# “智慧政务”中的文本挖掘应用

**摘要：**近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依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挑战。同时，随着大数据、云计算、人工智能等技术的发展，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。

针对问题一：对附件中市民的留言详情和留言主题进行基础的文本数据预处理，进行数据提取，除去重复留言，去除空留言，采用 jieba 精确模式下的分词，停用词过滤，向量矩阵转换等，对处理后的向量数据集训练，并对训练集中的各个类别项的各个特征属性有清晰的了解，对其进行去噪处理等操作，利用朴素贝叶斯算法生成分类器并构建模型。

针对问题二：首先我们从文件中读取数据，对数据进行预处理，对文本数据去除分词之后去除停用词，之后开始对留言主题中的人物地点作提取操作，在提取这里我们是依然是通过 jieba 对处理好的分词做词性的标注，然后我们把名词类的分词提取出来即可，接下来我们可以向量化提取出来的名词，通过余弦相似度的度量方法来计算相似度并将相似度高的文本数据分在同一组，以此来作为分组的依据，之后我们可以根据文本出现的次数、点赞数和反对数来合理的设计热度指数的计算公式，然后我们就可以对计算好的热度指数进行排名，最后只需对时间范围等稍作调整即可。

针对问题三：我们提出了基于词嵌入的文本摘要(WETS)方法，用于识别、排列和连接突出的 top-y 句，作为简明摘要。最常用的句子关联判断方法是对出现频率高、语用频繁的词语进行检测。鉴于此，由这些词组成的句子被认为是最重要的。然而，我们认为这种技术有两个主要缺点:首先，它在新的摘要中鼓励冗余;其次，对于由其他单词组成的非常重要的句子，它没有给出适当的分数。因此，建立冗余处理机制和面向意义的句子关联评价技术至关重要。因此，初步的任务是从原始文档中删除不相关的标记，如停止词。然后，将第一句的单词和常用词结合起来作为关键词。倾向于关注第一句话的单词的主要原因是，语言学文献显示了一个明确的论题陈述，大部分位于段落的开头，这表明重要的单词可能存在于第一句话中。此外，相对而言，标题中至少有几个词也可能出现在第一句话中。比较分析可以通过比较中间句中的词和最后一句中的词来进行。

**关键词：**中文分词、去重、Tf-IDF 算法、Word2vec 算法、相似度提取、KNN 算法、支持向量机

# Text Mining Application in "Smart Government Affairs"

**Abstract:** in recent years, along with WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government.

In view of the problem a: citizens' message in the attachment for details and the message subject on the basis of text data preprocessing, data extraction, eliminate duplicate messages, remove the empty message, USES the jieba mode of participles, stop words filtering, vector matrix transformation, etc., after processing of the vector data sets, training, and focus on training of each feature attribute of each category is a clear understanding, to deal with the noise of its operation, such as using the naive bayes algorithm to generate a classifier and build models.

For question two: First of all, we read from the file data, data pretreatment, the text data to remove participle after stop words, after the start of the characters in the message subject location do extract, here we are extracted by jieba do to deal with good participle of part of speech tagging, and then we put the word class of word segmentation can be extracted, and then we can be extracted to quantify nouns, by cosine similarity measurement method to calculate the similarity and give high similarity of text data points in the same group, to as a basis for the group, Then we can reasonably design the calculation formula of heat index according to the number of times the text appears, the number of thumb up and the number of opposition. Then we can rank the calculated heat index, and finally we only need to adjust the time range.

For question 3: we propose a text summary (WETS) method based on word embedding, which is used to identify, arrange and connect prominent top-y sentences as concise summaries. The most commonly used sentence association judgment

method is to detect the words with high frequency and frequent pragmatics. In view of this, sentences made up of these words are considered to be the most important. However, we believe that this technique has two major disadvantages: first, it encourages redundancy in the new summary; Second, it does not give a proper score for very important sentences made up of other words. Therefore, it is very important to establish a redundancy processing mechanism and a sense-oriented sentence correlation evaluation technique. Therefore, the initial task is to remove irrelevant tags, such as stop words, from the original document. Then, combine the words of the first sentence with common words as the key words. The main reason for the tendency to focus on words in the first sentence is that linguistic literature shows a clear thesis statement, mostly at the beginning of a paragraph, indicating that important words may exist in the first sentence. In addition, relatively speaking, at least a few words in the title are also likely to appear in the first sentence. Comparative analysis can be carried out by comparing the words in the middle sentence with the words in the last sentence.

**Keywords:** Chinese word segmentation, deduplication, tf-idf algorithm, Word2vec algorithm, similarity extraction, KNN algorithm, support vector machine

# 目 录

## 1. 挖掘目标

对现代热门的问政平台建立基于自然语言处理技术的智慧政务系统，统计市民意见问题，利用 jieba 中文分词库提取主要描述性词语，用 knn 算法、朴素贝叶斯等模型对其建立分类模型，利用 jieba 分词判断其词性的判断统计等完成以下几点目标要求：

1. 利用文本分词和 jieba 分词库提取词性的方法对非结构化的数据进行文本挖掘，通过建立分类模型对附件二中居民的留言主题和留言详情进行分析，通过 jieba 库建立对附件三的留言主题的词性分类。
2. 根据所提供的数据来对其留言分类，并返回给相应的部门，对热点问题进行提取排名并返回评价表。

## 2. 群众留言分类的流程与步骤

### 2.1 流程图

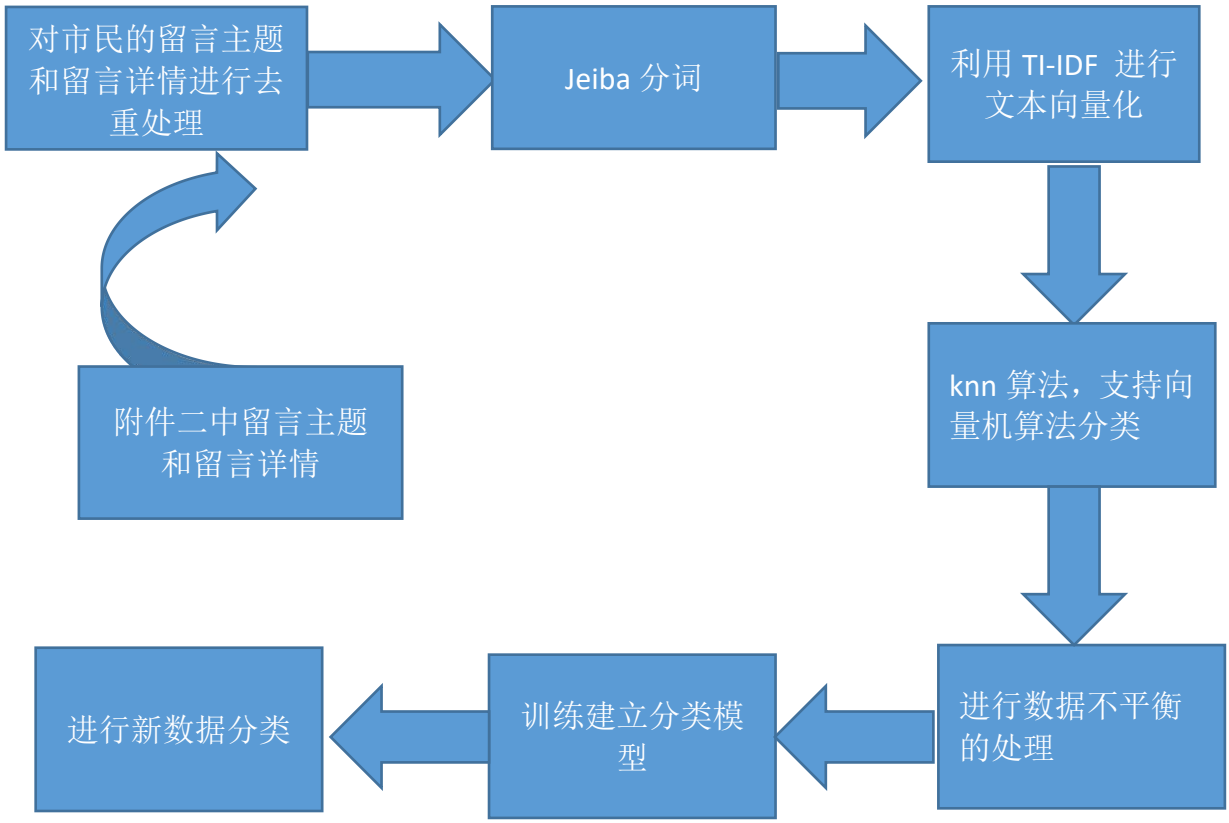


**图 1 总体流程图**

## 2.2 数据预处理

对题目附件二中给出的市民留言主题和详情进行提取，在原始文本语料上进行预处理，为文本挖掘或 NLP 任务做准备。通常，我们会选取一段预先准备好的文本，对其进行基本的分析和变换，遗留下更有用的文本数据，方便之后更深入、更有意义的分析任务。接下来将是文本挖掘或自然语言处理工作的核心工作。文本预处理的三个主要组成部分：

- ▼ 标记化（tokenization）
- ▼ 归一化（normalization）
- ▼ 替换（substitution）

标记化是将文本中的长字符串分割成小的片段或者 tokens 的过程。大段文字可以被分割成句子，句子又可以被分割成单词等等。只有经过了 Tokenization，才能对文本进行进一步的处理。Tokenization 同样被称作文本分割或者词法分析。有时，分割（segmentation）用来表示大段文字编程小片段的过程（例如段落或句子）。而 tokenization 指的是将文本变为只用单词表示的过程。

再进一步处理之前，文本需要进行归一化。归一化指的是一系列相关的任务，能够将所有文本放在同一水平区域上：将所有文本转化成同样的实例，删除标点，将数字转换成相应的文字等等。对文本进行归一化可以执行多种任务，但是对于我们的框架，归一化有 3 个特殊的步骤：

- 词干提取（stemming）
- 词形还原（lemmatizatiion）
- 其他

噪声消除延续了框架的替代任务。虽然框架的前两个主要步骤（标记化和归一化）通常适用于几乎任何的文本或项目，噪声去除是预处理框架中一个更加具体的部分。我们处理的过程必须以特定的顺序进行，视具体情况而定。因此，噪声消除可以发生在上述步骤之前或之后，或者是某个时刻。

## 2.3 市民留言主题和详情的去重

所提取的文本数据，对一些文本词汇出现较多的重复现象，基于对题一的要求描述利用正则表达式进行所有数据去重，并同时将所有非文本字符去除。

## 2.4 对于去重后的中文分词

对于中文文本的数据分词，是基于对文本词性和划分成词的可能性进行判断，对此我们采用 Python 中的中文分词库 jieba 分词，在此基础上对描述关键性的词汇语句进行分词，对于 jieba 算法的三种分词模式：

1.精确模式，试图将句子最精确的切开，适合文本分析；

2.全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解释歧义；

3.搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词；

Jieba 分词算法采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词语情况所构成的有向无环图（DAG），同时采用了动态规划查找最大概率路径，找出基于词频的最大切片组合。在 jieba 分词的同时采用了 TF-IDF 算法的关键词抽取。

```
0                    [西湖, 建筑, 集团, 占, 道, 施工, 安全隐患]
1              [在水一方, 大厦, 人为, 烂尾, 多年, 安全隐患, 严重]
2                  [杜鹃, 文苑, 外, 非法, 汽车, 检测站, 开业]
3      [民工, 明发, 国际, 工地, 受伤, 工, 地方, 拒绝, 支付, 医疗费]
4                                [丁字街, 商户, 乱, 摆摊]
                              ...
490                [卫生局, 药品, 监督局, 乱收费, 邓局, 明查]
491                  [山门, 人民, 医院, 存在, 乱收费, 现象]
492                [人民, 医院, 医务人员, 职称, 晋升, 好难]
493                          [医务, 工作者, 安全, 重要]
494                  [县乡镇, 通, 医务人员, 节假日, 何方]
Name: 留言主题, Length: 495, dtype: object
```

图 2 TF-IDF 算法的关键词抽取图

上图即为分词之后的相应的文本的结果，可以看见有很多对于此次关键词提取无关的无用词语，会对后续的特征词提取，模型建立会造成很大的影响，因此要对接下来的文本作去除停用词处理。

## 2.5 停用词过滤

停用词过滤，是文本分析中一个预处理方法。它的功能是过滤分词结果中的噪声（例如：的、是、啊等）。同时，该类词语的一些特征为：这些功能词极其普遍。记录这些词在每一个文档中的数量需要很大的磁盘空间；由于它们的普遍性和功能，这些词很少单独表达文档相关程度的信息。如果在检索过程中考虑每一个词而不是短语，这些功能词基本没有什么帮助。

这些功能词也可能是高频词，而高频词与高噪声词具有相关性，因此可以采用消除高噪声词对其进行过滤，对于该类词语在不同文本中可以以出现的频率来作为噪声词的衡量标准，但是基于文本数据的数据量以及关键词在文本数据中的高频比例可能会被误判为噪声词，因此我们可以通过停用词表对该文本进行过滤。

```
0      [大道, 西行, 便道, 未管, 路口, 加油站, 路段, 人行道, 包括, 路灯, 杆, ...
1      [位于, 书院, 路, 主干道, 在水一方, 大厦, 一楼, 四楼, 人为, 拆除, 水电,...
2      [市政府, 交警支队, 安监局, 环保局, 区政府, 杜鹃, 文苑, 业主, 涉及, 严重,...
3      [胡书记, 您好, 感谢您, 百忙之中, 查看, 这份, 留言, 父亲, 金星, 北路, 明...
4      [丁字街, 商户, 乱, 摆摊, 前段时间, 丁字街, 交通, 好, 几天, 最近, 丁字街...
                                ...
490    [邓, 局长, 卫生局, 卫生, 监督, 药品, 监督局, 原来, 国家, 不得, 卫生室,...
491    [山门, 人民, 医院, 挂号费, 这几年来, 一直, 高额, 收取, 知道, 是从, 哪一...
492    [人民, 医院, 普通, 医务人员, 想, 一下, 人事科, 领导, 正常, 职称, 晋升,...
493    [张, 厅长, 您好, 一名, 医务, 工作者, 目前, 医患, 系, 尤, 以市, 县级,...
494    [书记, 你好, 县乡镇, 卫生院, 一名, 普通, 医务人员, 医改, 工资, 卫生局, ...
Name: 留言详情, Length: 495, dtype: object
```

**图 3 停用词表过滤图**

## 2.6 TF-IDF 算法

对市民的留言主题和详情进行去重等操作处理之后还需要将文本的内容准换为向量矩阵以便之后的分析模型。因此，我们采用 TF-IDF 算法：TI-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。

**（1）TF 是词频(Term Frequency)：**

词频（TF）表示词条（关键字）在文本中出现的频率。这个数字通常会被归一化（一般是词频除以文章总词数），以防止它偏向长的文件。

$$\text{公式：} \quad tf_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}}$$

$$\text{即：} \quad TF_w = \frac{\text{在某一类中词条 w 出现的次数}}{\text{该类中所有的词条数目}}$$

其中 ni,j 是该词在文件 dj 中出现的次数，分母则是文件 dj 中所有词汇出现的次数总和；

**（2）IDF 是逆向文件频率(Inverse Document Frequency)逆向文件频率 (IDF)：**

某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

公式：

或，逆文档频率（IDF） = $\log(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数+1}})$

其中，|D|是语料库中的文件总数.|{j:ti∈dj}|表示包含词语 ti 的文件数目（即 ni，j≠0 的文件数目）。如果改词语不在语料库中，就会导致分母为 0，因此一般情况下使用 1+|{j:ti∈dj}|.

**（3）TI-IDF 实际上是：TF * IDF：**

某一特定文件内的高词语频率，以及改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 TI-IDF。因此，TI-IDF 倾向于过滤掉常见的词语，保留重要的词语。

公式： TF – IDF = TF * IDF

## 2.7 朴素贝叶斯

在上述将具体的文本数据转换成向量矩阵之后，要利用朴素贝叶斯算法进行判定多个所属类别特征。根据向量空间模型，将每一类别文本训练后得到的该类别的中心向量记为：Cj={ a1,a2,….,an }，将待分类的文本记为 n 维向量的形式：x = {a1,a2,….,an}，a 为 x 的一个特征属性。朴素贝叶斯的表述形式：

$Y=f(x)=\text{argmax}P(Y=c_k|X=x) = \text{argmax}_{ck}\frac{P(Y=c_k\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}$，

当特征项为 x 时，计算所有类别的条件概率，选取条件概率最大的类别作为待分类的类别，根据给出的待分类项，求解在此项出现的条件下，各个类别出现的概率哪个最大，就认为此待分类别属于哪个类别。

### 2.7.1 第一阶段——分类器训练阶段

对于朴素贝叶斯分类的定义步骤：

（1）设 x = {a1,a2,….,an}为一个待分类项，而每个 a 为 x 的一个特征属性，而且特征属性之间相互独立。

（2）设 C={y1,y2….,yn}为一个类别集合。

（3）根据有类别集合 C = {y1,y2,y3….yn}，计算 P(y1|x)，P(y2|x)……P(yn|x)。

（4）根据上述返回的各个类别在其相对的特征属性的条件概率下，生成分类器，计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的

条件概率，如果　P（yk|x） = max{ P(y1|x)，P(y2|x)……P(yn|x)}，则就直接将该类别归为相应的那一类。

可以看出，计算各个划分的条件概率时朴素贝叶斯分类的关键西步骤，当特征属性为离散值时，只要很方便的统计训练样本中各个划分在每个类别中出现的频率即可用来估计 P(a|y)。

### 2.7.2 第二阶段——应用阶段

此阶段是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。

### 2.8 分析模型定义标签所属范围

对于附件二的所给出的有类别集合，进行分类得到相应的特征属性并给予标签并返回其相应的类别的所属范围，所属为那个政府部门的管理，即可以通过传入待分类类别项进行概率估计，根据生成器计算其中各个类别中的各个特征属性的概率估计，判断所属的标签，返回市民的留言是哪个部门所管理的范围，并给予处理。

## 3. 热点问题挖掘

## 3.1 数据描述

通过观察附件 3 所给出的数据，可以看到数据量有 4000 多条，属性有留言编号、留言用户 、留言主题 、留言时间、留言详情、反对数和点赞数，其中留言编号、点赞数和反对数都是数值型（int64），其余均为 object 类型，因此我们需要把留言时间转换成时间序列类型，其余转换成字符串类型。而且数据中含有较多的陌生的地址信息，因为这些地址名词并没有什么规律可言，这对地址的识别有着很大阻碍(类似的还有人物名称，专有名词等)。在附件 3 中的留言主题这一个属性上都是文本数据，因为后面需要对其进行相似度的一些计算，所以需要将其量化成数值形式，因此在挖掘热点问题之前，我们需要对数据做预处理。

## 3.2 流程图

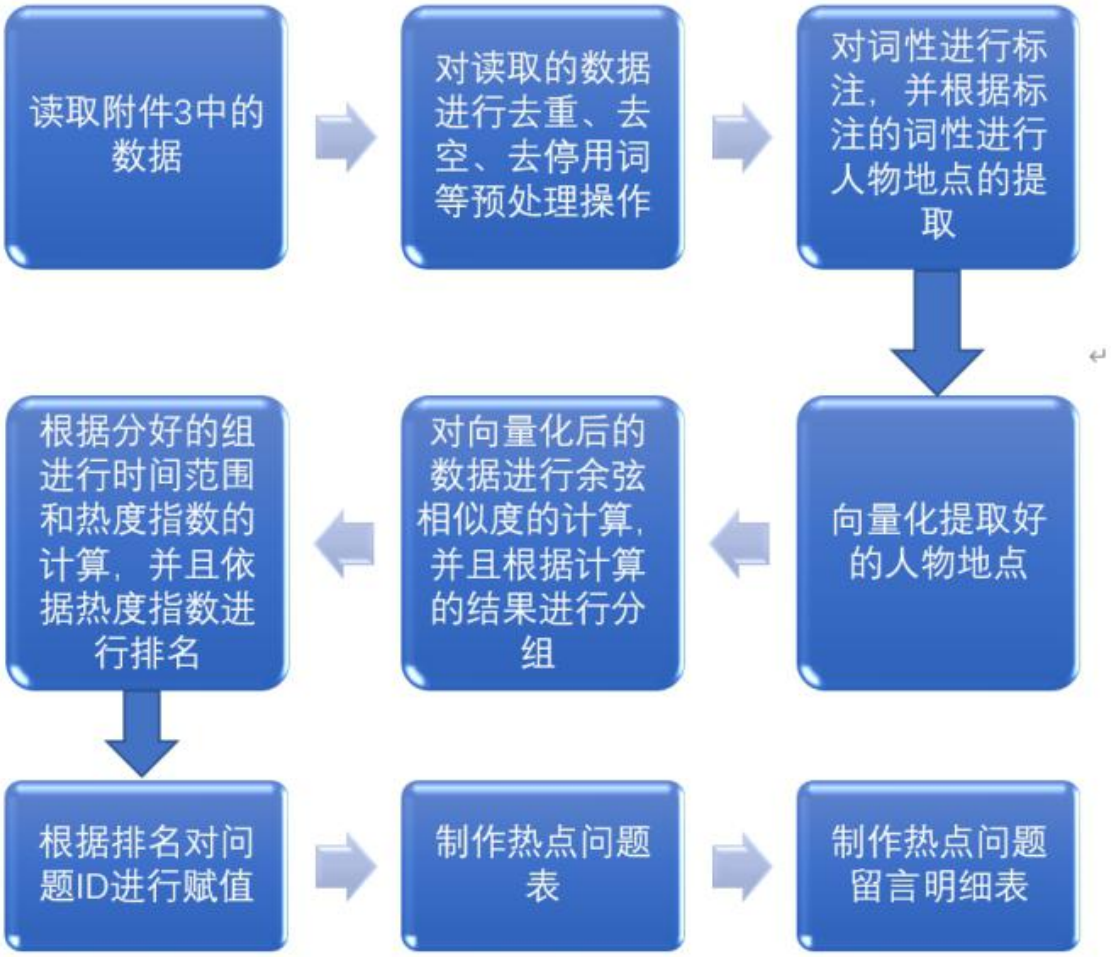


**图 4 总体流程图**

## 3.3 数据预处理

我们可以把这些数据预处理的操作分为以下三个部分：

**（1）去除带有空值的数据：**经过测试之后发现并无空值，因此在这里并没有需要剔除的数据。

**（2）把中文进行分词：**因为中文文本的词与词是相连的，并没有明确的界限，所以我们需要对其进行词。

在这里我们用到的是 python 中的 jieba 库，Jieba 库分词的基本原理：

1、利用中文词库，分析汉字与汉字之间的关联几率；

2、还有分析汉字词组的关联几率；

3、还可以根据用户自定义的词组进行分析；

Jieba 库分词用到的算法：

1、基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；

2、采用了动态规划查找最大概率路径，找出基于词频的最大切分组 3、对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

部分结果如图所示：

```
0                [A3区，一米阳光，婚纱，艺术摄影，是否，合法，纳税，了，？]
1        [咨询，A6区，道路，命名，规划，初步，成果，公示，和，城乡，门牌，问题]
2            [反映，A7县，春华，镇金鼎村，水泥路，、，自来水，到户，的，问题]
3             [A2区，黄兴路，步行街，大，古道，巷，住户，卫生间，粪便，外排]
4       [A市，A3区，中海，国际，社区，三期，与，四期，中间，空地，夜间，施...
                                ...
4321          [A市，经济，学院，寒假，过年，期间，组织，学生，去，工厂，工作]
4322              [A市，经济，学院，组织，学生，外出，打工，合理，吗，？]
4323                        [A市，经济，学院，强制，学生，实习]
4324                    [A市，经济，学院，强制，学生，外出，实习]
4325                [A市，经济，学院，体育，学院，变相，强制，实习]
```

**图 5 Viterbi 算法部分结果图**

**（3）去除停用词**

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。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。

去除停用词后的部分结果如图：

```
0                    [A3区，一米阳光，婚纱，艺术摄影，合法，纳税]
1            [咨询，A6区，道路，命名，规划，初步，成果，公示，城乡，门牌]
2                    [A7县，春华，镇金鼎村，水泥路，自来水，到户]
3             [A2区，黄兴路，步行街，古道，巷，住户，卫生间，粪便，外排]
4       [A市，A3区，中海，国际，社区，三期，四期，空地，夜间，施工，噪音，扰民]
```

**图 6 去停用词结果图**

## 3.4 文本人物地点提取

**（1）词性标注**

由题目所给实例表格可知，我们需要根据相同的地点或人物来划分数据，因此我们需要将人物地点提取出来作为划分数据的依据。通过总结分析可以得知人物名称和地点名称均为名词，因此我们可以根据词性来提取人物和地点。（词性（part-of-speech）是词汇基本的语法范畴，通常也称为词类，主要用来描述一个

词在上下文的作用）。目前采用的词性标注方法主要有基于统计模型的标注方法、基于规则的标注方法、统计方法与规则方法相结合的方法、基于有限状态转换机的标注方法和基于神经网络的词性标注方法。在标注词性这里我们依旧可以选择 jieba 来进行标注，jieba 分词中提供了词性标注功能，可以标注句子分词后每个词的词性，词性标注集采用北大计算所词性标注集，属于采用基于统计模型的标注方法。

jieba 库中名词的词性标注如下图所示:

名词分为以下子类:
n 名词
nr 人名
nr1 汉语姓氏
nr2 汉语名字
nrj 日语人名
nrf 音译人名
ns 地名
nsf 音译地名
nt 机构团体名
nz 其它专名
nl 名词性惯用语
ng 名词性语素

**图 7 jieba 词性标注图**

**（2）添加自定义词典**

由于给出的数据中包含大量的陌生地址、人物名称并且这些名称的命名没有规律，虽然 jieba 有新词识别能力，但是自行添加新词可以保证更高的正确率，因此，我们可以指定自己自定义的词典，以便包含 jieba 词库里没有的词。（在该自定义字典中词的频率设置较高，是为了防止受到 jieba 自带词库的影响）自定义的词典包含的部分词语如下图所示:


**图 8 自定义词典图**

通过以上对数据的处理后，提取出的人物地点效果如下图所示：

```
0                A3区 一米阳光 婚纱 艺术摄影 纳税
1              A6区 道路 命名 规划 成果 城乡 门牌
2                  A7县 春华 镇金鼎村 水泥路
3          A2区 黄兴路 步行街 古道 住户 卫生间 粪便
4              A市 A3区 中海 国际 社区 空地 噪音
                      ...
4321           A市 经济 学院 寒假 学生 工厂
4322                 A市 经济 学院 学生
4323                 A市 经济 学院 学生
4324                 A市 经济 学院 学生
4325                 A市 经济 学院 学院
```

**图 9 提取人物地点效果图**

## 3.5 对数据进行分类

我们可以把对数据分类分为以下三个步骤：

## （一）文本向量化

文本向量化就是将文本表示成一系列能够表达文本语义的向量，是文本表示的一种方式，由于计算机不能够直接处理文本信息，所以我们需要对文本进行处理，将文本表示为计算机能够直接处理的形式，即文本数字化。向量化后的矩阵如下图所示：

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 4346 | 4347 | 4348 | 4349 | 4350 | 4351 | 4352 | 4353 | 4354 | 4355 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4321 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4322 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4323 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4324 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4325 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**图 10 向量化后矩阵图**

## （二）计算相似度

相似度是衡量两个文本之间相似程度的标准。我们可以通过相似度的计算来对上面的数据进行分类，相似度较高的可以分为一类，较低的可以分为不同类。目前相似度计算方法分为距离度量和相似度度量。本文采用的是基于相似度度量的余弦相似度计算。

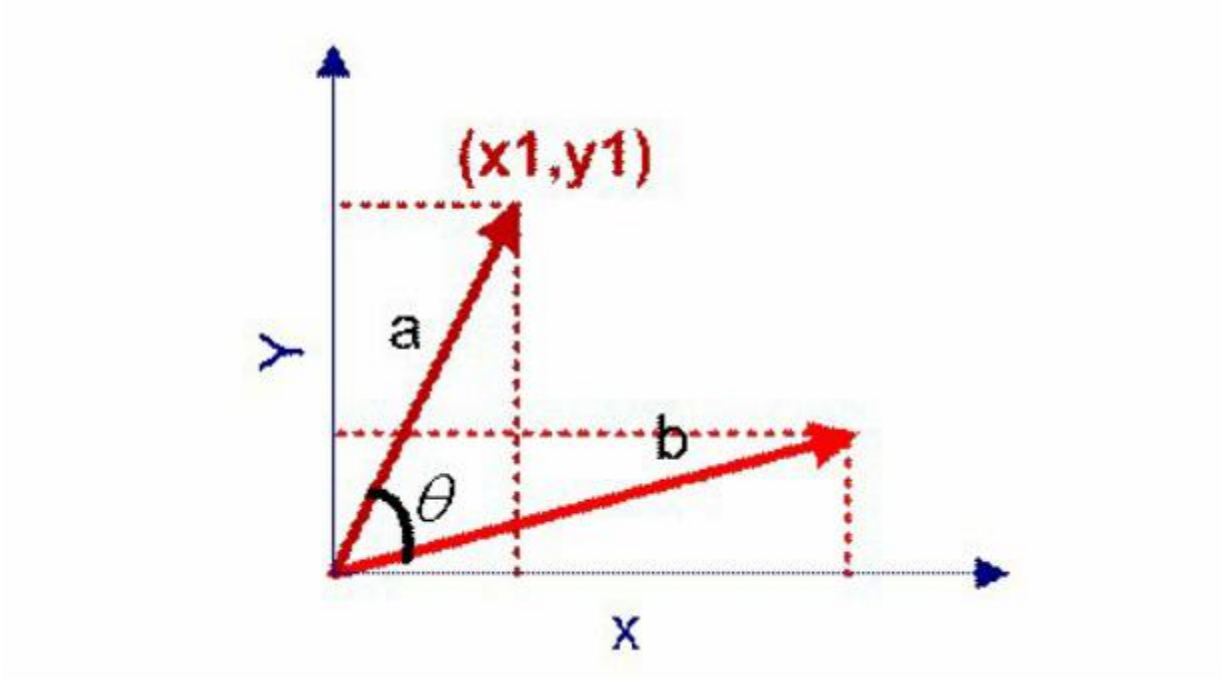


**图 11 余弦相似度计算**

余弦相似度算法：一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值

接近于 0，夹角趋于 90 度，表明两个向量越不相似。以下为余弦计算相似度的公式，其中 Ai 和 Bi 是两个文本向量化后的向量。

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



## （三）分类

经过对相似度阈值的调整，多次测试，设置一个较为合适的阈值，计算出的相似度一旦大于这个阈值，那么我们就可以将其归为同一个地点或者人物的问题，否则就不是同一个地点或者人物的问题。（我们可以通过留言编号的唯一性来区别不同的文本数据）得到结果类似下图：

| 留言编号 |
| --- |
| [194343, 217032, 218132, 220711, 234320, 24055... |
| [208636] |
| [223297] |
| [191951, 202575, 243551, 246974, 263672, 266931] |
| [188119, 188409, 191580, 192440, 193893, 19426... |
| [193091, 226251, 264925] |
| [218442, 231773, 234885, 254865, 262052, 26825... |
| [233542, 239670, 247982, 256358, 261625, 275990] |

图 12 留言编号区分图

## 3.6 时间范围的计算

在我们从文件里读入的留言时间是字符串类型的，因此，如果我们想要对其进行计算需要将其转换为时间类型的数据，在上面我们对数据进行分类的同时，时间类型的数据也同时被我们给分成了不同的组，和上面划分留言编号一样，相同（人物地点）的在一组，之后我们可以通过 pandas 模块中处理时间类型数据的方法来对其经行处理，通过同一组内之间的比较获得最大的时间和最小的时间，之后对其进行格式化，效果如下图所示：

| 时间范围 |
| --- |
| 2019-1-16至<br>2019-7-8 |
| 2019-8-19 |
| 2019-4-11 |
| 2019-8-23至<br>2019-9-6 |
| 2019-1-3至<br>2020-1-7 |
| 2019-6-19至<br>2019-11-12 |

图 13 格式化后时间表

## 3.7 计算热度指数

热度在不同情况下有着不同的含义，在这里主要指的是受到群体大众关注的程度，热度越高说明某件事就越受到大众的关注，可能是某一个问题大家都碰到了，一般情况下是急需解决的，热度低的话则说明关注的人则较少。热度指数是为了我们可以明确的对热度高低进行比较，进而对其进行一个排名。通过观察原始数据我们可以看到有点赞数、反对数，通过上面的分类之后，我们还可以得到相同问题出现的次数，如果我们可以对其进行合理的分类，这些都可以作为量化指标来对热度指数进行计算。在这里，我们把出现的次数用 s 表示，点赞数用 y

表示，反对数用 n 表示，热度指数用 r 表示，我们可以设计以下公式来对热度进行计算：

$$r = s+(y-n)*0.5$$

在计算完热度指数之后，我们需要再根据其大小进行热度的排名，再根据排名对问题 ID 进行赋值，类似效果如下图所示：

| 热度排名 | 问题ID | 热度指数 |
|---|---|---|
| 1 | 1 | 1182.0 |
| 2 | 2 | 1049.5 |
| 3 | 3 | 879.5 |
| 4 | 4 | 350.5 |
| 5 | 5 | 134.5 |
| 6 | 6 | 124.0 |
| 7 | 7 | 92.5 |

**图 14 热度排名与 id 赋值效果图**

## 3.8 制作热点问题表

我们虽然已经对热点问题进行了分类、提取、计算热度指数，但是直接观看这些数据并不是特别的方便，因此，为了观看更加的直观，我们需要制作一张热点问题表来方便我们的观看。在经过数据预处理、人物地点提取、时间范围计算等对数据的操作之后，我们已经准备好了制作热点问题表的相关操作，最后得到的效果如下图所示：

| 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 | 1182.0 | 2019-1-16至2019-7-8 | A市车贷案警官 | 承办A市58车贷案警官应跟进关注留言 |
| 2 | 2 | 1049.5 | 2019-8-19 | 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 |
| 3 | 3 | 879.5 | 2019-4-11 | A市金毛湾入学 | 反映A市金毛湾配套入学的问题 |
| 4 | 4 | 350.5 | 2019-8-23至2019-9-6 | A4区绿地小区渝长厦高铁 | A4区绿地海外滩小区距渝长厦高铁太近了 |
| 5 | 5 | 134.5 | 2019-1-3至2020-1-7 | A市地铁用工 | 对A市地铁违规用工问题的质疑 |
| 6 | 6 | 124.0 | 2019-6-19至2019-11-12 | A市富绿物业丽发新城业主家水 | A市富绿物业丽发新城强行断业主家水 |

图 15 热点问题表

## 3.9 制作热点问题留言明细表

在之前我们已经制作了热点问题表，但是省去了很多留言的细节，如留言编号、留言信息、每一条数据的点赞数和反对数等等，为了方便以后对问题的研究，我们需要制作一张热点问题留言明细表。在之前我们已经对留言编号进行了分组，因此我们需要给在同一组里的数据赋上相同的问题 ID，得到的效果如下图所示：

| | 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 |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 272858 | A00061787 | A市58车贷恶性退出案件为什么不发布案情进展通报? | 2019/1/16 23:21:21 | 唐局长，您好。我是A市58车贷恶性退出案件的受害人，我认为您是知道58车贷案件的，因为在20... | 0 | 0 |
| 1 | 1 | 240554 | A00029163 | A市58车贷老板跑路美国，经侦拖延办案 | 2019/2/10 20:58:40 | A4区经侦毛浚涉嫌58车贷保护伞2018年8月6日，A市58车贷(西地省展星投资有限公司)爆... | 6 | 0 |
| 2 | 1 | 234320 | A000106592 | 不要让A市因为58车贷案件而臭名远扬 | 2019/7/8 17:16:57 | 胡书记：您好反映关于西地省A市58车贷恶性退出案件的进展情况，到现在还没收到回复。因此再次留... | 0 | 0 |
| 3 | 1 | 220711 | A00031682 | 请书记关注A市A4区58车贷案 | 2019/2/21 18:45:14 | 尊敬的胡书记：您好!A4区p2p公司58车贷，非法经营近四年。在受害人要求下，于去年8.20... | 821 | 0 |
| 4 | 1 | 218132 | A000106090 | 再次请求过问A市58车贷案件进展情况 | 2019/1/29 19:15:49 | 尊敬的胡书记：您好！西地省A市58车贷是一家P2P平台，2018年8月6日在平台公告说良性退... | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 95 | 7 | 231773 | A00010141 | 反对A6区月亮岛路架设高压电线，强烈要求重启环境评估 | 2019/4/12 14:59:14 | A市电力局在月亮岛路绿化带，架设110kv高压电缆。长郡月亮岛的学生每天上学都需要穿越高压电... | 1 | 0 |
| 96 | 7 | 262052 | A00072424 | 关于A6区月亮岛路沿线架设110kv高压线杆的投诉 | 2019/3/26 14:33:47 | 联名信——坚决要求A市润和又一城、三润城、润和紫郡、润和长郡、润和美郡、润和星城、润和滨江府... | 78 | 0 |
| 97 | 7 | 234885 | A00060375 | A6区月亮岛路11万伏高压线没用地埋方式铺设 | 2019/4/5 13:01:17 | 月亮岛路两旁小区较多，都是高楼，架空线路不仅存在安全隐患，而且破坏市容！从长远发展看，建议用... | 2 | 0 |

图 16 热点问题留言明细表

从图中可以看出相同的问题 ID 有着类似的地点或者人物，可以直观的看到每一条留言的留言时间、点赞数和反对数。

# 4 回复意见评价模型

及时对回复意见进行评价，有助于相关部门迅速反馈改良回复文本，有针对性的提升服务质量，从而提高群众满意度。通过生成简明摘要并对比与原留言主题之间的相关性、完整性和可解释性等质量评价标准，在自动评价指标的基础上，建立一种基于词嵌入的文本自动摘要和对比评价框架，以满足任务要求。

- 提出了一个名为 WEEM4TS 的自动评估指标，用于评估回复意见与原文相关性的系统性能。WEEM4TS 的目的是根据原始文档的保留意义来评估简明摘要的质量。因此，认为它代表了适用于所有类型系统总结的评估指标:提取、抽象和混合
- 提出了一种称为 WETS 的方法，用于确定原始文档中最重要的句子，以评估回复意见的完整性和可解释性。

## 基于字嵌入的文本摘要

**（1）算法解释**

我们提出了基于词嵌入的文本摘要(WETS)方法，用于识别、排列和连接突出的 top-y 句，作为简明摘要。

最常用的句子关联判断方法是对出现频率高、语用频繁的词语进行检测。鉴于此，由这些词组成的句子被认为是最重要的。然而，我们认为这种技术有两个主要缺点:首先，它在新的摘要中鼓励冗余;其次，对于由其他单词组成的非常重要的句子，它没有给出适当的分数。因此，建立冗余处理机制和面向意义的句子关联评价技术至关重要。

因此，初步的任务是从原始文档中删除不相关的标记，如停止词。然后，将第一句的单词和常用词结合起来作为关键词。倾向于关注第一句话的单词的主要原因是，语言学文献显示了一个明确的论题陈述，大部分位于段落的开头，这表明重要的单词可能存在于第一句话中。此外，相对而言，标题中至少有几个词也可能出现在第一句话中。比较分析可以通过比较中间句中的词和最后一句中的词来进行。

**（2）算法步骤**

输入:原始文本、词嵌入模型、停止词

输出:Top-y 突出句子

1: for i in range (原始文本长度):

2:　设原始文本中的对应句子为第 i 句，如为原始文本起始句则作为第一句。

3:　for sentence in sentences:

简单预处理以保存句中词向量实现标记化，

标记化后移除停止词实现不包含停止词的句子

4:  for m in range(第一句话的长度):

5:  如果第一句话不包含停止词，则将第一句话作第一个关键词

6:  原始文本中的高频词作为第二关键词

7:  由第一关键词和第二关键词构成关键词库

8:  设定初始权重为 0，自权重为 0

9:  for n in range (无停止词句子长度):

由标记和关键词的余弦相似值得出最大权重

句子权重=原权重+计算出的最大权重

10: 相关性分值= 句子权重 / 无停止词句子长度

11: 根据相关性分值将句子排序得到 top-y

12:返回 top-y

将得到的每个句子中所有单词的余弦相似值(权值)相加，然后除以相应句子更新后的长度。更新后的长度是去掉冗余词和停止词后的句子长度。根据相关度得分，将句子从上到下排序。最后，按照所需的长度连接第 y 个句子。值得注意的是，虽然句子关联分数是基于语义相似度计算的，但 WETS 是一个提取文本摘要的系统。在本研究中，文本摘要的长度是可变的，即，可根据所需的系统摘要长度进行调整。这有助于比较不同长度的系统摘要。例如，假设原始文本由 200 个字词组成;文本摘要系统的 if 摘要:A 和 B 分别由 150 和 100 个字词组成。然后，通过从上到下选择所需的字数，可以将文本摘要的长度调整为系统 A 的 150，系统 B 的 100。然而，更长的系统概要可能更受欢迎。为了控制这一点，系统摘要的长度应该符合预先确定的阈值或压缩比。

## 基于词嵌入的回复文本自动评价指标

**（1）算法解释**

提出了一个基于词嵌入的文本摘要评价指标(WEEM4TS)，参见算法如下。WEEM4TS 由三个部分组成:预处理、单词加权、计算修改后的单字符回忆

**（2）算法步骤**

输入: 参考文本，系统总结，单字调用，词嵌入模型，停止词

输出：指标得分

1: 句子权重 ，权重，计数清零

2:  for i in range(参考文本长度):

3:  for n in range(系统总结):

if 系统总结[i][n] in 参考文本[i]:

权重为 1

句子权重=原句子权重+权重

else if 系统总结[i][n] in 词汇表(词嵌入模型):

权重→最大值(余弦相似度(系统总结[i][n], 词汇表))

句子权重=原句子权重+权重

else:

权重为 0

句子权重=原句子权重+权重

4: if 参考文本[i]长度>0:

单字调用 句子权重/参考文本[i]长度

5: else:

单字调用为 0

单位片段精确度值 (相邻双字技术/(系统总结[i]长度− 1))*100

单字调用 单字调用*100

WEEM4TS 得分 (α*单字调用 ) + (β*单位片段精确度值)

6: return WEEM4TS 得分

从系统和参考摘要中删除不相关的单位，如停止单词和字符。利用词嵌入模型，计算系统中词与参考文献的余弦相似度。因此，如果一个单词被系统和简介摘要共享，它将获得+ 1 分。如果系统摘要中的目标词没有出现在参考摘要中，而是存在于嵌入词汇表中，则将该词与向量空间中最近的词之间的余弦相似度值视为目标词的权值。如果两者都没有发生，则目标单词得分为 0

## 5.参考文献

[1]路永和，李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作, 2013(03):92-97.
[2]邓乃扬 田英杰. 数据挖掘中的新方法 ：支持向量机[M]. 科学出版社, 2004.
陆尹浩. 一种基于 Word2Vector 与编辑距离的句子相似度计算方法[J]. 电脑知识与技术, 2017(5).
[3]王晓宇，熊方，凌波,等. 一种基于相似度分析的主题提取和发现算法[J]. 软件学报, 2003(09):79-86.
[4]付慧，刘峡壁，贾云得. 基于最大-最小相似度学习方法的文本提取[J]. 软件学报, 2008(03):149-157.
[5]蒙晓燕，殷雁君. 基于 word2vec 的中文歌词关键词提取算法[J]. 内蒙古师范大学学报(自然科学汉文版), 2018, v.47;No.190(02):50-53.
[6]施聪莺，徐朝军，杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 029(B06):P.167-170,180.
[7]田瑞，闫丹凤. 针对特定主题的短文本向量化[J]. 软件, 2012(11):210-213.
[8]于政. 基于深度学习的文本向量化研究与应用[D]. 2016.
[9]解宇涵. 基于深度学习的中文分词模型应用研究[D].
[10]高岩. 朴素贝叶斯分类器的改进研究[D]. 华南理工大学.
[11]王煜，王正欧，白石. 用于文本分类的改进 KNN 算法[J]. 中文信息学报, 2007, 21(3):76-82.
[12]亚力青·阿里玛斯，哈力旦·阿布都热依木，陈洋. 基于向量空间模型的维吾尔文文本过滤方法[J]. 新疆大学学报(自然科学版), 2015(02):99-104.