“智慧政务”中的文本挖掘应用

摘要

网络问政随着互联网的发展日渐成为民众参政议政的新渠道，各问政平台留言的文本信息数量也飞速增长。本文结合文本挖掘和自然语言处理等技术，建立了留言分类模型，并设计了热点问题和答复意见评价体系，能够极大程度上缓解政府部门处理问政平台信息的压力，推进政务处理智慧发展。

针对任务一，首先对文本进行预处理，借助 jieba 并结合构建的停用词表对留言详情进行分词，接着利用 word2vec 将分词转化为稠密向量，进行特征工程处理。为了避免训练集各分类文本下数据不均衡对分类模型的影响，我们进行了 SMOTE 过采样处理。最后利用训练好的逻辑回归（LR）模型对测试集文本进行分类，使用 F-Score 方法对分类效果进行评价，得分为 0.865698。

针对任务二，在识别相似热点问题前，我们先用点赞数和反对数进行模糊 C 均值聚类，缩小热点问题的集中范围。接着用计算杰卡德相似度来识别出同一热点问题下的留言。最后结合居民认可度、相关留言数、问题持续时间、内容充实度建立热度评价指标体系，进行灰色关联度分析，根据得分确定排名前五的热点问题。他们分别是 A 市 A4 区 58 车贷案、A 市丽发新城搅拌站扰民、西地省聚利人普惠投资有限公司涉嫌诈骗、A 市 A5 区五矿万境 k9 县存在一系列问题、A 市万润滨江房屋质价不符。

针对任务三，选取相关性、完整性、可解释性、时效性作为一级指标，类别相关性、内容完整度、礼貌规范性、内容充实度、文献性、针对性、回复及时性作为二级指标，建立官方答复评价指标体系。利用灰色关联度分析对官方答复情况进行评分，并按不同城市和不同类别进行了分组分析，给出了相应的评价与建议。

关键字： 逻辑回归模型；模糊 c 均值聚类；杰卡德相似度；灰色关联度分析

目录

一、 问题背景

随着网络的覆盖力和影响力日益扩大，互联网在人们生活中发挥的作用日益明显。中国公民通过互联网行使个人的监督权、知情权、表达权和参与权，进行网络问政的现象也变得更加普遍。通过网络问政，普通公民可以传递民意、建言献策；政府部门可以及时了解民情、倾听民心，从而实现科学决策，民主决策。

网络问政的飞速发展，使得公民表达民意的文本数据急速上升，给政府部门依靠传统人工对留言进行分类划分和梳理热点问题带来了很大挑战。但应运而生的大数据，云计算等技术为这些问题的解决提供了契机。因此我们希望能够通过文本挖掘，从中获取有价值的信息和知识，将文本表达转化为数据表达。利用自然语言处理技术，对文本分类、聚类、热度排序和回复评价等问题进行较好的解决，以帮助政府部门更好的处理问政平台上民众反馈的复杂多样的信息。

二、 问题分析

- **群众留言分类**

  政府部门在处理问政平台的留言时，需要先根据已建立的内容分类三级标签体系对留言进行分类，再分派到相对应的工作部门进行回复处理。这个过程要求工作人员迅速阅读留言，了解主要内容并确定所属分类。对人工的要求极高，耗时长且分类错误不可避免。因此我们希望借助计算机提取附件 2 中留言详情的主要内容，根据附件 1 提供的一级标签，建立分类模型，由此缓解依靠人工分类出现的问题。

- **热点问题探讨**

  热点问题即在某段时间内群众集中讨论的问题。对热点问题的及时发现及针对性处理，能够提高政府的政务管理水平。我们需要对附件 3 中的热点问题进行归类与提取并确定合理的评价指标，建立热度评价体系，对热点问题的热度进行评分。最终给出排名前 5 的热点问题及对应的留言明细表，以帮助政府部门迅速及时处理，对民众强烈反映的问题做出高效的回复与解决。

- **答复意见评价**

  对于民众反映的问题，相关部门是否做出了积极有效的答复关系到民众是否满意政府的工作。附件 4 给出了相关部门对留言的答复意见，我们需要从中进行探索，了解不同答复意见对相应留言问题的解决情况。因此要建立一套综合评价体系，从答复意见与留言问题的相关性、答复完整性、可解释性、时效性等方面进行评分，尽可能从不同角度分析政府部门的答复质量并尝试实现该套评价体系。

三、 符号说明

符号	意义
x	文本向量的特征值
y	一级标签编码出的预测目标值
P_i	第 i 类的查准率
R_i	第 i 类的查全率
A,B	样本的分词集合
$J(A,B)$	杰卡德相似度
c	聚类数目
b	加权指数
$s_i(i=1,2,\cdots,n)$	聚类的第 i 个样本
c_j	第 j 个聚类的中心
$f_j(s_i)$	样本 s_i 属于 f_j 类的隶属度
$x_{ij}(i=1,2,3,4)$	第 i 个热度评价指标下的第 j 个问题
y_0	参考数列
y_n	比较数列
$\eta_n(k)$	第 k 个指标下第 n 个热点问题的灰色关联系数
S_n	灰色加权关联度的得分
p_i	留言属于第 i 类的概率
p_i^*	答复属于第 i 类的概率
R	相关性

四、 任务一

4.1 数据清洗

利用 CLEAN，TRIM，SUBSTITUTE 函数对各附件留言详情的文本中出现的、由空格或非空格键等造成的空白内容进行清除，并利用查找和替换将上述操作无法替换的空白内容以及网页地址进行清除，避免对任务二、三中所选取的“内容充实度（留言字数）”指标的计算产生影响。

4.2 分词与去停用词

中文文本中往往包含很多功能词，帮助进行描述或表达概念（例如语气词、连词、代词等）。但与其他词相比，这些词并没有实际含义反而使文本显得更加繁冗。根据这些功能词构建停用词表，在后续文本处理中，遇到这些无意义的停用词即将其剔除，提高处理效率。同时也可以将标点符号等特殊字符放入停用词表中，使得文本在后续处理中更加精简。

jieba 可以对中文文本进行分词，词性标注等。为了避免中文表达中词组交叉导致的文本分词带来的语义不同的问题，利用 jieba“全模式”的分词模式并结合构建的停用词表对留言详情的文本内容进行切割分词，最终得到留言详情的分词表。

4.3 特征工程

特征工程是将原始数据转化为特征，更好表示预测模型处理的实际问题，提升对于未知数据的准确性。良好的特征能够对数据的固有结构进行较好的描述，可以使构建的模型更加简单，也可以提升模型的性能。

4.3.1 word2vec 词向量

基于词袋模型的词向量构建方法往往生成很稀疏的矩阵，对于词与词之间关系的表达能力有限，因此我们选择 word2vec 构建词向量，它可以将得到的分词表中的每个分词映射到向量空间，即将中文文本转化为计算机能够理解的稠密向量，并以此来表示词与词之间的对应关系。

4.3.2 特征选取

特征的选取对模型性能有着至关重要的影响。为了获取每条留言的特征向量，进行以下处理:

- 针对每条留言，获取其中每个分词在 word2vec 中的词向量；
- 计算每条留言下各分词在 word2vec 中词向量的平均数，即该留言对应的特征向量；

- 最终得到所有一级标签下每条留言的特征向量，共同构成特征矩阵。

4.3.3 SMOTE 过采样处理

我们分别从每类一级标签下随机选出 100 条作为测试集，余下的全部数据作为训练集。此时训练集中的文本数据分布情况如下：

表 **1** 训练集文本数据分布

交通运输	劳动和社会保障	卫生计生	商贸旅游	城乡建设	教育文体	环境保护
513	1869	777	1115	1909	1489	838

各一级标签下的留言数量不均衡，在一定程度上会对模型预测分类的结果有影响。为了减小样本数据不平衡带来的影响，使用 SMOTE 随机采样的方法对转化为特征向量的数据进行过采样处理。

SMOTE 过采样方法是在少数类样本周围随机选择一个与之最近邻的样本, 然后从两者连线上随机选取一个点作为新的少数类样本。

4.4 LogisticRegression 分类模型

对于多分类的文本问题，我们将其转化为对单条留言详情内容进行的二分类问题。即对于一条留言做出属于或不属于某一类一级标签的判断。LogisticRegression 模型可以看作被 Sigmoid 函数归一化之后的线性回归模型。其中 Sigmoid 函数引入了非线性映射，可以将一般线性回归（$-\infty, +\infty$）的值域缩小映射至 0-1 之间：大于 0.5 表示阳性，小于 0.5 表示阴性。由此可对二分类问题做出预测。

LR 模型假设函数为:

$$P(y = 1|x; \theta) = \frac{1}{1 + e^{-\theta^T x}} \tag{1}$$

它表示在给定 x 的条件下，事件 y（将因变量预测成 1（阳性））的条件概率。

若记

$$h_\theta(x) = \theta^T x \tag{2}$$

则 LR 分类的本质是通过学习确定一组权值 θ，当有测试样本 $x_1, x_2, ..., x_n$（测试样本的 n 个特征值）输入时，通过加权得到：

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n \tag{3}$$

进而求出 $P(y=1|x;\theta)$ 来判断每个测试样本所属的类别。

首先利用 LabelEncoder 对一级标签进行编码作为模型的预测目标值，再对训练集进行特征工程处理后，结合预测目标值，调用 Python 中 sklearn 库的 LogisticRegression 方法，训练 LR 模型。

4.5 模型评估

4.5.1 交叉验证

由于样本数据有限，为了检测训练好的 LR 模型在新数据上的预测效果，对构建好的 LR 模型进行交叉验证。调用 sklearn 库中的 ShuffleSplit 方法，对样本中的原始数据随机打乱后按照一定比例划分训练集与测试集，进行 K-fold 交叉验证。最终得到该 LR 模型的分类准确率将基本稳定在 86%~88%，对新的文本内容将有较好的预测分类效果。

4.5.2 F-Score 评分

我们用 F-Score 来对分类方法进行评价。记 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。F-Score 的表达式为：

$$F_1 = \frac{1}{n}\sum_{1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{4}$$

word2vec 构建词向量的目标是使得上下文相似的词距离相近，但并没有信息可以指导每个词向量的绝对位置，所以每次训练出来的词向量不相同。经过多次训练，我们结合 F1 评分情况，保存训练效果较好的 LR 模型，在对测试集文本进行分类时直接对其进行调用，最终得到各分类下的 F_1 得分和平均得分如下:

	Label	Precision	Recall	F1	Support
0	交通运输	0.855670	0.830000	0.842640	100
1	劳动和社会保障	0.836364	0.920000	0.876190	100
2	卫生计生	0.943182	0.830000	0.882979	100
3	商贸旅游	0.794118	0.810000	0.801980	100
4	城乡建设	0.780000	0.780000	0.780000	100
5	教育文体	0.913462	0.950000	0.931373	100
6	环境保护	0.949495	0.940000	0.944724	100
999	总体	0.867470	0.865714	0.865698	700

图 **1** **F-Score** 评分表

五、 任务二

5.1 热点问题识别

当某一时段内突然出现许多反映相同问题的留言时，我们就应该意识到这可能是个热点问题。通常来说，反映同一问题的留言内容会比较接近。但是反过来，内容接近的留言不一定是反映同一问题的。举个例子，“A 市餐饮店扰民”和“B 市餐饮店扰民”的内容非常接近，但显然不是同一件事。因此，我们先确定一点，如果两个留言是反映同一问题的，那么它们反映问题的地点和事件一定是一样的。

5.1.1 杰卡德相似度

我们利用杰卡德相似度来确定同一热点问题下的留言。设 A 为一个样本的分词集合，B 为另一个样本的分词集合，$J(A,B)$ 为杰卡德相似度。

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \tag{5}$$

杰卡德相似度越接近 1，则两个样本越有可能反映的是同一热点问题。

5.1.2 模糊 C 均值聚类

一方面，我们需要计算每条留言与其他留言的杰卡德相似度，从而选出最后热度 top5 的热点问题。巨大的样本数量会导致工作量过大且重复的操作过多。另一方面，后续采用的评价模型是灰色关联度分析模型。如果样本间的指标数值差距过大，会导致评价结果区分度不够。因此，我们在计算杰卡德相似度前，先对所有的留言进行模糊 c 均值聚类处理。

我们把附件三中所有的留言作为样本，根据留言的赞成数和反对数来进行模糊 c 均值聚类。显然，赞成数越高的留言所反映的问题，越可能成为热点问题。$s_i(i=1,2,\cdots,n)$ 是第 n 个样本。c 是聚类数目，$f_j(s_i)$ 是样本 s_i 属于 f_j 类的隶属度，b 是加权指数。所以，我们的目标函数为：

$$min\quad \sum_{j=1}^{c}\sum_{i=1}^{n}[f_j(s_i)]^b\|s_i-c_j\|^2 \tag{6}$$

根据 Bezdek 等人的经验，我们有如下约束条件：

$$\begin{cases} b=2 \\ c=3 \\ \sum\limits_{j=1}^{c} f_j(s_i)=1 \end{cases} \tag{7}$$

通过 MATLAB 编译模糊 c 均值聚类算法，得到聚类的中心如下：

表 **2**　聚类中心

类别	数量	赞成数	反对数
1	2	1932.69	2.45
2	4	753.28	0.001
3	88	24.17	0.68
4	4232	0.69	0.09

可见第一类、第二类和第三类的留言的赞成数远多于反对数。因此，我们认为这三类留言所反映的问题有可能成为潜在的热点问题。

5.2 热度评价指标体系

借鉴微博热度和视频热度的评价体系，我们的热度评价指标体系由四个一级指标构成。

- 相关留言数

  即反映同一问题的留言数目。相关留言数越多，就说明越多的人发帖反映该问题。

- 居民认可度

  即所有相关留言的点赞数与反对数的差值。点赞的人分两种：问题的经历者，认同留言的说法并希望有关部门的答复；普通群众，认为留言反映的问题确实存在并希望得到解决。因此，留言的点赞数越多，反对数越少，也会从侧面说明问题的影响力和真实性。

- 问题持续时间

  即相关留言的时间跨度。一个热点问题持续的时间越长，该问题的影响力和严重性也越大。

- 内容充实度

  即该热点问题下的相关留言所含的平均字数。字数越多，在一定程度上也能反映民众是否有足够的内容来说明该问题。这从侧面反映出问题的严重性和真实性。

表 **3**　热度评价指标体系

	一级指标	指标内涵
热度评价指标体系	相关留言数	反应同一问题的留言数目
	居民认可度	留言问题的点赞数与反对数的差值
	问题持续时间	最早的留言时间与最迟的留言时间的差值
	内容充实度	相关留言的平均字数

5.3 灰色关联度分析

为了衡量比较每个热点问题的热度，我们建立了一个灰色关联度分析模型。我们取热度评价指标体系中的四个指标作为评价指标，记为 $x_i, i = 1, 2, 3, 4$。评价指标一般分为效益型指标和成本型指标，顾名思义，效益型指标即越大越好的指标，成本型指标即越小越好的指标。

我们先将两种指标分别标准化:

$$x_{ij} = \begin{cases} \frac{x_{ij}-x_{imin}}{x_{imax}-x_{imin}}, \text{效益型指标} \\ \frac{x_{imax}-x_{ij}}{x_{imax}-x_{imin}}, \text{成本型指标} \end{cases} \tag{8}$$

j 是热点问题的数量。而每个指标对应的权重 ω_i 我们均取为 0.25。参考数列记为 $y_0 = \{y_0(k)|k = 1, 2, \cdots, i\}$，均取各标准最大值组成。比较数列记为 $y_n = \{y_n(k)|k = 1, 2, \cdots, i\}, n = 1, 2, \cdots, j$。接着我们要确定每个指标的灰色关联系数：

$$\eta_n(k) = \frac{\min\limits_s \min\limits_t |y_0(t) - y_s(t)| + \rho \max\limits_s \max\limits_t |y_0(t) - y_s(t)|}{|y_0(k) - y_n(k)| + \rho \max\limits_s \max\limits_t |y_0(t) - y_s(t)|} \tag{9}$$

最后，用 MATLAB 编译相应程序来计算灰色加权关联度作为评分 S_n:

$$S_n = \sum_{k=1}^{i} \omega_n \eta_n(k) \tag{10}$$

5.4 热点问题得分排名

我们的评价指标均为效益型指标。利用上述模型，我们得到了共计 75 个热点问题的得分。以下为部分结果：

表 **4**　热点问题得分排名

排名	问题	得分
1	A 市 A4 区 58 车贷案	0.59540383
2	A 市万家丽南路丽发新城居民区附近搅拌站扰民	0.521067807
3	西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金	0.514925403
4	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	0.511734801
5	A 市无良万润滨江天著牟取精装暴利	0.508442979
6	西地省需要尽快外迁京港澳高速城区段至远郊	0.507933564
7	A 市长房云时代多栋房子现裂缝，质量堪忧	0.457205331
8	A 市月湖市场的传销需要政府整治	0.455967982
9	A 市绿地城际空间站建筑质量太差了	0.448771923
10	A7 县公交线路需要接入市区统一实时公交查询 app	0.445646389
11	A7 县诺亚山林小区门口不应设置医院	0.439000581
12	A 市金毛湾配套入学的问题	0.43080897
13	A4 区洪山公园的建设计划不知何时才能正式启动	0.407355531
14	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到	0.400384405
15	A 市社保局 12333 电话一直没人接	0.39755437
...	...	...

六、 任务三

6.1 官方回复评价指标体系

官方回复评价指标体系由 4 个一级指标组成，4 个一级指标可再细分为 7 个二级指标。

- 类别相关性

  即留言与答复所属类别的相关性。设留言属于第 i 类的概率为 p_i，答复属于第 i 类的概率为 p_i^*。则定义相关性 R 的计算方法为：

$$R = 1 - \sqrt{\sum_{i=1}^{7}(p_i - p_i^*)^2} \tag{11}$$

  相关性 R 越大，则越能说明官方对留言分类的准确。针对留言所属的类别，派出相应部门给予回答，避免出现答非所问的情况。

- 内容完整性 & 内容充实度

  即回复的字数。我们认为字数越多，官方回复就越完整和充实。官方的态度以及是否言之有物也是评价回复的指标。

- 礼貌规范性

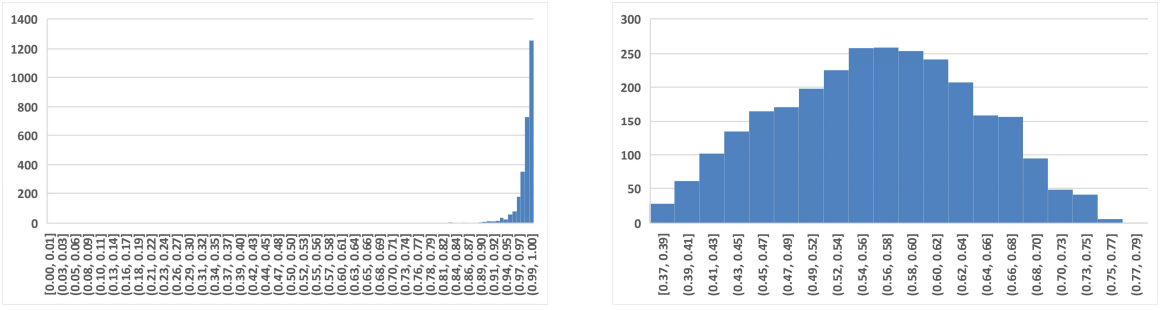
  即表明官方对用户的问候（“你好”或“您好”）、表明官方已知晓反映的问题（“收悉”或“据悉”）、表明官方对用户反映情况的感谢（“感谢”或“谢谢”）。我们认为官方的回复不能太过随意，因此需要用这三点来构建一套回复的模板，即一个标准的官方回复必须拥有这三点。这三个指标均为 01 变量。

- 文献性 & 针对性

  即官方是否在回复过程中针对留言中的某些问题来给出答复, 是否在回复过程中引用官方文献（如“《xxxx 公告》”或”《xxxx 标准》“等）。

- 回复及时性

  即官方回复与留言发布的时间差。时间差越短，则官方回复越及时，评分就会越高。但部分留言的时间差过大，会对评分造成干扰，故对其进行对数化处理来达到数值平滑的效果。



原分布　　对数化处理后分布

图 **2**　对数值归一化后的分布

表 **5**　官方答复评价指标体系

	一级指标	二级指标	指标内涵
官方答复评价指标体系	相关性	类别相关性	答复与留言所属类别的吻合程度
	完整性	内容完整度	官方回复的字数
		礼貌规范性	是否含“您好”或“你好”
			是否含“收悉”或“据悉”
			是否含“感谢”或“谢谢”
	可解释性	内容充实度	官方回复字数
		文献性	是否含有“《xxx》”
		针对性	是否含有““xxxx””
	时效性	回复及时性	官方回复与留言发布的时间差

6.2 分组分析

对字数较少（因为数据挖掘过程中导致的大块语句缺失）的样本进行清洗。相关性、完整性和可解释性均为效益型指标，时效性为成本型指标。对 2812 条回复进行灰色关联度分析。

6.2.1 不同城市

用 EXCEL 筛选出留言标题中含各市的留言所对应的官方回复，得到各市在相关性、完整性、可解释性、时效性以及综合得分上的评分。D 市、E 市、H 市、I 市样本较少，具有偶然性，所以不予评论。除 A 市、B 市外，其他市的得分情况大致为：相关性 > 完整性 > 可解释性 > 时效性。结合图表，我们大致给予以下建议：

- 提高官方回复的完整性

  A 市、B 市得分偏高，主要是因为它们大多的回复都有一套模板，即包含了对用户的问候、对消息的收悉和对用户的感谢。而其他市在这方面做的并不好，回复的格式都不相同，且有些随意。因此，我们建议其他市的官方回复可以建立一套模板，逐渐规范化。

- 提高官方回复的可解释性

  官方回复可以多针对留言的具体内容来给出解答，而不是针对留言反映的问题给出一个笼统的回答。这样会更有利于向用户表达官方的正面态度，增加用户对于

官方的好感。另外，在回复时多引用官方文件的内容，会更好的增加回复在用户心中的可信度。

- 提高官方回复的时效性

所有市在这方面做的都不好。虽然在回答一个留言时，可能需要翻阅详细的资料或向上级或下级询问从而耗费一定的时间，但让留言的用户等待官方的回复等上几个月甚至几年，我认为是不太好的，可能会让民众对有关部门产生不好的印象。所以在力所能及的范围，如果能提高一定的效率，尽快回复用户，对官方而言是非常有利的。

表 **6**　不同城市的得分情况

城市	数量	相关性	完整性	可解释性	时效性	综合得分
A 市	364	66.25	80.97	61.08	38.91	60.29
B 市	107	63.72	78.22	54.56	57.00	60.48
C 市	59	72.59	63.49	53.95	46.53	55.92
D 市	5	35.90	46.81	37.97	44.37	42.52
E 市	1	27.00	33.63	78.18	41.98	45.21
F 市	29	57.66	58.35	49.84	41.73	50.06
G 市	87	68.20	65.57	47.70	40.84	53.02
H 市	8	43.02	56.38	47.39	46.86	47.31
I 市	3	39.84	55.62	40.83	27.31	42.78
J 市	31	66.85	56.82	53.89	40.30	51.76
K 市	64	65.26	61.46	59.04	34.74	52.43
L 市	49	63.59	59.01	43.07	48.77	51.37
M 市	39	70.17	67.79	50.79	41.79	55.00

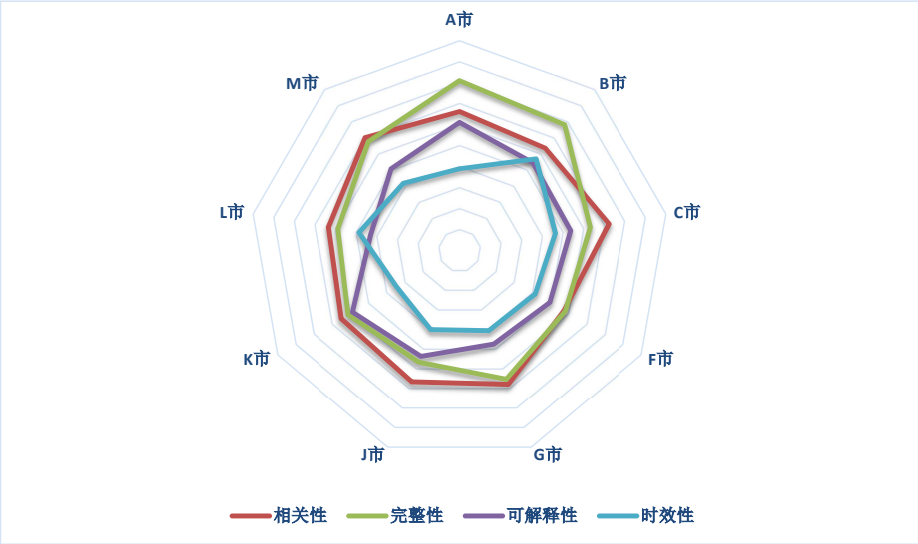


图 **3**　不同城市的得分情况

6.2.2 不同类别

与不同城市类似，用任务一中的 LR 模型来判断留言的类别。根据类别，观察同一类别下的官方回复在各个方面的得分情况。观察图表可以发现，各个类别的官方回复并没有差距很大，尤其在可解释性和时效性的得分上。

表 **7**　不同类别的得分情况

类别	数量	相关性	完整性	可解释性	时效性	综合得分
交通运输	404	61.78	69.23	54.04	44.22	54.65
劳动和社会保障	322	68.34	65.11	55.62	44.19	55.59
卫生计生	141	58.60	61.56	59.46	43.80	53.20
商贸旅游	275	60.91	68.80	56.79	43.19	54.67
城乡建设	1119	68.17	71.13	55.67	43.56	56.93
教育文体	328	77.02	69.19	58.85	45.13	59.58
环境保护	222	58.22	66.29	54.55	42.82	53.37

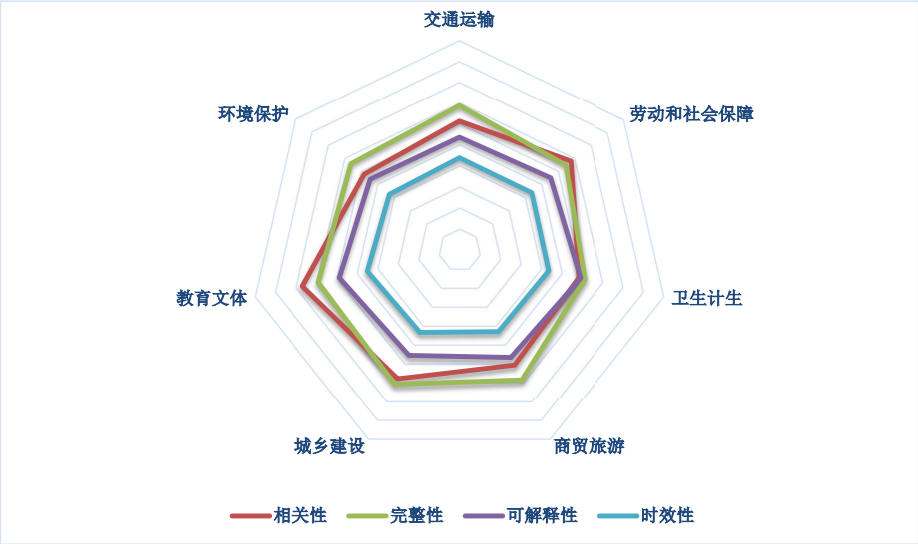


图 **4**　不同类别的得分情况

结合两个分组分析，我们基本上可以得出结论。官方回复的得分更多受所在城市的影响，而不是所属类别的影响。如果一个城市注重官方回复，那么不管用户提出哪一类的问题，大多都能得到一个好的回答。

参考文献

[1] 汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究 [J]. 计算机系统应用,2018,27(05):209-215.

[2] 戴维. 逻辑回归解决文本分类问题 [J]. 通讯世界,2018(08):266-267.

[3] https://www.jianshu.com/p/6c8588d40d59

[4] https://www.cnblogs.com/kukri/p/8566287.html

[5] https://www.cnblogs.com/xiaosongshine/p/10557891.html

[6] 梁昌明，李东强. 基于新浪热门平台的微博热度评价指标体系实证研究 [J]. 情报学报，2015，34(12):1278-1283