

$$l \leq |D(X) - D(Y)|$$

操作者视为不满足需求，当

$$l > |D(X) - D(Y)| \quad (2)$$

操作者视为满足条件。

同理假设操作者能接受的相关系数最低为 m(m>0),当

$$\frac{Cov(X,Y)}{\sqrt{D(X)} * \sqrt{D(Y)}} \leq m \quad (3)$$

操作者视为满足条件，当

$$\frac{Cov(X,Y)}{\sqrt{D(X)} * \sqrt{D(Y)}} > m$$

操作者视为不满足需求。

综上所述，当同时满足（1）（2）（3）时，操作者将机器打分视为与人工打分同等有效，若其中一项不满足，操作者将机器打分视为无效。

5.3.5 可行性检验（方法二）

重复 5.3.4 中的部分步骤，操作者能得到机器打分的样本 X 与人工打分的样本 Y 并且得知他们的期望和方差。X 来自于机器打分总体，Y 来自于人工打分总体（人工目前未对总体打分）。

操作者假设原假设 H0 为两总体均值无显著差异，备选假设 H1 两总体均值有显著差异。设显著水平为 α 。

当 F≥30 时：

$$Z = \frac{[E(X) - E(Y)] - 0}{\sqrt{\frac{D(X)}{F} + \frac{D(Y)}{F}}}$$

当$Z > Z_{\frac{\alpha}{2}}$时，拒绝 H0 即均值有显著差异

当$Z \leq Z_{\frac{\alpha}{2}}$时，接受 H0 即均值无显著差异

当 0≤F＜30

$$t = \frac{[E(X) - E(Y)] - 0}{\sqrt{\frac{D(X)}{F} + \frac{D(Y)}{F}}}$$

其中 t 的近似自由度为 f

$$f = \frac{\left(\frac{D(X)^2}{F} + \frac{D(Y)^2}{F}\right)^2}{\frac{(\frac{D(X)^2}{F})^2}{F-1} + \frac{(\frac{D(Y)^2}{F})^2}{F-1}}$$

当$t > t_{\frac{\alpha}{2}}$时，拒绝 H0 即均值有显著差异

当$t \leq t_{\frac{\alpha}{2}}$时，接受 H0 即均值无显著差异

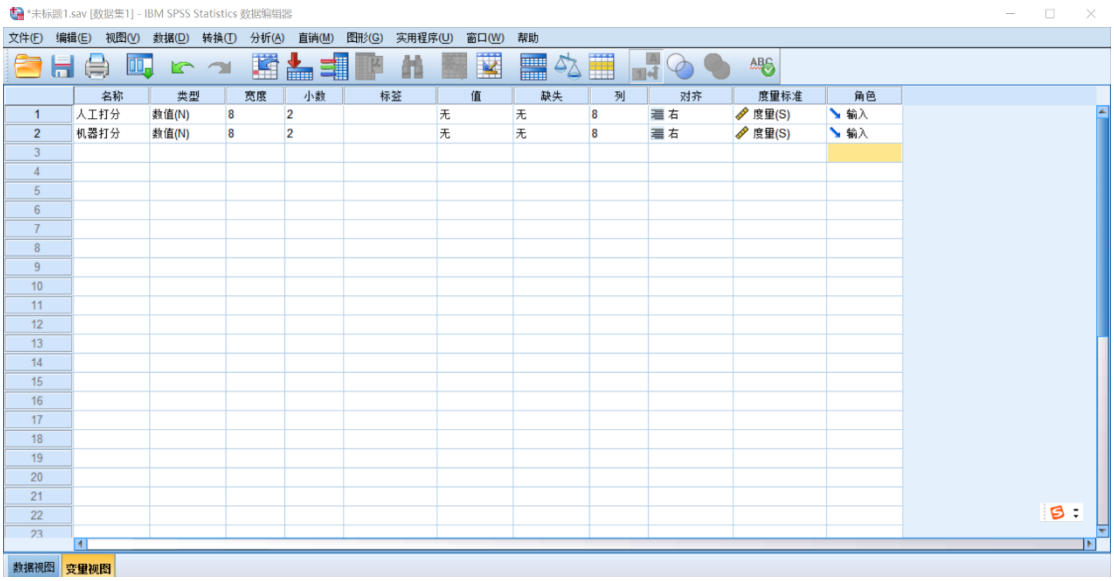
5.3.6 举例说明

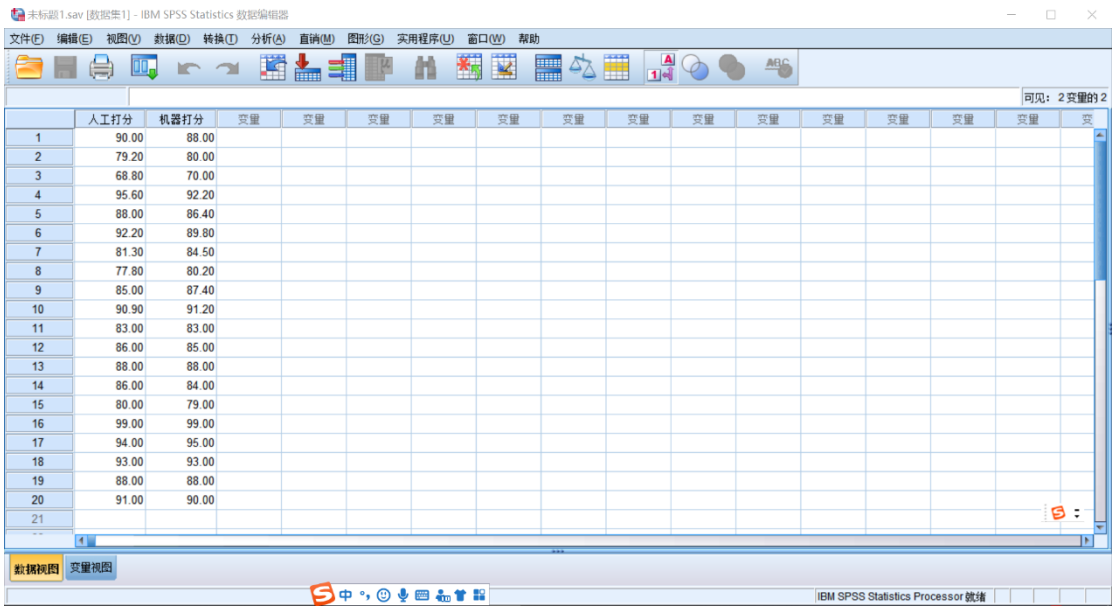
假设 H0 为两总体均值无显著差异，备选假设 H1 两总体均值有显著差异。

假设样本 X 中含有二十条经过机器打分的答复意见的分数，分别为 88.00 80.00 70.00 92.20 86.40 89.80 84.50 80.20 87.40 91.20 83.00 85.00 88.00 84.00 79.00 99.00 95.00 93.00 88.00 90.00。

假设样本 Y 中含有二十条经过人工打分的意见答复的分数，分别为 90.00 79.20 68.80 95.60 88.00 92.20 81.30 77.80 85.00 90.90 83.00 86.00 88.00 86.00 80.00 99.00 94.00 93.00 88.00 91.00。

设 α =0.05，代入 SPSS 得到：





成对样本统计量

		均值	F	标准差	均值的标准误
对 1	人工打分	86.8400	20	7.08188	1.58356
	机器打分	86.6850	20	6.45040	1.44235

成对样本相关系数

		F	相关系数	Sig.
对 1	人工打分 & 机器打分	20	.972	.000

成对样本检验

	成对差分				t	df	Sig.（双侧）
	均值	标准差	均值的标准误	差分的 95% 置信区间			

				下限	上限			
对 1	人工打分 - 机器打分	.15500	1.70648	.38158	-.64366	.95366	.406	19	.689

得到 t=0.406.通过查表我们可以得到$t_{\frac{0.05}{2}}$（19）=2.0930，$t < t_{\frac{0.05}{2}}$（19）故我们接受原假设即两总体均值无显著差异即机器打分和人工打分无显著差异。

6.模型优点和缺点

6.1 模型优点

（1）、通过 python 过滤数据消除了无关的条件，减少了工作量。

（2）、采用简单模型计算复杂模型，大大减少了运算量。

（3）、运用了 F-score 能够准确判断出分类方法的优劣。

（4）、通过运用相关系数能够较为准确判断出机器打分和人工打分是否具有线性相关性。

（5）、运用假设检验模型能较为精准地判断人工打分和机器打分是否具有显著性差异。

6.2 模型缺点

（1）、通过出现频率较高的词语集合代替留言仍会出现较大的误差以及遗漏留言中重要的信息。

（2）、模型中需要进行机器打分，但是目前机器打分是否靠谱仍需进一步商榷。

（3）、模型中提到用相关系数进行检验，但是相关系数只能检验其是否具有线性相关性，有可能出现两组数据具有相关性但是不是线性相关而是非线性相关等等的情况。

7.参考文献

[1] 盛骤.概率论与数理统计 [第四版].浙江大学

[2] 贾俊平.统计学 [第七版].中国人民大学

[3] 薛薇.统计分析与 SPSS 的应用 [第五版].中国人民大学

[4] 孙海锋,郑中枢,杨武岳的网络招聘信息的分析和挖掘.北京林业大学.2017

[5]梁昌明，孙冬强.基于新浪热门平台的微博热度评价指标体系实证研究.山东师范大学.2015