C 题

摘要：

智慧政务系统凭借其结合人工智能、大数据、云计算等先进技术的优势，与高效、快捷、便利的特点，成为社会发展的新亮点，有助于搭建更贴近民生、更关注热点的线上问政平台。

本次建模目标是利用已有留言数据，分别作以下挖掘与分析：

问题 1，对文本数据进行中文分词、除停用词等预处理后，使用 TF-IDF（Term Frequency-Inverse Document Frequency 词频-逆文档频率）进行特征提取。根据提取的关键词，对留言进行分类。

问题 2，根据题目，热点问题归类的核心在于特征提取与问题聚类，使用 LDA（Latent Dirchlet Allocation）模型训练，根据聚类结果将留言归类。根据留言的点赞数、反对数及对应问题下留言条数，制定热点问题评价方案，筛选出排名前五的热点问题。

问题 3，从答复的完整性、相关性、可解释性三个角度，定义答复意见的评价方案，从而评估实际答复意见内容，并作出打分。

关键词：文本挖掘 特征提取 TF-IDF 算法 LSTM 模型 LDA 模型

Abstract

The intelligent government system has become a new highlight of social development and also is conductive to build an online political platform wich is closer to people's livelihood and more concerned with hotspots. This achievement is based on the advantages of advanced technologies such as artificial intelligence, big data and cloud computing, and the features of high efficiency, speed and convenience.

The goal is to use the existing message data to mine and analyze the following:

Aiming at the issue of the first, after preprocessing the text data with Jieba Chinese word segmentation tools, TF-IDF (Term frequency-inverse Document Frequency) is used to extract the key words from which the comments were classified.

Aiming at the issue of the second, according to the theme, the core of hotspots classification lies in key words extraction and problem clustering. The LDA (Latent Dirchlet Allocation) model was used for training. The messages were classified according to the clustering results. According to the number of likes, the number of dislikes and the number of messages under the corresponding topics, the hot spots evaluation scheme was developed. Based on it, the top five hot spots were selected.

Aiming at the issue of the third, from the perspectives of the completeness, relevance and interpretability of the replies, the evaluation scheme is defined to evaluate the content of the replies and rank them.

Keywords: text mining features extraction TF-IDF algorithm LSTM model LDA model

目录

1. 挖掘目标

本次建模目标是利用已有留言数据，基于 Jieba 中文分词工具，TD-IDF（Term Frequency-Inverse Document Frequency 词频-逆文档频率）算法、深度学习 LSTM 模型、LDA（Latent Dirchlet Allocation）模型，实现以下目标：

（1）根据留言主题和详情进行挖掘分析，对应归类到不同主题下；

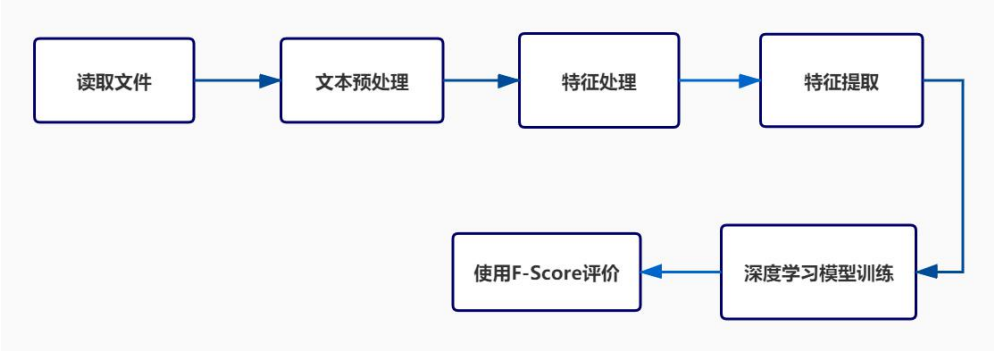
（2）基于已有留言数据，制定热点问题评价方案，捕捉热点并集中提取相关留言；

（3）从完整性、相关性、可解释性定义答复意见评价方案，根据规范考察答复意见，并进行质量评分。

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 文本预处理

2.1.2.1 对留言进行中文分词

由于计算机无法直接处理非结构化的文本信息，本文采用中文分词模块——jieba 分词，对留言详情文本进行切分，转换为结构化信息。

Jieba 分词过程如下：

（1）去除特殊字符，通过 Trie 树分词模型，建立 DAG 中文分词有向无环图；

（2）使用动态规划法查找最大概率路径，找出基于前缀辞典的词频最大切分组合，对于登录词，按字典标注标识；

（3）对于未登录词，首先使用 Token 识别中文和英文部分，中文部分加载 HMM 概率模型图，基于 Viterbi 算法动态规划取得分词和标注。

为了提高后续特征处理与提取的效率，节省存储空间，方便文本挖掘与分析，去除分词结果中的标点符号，并过滤停用词。

2.1.3 特征处理与提取

实现留言分类，需要提取留言详情中的关键信息。因此，接下来针对详情内容进行特征处理和特征提取。

2.1.3.1 特征处理

考虑到人们日常书写表达习惯，超高频词（词语出现频率过大）和超低频词（词语出现频率过小）含有文段关键信息的概率较低。因此在不改变留言核心内容的前提下，本方案统计详情内容各个词语的出现频率，除去超高频词和超低频词，尽量减少要处理的词语数量，简化计算复杂度，提高效率。

2.1.3.2 特征提取

自然语言挖掘与处理中，特征提取常用方法包括 TF-IDF（Term Frequency-Inverse Document Frequency 词频-逆文档频率）、textRANK 等，本文采用 TF-IDF 计算词频，提取各条留言详情中的关键词。

TF-IDF 原理如下：

词频（Term Frequency, TF），表示某个特定词语 t 在给定文档 d 中出现的频率。当频率 TF 值越高，说明这个词与文本内容关系越密切，相反 TF 值越低，则说明词 w 在文本中越不重要。TF 计算公式为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

逆向文件频率（Inverse Document Frequency, IDF）的大小与包含词 t 的文档多少有关。若包含词 t 的文档越多，则 IDF 越小；相反包含词 t 的文档越少，则 IDF 越大。从而可排除“了”、“你”、“我”等这类几乎不含关键信息的常用中文词语。

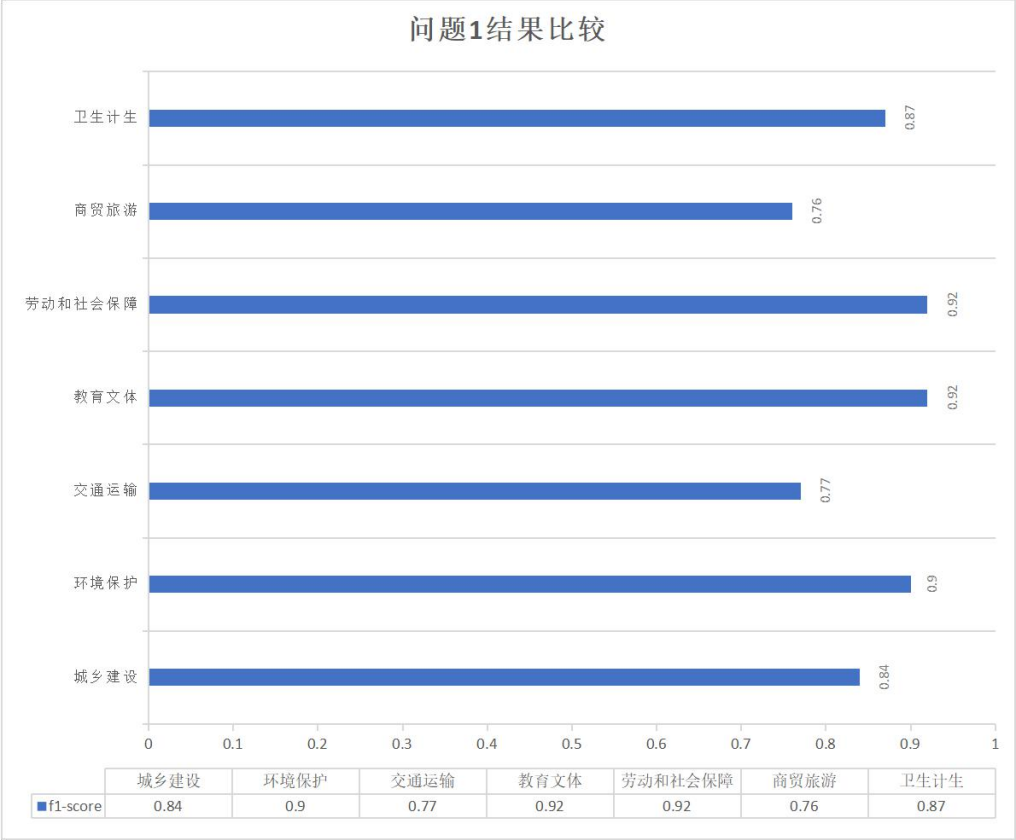
2.1.4 深度学习模型训练

在自然语言挖掘与文本分析中，常用的特征提取器有三类：RNN（Recurrent Neural Network，递归神经网络）、CNN（Convolutional Neural Networks，卷积神经网络）、Transformer。这里使用 RNN 中的 LSTM（Long-short Term Memory）模型。

2.1.5 结果评价

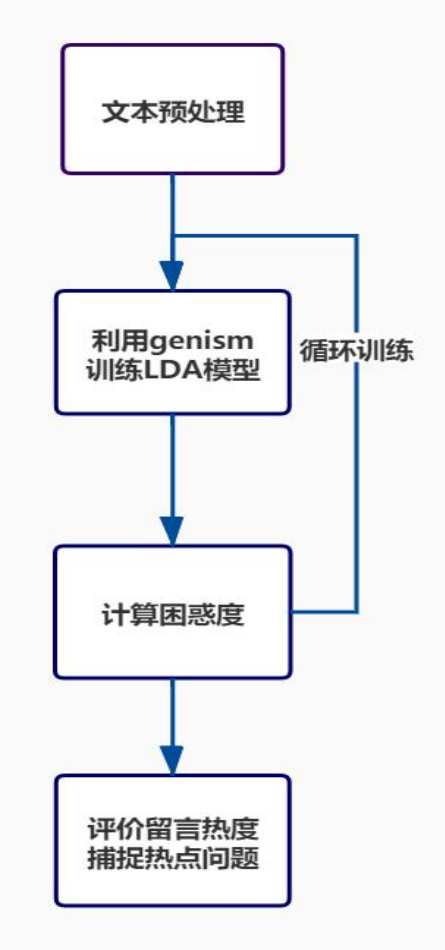
比较原始数据提供的分类结果与预测结果，评价这一分类方法。从下表可得（原始分类和预测结果图）

一级标签	precision	recall	f1-score	support
城乡建设	0.85	0.82	0.84	200
环境保护	0.91	0.88	0.9	102
交通运输	0.88	0.68	0.77	73
教育文体	0.92	0.93	0.92	170
劳动和社会保障	0.90	0.94	0.92	172
商贸旅游	0.71	0.8	0.76	114
卫生计生	0.88	0.87	0.87	90
Accuracy			0.86	921
Macro avg	0.86	0.85	0.85	921
Weighted avg	0.87	0.86	0.86	921



2.2 问题 2 分析方法与过程

2.2.1 流程图



2.2.2 文本预处理

对附件 3 文本数据进行预处理，包含中文分词、去停用词、做词性筛选等处理。将文本数据转为结构化信息。

2.2.3 LDA 模型训练

在自然语言挖掘与分析中，LDA（Latent Dirchlet Allocation）是一种主题模型，也被称为三层贝叶斯概率模型，即文档层、主题层和词层三层结构，常用于特征提取。LDA 模型生成流程如下：

（1）输入：文档集合 D。设文档集合为 D，$D=\{\mathrm{d}_1,\mathrm{d}_2,...,\mathrm{d}_\mathrm{n}\}$，主题集合为 T，$T=\{\mathrm{t}_1,\mathrm{t}_2,....,\mathrm{t}_\mathrm{k}\}$。D 中含有的词组成字典 DIC，$DIC=\{\mathrm{w}_1,\mathrm{w}_2,...,\mathrm{w}_\mathrm{m}\}$ 设字典含有 m 个单词；

（2）对应单个文档 d 对应每个主题的概率为 $\theta_d=\{p_{t_1},p_{t_2},....p_{t_k}\}$，对应每个主题 t，生成不同单词的概率为 $\omega_w=\{p_{w_1},p_{w_2},....,p_{w_m}\}$，初始对 θ_d 和 ω_w 随机赋值；

（3）以主题层作为中间层，根据 θ_d 和 ω_w 计算得出文档 d_i 出现单词 w_j 的概率。计算公式如下：

$$P(w_j \mid d_i) = P(w_j \mid t_z) * P(t_z \mid d_i)$$

（4）枚举主题 T 中的元素 t_z，计算并排序所有 $P(w_j \mid d_i)$，取最大值对应的 t_z。若词语 w_j 选择的主题 t 与初始主题不同，则对应的，θ_d 和 ω_w 也会发生变化；

（5）在 D 所有元素 d 中的词语 w 重复迭代（3）（4），直至 LDA 模型收敛。

2.2.4 热点评价与捕捉

在某一时间范围内，市民集中反映、关注的问题应是当下社会治理的焦点。我们以此为标准，制定对应的热点评价方案。本文以留言的点赞数、反对数和该类热点下留言数量之和的大小，决定该留言是否为热点问题。结果如下：

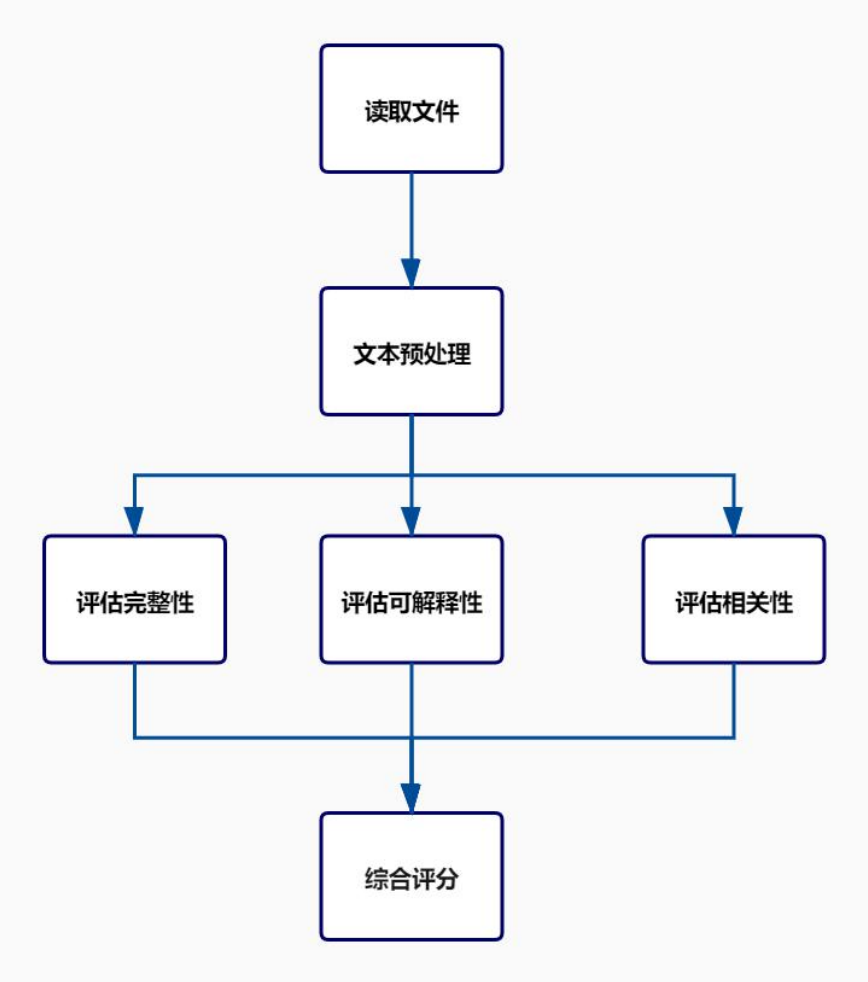
热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	2097	2019/8/19 11:34	A市A5区汇金路五矿万境	A市A5区汇金路五矿万境K9县存在一系列问题
2	2	1762	2019/4/11 21:02	梅溪湖金毛湾的一名业主	A市金毛湾配套入学的问题
3	3	790	2019/2/25 9:58	58车贷诈骗案受害者	严惩A市58车贷特大集资诈骗案保护伞
4	4	821	2019/2/21 18:45	A市A4区58车贷诈骗案受	请书记关注A市A4区58车贷案
5	5	733	2019/3/1 22:12	58车贷诈骗案受害人	承办A市58车贷案警官应跟进关注留言

根据留言的热度指数，按降序排列，提取热度前五的问题明细表。

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	208636	A00077171	A市A5区汇	2019/8/19 11:34	市 区汇 金路	2097	0
2	223297	A00087522	反映A市金	2019/4/11 21:02	书记 先生 您好	1762	5
3	217032	A00056543	严惩A市58	2019/2/25 9:58	胡 市长 您好	790	0
3	217032	A00056543	严惩A市58	2019/2/25 9:58	胡 市长 您好	790	0
4	220711	A00031682	请书记关注	2019/2/21 18:45	尊敬 胡书记 您	821	0
5	194343	A00010616	承办A市58	2019/3/1 22:12	胡书记 您好 5	733	0

2.3 问题 3 分析方法与过程

2.3.1 流程图



2.3.2 评价方案定义

一个高效惠民的网络问政平台除了实现分类留言，传派至对应政府职能部门处理外，也需要实现针对相关部门回复的评估功能，促进构造有参与、有反馈的线上问政氛围，市民与政府部门共同营建和谐法治社会。评价方案从以下三个角度进行评分，包括答复意见的完整性、相关性与可解释性。

2.3.2.1 完整性评估

是否具有完整性是衡量答复意见是否高质量的首要标准。对答复意见的完整性评估主要考察是否满足以下规范，并作出评价。

完整性规范与对应评分如下：

（1）具体回复日期，分值 0.2；

（2）回复内容，包括：

①表明收到，“收悉”、“收到”，分值 0.1；

②表明答复，“回复如下”、“答复如下”，分值 0.1；

③表达问候，“市民”、“网民*”、“你好”，分值 0.1；“您好”，分值 0.2;

④表达感谢，“感谢”、“谢谢”，分值 0.2。

2.3.2.2 相关性评估

对于广大网民而言，相关部门是否对留言作出相关答复，无疑是网民最关注的焦点。因

此，答复意见内容是否与留言相关，是评价答复质量是否过关的重要标准。

评估过程如下：

（1）文本预处理。对留言详情和答复意见进行去符号处理

（2）使用 text2vec 工具计算文本相似度：

①在答复意见中随机抽取八条，每四条一组，作为两组参照意见；

②使用 BM25 算法，分别计算留言主题与答复意见、任意一组参照意见的相似性；

③比较得分。若答复意见得分最高，则答复与留言主题相关。若答复意见得分不是最高，则选择另一组参照意见，重复步骤②。重复步骤②后，若答复得分最高，则与留言主题相关；否则答复意见与留言主题不相关。

（3）答复意见与留言主题相关，得分；若二者不相关，则不得分。

2.3.2.3 可解释性评估

可解释性，即答复意见是否对内容作出相关解释，是否指出对应规定或法律条文。

评估过程如下：

（1）检测是否含有书名号

（2）检测是否含有“根据”、“按照”、“依据”等类似词；

（3）检测是否含有“已转”、“转由”、“转交”等类似词。

（4）若以上规则满足任意一条或多条，则认为该答复具有可解释性，得分分值 1；否则，得分为 0。

2.3.3 综合评价

根据以上三类评估，整合得分作为该条答复意见的质量得分，并根据得分按降序排列。部分评价结果如下：

	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	intergrity	interpretability	答复相关性	答复质量
1420	39514	UU0081705	希望B9市4	2019/8/26 16:36	玉瓷俪景小区近60	网友：您好！您反映的“	2019/10/14 16:02	1	1	1	3
503	10724	UU0081866	希望A市的	2015/12/30 16:23	我是一名在外	网友“UU0081866” 您	2016/1/7 10:43	1	1	1	3
558	12708	UU0081902	呼吁请求解	2013/9/3 17:07	你好！百忙之	网友“UU0081902” 您	2013/11/22 16:39	1	1	1	3
559	12712	UU0081942	希望解决清	2013/9/2 20:32	易书记： 您好	网友“UU0081942” 您	2013/10/31 16:14	1	1	1	3
560	12724	UU008826	关于反映A	2013/8/28 14:49	尊敬的领导：	网友“UU008826” 您	2013/9/27 18:14	1	1	1	3
562	12755	UU0081083	A3区含浦镇	2013/8/14 12:04	尊敬的易书记：	网友“UU0081083” 您	2013/10/31 16:06	1	1	1	3
563	12768	UU008506	关于反对在	2013/8/8 11:10	尊敬的A市委市政府	网友“UU008506” 您	2013/9/27 18:09	1	1	1	3
564	12847	UU0081549	反映因建设	2013/6/17 18:24	易书记： 您好	网友“UU0081549” 您	2013/8/30 11:57	1	1	1	3
565	12864	UU008967	A5区龙锐尚	2013/6/3 22:18	龙锐尚苑中心	网友： 您好，您的留	2013/8/30 11:40	1	1	1	3
568	16172	UU0081431	咨询A7县城	2018/12/6 11:49	领导好！犹记得从	网友“UU0081431”您好！	2018/12/11 14:04	1	1	1	3
569	16176	UU008715	反映A7县百	2018/12/4 18:06	我是百熙学校的一	网友“UU008715”您好！	2018/12/12 10:31	1	1	1	3
571	16306	UU0081363	咨询A7县楚	2018/9/17 10:07	尊敬的曾书记：感	网友“UU0081363”您好！	2018/9/26 11:50	1	1	1	3
572	16329	UU008422	咨询A7县办	2018/9/6 8:38	上个月我和我女朋	网友“UU008422”您好！	2018/9/7 11:04	1	1	1	3
573	16345	UU0082434	咨询A7县山	2018/8/29 13:40	我现在居住在山水	网友“UU0082434”您好！	2018/9/5 15:00	1	1	1	3
574	16351	UU0081733	举报A7县横	2018/8/28 14:39	此处违章建筑已经	网友“UU0081733”您好！	2018/9/3 9:41	1	1	1	3
576	16366	UU0081112	咨询A7县小	2018/8/18 19:54	作为安沙镇一名底	网友“UU0081112”您好！	2018/8/30 9:24	1	1	1	3
577	16405	UU0081004	反映A7县松	2018/7/29 7:15	最近天气炎热，清	网友“UU0081004”您好！	2018/8/2 15:32	1	1	1	3
579	16440	UU0081233	关于A7县新	2018/7/12 9:24	曾书记您好。县图	网友UU0081233 ：您好！	2018/7/13 15:57	1	1	1	3
580	16443	UU008227	咨询A7县城	2018/7/10 8:34	1:A7县东方航标东	网友“UU008227”您好！	2018/7/11 9:19	1	1	1	3
581	16450	UU008976	再次请求A7	2018/7/5 21:50	A7县曾书记及教育	网友“UU008976”您好！	2018/7/13 10:24	1	1	1	3
582	16451	UU008149	希望在A7县	2018/7/5 9:49	A7县星沙街道东一	网友“UU008149”您好！	2018/7/30 16:42	1	1	1	3
583	16463	UU0081807	举报A7县都	2018/6/30 9:40	尊敬的领导：我是	网友“UU0081807”您好！	2018/7/9 10:42	1	1	1	3
1366	37428	UU0082439	反映关于B9	2019/3/9 23:47	关于西地省B9市均	网友：您好！收到您反映	2019/3/12 15:05	1	1	1	3
556	12669	UU0081548	强烈呼吁采	2013/10/9 17:40	近段时间以来	网友“UU0081548” 您	2013/12/27 11:23	1	1	1	3

3. 结论

总结本次比赛，我们使用 jieba 分词模块预处理文本数据，通过去除超高频词、超低频词和词义筛选对词汇-文本矩阵进行降维，基于 TF-IDF（词频-逆文档频率）特征提取，使用深度学习 LSTM 模型，根据留言详情分类，并使用 F-score 分析模型训练结果。针对热度问题，我们采用 LDA 模型，对文档——词语——主题三层进行。最后从完整性、相关性、可

解释性三个角度，制定答复意见评价方案。根据评分标准，分析答复意见并进行打分排序。

但是我们得到的结果与预期目标有一些出入。其中问题 2 中，热点问题明细表与题目预期结果有一些差距。这可能是由于热点评价方案中点赞数在同类留言下差距较大，影响聚类结果导致。下一步我们将从特征整理与提取、热点评价方案等角度作进一步的改善，也会对文本挖掘作更深入细致的探讨。

4. 参考文献

【1】唐晓波.基于句子主题发现的中文多文档自动摘要研究.情报科学，2020
【2】聂文汇.基于热度矩阵的微博热点话题发现.计算机工程，2017
【3】邹运怀.基于文本挖掘的道岔故障分类研究.北京交通大学，2016
【4】王子牛.基于语义强化和特征融合的文本分类.软件，2020
【5】黄九鸣.面向舆情分析和属性发现的网络文本挖掘技术研究.国防科技大学，2011
【6】黄钊炜.面向主题的文本挖掘研究与应用.华中科技大学，2018
【7】张琳.文本分类中一种改进的特征项权重计算方法.福建师范大学学报（自然科学版），2020
【8】陈义.文本挖掘在网购用户评论中的应用研究，2018