C 题

摘　要：在各类社情民意相关的文本数据量不断攀升，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不便利的背景下。本文围绕政务平台上群众留言的分流、热点问题的筛选、答复意见的质量评估，以统计机器学习方法、深度学习方法为理论依据。使用基于 python 的 paddlepaddle、ernie、sklearn 等模块作为主要工具。建立了群众留言分类模型、热点问题的推送模型、答复意见的评估模型。

针对问题一，对文本类型数据进行分词、去标点后映射到 embedding；对交通运输等数据量少的标签稍加过采样后对短文本主题数据使用 CNN 网络来进行模型的训练；对长文本类型的留言详情数据使用 LSTM 网络来进行模型的训练。

针对问题二，为提取出热点问题设计了三种粗略的模型最后将各个模型的结果加权相加得到热度的评价指标。方案一，对 embedding 后的数据进行特征融合再暴力求解各样本间的 cos 相似度取大于 0.65 的个数；方案二，TFIDF 后 ernie 获取留言数据的句子粒度特征，使用 kmeans 聚类以二级标签总个数为簇中心取各簇中样本个数；方案三，对留言数据分词后去标点去停用词使用 TF-IDF 提取特征后以 LDA 进行按周按月聚类，取各类中样本的个数；最后综合点赞数与反对数对各样本的各模型得到的指标加权求和即为热度评价的指标。找到人群地点信息在留言主题上的规律利用正则表达式对人群地点信息进行提取。

针对问题三，相关性指标：对留言详情与留言回复特征处理后做相似度计算，以计算出的相似度作为相关性指标。完整性指标：对留言详情与留言回复分别进行 TF-IDF 以及 LDA 主题提取，对提取出能代表主题的前三个词汇交叉进行相似度计算，相似度大于 0.8 以上即满足留言详情的这一特征词，记录每条回复的主题词与留言详情主题词相似度大于 0.8 的个数与提取主题词的个数做比值，当作完整性指标评价。可解释性指标，利用 N-gram 来鉴别语句是否通顺，并给出评价。

关键词：LDA　TF-IDF　ernie　kmeans　embedding　LSTM　CNN 相似度

The thesis title

Abstract: In the context of the increasing amount of text data related to social situation and public opinion, it is inconvenient for the relevant departments to manually divide messages and sort out hot spots. Based on the theory of statistical machine learning and in-depth learning, this paper focuses on the diversion of public comments, the selection of hot issues, and the quality evaluation of replies on the government platform. Paddlepaddle, Ernie and sklearn modules based on Python are used as the main tools. This paper establishes the classification model of mass message, the push model of hot issues, and the evaluation model of reply opinions.

To solve the first problem, the text type data is segmented, de punctuated and mapped to embedding; the short text subject data is trained by CNN network after a little over sampling for tags with small amount of data such as transportation; the long text type message detail data is trained by LSTM network.

To solve the second problem, three rough models are designed to extract the hot spots. Finally, the results of each model are weighted together to get the evaluation index of heat degree. In scheme 1, feature fusion is carried out for embedded data, and then the number of COS similarity between samples is more than 0.65; in scheme 2, Ernie after TFIDF obtains sentence granularity feature of message data, and kmeans clustering is used to take the number of samples in each cluster with the total number of secondary tags as the cluster center; in scheme 3, after segmentation of message data, punctuation is removed to stop words, TF-IDF is used to extract features The LDA was used to cluster every week and every month, and the number of samples in each category was taken. Finally, the weighted sum of the indexes obtained by combining the likes and objections of each sample model was the index of heat evaluation.

For the third problem, correlation index: after processing the message details and message reply features, we do similarity calculation, and take the calculated similarity as the correlation index. Integrity index: TF-IDF and LDA subject extraction are carried out for message details and message replies respectively, and similarity calculation is carried out for the first three words that can represent the subject extracted. If the similarity is more than 0.8, then the feature word satisfying message details is met. The number of main title words and message details subject words with similarity greater than 0.8 and the number of extracted subject words are recorded as the ratio As an indicator of integrity. Interpretability index, using n-gram to identify whether the sentence is smooth, and give an evaluation.

Key words: LDA TF-IDF ernie kmeans embedding LSTM CNN Similarity

目　录

1. 挖掘目标

第一点对留言按照一定的划分体系进行分类，以便后续将群众留言分派至相应的职能部门处理。

第二点某一时段内群众集中反映的某一问题可称为热点问题。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，给出评价结果。

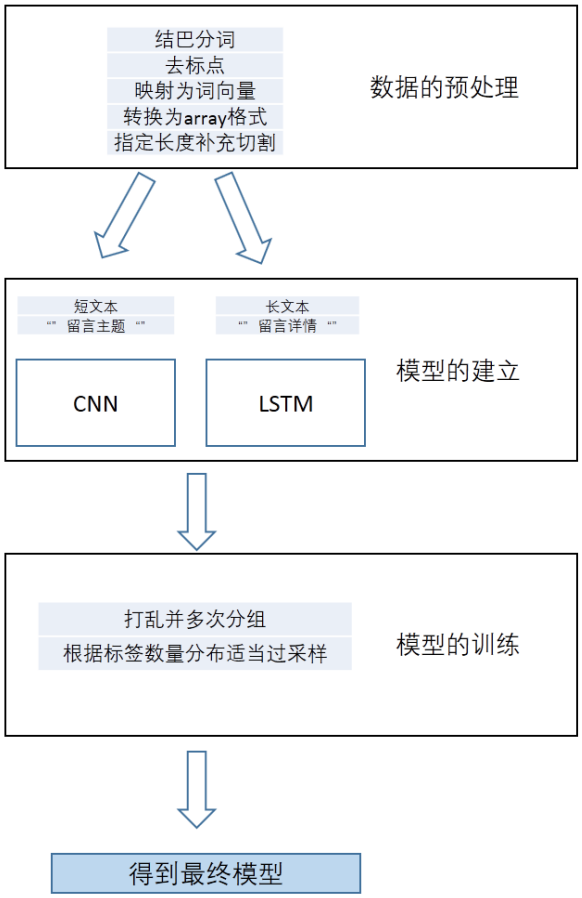
第三点对相关部门的留言答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

2. 分析方法与过程

2.1. 总体流程

2.1.1. 对问题一的分析过程

问题一需要对留言按照一定的划分体系进行分类，以便后续将群众留言分派至相应的职能部门处理。根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。



2.1.2. 对问题二的分析过程

某一时段内群众集中反映的某一问题可称为热点问题。问题二根据附件 3 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，给出评价结果。

在对问题二进行建模之前首先需要对问题二的数据做充分的了解。

A	B	C	D	E	F	G
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
188006	A000102948	阳光婚纱艺术摄影是否合法	2019/2/28 11:25:05	，因为地处居民楼内	0	0
188007	A00074795	路命名规划初步成果公示和城	2019/2/14 20:00:00	0年都未曾更换过，	0	1
188031	A00040066	华镇金鼎村水泥路、自来水	2019/7/19 18:19:54	，且天还没黑就开	0	1
188039	A00081379	路步行街大古道巷住户卫生间	2019/8/19 11:48:23	行清扫。没有解决	0	1
188059	A00028571	际社区三期与四期中间空地夜	2019/11/22 16:54:42	投诉业主，态度强	0	0
188073	A909164	单方面改变麓谷明珠小区6栋	2019/3/11 11:40:42	何政府调规、改建的	0	0
188074	A909092	区富绿新村房产的性质是什么	2019/1/31 20:17:32	让给业主了，然而因	0	0
188119	A00035029	对A市地铁违规用工问题的质	2019/5/27 16:04:44	加班还扣钱，扣身份	0	0
188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	该司机并未按地面车	0	0
188249	A00084085	桐梓坡路与麓松路交汇处地	2019/9/17 4:25:00	边邻居也是苦不堪言	0	0
188251	A00013092	东四路口晚高峰太堵，建议调	2019/10/19 11:02:40	下至少两到三个信号	0	0
188260	A00053484	小区乐果果零食炒货公共通道	2019/5/31 17:06:13	零食炒货公共通道摆	0	0
188396	A00047580	楚在西地省商学院宿舍旁安	2019/4/15 16:23:09	《中小学校校园环境	2	1

附件 3 共计 7 列数据。要求对某一时间段内群众集中反映的问题进行归类，并划分为热点问题。附件 3 数据中的列“留言编号”，“留言主题”，“留言时间”，“留言详情”都在不同程度上影响着对某一时间段内热点问题的聚类。而点赞数与反对数，其本身就是一个衡量热度与推送程度的指标。

	数据处理阶段	模型	指标	自定义权重	指标
方案一	对 embedding 后的数据进行特征融合	暴力求解各样本间的 cos 相似度	相似度大于 0.65 的个数	*20%	求和
方案二	TFIDF 后 ernie 获取留言数据的句子粒度特征	kmeans 聚类以二级标签总个数为簇中心	各簇中样本个数	*20%	
方案三	分词后去标点去停用词	TF-IDF 后 LDA 进行按周按月聚类	各类中样本的个数	*30%	
指标四	无	无	点赞数与反对数	*30	

对于时间范围的提取，以 LDA 模型的聚类为主。提取某一类的最早时间与最晚时间。

对于地点人群信息，利用正则表达式观察文本中存在的规律，提取进行提取。

2.1.3. 对问题三的分析过程

问题三需要对附件 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。

针对留言的答复意见质量的评价方案，重点在于提炼 user 的留言详情与留言主题中的关键词以及 admin 的留言答复中的关键词。

相关性指标：对留言详情与留言回复特征处理后做相似度计算，以计算出的相似度作为相关性指标。

完整性指标：对留言详情与留言回复分别进行 TF-IDF 以及 LDA 主题提取，对提取出能代表主题的前三个词汇交叉进行相似度计算，相似度大于 0.8 以上即满足留言详情的这一特征词，记录每条回复的主题词与留言详情主题词相似度大于 0.8 的个数与提取主题词的个数做比值，当作完整性指标评价。

可解释性指标，利用 N-gram 来鉴别语句是否通顺，并给出评价。

2.2. 具体步骤

2.2.1. 问题一的具体步骤

（1） 数据的理解

在进行留言分类之前首先要对待分类的数据做充分的了解，观察附件 2 表的结构：

A	B	C	D	E	F
留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A00074011	建筑集团占道施工有安	20/1/6 12:09:	围墙内。每天尤其	城乡建设
37	U0008473	大厦人为烂尾多年，安	20/1/4 11:17:	着，不但占用人行道	城乡建设
83	A00063999	市A1区苑物业违规收停	9/12/30 17:06	多次向物业和社区	城乡建设
303	U0007137	南路A2区华庭楼顶水箱	19/12/6 14:40:	，霉是一种强致癌	城乡建设
319	U0007137	2区华庭自来水好大一股	19/12/5 11:17:	，霉是一种强致癌	城乡建设
379	A00016773	市盛世耀凯小区物业无	19/11/28 9:08:	业不是为业主服务的	城乡建设
382	U0005806	询A市楼盘集中供暖一	9/11/27 17:14	月亮岛片区近年规划	城乡建设
445	A00019209	西路可可小城长期停水	9/11/19 22:39	帮助至今没有找到	城乡建设
476	U0003167	取城市垃圾处理费不	9/11/15 11:44	在的物业公司也未给	城乡建设
530	U0008488	A3区魏家坡小区脏乱差	9/11/10 18:59	让人好好休息一下	城乡建设
532	U0008488	A市魏家坡小区脏乱差	9/11/10 12:30	让人好好休息一下	城乡建设
673	A00080647	四届非法业委会涉嫌侵	9/10/24 11:29	责令B4区有关部门拆	城乡建设
994	U0005196	梅溪湖壹号御湾业主用	19/9/18 22:43:	别的城市都已经一门	城乡建设
1005	U0006509	翡翠湾强行对入住的业	19/9/18 13:36:	地产公司和金晖物业	城乡建设
1110	A00099772	市锦楚国际星城小区三	19/9/9 11:07:	是无通知，突然断电	城乡建设
1309	U0005083	和紫郡用电的问题能不	19/8/21 15:12:	之后，我们的用电	城乡建设
1440	A0003288	际新城从6月份开始停电	19/8/6 10:28:	的生活，而且我们	城乡建设

1）附件 2 中共计 6 列数据，问题一要求根据一级标签对留言数据进行分类。根据直观判断影响所属类别的变量是“留言主题”或“留言详情”列。

2）首先观察留言主题与留言详情的文本长度：

在对变量“留言详情”进行分词后，留言详情数据的平均长度为 554.379

而对变量“留言主题”进行分词后，留言主题数据的平均长度为 28.946

（2） 数据预处理

其中，留言主题或详情列的数据都是以中文文本格式作存储展示的。

以往所学习的机器学习算法无法直接对文本形式的数据进行分类，需要将这样的文本类数据转换成便于运算数值类型的数据。

考虑让数字代表词汇来带入计算，把每一个词汇替换成一个数字。完成这一步的前提需要对每条留言语句的词汇进行拆分，把独立的单词找出来形成一个词汇表。

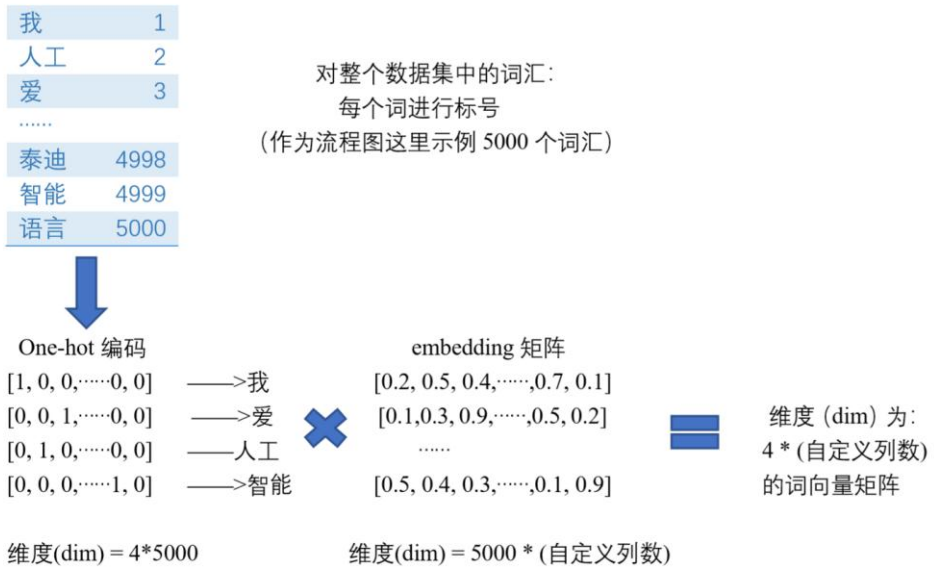
得到词汇表后，把给每一个词汇赋予一个数字进行表示。

然而单纯使用数字对词汇进行表示，无法表达词汇的重要性，数值大的词汇编号总是占取较大的权重影响整体的分类结果。

对使用数字进行编码的词汇进一步处理。

将数字替换为一个 one-hot 编码的形式。

最后将 one-hot 编码乘以 Embedding Matrix 得到对应的词向量



在将留言详情与留言主题处理成 word-Embedding 后，考虑建立什么样的模型对留言数据进行分类。

（3） 模型的选择

从数据的理解过程中，得知“留言详情”变量中的数据为长文本数据，“留言主题”中的数据为短文本数据。

对于长文本数据，使用普通的神经网络进行训练学习存在一个明显的缺陷，就是当阅读很长的序列时，网络内部的信息会变得越来越复杂，甚至会超过网络的记忆能力，使得最终的输出信息变得混乱无用。长短时记忆网络（Long Short-Term Memory，LSTM）内部的复杂结构正是为处理这类

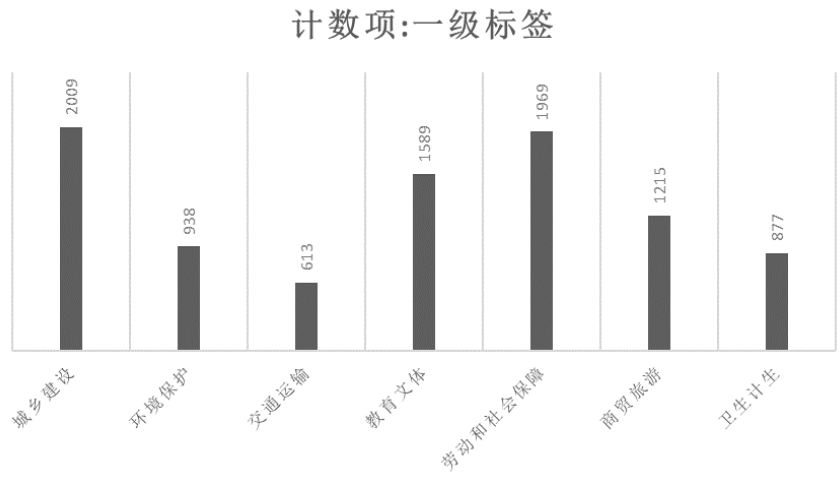
问题而设计的。因此对于变量“留言主题”进行分类时选用 LSTM 网络进行训练。

而对于短文本卷积神经网络通常比 LSTM 的效果要好一些，对于“留言主题”数据使用 CNN 来进行分类。

（4） 模型的修正

在使用深度学习来进行类别训练的时候，不同量级数据标签分布会影响整个模型的学习，模型往往会给数据量大的一类标签赋以更高的权重。因此在进行训练之前我们还需要观察一下各类别数据量的一个分布情况。

对数据量较少的交通运输类稍加过采样以保持模型对各类学习程度的平衡。



2.2.2. 问题二的具体步骤

(1) 数据的理解

在对问题二进行建模之前首先需要对问题二的数据做充分的了解。

A	B	C	D	E	F	G
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
188006	A000102948	阳光婚纱艺术摄影是否合法	2019/2/28 11:25:05	，因为地处居民楼内	0	0
188007	A00074795	路命名规划初步成果公示和城	2019/2/14 20:00:00	0年都未曾更换过，	0	1
188031	A00040066	华镇金鼎村水泥路、自来水	2019/7/19 18:19:54	，且天还没黑就开	0	1
188039	A00081379	路步行街大古道巷住户卫生间	2019/8/19 11:48:23	行清扫。没有解决	0	1
188059	A00028571	际社区三期与四期中间空地夜	2019/11/22 16:54:42	投诉业主，态度强	0	0
188073	A909164	单方面改变麓谷明珠小区6栋	2019/3/11 11:40:42	何政府调规、改建的	0	0
188074	A909092	区富绿新村房产的性质是什么	2019/1/31 20:17:32	让给业主了，然而因	0	0
188119	A00035029	对A市地铁违规用工问题的质	2019/5/27 16:04:44	加班还扣钱，扣身份	0	0
188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	该司机并未按地面车	0	0
188249	A00084085	桐梓坡路与麓松路交汇处地	2019/9/17 4:25:00	边邻居也是苦不堪言	0	0
188251	A00013092	东四路口晚高峰太堵，建议调	2019/10/19 11:02:40	下至少两到三个信号	0	0
188260	A00053484	小区乐果果零食炒货公共通道	2019/5/31 17:06:13	零食炒货公共通道摆	0	0
188396	A00047580	楚在西地省商学院宿舍旁安	2019/4/15 16:23:09	《中小学校校园环境	2	1

附件 3 共计 7 列数据。要求对某一时间段内群众集中反映的问题进行归类，并划分为热点问题。

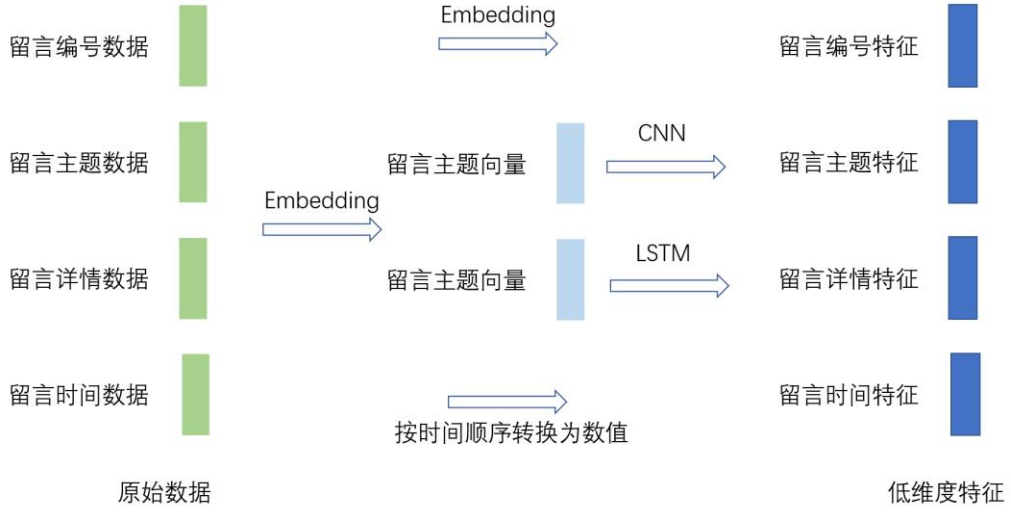
附件 3 数据中的列“留言编号”,“留言主题”,“留言时间”,“留言详情”都在不同程度上影响着对某一时间段内热点问题的聚类。而点赞数与反对数，其本身就是一个衡量热度与推送程度的指标。

(2) 方案设计

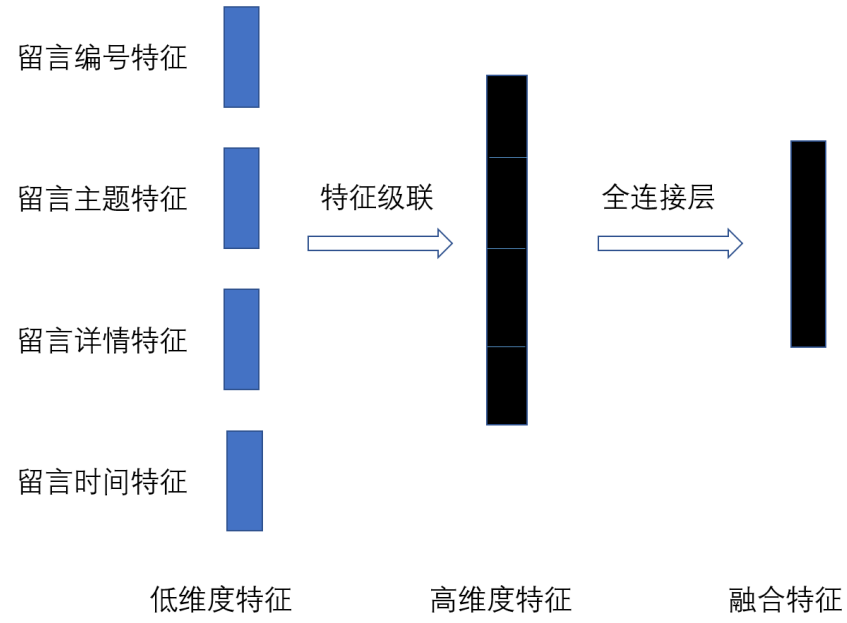
为提炼留言数据热度评价指标，这里提出从多方面进行思考，最后对结论进行融合的方式来提炼热度指标。

1） 方案一

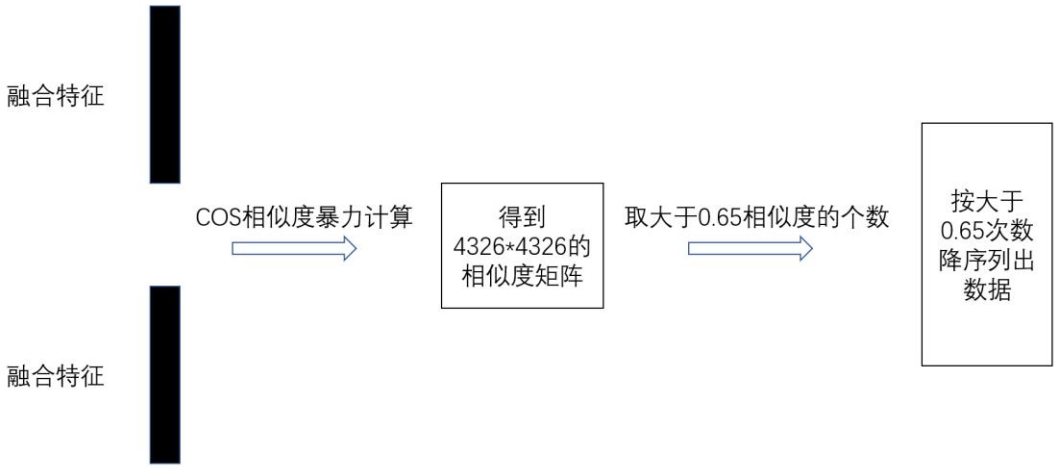
对留言数据的多个变量进行分词，去停用词等一系列处理后得到处理后各个变量处理后的词向量特征。



得到各变量的词向量特征后，采用特征级联的方式对各变量的特征融合成一个高维度的特征，再经过一个全连接层形成融合后的样本特征。



对样本的融合特征本身进行暴力的 cos 相似度计算，让一个样本与其他每一个样本做一次 cos 相似度计算，得到一个相似度矩阵。取相似度大于 0.65 的数据作为衡量热度指标的一个计数。

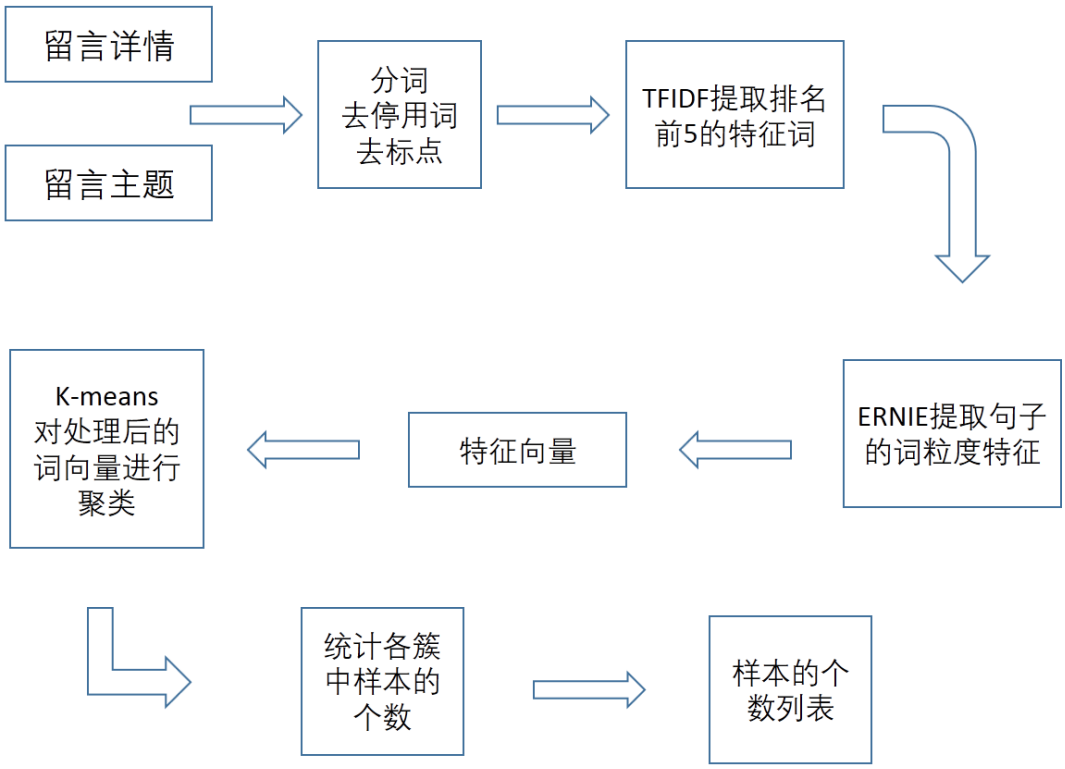


最终，得到各个样本的一个计数列表，作为评价热度指标的依据之一。

2） 方案二

对变量“留言详情”,“留言主题”使用 tf-idf 提取出每条留言中词汇的 tfidf 值排名前五的词汇。

直接使用 ernie 获取输入文本留言主题数据与留言详情数据的句子粒度特征。对得到的句子粒度特征向量使用 kmeans 进行聚类操作，统计被分到各类的词语的个数。



最终，得到各个样本的一个计数列表，作为评价热度指标的依据之一。

3） 方案三

核心是利用 LDA 主题模型进行主题获取，主题即是热点。

根据附件 1 给出的类别信息，二类标签共计 118 个三类标签共计 517 个，取中间位置的 300 作为 LDA 聚类的簇中心个数，对留言数据进行聚类。

记录没类留言数据的起止时间，对每类留言数据进行 LDA 主题提取，获取他的主要问题。

以 LDA 类中的个数作为排序依据。

4） 方案融合

要维护出一个地点、和人群的一个数据字典，在做特征抽取的时候。将地点和人群信息单独抽取出来。综合方案一、二、三的热度指标评价依据并结合点赞数对热点问题进行排序。

其中内容的抽取以 LDA 方法的特征提取为主。

2.2.3. 问题三的具体步骤

针对留言的答复意见质量的评价方案，重点在于提炼 user 的留言详情与留言主题中的关键词以及 admin 的留言答复中的关键词。

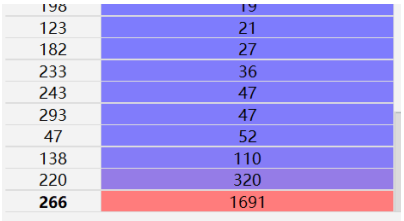
相关性指标：对留言详情与留言回复特征处理后做相似度计算，以计算出的相似度作为相关性指标。

完整性指标：对留言详情与留言回复分别进行 TF-IDF 以及 LDA 主题提取，对提取出能代表主题的前三个词汇交叉进行相似度计算，相似度大于 0.8 以上即满足留言详情的这一特征词，记录每条回复的主题词与留言详情主题词相似度大于 0.8 的个数与提取主题词的个数做比值，当作完整性指标评价。

可解释性指标，利用 N-gram 来鉴别语句是否通顺，并给出评价。

2.3. 结果分析

使用 LDA 对附件 3 留言详情数据进行聚类时按附件 1 的类别数据折中取 300 进行聚类，对各组内的数量进行统计：



198	19
123	21
182	27
233	36
243	47
293	47
47	52
138	110
220	320
266	1691

其中第 266 组内的数目过多需要进一步聚类。观察数据较多的其他几组关键词情况：

topic 220	topic 138	topic 266
开发商	噪音	小区
请问	不能	我们
本人	a1	没有
办理	扰民	业主
我们	休息	居民
学生	社区	领导
谢谢	住户	问题
情况	城管	部门
合同	单位	一个
为什么	街道	严重
孩子	凌晨	希望
业主	垃圾	相关
什么	是否	解决
自己	a2	影响
项目	执法	学校
拆迁	其他	政府
购房	规定	现在
还有	电梯	施工
交房	反映	
安置		

调整聚类中心数目，多次进行查看发现，“物业”关键词总是能够占据待分类的第一大类上，噪

音问题、为学生考虑的购房问题是共鸣度较高的两大类问题。

3. 完成情况

参与竞赛时间较晚，时间有限未完整的完成整个模型的建立。论文中的模型大部分在 AIstudio 平台上线上进行运行实现。

完成情况：问题一分类模型 LSTM 网络及 CNN 网络的建立

问题二特征级联模型的建立及相似度求解

问题二 ernie 提去句子粒度特征并应用 kmeans 进行聚类

问题二 LDA 方法进行聚类及各类别主题词的提取，各类起止时间的提取

问题三 LDA 方法进行主题词的提取，两变量对应相似度的计算

4. 参考文献

[1]paddlepaddle 官方文档