“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：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。

针对问题二：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。

针对问题三：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。并 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：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.

For problem 1: first, the data was preprocessed and the text was vectored. The principal component analysis method was used to reduce the dimension of the data. In the model training, the neural network MLPClassifier method was used to complete the supervised learning.

For problem 2: firstly, DBSCAN clustering algorithm was used to cluster the problem categories. Thumb up number and anti-number scatter graphs were drawn by extracting the document abstract, and the outliers in annex 3 were processed by raedah criterion. Based on the factors affecting heat, appropriate heat evaluation indexes were defined and hot issues were obtained.

Aiming at question 3: an automatic evaluation index named WEEM4TS is proposed to evaluate the system performance of the relevance between the response comments and the original text.The purpose of WEEM4TS is to evaluate the quality of a concise summary based on the retention meaning of the original document.Therefore, it is considered to represent the evaluation metrics applicable to all types of system summaries: extraction, abstraction, and mixing.A method called WETS is proposed to determine the most important sentences in the original document to assess the completeness and interpretability of the response comments.

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

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：

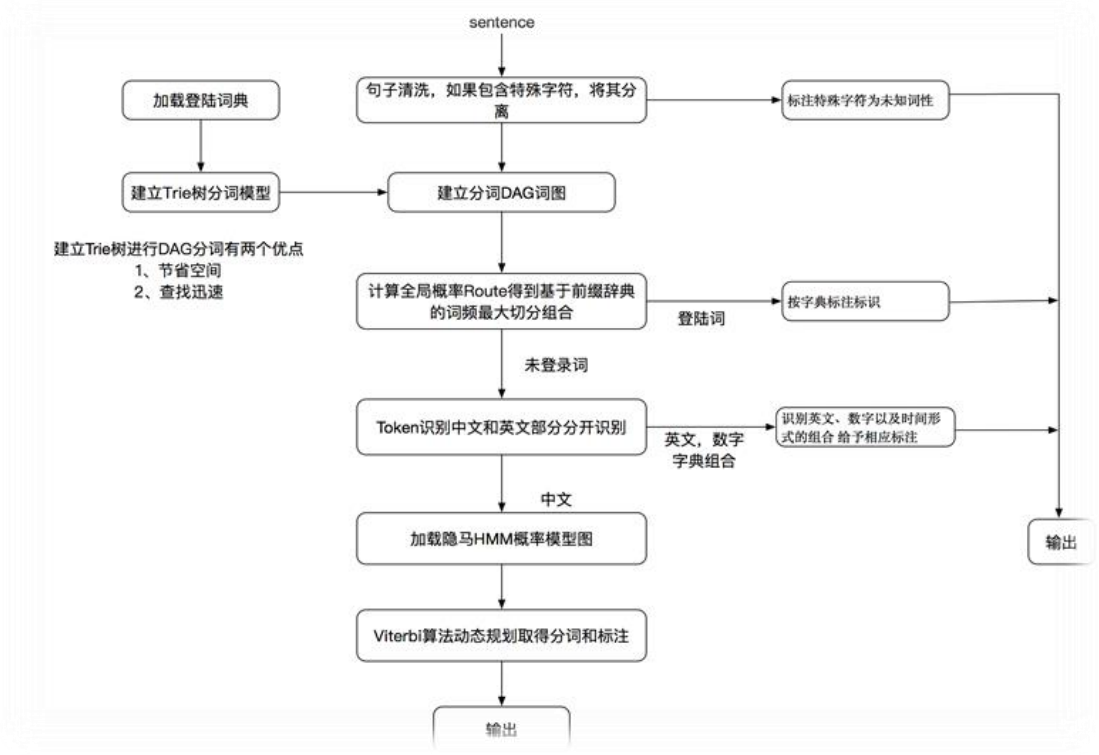

图 2 结巴分词原理图

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。

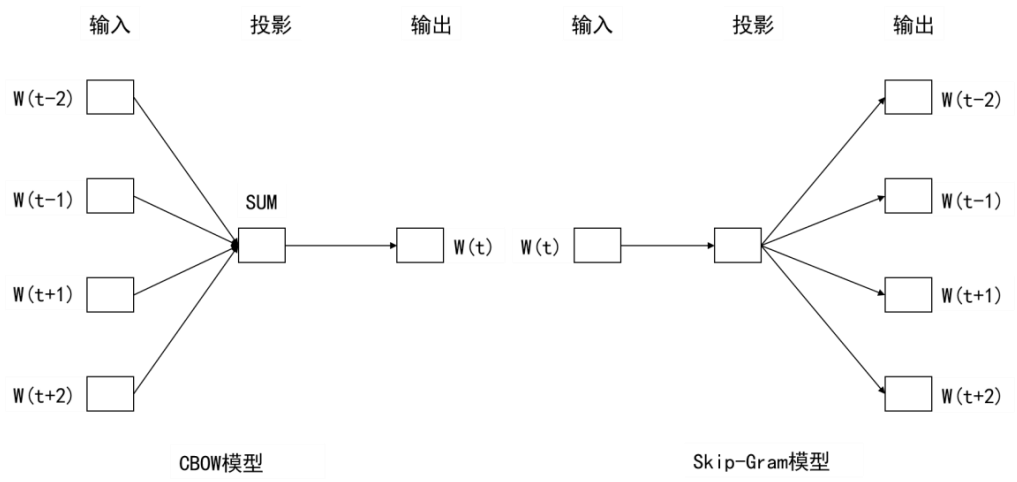

图 7 Word2vec 两种训练模型

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

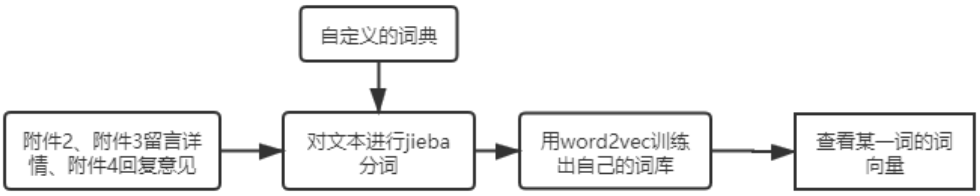

图 8word2vec 训练词向量

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。

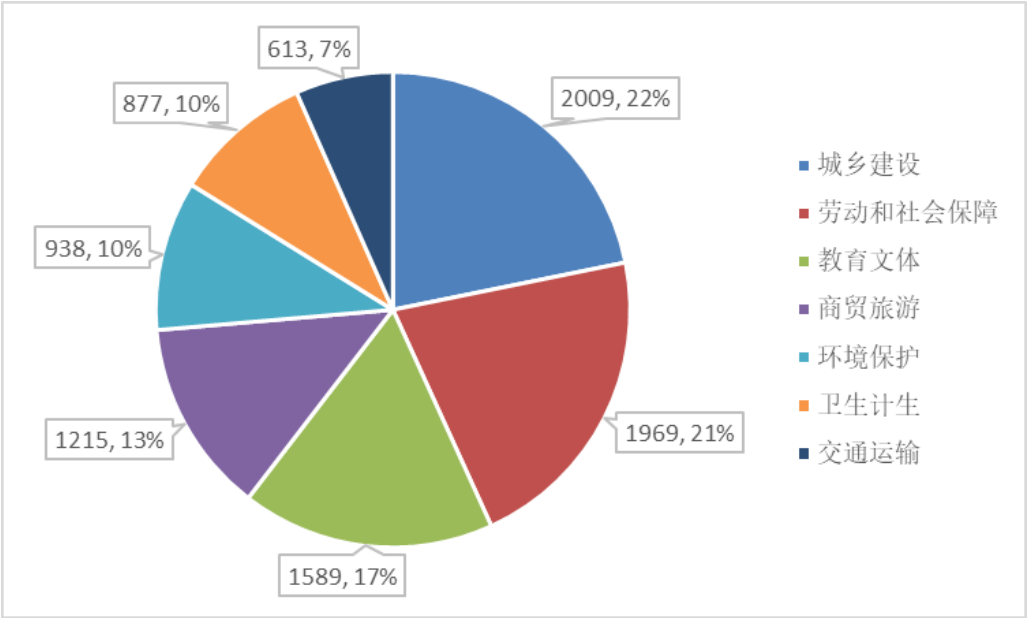

图 9 一级标签类别占比

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。

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

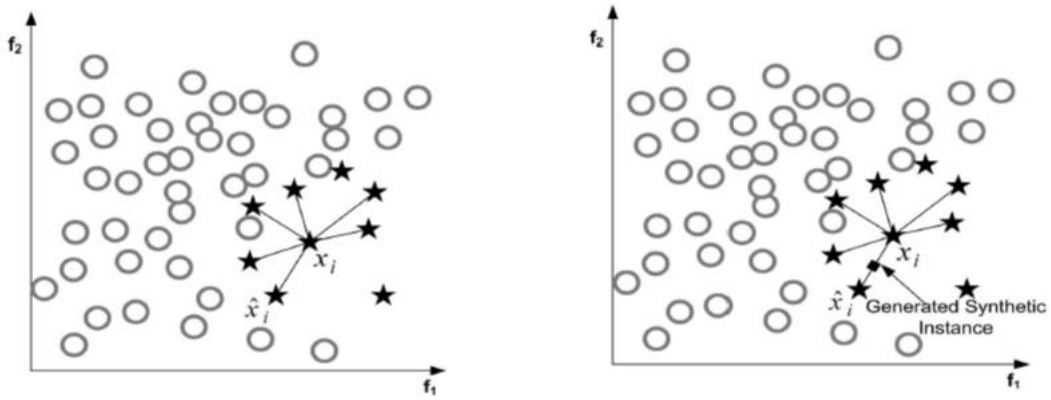

图 10 Smote 分析方法

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

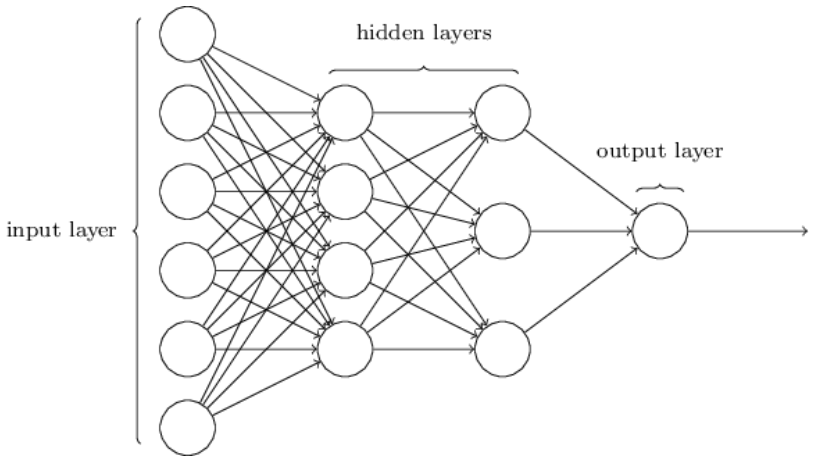

图 11 多层感知机三层模型

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

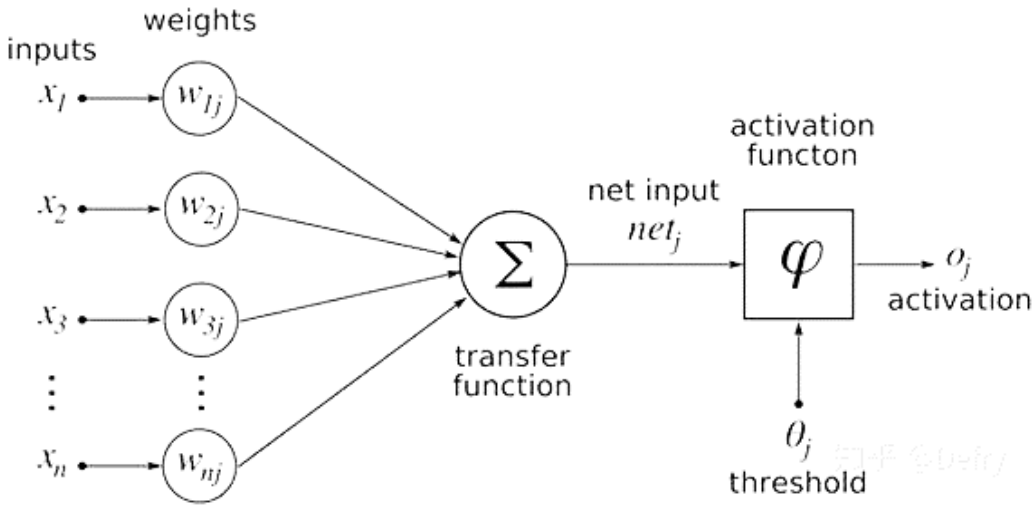

图 12 多层感知机原理图

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。

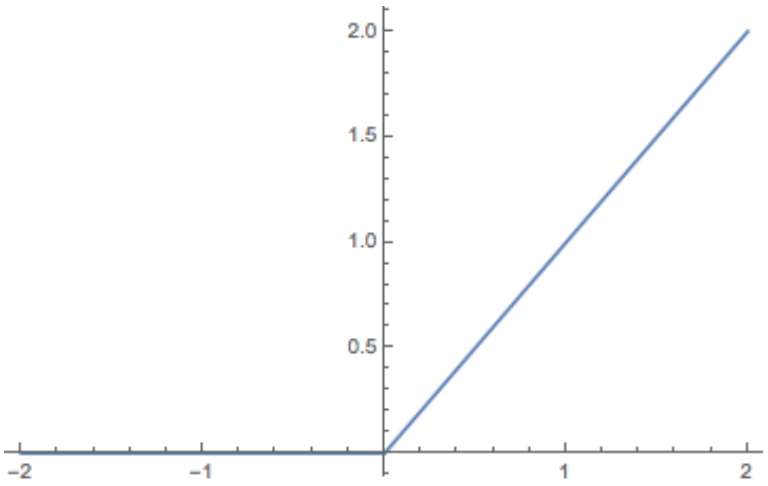

图 13 ReLu 激活函数图像

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、

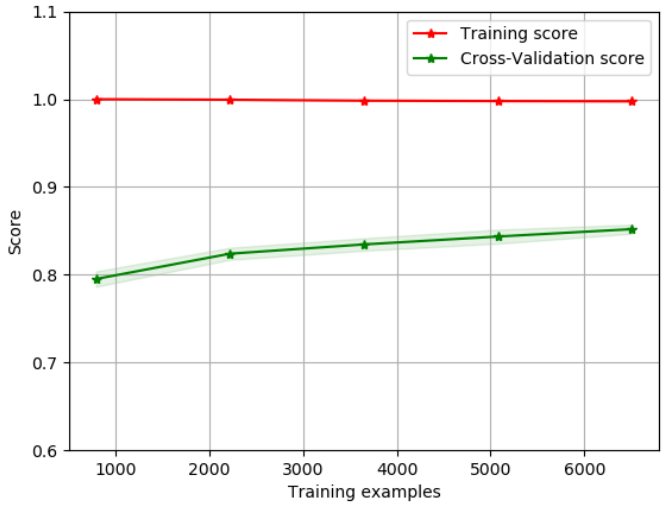

图 14 学习曲线

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：

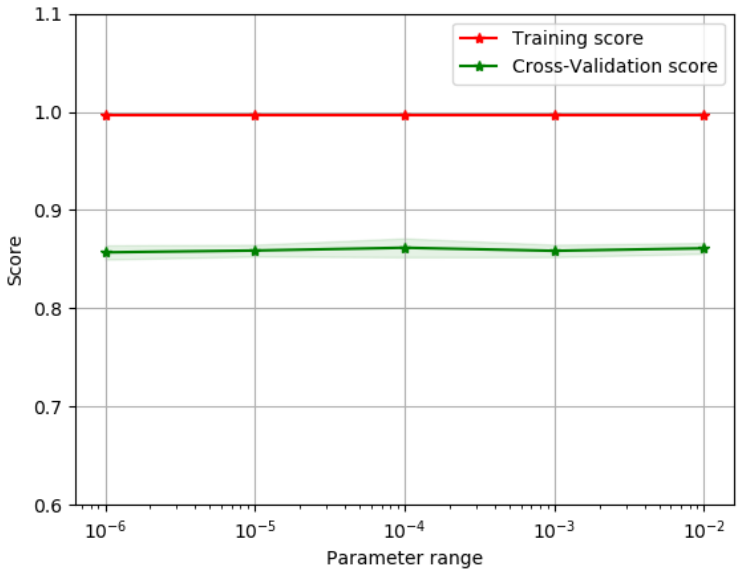

图 15 预测曲线

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：

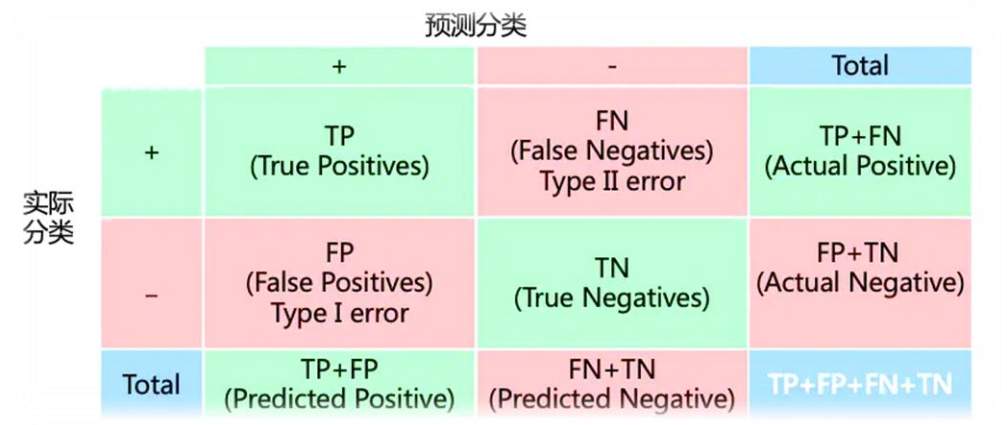

图 16 二分类混淆矩阵

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

（一）热点问题挖掘

（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。

（2）流程图

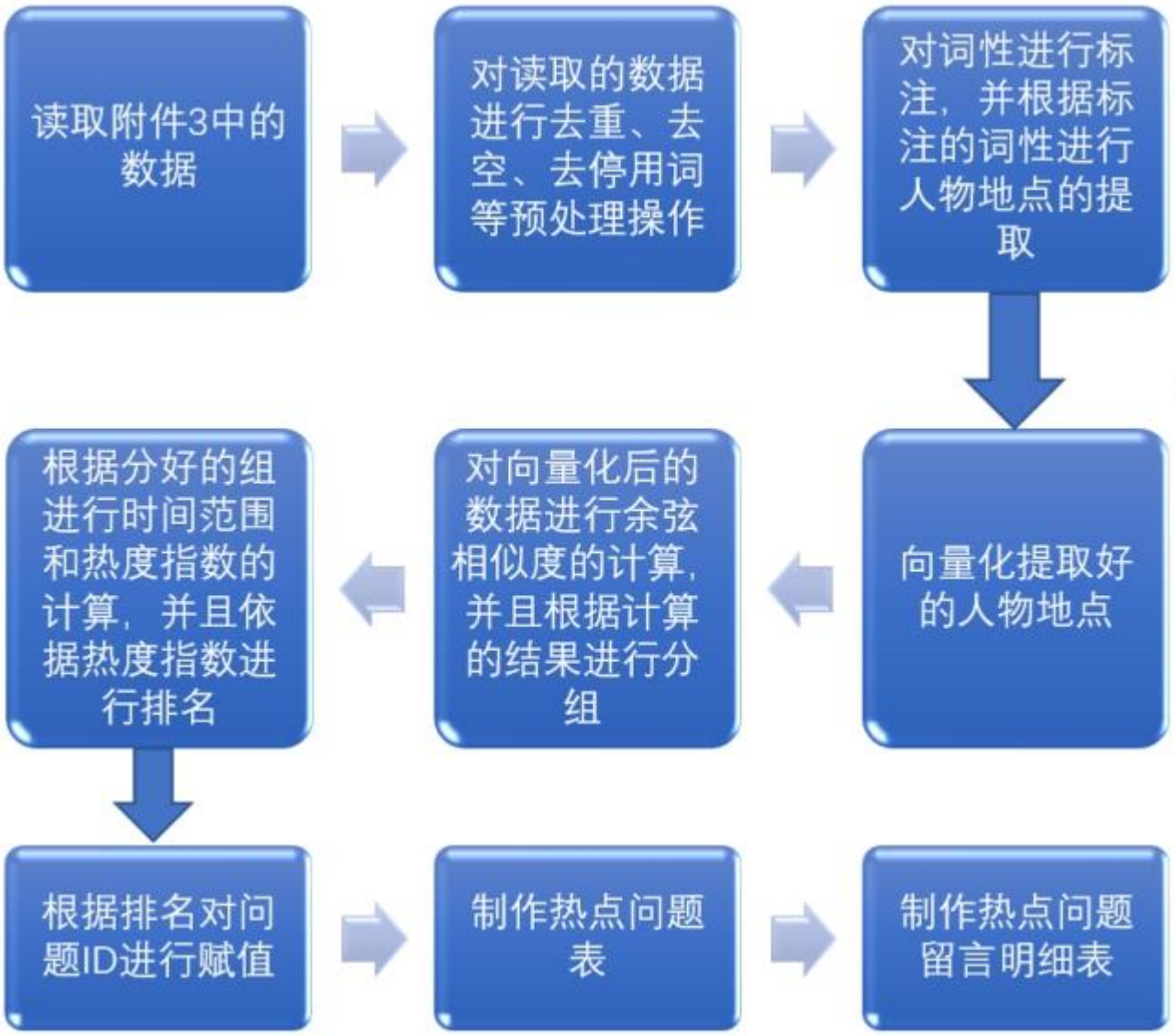

图 20 总体流程图

（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 21 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 22 去停用词结果图

（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 23 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 24 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 25 提取人物地点效果图

（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：

1）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 26 向量化后矩阵图

2）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。

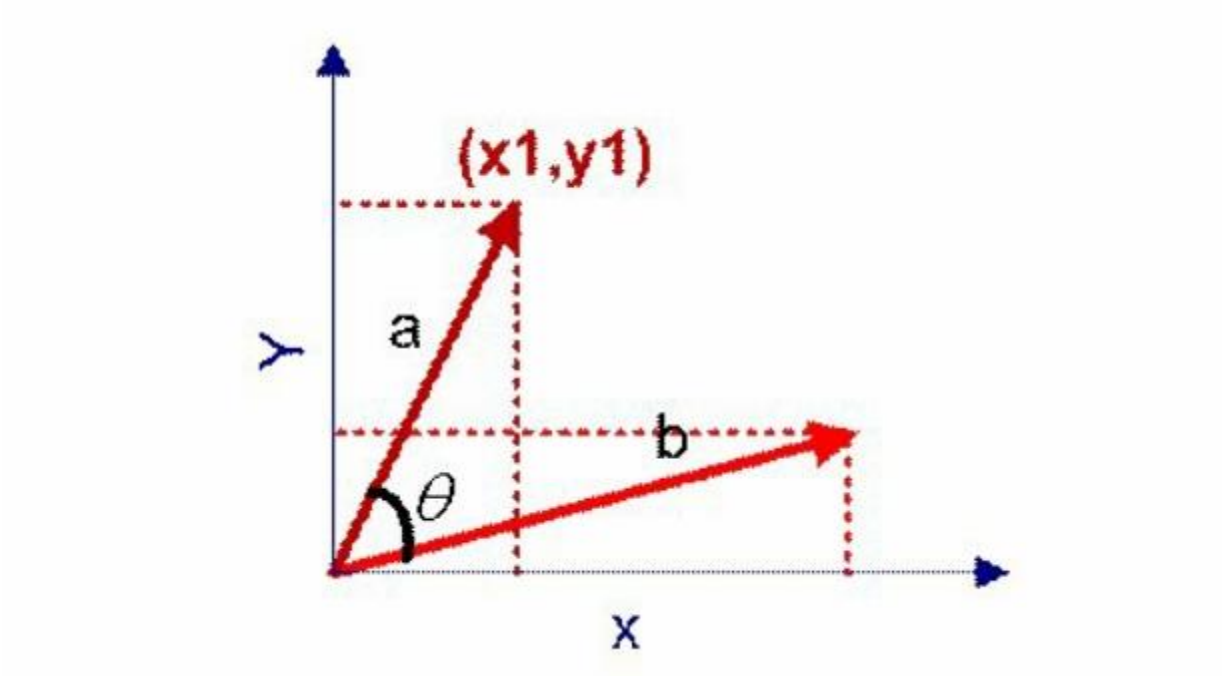

图 27 余弦相似度计算

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。

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

3）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 28 留言编号区分图

（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2019-7-8
2019-8-19
2019-4-11
2019-8-23至2019-9-6
2019-1-3至2020-1-7
2019-6-19至2019-11-12

图 29 格式化后时间表

（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 30 热度排名与 id 赋值效果图

（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 31 热点问题表

（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 32 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。

六、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.