摘要

随着信息时代的到来，政府收到的社情民意反馈不断攀升，而目前依靠人工检索的方式对庞大的数据进行留言划分和热点整理，显然是效率低下的。因此，我们需要构建一个文本挖掘应用模型，辅助政府人员从诉求类型、诉求热点和答复质量三方面，更快更高效的提高服务质量的目标。

第一步：通过 jieba 分词和去除停用词进行**数据预处理**，只留下重要的字段，提高模型的可行性和高效性。因为群众留言的原始文本字段过长，还有很多对文本挖掘没有帮助的词，如果我们不对原始数据进行预处理，不仅容易导致“维数灾难”，而且会影响我们的计算效率。

第二步：利用 **Fasttext 快速文本分类算法**进行有监督模型训练，计算出 F-scores 来评价模型。最终得到 **F1 评价指标等于 0.92**。

查全率	0.92056583242655
查准率	0.92056583242655
F1	0.92

第三步：运用 **TF-IDF 算法**计算词频和逆向文件频率，得到词向量。当两个文本之间相似度大于 0.9 时，将两个文本视为相似文本，相似文本以边相连，并以此建立图模型。然后利用 **DFS 深度优先搜索**，将文本进程查并集分类，把总文本看成图，这样我们就很容易得到多个子图，而当某个子图为所有子图中度最大的子图时，该子图对应的文本即为最热点的问题。

热度排名	涉及留言数	问题描述
1	314	A7 县土地拆迁
2	184	58 车贷案董事长外逃，经侦拖延办案
3	171	五矿万境小区具有一系列问题
4	167	A 市三一大道全线快速化改造何时启动？
5	161	小区边上建了个大型搅拌厂，噪音大粉尘大

第四步：在 TF-IDF 计算完词向量的基础上，通过问题词向量与答复词向量的**余弦值计算相似度**，夹角角度越接近 0°，也就是两个向量越相似，同时考虑答复的时效性，绘制问题与答复的时间差图像。最后将答案的时效性与答复的相关性进行加权考虑，综合得到答复的评价标准，结果如下所示（答复总数为 2816 条，50 天内答复视为答复快）：

	答复数	占比
答复相关	1986	0.70
答复快	2718	0.96
综合评价	1492	0.53

关键字：**jieba 分词 Fasttext 算法 TF-IDF 算法 DFS 算法 余弦相似度**

Abstract

With the advent of the information age, the government has received increasing feedback from public opinion on social conditions. At present, it is obviously inefficient to rely on manual retrieval to divide messages and sort out hot topics on the huge data. Therefore, we need to build a text mining application model to assist the government personnel to improve the service quality faster and more efficiently from the three aspects of appeal type, appeal hot spot and reply quality.

Step 1: pre-process the data through jieba word segmentation and removal of stop words, leaving only important fields, so as to improve the feasibility and efficiency of the model. Because the original text field of people's comments is too long, and there are many words that are not helpful to text mining. If we do not preprocess the original data, it will not only easily lead to "dimension disaster", but also affect our computing efficiency.

Step 2: use Fasttext classification algorithm to conduct supervised model training and calculate f-scores to evaluate the model. The final formula 1 evaluation index is equal to 0.92.

Recall ratio	0.92056583242655
Precision	0.92056583242655
F1	0.92

Step 3: the tf-idf algorithm is used to calculate the word frequency and the reverse file frequency to obtain the word vector. When the similarity between two texts is greater than 0.9, the two texts are regarded as similar texts, and the similar texts are connected by edges, and a graph model is established based on this. Then, DFS depth-first search is used to search and classify the text process into union sets and treat the total text as a graph. In this way, we can easily get multiple sub-graphs. When a certain sub-graph is the largest sub-graph in all sub-graphs, the corresponding text of the sub-graph is the hottest problem.

The heat level	Number of comments involved	Problem description
1	314	A7 county land demolition
2	184	58 car loan case chairman fled, economic investigation delay handling the case
3	171	Minmetals wanjing community has a series of problems
4	167	When will the rapid transformation of A city sany avenue start?
5	161	A large mixing plant was built on the edge of the community

Step 4: on the basis of tf-idf to calculate the word vector, the similarity is

calculated by the cosine value of the question word vector and the answer word vector. The closer the included Angle is to 0°, the more similar the two vectors are. Meanwhile, considering the timeliness of the reply, the time difference between the question and the reply is drawn.Finally, the timeliness of the answer and the completeness of the answer are weighted into consideration, and the evaluation criteria of the answer are integrated. The results are as follows (the total number of replies is 2816, and the reply within 50 days is considered as the reply is quick) :

	Answer number	Proportion
Complete reply	1986	0.70
Fast answer	2718	0.96
Comprehensive evaluation	1492	0.53

Key words: jieba、 Fasttext algorithm、 tf-idf algorithm、 DFS algorithm、 cosine similarity

目录

一、简介

1.1 挖掘意义

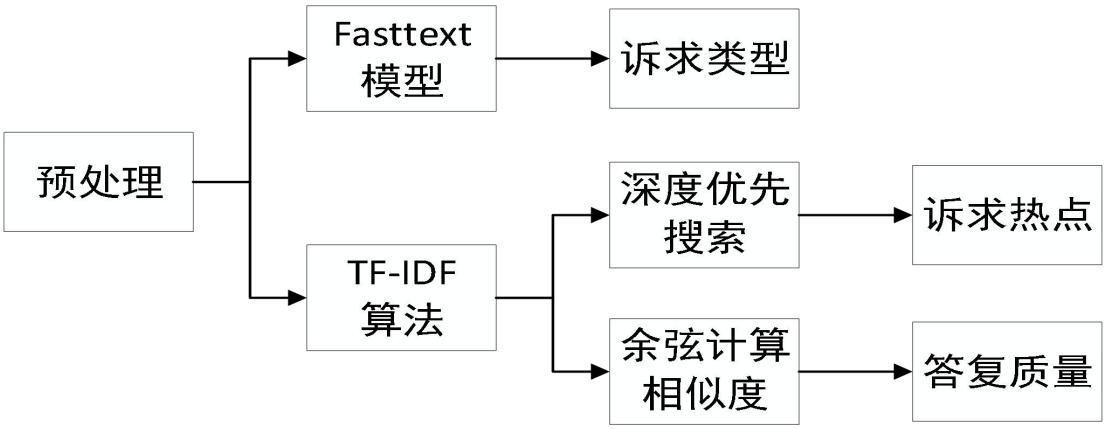
中国作为社会主义国家，对于民意反馈非常重视。如今，随着网络的普及和百姓法律意识的提高，政府收到的关于社情民意方面的文本数据量不断地攀升，以往依靠人工来进行留言划分和热点整理的方法已不再适用，建立基于自然语言处理技术的智慧政务应用将成必然趋势。

应用自然语言处理技术，智慧政务系统能在线上简单收集问题，并在第一时间分类整理出百姓的诉求热点，评判政府人员答复的质量，从而更快更高效的处理解决问题，缩短百姓和政府的沟通时间，实现政务数据实时化，响应及时化，服务高效化。因此，文本挖掘之类的自然语言处理技术，拥有重要的应用价值。

1.2 挖掘目标

本文主要利用 jieba 中文分词工具、TF-IDF 算法和 Fasttext 模型的，构建文本挖掘模型，从而达到辅助政府人员，从诉求类型、诉求热点和答复质量三方面，更快更高效的提高服务质量的目标。

1.3 挖掘流程



问题一：运用 JieBa 分词，利用 Extract_tags 去除停顿词，预处理后，通过 Fasttext 进行有监督模型训练，计算出 F-scores 来评价模型。

问题二：预处理后，使用 TF-IDF 计算词向量，建立图模型，通过 DFS 深度优先搜索将文本进程查并集分类，将总文本看成图，得到多个子图，子图中度最大的即为诉求热点的问题。

问题三：预处理后，使用 TF-IDF 计算词向量，计算问题词向量与答复词向量的欧几里得距离，同时考虑答案的时效性，绘制答复与答复的时间差图像，将答案的时效性与答复的相关性进行加权考虑，评判答复质量。

二、预处理

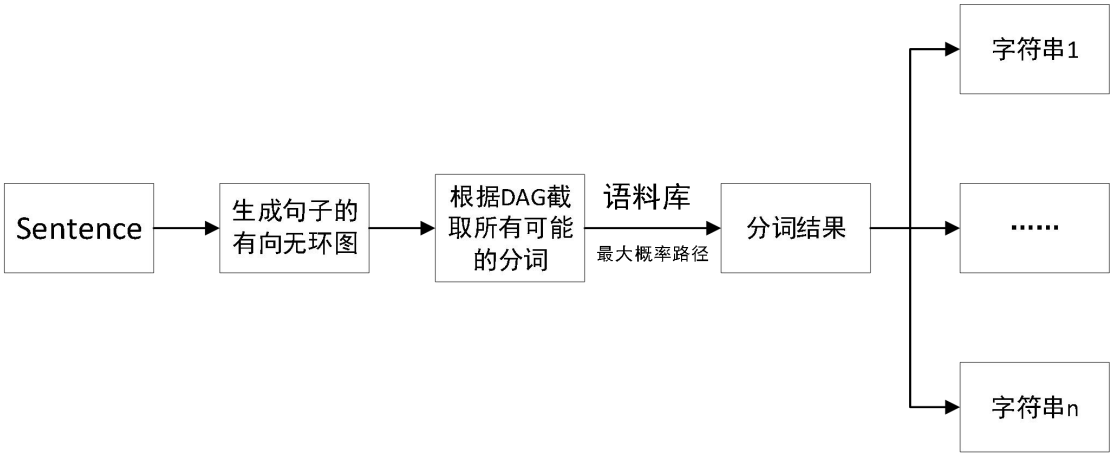
2.1 数据特点

观察附件数据不难发现，留言条数多达 16352 条，数据量庞大。所有留言字段都是文本形式，而且多数还为长文本，给文本挖掘带来了一定难度。除此之外，留言中还有许多停用词和噪声词存在，不进行处理过滤的话，不仅会加大计算量，还会影响结果精确度，因此，对原始数据进行预处理尤为重要。

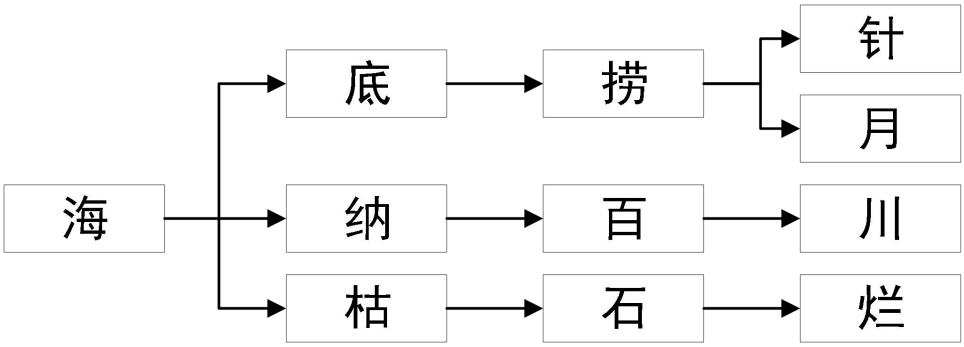
2.2 jieba 分词

因为我们并不能对整句话进行直接处理，所以我们需要使用 jieba 中文文本分词工具，将文字组成的词语分隔提取出来，数值化后构成一个个向量。

jieba 分词工具首先借助前缀词典进行词图扫描，通过前缀词典可以快速构建包含全部可能分词结果的有向无环图，这个图中包含多条分词路径。最后基于标注语料，使用动态规划的方法可以找到最大概率路径，作为最终的分词结果。



计算机根据词典来分词，但即使机器速度比人类效率快了很多，也会非常耗时间。所以 jieba 分词用到了数据结构中的前缀树或字典树对词语进行高效的分类，便于查找。例如，我们分词过程中出现了“海底捞针”的词语，传统方法需要将字典中的“海纳百川”、“海枯石烂”、“海底捞月”、“海底捞针”四个成语逐字探寻，而字典树只需从上到下搜索，每一次判定一个字，若某个结点不符合要求，那么就会停止这条线的继续搜索，高效快捷。



在实际运用中，Jieba 分词分为了三种模式，精确模式、全模式和搜索引擎模式。以附件二中的第 148836 条留言为例：

原文：	上半年，C市有没有什么教师招聘啊？我希望能在C市当一名教师。
全模式：	上半 上半年 半年，C 市 有没有 没有 什么 教师 招聘 啊 ？ 我 希望 能 在 C市 当 一名 教师。
精确模式：	上半年 ， C市 有没有 什么 教师 招聘 啊 ？ 我 希望 能 在 C市当 一名 教师。
搜索引擎：	上半 半年 上半年 ， C市 没有 有没有 什么 教师 招聘 啊 ？ 我 希望 能 在 C市当 一名 教师。

根据文本类型和测试结果分析，三种模式中全模式效果最好，故采用全模式进行分词。

2.3 去除停用词

停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。为了节省存储空间和提高搜索的效率，我们需要在处理时自动忽略这些词，例如中文中的“的”，“和”，“吗”，“是”等，英文中的“i”，“is”，“what”等，另外还可以去除一些不要的标点符号，这些词几乎在每个文档都会出现，查询这些词是没有意义的，还会降低搜索的效率，加大了工作量。

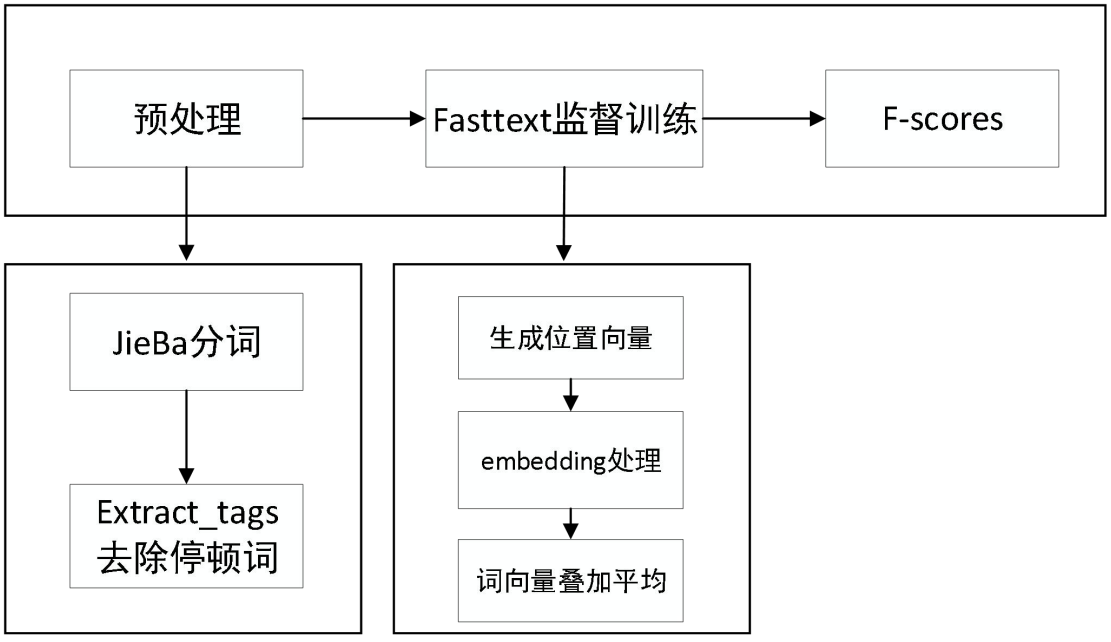
过滤停用词的基本步骤：首先读入停用词表文件，在正常分词后，从分词结果中取出停用词。

还是以附件二中的第 148836 条留言为例：

原文：	上半年，C市有没有什么教师招聘啊？我希望能在C市当一名教师。
停用词：	“，”“有没有”“什么”“啊”“？”“在”“。”
输出结果：	上半 上半年 半年 C 市 教师 招聘 我 希望 能 C市 当 一名 教师

三、问题一分析

3.1 问题一流程图

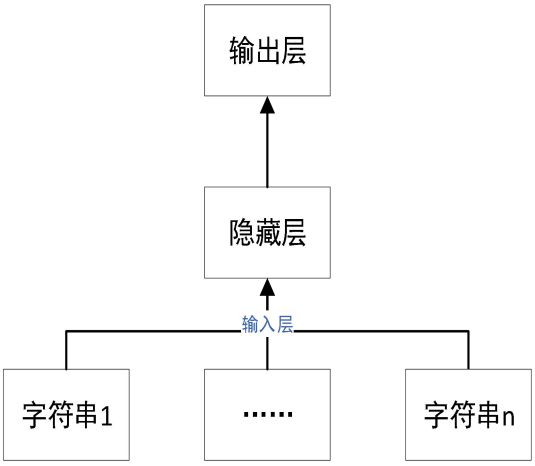


3.2 FastText 模型

（1） 特点：fastText 作为一个常用的快速文本分类算法， 专注于文本分类，能在许多文本问题上实现当下最好的表现，例如文本倾向性分析或标签预测。fastText 拥有很多优良的性质，能在保持高精度的情况下加快训练速度和测试速

度，不需要预训练好的词向量，因为 fastText 可以自己训练词向量，而且拥有两个重要的优化：Hierarchical Softmax、N-gram，号称能够训练模型“在使用标准多核 CPU 的情况下 10 分钟内处理超过 10 亿个词汇”。

（2）模型架构：fasttext 模型是有三层：输入层，隐藏层、输出层



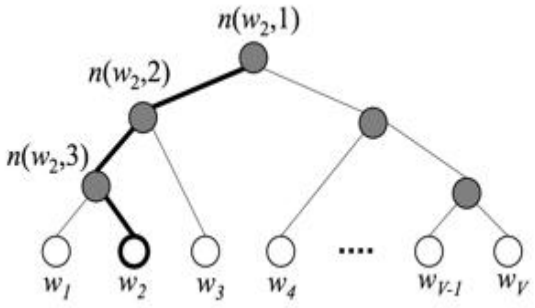
输入层：输入为要分类句子的每个词向量，这些词向量是由多个字符级别的 n-gram 向量表示的。

隐藏层：将输入层的输入生成表征文档的向量，提供给输出层进行分类。具体做法是：叠加这篇文档所有词及 n-gram 的词向量，然后取平均。叠加词向量背后的思想就是传统的词袋法，即将文档看成一个由词构成的集合。

输出层：一个 softmax 多类别线性分类器。输入为用来表征当前文档的向量。

整体核心思想：将一篇文档的词及其 n-gram 向量叠加后平均得到文档向量，然后使用 softmax 多分类器进行分类。为了提高效率，softmax 分了使用了层次 softmax。

（3）层次 softmax：softmax 函数可在 fasttext 的模型结构的输出层中充当激活函数，目的就是将输出层的值归一化到 0-1 区间，将神经元输出构造成概率分布，主要就是起到将神经元输出值进行归一化的作用。在标准的 softmax 中，计算一个类别的 softmax 概率时，我们需要对所有类别概率做归一化，在这类别很大情况下非常耗时，因此提出了分层 softmax(Hierarchical Softmax)，思想是根据类别的频率构造霍夫曼树来代替标准 softmax，通过分层 softmax 可以将复杂度从 N 降低到 logN。



在层次 softmax 模型中，叶子结点的词没有直接输出的向量，而非叶子节点都有响应的输在在模型的训练过程中，通过 Huffman 编码，构造了一颗庞大的 Huffman 树，同时会给非叶子结点赋予向量。我们要计算的是目标词 w 的概率，这个概率的具体含义，是指从 root 结点开始随机走，走到目标词 w 的概率。因此在途中路过非叶子结点（包括 root）时，需要分别知道往左走和往右走的概率：

$$p(n,left)=\sigma(\theta_n^T\times h)$$
$$p(n,right)=\sigma(-\theta_n^T\times h)$$

其中目标词为 w 的概率可以表示为：

$$p(w)=\prod_{j=1}^{L(w)-1}\sigma(sign(w,j)\times\theta_{n(w,j)}^T\times h)$$

其中 θ n(w,j)是非叶子结点 n(w,j)的向量表示（即输出向量）；h 是隐藏层的输出值，从输入词的向量中计算得来；sign(x,j)是一个特殊函数定义

$$sign(w,j)=\begin{cases}1，若n(w,j)的左子节点\\-1，若n(w,j)的右子节点\end{cases}$$

此外，所有词的概率和为 1，即

$$\sum_{i=1}^{n} p(w_i)=1$$

综上所述，参数更新公式为：

$$\theta_j^{new}=\theta_j^{old}-\eta(\sigma(\theta_j^T h)-t_j)h$$

（4）N-gram 特征

n-gram 是基于语言模型的算法，基本思想是将文本内容按照子节顺序进行大小为 N 的窗口滑动操作，最终形成窗口为 N 的字节片段序列。而且 n-gram 可以根据粒度不同有不同的含义，有字粒度的 n-gram 和词粒度的 n-gram。

例如对单个单词 data 来说，假设采用 3-gram 特征，那么 data 可以表示成图中 4 个 3-gram 特征，这五个特征都有各自的词向量，五个特征的词向量和即为 data 这个词的向其中“<”和“>”是作为边界符号被添加，来将一个单词的 ngrams 与单词本身区分开来：

<data>: “<da” “dat” “ata” “ta>”

使用 n-gram 有如下优点：

1、为罕见的单词生成更好的单词向量：根据上面的字符级别的 n-gram 来说，即是这个单词出现的次数很少，但是组成单词的字符和其他单词有共享的部分，因此这一点可以优化生成的单词向量

2、在词汇单词中，即使单词没有出现在训练语料库中，仍然可以从字符级 n-gram 中构造单词的词向量

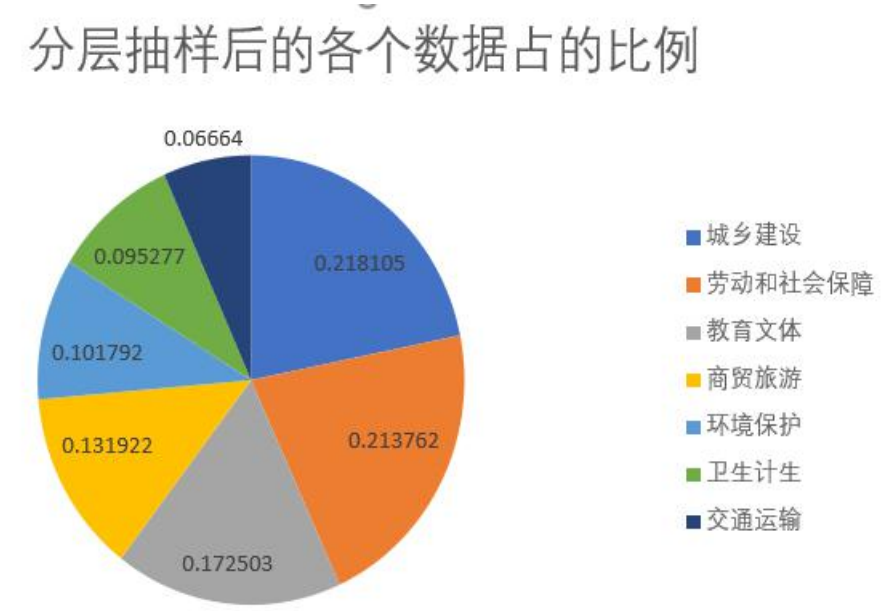
3、n-gram 可以让模型学习到局部单词顺序的部分信息，如果不考虑 n-gram 则便是取每个单词，这样无法考虑到词序所包含的信息，即也可理解为上下文信息，因此通过 n-gram 的方式关联相邻的几个词，这样会让模型在训练的时候保持词序信息

3.3 结果分析

留言里有许多无意义表达，所以我们将它去除停用词，空行，经过预处理再分词。

我们用 StratifiedShuffleSplit（）函数将数据集划分成了训练集与测试集，比例为 0.2 和 0.8，它提供分层抽样功能，确保每个标签对应的样本的比例。而数据集在进行划分之前，首先是需要用参数 random_state 控制将样本随机打乱处理，否则容易产生过拟合，模型泛化能力下降。

下图为分层抽样后的各个数据占的比例：



将附件二中的留言分词处理并去除停用词后，绘制词云图如下所示：

由上面两幅图共同得出，出现频率最多的是“规划”，“相关”，“政府”，“工作”，“国家”，“文化”等词，大家的留言数较多的是有关于“城乡建设”，“劳动建设”和“教育文化”的话题，其中“城乡建设”和“劳动建设”的话题占了总话题的 43%左右，说明政府在城乡建设和劳动建设方面还有较多地方值得改进。

最后根据 F-Score 评估法：

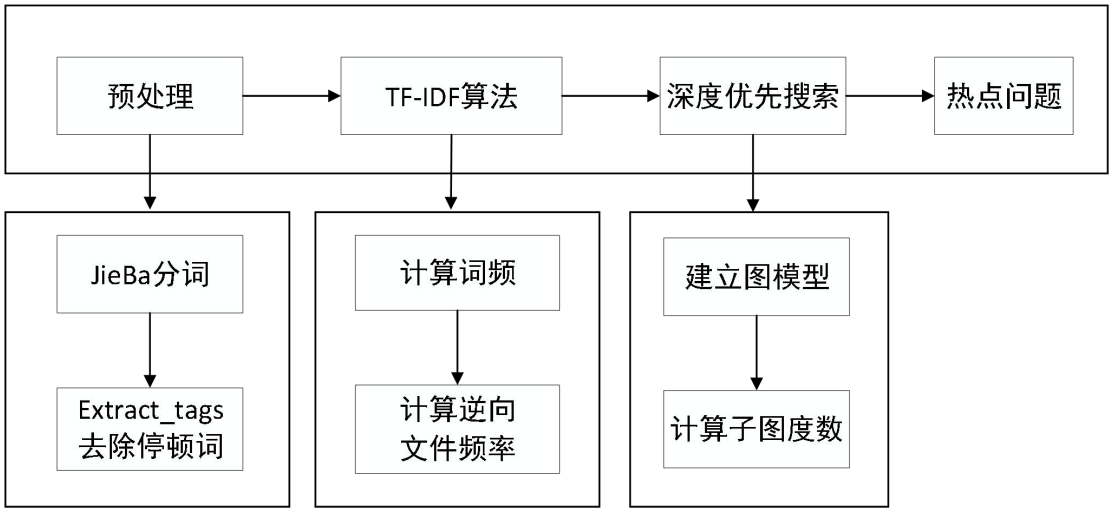
$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

查全率	0.92056583242655
查准率	0.92056583242655
F1	0.92

最终计算得出结果 F1=0.92，符合预期结果。

四、问题二分析

4.1 问题二流程图



4.2 TF-IDF 算法

对于民意反馈留言来说，一些出现频率较高的通用词，比如“国家”、“建议”、“影响”之类的词，往往对与留言热点的贡献并不是特别大，而一些出现频率较小的词，比如“《城乡规划法》”、“12315 热线”之类的专有名词，却能起到直接揭示热点的作用。

基于这个思想，本文采用 TF-IDF 算法来过滤掉噪声词，保留重要的词语，计算词向量。

（1）TF：即词频，某个特定的词语 w 在留言中出现的频数。通常认为词语 w 在留言中大量出现时，该词为噪声词。

由于留言的词数长短不一，所以我们使用如下公式，以达到词频标准归一化的目的。

$$TF_{\mathrm{w}}=\frac{\text{词语w出现的频数}}{\text{留言总词数}}$$

（2）IDF：即逆向文件频率，包含特定词语 w 的留言越少，则 IDF 越大，说明该词对主题具有很好的类别区分能力。通常认为词语 w 在大量留言中出现时，

该词为噪声词。

计算方法如下所示：

$$\mathrm{IDF} = \log(\frac{\text{语料库的留言总数}}{\text{包含词语w的留言数}+1})$$

（3）TF-IDF 向量生成：把两个词两两合并成一个集合，每个词的 TF-IDF 向量即权重向量等于词频乘以逆向文件频率。

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF}$$

4.3 DFS 算法

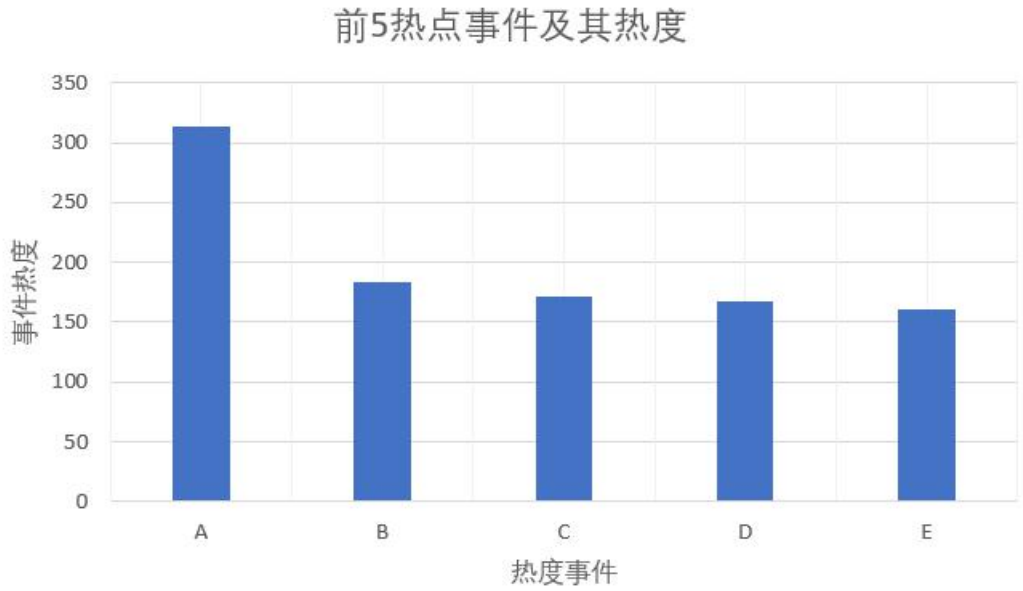
1.设置一个文本的访问数组。
2.选择一个起点文本 t，并设置一个栈 T，将 t 放入栈中。
3.通过计算 t 与文本 c 的相似度，认为相似度大于 0.9 的在一个子图里面，将 c 放入栈 T 中。
4.再将访问数组中文本对应的位置设置为 false。
5.遍历完一遍后再从文本中寻找未被访问文本更新 t。
6.重复上述步骤，直到所有文本都被访问。

4.4 结果分析

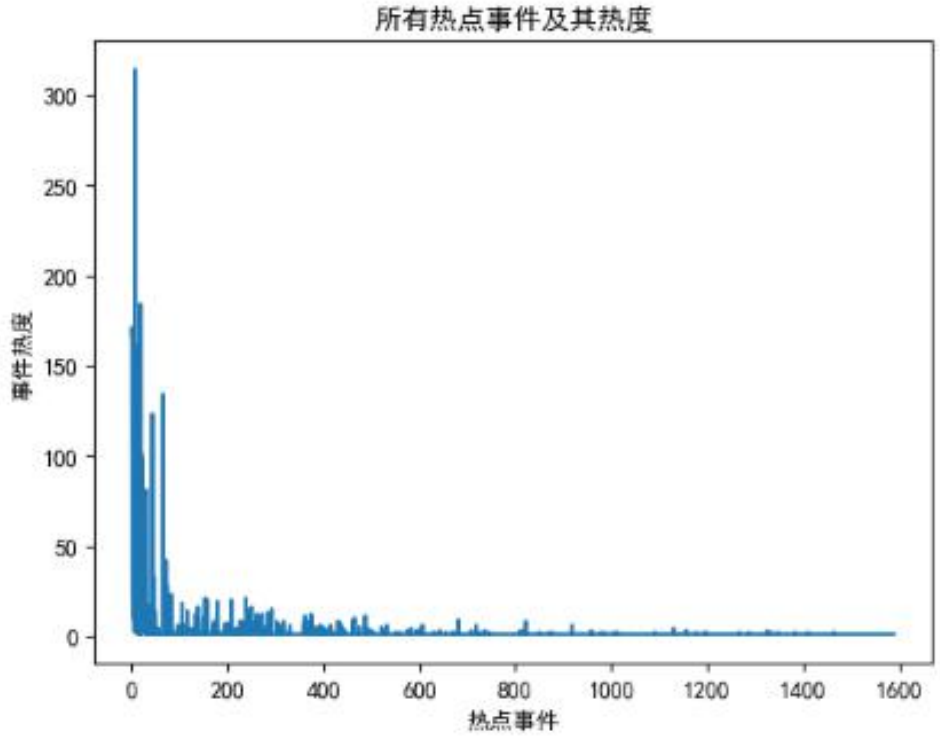
一千个读者，就有一千个哈姆雷特，每个人都有问题，如果每一个都单独处理的话，那肯定耗时耗力，效率低下。于是通过 DFS 深度优先搜索算法，巧妙的利用了图的性质，计算出了排名前五的热点，整理出了问题表和热点问题留言明细表，把热点问题留言明细表归为哪类问题都一一列举出来。可以看出，热度最高的五件事件分别为:“A7 县土地拆迁”“8 车贷案董事长外逃，经侦拖延办案”“五矿万境小区具有一系列问题”“A 市三一大道全线快速化改造何时启动”“小区边上建了个大型搅拌厂，噪音大粉尘大”。

热度排名	涉及留言数	地点/人群	问题描述
1	314	A7 县的市民	A7 县土地拆迁
2	184	58 车贷受害者	58 车贷案董事长外逃，经侦拖延办案
3	171	A 市五矿万境的业主	五矿万境小区具有一系列问题
4	167	A 市三一大道关注人群	A 市三一大道全线快速化改造何时启动？
5	161	A 市暮云街道丽发新城的业主	小区边上建了个大型搅拌厂，噪音大粉尘大

就像微博的热搜一样，热点问题表也直观的将某个时间段的特定人群问题给展现出来，热点问题表将当时热度高的话题排在前面，展现出来。如图，前五个热点事件 ABCDE 分别对应上面按热度排名的问题描述，从图像中可以很直观的看出来 A7 县土地拆迁的问题在当地引起了很大争议，事件热度为 300 多，远超其余的热点事件，所以急需政府前去调查。而剩下的事件热度相对较低，可以留在解决土地拆迁问题后面解决。



为了能更加直观的反映出不同热点事件的差距，以及他们的变化趋势，我们绘制了所有上报事件的热度的柱状图。

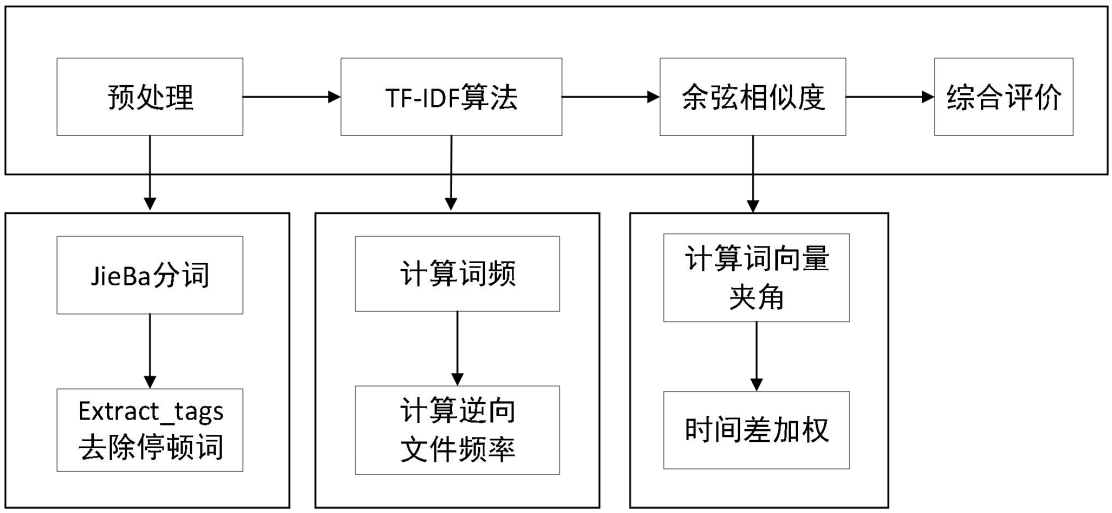


从这张图反应的结果来看，所有热点事件和它的热度，都与他涉及范围与

影响人数有关。有一些问题涉及范围广了，会受到许多人的关注以及留言，但一些问题涉及范围小的，热度就较小，关注人较少。通过这个柱状图来看，热度小的事件有许多，生活中经常每个人都有很多琐碎的事情，如留言编号为 229228 的用户留言主题是“妻子公派国外访学，我能否申请出国探亲假？”，受关注度就很小，因为该事只涉及到他和他的公司。而“58 车贷案”热度却很大，他牵涉的人很多，在社会上也引起了一定的影响，还有许多路人了解该事会产生的情绪都让这个事件热度闹得很大。就像这个热度排名，涉及人群范围广的热度一般更高。

五、问题三分析

5.1 问题三流程图



5.2 余弦相似度

余弦相似度原理：用向量空间中的两个向量夹角的余弦值作为衡量两个个体间差异大小的度量，值越接近 1，就说明夹角角度越接近 0°，也就是两个向量越相似，就叫做余弦相似。

经过数据预处理后，问题就变成了如何计算这两个向量的相似程度。我们可以把它们想象成空间中的两条线段，都是从原点（[0, 0, ...]）出发，指向不同的方向。两条线段之间形成一个夹角，如果夹角为 0 度，意味着方向相同、线段重合,这是表示两个向量代表的文本完全相等；如果夹角为 90 度，意味着形成直角，方向完全不相似；如果夹角为 180 度，意味着方向正好相反。因此，我们可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。

以留言 230004 和 241293 为例：

留言230004	A7县星沙一桥从早堵到晚，什么时候会动工加宽？
留言241293	A7县星沙一桥从早堵到晚不见整改

计算出词向量分别为：

202397词向量	[1 1 1 1 1]
241293词向量	[1 1 1 0 0]

利用公式

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(X_i \bullet Y_I)}{\sqrt{\sum_{i=1}^{n}X_i^2}\bullet\sqrt{\sum_{i=1}^{n}Y_i^2}}$$

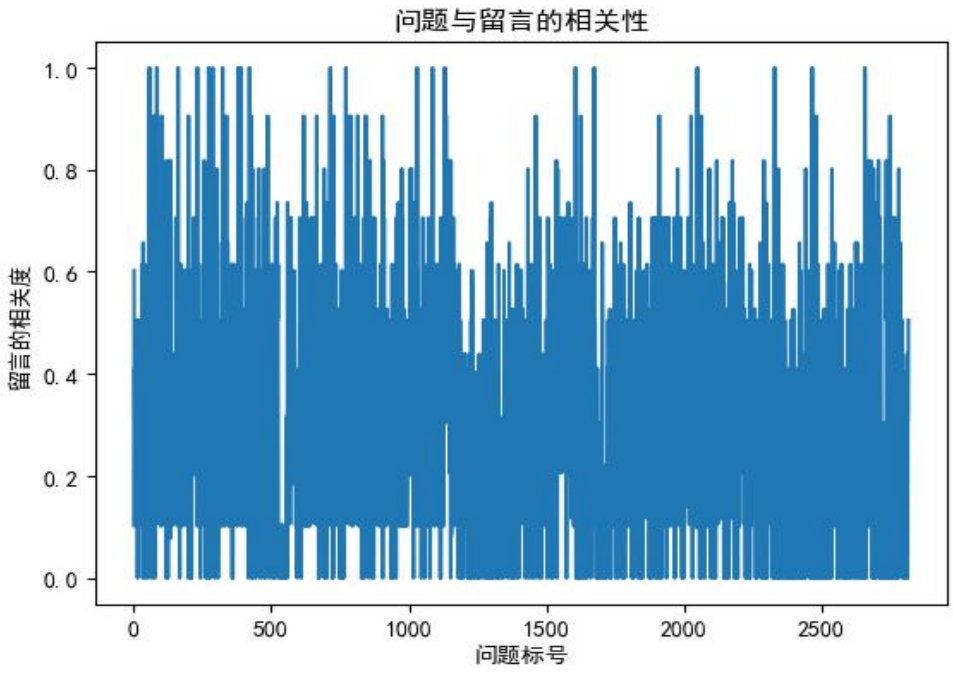
可以计算得出

$$\cos(\theta)=0.77$$

计算结果中夹角的余弦值为 0.77 非常接近于 1，所以，上面的留言 202397 和 241293 是基本相似的。

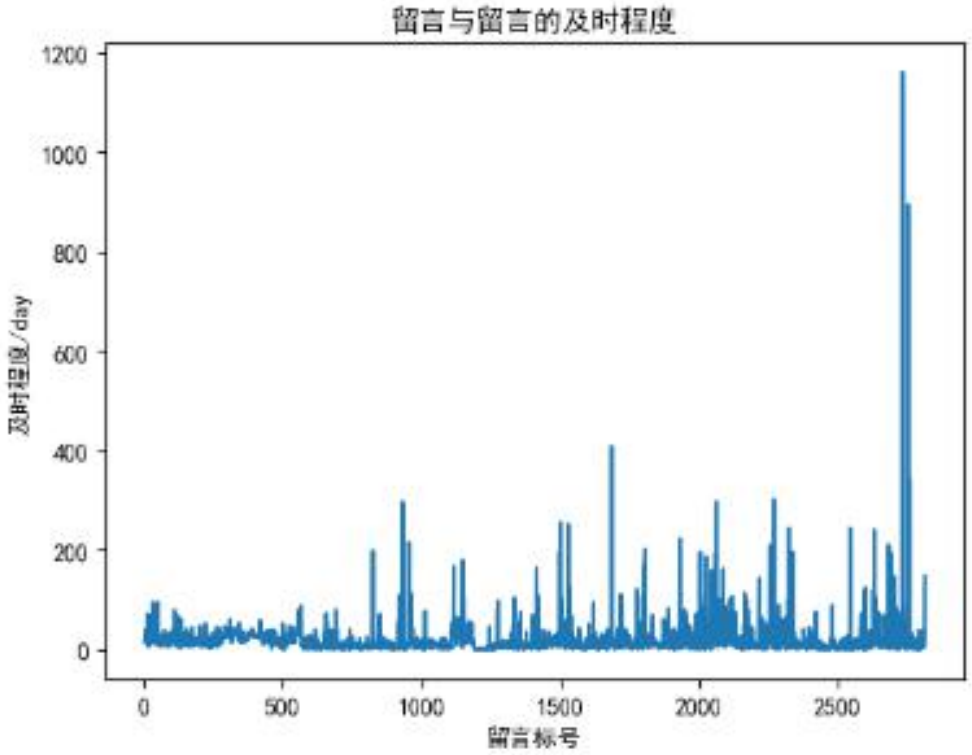
5.3 问题三结果分析

根据余弦相似度计算：



其中有部分的回复和问题没有任何相关性，相关度取值为 0，因为数据较为密集，出现了下面是空白上面相连的情况。分析上图，可以看见相关性总体落在（0.6，0.8）这个区间，政府回复总体上较为相关。

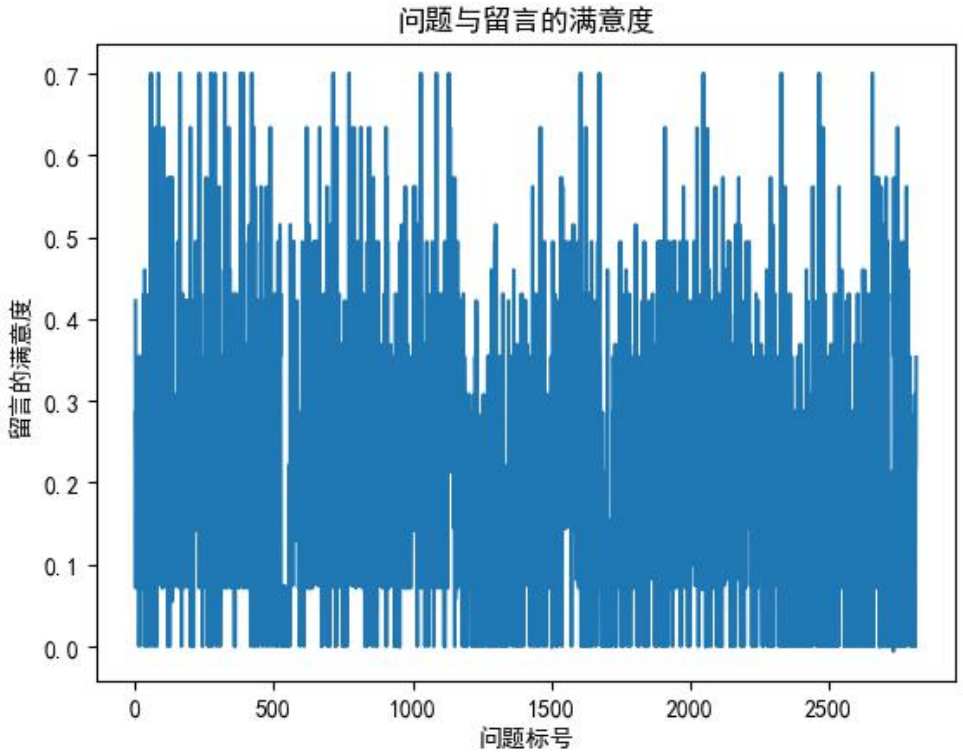
然后计算附件中的留言时间差，得到如下结果：



从图中可以看出，大部分回复都是在 50 天内完成的，只有约百分之 4 的回复拖到了 50 天后，其中最晚的回复时间已经接近 3 年左右。如果在现实生活中，诉求在三年后才得到回复，就算回复再完整，相关性再高，显然都不会让人满意的。所以在考虑回复标准时，也应该把回复时间考虑在内，查阅相关资料并结合实际情况得出：

$$综合评价=0.7\times回复完整性+0.3\times回复及时性$$

于是我们利用新指标——综合评价作为回复的评价标准，即满意程度，绘制出相应图形：



因为只有极少部分答复出现不及时回复的现象，所以最终的满意程度在图像上与相似性大致相同，与预期一致。与柱状图下方的空白类似，少量的不同点也因为横坐标数据量过大，导致其不能清晰呈现。

综上所述，总体结果如下所示：

	答复数	占比
答复相关	1986	0.70
答复快	2718	0.96
综合评价	1492	0.53

最终的综合评价占比 53%，比相关性降低 17%。与预期有略微出入，因为有些回复虽然相关，但是可能存在回复不及时的现象，所以综合评价会比相关性略低，但是预计值还是在 65%左右，因为答复快的达到了 96%，下降不会特别突出。经讨论，结果与预期有出入的原因可能有二：一、相关性与及时性的权重设置有待商榷；二、对于答复较快的时间阈值可能存在不合理，我们设置的是 50 天，每个人对于这个时间的设定可能存在差异，从而影响计算结果。

六、模型优化

6.1 语序丢失

我们使用 N-gram 模型的好处是快速便捷，但未考虑到词序带来的影响，比如“你欺负我”与“我欺负你”，我们认为是相同的，我们可以考虑使用启发式搜索引入 N-gram 模型中，从而减少未考虑词序带来的影响。

6.2 语境丢失

如果语句 A 自身有一层意思但在文本 T 中考虑又会蕴含另一层意思，也就是两层意思，我们只考虑了第一层意思，可以考虑下分词后抽样，将文本结合在一起减少因语句 A 在文本 T 中的含义损失。

6.3 语义结合

语句 A 与语句 B 各自都有含义，但两者结合产生新的含义，我们可以考虑引入遗传算法，考虑让词向量变异，从而减少因 A 与 B 结合含义的损失。

七、反思与展望

在本篇论文中，主要使用了 jieba 分词、Fasttext 算法、TF-IDF 算法、DFS 算法、余弦相似度等模型或工具，在处理政务问题中的诉求类型、诉求热点和答复质量三方面做出了解答，所得结果也符合实际，我们也对结果做了一定分析。但是要把我们建立的模型推广到现实生活中，我们还有很长的路要走：主要使用的 fasttext 模型虽然能在保证精确性的同时快速得出结果，但是只能对一些比较简单的文本进行分类，对于一些较为复杂的意见也会显得乏力；N-gram 中没有考虑词序，可能会对结果产生一定影响，改进措施在模型优化上有提及；DFS 也存在运行会大量占用内存，对计算机内存消耗高的问题。

在接下来的工作，我们将主要关注怎么样才能更高效地提高算法的能力，了解更多的知识，让论文结构更加实用，更具推广性。通过本次比赛，我们明白了在信息大爆炸时代数据挖掘的重要性，老一套的人工方法大多是需要耗时耗力的。就像当年非典期间，上报染病人数只能通过人工统计，效率较低，影响防疫工作，而随着科技越来越发达，利用大数据或数据挖掘之类的技术，我们可以对现在的新冠状病毒感染者实时上报，坐在家中就可以了解全球疫情，大大推动了防疫工作的前进，也为中国能马上控制住疫情提供了帮助。

七、参考文献

[1]祝永志,荆静.基于 Python 语言的中文分词技术的研究[J].通信技术,2019,52(07):1612-1619.
[2]艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.
[3]王光慈,汪洋.基于 FastText 的短文本分类[J].电子设计工程,2020,28(03):98-101.
[4]李泽龙. 基于 FastText 的长文本快速精确分类算法研究[D].浙江大学,2018.
[5]胡学钢, 董学春, 谢飞. 基于词向量空间模型的中文文本分类方法[J]. 合肥工业大学学报:自然科学版, 2007, 30(10):1261-1264.
[6]曹春萍,黄伟.基于用户权威度与热度分配聚类的微博热点发现[J].计算机工程与设计,2020,41(03):664-669.
[7]王美方, 刘培玉, 朱振方. 基于 TFIDF 的特征选择方法[J]. 计算机工程与设计,2007, 28(23):5795-5796.
[8]崔鹏程. 基于文本挖掘的学术文献内容智能识别方法研究[D].北京交通大学,2019.
[9]张洪,钟凯迪,柴源,魏济,吴艳,谭锦涛,叶文韬.基于 N-Gram 和动态滑动窗口的改进余弦相似度算法研究[J].成都大学学报(自然科学版),2019,38(02):163-166.
[10]王影,库婷婷,许书萍,李伟强,袁博.敬畏感的情绪成分分析:基于社交网络的文本挖掘[J].心理技术与应用,2020,8(04):235-242.
[11]孙洋,粟栗,张星,王峰生,杜海涛.基于子语义空间的挖掘短文本策略方法[J].电信科学,2020,36(03):83-92.
[12]石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.