基于快速文本的智慧政务文本挖掘应用

摘 要

在日常生活中，民众的一些意见需要反馈到政府平台，但是以往到各个部门提交匿名信已经成为过去式，随着大数据、云计算、人工智能等技术的发展，自然语言处理作为人工智能的一个重要领域得到了飞速发展。更多便捷的等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的平台。但是这给相关部门的工作人员带来挑战，如何有效地将留言归类，实现信息利用和数据共享，汇聚整理热点问题，如何更加高效的回复评价。本文采取构建基于自然语言处理技术的智慧政务系统，来解决新发展新趋势带来的问题。

我们首先对材料还有训练集进行数据的预处理，清晰的了解训练集中数据的特征。在这里使用了 jieba 分词以及去停用词，然后我们提取词向量利用 word2vec 负采样文本抽取技术进行取样。与原始数据进行合并处理，之后输入到 fasttext 模型进行快速训练。随后使用 F-Score 对分类方法进行评价，计算得到准确率为 99.9%以上，F1 score 为 0.99。

第二小问通过选取处理好的数据进行实验，使用 TF-IDF 模型和 K-means 聚类算法来解决热点问题，使用 TF-IDF 计算关键词权重矩阵，并将稀疏权重矩阵通过主成分分析（PCA）算法进行降维，维度需要通过多次计算，选取最优值。之后再通过 Kmeans 算法进行聚类，将分散的数据进行无监督学习聚类形成带有类别标签的数据。

第三小问利用层次分析法根据答复的相关性、完整性、可解释性等角度出发，来对相关部门对留言的答复意见给出一套评价方案。为了便于对留言答复的比较与判别，每层的元素个数在 3~7 之间为佳，超过 7 以后增加了比较判断的难度。建立评价模型后，根据经验对每层里的各个元素建立重要判别矩阵，从判别矩阵中可以得到某一层中各个指标的归一化权重。由层与层之间权重的传递可以得到最低层的综合权重，最终选取置最优的回复评价方案作为答案。

最后我们将模型在线下对测试集进行验证，得到了比较好的结果，也说明了模型的有效性，实用性。
关键词: TF-IDF，自然语言处理，负采样技术， FastText 模型，K-means 聚类算法，层次分析法

Abstract

In daily life, some opinions of the public need to be fed back to the government platform, but in the past, submitting anonymous letters to various departments has become the past style. With the development of big data, cloud computing, artificial intelligence and other technologies, natural language processing as an important area of artificial intelligence has been rapidly developed. More and more convenient and other online platforms have gradually become a platform for the government to understand public opinion, gather people's wisdom and gather people's spirit. But it brings challenges to the staff of relevant departments. How to effectively classify the messages, realize information utilization and data sharing, gather and sort out hot issues, and how to reply and evaluate more efficiently. In order to solve the problems brought by the new development trend, this paper constructs a smart government system based on natural language processing technology.

First of all, we preprocess the data of materials and training sets to clearly understand the characteristics of data in training sets. In this paper, we use the Jieba word segmentation and de stop words, then we extract the word vector and use the word2vec negative sampling text extraction technology to sample. After merging with the original data, it is input to the fasttext model for fast training. Then, F-score was used to evaluate the classification method, and the accuracy was over 99.9%, and F1 score was 0.99.

The second question is to select the processed data for experiments, use TF-IDF

model and K-means clustering algorithm to solve the hot issues, use TF-IDF to calculate the keyword weight matrix, and use PCA algorithm to reduce the dimension of sparse weight matrix. The dimension needs to be calculated many times to select the optimal value. After that, the kmeans algorithm is used to cluster the scattered data to form the data with category label.

The third question uses the analytic hierarchy process (AHP) to give a set of evaluation scheme for the reply of relevant departments to the message according to the relevance, integrity, and interpretability of the reply. In order to facilitate the comparison and discrimination of message replies, the number of elements in each layer is better between 3 and 7, which increases the difficulty of comparison and judgment after exceeding 7. After the evaluation model is established, an important discrimination matrix is established for each element in each layer according to experience, and the normalized weight of each index in a certain layer can be obtained from the discrimination matrix. The comprehensive weight of the lowest layer can be obtained from the weight transfer between layers. Finally, the optimal reply evaluation scheme is selected as the answer.

Finally, we verify the model under the online test set, and get better results, which also shows the validity and practicability of the model.

Keywords: TF IDF, natural language processing, negative sampling technology, fasttext model, K-means clustering algorithm, AHP

目录

一、引言

1.1 介绍

随着科技的发展和网络平台的应用，智慧政务的建设革新了各个部门、窗口、办公的传统模式。它以更加方便快捷的优势，赢得大众的接受认同。民众不再需要到各个部门提交匿名信，而是以更加便捷的方式汇集到政府部门手中。在大数据的环境下，留言划分和热点整理已经成为相关部门的主要工作。如何有效地处理网络问政平台的群众留言，真正实现信息利用和数据共享成为新的挑战。比如，有关教育的留言被发送到其他部门，那么要将其留言重新划分到教育局的平台。民众在近一个月反映的留言中，整理出相关统一问题，分析汇总整理出近一个月的热点问题。但是，在各类社情民意相关的文本数据量的不断攀升中，要人工准确无误的划分文本留言到不同部门无疑是一个难题。

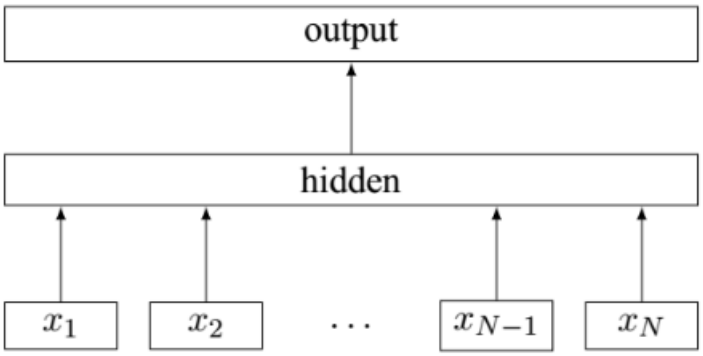
网络问政留言中智能分类的应用能够有效解决上述难题。随着大数据、云计算、人工智能等技术的发展，构建基于自然语言处理技术的智能的文本挖掘模型。模型可以起到辅助分类的作用，可以解决靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。在具体使用情景里面，对于用户输入的留言，模型可以直接将留言划分至不同职能的部门。

综上所述，对题目的问题集进行数据分析以及预处理后，我们构建了 fastText 精准快速文本分类模型，使用 TF-IDF 模型和 K-means 聚类算法进行无监督学习，利用层次分析法来建立重要判别矩阵得到归一化权重。该模型与其他主流方案相比，在准确率、查全率、查准率、F1-Score 以及泛化能力上都表现出优越的效果。本文包括引言、fastText 精准快速文本分类模型、预处理、实验、总结与展望五个部分。

二、fastText 精准快速文本分类模型

2.1 FastText 模型框架

FastText 的模型架构是由一层 word embedding 的隐藏层和输出层组成。X_1到X_N表示表示一个文本中的 n-gram 向量，每个特征是词向量的平均值。FastText 仅含有一层的隐藏层,它能够让训练速度变快，处理速度得到了创新，又能够符合网络的复杂性。FastText 在预测标签时使用了非线性激活函数，但在中间层不使用非线性激活函数，在一定程度上增加模型的复杂性。FastText 是用全部的 n-gram 去预测指定类别。



fastText 模型

FastText 是 Facebook AI Research 在 16 年开源的一个文本分类器，也算是 word2vec 所衍生出来的。FastText 模型的词向量模型沿用了 CBOW 的单层神经网络的模式，它的思路和架构组成和 CBOW 模型基本相同,但是 FastText 的处理速度得到了创新。CBOW 模型是将单词的高维稀疏的 one-hot 向量映射为低维稠密的表示的方法，FastText 模型使用 n-gram 特征代替单个词的特征，提取序列信息，所得到的效果与深度学习分类器无差别，语义信息也变得更加完整。在模型 CBOW 输出的是目标单词及其概率值，是根据上下文词条预测得出的。而 Fasttext 模型的输出是短文本的类别标签及其概率值。即 FastTest 预测标签，而 CBOW 预测的是中间词。FastTest 相对于 word2vec 来说，增加了 subwords 特性，并通过隐藏表征在类别间共享信息。相对于其他分类模型，它在保持分类效果的同时，大大缩短了训练时间。

对于 FastTest 算法, Softmax 回归可以处理多类别问题，是将可以判定为某

类的特征相加，然后将这些特征转化为判定是这一类的概率。Softmax 回归又被称作多项逻辑回归,它是逻辑回归在处理多类别任务上的推广。所以隐藏层到输出层可以看作一个 Softmax 回归。Softmax 的函数为：

$$h_\theta(x^i)=\begin{bmatrix} p(W^i=1|x^i;\theta) \\ p(W^i=2|x^i;\theta) \\ \vdots \\ p(W^i=k|x^i;\theta) \end{bmatrix}$$

在 Softmax 回归中，对于样本集，多分类模型的输出结果为该样本属于 k 个类别的概率，从这 k 个概率中我们选择最优的概率对应的类别来作为该样本的预测类别。因此，Softmax 有有 k 个类别，除以它们的累加和，这样做就实现了归一化，使得输出的 k 个数的和为 1，而每一个数就代表那个类别出现的概率。Softmax 的代价函数为:

$$J(\theta)=-\frac{1}{v}\left[\sum_{i=1}^{v}\sum_{j=1}^{k}1\{w^i=j\}\log\frac{e^{\theta_j^T x^i}}{\sum_{l=1}^{k}e^{\theta_j^T x^i}}\right]$$

矩阵中的每行为一个类别对应的分类器参数。对于有大量类别的数据集，FastTest 使用了一个分层分类器。在某些文本分类任务中类别很多，计算线性分类器的复杂度高。为了改善运行时间，数据集分布会产生不平衡,FastTest 模型结合 smote 过采样技术来提高不平衡数据中分类器的分类性能。

2.2 层次 SoftMax

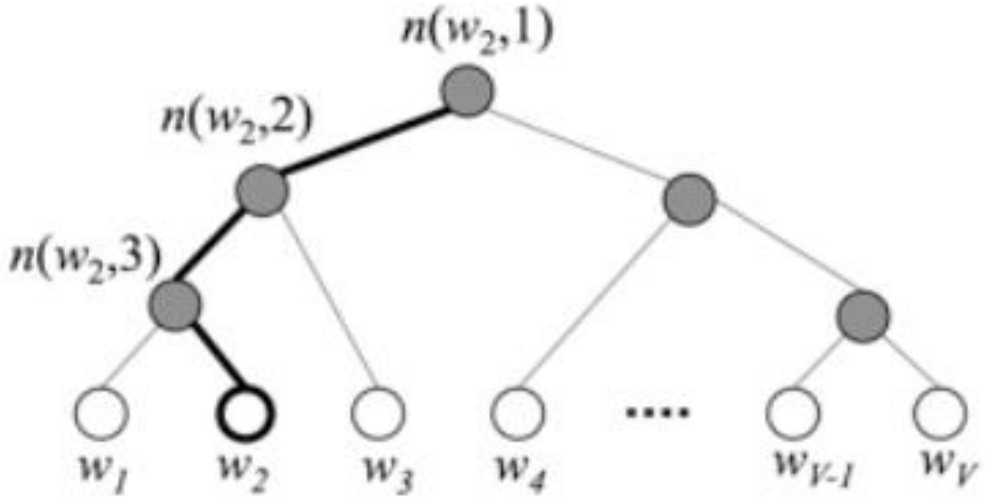
SoftMax 函数常在神经网络输出层充当激活函数，为了将神经元输出值进行归一化。将输出层的值归一化到 0-1 区间，将神经元输出构造成概率分布。Softmax 回归可以处理多类别问题，是将可以判定为某类的特征相加，然后将这些特征转化为判定是这一类的概率。类标 W 可以取 k 个不同的值。当训练集$\{X^1，W^1\}$，…,$\{X^I，W^i\}$，我们有 $W^i\in\{1,2,\dots\dots,k\}$。SoftMax 函数表示如下：

$$h_\theta(x)=\frac{1}{1+\exp(-\theta^T x)}$$

最小化代价函数:

$$J(\theta) = -\frac{1}{v}\left[\sum_{i=1}^{v} w^i \log h_\theta(x^i) + (1 - w^i)\log(1 - h_\theta(x^i))\right]$$

层次 SoftMax 技巧建立在哈弗曼编码的基础上，对标签进行编码，能够极大地缩小模型预测目标的数量。层次 softmax 相对于普通的 softmax 来说极大地提高了训练速度。在训练词向量时，Huffman 树中必存在一条唯一的从根节点到词 w 对应结点的路径。它的时间复杂度就是树的深度，这个深度是远远小于字典大小的。并且作为一个监督学习，在监督学习的过程中已经知道了要计算哪一条路径的概率。所以只需要计算霍夫曼树的一条路径，大大提高了计算效率。



分层 softmax 图

从根节点开始算概率，在这条路径上存在若干分支，将每个分枝看做一次二分类，每次分类产生一个概率，是目标词 w 的概率。从 root 结点开始随机走，走到目标词 w 形成一条路径,路径长度被表示为$L(W_i)$则$P(W_i)$表示如下：

$$P(W_i) = \prod_{i=1}^{L(W_i)-1} \sigma\left(\left|\left|n(W_j, l+1) = LC\left(n(W_j, l)\right)\right|\right) \cdot \theta n(W_j, l)^T X$$

其中$\theta n(W_j, l)$ 是非叶子结点$n(W_j, l)$ 的向量表示；X 是隐藏层的输出值，从输入词的向量中计算得来；$||A||$是一个特殊函数定义，表示如下：

$$||A|| = \begin{cases} 1 & n(W_j, l+1)是n(W_j, l)的左孩子 \\ -1 & n(W_j, l+1)是n(W_j, l)的右孩子 \end{cases}$$

通过分层的 Softmax,实现复杂度从|K|降低到 log|K|，在分层 softmax 图中能够看出从根节点到W_2的路径长度为 4，但是在这期间只做了三次的二次分类的逻辑回归。

2.3 word2vec 负采样文本抽取技术

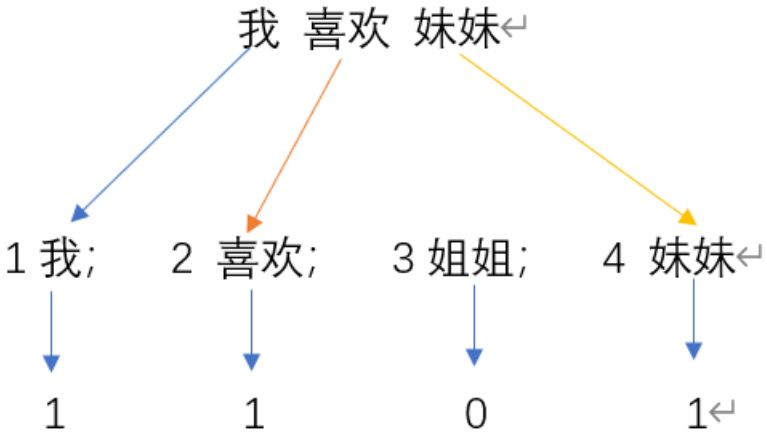
Fasttext 模型作为浅层神经网络，FastText 模型对于多数类的预测性比较高，对于少数类的预测性而言，存在一定性的不平衡数据。对于不平衡数据还是要降低其影响。One-Hot 编码是分类变量作为二进制向量的表示。每个整数值被表示为二进制向量，其中绝大多数元素为 0，只有一个维度的值为 1。One-hot 编码的维数由词典长度而定，One-hot 编码下常用的特征提取是词袋模型。就是将输入数据转化为对应的 Bow 形式。One-hot 编码是如何进行特征提取的，我们用一个例子来表示。语料中有三段话

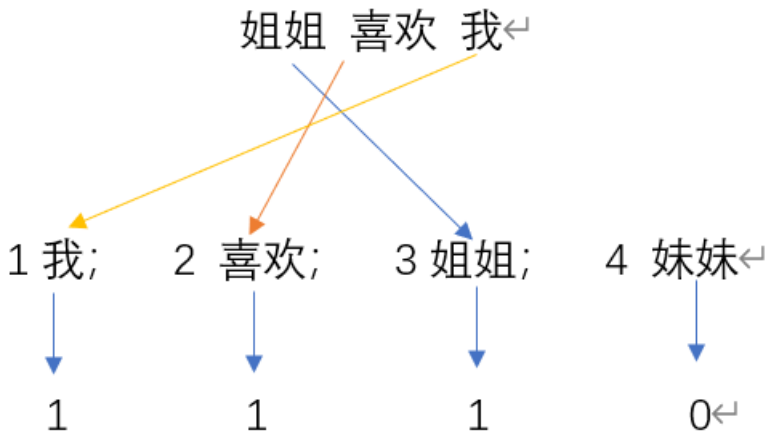
[“我喜欢妹妹”，“姐姐喜欢我”，“姐姐喜欢妹妹”]

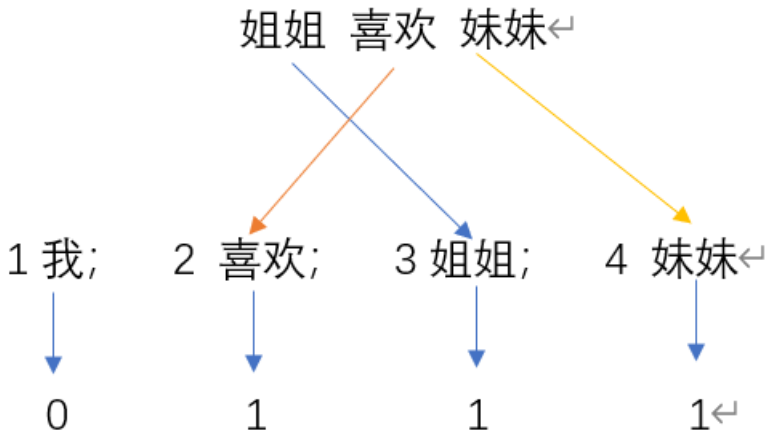
将语料库中的每句话分成单词，并编号：

1：我 2：喜欢 3：姐姐 4：妹妹

然后，用 one-hot 对每句话提取特征向量：







最终得到的每句话的特征向量就是：

我喜欢妹妹 -> [1，1，0，1]

姐姐喜欢我 -> [1，1，1，0]

姐姐喜欢妹妹 -> [0，1，1，1]

而利用负采样技术，一个词 v 的上下文是 context(v)，那么词 v 就是一个正例，其他词就是一个负例。在语料库里面，每个词语所出现的频率是不一样的，负采样技术就是我们采样的时候高频词选中的概率大，而低频词选中的概率小。比如说词典 D 中的每一个词 v 对应线段的一个长度，任何采样算法都应该保证频次越高的样本越容易被采样出来。基本的思路是对于长度为 1 的线段，根据词语的词频将其公平地分配给每个词语：

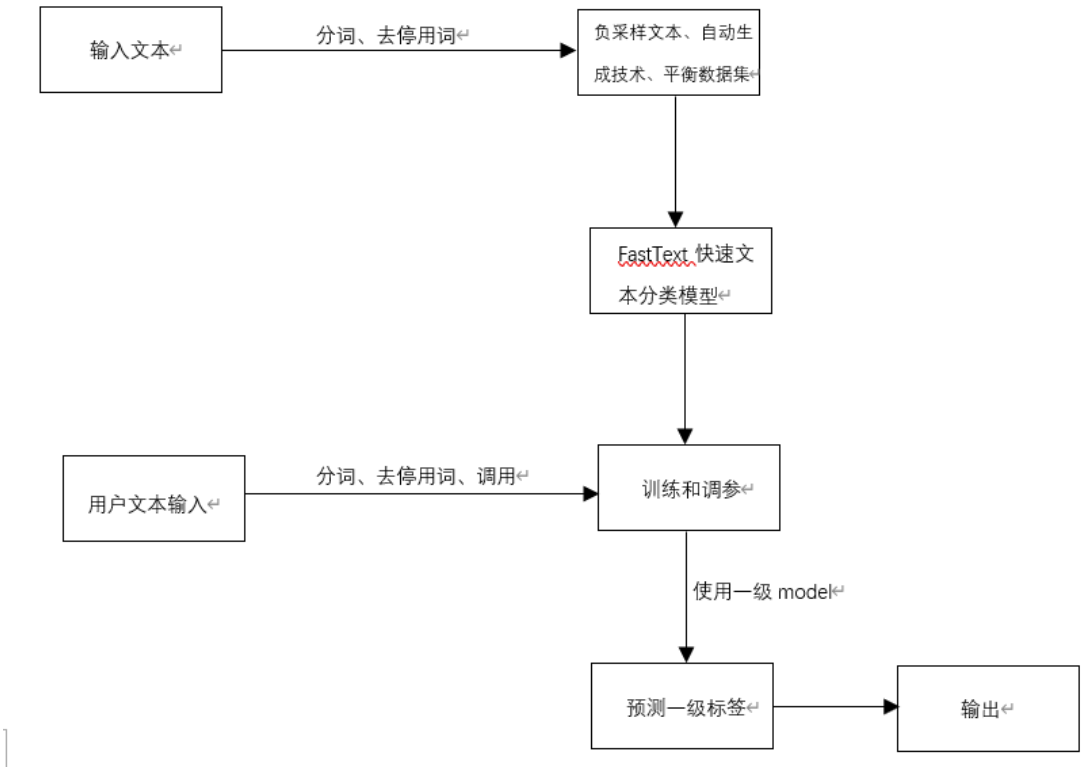
$$len(v) = \frac{counter(v)}{\sum_{u \in D} counter(u)}$$

counter 就是 w 的词频。

我们将该线段分配后生成 0-M 之间的整数，在 word2vec 中，去一查就能 table 数组抽中单词。公式如下：

$$len(v) = \frac{[counter(v)]^{0.75}}{\sum_{u \in D}[counter(u)]^{0.75}}$$

2.4 fastText 的精准快速分类模型



fastText 的精准快速分类模型

首先对每个文本进行分词、去停用词进行处理，提取词向量后得到句向量利用负采样文本抽取技术进行取样。将生成的中文分词与原始数据进行合并处理，之后输入到 fasttext 模型进行快速训练。对于输入的文本，使用一级 model 预测其一级标签，最后得到我们需要的结果。

三、预处理

3.1 数据清洗

在留言数据文本中存在特殊字符。为了更好的对数据进行处理，从而进行数据清洗。本文采用正则表达式[1]对特殊字符进行处理。正则表达式可以将匹配的子串进行替换，首先利用 findall 查找特殊字符，再用 split 对文本和特殊字符分割，最后 sub 对特殊字符进行替换，进而完成数据清洗。

3.2 分词和词性

对留言进行分词处理，在文本中提取词语要先进行分词，中文文本并不是以词进行断句，所以在句子中不能明显区分两个词语。本文对句子里面的中文文本进行词语分词，采用的是 Python 中文分词组件 jieba 分词[2]。 jieba 分词实现高效的词图扫描，利用字典将句子中汉字所有可能成词情况进行标识，构成有向无环图。每个有向边都赋予相应的权值，最后统计出有向无环图的最短路径，最短路径便是词频的最大切分组合。在分词完成之后，通过添加自定义词典，来标注句子分词后每个词的词性。自行添加新词可以保证更高的正确率。

3.3 去停用词

在留言中出现频率很高，但实际意义又不大的词称为停用词。在留言的长文本中，有很多是无意义的表达，只需保留一些关键词。停用词主要包括了语气助词、副词、介词、连词等，通常自身并无明确意义。本文的停用词词典取至哈工大停用词表，可以缩小搜索范围，同时提高搜索的效率。在数据预处理的时候，会先将停用词进行删除。比如标点符号和中文的“哈哈、的、了、啦”等。

3.4 文本特征选择

目前大多数中文文本分类系统都采用词作为特征项，作为特征项的词称

作特征词。如果把所有的词都作为特征项，那么特征向量的维数将过于巨大，会对分类系统的运算性能造成极大的压力。所以寻求一种有效的特征降维方法，不仅能降低运算复杂度，还能提高分类的效率和精度，是文本自动分类中一项重要技术。

文本应用使用基于数学方法进行特征选择比较精确，人为因素干扰少。这种方法通过构造评估函数，对特征集合中的每个特征进行评估，并对每个特征打分，让每个词语都获得一个权值。然后将所有特征按权值大小排序，提取预定数目的最优特征作为提取结果的特征子集。

文档频数(Document Frequency， DF)指的是在整个数据集中有多少个文本包含这个单词。将训练文本集中对每个特征计算它的文档频数，若该项的DF 值小于某个阈值则将其删除，若其 DF 值大于某个阈值也将其去掉。DF 的优点在于计算量小，速度快。它的时间复杂度和文本数量成线性关系，所以非常适合于超大规模文本数据集的特征选择。文档频数还非常地高效，在有监督的特征选择应用中当删除 90%单词的时候其性能与信息增益和卡方校验统计的性能相当。

3.5 文本表示

我们要将文本语言转换成计算机能够理解的数字形式，才能进一步进行神经网络进行训练。Ont-hot 编码简单易懂，它只有一个分量的值为 1，其他分量均为 0。但是 Ont-hot 编码太过于稀疏，不考虑词与词之间的顺序，并且它假设词与词相互独立。但是在大多数情况下，词与词是相互影响的。

FastText 是 word2vec 所衍生出来的，也是开源的一个文本分类器。FastText 将句子中的每个词先通过一个 lookup 层映射成词向量，然后对词向量取其平均值，将这个值作为整个句子的句子向量，再用线性分类器进行分类，从而实现文本分类。并且 FastText 结构简单，训练的更快。

3.6 文本相似度

在热点问题的挖掘上，涉及到如何度量两个文本的相似度问题。要计算相

似性，首先将文本转换成可计算的数。反复出现，出现的次数越多（即 term frequency 越大），这个词就可能越重要。出现的次数可能很多，但是这个词太普通了，在很多文本中，这个词都出现了（即 document frequency 较大）。

TF-IDF 方法就是从这两个角度出发定义的词语重要程度，由这种方法计算出来的词语重要程度就叫 tfidf 值。先求出每个词的 tfidf 值，再求出它们的相似度。比如 tf 除以最大的 tf 值，对 idf 值取对数等来达到规范我们得到的结果的目的。

四、实验

4.1 实验环境

本实验以 windows 10 系统下的 python3.7 作为软件基础；CPU 频率为 2.50GHz，核心数量为 4 个，显卡为 GTX 1050，内存为 8G。

使用到的 python 库有：fasttext、sklearn、jieba 等。

包/库	版本号
Python	**3.7**
Pandas	**0.24.2**
jieba	**0.39**
Fasttext	**0.9.1**
numpy	**1.8.3**
sklearn	**0.22.2**

CPU	CPU v4 @ 2.50GHz
内存	RAM 8GB
显卡	GTX 1050

4.2 实验数据

本实验实验数据来源选取出题方给予的附件 1.xlsx、附件 2.xlsx、附件 3.xlsx、附件 4.xlsx。

名称	修改日期	类型	大小
附件1.xlsx	2020/2/29 11:45	Microsoft Excel ...	22 KB
附件2.xlsx	2020/4/24 9:44	Microsoft Excel ...	5,328 KB
附件3.xlsx	2020/4/24 9:50	Microsoft Excel ...	2,363 KB
附件4.xlsx	2020/4/24 9:49	Microsoft Excel ...	2,587 KB

4.3 实验评价指标

对题目拆解可分为三个小题：

第一小题为“根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型”，评价指标为 F-Score 和准确率对分类模型进行评价。

第二小题为“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，并给出评价结果”。评价指标为聚类的轮廓系数值。

第三小题为“对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案”，此题为开放性题目，评价指标为方案的可行性。

4.4 实验过程及结果

第一小题：

#先运行 数据初始化.py 文件

（1）我们使用 pandas 的库打开附件二，对数据进行预处理操作

第一步，采用 jieba 分词库对文本进行分词。

第二步，在分词的过程中，进行过滤某些特殊字符。

第三步，对文本进行标记分类类别采用__label__的方式进行间隔

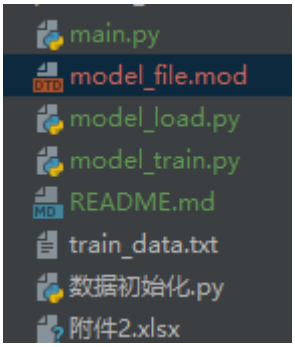
第四步，保存成文本文件，这个文本文件将作为训练集进行训练

```
A 市 西湖 建筑 集团 占 道 施工 有 安全隐患  __label__城乡建设
A 市 在水一方 大厦 人为 烂尾 多年 ， 安全隐患 严重  __label__城乡建设
投诉 A 市 A1 区苑 物业 违规 收 停车费  __label__城乡建设
A1 区 蔡锷 南路 A2 区华庭 楼顶 水箱 长年 不洗  __label__城乡建设
A1 区 A2 区华庭 自来水 好大 一股 霉味  __label__城乡建设
投诉 A 市 盛世 耀凯 小区 物业 无故 停水  __label__城乡建设
咨询 A 市 楼盘 集中 供暖 一事  __label__城乡建设
A3 区 桐梓 坡 西路 可可 小城 长期 停水 得不到 解决  __label__城乡建设
反映 C4 市 收取 城市 垃圾处理 费 不 平等 的 问题  __label__城乡建设
A3 区 魏家坡 小区 脏乱差  __label__城乡建设
A 市 魏家坡 小区 脏乱差  __label__城乡建设
A2 区 泰华 一村 小区 第四届 非法 业委会 涉嫌 侵占 小区业主 公共 资金  __label__城乡建设
A3 区梅 溪湖 壹号 御湾 业主 用水 难  __label__城乡建设
A4 区鸿涛 翡翠 湾 强行 对 入住 的 业主 关水 限电  __label__城乡建设
地铁 5 号线 施工 导致 A 市锦楚 国际 星城 小区 三期 一个月 停电 10 来次  __label__城乡建设
```

（2）进行模型的构建，以及模型的训练

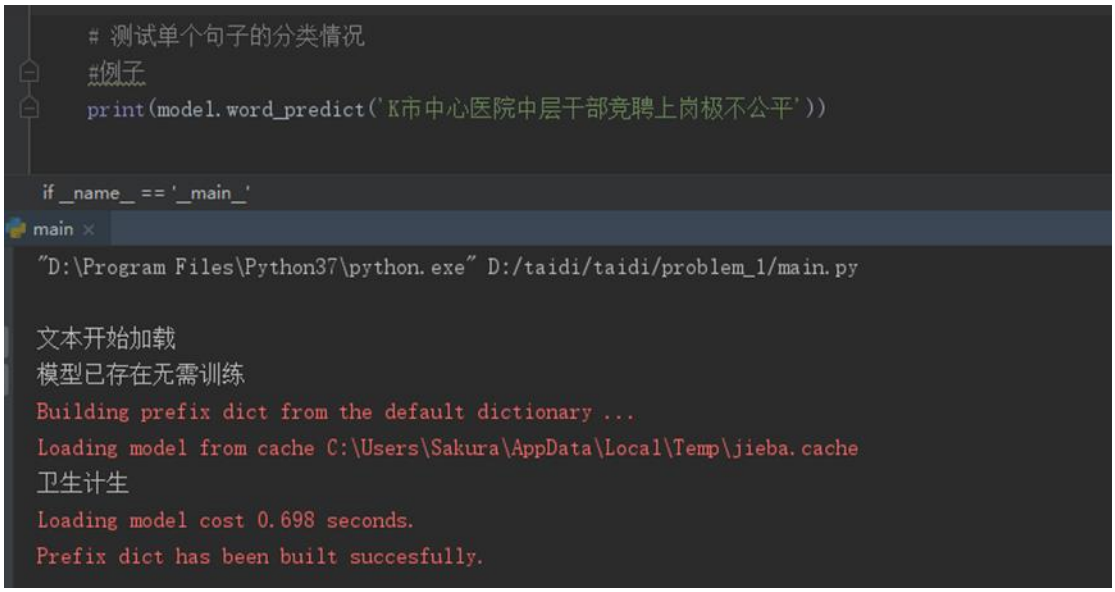
```
Read 0M words
Number of words:  15167
Number of labels: 7
Progress: 100.0% words/sec/thread:  332434 lr:  0.000000 loss:  0.056109 ETA:   0h 0m
```

这里我们能看到，模型训练出了 15167 个词，文本类别 7 种，为了模型的持久化，我们将以二进制文件形式保存到当前文件夹，文件名称是 model_file.mod



（3）加载模型，以及对文本的预测

我们使用'K 市中心医院中层干部竞聘上岗极不公平‘句子进行文本分类的预测，



我们看到模型对该句子进行了预测，预测结果是卫生计生，符合我们的预期结果。

使用 F-Score 对分类方法进行评价：

$$F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

模型使用 F-Score 对分类方法进行评价得到的结果，F1 值、正确率、查准率和查全率高达 0.99。

第二小题：

（1）本次小题采用 TF-IDF 将预处理好的留言标题进行权重计算

```
In[5]: m_array.shape
Out[5]: (4326, 7254)
```

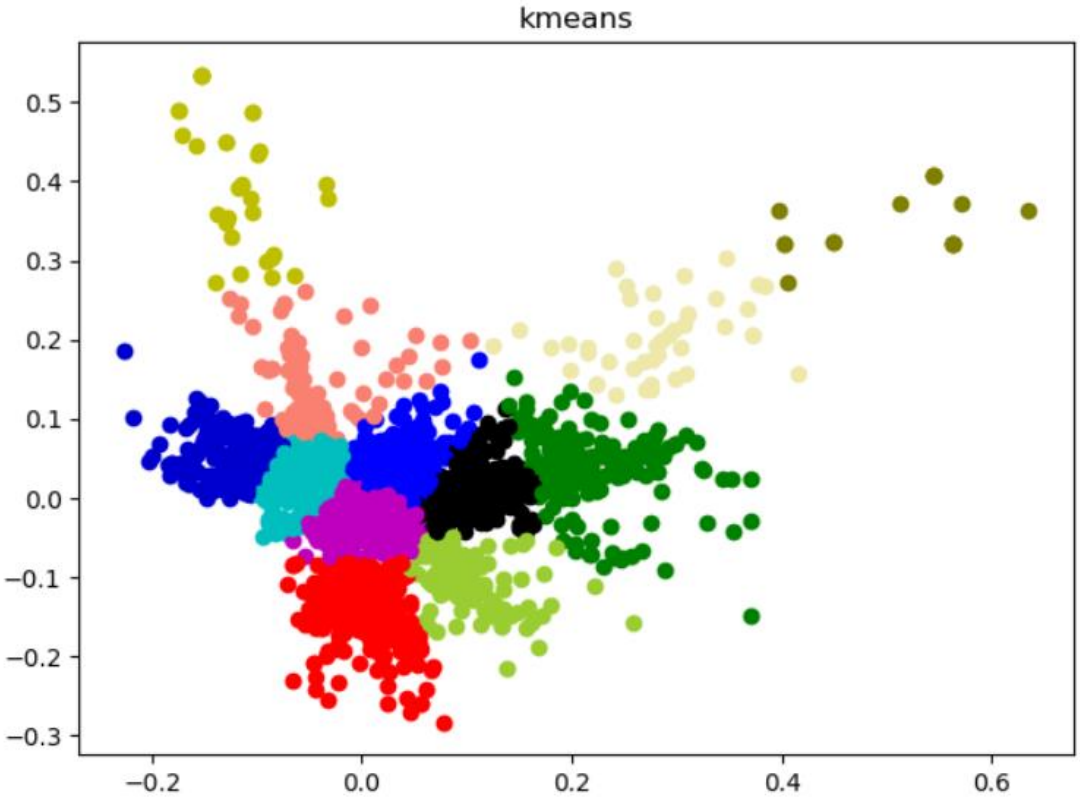
（2）处理好的 TF-IDF 权重矩阵是一个（4326，7254）的稀疏矩阵，需要通过 PCA 算法将稀疏举证进行降维。

```
In[8]: m_array2.shape
Out[8]: (4326, 2)
```

（3）使用 Kmeans 算法进行聚类，得到轮廓系数。

```
Out[19]: 0.46862501431699677
```

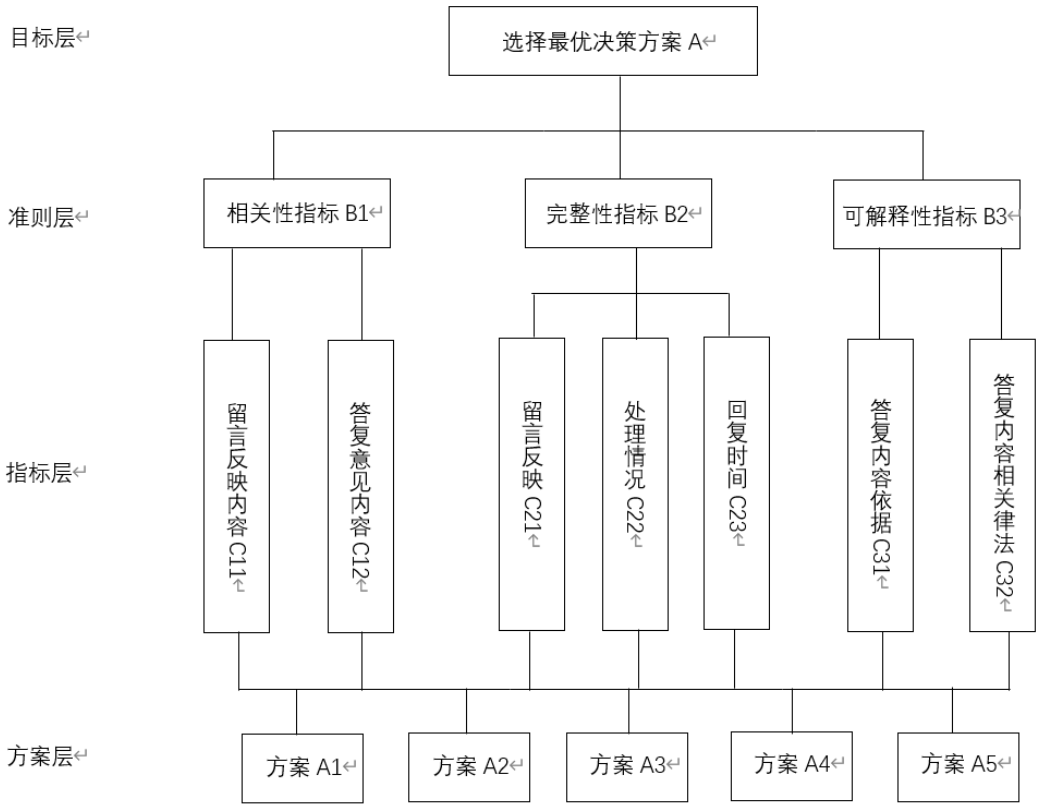
（4）将聚类结果进行绘图展示。（总共分为 12 类）



（5）选取聚类前五的标签按要保存为热点问题表.xls 和热点问题留言明细表.xls。

第三小题：

层次分析法：根据答复的相关性指标、完整性指标、可解释性指标可建立判别矩阵，从判别矩阵中可以得到指标的归一化权重（W 相、W 完、W 可）。



五、 总结和展望

综上所述，我们提出了一款 FastText 精准快速文本分类模型，它能够精准的对一级标签进行分类，我们的模型在线下测试集上，精准度和泛化能力比较好且 F1 值达到 0.99。并且能够很好的处理效率低，且差错率高问题。在有关热点问题采用 TF-IDF + K-means。先计算词语的重要程度，求出每个词的 tfidf 值，再求出它们的相似度。K-means 聚类算法采用距离作为相似性的评价指标，认为类簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。最终将选取聚类前五的标签按要保存为热点问题。在评价方案系统中，通过对答复的相关性、完整性、可解释性进行量化处理，使用层次分析法基于回复内容，对多个待评价方案进行评价，从而得到方案的重要性排序。利用建立重要判别矩阵，从判别矩阵中可以得到指标的归一化权重。最终得到的评价方案的最终综合权重，最终选取置最优的回复评价方案作为答案。接下来我们把如何提高聚类数据，降低困惑度作为主要关注点，进行更高更快速的训练。

六、参考文献

[1] https://www.runoob.com/regexp/regexp-syntax.html

[2]https://github.com/fxsjy/jieba

Bag of Tricks for Efficient Text Classification

Enriching Word Vectors with Subword Information

《机器学习》，周志华，清华大学出版社

龚千健. 基于循环神经网络模型的文本分类[D]. 2016.

屈渤浩. 基于改进 FastText 的中文短文本分类方法研究[D].辽宁大学,2018.

李少温. 基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D].华中科技大学,2019.

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.

吋义成,柯丽华,黄德育.系统综合评价技术及其应用[M].北京:冶金工业出版社,2006