基于 BERT 模型的文本挖掘应用

摘　要：近年来，随着网络问政平台逐步成为政府了解民意的重要渠道，因此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。

本文主要通过使用 BERT 的预训练模型从语料库生成的任务来学习建模句子之间的关系，从而完成文本分类等下游任务。通过 BERT 产生的句向量具有深度语义特性，能做到无关键词相同的情况下匹配同一事件的留言，进而结合 DBSCAN 聚类算法进行相似文本聚类，最后建立热度评价模型，提取出热点问题，从而完成热点问题挖掘的任务。通过提取问题与答复的关键词，在传统余弦度量法基础上进行改进，将问题与答复相关性等做出评价，建立一套完备的答复评价方案。

对于问题 1，首先将附件 2 中的数据进行预处理，去掉 HTML 字符、删除 URL 等，再使用 fast.ai 深度学习框架，使用 BERT 的预训练模型，训练处理好的数据，得到更适合该问题数据中的模型。

对于问题 2，将附件 3 中的数据进行预处理，运用 BERT 的预训练模型将文本的留言主题转化为句向量，文本间的相似度用余弦距离，然后用 DBSCAN 聚类算法对留言的句向量进行聚类，最后设计热点评价指标并建立相应的热度评价模型提取出热点问题。

对于问题 3，针对留言的答复意见，提取留言回复信息质量优劣的特征，研究如何从答复的相关性、完整性、可解释性等角度对回复信息的质量进行量化描述，建立衡量回复信息质量等方面的综合评价模型。

关键词：深度学习，NLP，迁移学习，BERT，中文分词，DBSCAN 聚类，Single-pass 聚类，余弦度量

Text mining application based on BERT model

Abstract: In recent years, as the online political inquiry platform has gradually become an important channel for the government to understand public opinion, the establishment of a smart government system based on natural language processing technology has become a new trend in the development of social governance innovation, which is extremely important for improving the management level and efficiency of government A big boost.

This article mainly learns the relationship between modeled sentences by using BERT's pre-training model to generate tasks from a corpus, thus completing downstream tasks such as text classification. The sentence vectors generated by BERT have deep semantic characteristics, which can match the messages of the same event without the same keywords, and then use the DBSCAN clustering algorithm to cluster similar texts, and finally establish a heat evaluation model to extract hot issues In order to complete the task of mining hot issues. By extracting the keywords of questions and answers, and improving on the basis of the traditional cosine metric method, the relevance of questions and answers is evaluated, and a complete set of answer evaluation plan is established.

Aiming as the problem of the first, process the data in Annex 2 to remove spaces, use the fastai deep learning framework, use the Transfer Learning method, import the pre-trained BERT model, train the processed data, and get the more suitable model for the problem.

Aiming as the problem of the second, process the data in Annex 3, use BERT's pre-training model to convert the subject of the text message into a sentence vector, use the cosine distance for the similarity between the text, and then use the DBSCAN clustering algorithm to the sentence vector of the message Perform clustering, and finally design hot spot evaluation indicators and establish corresponding hot evaluation models to extract hot spots.

Aiming as the problem of the third, extract the characteristics of the quality of the message response

information in response to the comments of the message, study how to quantitatively describe the quality of the response information from the perspective of relevance, completeness, and interpretability of the response, and establish a measure to measure the quality of the response information Comprehensive evaluation model of other aspects.

Key words: Deep Learning，NLP，Transfer Learning，BERT，Chinese Participle，DBSCAN clustering，Single-Pass clustering, Cosine metric

目录

1. 挖掘目标

本次建模目标是利用互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，采用深度学习、数据挖掘等技术，利用 BERT 模型，Hanlp 中文分词工具，达到以下三个目标：

（1）利用迁移学习技术，基于附件 2 的数据建立 BERT 模型，并期望在测试集达到较高的分类效果。

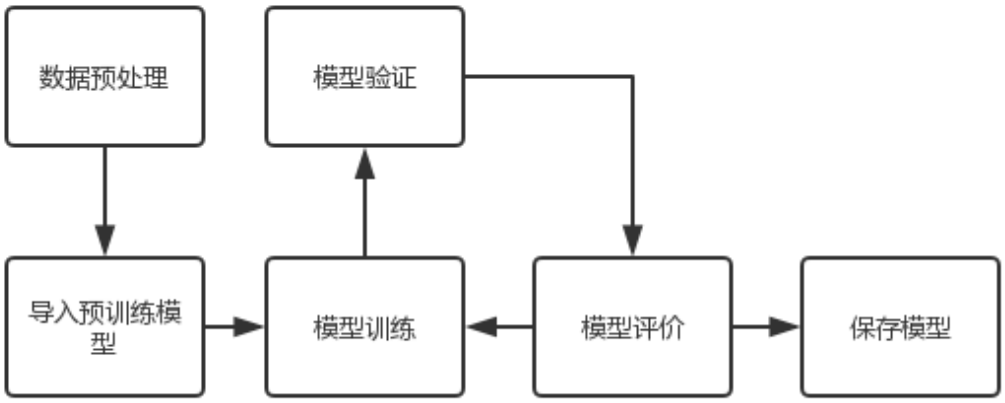
（2）使用 BERT 模型生成句向量，并对句向量聚类，根据聚类后类别中的样本数量、时间、点赞反对数来定义合理的热度评价指标，并得到评价结果。

（3）针对留言的答复意见，提取留言回复信息质量优劣的特征，研究如何从答复的相关性、完整性、可解释性等角度对回复信息的质量进行量化描述，建立衡量回复信息质量等方面的综合评价模型。

2. 分析方法与过程

2.1. 问题 1 分析方法与过程

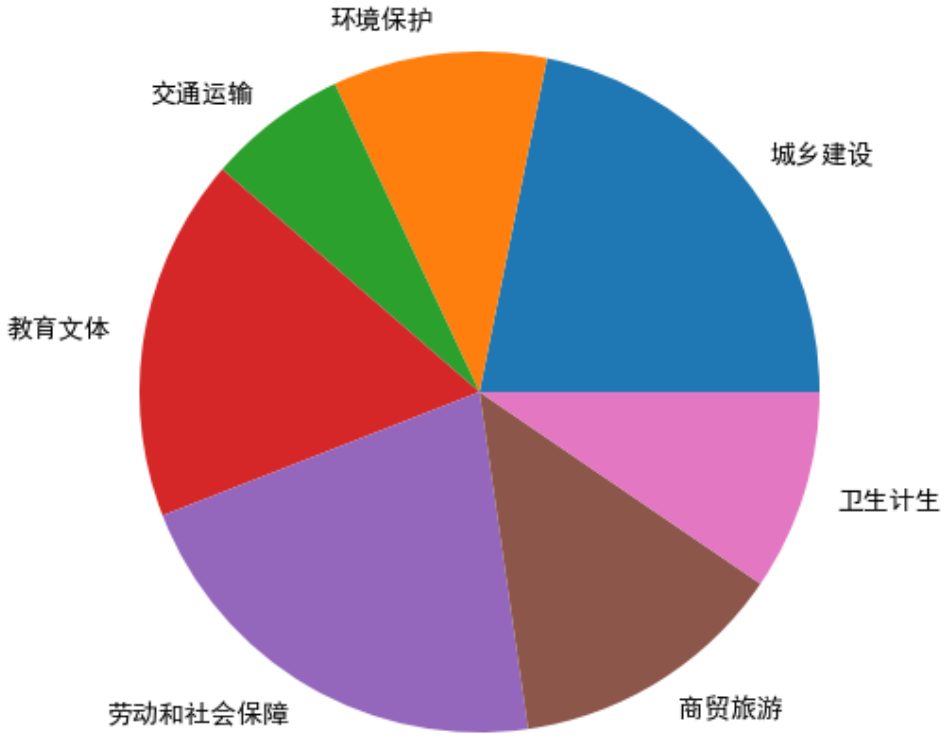
2.1.1. 流程图



2.1.2. 数据预处理

题目给出的附件 2 有用的信息是留言主题、留言详情以及一级标签，需要根据留言主题与留言详情的内容将所有数据样本按照一级标签进行分类，首先使用 Python 的 Pandas 库将数据导入，将留言主题与留言详情两栏提取并合并，由于数据中有很多非中文字符，如“\t”“\n”等，所以需要进行数据清洗，将这些字符看做空格即可，在 Python 中很容易实现这个操作。

首先统计一下标签的分布情况。



接下来将数据中的标签“城乡建设”、“劳动和社会保障”、“教育文体”等 7 个中文标签全部转换成计算机能够识别的数字 0,1,2,...,6，便于后续分类模型的训练。

上述过程处理的 python 程序见附件 preprocess.py。完成后将处理好数据保存到 preprocessed_data.csv。

可以看到处理过后的数据如下图，文本中已经没有 URL 等噪声，并且标签也已经转换为数字。

	Unnamed: 0	comment	label
0	0	A市西湖建筑集团占道施工有安全隐患A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯...	0
1	1	A市在水一方大厦人为烂尾多年，安全隐患严重位于书院路主干道的在水一方大厦一楼至四楼人为拆除水...	0
2	2	投诉A市A1区苑物业违规收停车费尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物...	0
3	3	A1区蔡锷南路A2区华庭楼顶水箱长年不洗A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗...	0
4	4	A1区A2区华庭自来水好大一股霉味A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在...	0
...	...	...	...
9205	9205	两孩子一个是一级脑瘫，能再生育吗？我们夫妻都是农村户口，大的是女9岁，小的是儿2岁半，才15...	5
9206	9206	B市中心医院医生不负责任，做无痛人流手术后结果还是活胚芽本人2015年2月16号在B市中心医...	5
9207	9207	西地省二胎产假新政策何时出台？我们是再婚，很想再要一个小孩，不知我省二胎新政策何时出，如果先...	5
9208	9208	K8县惊现奇葩证明！K8县惊现奇葩证明！我是西地省K8县人，想生二孩。被告知要开证明，即“没...	5
9209	9209	请问J4县卫计委社会抚养费到底该交多少钱？领导你好，我们属于未婚生子，但是在2013年已经接...	5

9210 rows × 3 columns

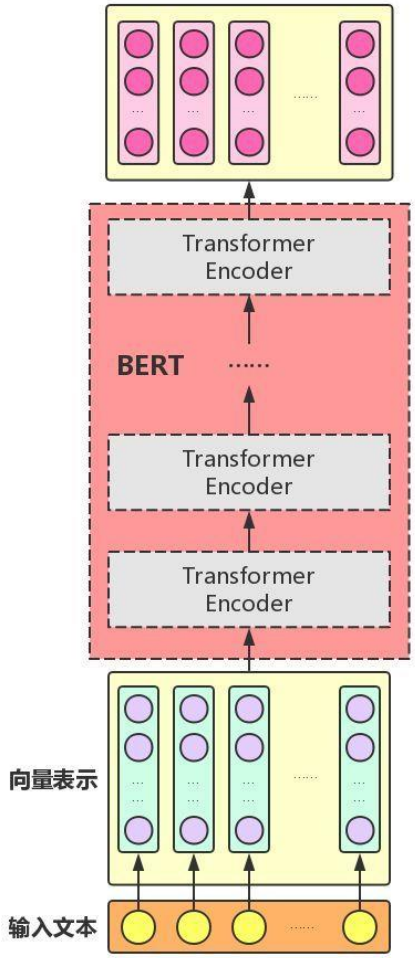
2.1.3. 导入预训练模型

问题 1 是一道文本分类的题目，近年来解决此类问题最好的模型莫过于 BERT 模型，BERT 模型是谷歌于 2018 年提出的 NLP 模型，从下图这个例子可以看出，BERT 在分类领域的效果超过了 XGBoost 分类器以及卷积循环神经网络。

BERT模型与三种对比方法的正面、负面、中立情感分类F1值如下：

方法	正面F1值	负面F1值	中立F1值
XGBoost	67%	60%	91%
Char-level CNN	69%	74%	92%
Attention-based RNN	66%	71%	91%
BERT	71%	76%	92%

下图简单介绍了一下 BERT 模型的架构，BERT 模型通过查询字向量表将文本中的每个字转换为一维向量，作为模型输入；模型输出则是输入各字对应的融合全文语义信息后的向量表示。

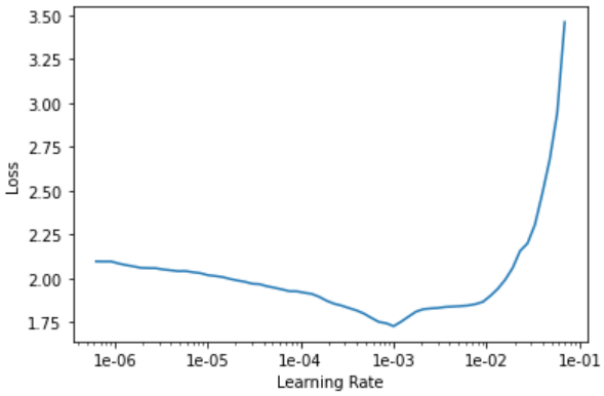


为了减少工作量，直接导入谷歌发布的预训练模型，问题 1 是中文文本分类，所以在具体模型选择上选择"bert-base-chinese"。

关于 BERT 模型具体架构，见文件“BERT 架构.png”。

2.1.4. 模型训练、模型验证、模型评价

在机器学习任务中，通常将数据集切分成训练集，测试集和验证集，这样有利于提高模型的泛化能力，防止过拟合。在训练的时候，为了寻找最优的学习速率，绘制学习率与 Loss 之间关系图，通常选取 Loss 下降最快时对应的学习率，通过图像观察并结合实际测试，最终选择的学习率大小为 2e-5。



训练过程如下：

epoch	train_loss	valid_loss	accuracy	time
0	1.531257	1.190692	0.664858	01:21
1	0.489912	0.333418	0.902307	01:19
2	0.267390	0.228459	0.932157	01:18
3	0.172858	0.240687	0.928087	01:19
4	0.091345	0.253052	0.934193	01:19
5	0.051785	0.258684	0.939620	01:20
6	0.021891	0.308486	0.930801	01:21
7	0.010994	0.291843	0.939620	01:21
8	0.008500	0.300818	0.939620	01:19
9	0.006544	0.302195	0.940977	01:21

超参数 epoch 设置为 10，迭代 10 轮，在第 8 轮的时候模型收敛，上图给出了准确率，下图使用 sklearn 中的 metrics.classification_report()方法，来计算精确率、召回率以及 F1-Score。可以看出模型在测试集上的表现较为理想。

	precision	recall	f1-score	support
0	0.93	0.90	0.91	413
1	0.95	0.95	0.95	369
2	0.95	0.96	0.95	307
3	0.87	0.92	0.90	234
4	0.94	0.94	0.94	210
5	0.93	0.93	0.93	176
6	0.93	0.86	0.89	133
accuracy			0.93	1842
macro avg	0.93	0.92	0.93	1842
weighted avg	0.93	0.93	0.93	1842

将训练好的模型保存为 trained_model.pth，以便下次测试数据时直接导入训练好的模型即可。

2.1.5. K 折交叉验证

交叉验证（Cross validation)，交叉验证用于防止模型过于复杂而引起的过拟合.有时亦称循环估计， 是一种统计学上将数据样本切割成较小子集的实用方法。于是可以先在一个子集上做分析， 而其它子集则用来做后续对此分析的确认及验证。 一开始的子集被称为训练集。而其它的子集则被称为验证集或测试集。交叉验证是一种评估统计分析、机器学习算法对独立于训练数据的数据集的泛化能力[1]。

k-folder cross-validation:k 个子集，每个子集均做一次测试集，其余的作为训练集。交叉验证重复 k 次，每次选择一个子集作为测试集，并将 k 次的平均交叉验证识别正确率作为结果。

我们采用 5 折交叉验证，将数据分成 5 份，轮流将其中 4 份作为训练集，另外 1 份作为测试集，进行实验。5 次的结果的正确率（或差错率）的平均值作为对算法精度的估计。

2.1.6. 模型融合

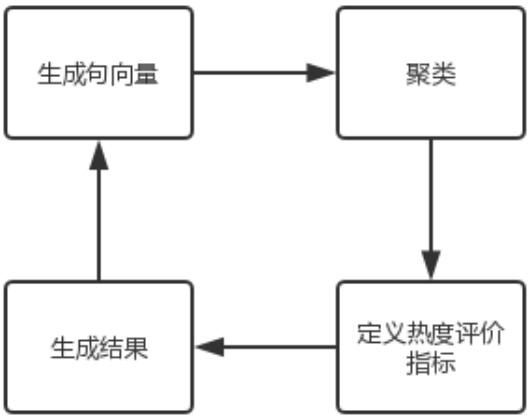
使用的模型有 BERT、BERT-wwm、RoBERTa，具体使用的代码是基于 Huggingface 出品的 PyTorch-Transformers 上进行修改，不同的模型只需输入不同的预训练模型即可。使用的预训练模型有：

BERT-base_Chinese, BERT-wwm_Chinese, BERT-wwm-ext_Chinese, RoBERTa-wwm-ext_Chinese, RBT3_Chinese, RBTL3_Chinese。

本任务使用的基模型为 BERT，该模型虽然拥有非常强大的表征建模能力，但同时 BERT 的网络结构复杂，包含的参数众多，计算复杂度很高，即使使用了专用的 GPU 计算资源，其训练速度也是比较慢的，因此这就要求在对 BERT 模型融合时不能直接使用 Stacking 这种高计算复杂度的技术，因此我们选择了简单暴力，融合效果相对较好的投票方法对基模型 BERT 做融合的投票的方法,我们采用非等权重投票方法，对最好的三个模型设置 4、3、2 这样的权值。

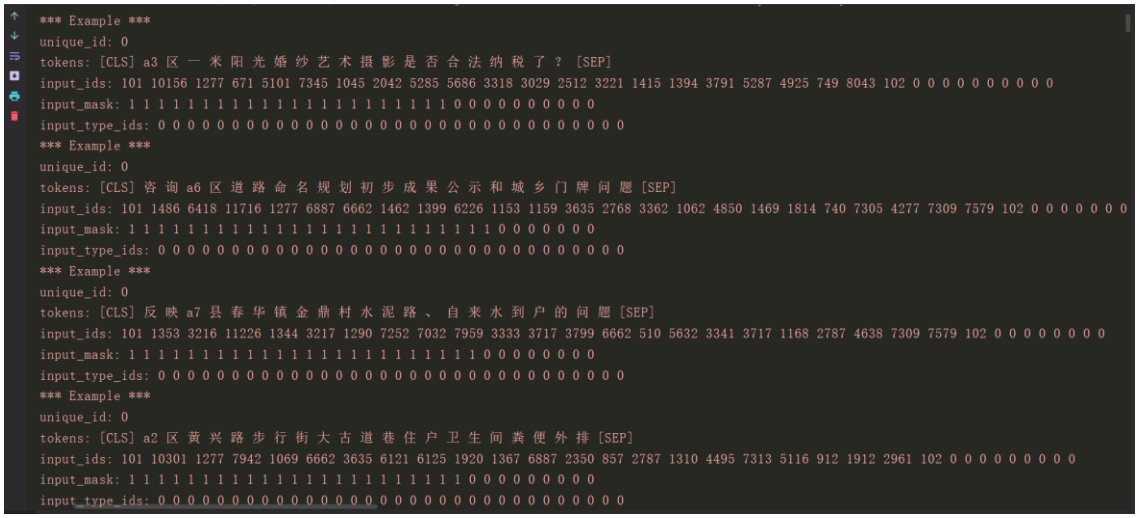
2.2. 问题 2 分析方法与过程

2.2.1. 流程图



2.2.2. 生成句向量

使用与第 1 问相同的 BERT 模型，将中文文本转换成计算机能够识别的向量，这里做了两种尝试，分别是将留言主题单独转换和将留言主题、留言详情合并后共同转换，下图是将留言主题单独转换的过程示意图。



在处理文本问题时通常使用余弦相似度算法计算两段文本的相似度，余弦相似度的定义如下：

$$\mathbf{similarity} = \cos(\theta) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}.$$

这里的 A_i，B_i 分别代表向量 A 和 B 的各分量。

如果两个句向量的余弦相似度越大，即越接近于 1，那么这两个句向量代表的文本内容越相似，这为后面的聚类算法奠定了基础。

利用上述方法可以的得到所有样本的余弦相似度矩阵 C，这是一个对角线元素为 1 的对称矩阵，其中 C_{ij} 是第 i 个句向量与第 j 个句向量的余弦相似度，

由于使用了两种不同的方式获得句向量，现将二者的余弦相似度矩阵进行比较，

C' =

```
[[0.9999999  0.8982246  0.87938285 ... 0.8613156  0.8608216  0.8698715 ]
 [0.8982246  0.99999964 0.9133814  ... 0.8553964  0.85025525 0.8698029 ]
 [0.87938285 0.9133814  1.         ... 0.84875035 0.8479595  0.85681   ]
 ...
 [0.8613156  0.8553964  0.84875035 ... 1.0000001  0.990154   0.9724798 ]
 [0.8608216  0.85025525 0.8479595  ... 0.990154   1.0000002  0.9640455 ]
 [0.8698715  0.8698029  0.85681    ... 0.9724798  0.9640455  0.9999999 ]]
```

C'' =

```
[[1.         0.94232637 0.9399134  ... 0.9277121  0.9432893  0.9420985 ]
 [0.94232637 1.         0.9494742  ... 0.9203919  0.9278579  0.9276342 ]
 [0.9399134  0.9494742  1.0000004  ... 0.9195499  0.920164   0.9268118 ]
 ...
 [0.9277121  0.9203919  0.9195499  ... 0.9999996  0.9649271  0.96828425]
 [0.9432893  0.9278579  0.920164   ... 0.9649271  0.9999999  0.9708476 ]
 [0.9420985  0.9276342  0.9268118  ... 0.96828425 0.9708476  0.99999964]]
```

C' 是将留言主题转换后得到的余弦相似度矩阵，C'' 是将留言主题和留言详情合并转换后得到的余弦相似度矩阵，可以看出 C'' 中所有的值几乎都是 0.9 以上，对于后续

的阈值选择困难较大，进而影响聚类效果，换个角度思考，如果将留言主题与留言详情共同作为样本来提取特征，由于留言详情内容较多，噪声也多，相对来说会降低特征提取的准确性、有效性，进而影响聚类的效果。因此选择第一种方式，仅仅采用留言主题作为样本，进行句向量转换，并将转换后的句向量保存为 data.npy，以便后续使用。

2.2.3. 聚类

聚类是一种无监督学习算法，最常见的是 k-means 算法，但是必须要提前选择有多少类别，很明显本题并不能事先知道有多少类别，故不能采用 k-means 算法。

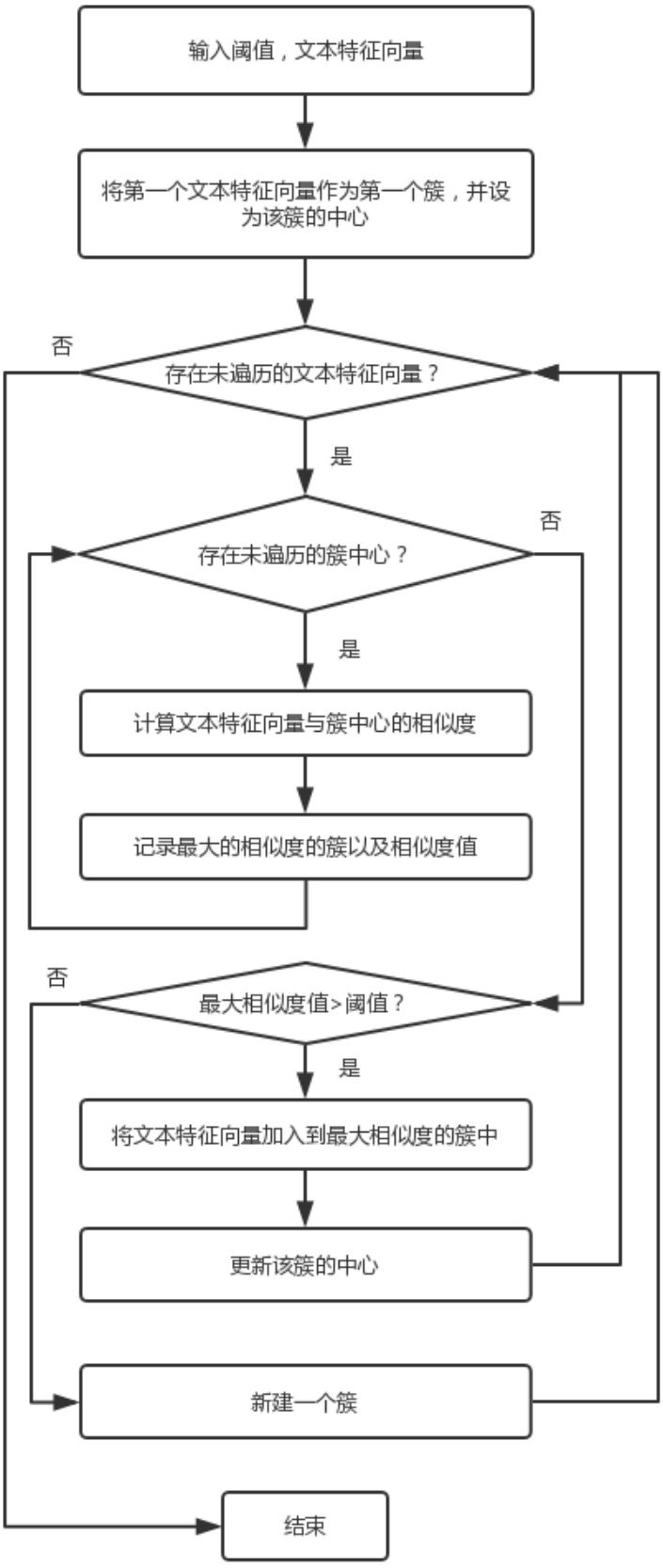
以下将使用两种聚类算法 Single-Pass 以及 DBSCAN 来解决本题。

2.2.3.1. Single-Pass

Single-Pass 算法是简单的文本聚类算法，将文本特征向量做相似度比较度相似度值大于阈值的文本归为一类文本。从而使主题更容易被发现，使计算更精准。

它的主要思想是，依次输入一个文本，判断当前文本与已有簇的匹配程度，如果当前文本与已有的某个簇相匹配，则把当前文本归入到该簇，反之则创建新的簇。

算法流程见下图：

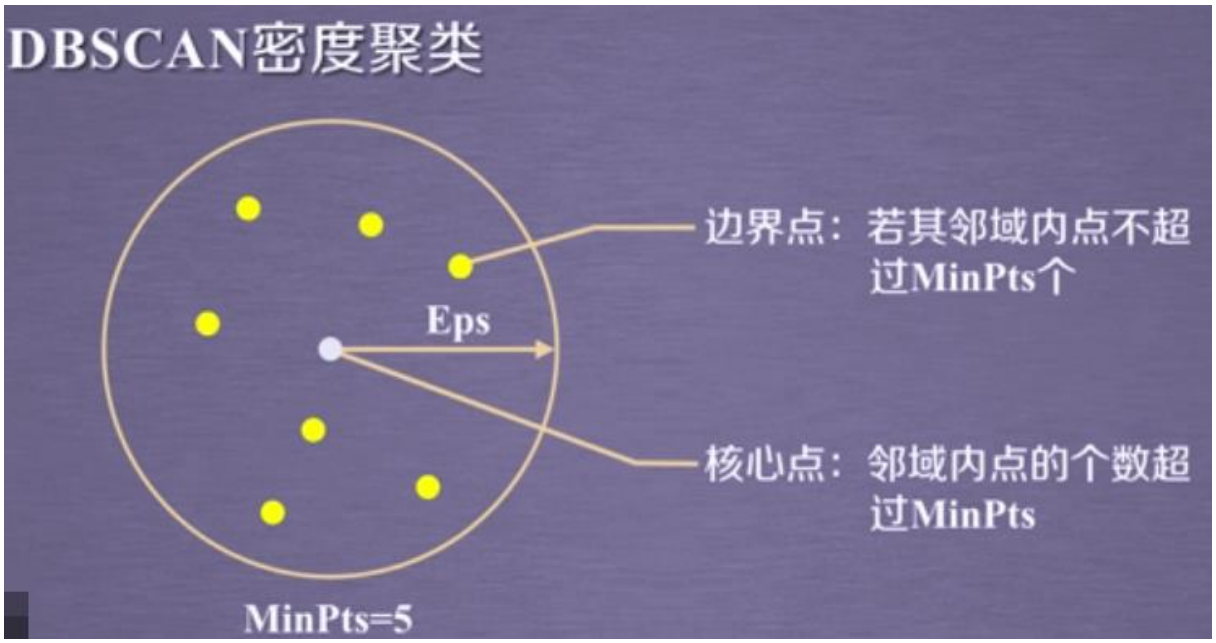

输入阈值，文本特征向量
将第一个文本特征向量作为第一个簇，并设为该簇的中心
否
存在未遍历的文本特征向量？
是
否
存在未遍历的簇中心？
是
计算文本特征向量与簇中心的相似度
记录最大的相似度的簇以及相似度值
否
最大相似度值>阈值？
是
将文本特征向量加入到最大相似度的簇中
更新该簇的中心
新建一个簇
结束

具体代码见 Single-Pass.py 文件，最后得到排名第一的热点问题如下

	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
88	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019-08-23 12:22:00	投诉伊景园.滨河苑开发商捆绑销售车位！A市武广新城片区下的伊景园.滨河苑是广铁集团铁路职工的...	0	0
318	196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52:14	\n\t\t\t\t\t\n\t\t\t\t\tA市伊景园.滨河苑现强制要求购房者捆绑购买1...	0	0
450	199190	A00095080	关于A市武广新城违法捆绑销售车位的投诉	2019/8/1 22:32:26	\n\t\t\t\t\t\n\t\t\t\t\t武广新城为铁广集团的定向商品房，在未取得预售...	0	0
1482	223247	A00044759	投诉A市伊景园滨河苑捆绑销售车位	2019/7/23 17:06:03	\n\t\t\t\t\t\n\t\t\t\t\t关于铁广集团铁路职工定向商品房伊景园滨河苑项...	0	0
1777	230554	A909174	投诉A市伊景园滨河苑捆绑车位销售	2019-08-19 10:22:44	投诉A市伊景园滨河苑开发商强制要求购房者捆绑购买12万车位，如果不买就取消员工的购房资格，国...	0	0
2345	244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	2019-08-24 18:23:12	广铁集团铁路职工定向商品房伊景园滨河苑项目非要捆绑销售车位！现在才知道车位是没有产权的只有7...	0	0
2658	251844	A909167	投诉伊景园滨河苑项目违法捆绑车位销售	2019-08-20 13:34:12	投诉广铁集团强制要求职工捆绑购买12万的车位费，不购买就取消购房资格！诺大的广铁集团，竟不顾...	0	1
2714	253253	A00051439	投诉A市北大资源理想家园小区人防车位涉嫌欺诈销售	2019/9/26 8:27:31	\n\t\t\t\t\t\n\t\t\t\t\t近几年来，A市的房地产业发展迅速，高楼大厦鳞...	0	0
2936	258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019-08-23 11:46:03	尊敬的领导:\n我是广铁集团铁路职工,本来武广新城片区的伊景园滨河苑是我们的定向限价商品房,...	0	0
3050	260254	A909173	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车位	2019-08-30 18:10:23	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车位，将购买车位强加在认购的职工头上，如果名业主...	0	0
3839	279070	A00095080	投诉A市伊景园滨河苑开发商违法捆绑销售无产权车位	2019/8/31 6:33:25	\n\t\t\t\t\t\n\t\t\t\t\t投诉A市伊景园.滨河苑开发商违法捆绑销售无产...	0	0
4137	285897	A909191	武广新城伊景园滨河苑违法捆绑销售车位,求解决	2019-08-01 20:06:52	我们是广铁集团铁路职工,武广新城片区的伊景园滨河苑楼盘是我们的定向限价商品房,\n在现在楼盘...	0	0
4297	289950	A00044759	投诉A市伊景园滨河院捆绑销售车位	2019/7/7 7:28:06	\n\t\t\t\t\t\n\t\t\t\t\t提问A市政府就广州局集团公司与A市政府及A市...	0	0

2.2.3.2. DBSCAN

DBSCAN 是一种非常著名的基于密度的聚类算法。其英文全称是 Density-Based Spatial Clustering of Applications with Noise，意即：一种基于密度，对噪声鲁棒的空间聚类算法。直观效果上看，DBSCAN 算法可以找到样本点的全部密集区域，并把这些密集区域当做一个一个的聚类簇[2]。



DBSCAN 算法的流程：

1.将所有点标记为核心点、边界点或噪声点；
2.删除噪声点；
3.为距离在 Eps 之内的所有核心点之间赋予一条边；
4.每组连通的核心点形成一个簇；
5.将每个边界点指派到一个与之关联的核心点的簇中（哪一个核心点的半径范围之内）。

具体代码见 DBSCAN.py 文件，最后得到排名第一的热点问题如下

	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
832	208285	A909205	投诉小区附近搅拌站噪音扰民	2019-12-15 12:32:11	尊敬的领导，我是A市暮云街道丽发新城的一名业主，最近遇到了意见特别烦心的事情，我是做小区安保...	0	24
1050	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019-12-10 12:34:21	我是暮云街道丽发新城小区的业主，我要投诉开发商在小区附近违建大型搅拌站。该搅拌站的设备太吵了...	0	0
2808	255008	A909208	投诉小区附近搅拌站噪音扰民	2019-11-18 12:23:22	暮云街道丽发新城边上在建大型搅拌站，听说是从别的地方搬过来的，体会最深的就是噪音很大，扬尘污...	0	0
3075	261072	A909207	投诉小区附近搅拌站噪音扰民	2019-11-23 23:12:22	投诉A市暮云街道丽发新城附近大型搅拌站水泥厂噪音严重扰民，扬尘污染环境，希望有关部门回复，在...	2	9
3321	266665	A00096279	投诉小区附近搅拌站噪音扰民	2019-12-04 17:23:22	开发商把特大型搅拌站，水泥厂从绿心范围内搬迁到A市暮云街道丽发新城的居民区，产生的噪音和扬尘...	0	0
68	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	我是A2区丽发新城小区的一名业主，我要投诉同发投资有限公司在未经小区业主同意的情况下，在离小...	0	0
1797	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-12-02 11:20:21	尊敬的领导，我是A2区丽发新城小区的一名业主，再次投诉同发投资有限公司在未经小区业主同意的情...	0	0
2319	243692	A909201	丽发新城小区附近的搅拌站噪音严重扰民	2019-11-15 11:23:21	领导您好！我是暮云街道丽发新城小区的业主，我投诉小区附近建的搅拌站，机器不分昼夜的运行，各种...	0	2
2707	253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-12-04 12:10:21	投诉A2区丽发新城小区附近违建搅拌站！该站每天早6点一直到晚8点设备都在运行，每天耳边都是各...	0	0
3562	272361	A909242	丽发新城小区旁建搅拌厂严重扰民	2019-12-04 08:46:20	本人是丽发新城小区的居民，最近小区附近居然新建了一个搅拌站，严重影响的我的生活，平常上班不在...	0	1
1089	214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和污染环境	2020-01-25 09:07:21	你们管不管A2区丽发新城小区啊！这个附近建了个搅拌厂啊，天天吵天天吵，烦死了不仅吵还臭！说好...	0	0
1236	217700	A909239	丽发新城小区旁的搅拌站严重影响生活	2019-12-21 02:33:21	开发商把特大型搅拌站，水泥厂从绿心范围内搬迁到丽发新城小区旁边不到百米的地方，灰尘，噪音污染严重	0	1
3254	264944	A0004260	A2区丽发新城附近修建搅拌厂噪音、灰尘污染	2019-11-02 14:23:11	A市A2区丽发新城小区附近，作为长株潭绿心地带，还是人口众多的小区，竟然堂而皇之修建搅拌厂，...	0	0
3389	268300	A909225	A2区丽发新城附近修建搅拌厂噪音污染导致生活不正常	2019-11-25 10:17:58	人口众多的丽发新城小区附近，堂而皇之修建搅拌厂，请问是谁审批的，这不是正常现象，请还小区居民...	0	0
3626	274004	A00026895	A市暮云街道丽发新城社区附近搅拌站噪音污染严重	2019-12-21 10:11:09	我是A市暮云街道丽发新城小区的一名业主,向领导反映开发商在居民区附近建搅拌站,每天噪音影响根...	0	0
1881	233158	A909242	丽发新城小区旁建搅拌厂严重扰民！	2019-12-05 08:46:20	本人是丽发新城小区的居民，最近小区附近居然新建了一个搅拌站，严重影响的我的生活，平常上班不在...	0	0
105	190802	A00072636	A市丽发小区建搅拌站，噪音污染严重	2019/11/25 18:58:05	\n\t\t\t\t\t\n\t\t\t\t\t发同投资有限公司在未经业主同意的情况下，在丽...	0	0
2154	239648	A909211	A市A2区丽发新城小区附近搅拌站明目张胆污染环境	2020-01-06 22:41:31	丽发新城小区附近近日突然建起了搅拌厂！特别扰民，机器一天到晚的响，吵得人不得休息，还有灰尘颗...	0	0
2350	244335	A909135	A市暮云街道丽发新城社区搅拌站灰尘，噪音污染严重	2019/12/2 12:11:23	\n\t\t\t\t\t\n\t\t\t\t\t您好！我是A市暮云街道丽发新城社区（丽发新城...	0	0
2468	247160	A000104195	A市丽发小区建搅拌站，噪音污染严重	2019/11/25 19:08:41	\n\t\t\t\t\t\n\t\t\t\t\t发同投资有限公司在未经业主同意的情况下，在万...	0	0
2945	258242	A909220	A市暮云街道丽发新城社区搅拌站灰尘，噪音污染严重	2019-12-02 12:23:11	A市暮云街道丽发新城小区附近的搅拌站灰尘、噪音污染严重，严重影响附近居民休息，使其不能以最佳...	0	0
463	199379	A00092242	A2区丽发新城附近修建搅拌厂，严重污染环境	2019/11/25 10:17:56	\n\t\t\t\t\t\n\t\t\t\t\tA市A2区丽发新城小区近期百米范围内修建搅拌...	0	0
848	208714	A00042015	A2区丽发新城附近修建搅拌站，污染环境，影响生活	2020-01-02 00:00:00	尊敬的领导：\n 您好！作为一名居住在A2区丽发新城的业主，和小区内的每一位业主一样，...	0	4
1557	225217	A909223	A2区丽发新城附近修建搅拌厂严重影响睡眠	2019-11-15 09:17:36	我已经好久没睡过安稳觉了,A市暮云街道丽发新城小区开发商在小区附近建一搅拌站,每天尘土飞扬,...	0	0
2948	258378	A00084226	丽发新城社区附近搅拌站修建严重影响居民生活	2019-11-23 00:00:00	开发商在A市暮云街道丽发新城社区附近百米范围内修建搅拌厂，整天尘土飞扬，噪声嘈杂，不仅污染了...	0	0
1206	216824	A909214	搅拌站大量加工砂石料噪音污水影响丽发新城小区环境	2019-12-25 12:15:57	最近一段时间以来，A2区丽发新城小区一带的居民，深夜经常被刺耳的机器轰鸣声惊醒，经调查了解，...	0	0
3595	273282	A909226	A2区丽发新城附近修建搅拌厂烟尘滚滚，声音刺耳	2019-12-25 10:17:59	A2区丽发新城附近修建搅拌厂烟尘滚滚，声音刺耳，请问政府，我们该怎么生活。审批的、监管的、城...	0	0
3024	259788	A909221	A市暮云街道丽发新城社区搅拌厂危害居民健康	2019-12-07 00:00:00	A市暮云街道丽发新城社区（丽发新城小区）搅拌站灰尘，噪音污染严重，危害附近小区里的居民身心健...	0	0
632	203393	A00053065	A市丽发新城小区侧面建设混泥土搅拌站，粉尘和噪音污染严重	2019/11/19 14:51:53	\n\t\t\t\t\t\n\t\t\t\t\t发同投资有限公司在未经业主同意的情况下，在丽...	0	2

2.2.3.3. 算法比较

Single-Pass 算法作为一种增量式的聚类方法，其具有流程清晰、结构简单、运算

速度快等几个重要优点，并且非常适用于大数据环境之下来使用。并且 Single-Pass 算法具有较低的时间复杂度，因此在话题发现和话题跟踪领域使用较多，其性能和效果也得到了认可。但是该算法的聚类精度较差，主要原因是初始聚类阶段的样本点较少，此时的类别信息还不够充分和准确。而随着样本点的增加，早期由于信息量不足而导致的错误聚类，到系统后期可能产生巨大影响。而且 Single-Pass 算法在增量式子处理数据时每次只处理一个样本点，在大数据环境下是不可容忍高时间复杂度的运算。

DBSCAN 算法可以对任意形状的样本簇进行很好的聚类，同时又不需要像 K-Means 等算法一样,需要提前估计样本簇的个数,在新闻等样本集中,DBSCAN 算法的最大优势是:可以在聚类的过程中寻找到异常点(本文中的异常点即留言数很少的话题)，对具有大量噪声的样本集有着优秀的性能。

通过上述两种算法的结果也可以看出，DBSCAN 得到的效果更好，排名第一的热点问题中错误数据比 Single-Pass 的更少，而且得到的类别数量更全，所以最后选择了 DBSCAN 算法来解决该问题。

2.2.4. 定义热度评价指标

对于一个群众问题而言，问题的数量可以作为热度的一个评价指标，同时，持久性也具有一定代表性。基于上述思考，提出了基于广度和深度的热度评价指标。

从广度上来说，一个群众问题的热度代表该问题一段时间内的反映数量。从深度上说，一个群众问题的热度代表该问题的持续时长，即使两个问题在某一天的数量差不多，但是从深度上来说数天内数量更多的问题热度值更高。基于上述思想，设计一个热度指数计算公式：[3]

$$\mathrm{HotPts(T,\ date)} = \lg(\mathrm{Num}_{\mathrm{T,\ date}}) + \alpha HotPts(T, date-1)$$

其中 T 为任意一个指定的群众问题； $\mathrm{Num}_{\mathrm{T,\ date}}$ 为指定的群众问题 T 在日期 date 下所对应聚类簇下的问题数目；α 为前一日的问题热度值消退程度的超参数，$0<\alpha<1$，一般取 α=0.1；date 为群众问题持续时间，默认 HotPts(T,0)=0。

2.2.5. 生成结果

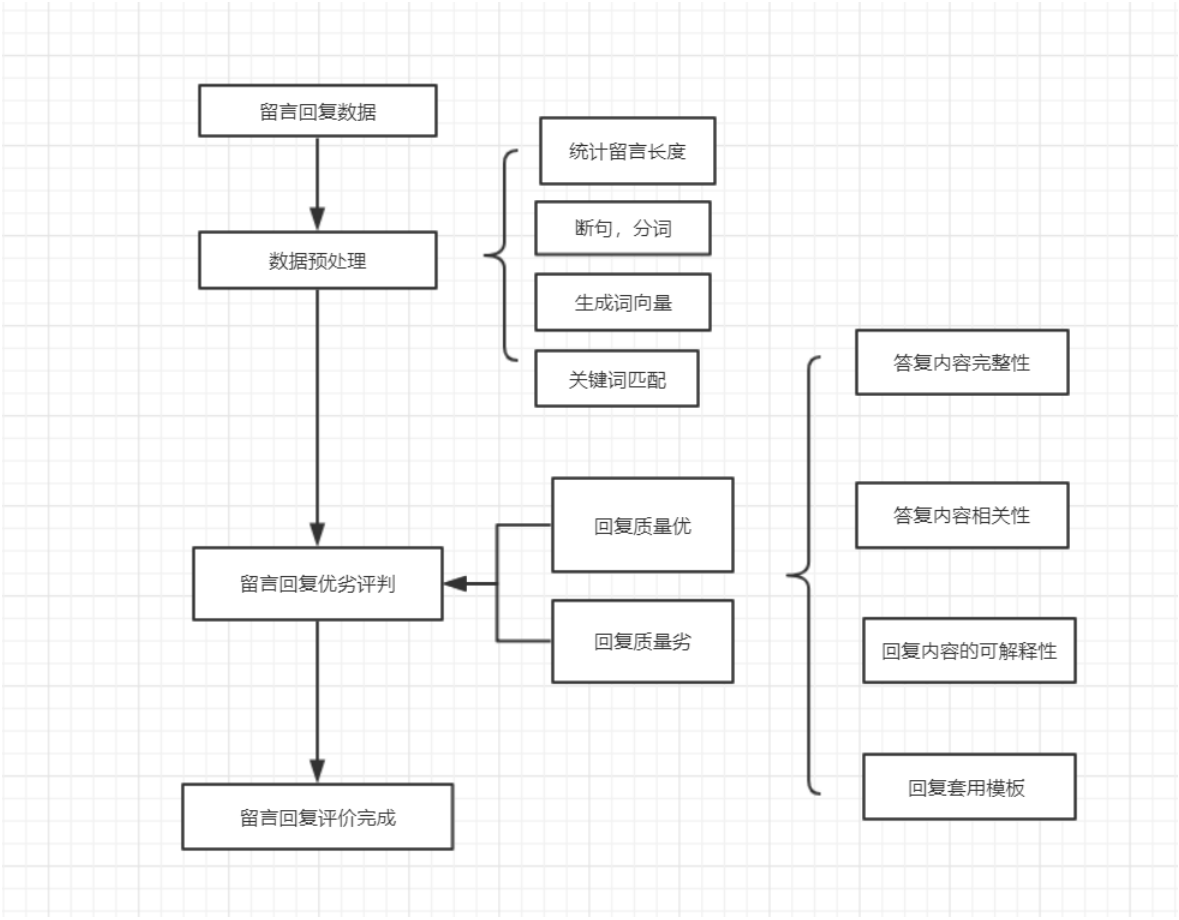
将 DBSCAN 算法求得的最终排名前五的热点问题保存到“热点问题留言明细表.xls”中，并将每一个类别里的留言时间范围、地点人群找出并计算热度指数，将上述信息保存到“热点问题表.xls”。

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	3.37	2019/11/02至2020/01/25	A市A2区丽发新城小区	小区附近搅拌站噪音扰民
2	2	3.21	2019/01/15至2019/12/24	A市	人才购房补贴
3	3	2.83	2019/07/07至2019/10/24	A市伊景园滨河苑	违法捆绑销售无产权车位
4	4	2.64	2019/04/18至2020/01/03	A2区福满新城	施工噪音扰民
5	5	2.30	2019/03/04至2019/11/04	A8县星河湾	房屋有严重的质量问题

	A	B	C	D	E	F	G	H
1	问题ID	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
2	1	208285	A909205	投诉小区附	2019-12-15 12:32:11	尊敬的领导	0	24
3	1	213464	A909233	投诉丽发新	2019-12-10 12:34:21	我是暮云街	0	0
4	1	255008	A909208	投诉小区附	2019-11-18 12:23:22	暮云街道丽	0	0
5	1	261072	A909207	投诉小区附	2019-11-23 23:12:22	投诉A市暮	2	9
6	1	266665	A0009627	投诉小区附	2019-12-04 17:23:22	开发商把特	0	0
7	1	189950	A909204	投诉A2区丽	2019-11-13 11:20:21	我是A2区丽	0	0
8	1	231136	A909204	投诉A2区丽	2019-12-02 11:20:21	尊敬的领导	0	0
9	1	243692	A909201	丽发新城小	2019-11-15 11:23:21	领导您好！	0	2
10	1	253040	A909202	投诉A2区丽	2019-12-04 12:10:21	投诉A2区丽	0	0

2.3. 问题 3 分析方法与过程

2.3.1. 流程图



针对附件 4 相关部门对留言的答复意见，提取留言回复信息质量优劣的特征，研究如何从答复的相关性、完整性、可解释性等角度对回复信息的质量进行量化描述，建立衡量回复信息质量等方面的综合评价模型。

2.3.2. 留言回复质量优劣的特征

如何识别出最佳答复，如何衡量答复的质量，是政府有关部门亟需解决的重要问题。根据日常回复的标准和规范，我们对留言回复质量优劣的特征进行归纳，得到以下优劣评判特征。

2.3.2.1. 良好回复的参考特征：

1）回复内容与问题相关性较高，精确地解答网民。

2）回复内容和政府有关领域词语的语义相似度高；

可解释性高，回复内容引用相关法规、文件及政府通知，例如“《西地省人口与计划生育条例》第 17 条规定”，“《E9 县 2019 年城乡居民医疗保险参保缴费工作方案》的通知”；

回复内容充实完整，对群众问题的答复比较详细；

回复态度良好，对网民的问题耐心地回答。

2.3.2.2. 较差回复的参考特征：

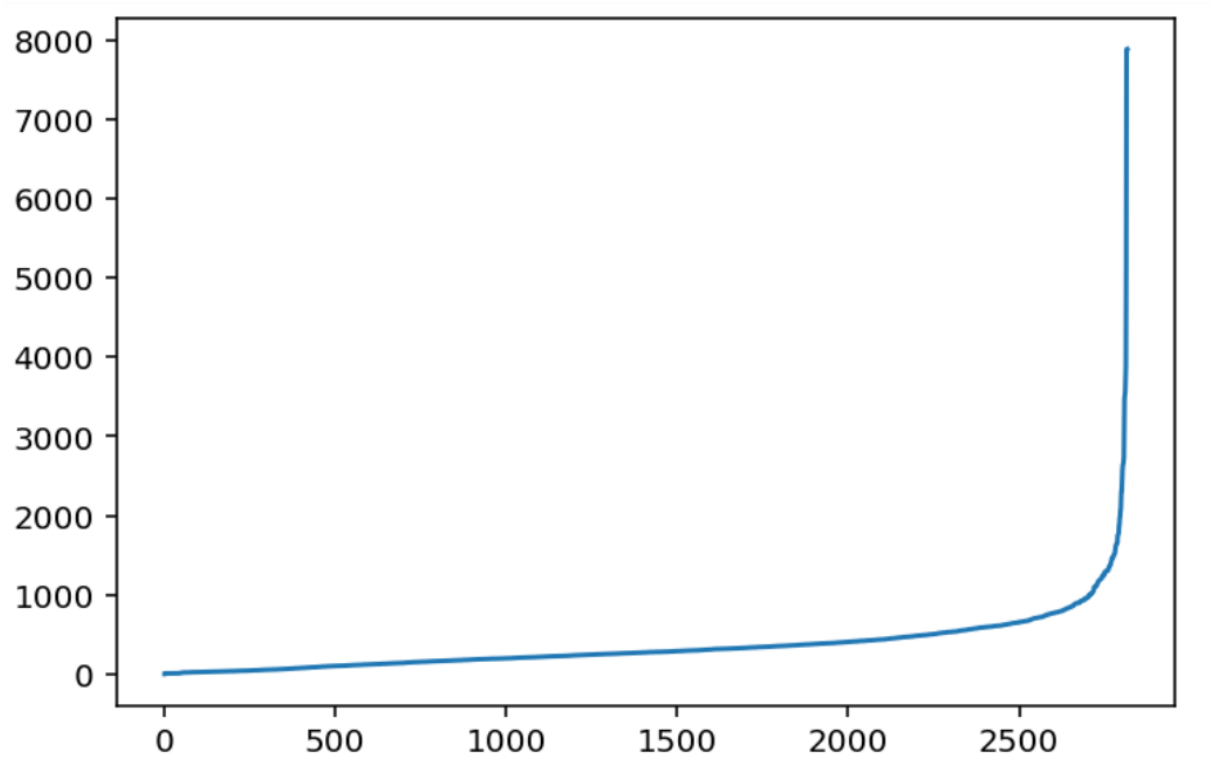
1）回复内容过于简短，例如“已收悉”“已调查”等；
2）回复内容套用固定模板，例如“您的留言已收悉，关于您反映的问题，已转……查处理。”等；
3）对不同问题使用同一回复，或回复内容的相似度很高。例如大量出现“您好，您所反映的问题，已转交相关部门调查处置。”等类似语句；
4）语句毫无逻辑或者全为一串乱码，例如“你好！2019 年 6 月 13 日”，“UU0081182”。

2.3.3. 答复信息的量化方法

2.3.3.1. 答复内容完整性

在许多情况下，无效答复的文本长度较短，且与一般问题的文本长度有明显差距。利用 pandas 读入附件 4 的答复意见，并按照答复的字数进行排序，如下图所示，字数相对较少的答复很明显是无效的。所以根据此方法可以初步将一些无效答复筛选出来。

```
0 已收悉
1 已收悉
2 已收悉
3 2016年6月12日
4 2019年1月14日
5 请咨询K市人社部门。
6 网友：您好！留言已收悉
7 网友：您好！留言已收悉
8 网友：您好！留言已收悉
9 网友：您好！留言已收悉
10 网友：您好！留言已收悉
11 网友：您好！留言已收悉
12 网友：您好！留言已收悉
13 网友：您好！留言已收悉
14 网友：您好！留言已收悉
15 网友：您好！留言已收悉
16 网友：您好！留言已收悉
17 2018年12月12日
18 “UU0081182”
19 网友：您好！留言已收悉
20 网友：您好！留言已收悉
21 网友：您好！留言已收悉
22 网友：您好！留言已收悉
23 网友：您好！留言已收悉
24 网友：您好！留言已收悉
25 网友：您好！留言已收悉
26 网友：您好！留言已收悉
```



针对答复意见，答复内容的详细程度与答复的文本长度有直接的关系。简短内容的答复信息量一般不够，评分应该较低；同时，较长文本的答复评分不应该过高。因此，可以考虑使用对数函数来量化答复的文本长度与评分的关系，建立“答复内容是否充实”的评价项 F_1[4]：

$$F_1 = \log_m L_i$$

其中，L_i 为针对第 i 个问题回复的文本长度，m 为常数。

2.3.3.2. 答复内容相关性

某一个部门答复不同类问题时采用相同的模板或者复制之前的答复，导致问题与答复的相关性不高，可以通过建立文本相关性评价函数 TEXTREL 来解决，将 $TEXTREL(P_i, A_i)$ 作为问题 Pi 与答复 Ai 的一个相关性指标，该值越大，表示二者越相关。下面介绍 TEXTREL 的计算方法。

一条问题或答复由许多关键词构成，利用 Hanlp 对每一组问题与答复分词，但二者大部分词语间相似度较低，所以只考虑二者间相似度最高的词语间的相似度。

利用 BERT 模型生成词向量后计算词向量间的余弦相似度矩阵 S，与问题二类似，判断问题与答复间的相关性，令 m1 为第 1 行的最大值，去掉 m1 所在的行与列得到余子阵 M_{1i}^1，找出余子阵 M_{1i}^1 第 1 行的最大值 m2，并去掉 m2 所在的行与列得到余子阵 M_{1i}^2，采用上述方法直到得到的余子阵为空矩阵为止。则

$$F_2 = TEXTREL = \frac{\mathrm{m}_1 + \mathrm{m}_2 + ... + \mathrm{m}_i}{l + \dfrac{\max(k, p) - l}{l}}$$

其中 k，p 分别是问题与答复生成的关键词数量，$l = \min(k, p)$ 为查找的次数。

这种方法相比于传统的余弦度量法可以解决文本数据中存在的自然语言问题，即同义词和多义词的情况。当 TEXTREL 越接近于 1，说明问题与答复的相关性越高。

2.3.3.3. 回复内容的可解释性

回复的内容需要有严密的逻辑和准确的表达，即回复信息是否有理有据。假设文本数据中共有 N_0 条语句。将引用政府文件通知时出现的关键词 （如“《……》第…条规定”，“根据……通知”）和回复中的文本进行语义匹配。若在回复的文本语句中匹配到相应的关键字，则认为该条回复引用了政府文件通知。统计回复信息引用法律条文的回答数目 N_{low}，计算出现频率，即 N_{low} 与 N_0 的比值。该频率值可以作为“是否有理有据”的评价项，记为： $F_3 = \dfrac{N_{T_{key}}}{N_0}$

2.3.3.4. 回复套用模板

有些较差回复中会出现很多固定词语，如“详情咨询”“已收悉”“已转交相关部门”等。本文建立了一个较差回复的关键词组成的集合 T 。对回复 T_{key} 进行关键词匹配，若一条问答中出现特别多的 T_{key}，则认为 该条回答为较差回答。统计此类回复的语句数量 N_{key} ，以其出现频率作为评价项： $F_4 = \dfrac{N_{T_{key}}}{N_0}$

2.3.4. 回复评价模型

评价问题回复的关键在于如何建立对回复质量的量化评分模型，根据上述量化方法，给 出问题回复评价模型，其流程如图 1 所示。

下面对 4 项量化指标进行整合，以计算留言回复信息的得分情况，建立留言回复信息的质量评价函数 F 。

$$F = \lambda_1 * F_1 + \lambda_2 * F_2 + \lambda_3 * F_3 + \lambda_4 * F_4$$

3. 结论

利用 BERT 模型处理中文文本问题时具有很大的优势，面对本次数据挖掘任务时，有着高于传统方法的准确率与效率；利用 BERT 模型生成句向量，解决文本间的相关性的问题也显得游刃有余。针对答复质量评价问题，通过将传统的算法改进，使之更适合于处理有近义词多义词的自然语言文本问题。

从分析结果来看，群众反映的热点问题主要是日常生活中的事件，例如噪声扰民、购房落户等。相关部门需要重点关注一下这些民生问题，针对广大群众反映的问题，做出合理、有效的回复。这样才能充分发挥各部门的职能，有助于相关部门进行有针对性地处理，提升服务效率。

4. 参考文献

[1] 胡局新，张功杰. 基于 K 折交叉验证的选择性集成分类算法. 江苏师范大学. 2013

[2] 夏鲁宁，荆继武. SA-DBSCAN:一种自适应基于密度聚类算法. 中国科学院大学. 2009

[3] 童昱强. 基于数据挖掘的网络新闻热点发现系统的设计与实现. 北京邮电大学. 硕士学位论文. 2019

[4] 杨开平，李明奇，覃思义. 基于网络回复的律师评价方法. 电子科技大学. 2018