基于 GRU 循环神经网络、k-means 聚类、Textrank 摘要提取的民意分析

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，各类与民意相关文本数据往往主要依靠人工来进行留言划分和热点整理，这给相关部门的工作带来极大的挑战。随着大数据、人工智能等技术发展，我们可以使用机器学习中监督学习学会留言分类，通过自然语言处理和无监督学习在众多留言中提取热点信息，最终也需要给出模型评价指标，反映政府部门反馈的准确性、全面性等。

针对第一问，在本次数据挖掘中，本文首先对得到的留言数据利用 python 进行数据预处理，如主题与详情的融合、文本 jieba 分词、文本去重、特殊符号的删除、词语泛化、样本规范化。使用北京师范大学中文信息处理研究所与中国人民大学 DBIIR 实验室的研究者开源了中文词向量语料库训练的词向量进行**词嵌入**，最后使用 **GRU 循环神经网络**，确定一级分类模型的参数。训练集准确率达到 89.35%，验证集准确率为 88.68%。

针对第二问，对留言进行第一问中的预处理得到词向量，使用题一 **GRU 模型**得到留言的一级分类。对每条留言中的词向量加权平均且单位化得到模长为 1 的句向量。构建每条留言的热度指数，同时对每个一级分类使用 **k-means 聚类**，结合热度指数和各一级标签之间的聚类，分析各自的热点事件。最终使用 **Textrank 算法**分别对留言提取摘要。

针对第三问，我们从留言详情与答复意见中得到分词结果，利用答复与留言的匹配程度构建**完整性指标**；在完整性指标基础上，利用答复的详细程度构建**解释性指标**；将留言详情与答复意见转化为题二中的句向量，通过欧几里得平均距离构架两者的**相关性指标**；将答复时间与留言时间的间隔作为答复的**及时性指标**。

关键词：Word2Vector　GRU 循环神经网络　k-means 聚类　Textrank 算法

目录

1. 挖掘目标

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。

第一、在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系，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。

第二、某一时段内群众集中反映的某一问题可称为热点问题，需要定义合理的热度评价指标，及时发现热点问题，提升服务效率，并给出评价结果

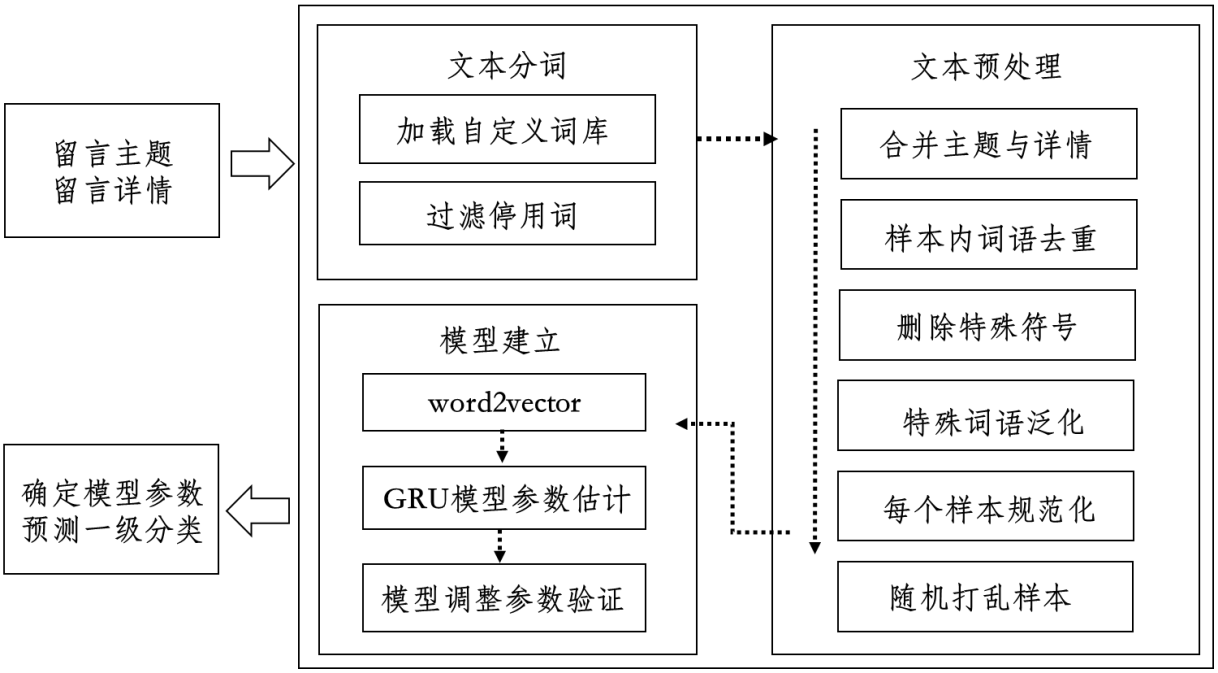
第三、针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

2. 分析方法与过程

2.1 问题 1 分析方法与过程

总体流程如图 1，使用 python 编程工具，对文本进行分词、预处理、估计 GRU 模型参数，建立一级分类模型，进行预测分类。

2.1.1 总体流程



图表 1

2.1.2 具体步骤

数据介绍

本题使用的实验数据来自互联网公开来源的群众问政留言记录，使用附件一的一到三级标签体系以及附件二中留言主题、留言时间、留言详情以及分类标签。

文本分词

使用 jieba 分词库进行分词

Pkuseg 是由北京大学语言计算与机器学习研究组研制推出的全新中文分词工具包，具有高分词准确率，能支持多领域分词，在不同领域的数据上都大幅提高了分词的准确率。与 jieba 相比，对地点的分词准确率较高，但是时间效率与实用性极低，实验对比，保留使用 jieba 进行分词，并适当扩充 jieba 分词的词库

以达到更好的效果。

停用词过滤

在中分分词之后，将分词结果处理成词的集合，应该删除文本表达含义无意义的词语，减小它们对后续工作产生的消极影响。这类词为停用词，停用词其特征是具有普遍性、高频性、且信息量非常低，如留言中的“的”、“了”、“啊”等。

文本预处理

因为文本评论中存在大量价值含量低的信息，如果直接对数据进行分词及词频统计等操作，会对很大程度上影响分析结果，得到不准确、存在问题、甚至不合理的结果，所以在分析之前先对文本进行预处理。

合并主题与详情

考虑到留言主题和留言详情都包含重要且相似的信息，所以将每一个样本的留言主题与留言详情的合并为同一文本，进行统一处理。

文本去重

文本去重，就是去除同一留言中重复的部分。实际上因为文本的性质是网络问题留言，不同留言中完全或接近重复的留言极少，所以本文的去重工作放在分词之后，对同一文本重复词去除，不需要使用其他文本去重算法。

特殊符号去除

留言中有大量包含半角与全角的无用标点符号等特殊符号，删除特殊符号，利用 re 正则化程序包去除。

文本泛化

将文本中的大量地址转化为同一个特征，如“A3 市”化为“某市”、将所有网址化为“域名”等。

文本规范化

为提高训练效率，在总体词集合中根据频率选出前8000个词语。将每个样本的分词结果统一长度，取所有长度的最大值为统一长度，对样本词数未满的数据进行补充，补充元素为 0。具体步骤如下：

每一个样本形如[*词语* 1, *词语* 2, … , *词语K*]的向量，均映射至一个p维向量，其中p为样本最大长度，若第i个样本在分词处理后有K_i个词，每一个词在词库中对应的标签记为a_j，则样本i映射为p维向量如下：

$$Sentence_i = \left[0, \dots, 0, a_{i1}, a_{i2}, \dots, a_{iK_i}\right]_{1\times p},\ p = MAX\{K_i\}$$

模型训练

Word2Vector

北京师范大学中文信息处理研究所与中国人民大学 DBIIR 实验室的研究者开源了中文词向量语料库，该库包含经过数十种用各领域语料训练的词向量[i]。

在经过文本预处理后，需要将词语转化为词向量，需要对每个样本词使用知乎语料库中训练的词向量匹配，其中已训练的词向量为1×300维。若匹配不到则词向量记为 0，具体如下：

已知预处理后第i个样本如下：

$$Sentence_i = \left[0, \dots, 0, a_{i1}, a_{i2}, \dots, a_{iK_i}\right]_{1\times p},\ p = MAX\{N_i\}$$

对每一个词匹配为一个 300 维的词向量，得到以下结果：

$$Word_{a_{ij}} = [b_1, b_2, \dots, b_{300}]$$

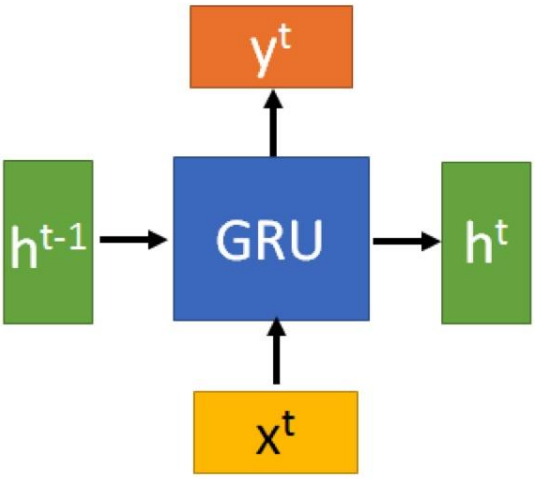
最终得到一个维数为$p \times N \times 300$的样本空间，N为样本数量，p为样本最大长度，300为词向量长度。

GRU 模型估计

对文本分类使用传统神经网络效果较差，故考虑循环神经网络（RNN），引入有记忆单元和门控记忆单元保存历史信息、长期状态的循环神经网络。而循环神经网络也有多种变体，通过实验可以发现, 基于门控结构的循环网络比传统的 RNN 的性能更好，其中长短期记忆网络（LSTM）与 LSTM 变体 GRU 的表现基

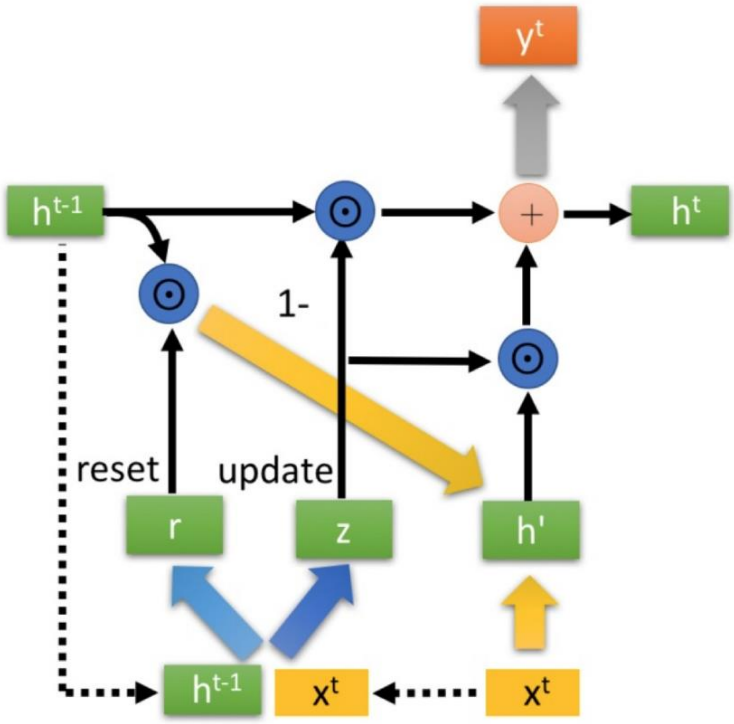
本相当，且 GRU 效率更高[2]。所以为提高效率采用 GRU 循环神经网络。

GRU 输入输出结构如图 2：



图表 2

GRU 单元结构如图 3：



图表 3

具体步骤如下：

1.输入以转化为词向量的样本X：

$$X = \begin{bmatrix} Word_{a_{11}} & Word_{a_{12}} & \cdots & Word_{a_{1p}} \\ Word_{a_{21}} & Word_{a_{22}} & \cdots & Word_{a_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ Word_{a_{n1}} & Word_{a_{n2}} & \dots & Word_{a_{np}} \end{bmatrix}$$

其中$Word_{a_{ij}} = [b_1, b_2, \dots, b_{300}]$

3.建立 GRU 模型，设置 32 个 GRU 单元。

4.在 GRU 后构建一层 256 单元全连接，激活函数选择为 relu 函数。

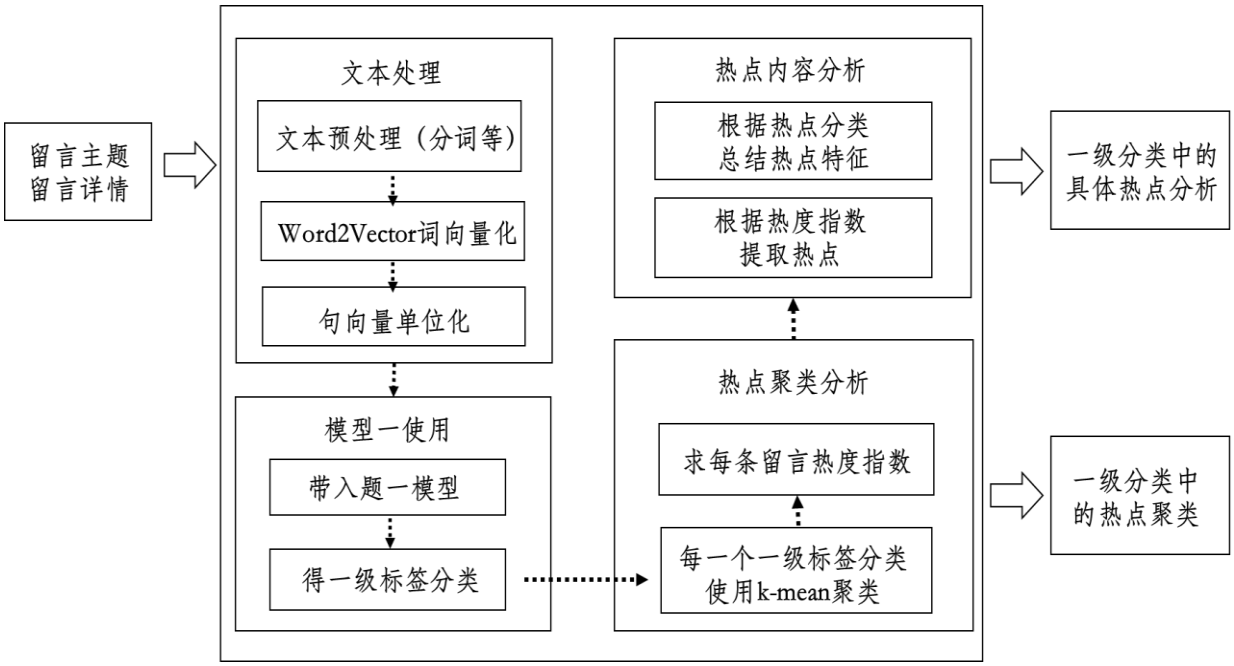
5.输出层选择 softmax 函数分类，选择交叉熵损失函数和随机梯度下降。

6.根据训练集与验证集的结果调整得到模型。

2.2 问题 2 分析方法与过程

2.2.1 总体流程

本题总体流程如图 4



图表 4

分析流程大致分为以下：

第一步：对留言进行分词、向量化、加权平均且单位化得到句向量（模长为 1）。

第二步：使用题一的模型对所有样本进行一级标签分类。

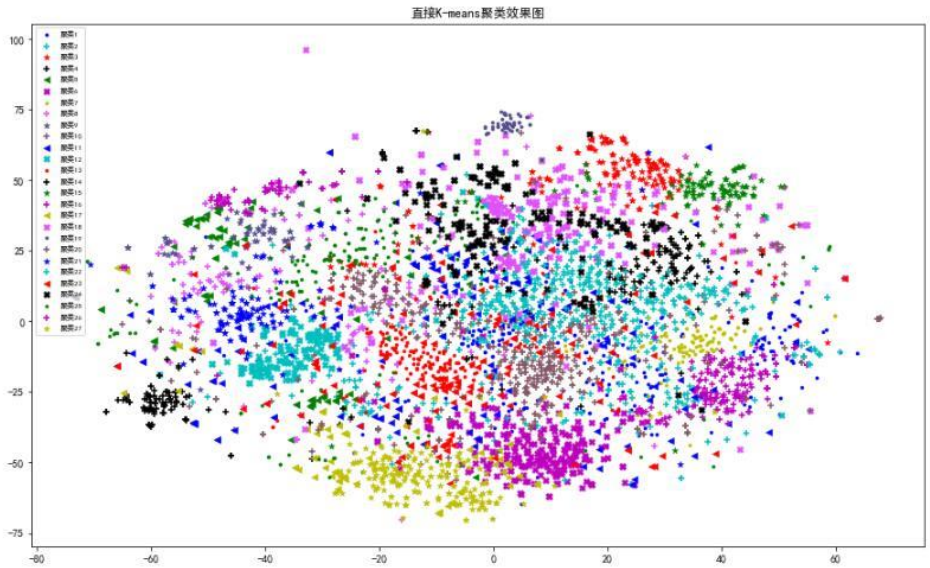
第三步：对每个一级标签分类使用 k-means 聚类，且计算每个样本的热度指数。

第四步：通过每个一级标签的分类，分别根据热度指数提取热点内容，分析热点。

2.2.2 GRU 循环神经网络一级标签分类

第一步是文本处理，在第一问中将每一个留言都映射到一个向量上。分词与词向量化具体步骤题一已表明，不赘述。将单个留言的词向量加权平均得到句向量，且对句向量单位化（即把句向量化为模长为 1 的向量）。

第二步是题一模型的使用，如果直接对总体样本进行聚类，则过多因素混杂，效果不理想，尽管分成 27 类分数最高，但聚类内容比较混乱，**如图 5**。所以需要对总体样本进行一级标签分类（借用题一的一级分类模型），分类之后再对单独一类进行聚类，此举聚类效果比总体聚类更有针对性，效果更好，热点信息更为准确，第二步的模型具体步骤在题一中表明。



图表 5

2.2.3 k-means 热点聚类与热度指数

第三步中需要对每个一级标签类别数据进行聚类，得到相应的热点类别。再定义计算每个留言的热度分数。

K-means 聚类

队伍标签样本$X_i = (x_1, x_2, \dots, x_{300})'$，生成$k$个簇，损失函数如下：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

其中μ_i为簇C_i的中心点

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

再聚类过程中利用轮廓系数（Silhouette Coefficient），判断聚类是否合理，是否有效，最终确定系数较大的对应的k值。

热度指数

因为每一条留言信息都包含赞成数量与反对数量，同时也有其他留言的观点与之相似。因此热度指标构造如下：

设一个一级标签分类下总体留言为m条，已分为k类，则第t类的留言数量n_t与总体之比$\theta = \frac{n_t}{m}$为类规模，标准化后得到类间热度指数为

$$v_t = \theta(\alpha x_{\text{赞成}} + \beta x_{\text{反对}} + 1)$$

其中α为赞成的权重，β为反对的权重，x为赞成或反对的数量，θ为该留言所在类的规模。本题中赞成的权重与反对的权重应各位0.5，但考虑到反对者越多及比普通热点可能引起矛盾更大，需要更加关注，故将反对者的权重调高至0.7。

2.2.4 Textrank 摘要提取

考虑到题目要求生成留言的简单摘要，本题使用 textrank 提取关键词且提取摘要

关键词提取基本算法如下：

1)把给定的文本 Text 按照完整句子进行分割，即

$$Text = [S_1, S_2, \dots, S_m]$$

2)对于每个句子$S_i \in Text$，进行分词和词性标注处理，并过滤掉停用词，只保留指定词性的单词，如名词、动词、形容词，即

$$S_i = [t_{i1}, t_{i2}, \dots, t_{in}], \qquad t_{ij}\text{为保留的候选关键词}$$

3)构建候选关键词图$G=(V,E)$其中V为节点集，由（2）生成的候选关键词组成，然后采用共现关系（co-occurrence）构造任两点之间的边。根据以下公式迭代传播各节点的权重，直至收敛。

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

4)对节点权重进行倒序排序，从而得到最重要的T个单词，作为候选关键词。

5)由 4 得到最重要的T个单词，在原始文本中进行标记，若形成相邻词组，则组合成多词关键词。

TextRank 生成摘要

将文本中的每个句子分别看做一个节点，如果两个句子有相似性，那么认为这两个句子对应的节点之间存在一条无向有权边。考察句子相似度的方法是下面这个公式：

$$Similarity(S_i,S_j)=\frac{|\{w_k|w_k\in S_i\&w_k\in S_j\}|}{\log(|S_i|)+\log(|S_j|)}$$

公式中，S_i,S_j分别表示两个句子词的个数总数，W_k表示句子中的词，分子是同时出现在两个句子中的同一个词的个数，分母是对句子中词的个数求对数之和。分母这样设计可以遏制较长的句子在相似度计算上的优势。

我们可以根据以上相似度公式循环计算任意两个节点之间的相似度，根据阈值去掉两个节点之间相似度较低的边连接，构建出节点连接图，然后计算 TextRank 值，最后对所有 TextRank 值排序，选出 TextRank 值最高的几个节点对应的句子作为摘要。

2.3 问题 3 分析方法与过程

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。

在变量定义前对数据进行预处理，对附件 4 的留言详情与答复意见分别进行分词、去词、去重等第一问的基本操作后，得到词集合，下面给出评价方案。

2.3.1 变量定义

第i条答复的词集合$R_i = \{word_1^{(r)}, \ldots, word_{k_i}^{(r)}\}$，$k_i$为第$i$条答复的关词数量

第i条问题的词集合$Q_i = \{word_1^{(q)}, \ldots, word_{l_i}^{(q)}\}$，$l_i$为第$i$条问题的词数量

第i条问题与答复的全集$C_i = Q_i \cup R_i$

2.3.2 完整性模型

答复完整性定义为：其中$|R_i \cap Q_i|$为第i个样本留言问题与留言答复共词的数量。$|Q_i|$为第i个样本留言问题的关键词数量。I_i为第i个样本答复率。

$$I_i = Intergrety_i = \frac{|R_i \cap Q_i|}{|Q_i|}$$

因为答复率在严格意义上可以达到100%，因此指标I的构造过程需要经过Box-Cox 幂变换如下式，并给出完整性分布。

$$I_i^* = \frac{I_i^\lambda - 1}{\lambda}$$

2.3.3 可解释性

可解释性为答复对问题的解释程度，即答复的详细程度：

$$E_i = Explain_i = I_i^* \frac{|R_i \cup Q_i - Q_i|}{|R_i|} = \frac{I_i^\lambda - 1}{\lambda} \frac{|R_i \cup Q_i - Q_i|}{|R_i|}$$

$$E_i^* = Explain_i^* = \frac{E_i^\lambda - 1}{\lambda}$$

其中$\frac{|R_i \cup Q_i - Q_i|}{|R_i|}$为答复中除了问题关键词之外的词语比例，反映了答复内容的多少.故E_i结合了完整性和答复内容的多少，反映了答复的详细程度.对此进行Box-cox 变换后给出可解释性分布。

2.3.4 相关性

本题使用第一问中已训练过的词向量匹配附件 4 的文本词语，即重复第一问的 Word2Vector 的步骤，每个词语都化为以下词向量：

$$Word = (a_1, a_2, \dots, a_{300})$$

在留言详情与答复意见文本中，分别把各自的词向量加权平均得以下句向量：

$$V = \sum a_k Word_k = (b_1, b_2, \dots, b_{300})$$

对留言详情与答复意见的句向量求其欧几里得平均距离，如下式：

$$S = \sqrt{len(V)^{-1}\left(V_{答复意见} - V_{留言详情}\right)\left(V_{答复意见} - V_{留言详情}\right)'}$$

通过幂变换后能反映答复意见与留言详情的相关性，若距离越小，则相关性越大，距离越大，则相关性越小。给出相关性分布。

2.3.5 及时性

答复的及时性可通过答复时间与留言时间的差值反映答复的及时性，将时间数值化后得到及时性指标：

$$\Delta Time = T_r - T_q$$

其中T_r，T_q分别为答复时间与留言时间。

2.3.6 完整性模型改进

构造Q_i的满足以下条件的若干(p_i个)真子集$Q_i^{(j)}$，即将问题集合切分

$$\bigcup_{j=1}^{p_i} Q_i^{(j)} = Q_i \ ， \ Q_i^{(j)} \bigcap_{j \neq k} Q_i^{(k)} = \phi$$

给定示性变量表示第i个回答的内容是否与第i个问题的第j个真子集共词

$$X_i^{(j)} = \begin{cases} Q_i^{(j)} & \text{for} \quad Q_i^{(j)} \cap R_i \neq \phi \\ \phi & \text{for} \quad Q_i^{(j)} \cap R_i = \phi \end{cases} \quad j = 1,2,\dots,p \qquad (5)$$

答复完整性定义如下，即真子集被答复的比例：

$$I_i = Integrety_i = \frac{|\cup_{j=1}^{p} X_i^{(j)}|}{|Q_i|} \qquad (6)$$

3. 结果分析

3.1 第一题结果

模型如图 6：

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 4212, 300)         22912800
_________________________________________________________________
dropout_1 (Dropout)          (None, 4212, 300)         0
_________________________________________________________________
gru_1 (GRU)                  (None, 32)                31968
_________________________________________________________________
dropout_2 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               4224
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 903
=================================================================
Total params: 22,949,895
Trainable params: 37,095
Non-trainable params: 22,912,800
_________________________________________________________________
```

图表 6

GRU 循环神经网络训练结果如表 1：

表格 1 GRU **训练结果表**

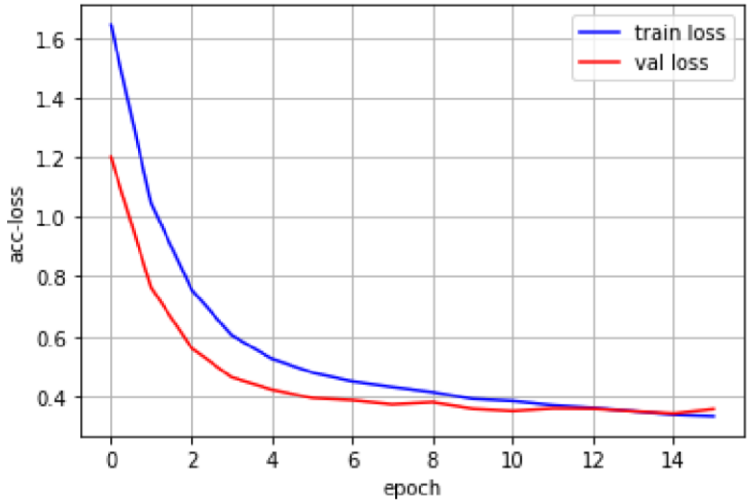
训练集损失	训练集准确率	验证集损失	验证集准确率	F1-Score
0.3336	0.8935	0.3572	0.8868	0.8802

运行 14 小时，得到上述结果。在使用评论集中的 80%的数据作为训练集，20%作为验证集。可以看出在 32 个 GRU 单元与一层 128 单元全连接下，模型自动学习分类能力优良，采用交叉熵损失函数，训练集损失达到 0.3336，验证集损失达到 0.3572，训练正确率达到 89.35%，测试准确率达到 88.68%。

因为模型在词向量基础上，使用 GRU 循环神经网络，与全连接普通神经网络结合后，随着迭代次数增加而渐优。但是训练效果变好的同时需要考虑验证集的效果。模型精度还有待提高，尤其是电脑硬件设备不足，可以考虑更为高效准

确的模型，而且语料库的正确选用对词向量的影响能进一步提升 GRU 模型的精度。

首先我们观察损失函数随着对数据的迭代次数的而减小（即 epoch 的增加），训练损失函数与验证损失函数变化图如图 7。可以见到当 epoch 达到 14 左右，验证集的损失开始增大，容易出现过拟合。根据该特征选择 epoch=16。

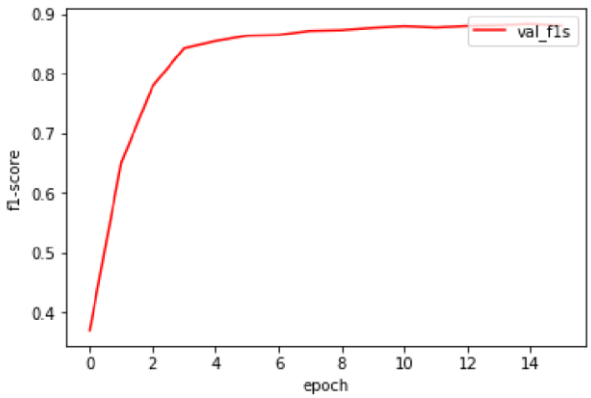


图表 7

求解模型结束后使用问题的评价模型，即对留言内容进行标签分类模型的评价，使用 F-Score 公式

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

其中P_i为第i类的查准率，R_i为第i类的查全率。随着迭代次数增加，模型的评价分数越高，当 epoch 取 X=16, F-Score 分数为 0.88(实例)，如图 8。



图表 8

3.2 第二题结果

对每一个留言计算热度指数得到以下热点问题结果如表 2

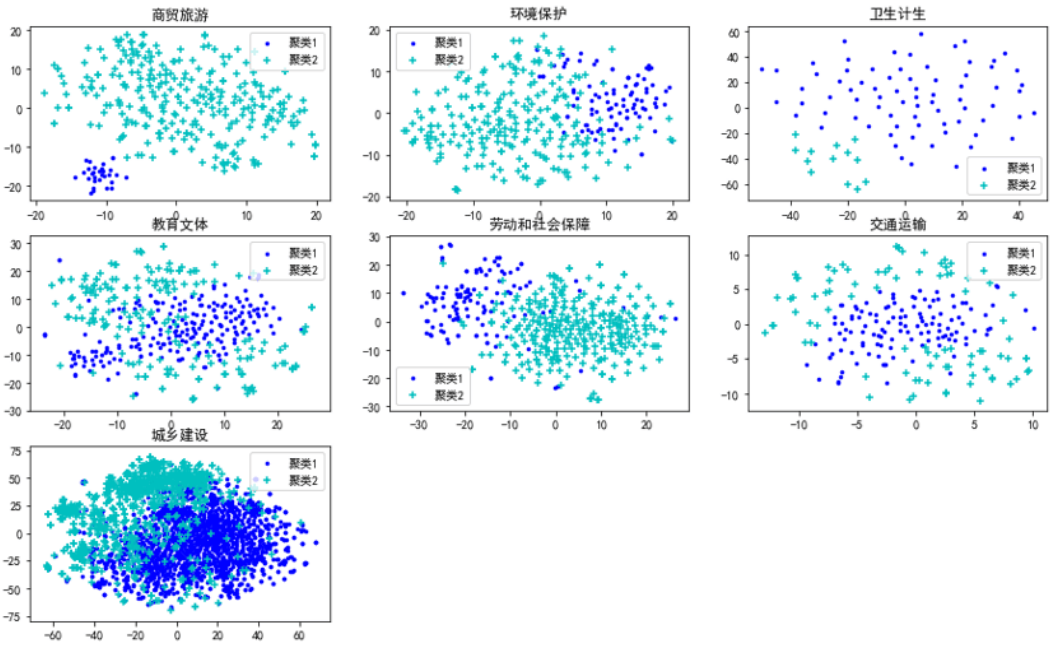
表格 2 热点问题表

热度排名	问题ID	热度指数	时间范围	地点人群	问题描述
1	1	4015.9	2018/10/26 至 2020/01/06	A 市地铁	请 A 市加快地铁的建设
2	2	3723.5	2018/12/31 至 2020/01/25	A 市区柏家小区	举报 A 市 A3 区柏家塘小区私营 ktv 严重扰民
3	3	2557.7	2017/06/07 至 2020/01/06	A 市区稻田中学	A 市 A5 区稻田中学午餐伙食太差劲
4	4	1009.1	2019/01/06 至 2020/01/05	西地省 A 市宏源有限公司	西地省 A 市盛天宏源有限公司诱骗消费者
5	5	608.9	2019/01/02 至 2020/01/06	A 市区省鸟玄科技有限公司	恳请帮忙解决 A 市 A3 区西地省鸟玄科技有限公司拖欠员工薪资问题

利用分类模型分类后共有七大类，即城乡建设、环境保护、交通运输、教育文体、劳动和社会保障商贸旅游、卫生计生。对这七大类分类使用 k-mean 聚类，聚类过程中采用轮廓系数判断$k = 2$的时候，将商贸旅游、环境保护、劳动和社会保障和城乡建设分成 2 类热点、其余一级分类不进行分类，总计十一个热点。**如图 9，**其中热点 1 问题留言明细表如表 3，其余热点明细参见附录。

表格 3 热点 1 问题留言明细表（部分）

留言编号	留言用户	留言主题	留言时间	留言详情	反对	点赞
188170	A88011323	A 市 6 路公交车随意变道通行	2019/12/23 8:50:24	12 月 21 日下午 17 时 52 分许，6 路公交车在 A3 区大道金星路口由西往东辅道左拐通行时，该司机并未按地面车辆指示导向通行……	0	0
188251	A00013092	A7 县特立路与东四路口晚高峰太堵，建议调整信号灯配时	2019/10/19 11:02:40	近来，下午晚高峰五点半左右，经过特立路与东四路口时，东往西方向车越来越多……	0	0
188409	A0003274	A 市地铁 3 号线星沙大道站地铁出入口设置极不合理！	2019/6/19 10:14:39	我是 A7 县星沙街道的一个上有老下有小的普通居民，更有部分居民直斥“A 市地铁是跟钱走”…,,,	0	4

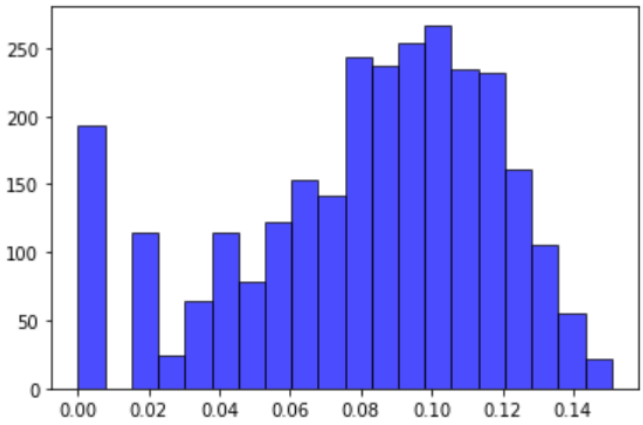


图表 9 聚类图

3.3 第三题结果

完整性指标

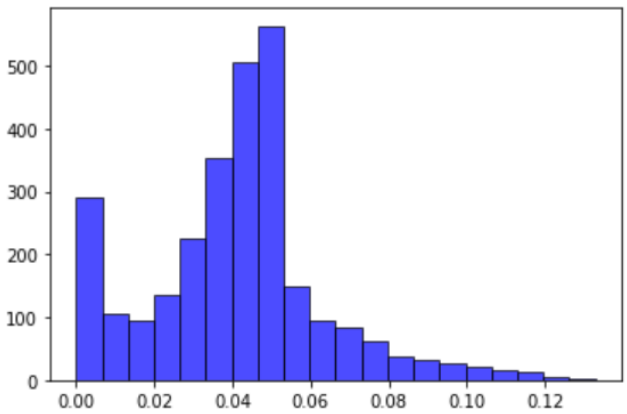
对问题答复的完整度越高即正确率越高，首先根据结果观察答复的完整性，我们从完整性的分布可知，65%答复比较完整，25%的答复相对较为完整，不过也有一部分正确率较低，而正确率最低的答复相对集中。可见相关部门再答复问题的时候答复完整性较强，但部分答复出现答非所问的情况。如图 10。



图表 10 答复完整性

可解释性指标

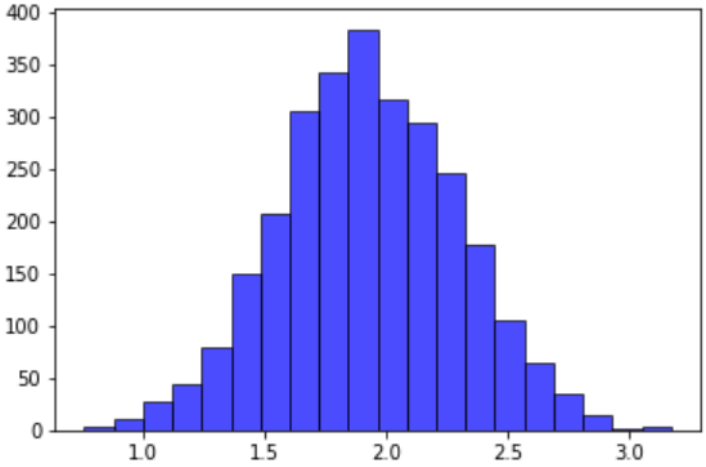
解释性定义为对问题的答复详细程度，如图 11 可见，大多数答复意见在图 10 给出完整性后均有所下降，答复的解释性稍有下降；而原本低完整性答复的解释性变化不大，可见在答复问题上虽然答复比较完整，但详细程度不足；而在完整性低的答复中解释内容却比较高，虽然答复问题不全，但就已答复的问题中答复比较详细。



图表 11 答复可解释性

相似性指标

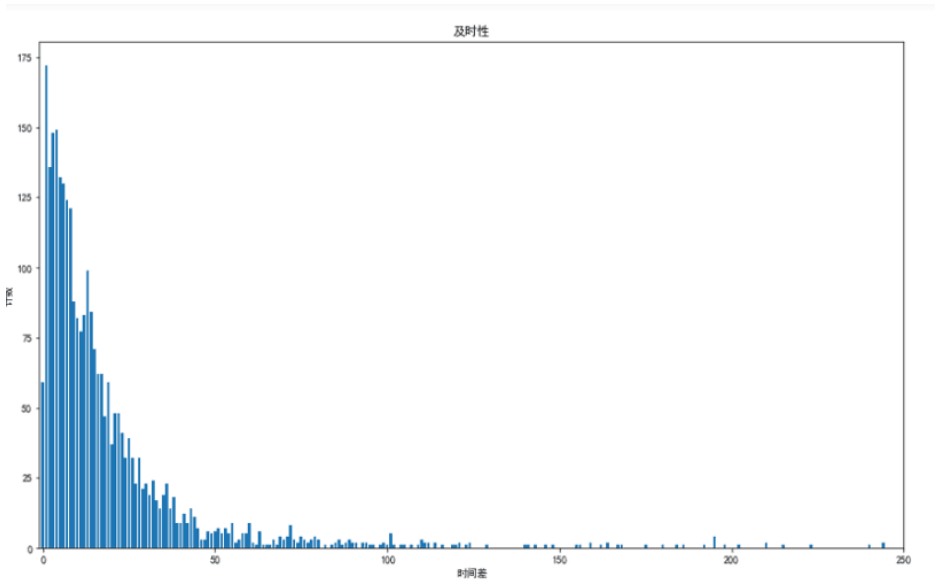
观察留言详情与答复意见的相似性，如图 12,留言详情与答复意见相似度近似正态分布，根据2σ原则可知，有95%的留言与答复相似度相近且较高。



图表 12 答复相似性

及时性指标

如图 13，可见对留言的答复时间间隔呈现递减趋势，符合常理，一般部门回复存在周期性，在周期之内的问题一般能得到答复，有95.3%的留言在两个月内供给与了答复。有41.2%留言在两周内给予回复。只有少数问题可能复杂，存在两个月以上最长长达三年之久才给予回复。可见回复的及时性比较好，部分问题有待提高效率。分布图如图 13。



图表 13 答复及时性

每一条留言对应的完整性指标、可解释性指标、相似性指标、及时性指标具体明细参见附录。部分结果如表 4：

表格 4

留言用户	留言主题	答复意见	完整性	可解释性	相似性	时间间隔
A00045581	A2 区景蓉华苑物业管理有问题..	现将网友在平台《问政西地省》栏目向胡华衡书记留言……	0.795005	0.396666	0.459401	15
A00023583	A3 区潇楚南路洋湖…	您好！针对您反映 A3 区……	0.575545	0.308317	0.487549	15
A00031618	请加快提高 A 市民营幼儿园….	市民同志：你好！您反映的“请加快提高民营幼儿园教师……	0.575545	0.308317	0.530605	15
A000110735	在 A 市买公寓……	您好！您在平台《问政西地省》……	0.834526	0.678398	0.525457	15

4. 参考文献

[1] 杨丽,循环神经网络研究综述[J],计算机科学, 2018, S2

[2] 牛伟农,一种基于词聚类信息熵的新闻提取方法[J],软件导刊,2020,01

[3] GitHub, Keras:基于 Python 的深度学习库[OL], https://keras.io/zh/, 2020

[4] 刘晓坤,Chinese Word Vector[OL], https://www.jiqizhixin.com/articles/2018-05-15-10

[5] Shen Li, Zhe Zhao, Renfen Hu, Analogical Reasoning on Chinese Morphological and Semantic Relations [M] Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,2018,138-143