“智慧政务”中的文本挖掘应用

摘要：近年来，随着网络问政平台逐步成为政府了解民意、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。因此，对民意相关的文本数据进行分析研究，快速分类并处理群众的留言，及时了解群众的热点问题，评价相关部门对群众留言的答复意见，对提升政府的管理水平和施政效率具有极大的推动作用具有非常重要的意义。本作品根据赛题要求，分别对群众留言分类、热点问题挖掘、答复意见评价三个问题提出解决方法并对结果进行评估与验证。在问题一中，首先对数据进行预处理，并将文本向量化，利用主成分分析方法对数据降维，在模型训练中利用神经网络 MLPClassifier 方法完成监督学习，最后采用 F1 分数（F1 Score）进行方法检验，利用 k 折交叉辅助检验。在问题二中，首先采用了 DBSCAN 聚类算法对问题类别进行聚类，通过提取文档摘要绘制点赞数、反对数的散点图，利用拉依达准则处理附件 3 中离群值，结合影响热度的因素，定义合适的热度评价指标，得出热点问题。在问题三中，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。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。

关键词：留言分类、主成分分析、热点问题挖掘、意见评价

Text Mining Application in "Smart Government Affairs"

Abstract: In recent years, as the online political inquiry platform has gradually become an important channel for the government to understand public opinion, gather people's wisdom, and gather people's popularity, the amount of text data related to various social conditions and public opinion has been increasing. Therefore, the analysis and research of text data related to public opinion, rapid classification and processing of mass messages, timely understanding of hot issues of the masses, and evaluation of relevant departments ’responses to mass messages have great impact on improving the government ’s management level and governance efficiency The role of promotion is very important. According to the requirements of the competition questions, this work proposes solutions to the three questions of the people's message classification, hot spot problem mining, and answer evaluation, and evaluates and verifies the results. In the first problem, first preprocess the data, vectorize the text, use principal component analysis to reduce the dimension of the data, use neural network MLPClassifier method to complete the supervised learning in the model training, and finally use F1 Score Method test, using k-fold cross auxiliary test. In the second question, the DBSCAN clustering algorithm is used to cluster the question categories. The scatterplot of the likes and antilogs is drawn by extracting the document summary. Factors of heat, define appropriate heat evaluation index, and draw hot issues. In question three, an automatic evaluation indicator called WEEM4TS was proposed to evaluate the system performance of the relevance of the response to the original text. A method called WETS is proposed to determine the most important sentences in the original document to evaluate the completeness and interpretability of the reply comments

Keywords: message classification, principal component analysis, hot spot problem mining, opinion evaluation

目 录

一、前言

（一）研究背景与意义

目前，我们已经步入了“互联网+”的生活时代，网络问政平台逐步成为政府了解民意的重要渠道，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。

然而，传统的对信息的处理方式无法满足文本数据量不断攀升和快速发展的要求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，导致相关部门对部分群众留言的答复意见无法做出及时和精准的回复、给出相应的解决方案。

同时，社情民意的动态监测要尽可能早的主动去发现民意的变化情况，一遍组织及时做出相应的调整，及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。因此，如何高效和准确的对群众留言进行分类和处理、掌握某一时段群众集中反映的热点问题成为提高，对提升政府的管理水平和施政效率的新问题。

在本文中，我们对传统的文本数据处理方式进行了改进，运用文本分析技术取代依靠人工经验的处理方式，在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派到相应的职能部门处理，同时对于某一时段群众集中反应的某一问题，及时发现，有助于相关部门进行有针对性地处理，提升服务效率。同时制定一套答复意见的评价方案，规范留言回复。

（二）需求分析

本文的首要目标是构建基于文本内容建立识别和挖掘模型，建立基于自然语言处理技术的智慧政务系统，为网络问政平台及相关部门提供快速处理留言、热点挖掘、意见评价等服务。现今对新闻专题的研究，大多都提出新闻热点的发现，较少的对新闻内容的问题分析。本文提出对新闻内容关注的问题进行分析，实现了解当前的舆论焦点和民意的目的。解决各类社情民意相关的文本数据量不断攀升和快速发展带来的大数据时代的信息超载问题，因此，对于网络问政平台而言面临着以下问题：

（1）群众留言的多样性

对于不同类别问题的留言信息，对于同一类别问题的不同表达形式，需要对其进行处理，并按照一定的划分体系对留言进行较为准确的分类，以便后续将群

众留言的问题分派到相应的职能部门高效处理。

（2）热点问题的实时性

某一时段内群众集中反映某一问题，社情民意的动态主动监测要求尽可能早的主动去发现社情民意的变化情况，再对热点进行分析，通过对某一热点相关词汇的聚类，得到热点问题所涉及的人物、行业或组织等，实现了解当前的舆论焦点和民意的目的，相关部门进行有针对性地处理，提升服务效率。

（3）答复意见的精确性

针对于相关职能部门对群众留言问题答复，要建立一定的模型对答复信息校对其完整性、准确性、相关性、时效性、可信性和可解释性，从而实现答复意见评估体系，以此来检验相关职能部门的政策落实情况，实现对不同职能部门的政绩评价体系。

二、分析方法与过程

本文主要通过文本挖掘技术进行新闻热点问题分析，为网络问政平台及相关部门提供快速处理留言分类、热点问题挖掘、意见评价等服务。总体流程图如图 1 所示：

图 1 总体流程图

步骤一：对附件 2、附件 3 中留言主题做去除符号、重复项处理，jieba 中文分词，载入自建词库、并用停用词过滤，TF-IDF 特征提取，并将数据划分为训练集和测试集，再文本向量化。

步骤二：预处理后，通过 smote 对数据平衡化后，对数据进行标准化并利用 PCA 法对数据降维，神经网络中的 MLP Classifier 结合学习曲线和训练曲线，进行模型训练，利用 F1 Score 综合评估模型。

步骤三：根据据文本向量，基于 DBSCAN 聚类算法对各个问题类别进行聚类，利用 snownlp 提取文档摘要，建立合适的热度评价指标，来得出热点问题。

步骤四：针对附件 4 相关部门对留言的答复意见，得到相关度分值，从相关性、完整性、可解释性等角度对答复意见进行评价。

三、数据探索与预处理

（一）数据探索

通过观察所给数据，可以发现数据量比较大，且附件 2、附件 3 中的字段大多为文本格式(附件二与附件三预处理部分大致相同，下文以处理附件二为例)，并且文本信息存在大量噪声特征，如果不做处理会对后续分析造成影响，则必然会对结果的质量造成很大的影响，于是本文首先要对数据进行预处理。同时对附件文本完全一致的样本，做去重处理。

（二）数据预处理

（1）剔除符号

对文本进行结巴分词之前，利用正则表达式，对数据进行空行、符号删除，例如：“，。！？”使数据保留有价值的信息同时让数据更加简洁，去除掉这些非文本的内容后，我们就可以进行真正的文本预处理了。

（2）jieba 分词

我们开始进行分词。由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，本文采用 Python 开发的一个中文分词模块——jieba 分词，jieba 分词有着较好的分词精度，我们采用改分词系统进行分词达到的效果也较为理想。jieba 分词原理如图 2 所示：

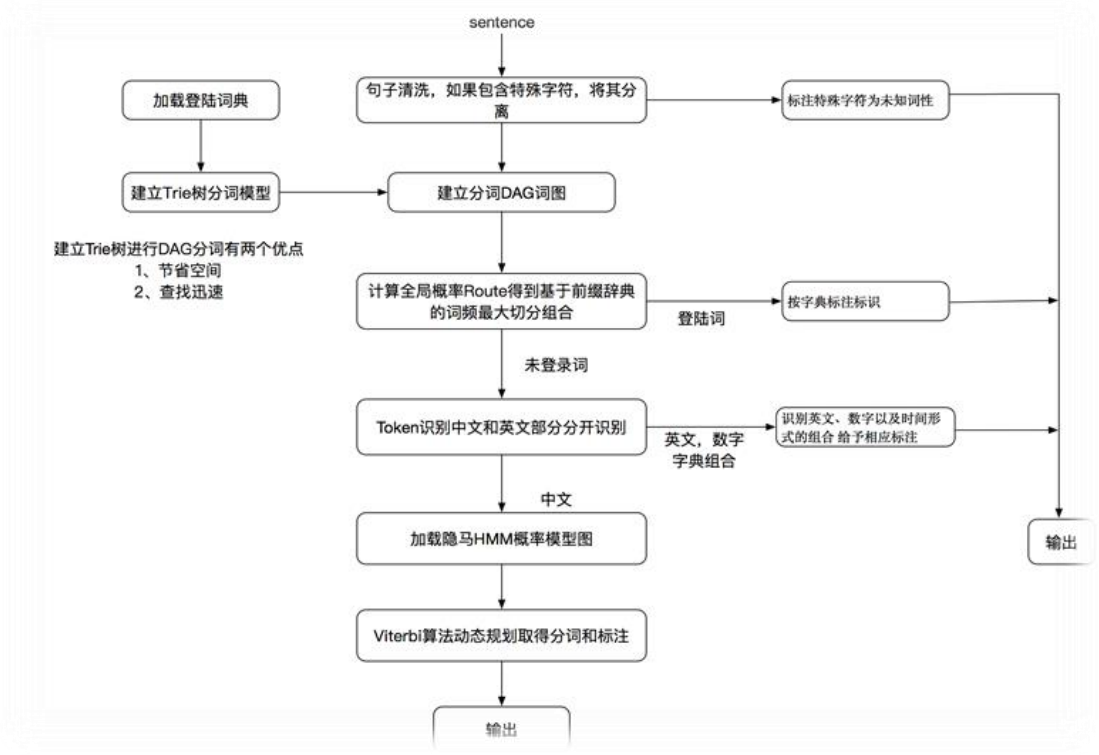

图 2 结巴分词原理图

分词结果如图 3 所示：

```
0                    [A，市，西湖，建筑，集团，占，道，施工，有，安全隐患]
1                  [A，市，在水一方，大厦，人为，烂尾，多年，安全隐患，严重]
2                      [投诉，A，市，A1，区苑，物业，违规，收，停车费]
3                [A1，区，蔡锷，南路，A2，区华庭，楼顶，水箱，长年，不洗]
4                      [A1，区，A2，区华庭，自来水，好大，一股，霉味]
                                    ...
9205                   [两，孩子，一个，是，一级，脑瘫，能，再，生育，吗]
9206    [B，市中心，医院，医生，不负责任，做，无痛，人流，手术，后，结果，还..
9207                        [西地省，二胎，产假，新，政策，何时，出台]
9208                                [K8，县惊现，奇葩，证明]
9209             [请问，J4，县卫，计委，社会，抚养费，到底，该交，多少，钱]
Name: topic, Length: 9210, dtype: object
```

图 3 结巴分词图

由于留言存在这一定的的特殊性，其中存在着大量口语化的词语以及一些热点话题所特有的人名、事件名称和地名，直接使用 jieba 分词无法达到较好的分词效果，而分词的效果对后续的建模分析影响较大。自行添加新词可以保证更高的正确率，对此以及我们从清华大学官网、国家统计局等平台中获取了常用的的政府机关、高校名称等词典，根据文本内容我们自定义了 lexicon.txt 词典。添加词典后效果如图 4 所示：

```
0                      [A市，西湖，建筑，集团，占道，施工，有，安全隐患]
1                    [A市，在水一方，大厦，人为，烂尾，多年，安全隐患，严重]
2                        [投诉，A市，A1区，苑，物业，违规，收，停车费]
3                    [A1区，蔡锷南路，A2区，华庭，楼顶，水箱，长年，不洗]
4                        [A1区，A2区，华庭，自来水，好大，一股，霉味]
                                    ...
9205                   [两，孩子，一个，是，一级，脑瘫，能，再，生育，吗]
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...
9207                        [西地省，二胎，产假，新，政策，何时，出台]
9208                                [K8县，惊现，奇葩，证明]
9209             [请问，J4县，卫，计委，社会，抚养费，到底，该交，多少钱]
Name: topic, Length: 9210, dtype: object
```

图 4 添加词典后分词图

（3）停用词过滤

观察分词结果可知，为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，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，比如：啊，哦，的，地，得，我，你等。通过配置 stop_word 文件，我们利用文件停用词来过滤停用词，将分词结果与停用词表中的词语进行匹配，若匹配成功，则进行删除处理。

去除停用词后的部分结果如图 5 所示：

```
: 0              [A市，西湖，建筑，集团，占道，施工，安全隐患]
  1            [A市，在水一方，大厦，人为，烂尾，多年，安全隐患]
  2                          [A市，A1区，物业，停车费]
  3        [A1区，蔡锷南路，A2区，华庭，楼顶，水箱，长年，不洗]
  4             [A1区，A2区，华庭，自来水，好大，一股，霉味]
                            ...
  9205                     [孩子，一个，一级，脑瘫，生育]
  9206       [B市，中心医院，医生，不负责任，无痛人流，手术，胚芽]
  9207                     [西地省，二胎，产假，政策，出台]
  9208                       [K8县，惊现，奇葩，证明]
  9209             [J4县，卫计委，抚养费，到底，该交，多少钱]
  Name: topic, Length: 9210, dtype: object
```

图 5 添加停用词后分词图

（4）基于 TF-IDF 算法的文本特征提取技术

经过上述处理后，虽然已经去掉部分停用词，但还是包含大量词语，给文本向量化过程带来困难， 特征抽取的主要功能是在不损伤文本核心信息的情况下尽量减少要处理的单词数，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。

常用的传统方法 ：词频-逆向文档频率(TF-IDF)，信息增益(IG)，互信息(MI)，X^2(CHI)统计法等等。上述方法各有利弊。因此本文采用目前公认的比较有效的 TF-IDF 算法抽取特征词条。

这里我们就用 scikit-learn 的 TfidfVectorizer 类来进行 TF-IDF 特征处理。

TF-IDF 算法：TF-IDF 技术采用一种统计方法，根据字词的在文本中出现的次数和在整个语料中出现的文档频率来计算一个字词在整个语料中的重要程度在 TF-IDF 中，单词的重要性由两个因素共同决定，它与它在文档中出现的次数成正比，但它随着语料库中出现该词的频率越多而下降。

- TF(term frequency) 词频统计：

区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，因此引入 TF，计算单词的词频

$$\text{词频(TF)} = \frac{\text{某个词再文章中的出现次数}}{\text{文章的总次数}}$$

- IDF(inverse document frequency) 逆文本频度：

一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度 IDF 的概念。如果一个词越常见，那么分母就越大，逆文档频率就越小越接近 0，说明这个词不那么重要。为了避免某词可能从来都没出现在所有的文档中，而导致被除数为零，一般分母用（包含该词的文档数+1）代替。

$$\text{逆文档频率(IDF)} = \log\left(\frac{\textit{总样本数}}{\textit{包含有该词的文档数}+1}\right)$$

- TF-IDF

TF 和 IDF 的乘积作为特征空间坐标系的取值测度。使用 IDF 作为权重乘以 TF，实现对单词权重的调整，调整权值的目的在于突出重要单词，抑制次要单词。

TF-IDF = 词频(TF)*逆文档频率(IDF)

同时我们利用标注词性的方式，删除了指定的词性词、去除了无用单字对于附件 2 删除了地名。(代码见特征提取.py 文件)

特征提取后结果如图 6 所示：

```
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]
2                              [物业, 停车费]
3            [蔡锷南路, 华庭, 楼顶, 水箱, 长年, 不洗]
4                 [华庭, 自来水, 好大, 一股, 霉味]
                        ...
9205                [孩子, 一个, 一级, 脑瘫, 生育]
9206    [中心医院, 医生, 不负责任, 无痛人流, 手术, 胚芽]
9207              [西地省, 二胎, 产假, 政策, 出台]
9208                        [惊现, 奇葩, 证明]
9209           [卫计委, 抚养费, 到底, 该交, 多少钱]
Name: topic, Length: 9210, dtype: object
```

图 6 特征提取后分词图

上图是采用 TF-IDF 算法后的得到的特征词，从结果可以看出，特征词与附件 2 的留言主题接近，说明特征词采取效果良好。

（5）划分训练与测试集

为了防止过度拟合，同时为通过实验测试对学习器的泛化误差进行评估，然后以测试集上的“测试误差”作为泛化误差的近似。最直接的方法是从一堆的数据集中直接划分出两部分，一部分是训练集，另一部分就是测试集，在机器学习中，我们从 sklearn.model_selection 中调用 train_test_split 函数，将原始数据集按照按照一定比例划分为测试集和训练集，运用机器学习传统方法的时候，一般将训练集和测试集划为 7：3。

```
def get_split(keys, test_size=0.3):
    """
    划分训练测试集
    Parameters
    ----------
    keys: tuple, 需要进行划分的数据及其对应的标签
    test_size: float, 测试集所占比例

    Returns
    -------
    topic_train: Series, 划分的训练集
    topic_test: Series, 划分的测试集
    genre_train: Series, 训练集对应的标签
    genre_test: Series, 测试集对应的标签
    """
    topic, genre = keys
    topic_train, topic_test, genre_train, genre_test = model_selection.train_test_split(topic, genre,
                                                                                        test_size=test_size, stratify=genre)
    return topic_train, topic_test, genre_train, genre_test
```

（5）文本向量化

文本表示是自然语言处理中的基础工作，文本表示的好坏直接影响到整个自然语言处理系统的性能。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，是文本表示的一种重要方式。我们结合部分腾讯 AI Lab 开源的词向量，对于腾讯词向量不包括的，我们用 word2vec 训练词向量得到词向量模型。

腾讯 AI Lab 词向量超过 800 万中文单词和短语提供了 200 维矢量表示它使用 Directional Skip-Gram（Skip-Gram 的改进版）训练而成，可使用 ginsim 调用，这些词和短语都经过大规模高质量数据的预培训，可广泛应用于许多下游的中文处理任务和进一步研究。与现有的中文嵌入语种相比，AI Lab 语库的优势主要在于覆盖面、新鲜度和准确性。

1) 覆盖率（Coverage）

该词向量数据包含很多现有公开的词向量数据所欠缺的短语，比如“不念僧面念佛面”、“冰火两重天”、“煮酒论英雄”、“皇帝菜”、“喀拉喀什河”等。

2) 新鲜度（Freshness）

该数据包含一些最近一两年出现的新词，如“恋与制作人”、“三生三世十里桃花”、“打 call”、“十动然拒”、“供给侧改革”、“因吹斯汀”等。

3) 准确性（Accuracy）

由于采用了更大规模的训练数据和更好的训练算法，所生成的词向量能够更好地表达词之间的语义关系。腾讯 AI Lab 采用自研的 Directional Skip-Gram 算法作为词向量的训练算法。DSG 算法基于广泛采用的词向量训练算法 Skip-Gram 在文本窗口中词对共现关系的基础上，额外考虑了词对的相对位置，以提高词向量语义表示的准确性。

Word2vec 是 Google 于 2013 年开源的一个用于训练获取词向量的工具包，它简单、高效，因此备受欢迎和关注。word2vec 主要分为 CBOW（连续词袋模型）和 Skip-Gram（跳字模型）两种训练模式：

1) CBOW 模型：用 context(w)去预测 w，目标是最大化概 p(w|context(w))
2) Skip-gram 模型：用 w 去预测 context(w)，目标是最大化概 p(context(w)|w)
   context(w)是指词汇 w 的上下文，如果设置阈值为 N，那么 context(w)指的就是句子中 w 的前 N 个词和 w 之后的 N 个词。

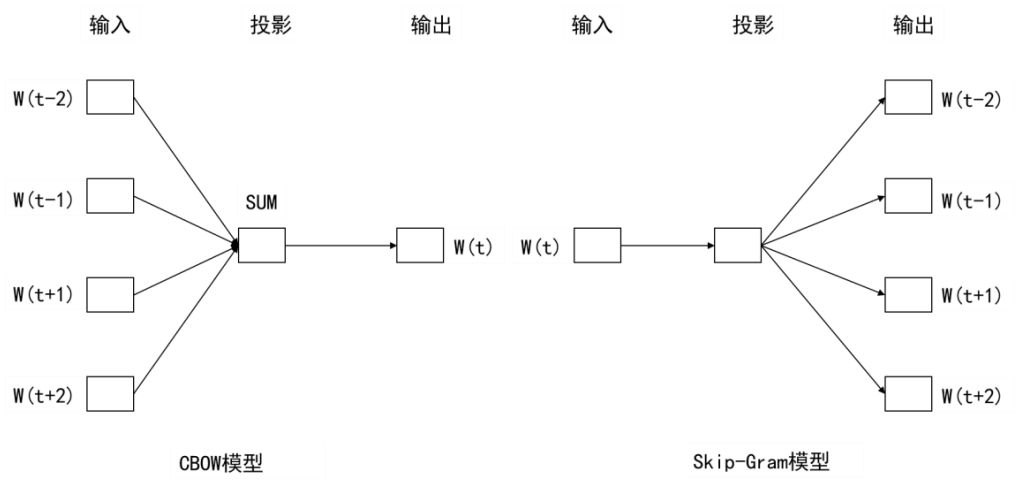

图 7 Word2vec 两种训练模型

由图 7 可见，word2vec 的两种训练模式其实是很类似的，CBOW 模型是将单词 W(t)的上下文作为输入，从而预测单词 W(t)；而 Skip-Gram 模型是由单词 W(t)作为输入，从而来预测出单词 W(t)的上下文。

由附件训练 word2vec 词向量模型的过程如图 8 所示

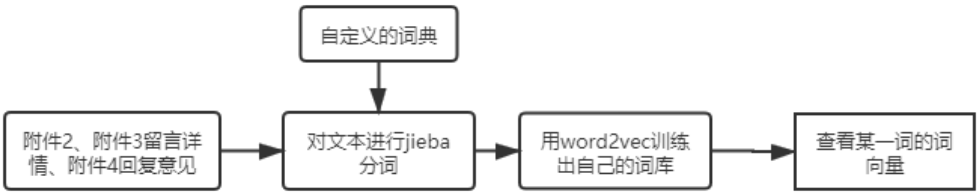

图 8word2vec 训练词向量

由图可见，本文运用附件 2、附件 3、附件四进行分词，接着进行训练，得到词向量模型。

四、群众留言分类模型

（一）数据平衡化

（1）引言

对附件二给出的一级标签统计发现，如图 9 所示数据集中样本类别不均衡，即数据倾斜。

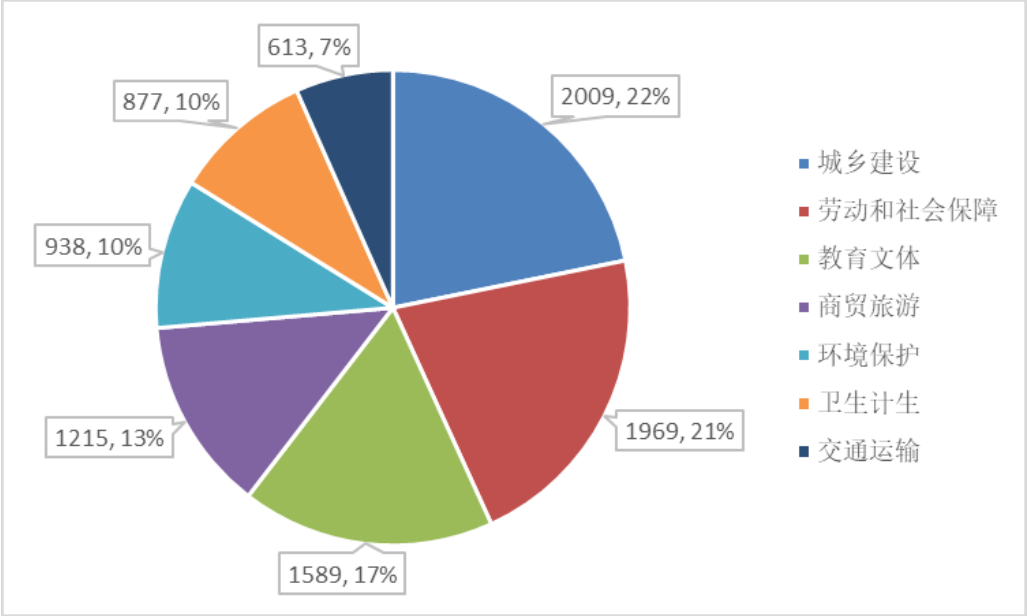

图 9 一级标签类别占比

针对这一问题，主要有两个方法，一是欠抽样，顾名思义就是删除正样本（以正样本占绝大多数为例）中的样本，会删除正样本所带的信息，当正负样本的比例悬殊时，需要删除较多的正样本数量，这会减少很多正样本携带的信息。一种过抽样的方法是随机采样，采用简单随机复制样本来增加负样本的数量。这样容易产生模型的过拟合问题，即使得模型学习到的信息过于特别而不够泛化。

因此我们采用 SMOTE（Synthetic Minority Oversampling Technique）即合成少数类过采样技术来解决数据不平衡问题。

（2）SMOTE 方法介绍

SMOTE 是一种对普通过采样(oversampling)的一个改良。普通的过采样会使得训练集中有很多重复的样本，SMOTE 没有直接对少数类进行重采样，而是设计了算法来人工合成一些新的少数类的样本。

样本本身就是在特征空间的一些点，所以该算法用于增加样本的方法就是在特征空间中两个同类点之间随机选取一个点，这个点就是一个新样本了，和另外两个点具有相同的类别。算法流程如下：

1) 对于少数类中每一个样本 x，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其 k 近邻。

2) 根据样本不平衡比例设置一个采样比例以确定采样倍率 N，对于每一个少数类样本 x，从其 k 近邻中随机选择若干个样本，假设选择的近邻为 Xn。

3) 对于每一个随机选出的近邻 Xn，分别与原样本按照如下的公式构建新的样本。

$$x_{new} = x_i + random(0,1) * (\hat{x} - x_i)$$

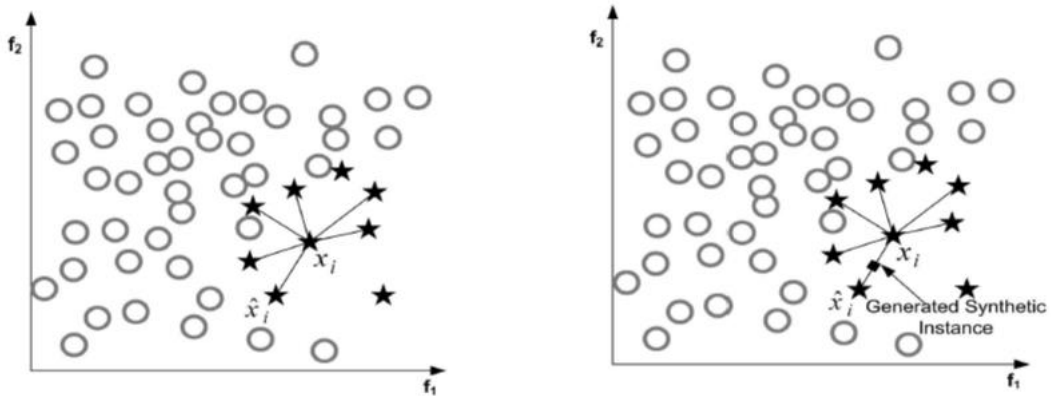

图 10 Smote 分析方法

代码如图所示:

```
def balance(vec, genre):
    """
    数据平衡化
    Parameters
    ----------
    vec: array, 需要平衡化处理的文本向量
    genre: Series, 文本对应的标签

    Returns
    -------
    平衡化处理后的文本向量及其对应的标签
    """
    smote = over_sampling.SMOTE()
    # ros = over_sampling.RandomOverSampler()
    return smote.fit_sample(vec, genre)
```

（二）主成分分析方法

数据中包括很多属性，有些是没意义的，有些是重复的，有些组合后意义更明显。此时，我们需要简化属性节约算力，去噪，去冗余，求取更典型的属性，同时又希望不损失数据本身的意义。

在高维空间中研究样本的分布规律比较复杂，势必增加分析问题的复杂性。自然希望用较少的综合变量来代替原来较多的变量，而这几个综合变量又能尽可能多的反映原来变量的信息，并且彼此之间互不相关。

降维技术使得数据变得更易使用，并且能够去除数据中的噪音，使得其他机器学习任务更加精确，PCA(主成分分析)将多指标转化为少数几个综合指标的一

种同计分方法，为此我们采用主成分分析来降低数据维度。

其主要步骤如下：

1) 将原始数据按列组成 m 行 n 列矩阵 X
2) 将 X 的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值
3) 求出协方差矩阵
4) 求出协方差矩阵的特征值及对应的特征向量
5) 将特征向量按对应特征值大小从上到下按行排列成矩阵
6) 在需要截取 k 维时，取 P 的前 k 行即可即为降维到 k 维后的数据

主成分分析优点：降低数据复杂性，识别最重要的 N 个特征。即把高维数据在损失最小的情况下转换为地位数据，我们调用 sklearn.decomposition.PCA(n_components=None, copy=True, whiten=False)

（三）MLP 模型训练

```
model = neural_network.MLPClassifier(solver='sgd', activation='relu', learning_rate='adaptive',
                                     learning_rate_init=0.05, max_iter=10000, alpha=np.float_power(10, -4))
```

（1）MLP 神经网络的结构

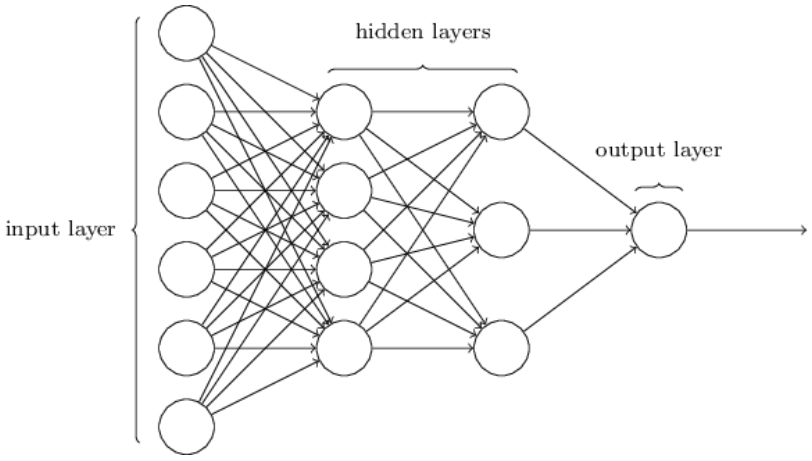

图 11 多层感知机三层模型

多层感知器（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network），基于生物神经元模型可得到多层感知器 MLP 的基本结构，最典型的 MLP 包括三层：输入层、隐层以及输出层，MLP 神经网络不同层之间是全连接的（全连接的意思就是：上一层的任何一个神经元与下一层的所有神经元都有连接），如图 11 所示

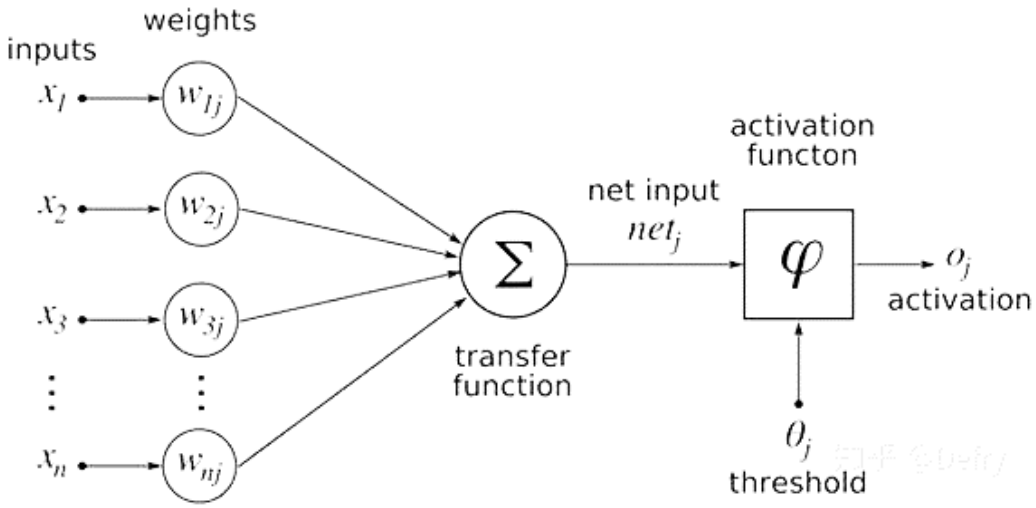

图 12 多层感知机原理图

由图 12 可知，神经网络主要有三个基本要素：权重、偏置以及激活函数

1) 权重：神经元之间的连接强度由权重表示，权重的大小表示可能性的大小
2) 偏置：偏置的设置是为了正确分类样本，是模型中一个重要的参数，即保证通过输入算出的输出值不能随便激活。
3) 激活函数：起非线性映射的作用，其可将神经元的输出幅度限制在一定范围内，一般限制在（-1~1）或（0~1）之间。

我们采用 ReLu（线性整流函数）激活函数，ReLu 是近来比较流行的激活函数，当输入信号小于 0 时，输出为 0；当输入信号大于 0 时，输出等于输入，避免了梯度爆炸和梯度消失问题，没有了其他复杂激活函数中诸如指数函数的影响；同时活跃度的分散性使得神经网络整体计算成本下降。

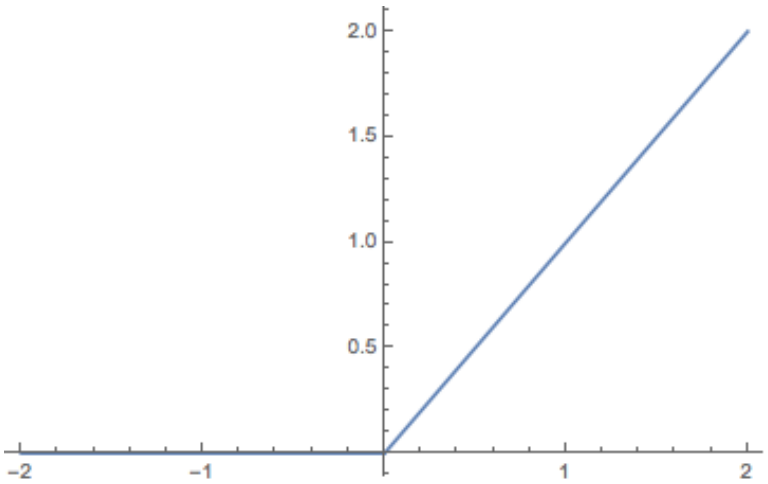

图 13 ReLu 激活函数图像

参数说明：MLP 中存在 alpha 参数

alpha	float，可选，默认为0.0001。L2惩罚（正则化项）参数。

在机器学习称作正则化；统计学领域称作惩罚项；数学界会称作范数。L2 范数:L2 就是欧式距离，向量元素绝对值的平方和再开平方$\|x\|_2 = \sqrt{\sum_{i=1}^{N} x_i^2}$

在机器学习中， L2 范数是通过使权重衰减，进而使得特征对于总体的影响减小而起到防止过拟合的作用的。L2 的优点在于求解稳定、快速。

（2）学习曲线和验证曲线

两个非常有用的诊断方法，可以用来提高算法的表现。他们就是学习曲线(learning curve)和验证曲线(validation curve)。

学习曲线可以判断学习算法是否过拟合或者欠拟合。如图 14 所示、

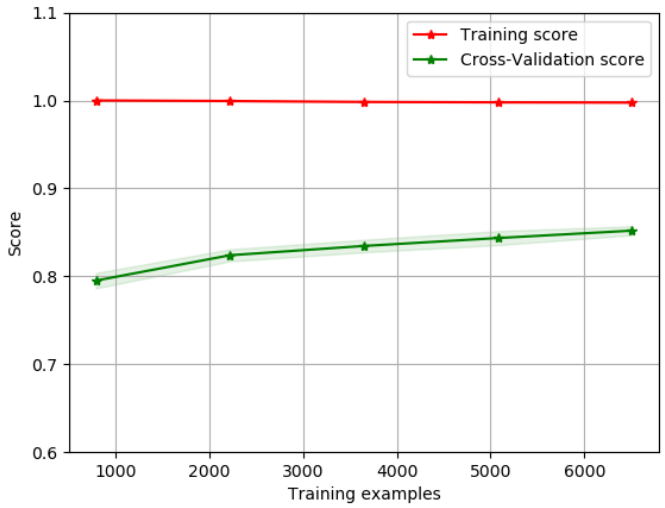

图 14 学习曲线

learning_curve 中的 train_sizes 参数控制产生学习曲线的训练样本的绝对/相对数量，我们设置的 np.linspace(0.1, 1, 5, endpoint=False)，learning_curve 默认使用分层 k 折交叉验证计算交叉验证的准确率，我们通过 cv 设置 k。

上图中可以看到，模型在测试集表现很好，不过训练集和测试集的准确率还是有一段小间隔，附件 2 还是依靠人工根据经验处理由于可能是模型有点过拟合。

验证曲线是非常有用的工具，他可以用来提高模型的性能，原因是他能处理过拟合和欠拟合问题。验证曲线和学习曲线很相近，不同的是这里画出的是不同参数下模型的准确率而不是不同训练集大小下的准确率，如图 15 所示：

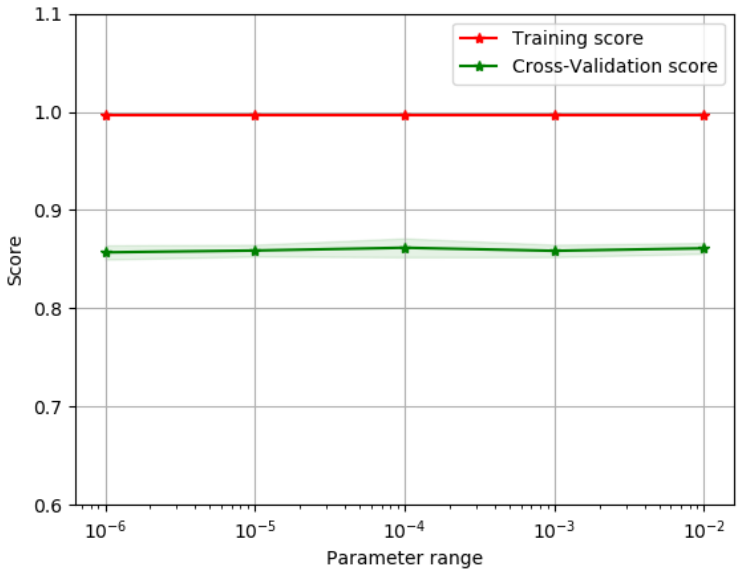

图 15 预测曲线

我们得到了参数 alpha 的验证曲线。

和 learning_curve 方法很像，validation_curve 方法使用采样 k 折交叉验证来评估模型的性能。观察上图，我们选择 alpha 值是10^{-4}。

（四）模型评估

精确率(Precision)和召回率(Recall)是常用的评价模型性能的指标，但事实上这两者在某些情况下是矛盾的。训练的机器学习模型过程中，你往往希望能够兼顾精确率和召回率，并使用一个统一的单值评价指标来评价你的机器学习模型的训练效果。

我们之所以使用调和平均而不是算术平均，是因为在算术平均中，任何一方对数值增长的贡献相当，任何一方对数值下降的责任也相当；而调和平均在增长的时候会偏袒较小值，也会惩罚精确率和召回率相差巨大的极端情况，很好地兼顾了精确率和召回率。

在机器学习领域，混淆矩阵用于衡量一个分类器的准确程度。对于二分类问题，将其样例根据真实类别和分类器的预测类别的组合划分为真正例(True Positive)、假正例(False Positive)、真反例(True Negative)假反例(False Negative)四种情形。

二分类的混淆矩阵如图 16 所示：

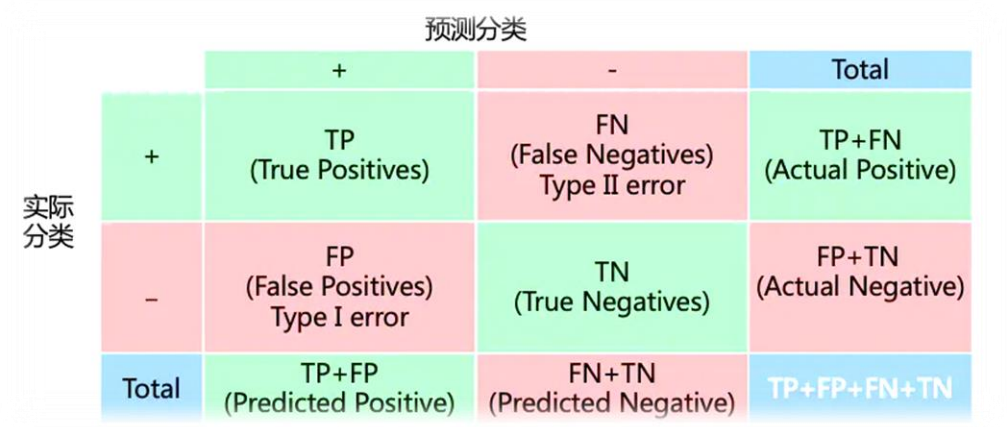

图 16 二分类混淆矩阵

1) 通过第一步的统计值计算每个类别下的 precision 和 recall

精准度/查准率 (precision)：指被分类器判定正例中的正样本的比重:

$$precision_k = \frac{TP}{TP+FP}$$

召回率/查全率(recall)： 指的是被预测为正例的占总的正例的比重

$$recall_k = \frac{TP}{TP+FN}$$

2) 通过第二步计算结果计算每个类别下的 f1-score，计算方式如下：

$$f1_k = \frac{2* \ precision_k \ * recall_k}{precision_k \ + recall_k}$$

3) 通过对第三步求得的各个类别下的 F1-score 求均值，得到最后的评测结果，计算方式如下:

$$F1 = (\frac{1}{n}\sum f1_k)^2$$

同时我们利用 k 折交叉验证，用于评估模型的预测性能，对模型性能进行无偏估计，可以在一定程度上减小过拟合。

运行结果如下：

混淆矩阵如图 17 所示：

混淆矩阵

```
[[521  18  16  15  18  24   1]
 [ 28 243   1   3   1   8   1]
 [ 12   1 150   1   4  17   1]
 [ 20   2   3 416  32   9   2]
 [ 26   2   2  16 529  11  12]
 [ 42   8   2  17   6 285   9]
 [  7   1   0   6  26  17 210]]
```

图 17 混淆矩阵

predicted actual	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生	All
城乡建设	521	18	16	15	18	24	1	613
环境保护	28	243	1	3	1	8	1	285
交通运输	12	1	150	1	4	17	1	186
教育文体	20	2	3	416	32	9	2	484
劳动和社会保障	26	2	2	16	529	11	12	598
商贸旅游	42	8	2	17	6	285	9	369
卫生计生	7	1	0	6	26	17	210	267
All	656	275	174	474	616	371	236	2802

图 18 混淆矩阵可视化

在图混淆矩阵可视化中，其每一列表示预测值，每一行表示实际值，通过图标可以直观的看出，该分类器预测能力良好。

```
模型测试集上的分数: 0.8586723768736617
模型训练集上的分数: 0.9987003898830351

K折交叉验证:
MLPClassifier: average: 0.8519628836545323

              precision    recall  f1-score   support

      城乡建设       0.81      0.85      0.83       613
      环境保护       0.89      0.84      0.86       285
      交通运输       0.85      0.88      0.86       186
      教育文体       0.90      0.88      0.89       484
    劳动和社会保障       0.90      0.88      0.89       598
      商贸旅游       0.81      0.83      0.82       369
      卫生计生       0.87      0.82      0.85       267

    accuracy                           0.86      2802
   macro avg       0.86      0.86      0.86      2802
weighted avg       0.86      0.86      0.86      2802
```

图 19 模型评估

如图 19 所示，K 折交叉验证的均值较高，表明模型的泛化能力良好，通过多次运行 F1-score 稳定在 0.86 左右，且总体高指标必须建立在同时满足高精确率和高召回率的情况之上，表明改类别划分模型评价良好。

五、热点问题模型

及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。通过文章文本的清洗、分词以及特征提取等预处理后，并借用第一问中的主成分分析法对留言主题进行降维，然后使用聚类的方式来聚集同一事件的新闻，这样便能得到新闻的热点，定义合理的热度评价指标，对热点问题进行排序。

（一）基于 DBSCAN 聚类算法的聚类分析

和传统的K-Means算法相比，DBSCAN最大的不同就是不需要输入类别数k，当然它最大的优势是可以发现任意形状的聚类簇，而不是像 K-Means，一般仅仅使用于凸的样本集聚类。同时它在聚类的同时还可以找出异常点，对数据集中的异常点不敏感。

（1）算法解释

DBSCAN 的聚类定义很简单：由密度可达关系导出的最大密度相连的样本集合，即为我们最终聚类的一个类别，或者说一个簇。这个 DBSCAN 的簇里面可以有一个或者多个核心对象。如果只有一个核心对象，则簇里其他的非核心对象样本都在这个核心对象的 ϵ-邻域里；如果有多个核心对象，则簇里的任意一个核心对象的 ϵ-邻域中一定有一个其他的核心对象，否则这两个核心对象无法密度可达。这些核心对象的 ϵ-邻域里所有的样本的集合组成的一个 DBSCAN 聚类簇。

（2）基本概念

- Eps 邻域：给定对象半径 Eps 内的邻域称为该对象的 Eps 邻域;
- 核心点（core point）：如果对象的 Eps 邻域至少包含最小数目 MinPts 的对象，则称该对象为核心对象;
- 边界点（edge point）：边界点不是核心点，但落在某个核心点的邻域内;
- 噪音点（outlier point）：既不是核心点，也不是边界点的任何点;
- 直接密度可达(directly density-reachable)：给定一个对象集合 D，如果 p 在 q 的 Eps 邻域内，而 q 是一个核心对象，则称对象 p 从对象 q 出发时是直接密度可达的;
- 密度可达(density-reachable)：如果存在一个对象链　p1, …,pi,.., pn，满足 p1 = p 和 pn = q，pi 是从 pi+1 关于 Eps 和 MinPts 直接密度可达的，则对象 p 是从对象 q 关于 Eps 和 MinPts 密度可达的;
- 密度相连(density-connected)：如果存在对象 O∈D，使对象 p 和 q 都是从 O 关于 Eps 和 MinPts 密度可达的，那么对象 p 到 q 是关于 Eps 和 MinPts 密度相连的

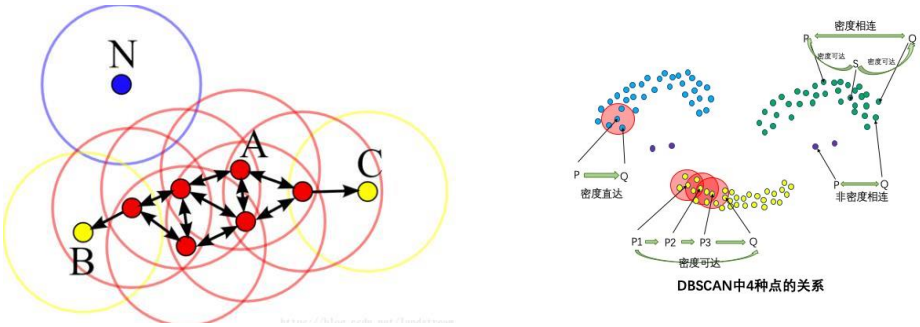

图 20 核心点、边界点、噪音点、四种点的关系

图中红色为核心点，黄色为边界点，蓝色为噪音点，minPts = 4，Eps 是图中圆的半径大小有关“直接密度可达”和“密度可达”

（3）算法步骤

1) 解析样本数据文件
2) 计算每个点与其他所有点之间的欧几里德距离
3) 计算每个点的 k-距离值，并对所有点的 k-距离集合进行升序排序，输出的排序后的 k-距离值
4) 将所有点的 k-距离值，在 Excel 中用散点图显示 k-距离变化趋势
5) 根据散点图确定半径 Eps 的值
6) 根据给定 MinPts=4，以及半径 Eps 的值，计算所有核心点，并建立核心点与到核心点距离小于半径 Eps 的点的映射
7) 根据得到的核心点集合，以及半径 Eps 的值，计算能够连通的核心点，并得到离群点
8) 将能够连通的每一组核心点，以及到核心点距离小于半径 Eps 的点，都放到一起，形成一个簇
9) 选择不同的半径 Eps，使用 DBSCAN 算法聚类得到的一组簇及其离群点，使用散点图对比聚类效果

（4）算法实现

我们调用 sklearn.cluster.DBSCAN (eps=0.15, min_samples=1)来实现。分类部分结果如图 21 所示 ：

问题ID	index_	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	0	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定向商...	0	0
	1	196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019-08-07 19:52:14	A市伊景园滨河苑现强制要求购房者捆绑购买12万车位一个不买就取消购房资格国家三令五申禁止捆绑...	0	0
	2	199190	A00095080	关于A市武广新城违法捆绑销售车位的投诉	2019-08-01 22:32:26	武广新城为铁广集团的定向商品房在未取得预售资格强行逼迫职工缴纳185万购房款且不签正式购房合...	0	0
	3	205277	A909234	伊景园滨河苑捆绑车位销售合法吗	2019-08-14 09:28:31	广铁集团强制要求职工购买伊景园滨河苑楼盘时捆绑购买12万一个的车位不买车位不能购买房子这种做...	1	0
	4	205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019-08-03 10:03:10	我坚决反对伊景园滨河苑捆绑销售车位原本广铁集团与市政府和开发商协议可以给铁路职工优惠定向购买...	2	0
...	...	...	...	...	...	...	...	...
3848	0	226245	A00049865	A市A1区万国城新建养老院是否存在违规	2019-01-01 21:10:17	本人A1区万国城业主对于新建养老院有几点需要投诉的问题1关于建之前所谓的民意调查是以什么方式...	0	0
3849	0	238266	A00063188	冰雪天气A7县校车停运合理吗	2019-01-01 12:40:33	刚接通了明天校车停运学生家长自行接送请问教育局校车怕出事承担责任家长呢	0	0
3850	0	283135	A00057709	A市A1区万国城moma未经业主同意强建养老院	2019-01-01 10:26:56	A1区万国城moma三期业主于2018年12月26日晚间发现位于该小区7栋楼栋内一至二层有一...	0	0
3851	0	285107	A00024520	A市电建星湖湾强制要求业主收房	2019-01-01 02:20:10	关于星湖湾洋房二期首批质量及小区品质问题及其他诉求书我们是星湖湾洋房二期购房者我们怀着对生活...	0	0
3852	0	289408	A0012413	在A市人才app上申请购房补贴为什么通不过	2018-11-15 16:07:12	我叫朱琦梦是2017年12月落户并于2018年初在A市A6区首次购房的硕士毕业生符合A市人才...	0	0

图 21 聚类部分结果

（二）热点问题排行

上一节实现了对文本的聚类，但是要想直观的观察新闻聚类的效果，还需要进行统计分析，为了清楚的了解每处热点的话题信息，可以把留言主题作为留言的话题，再进行排行，得出话题的排名。

（1）基于 TextRank 的自动文摘算法

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。类似于网页的排名，对于词语可得到词语的排行，对于句子也可得到句子的排名，所以 TextRank 可以进行关键词提取，也可以进行自动文摘。其用于自动文摘时的思想是：将每个句子看成 PageRank 图中的一个节点，若两个句子之间的相似度大于设定的阈值，则认为这两个句子之间有相似联系，对应的这两个节点之间便有一条无向有权边，边的权值是相似度，接着利用 PageRank 算法即可得到句子的得分，把得分较高的句子作为文章的摘要。 TextRank 算法的主要步骤如下：

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：在原论文中采用如下公式进行计算句子 1 和句子 2 的相似度：

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

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：

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

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。

（2）热度评价

根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标

1）拉依达准则去除离群值

通过 excel 分别绘制附件 3 中点赞数、反对数的散点图，如表 22 所示：

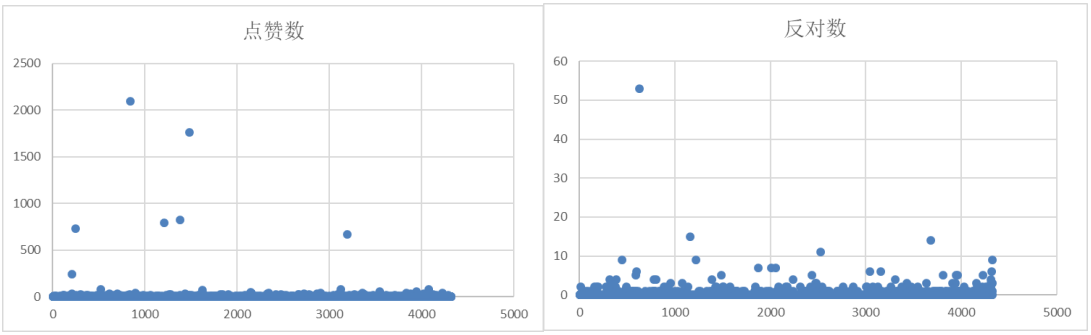

图 22 留言点赞、留言反对图

由上图可知，对于留言的点赞数、反对数个别数据与平均值差别较大，我们把此数据视为可疑值，也称离群值。如果在统计学中认为应该舍弃的数据留用了，势必会影响其均值的可靠性。相反，本应该留用的数据被遗弃，虽然精度提高，却夸大了均值的可靠性。由此可见，用恰当方法定量确定离群值的取舍在分析中有重要的意义。

点赞、反对数值的分布情况，如图 23 所示：

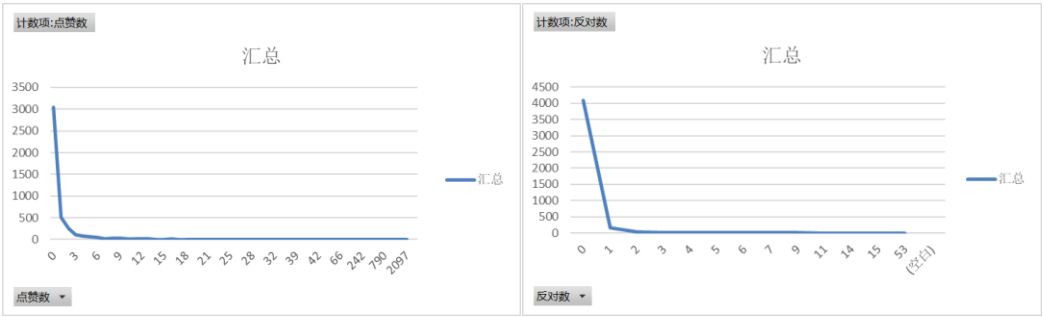

图 23 点赞数分布、反对数分布情况图

```
def revise_outlier(data, value):
    """
    以均值代替离群值
    Parameters
    ----------
    data: Series, 计算 sigma 的数据
    value: float, 需要检验的数值

    Returns
    -------
    修正后的数值
    """
    mean = data.mean()
    std = data.std()
    bool_ = (mean - 3 * std) < value < (mean + 3 * std)
    return value if bool_ else mean
```

拉依达准则又称 3σ 准则：数值分布在（μ-3σ，μ+3σ）中的概率为 0.9974；由此可见 X 落在（μ-3σ，μ+3σ）以外的概率小于千分之三，在实际问题中常认为相应的事件是不会发生的，基本上可以把区间（μ-3σ,μ+3σ）看作是随机变量 X 实际可能的取值区间，这称之为正态分布的“3σ”原则。

数据不服从正态分布，也可以用远离平均值的多少倍标准差来描述（这就使该原理可以适用于不同的业务场景，只是需要根据经验来确定 k sigma 中的 k 值，这个 k 值就可以认为是阈值），我们以均值代替离群值。

2）定义热度评价指标

考虑到影响热度的指标：点赞数、反对数、时间等因素，我们在魔方秀热度算法的基础上，加以改进。

魔方秀热度算法：

$$\left(\frac{\text{总赞数} * 0.7 + \text{总评论数} * 0.3}{\text{发布时间距离当前时间的小时差} + 2}\right)^{1.2}$$

由于点赞数和反对数在一定程度上都可以来表示留言主题的热度，但其所占的影响力是不同的，基于此，我们结合魔方秀热度算法，我们对点赞数和反对数按照 7：3 的进行加权，综合评定总赞数的影响力。在任一时刻，留言内容都有一个"当前温度"随着时间流逝，所有留言的温度都逐渐"冷却"，考虑到时间对热度的影响。我们将总赞数和时间热度综合评定留言热度。如时间热度图如 24 图所示：

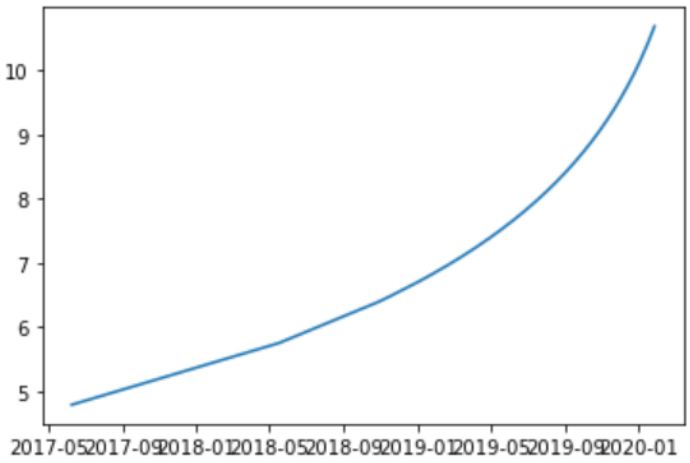

图 24 时间热度图

代码如图所示：

```
def compute_score(data):
    """
    计算热度
    Parameters
    ----------
    data: DataFrame, 需要进行计算的数据

    Returns
    -------
    data: DataFrame, 计算后的数据
    """
    data = data.assign(
        interact=lambda df: np.multiply([df['down'], df['up']], np.reshape([0.3, 0.7], (2, -1))).sum(axis=0))
    temp = data['interact']
    data['interact'] = data['interact'].apply(lambda r: revise_outlier(temp, r))
    data = data.assign(
        score=lambda df: df['interact'].apply(lambda r: np.log(r + 1.5) / np.log(1.5)) +
                         df['time'].apply(lambda r: np.log(1 / max(get_sec(r), 30 * 24 * 3600)) / np.log(1.5) + 50))
    data.drop(columns=['down', 'up', 'interact'], inplace=True)
    return data
```

2）热点问题排行实验结果

排名 Top5 热点问题的排行结果如图 25、图 26 所示：

		热度指数	时间范围	地点/人群	问题描述
热度排名	问题ID				
1	1	73.360745	2019-07-07 07:28:06 至 2019-09-01 14:20:22	伊景园 滨河苑 A市 新城 广铁集团 滨河 业主	伊景园滨河苑捆绑车位销售合法吗
2	2	33.121027	2019-01-16 11:58:48 至 2019-12-02 11:57:49	A市	反映A市人才租房购房补贴问题
3	3	31.758287	2019-11-13 11:20:21 至 2020-01-25 09:07:21	A2区 丽发新城 A市 搅拌站 小区	投诉小区附近搅拌站噪音扰民
4	4	26.954396	2019-06-10 10:24:59 至 2019-09-09 08:20:47	A7县 凉塘路 旧城 星沙 四区 星沙街道	A7县星沙四区凉塘路的旧城改造要拖到何时
5	5	25.693314	2019-01-18 09:55:02 至 2019-12-30 15:27:00	A7县 A市 A8县	A市365公交车经常等很久不来

图 25 部分热点问题表

A	B	C	D	E	F	G	H	I
问题ID	index_	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	0	190337	A0009051	关于伊景	2019-08-23	投诉伊景园	0	0
	1	196264	A0009508	投诉A市伊	2019-08-07	A市伊景园滨	0	0
	2	199190	A0009508	关于A市武	2019-08-01	武广新城为	0	0
	3	205277	A909234	伊景园滨	2019-08-14	广铁集团强	1	0
	4	205982	A909168	坚决反对	2019-08-03	我坚决反对	2	0
	5	207243	A909175	伊景园滨	2019-08-23	您好A市武广	0	0
	6	209506	A909179	A市武广新	2019-08-02	您好由A市广	0	0
	7	209571	A909200	伊景园滨	2019-08-28	广铁集团铁	0	0
	8	213584	A909172	投诉A市伊	2019-07-28	投诉A市伊景	0	0
	9	218709	A0001066	A市伊景园	2019-08-01	伊景园滨河	1	0
	10	222209	A0001717	A市伊景园	2019-08-28	广铁集团与	0	0
	11	223247	A0004475	投诉A市伊	2019-07-23	关于铁广集	0	0
	12	224767	A909176	伊景园滨	2019-07-30	伊景园滨河	0	0
	13	230554	A909174	投诉A市伊	2019-08-19	投诉A市伊景	0	0
	14	234633	A909194	无视消费	2019-08-20	伊景园滨河	0	0
	15	236301	A909197	和谐社会	2019-08-30	广铁集团与A	0	0
	16	239032	A909169	请维护铁	2019-09-01	广铁集团和A	1	0
	17	244243	A909198	关于伊景	2019-08-24	广铁集团铁	0	0
	18	244342	A0008948	投诉A市伊	2019-07-28	A市伊景园滨	0	0
	19	246407	A0009959	举报广铁	2019-09-01	我要举报广	0	0
	20	251844	A909167	投诉伊景	2019-08-20	投诉广铁集	1	0
	21	255507	A909195	违反自由	2019-08-20	广铁集团铁	0	0
	22	258037	A909190	投诉伊景	2019-08-23	尊敬的领导	0	0
	23	260254	A909173	投诉A市伊	2019-08-30	投诉A市伊景	0	0
	24	268299	A909193	惊A市伊景	2019-08-21	伊景园滨河	0	0
	25	276460	A909170	A市伊景园	2019-08-24	尊敬的领导	0	0
	26	279070	A0009508	投诉A市伊	2019-08-31	投诉A市伊景	0	0
	27	283879	A0004475	A市伊景园	2019-07-18	关于铁广集	0	0
	28	285897	A909191	武广新城	2019-08-01	我们是广铁	0	0
	29	286304	A909196	无视职工	2019-08-23	广铁集团与A	0	0
	30	289473	A0001034	反对滨河	2019-08-22	现有伊景园	0	0
	31	289950	A0004475	投诉A市伊	2019-07-07	提问A市政府	0	0

图 26 部分热点问题明细表

由图可见，每条留言都通过类别大小进行排序，在列表中，通过增加一列“热点指数”来代表留言的热度。

六、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。

（一）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。

（二）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

七、参考文献

[1] 王馨.网络新闻热点发现研究[D].河北大学,2015.

[2] 彭卫华.互联网新闻热点挖掘系统的研究与实现[D].哈尔滨工业大学,2010.

[3] 叶建成.利用文本挖掘技术进行新闻热点关注问题分析[D].广州大学,2018.

[4] 蒲梅,周枫,周晶晶,严馨,周兰江.基于加权 TextRank 的新闻关键事件主题句提取[J].计算机工程,2017,43(08):219-224.

[5] 周练.Word2vec 的工作原理及应用探究[J].科技情报开发与经济,2015,25(02):145-148.

[6] 刘妍东.大数据中数据的质量问题探析[J].现代商贸工业,2020,(4):193. DOI:10.19311/j.cnki.1672-3198.2020.04.092.

[7] 蔡元萃,陈立潮.聚类算法研究综述[J].科技情报开与经济,2007(01):145-146.

[8] 素质.基于腾讯 AILab 词向量进行未知词、短语向量补齐与域内相似词搜索[EB/OL].https://cloud.tencent.com/developer/user/1051732/articles,2019.

[9] Jiawei Han,Micheline Kamber. Data Mining Concepts and Techniques[M].范明，孟晓峰译，机械工业出版，2001.