“智慧政务”文本挖掘分析

摘要

本文以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为研究对象，首先旨在建立关于留言内容的一级标签分类模型。其次，通过定义合理的热度指标，挖掘出某一时段内关于特定地点或特定人群的热点问题。再者，设计相关部门对于留言的答复意见的评价方案。为政府作出正确的工作指示决策提供了方便，有利于大大的提升服务效率，并且对政府了解民意、汇聚民智、凝聚民气具有重大的意义。

针对问题一，首先，把附件2中“留言主题”这一列作为特征，“一级分类”这一列作为标签，建立了BERT模型。通过对数据的清洗、划分以及训练得到模型的准确率为87.25%，F1得分函数为0.8645。由于BERT模型存在一定的缺点，所以利用改进后的ERNIE模型进行再次训练，训练后得到模型的准确率为90.06%，F1得分函数为0.8935。可见，在ERNIE模型上准确率有所提高。

针对问题二，提取附件3中的“留言主题”以及“留言详情”，构建DBSCAN聚类模型对其进行聚类，得到33类。通过建立热度指标（热度指标=相同话题条数*0.5+赞同数*0.25+反对数*0.25）对每个类进行排名，得到了前5个热点问题，从而得出热点问题详细表。再利用TextRank算法计算每个标题的权重，用权重最高的标题当成热点的问题描述，得到热点问题表。

针对问题三，根据附件4中的“留言详情”、“答复意见”，构建gensim模型计算两者之间的文本相似度，并计算答复时间间隔，按照质量评分指标（质量评分=0.7*文本相关性+0.1*时间间隔标准化+0.1*平均字符长360）对答复意见的质量进行评价，分为低、中、高质量的评价。

关键词：智慧政务；BERT模型；ERNIE模型 ；DBSCAN聚类；TextRank算法；gensim算法

Abstract

Key words：

This paper takes the record of public political messages from open sources on the Internet
and the comments made by relevant departments on some public messages as the research object.First, it aims to establish a primary label classification model for the message content. Secondly, by defining a reasonable heat index, hot issues about a specific place or a specific group of people can be found in a certain period of time. Furthermore, the design of the relevant departments for the comments on the comments of the evaluation program. It is of great significance for the government to understand public opinion, pool people's wisdom and pool people's spirit. It also provides convenience for the government to make correct work instructions and decisions, which is conducive to greatly improving service eficiency.

For question 1, first of all, BERT model was established by taking the column of "message subject" in annex 2 as the feature and the column of "primary classification" as the label.The accuracy of the model was 87.25%, and the F1 score function was 0.8645.Due to the shortcomings of BERT model, the improved ERNIE model was used for retraining. After training, the accuracy of the model was 90.06% and the F1 score function was 0.8935.It can be seen that the accuracy of ERNIE model is improved.

For question 2, extract the "message subject" and "message details"' in attachment 3,construct DBSCAN clustering model to cluster them, and get 33 categories. By establishing heat index (heat index = number of topics *0.5+ number of pros *0.25+ number of cons *0.25) to rank each category, the top 5 hot issues were obtained, and then the detailed table of hot issues was obtained. TextRank algorithm is used to calculate the weight of each title, and the title with the
highest weight is used as the description of hot issues to obtain the hot issues table.

For question 3, according to annex 4 of the "message detais", "reply", build the model, carries on the text to quantify, use of computing text similarity between, and reply to interval statistics, according to the quality grading index (quality score = similarity + answer time * 0.2 * 0.8) to evaluate the quality of the views on reply, divided into low, medium and high quality.

Key word:intelligent government;BERT model;Groeb model;DBSCAN clustering;TextRank algorithm;gensim

目录

第一章 问题重述

1. 问题背景

近年来，随着大数据、云计算、人工智能等新兴技术的发展，越大越多的技术被应用于日常生活中。随着微信、微博、市长信箱、阳光热线等网络问政平台的兴起，各类社情民意相关的文本数据量不断攀升，利用新兴技术解决民生民意问题逐渐成为趋势，基于自然语言处理技术的智慧政务系统的应用成为新时代的趋势，能够极大的减轻了以往依靠人工进行留言划分和热点整理的相关部们的工作压力并极大的推动了政府的管理水平和施政效率。

2. 要解决的问题

（1）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）根据附件 3 给出的数据，建立合理的热度评价指标，得出排名前 5 的热点问题，并建立“热点问题表.xls”、“热点问题明细表.xls”。

（3）根据附件 4 中的答复意见，从答复的相关性、完整性、可解释性等对其质量给出评价方案。

第二章 群众留言分类

1.问题分析

对于附件 2 给出的数据（包含“留言编号”、“留言用户”、“留言主题”、“留言时间”、“留言详情”、“一级标签”），需要建立关于留言内容对应的一级标签的分类模型。通过数据预处理，剔除暂时无需用到的数据列，考虑到“留言主题”对“留言详情”内容有较好的概括性，提取出“留言主题”与“一级标签”列作为待处理数据，进行模型训练。

2.数据预处理

本文选取附件 2“留言主题”与“一级标签”作为研究对象。数据总量有 9210 条。经过去重去空后，有效数据量为 8908 条。 “一级标签”列共 7 类。用数字 id“0~6”分别表示，见表 1。转化后的“留言主题”与标签“id”列数据保存在 data.txt（见附录）中。

表 1 一级标签 id 转换表

Category	count	Id
城乡建设	1974	0
劳动和社会保障	1899	1
教育文体	1530	2
商贸旅游	1148	3
环境保护	909	4
卫生计生	856	5
交通运输	592	6

为了保证每次随机划分数据结果一致，本文在代码设置了随机种子（random.seed）。随机打乱数据生成文件 new_Data.txt（见附录）。

将随机打乱的数据 new_Data.txt 按划分比例 6:2:2 划分训练集(train.txt)、测试集(test.txt)、验证集(dev.txt)。（见附录）

数据预处理框架图如图 1 所示：

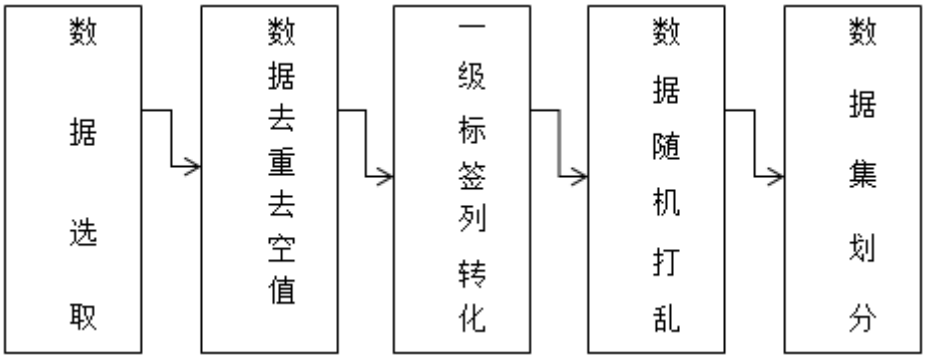

图 1 留言主题数据预处理

3. 模型的建立

3.1 BERT 模型的简介

BERT 模型，全称是 Bidirectional Encoder Representation Transformers，它是基于 Transformer 的双向编码器表征。BERT 模型的根基就是 Transformer，来源于 attention is all you need。其中双向的意思表示它在处理一个词的时候，能够考虑到该词前面和后面单词的信息，从而获取上下文的语义。可以理解为这是一个通用的 NLU（Natural Language Understanding）模型，为不同的 NLP 任务提供支持。

3.2 BERT 模型的网络架构

BERT 的模型架构基于 Transformer，实现了多层双向的 Transformer 编码器，模型结构如图 2。

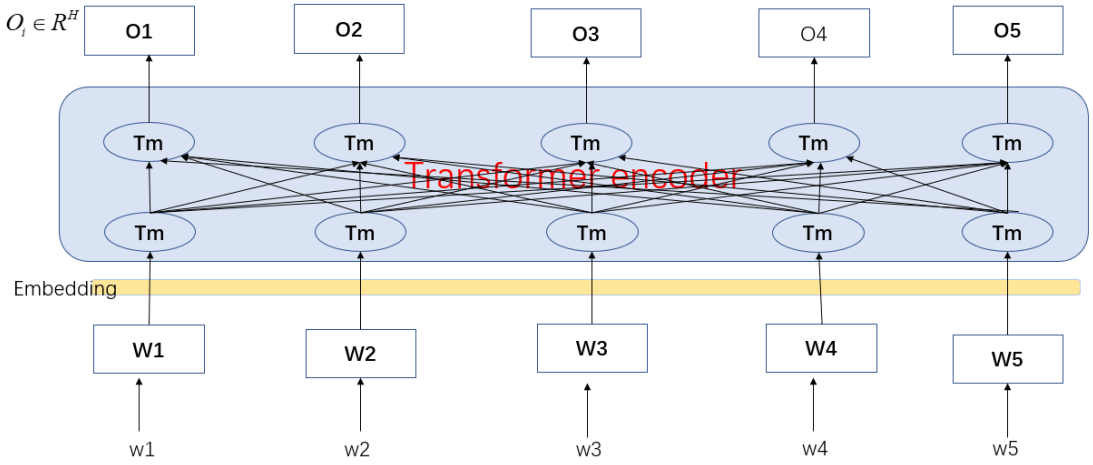

图 2 BERT 模型结构

BERT 有两个模型，一个是 1.1 亿参数的 base 模型，一个是 3.4 亿参数的 large 模型，而在本文当中选择第一个模型架构进行训练。里面所设置的参数如表 2 所示：

表 2 BERT 模型有关的参数表

Model	Transformer 层数（L）	Hidden units(H)	Self-attention heads(A)	总参数
BERT(base)	12	768	12	1.1 亿
BERT（large）	24	1024	16	3.4 亿

3.3 ENRT 模型的简介

ERNIE 模型通过对训练数据中的词法结构、语法结构、语义信息进行统一建模，极大地增强了通用语义表示能力，在多项任务中均取得了大幅度超越 BERT 模型的效果。

3.4 ERNIE 模型的网络架构

1. ERNIE 初探

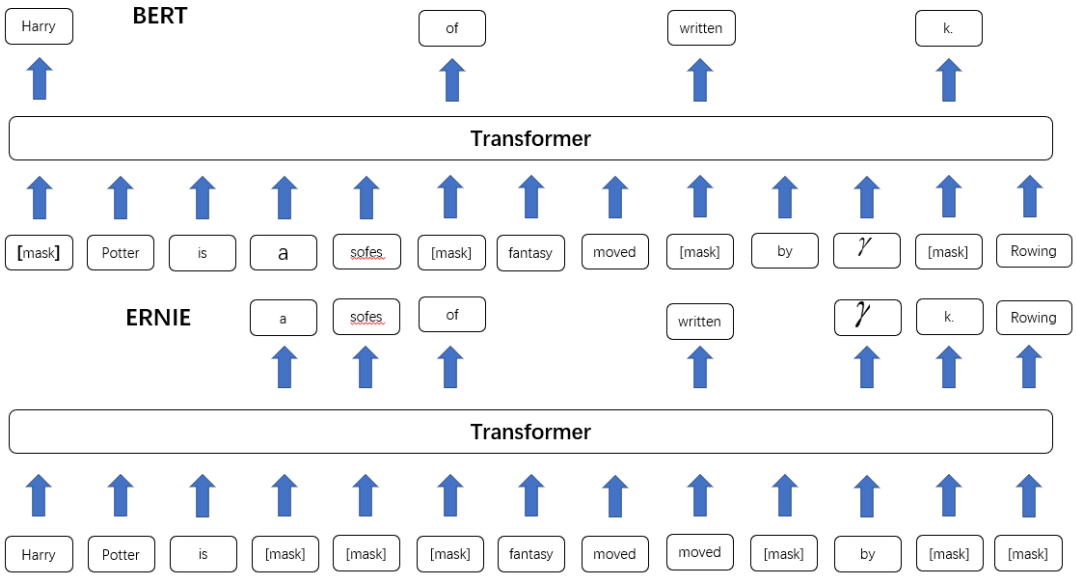

图 3 ERNIE 初探

2. ERNIE 结构详解

ERNIE 基本上是 transformer 的 encoder 部分，并且 encoder 在结构上是全部一样的，但是并不共享权重，具体区别如下：

（1）Transformer: 6 encoder layers, 512 hidden units, 8 attention heads。

（2）ERNIE Base: 12 encoder layers, 768 hidden units, 12 attention heads。

（3）ERNIE Large: 24 encoder layers, 1024 hidden units, 16 attention heads。

从输入上来看第一个输入是一个特殊的 CLS，CLS 表示分类任务就像 transformer 的一般的 encoder，ERINE 将一序列的 words 输入到 encoder 中. 每层使用 self-attention，feed-word network，然后把结果传入到下一个 encoder。

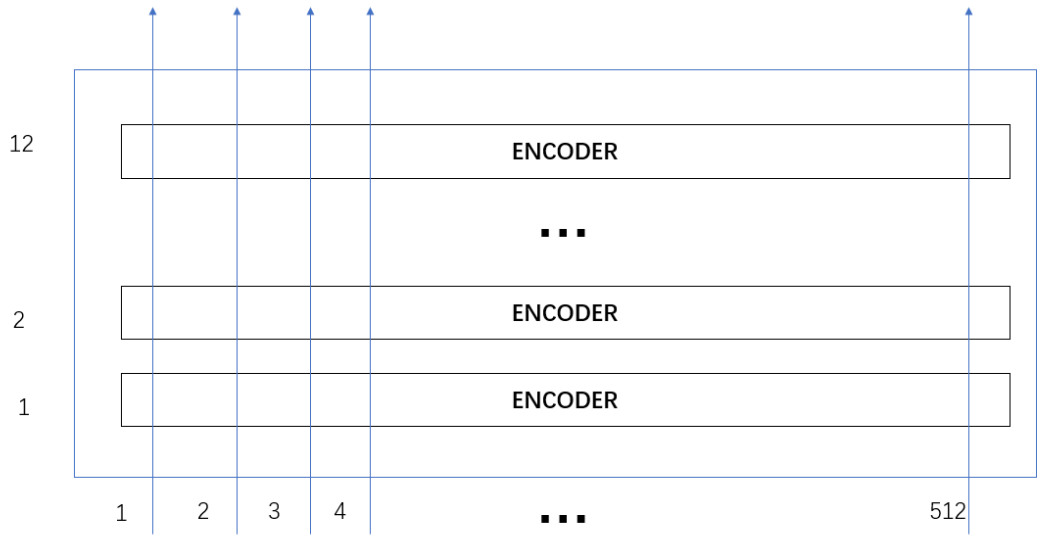

图 4 ERNIE 的 encoder 结构详解

3. ERNIE encoder 说明

encoder 由两层构成，首先流入 self-attention layer，self-attention layer 输出流入 feed-forward 神经网络。

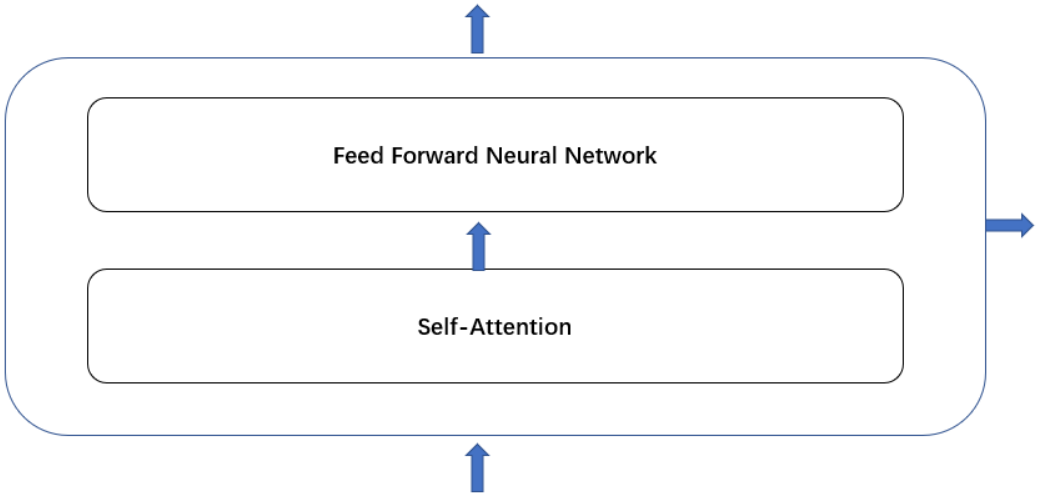

图 5 encoder 说明

4. encoder 结构详解

最下层的 encoder 的输入是 embedding 的向量，其他的 encoder 的输入，便是更下层的 encoder 的输出，一般设置输入的 vectors 的维度为 512。

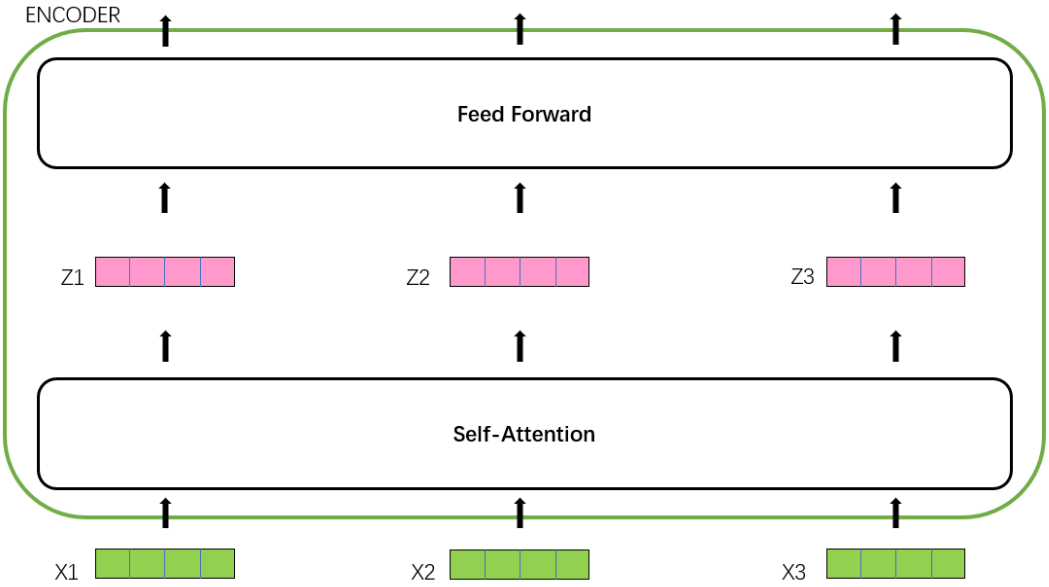

图 6 encoder 详解

4.模型的求解

4.1 实验条件

该模型训练所需要的实验配置，如表 2 所示：

表 3 实验环境配置

CPU	Intel i5
内存	16GB
显卡	RTM 540
操作系统	Window 16.04

Python	3.6.2
CUDA	9.0
cuDNN	7.1.3

4.2 BERT 算法流程

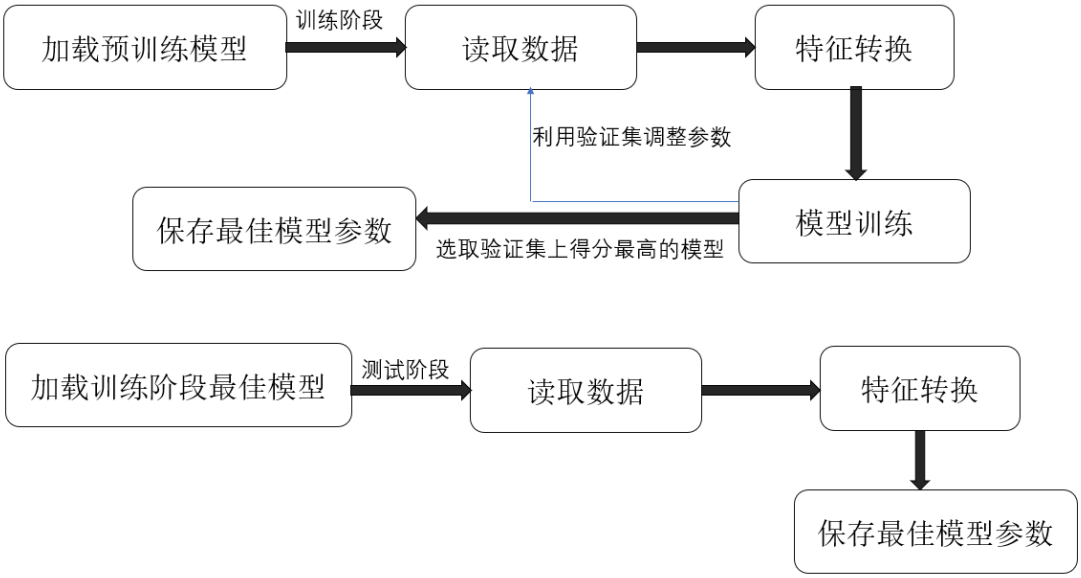

图 6 BERT 算法流程图

4.3 BERT 程序框架

该程序框架图如图 7 所示：

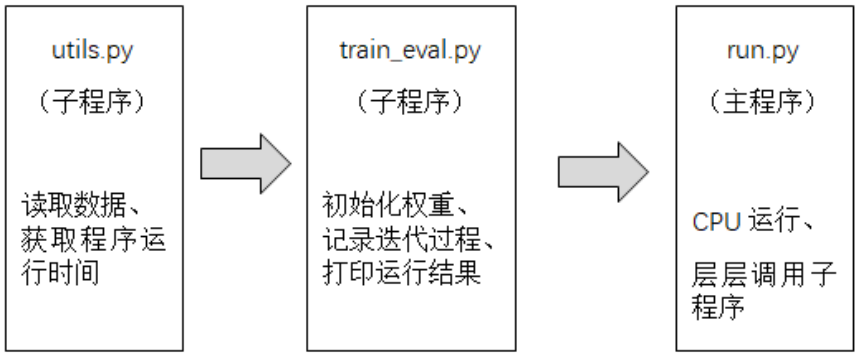

图 7 程序框架图

预训练语言 bert 模型放在 bert_pretain 目录下，目录下有 3 个文件。在 CPU 上运行 python run.py --model bert 并输出结果。运行 run.py 层层调用相应方法，同级目录下 utils.py 读取文件夹 Data 下划分好的数据集 dev.txt、test.txt、train.txt 以及获取程序运行时间，调用 models 文件夹下的 bert.py 进行训练，同时同级目录下 utils.py 获取程序运行时间，train_eval.py 进行初始化权重，记录迭代过程和损失函数并输出训练集、验证集、测试集上的准确率及程序运行时间，最后打印测试集的混淆矩阵与 Precision，Recall 和 F1-Score。

4.4 BERT 模型训练

在中文文本分类任务中，由于最佳超参数是特定于任务的，但是以下范围的可能值可以在所有任务中很好地工作。

表 4 模型主要参数

参数	值
Batch	16,32
学习率（rate）	5e-5，3e-5，2e-5
周期（epoch）	3，4

于是，分别调试以上参数，得出结果如表 4：

表 5

参数（batch；epoch；rate）			验证集准确率	测试集准确率
16	3	3e-5	89.79%	87.25%
16	4	2e-5	87.54%	89.56%
16	4	3e-5	89.45%	88.66%

16	4	5e-5	87.82%	86.92%
32	3	5e-5	87.54%	87.65%
32	3	3e-5	89.79%	87.25%
32	4	5e-5	88.55%	87.37%

根据表 4 所得出的结果，可以发现 rate 为 3e-5 在 BERT 模型上的验证集准确率最好，从而固定学习率不变。变化 epoch 参数，发现 epoch 为 3 时的验证集准确率最好，所以选定 epoch 为 3；那么，当改变 batch 参数时，发现所运行的结果一样，从而得出结果不受 batch 参数的影响。所以，可以确定验证集准确率为 89.79%，因此 BERT 模型准确率为 87.25%。

部分训练过程展示如下表：

表 6 BERT 模型的部分训练过程

```
stu07@ubuntu:~/q$ python run.py --model bert
Loading data...
5345it [00:00, 6248.74it/s]
1782it [00:00, 7906.68it/s]
1781it [00:00, 7853.59it/s]
Time usage: 0:00:01
Epoch [1/3]
Iter:      0,  Train Loss:   2.1,  Train Acc: 12.50%,  Val Loss:   2.0,  Val Acc: 12.91%,  Time: 0:00:05 *
Iter:    100,  Train Loss:  0.39,  Train Acc: 87.50%,  Val Loss:  0.62,  Val Acc: 81.20%,  Time: 0:00:35 *
Iter:    200,  Train Loss:  0.96,  Train Acc: 81.25%,  Val Loss:  0.54,  Val Acc: 82.83%,  Time: 0:01:01 *
Iter:    300,  Train Loss:  0.19,  Train Acc: 93.75%,  Val Loss:  0.46,  Val Acc: 85.35%,  Time: 0:01:25 *
Epoch [2/3]
Iter:    400,  Train Loss:  0.49,  Train Acc: 81.25%,  Val Loss:  0.43,  Val Acc: 86.87%,  Time: 0:01:51 *
Iter:    500,  Train Loss: 0.032,  Train Acc: 100.00%,  Val Loss:  0.37,  Val Acc: 88.78%,  Time: 0:02:17 *
Iter:    600,  Train Loss:  0.31,  Train Acc: 93.75%,  Val Loss:  0.41,  Val
```

```
Acc: 88.05%,  Time: 0:02:41
Epoch [3/3]
Iter:    700,  Train Loss: 0.067,  Train Acc: 100.00%,  Val Loss:  0.38,  Val
Acc: 88.44%,  Time: 0:03:05
Iter:    800,  Train Loss:  0.03,  Train Acc: 100.00%,  Val Loss:  0.38,  Val
Acc: 89.51%,  Time: 0:03:29
Iter:    900,  Train Loss:  0.13,  Train Acc: 93.75%,  Val Loss:  0.38,  Val
Acc: 89.62%,  Time: 0:03:53
Iter:   1000,  Train Loss:  0.23,  Train Acc: 93.75%,  Val Loss:  0.38,  Val
Acc: 89.79%,  Time: 0:04:17
Test Loss:  0.41,  Test Acc: 87.25%
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

          城乡建设      0.8589    0.8875    0.8730       391
          环境保护      0.8412    0.8314    0.8363       172
          交通运输      0.8696    0.8333    0.8511       120
          教育文体      0.9182    0.9238    0.9210       328
       劳动和社会保障      0.9313    0.8851    0.9076       383
          商贸旅游      0.7450    0.8462    0.7924       221
          卫生计生      0.9184    0.8133    0.8626       166

    accuracy                         0.8725      1781
   macro avg      0.8689    0.8601    0.8634      1781
weighted avg      0.8758    0.8725    0.8733      1781

Confusion Matrix...
[[347  16   5   3   1  19   0]
 [ 18 143   0   2   2   6   1]
 [ 12   0 100   1   2   5   0]
 [  5   2   2 303   4  12   0]
 [  8   4   0  18 339   8   6]
 [ 12   5   7   3   2 187   5]
 [  2   0   1   0  14  14 135]]

Time usage: 0:00:03
```

4.5 bert 模型的测试

通过对该模型的测试，得到的混淆矩阵如表 6：

表 7 BERT 模型的混淆矩阵

	真实类别						
预测类别	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**347**	16	5	3	1	19	0
1	18	**143**	0	2	2	6	1
2	12	0	**100**	1	2	5	0
3	5	2	2	**303**	4	12	0
4	8	4	0	18	**339**	8	6
5	12	5	7	3	2	**187**	5
6	2	0	1	0	14	14	**135**

5.bert 模型的评价

通常利用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。F-Score 越高，模型越稳健。BERT 模型的 F-Score 为 0.8645。

1、**Bert 模型的优点：**

（1）BERT 自带双向功能。

（2）为了获取比词更高级别的句子级别的语义表征，BERT 加入了 Next Sentence Prediction 来和 Masked-LM 一起做联合训练。

（3）为了适配多任务下的迁移学习，BERT 设计了更通用的输入层和输出层。

（4）微调成本小。

2、**BERT 模型的缺点：**BERT 模型主要是聚焦在针对字或者英文 word 粒度的完形填空学习上面，没有充分利用训练数据当中词法结构、语法结构，以及语义信息去学习建模。

6.模型的改进

针对 BERT 模型的不足，ERNIE 模型通过对训练数据中的词法结构、语法结构、语义信息进行统一建模，极大的增强了通用语义表示能力，使得模型对语义

知识单元的表示更贴近真实世界。在多项任务中取得了大幅度超越 bert 的效果。因此，本文将引入 ERNIE 模型，解决 BERT 模型的不足。

6.1 ENRT 模型的训练

表 8

参数（batch；epoch；rate）			验证集准确率	测试集准确率
16	3	3e-5	90.68%	90.51%
16	4	2e-5	90.91%	90.06%
16	4	3e-5	90.52%	89.78%
16	4	5e-5	21.55%	21.95%
32	3	3e-5	83.28%	84.11%
16	3	2e-5	90.24%	90.06%

根据表 7 所得出的结果，可以发现 rate 为 2e-5 在 ERNIE 模型上的验证集准确率最好，从而固定学习率为 2e-5 不变。变化 batch 参数，发现 batch 为 16 时的验证集准确率最好，所以选定 batch 为 3；那么，再调试参数 epoch=3，经过综合比较，当 batch=16，epoch=4，rate=2e-5 时的验证集准确率能达到 90.74%，这时评估模型最好，因此 ERNIE 模型准确率为 90.06%。

部分训练过程展示如下表：

表 9 ERNIE 模型的部分训练过程

```
stu07@ubuntu:~/q$ python run.py --model ERNIE
<pytorch_pretrained.tokenization.BertTokenizer object at 0x7feda3ec2128>
Loading data...
5345it [00:00, 6022.82it/s]
1782it [00:00, 8022.92it/s]
1781it [00:00, 8000.57it/s]
Time usage: 0:00:01
Epoch [1/4]
```

Iter: 0, Train Loss: 2.2, Train Acc: 12.50%, Val Loss: 2.1, Val Acc: 10.83%, Time: 0:00:05 *
Iter: 100, Train Loss: 0.74, Train Acc: 75.00%, Val Loss: 0.59, Val Acc: 80.30%, Time: 0:00:29 *
Iter: 200, Train Loss: 0.89, Train Acc: 68.75%, Val Loss: 0.42, Val Acc: 86.31%, Time: 0:00:53 *
Iter: 300, Train Loss: 0.26, Train Acc: 87.50%, Val Loss: 0.36, Val Acc: 88.33%, Time: 0:01:18 *
Epoch [2/4]
Iter: 400, Train Loss: 0.67, Train Acc: 75.00%, Val Loss: 0.34, Val Acc: 88.72%, Time: 0:01:39 *
Iter: 500, Train Loss: 0.072, Train Acc: 100.00%, Val Loss: 0.33, Val Acc: 89.11%, Time: 0:02:02 *
Iter: 600, Train Loss: 0.27, Train Acc: 87.50%, Val Loss: 0.33, Val Acc: 89.73%, Time: 0:02:25 *
Epoch [3/4]
Iter: 700, Train Loss: 0.058, Train Acc: 100.00%, Val Loss: 0.32, Val Acc: 90.07%, Time: 0:02:48 *
Iter: 800, Train Loss: 0.045, Train Acc: 100.00%, Val Loss: 0.31, Val Acc: 90.52%, Time: 0:03:10 *
Iter: 900, Train Loss: 0.067, Train Acc: 100.00%, Val Loss: 0.31, Val Acc: 90.91%, Time: 0:03:33 *
Iter: 1000, Train Loss: 0.24, Train Acc: 93.75%, Val Loss: 0.32, Val Acc: 90.74%, Time: 0:03:54
Epoch [4/4]
Iter: 1100, Train Loss: 0.015, Train Acc: 100.00%, Val Loss: 0.33, Val Acc: 90.35%, Time: 0:04:15
Iter: 1200, Train Loss: 0.04, Train Acc: 100.00%, Val Loss: 0.33, Val Acc: 90.40%, Time: 0:04:37
Iter: 1300, Train Loss: 0.023, Train Acc: 100.00%, Val Loss: 0.32, Val Acc: 90.63%, Time: 0:04:59
Test Loss: 0.33, Test Acc: 90.06%
Precision, Recall and F1-Score...

	precision	recall	f1-score	support
城乡建设	0.9138	0.8951	0.9044	391
环境保护	0.8619	0.9070	0.8839	172
交通运输	0.8689	0.8833	0.8760	120
教育文体	0.9354	0.9268	0.9311	328
劳动和社会保障	0.9263	0.9191	0.9227	383
商贸旅游	0.8493	0.8416	0.8455	221
卫生计生	0.8772	0.9036	0.8902	166
accuracy			0.9006	1781

```
      macro avg       0.8904    0.8967    0.8934       1781
weighted avg       0.9010    0.9006    0.9007       1781

Confusion Matrix...
[[350  14   9   4   6   8   0]
 [  9 156   0   2   2   1   2]
 [  6   2 106   0   0   6   0]
 [  3   2   1 304   9   9   0]
 [  4   2   1   9 352   3  12]
 [ 11   5   5   4   3 186   7]
 [  0   0   0   2   8   6 150]]
Time usage: 0:00:02
```

6.2 ENRT 模型的测试

通过对该模型的测试，得到的混淆矩阵如下表：

表 10 ERNIE 模型的混淆矩阵

	真实类别						
预测类别	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	350	14	9	4	6	8	0
1	9	156	0	2	2	1	2
2	6	2	106	0	0	6	0
3	3	2	1	304	9	9	0
4	4	2	1	9	352	3	12
5	11	5	5	4	3	186	7
6	0	0	0	2	8	6	150

6.3 ENRT 模型的评价

根据所得到的模型准确率，在该数据集上 ENRT 模型的准确率为 90.06%，模型的 F-Score 为 0.8935。相比于 BERT 模型，模型的准确率有所提高。

ERNIE 模型的优势在于：

（1）对实体概念知识的学习来学习真实世界的完整概念的语义表示。

（2）对训练语料的扩展尤其是论坛对话语料的引入来增强模型的语义表示能力。

第三章 热点问题挖掘

1.问题分析

根据附件 3 所给出的留言详情以及留言题，可以反馈出群众在某一时间段所反映的问题。首先，对留言主题与留言详情进行数据预处理，然后利用 jieba 分词对文本进行分词，并使用 TF-IDF 对文本进行向量化，再使用 DBSCAN 聚类。根据热度公式（热度指标=相同话题条数*0.5+赞同数*0.25+反对数*0.25）进行热度排名，并使用 TextRank 算法提取出每一类得分最高的留言主题作为热点问题描述，从而得到热点问题表以及热点问题留言明细表。

2.模型的建立

2.1 DBSCAN 模型简介

DBSCAN 是一种基于密度的聚类算法，其基本假设是一个集群的密度要显著高于噪声点的密度，因此，该方法的基本思想是对于集群中的每一个点，在给定的半径范围内，其相邻点的数量必须超过预先设定的某一个阈值。假设样本集 $D=(x_1,x_2,\cdots,x_m)$，则 DBSCAN 具体的密度描述定义如下：

Eps 领域（Eps-neighborhood）：对于一个点，记其 Eps 领域为 $N_{Eps}(p)$，则 $N_{Eps}(p)=\{q\in D\,|\,dist(p,q)\le Eps\}$，其中 $dist(p,q)$ 表示点 p 和 q 的距离。

直接密度可达（directly density-reachable）：称一个点直接密度可达点 q，如果满足以下条件：$p\cup N_{Eps}(q)$，$|N_{Eps}(q)|\ge MinPts$，其中，$MinPts$ 表示一个中心点的 Eps 领域必须包含的最小数量，需要事先确定。当点 p 和 q 都是一个集群的中心点时，则此时直接密度可达对两个点来说都是对称的。当 p 是边界点时，则此时直接密度可达是不对称的，如图 8 所示：

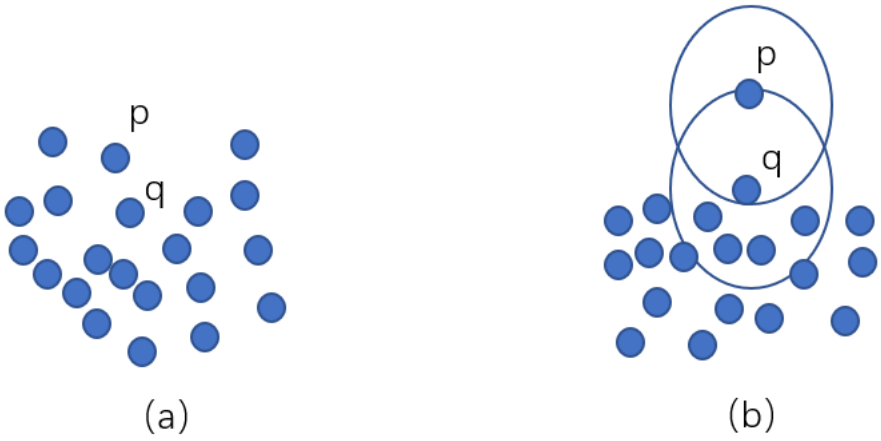

图 8

密度可达（*density – reachable*）：如果存在一串点 $p_1, p_2, \cdots, p_n; p_1 = q, p_n = p$，有 p_{i+1} 从 p_i 直接密度可达，那么就称 p 从 q 密度可达。

密度相连（*density– connected*）：如果存在一个点 o，使得 p 和 q 都从 o 密度可达，那么称点 p 和 q 密度相连。密度相连则是对称的。

类簇（*cluster*）：对于样本集 D，称子集 C 为 D 的一个类簇，如果其满足以下条件：①对于任意的点 p、q，如果 $p \in C$，并且 q 从 p 密度可达。② $\forall p, q \in C$, p, c都密度相连。那么，$q \in C$。

噪声（*noise*）：记 $C_1, C_2, \cdots, C_K$ 为数据集 D 中的 k 个簇，则噪声点则为 D 中那些不属于任何一个类簇的点，即 $noise = \{p \in D \mid \forall i : p \notin C_i\}$。

2.2 TextRank 算法

TextRank 是一种基于图的用于文本的排序算法，基本思想来自于 Google 的 PageRank 算法。可得到词语的排行、句子的排名，所以 TextRank 可以进行关键词提取，也可以进行自动文摘。通过文本分割成若干个组成单元（句子），构建节点连接图，用句子之间的相似度作为边的权重，通过循环迭代计算句子的 TextRank 值，最后抽取排名高的句子组成文本摘要。

3.模型的求解

3.1 数据预处理

由于附件 3 所提取的留言详情当中可能存在一些重复的留言，也可能存在一些对反映留言详情内容没有作用的字词，所以需要对其进行清洗，清洗后再分词，提取出重要的词进行特征提取，再进行对留言详情的聚类。留言详情预处理框架图如图 9 所示：

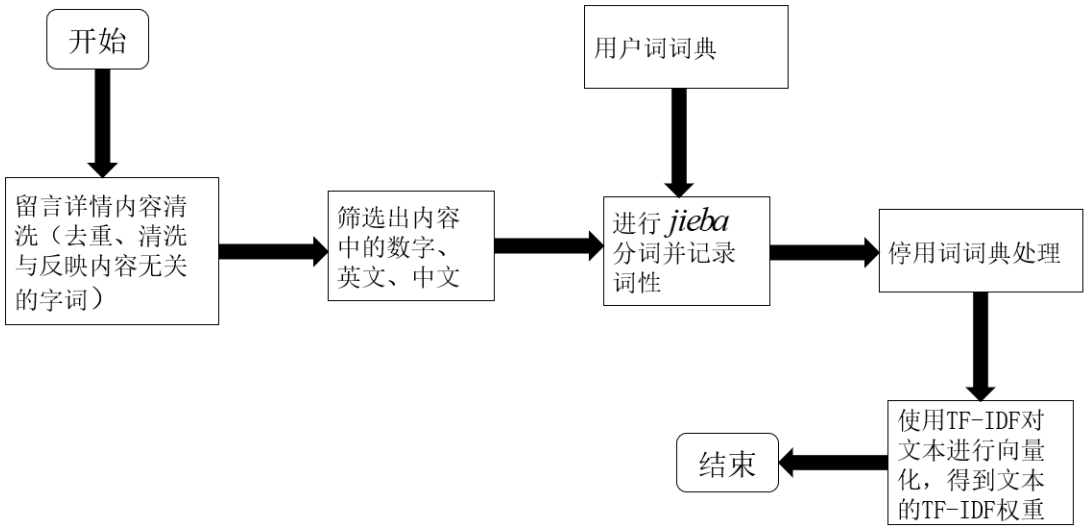

图 9 留言详情预处理

由上图可见，预处理过程可以分为三个部分，分别是：内容的清洗、分词、词汇提取以及特征提取。

1. 留言详情内容的清洗

观察从附件 3 中提取出来的留言详情，可以看到在留言详情当中，几乎每一条留言前面都有：“尊敬的 xxx”、“领导您好”等文字，留言详情中也有多处“我是网友 xxx”、“身份证号码*******”等文字，同时还存在一些网址、留言的出处等，为了清洗这些和反映问题的无关词汇，本文利用正则表达将其匹配并替换成空字符。对于文本聚类，标点符号会影响聚类的结果，所以本文筛选出留言详情中的数字、英文和中文。

2. 基于 *jieba* 分词的留言详情内容分词及重要词汇提取

需要清洗后的留言详情对进行分词操作，本文通过调用 *jieba* 词库对语料进行分词，并记录分词结果中每个词的词性。为了得到更好的聚类效果，所以本文在提取重要词汇的时候，只提取名词、动词、时间、地点等词性的词语，并且设置了停用词词典，把在停用词词典中不需要的词在分词后的词中去掉。

3.基于 TF-IDF 的留言详情内容特征提取

（1）TF-IDF 算法

TF-IDF（term frequency－inverse document frequency，词频-逆向文件频率)是一种用于信息检索(information retrieval)与文本挖掘(text mining)的常用加权技术。

TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

①TF 是词频（Term Frequency）：词频表示词条（关键词）在文本中出现的频率，计算公式为：$TF=\frac{\text{词在文档中的出现次数}}{\text{文档的总词数}}$。

②IDF 是逆文档评率，是衡量单词总体重要性的指标，计算公式为：$IDF=\log(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}})$，如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用（包含该词的文档数+1）代替。

③TF-IDF 实际上是 TF*IDF：某一特定文件内的高词语评率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。计算公式为：$TF-IDF=TF*IDF$。

（2）留言详情内容特征提取

前面得到了分词的结果，并对词性进行了记录，接下来可以针对不同词汇的词性，给予其 TF-IDF 权重以不同的乘数，这样可以突出某些类型的词汇的重要性，在一定程度上有助于聚类的效果。

3.2 算法流程图

1.DBSCAN 聚类过程

输入：样本集 $D=(x_1,x_2,\cdots x_m)$，邻域参数 $(\varepsilon,MinPts)$，样本距离度量方式

输出：簇划分 C

1）初始化核心对象集合 $\Omega=\varnothing$，初始化聚类簇数 $k=0$，初始化未访问样本集合 $\Gamma=D$，簇划分 $C=0$

2）对于 $j=1,2,\cdots,m$，按下面的步骤找出所有的核心对象：

a）通过距离度量方式，找到样本 x_j 的 ε- 邻域子样本集 $N_\varepsilon(x_j)$

b）如果子样本集样本个数满足 $|N_\varepsilon(x_j)|\geq MinPts$，将样本 x_j 加入核心对象样本集合：$\Omega=\Omega\cup\{x_j\}$。

3）如果核心对象集合 $\Omega=0$，则算法结束，否则转入步骤 4.

4）在核心对象集合 Ω 中，随机选择一个核心对象 o，初始化当前簇核心对象队列 $\Omega_{cur}=\{o\}$，初始化类别序号 $k=k+1$，初始化当前簇样本集合 $C_k=\{o\}$，更新未访问样本集合 $\Gamma=\Gamma-\{o\}$。

5）如果当前簇核心对象队列 $\Omega_{cur}=\varnothing$，则当前聚类簇 C_k 生成完毕，更新簇划分 $c=\{C_1,C_2,\cdots,C_k\}$，更新核心对象集合 $\Omega=\Omega-C_k$ 转入步骤 3。否则更新核心对象集合 $\Omega=\Omega-C_k$。

6）在当前簇核心对象队列 Ω_{cur} 中取出一个核心对象 o'，通过邻域距离阈值 ε 找出所有的 ε- 邻域子样本集 $N_\varepsilon(o')$，令 $\Delta=N_\varepsilon(o')\cap\Gamma$，更新当前簇样本集合 $C_k=C_k\cup\Delta$，更新未访问样本集合 $\Gamma=\Gamma-\Delta$，更新 $\Omega_{cur}=\Omega_{cur}\cup(\Delta\cap\Omega)-o'$，转入步骤 5。

输出结果为：簇划分 $c=\{C_1,C_2\cdots,C_K\}$

在计算过程中，使用“余弦相似度”来计算距离，所以本文在设置 ε 参数时，一般都设置为 0.3-0.5 之间，因为超过 0.5 的半径值会使不属于同类的留言内容聚类在一起，小于 0.3 则无法识别相同的留言内容。而在设置 $MinPts$ 参数时，可以人为假定一个阈值。

2. TextRank 算法过程

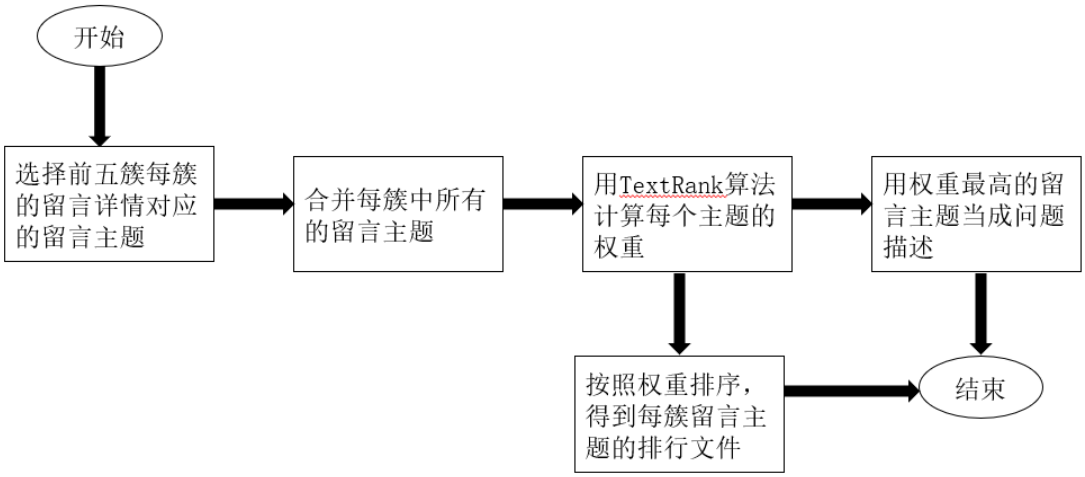

图 10 TextRank 算法流程

（1）预处理：分割原文本中的句子得到一个句子集合，然后对句子进行分词以及去停用词处理，筛选出候选关键词集。

（2）计算句子间的相似度：在本文中采用如下公式进行计算句子 1 和句子 2 的相似度：

$$\text{句子的相似度}=\frac{\text{两个句子都出现的词的数目}}{\log(\text{句子1中的词的数目})+\log(\text{句子2中的词的数目})}$$

对于两个句子之间的相似度大于设定的阈值的两个句子节点用边连接起来，设置其边的权重为两个句子的相似度。

（3）计算句子权重：

$$\text{句子1的权重}=(1-\text{阻尼系数})+\text{阻尼系数}*\sum_{\text{与句子1相连的所有句子}}\frac{\text{句子1和句子2的相似度}*\text{句子2的权重}}{\text{所有与句子2相连的句子的边的权重和}}$$

多次迭代计算直至收敛稳定之后可得各句子的权重得分。

（4）形成文摘：将句子按照句子得分进行倒序排序，抽取得分排序最前的几个句子作为候选文摘句，再依据字数或句子数量要求筛选出符合条件的句子组成文摘

3. 工作流程图

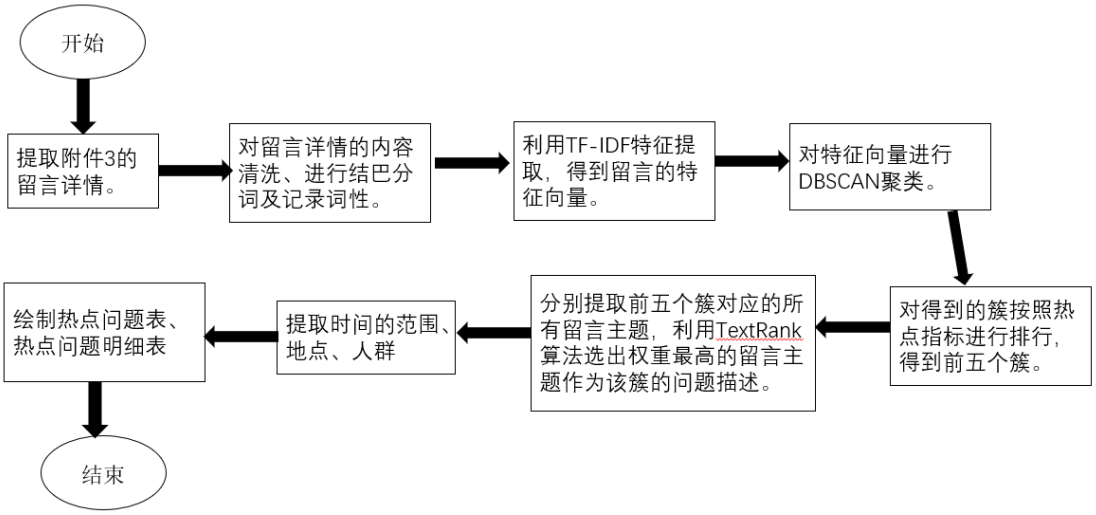

图 11 工作流程图

3.4 实验结果

1. 留言文本聚类结果

通过留言文本 DBSCAN 聚类的结果如表所示。

表 11 DBSCAN 聚类结果表

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	标签	标签类别
198750	A00024913	左家塘街道嘉园社区麻将馆	2019/12/6 22:30:56	馆.他经常营业到凌	0	0	43	43
201463	A00012757	城乐享健身突然关门，员工	2019/10/12 11:41:18	知突然关门，员工工	0	0	43	42
208841	A000111904	什么A市橘子洲的焰火又来了	2019/2/18 18:08:43	了。请求市委市政府	0	0	43	41
276954	A00044916	p随意冻结用户账户，公司没	2019/5/17 12:00:46	城购买了三个运动手	0	4	43	40
254536	A00023605	七辰三角洲入户楼道设计不合	2019/4/19 6:51:24	大型家具想要搬进户	0	0	42	39
265248	A00013897	共享汽车爆胎，告知要我赔	2019/6/28 15:29:59	超速，第三发现问	2	0	42	38
277066	A00075486	A7县时中路何时拉通至龟山	2019/11/4 21:01:23	到盼盼路已开始建设	0	2	42	37
289096	A00049370	公司成为A7县农业建设“全	2019/5/16 22:30:34	农业领域的“全能战	0	1	42	36
220576	A00040576	中贸城欺诈业主、拖欠业主资	2019/3/25 13:09:54	30之前退换全部款项	0	0	41	35
247302	A000911	少街道公共停车区域的僵尸车	2019/10/23 8:15:38	辆长期占据停车位	0	4	41	34
285601	A00016213	福安置小区门面经营汽车钣	2019/12/11 9:59:12	们周边的居民很多都	1	1	41	33
289350	A00083439	省现代雷德工程有限公司不按	2019/1/16 10:08:33	之前需要打匾铁到地	0	0	41	32
202119	A00086764	区西湖文化园偷钓者十分猖	2019/7/16 9:40:54	这时保安在何处呢，	0	0	40	31
245613	A00073278	A7县红树湾小区内安全隐患多	2019/8/12 9:21:41	电、公水，物业从未	0	0	40	30
259020	A00021812	盘湾13号房子还没协调好就	2019/11/29 10:56:25	。谁知天有不测风	0	0	40	29
262310	A00090026	村柏树组修路收取农民200元	2019/3/19 10:31:36	跃一条河，河上的	0	0	40	28
244528	A909235	景园滨河苑开发商强买强卖	2019-08-21 19:05:34	项目，但现在却要求	0	2	39	27
260766	A00042274	吾（梅溪新天地校区）店大服	2019/7/11 10:24:11	过去了，退款仍未到	0	0	39	26
270330	A00083527	省站首开，为什么没有直通	2019/5/20 7:44:49	加，特别在地铁3号	0	5	39	25
273867	A00083692	市明达中学附近快递点收费	2019/1/14 15:44:05	放，常常造成快递	0	0	39	24
275902	A000103972	管理部门人性化管理旅游客	2019/5/15 17:43:58	司去拿牌子，有的车	0	0	39	23

由表可见，通过 DBSCAN 聚类得到了 43 簇，每条留言都得到一个标签，标签为-1 的属于离群点，即为该留言反映的事件无法构成热点，而其他标签的留言则为一个热点。

2. 基于热点指标的总体留言的热点排行

在本文中利用 DBSCAN 聚类已经获得了每条留言所在的簇，根据热点指标（热度指标=相同话题条数*0.5+赞同数*0.25+反对数*0.25），从而获得每个簇留言的排行，将权重最高的留言所在的簇当成最热热点，以此类推，筛选的出前 5 名热点。如下表：

表 12 热点排行表

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	标签
202689	A000107068	映A市购第二套房的困	2019/12/17 18:52:17	享受该政策，A市和A7县	0	0	19
207472	A000102290	人行横道问题，建议	2019/1/30 10:08:10	隐患。并且在人流量最	0	0	19
210267	A00049584	请A7县教育局解答下	2019/3/2 11:45:53	育局人社局有大概的工作	0	3	19
214247	A00015366	小区一空调水处理设	2019/7/2 10:56:59	湖第二小学，影响到学生	0	0	19
221732	A00032672	的星沙周末有戏”演出	2019/12/3 14:52:45	传统花鼓戏偏多，建议多	0	8	19
225421	A00031618	重视物价过快上涨的	2019/4/8 12:51:48	全国垫底，百姓生活及	0	0	19
234821	A00051742	云花园岭社区交通出	2019/8/4 10:37:37	，没有栏杆，极不安全，	0	2	19
245119	A00031618	加快严查学校违规上课	2019/6/26 16:15:13	也没有休息地方且还不让	0	0	19
252786	A000112191	万境水岸小学堵车，	2019/5/16 9:40:22	安排家长义工在学校马路	0	3	19
256990	A00074594	乔庄巷摊贩长期占道	2019/6/19 10:19:07	堵在路上车子都不好进	0	0	19
260119	A0005325	映A7县时代星城小区乱	2019/6/4 11:22:23	小区的混乱/物业的不	0	3	19
266095	A00087773	局是否有职业培训的	2019/5/29 14:03:57	就业中心负责具体培训	0	0	19
268169	A000108008	加强景秀路流动活禽销	2019/4/6 17:59:09	染H7N9禽流感疫情的省	0	0	19
275761	A00013807	网上服务控件升级导致	2019/8/28 9:46:57	用，对我的工作影响极大	0	0	19
276125	A000101616	更新项目六标段施工噪	2019/5/20 17:37:15	1点仍然在施工，严重影	0	0	19
281831	A00052150	育津贴和医院生产费	2019/4/19 9:19:55	工作单位联系时又让我	0	0	19
283482	A909232	城小区附近搅拌站的一	2019-12-07 09:21:32	半站,生产产生了灰尘和	0	0	19
283862	A00097212	2区步行街市中心的老	2019/8/4 1:54:05	改造部正在将育英街部	0	0	19
285783	A00084296	社区还有几栋是用公	2019/12/30 12:07:45	小区在搞电梯改造，可	0	0	19
287492	A0003274	县星沙实验小学的几	2019/9/3 11:06:25	老师的车辆还要在本已拥	0	0	19
287566	A00076390	B2栋3单元羽婕钢材扌	2019/6/25 9:28:50	一楼门面管理及差，但	0	0	19

3. 基于 TextRank 算法的热点内部留言主题排行

热点内部的留言主题排行结果如下面各图所示：

第一热点标题排行结果：

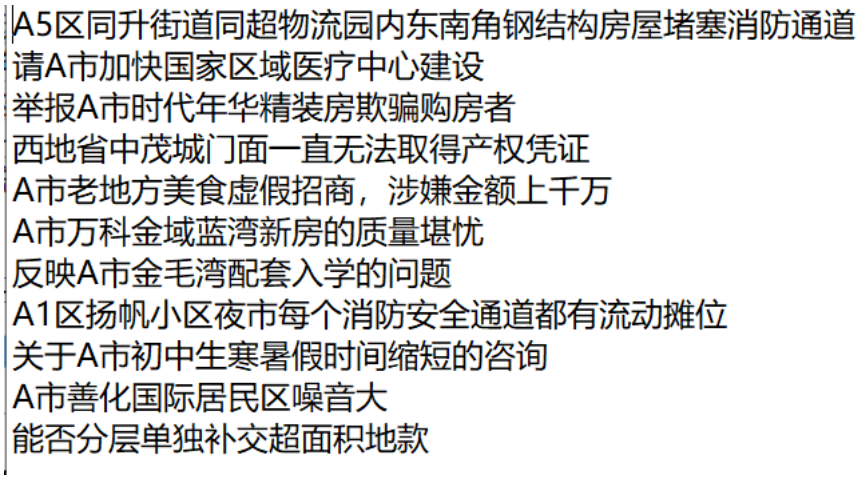

图 13 第一热点标题排行结果

第二热点标题排行结果：

2019年教师招聘即将开始，请A7县教育局解答下广大网友热切关注的问题
A3区岳北社区B2栋3单元羽婕钢材批发噪音太大
A1区西学巷有机更新项目六标段施工影响考生休息学习
A3区大道人行横道问题，建议早日解决
A3区望月湖小区一空调水处理设备制造噪音
请A市重视物价过快上涨的问题
A2区暮云花园岭社区交通出行不便
请修缮A2区步行街市中心的老旧房屋
反映A7县时代星城小区乱象
A5区砂子塘万境水岸小学堵车，交通不安全
请A市加快严查学校违规上课时间
对A7县“雅韵星沙周末有戏”演出活动的建议
A市普通上班族的生育津贴和医院生产费一分都不能报销吗

图 14 第二热点标题排行结果

第三热点标题排行结果：

请求解决A7县圣力华苑小区房屋两证合一的问题
西地省津楚投资责任有限公司资金链断裂，投资者血本无归
A2区南湖路森宇家园餐馆油烟扰民
投诉A市保险职业学院校园一卡通捆绑消费联通卡
A7县利保香槟国际协同中介公司违规收取业主服务费
A8县市大屯营乡居民家房屋后架设了一条220万伏的高压线
A7县未来漫城物业不作为，还能评为五星级
对西地省高速建设开发总公司工资制度的质疑

图 15 第三热点标题排行结果

第四热点标题排行结果：

A市A3区含浦镇含浦街道芝字港村可以开通公交车吗
反映A市地铁3号线松雅西地省站地下通道建设问题
A市物业公司进入企业老旧职工小区，业主有无选择权与知情权
寻找A7县退伍人员的下落
A3区观沙岭岳北安置区的露天菜市场阻塞消防通道
A市楚雅二医院正式员工无购房资格
A市在水一方大厦人为烂尾多年，安全隐患严重
A2区桂花坪街道农民的安置地到底安置了谁

图 16 第四热点标题排行结果

第五热点标题排行结果：

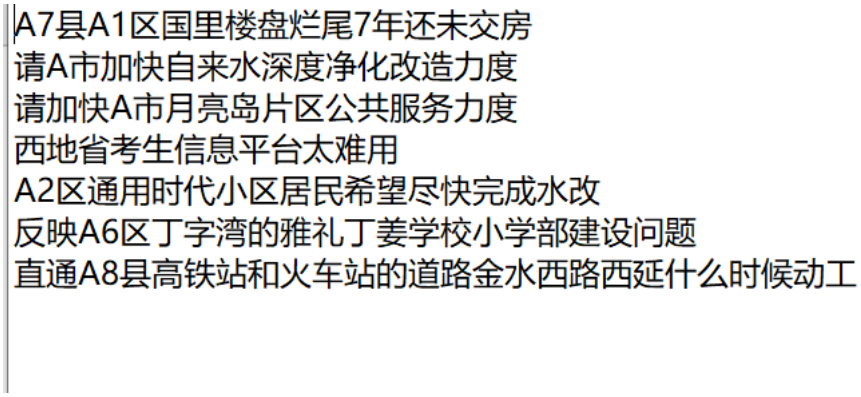

图 17 第五热点标题排行结果

4. 热点问题留言明细表

表 13 热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	200944	A00025724	中茂城门面一直无法取得产	2019/6/4 14:04:25	答复是含糊不清，一下说是2019年	0	0
1	213101	A00023847	方美食虚假招商，涉嫌金额	2019/12/16 21:58:15	直没有结果，2019年12月16日我们	0	0
1	214672	A00096117	万科金域蓝湾新房的质量堪	2019/5/10 17:21:25	通电后，客厅空调柜机插座至今没	0	0
1	217938	A000106629	市初中生寒暑假时间缩短自	2019/7/18 17:40:52	更是直接缩水半个月，7月14才正	0	1
1	218648	A00062710	区夜市每个消防安全通道都	2019/8/22 23:39:25	动摊位，车都进不了？希望领导严	0	0
1	219838	A00020614	A市善化国际居民区噪音大	2020/1/3 16:48:00	的货车经过减速带时发出哐哐的异	0	0
1	223297	A00087522	映A市金毛湾配套入学的问	2019/4/11 21:02:44	毛湾楼盘纳入配套入学，A3区教育局	1762	5
1	238397	A00031618	市加快国家区域医疗中心建	2019/9/9 13:45:52	区域医疗中心城市建设实现医疗配	0	0
1	257402	A00074795	A市时代年华精装房欺骗购	2019/5/15 8:30:29	。公摊每户就有三万八千元，在看	2	0
1	259907	A00061126	超物流园内东南角钢结构房	2019/12/12 13:53:55	筑后面就是美食街旁边就是仓库。	0	0
1	337458	A078325	否分层单独补交超面积地款	2019-09-16 18:48:29	地，得知该房占地属划拨用地而且	0	0
2	202689	A000107068	反映A市购第二套房的困难	2019/12/17 18:52:17	可以享受该政策，A市和A7县都属于	0	0
2	207472	A000102290	道人行横道问题，建议早日	2019/1/30 10:08:10	安全隐患。并且在人流量最大的几	0	0
2	210267	A00049584	始，请A7县教育局解答下广	2019/3/2 11:45:53	，教育局人社局有大概的工作计划	3	0
2	214247	A00015366	月湖小区一空调水处理设备	2019/7/2 10:56:59	望月湖第二小学，影响到学生上课	0	0
2	221732	A00032672	雅韵星沙周末有戏”演出活	2019/12/3 14:52:45	。3.传统花鼓戏偏多，建议多考虑	8	0
2	225421	A00031618	A市重视物价过快上涨的问	2019/4/8 12:51:48	却是全国垫底，百姓生活及其艰难	0	0
2	234821	A00051742	暮云花园岭社区交通出行	2019/8/4 10:37:37	高筑，没有栏杆，极不安全，有时	2	0
2	245119	A00031618	市加快严查学校违规上课时	2019/6/26 16:15:13	学校也没有休息地方且还不让学生	0	0
2	252786	A000112191	塘万境水岸小学堵车，交	2019/5/16 9:40:22	都要安排家长义工在学校马路边上	3	0
2	256990	A00074594	区乔庄巷摊贩长期占道经	2019/6/19 10:19:07	吵闹，堵在路上车子都不好通行，	0	0
2	260119	A0005325	反映A7县时代星城小区乱象	2019/6/4 11:22:23	所述，小区的混乱/物业的不作为	3	0
2	266095	A00087773	人社局是否有职业培训的项	2019/5/29 14:03:57	学校或就业中心负责具体培训。符合	0	0
2	268169	A000108008	应加强景秀路流动活禽销售	2019/4/6 17:59:09	过人感染H7N9禽流感疫情的省会城	0	0

5. 热点问题表

表 14 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	448	2019/4/11至2020/1/3	A5区同升街道同超物流园	园内房屋堵塞消防通道
2	2	15.75	2019/1/30至2019/12/30	A3区大道	大道人行横道出现问题，建议早日解决
3	3	11.25	2019/1/3至2019/11/21	A7县圣力华苑小区	请求小区房屋两证合一的问题
4	4	8	2019/1/4至2019/10/12	A市A3区含浦镇含浦街道芝字港村	村询问是否可以开通公交车
5	5	7	2019/1/16至2019/9/30	A6区丁字湾雅礼丁姜学校	学校小学部建设有问题

4.模型的评价

4.1 DBSCAN 模型的评价

1. DBSCAN 的主要优点有：

（1）DBSCAN 可以发现任意形状的簇，对噪声不敏感，可以自动发现簇的数量。

（2）计算速度快，对内存占用小，适用于大型数据集。

2.DBSCAN 的主要缺点有：

（1）如果样本集的密度不均匀、聚类间距差相差很大时，聚类质量较差，这时用 DBSCAN 聚类一般不适合。

（2）如果样本集较大时，聚类收敛时间较长。

（3）需要对距离阈值和邻域样本数阈值联合调参，不同的参数组合对最后的聚类效果有较大影响。

第五章 答复意见的评价

1.问题分析

根据附件4所给出的数据，按照质量评分指标（质量评分=0.7*文本相关性+0.1*时间间隔标准化+0.1*平均字符长360）对答复意见的质量进行评价。首先，提取出附件4的“留言详情”、“答复意见”进行数据的预处理、利用gensim模型计算“留言详情”、“答复意见”两者之间的文本相似度，再计算答复的时间间隔。通过计算质量评分，得到高、中、低质量的评价。

2.模型的建立

2.1 gensim 算法的简介

gensim（generate similarity）是一个简单高效的自然语言处理 python 库，用于抽取文档的语义主题（semantic topics）。gensim 的输入是原始的、无结构的数字文本（纯文本），gensim 算法通过在语料库的训练下检验词的统计共生模式(statistical co-occurrence patterns)来发现文档的语义结构。这些算法是非监督的，也就是说只需要一个语料库的文档集。当得到这些统计模式后，任何文本都能够用语义表示(semantic representation)来简洁的表达，并得到一个局部的相似度与其他文本区分开来。

2.2 建立质量评价指标

本文建立了有关文本相似度、时间间隔的质量评分标准，对其进行加权求和，得到答复留言质量的评价方案，公式为：

质量评分=0.7*文本相关性+0.1*时间间隔标准化+0.1*平均字符长 360

（1）答复意见和留言详情中反馈的问题是否相关，回复意见是否能够解决民众反馈的问题。与留言详情越相关的答复越有可能成为优质答复。运用 gensim 算法计算文本相。

（2）答复的时间是否及时，及时的答复能够体现问政的效率高低，答复时间与留言时间的时间间隔也是评价答复质量的指标，其公式为：

答复时间间隔=答复时间-留言时间

并对数据进行 Min-max 标准化，利用公式：

$$新数据=\frac{（原数据-最小值）}{（最大值-最小值）}$$

（3）答复的完整性：本文采取答复意见文本长度作为衡量完整性的指标，如果答复字数过少，则应是一个应付性答复，这可能是一个低质量答复。

3.模型的求解

3.1 工作流程图

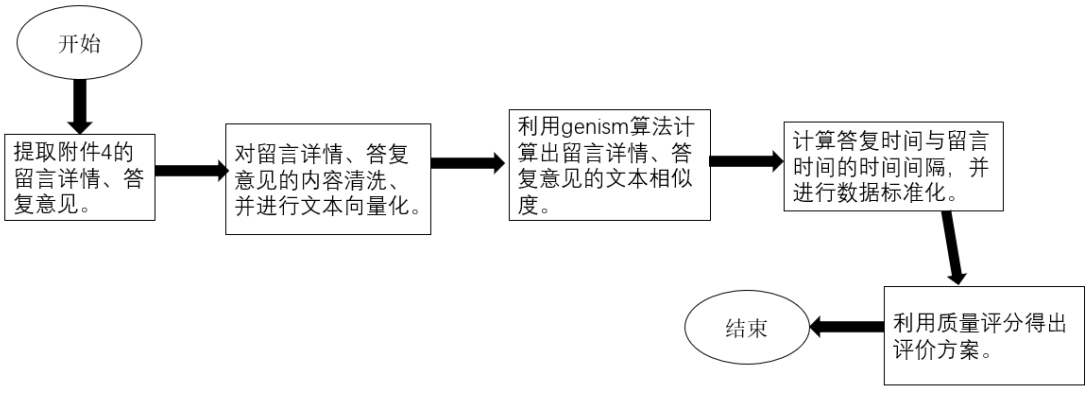

图 18

3.2 实验结果

得出部分结果结果如下表：

表 15

留言详情	答复意见	答复时间	相差天数	天数标准化			质量评分
业公司却以交20万保证金，不	收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	15	0.50	0.06	59.25	77.53
面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	15	0.50	0.00	0.00	36.05
同时更是加大了教师的工作	办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	15	0.50	0.00	0.00	36.05
落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	15	0.50	0.00	0.00	36.05
“马坡岭小学”，原“马坡岭	采留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	16	0.53	0.01	9.34	42.59
再把泥巴冲到右边，越是上下	于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	1	0.03	0.01	11.57	44.10
为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	10	0.33	0.02	24.34	53.07
好远，天寒地冻的跑好远，目	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	29	0.97	0.02	21.27	50.98
也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	16	0.53	0.00	0.00	36.05
立交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	16	0.53	0.01	7.35	41.20
规划局审批通过《温室养殖大	支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	9	0.30	0.02	19.06	49.38
区安置房地下室近两万平方米	续，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	0	0.00	0.01	8.84	42.19
逢，大量从小区开车出去的业	分局配合进行具体选址，招标（邀标）进行	2019/1/14 14:34:58	16	0.53	0.02	18.46	48.98
贵省相关政府部门的大力支持	映的相关警情，已由银盆岭派出所立刑事案	2019/1/3 14:03:07	6	0.20	0.03	29.81	56.89
小时以上！天寒地冻，其他公	常。由于驾驶员工作时间长，劳动强度大，	2019/1/14 14:33:17	17	0.57	0.01	9.59	42.77
址：https://baidu.com/。	的“披塘路路口两端各拆除20米中间花坛，	2019/3/6 10:26:14	7	0.23	0.01	7.02	40.94
便以各种理由拒绝退货，并	根据您提供的信息进行投诉信息的登记分送	2019/1/3 14:02:47	7	0.23	0.01	7.92	41.57
称。建议在艺术中心先期借	营业。梅溪湖二期金菊路与雪松路东南角规	2019/1/14 14:32:40	19	0.63	0.01	7.32	41.19

高质量答复：质量评分在 75 分以上。

中质量答复：质量评分 60 分以上。

低质量答复：质量评分低于 60 分。

参考文献

[1]胡鹏辉.基于多模型的问答社区答案质量评价研究[D].南京：南京师范大学，2019.
[2]宗成庆.统计自然语言处理[M].北京：清华大学出版社，2008.
[3]游丹丹，陈福集.我国网络舆情热点话题发现研究综述[J].现代情报，2017，(37):165-171.
[4]段学睿.研究热点的抽取和演变[D].太原：太原理工大学.

附录

详细代码以文件见附加文件。