“智慧政务”中的文本挖掘应用

摘要

近年来，随着科技的发展，对于民众的心声我们不需要再像以前一样必须去政府提交我们的匿名信才能反映，现在我们可以通过合理运用网络平台（微信、微博、市长信箱、阳光热线等）来向政府表达我们的心声，也使得我们日常的民情民意更方便化、更快捷化和更多样化。同时，民众的心声也更多的传递到了政府的手里。然而，带来的问题也层出不穷，例如文本数据的不断攀升，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。因此，运用自然语言处理技术和数据挖掘技术对“智慧政务”中文本挖掘的应用具有重大的意义。

针对问题一，根据所收集的群众留言数据，为了便于将群众留言分派到相对应的职能部门进行处理，建立关于留言内容的一级标签分类模型。通过对群众留言进行数据清洗，jieba中文分词工具分词，利用 TF-IDF 算法提取群众留言内容中的关键词，使用 K-means 聚类算法对 TF-IDF 权重向量进行聚类，再根据 KNN 算法为群众留言信息加上标签。最后使用 F-Score 对分类方法进行评价，得出最优分类模型。

针对问题二，利用 python 对 Excel 表格中数据进行特殊字符处理、利用正则表达式对文本替换、删除、查找，以及分词和停用词过滤操作，实现数据优化。再运用相关数字方法进行文本特征选择。随后将表中数据的时间、地点/人群、问题这三要素通过 Word Embedding 分词构建词向量，将自然语言数值化。然后根据卷积神经网络对文本数据分类。最后给出合理的热度评价指标。

针对问题三，针对问题三，依照有关部门对群众的留言给出答复，然后以答复的相关性、完整性已经可解释性作为参考标准来对答复的意见质量作出相关的评价方案并试着去实现完成。即是从答复意见的内容是否与问题相关、是否满足某种规范、答复意见中内容的相关解释的角度来衡量答复意见。通过利用 python 对答复的相关性、完整性、可解释性进行量化，最后构建出指标来计算和评价相关部门的答复内容是否成熟可行。

关键词：Python　TF-IDF 算法　文本分类　K-means 聚类算法　KNN 算法　TextCNN

Application of text Mining in Intelligent Government Affairs

Abstract

In recent years, with the development of science and technology, we don't need to submit our anonymous letters to the government to reflect the people's voice as before. Now we can express our voice to the government through reasonable use of network platforms (wechat, microblog, mayor's mailbox, sunshine hotline, etc.), which also makes our daily public opinion more convenient, faster and More variety. At the same time, the voice of the people is more transmitted to the hands of the government. However, there are also endless problems, such as the continuous rise of text data, which has brought great challenges to the work of the relevant departments which mainly rely on human to divide messages and sort out hot spots in the past. Therefore, using natural language processing technology and data mining technology is of great significance to the application of text mining in "smart government".

To solve the problem one, according to the collected data of the public message, in order to facilitate the distribution of the public message to the corresponding functional departments for processing, the first level label classification model of the message content is established. Through the data cleaning of the mass message, the Chinese word segmentation tool of Jieba is used to segment the words, TF-IDF algorithm is used to extract the keywords in the mass message content, K-means clustering algorithm is used to cluster TF-IDF weight vector, and then KNN algorithm is used to tag the mass message information. Finally, F-score is used to evaluate the classification method and get the optimal classification model.

In order to solve the second problem, we use Python to process the data in Excel

table with special characters, use regular expression to replace, delete and search the text, and filter the segmentation and stop words to optimize the data. Then we use the relevant digital methods to select the text features. Then, the time, place / population and problem of the data in the table are used to construct the word vector through word embedding segmentation, and the natural language is digitized. Then text CNN (convolutional neural network) is used to classify the text. Finally, a reasonable heat evaluation index is given.

In response to question 3, in response to question 3, according to the comments of the relevant departments to the masses, and then take the relevance and integrity of the replies as the reference standard to make relevant evaluation plans for the quality of the replies and try to achieve the completion. That is, from the perspective of whether the content of the reply is related to the question, whether it meets some norms, and the relevant interpretation of the content of the reply. Through the use of Python to quantify the relevance, integrity and interpretability of the response, and finally build indicators to calculate and evaluate whether the response content of relevant departments is mature and feasible.

Keywords: Python TF-IDF weighted text classification K-means clustering KNN algorithm TextCNN

目录

1. 挖掘目标

本次建模目标是利用网络问政平台系统发布的群众留言信息数据，基于Python语言、结合jieba中文分词工具、K-means聚类的方法、KNN算法及TextCNN，达到以下三个目标：

（1） 采用文本预处理、特征选择、特征权重计算和分类对群众留言数据进行文本挖掘，根据分类模型结果，使用 F-Score 对分类方法的所得出的模型进行评价，选取最优模型，降低传统人工处理的工作量和差错率，提高工作效率。

（2） 利用python和Word Embedding分词构建词向量将自然语言数值化，根据文本归类得出合理的热度评价指标，结合评价结果，有利于相关部门有效地处理问题，提高服务效率。

（3） 分析相关部门对群众留言给出的相应回答，要依据这个答复从多种角度分析给予适合的方案并真正的去完成。

2. 分析方法与过程

2.1 总体流程图

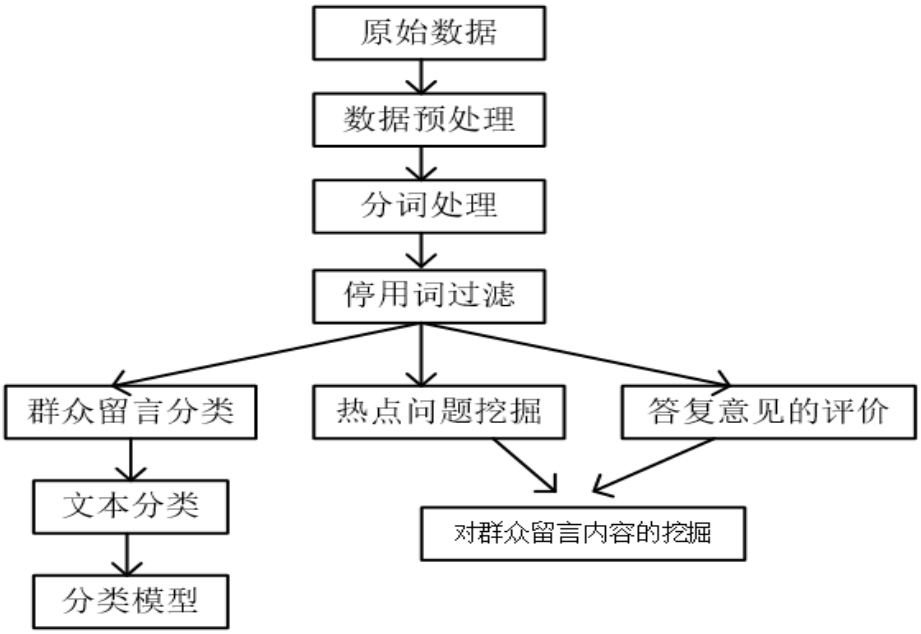


图 1：总流程图

本文主要分为四个步骤：

步骤一：获取准备分析的原始数据（群众留言）；

步骤二：数据预处理，对于群众留言内容，出现了许多重复的留言内容，需要对重复的留言进行分词处理和停用词过滤等方法做相应处理；

步骤三：群众留言数据经过处理后，开始针对三个问题分别从不同角度进行分析解决；

步骤四：对所给出模型进行评价优化，获取有价值的内容。

2.2 具体步骤

步骤一：获取准备分析的原始数据（群众留言）。

本文所使用的数据主要由留言主题、留言时间、留言详情及分类标签组成，所需数据均可通过筛选得到。随后即可根据筛选所得数据进行分析解答。

步骤二：数据预处理，对于群众留言内容，出现了许多重复的留言内容，在原始数据上进行分词处理和停用词过滤等操作。

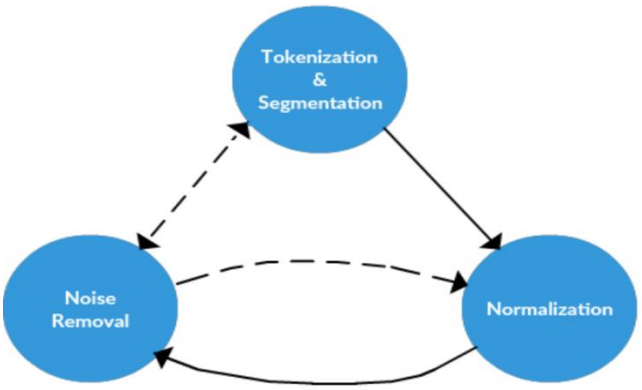


图 2：数据预处理基本框架

（1） 数据收集（在步骤一种已收集到相关数据）。

（2） 去除数据中非文本部分（数字、标点、字母），让文本数据只保留汉字部分。

对于表中给出的数据中的数字、标点、字母部分我们可以使用 Python 中的正则表达式（Re 模块）进行删除，删掉这些不是相关内容的数据后，则可以真正的

进行文本预处理部分。

（3） 对文本数据进行 jieba 分词。在此过程中，我们使用了 jieba 分词中的精简模式对文本内容进行操作，并对词性进行了相应的标注。

（4） 去停用词，过滤掉无意义词。在表格所给出的文本内容中，我们可发现其中有很多无意义的词以及无需在分析时引入的词，所以我们采用了停用词词典过滤掉无意义的词语。

（5） 用数字方法选取出最具有分类信息的内容。在此通过自定义函数 doc2onehot_matrix()实现 one-hot 编码向量化文本对数据进行特征处理，从而选取出含有特征性的分类信息。

（6） 建立分析模型。

步骤三：群众留言数据经过处理后，开始针对三个问题分别从不同角度进行分析解决。

首先我们由 TD-IDF 算法提取群众留言中的关键词，利用 K-means 算法为数据分类，KNN 算法负责找出与各个中心类似的元素，即：

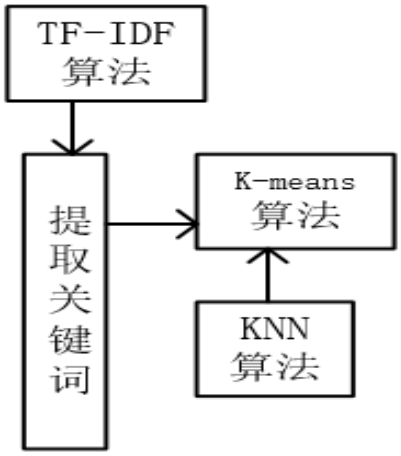


图 3：算法步骤

a.TF-IDF 算法的原理如下：

在对群众留言进行分词以后，需要将分词后得到的这些词语转换成向量，方便需要的时候可以使用。现在采用到的 TF-IDF 算法，就可以把群众留言的信息描述转换成权重向量。TF-IDF 算法的原理如下：

第一步，计算 TF 权重。

词频（TF）=某个词在文本中出现的次数

由于留言内容有长的也有短的，所以为了方便比较不同留言内容，需要进行“词频”对比化，除以留言文本的所有词条数目。

公式：$tf_{ij}=\frac{n_{ij}}{\sum_k n_{kj}}$

即： $TF=\frac{在某个词在文本中出现的次数}{该文本中所有的词条数目}$

其中，ni,j 是该词在文本 aj 中出现的次数，分母是文本 aj 中所有群众留言文本出现的次数总和；

第二步，计算 IDF 权重，|D| 是语料库中的文本总数。 |{j:ti∈cj}| 表示包括词语 ti 的文本数量（即 ni,j≠0 的文本数量）。如果这个词语不在语料库中，就要使用 1+|{j:ti∈cj}|，即：

$IDF=\log(\frac{语料库的文本总数}{包含该词的文本数+1})$，分母加 1 是为了避免分母为 0

第三步，计算 TF-IDF 值，某一特定留言内的高词语频率，以及该词语在整个留言文本集合中的低文本频率，可以产生出高权重的 TF-IDF。

公式：TF-IDF=TF*IDF

b.K-means 聚类算法原理：

K-means 算法是一种关于距离的算法，若用数据表达式，就比如有 K 个簇，簇就划分为$(C_1, C_2, \ldots C_k)$，那么我们的目标就是最小化平方误差 E：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$$

其中 μ_i 是簇的均值向量，也被叫做质心，：

公式：

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

K-means 聚类算法有 5 个步骤：

(1)从文本数据中选择 k 个簇作为最开始各自的中心；

(2)计算每个簇到聚类中心的距离来划分；

(3)再次计算各个聚类中心；

(4)计算标准测度函数，如果达到了最大迭代次数，就可以停止了，否则，继续操作，直到聚类得出的结果不变化；

（5）将结果输出。

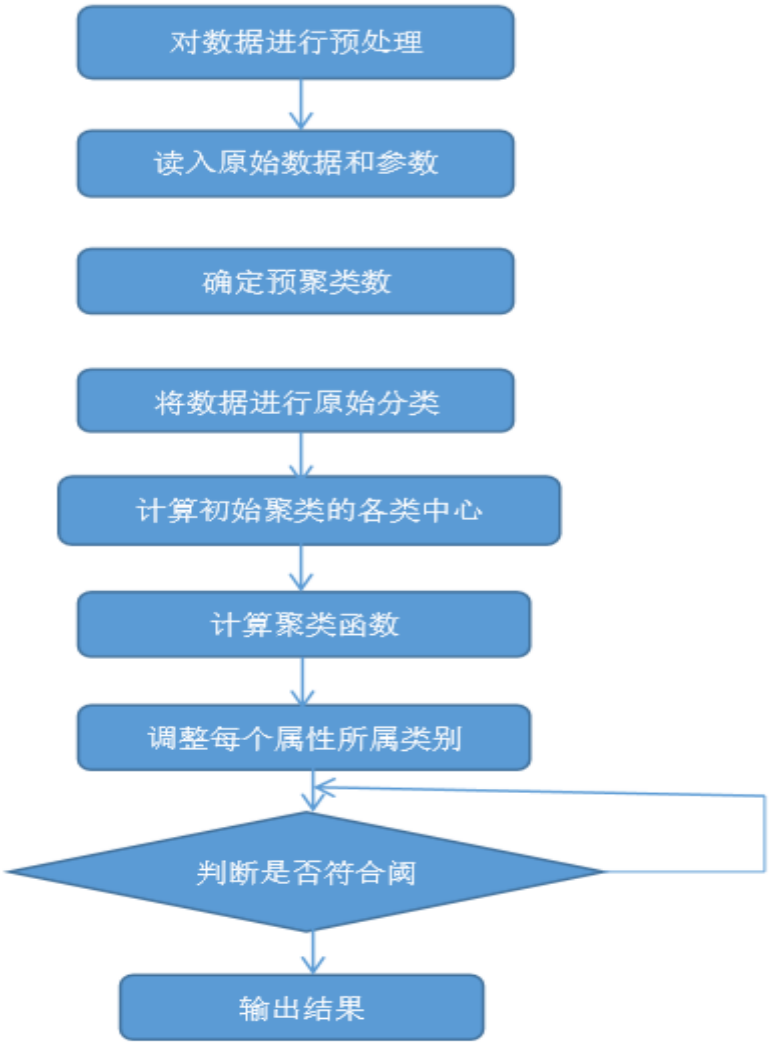


图 4：K-means 聚类算法

c.KNN 算法原理：

KNN 算法可以概括为将一个样本空间的部分样本分为几个类别，接着，给出一个待分类的数据，通过计算找出与自己最接近的 K 个样本，由这 K 个样本来决定得出分类数据归为哪一类。KNN 算法在类别决策时，只与极少量的相邻样本有关。由于 KNN 算法主要靠相邻的样本，而不是靠辨别类域的方法来确定所属类别的，所以说相对于交叉和重叠比较大的数据来看 KNN 算法是再合适不过的了。

对于 KNN 算法，我们通过计算每个数据样本间的距离来表示每个数据样本间的指标，从而避免了对象之间的匹配问题，在这里我们计算距离的时候一般使用欧氏距离或曼哈顿距离：

欧式距离：d(x,y)=$\sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$

曼哈顿距离：d(x,y)=$\sqrt{\sum_{k=1}^{n}|x_k - y_k|}$

KNN 算法是通过测量不同样本之间的距离进行分类。它的的思路是：如果一个样本在样本空间中的 k 个最相似(即样本空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。K 通常是不大于 20 的整数。KNN 算法中，所选择的邻居都是已经正确分类的对象。这个方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。我们可以画出一个图来描述。

如图所示：

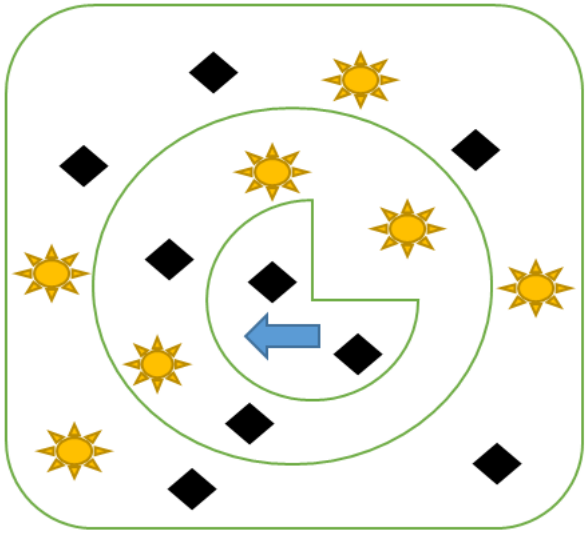


图 5：KNN 算法

从图中我们可以看到两个类型的数据，即黄色小太阳和黑色菱形两种来分别表示，二蓝色小箭头表示待分类数据。

当 K=2 时，图中的蓝色小箭头离得最近的就是两个黑色菱形，所以此时蓝色小箭头离属于黑色菱形这一类。

当 K=5 时，图中的蓝色小箭头离得最近的是两个黑色菱形和三个黄色小太阳，所以此时蓝色小箭头离属于黄色小太阳这一类。

当 K=9 时，图中的蓝色小箭头离得最近的是五个黑色菱形和四个黄色小太阳，所以此时蓝色小箭头离属于黑色菱形这一类。

因此，通过 KNN 算法的计算和评估，我们就可以将群众留言成功进行分类。

d. Text-CNN 模型

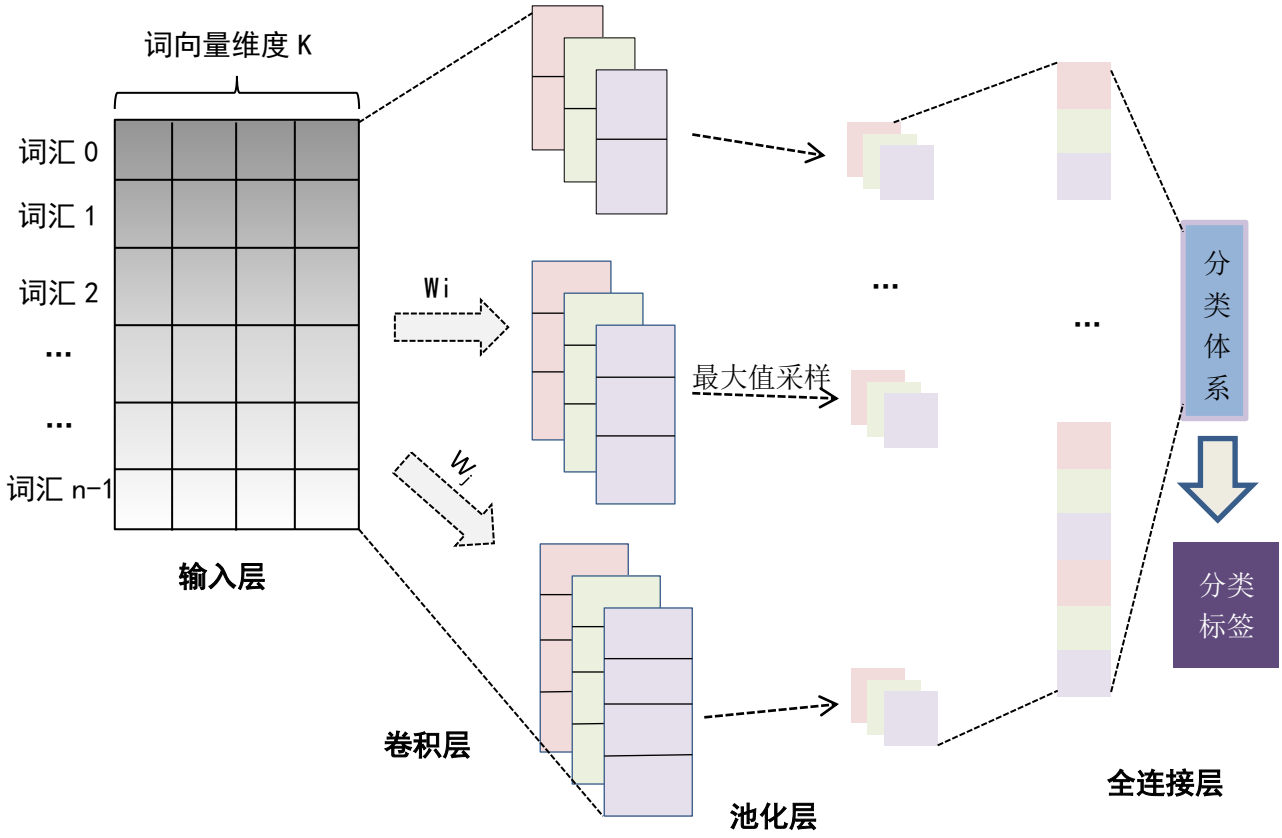


图 6 Text-CNN 框

上图为 Text-CNN 模型框架图。首先假设一个句子由n维向量组成，则我们输入的矩阵大小为x * y（x 为句子的长度），此时 CNN 需对输入样本进行卷积，最后可得到卷积后的向量，然后对所得向量进行池化并连接池化值，得到特征表示，随后将其进行分类，得到结果。具体算法如下：

第一步：利用 Word Embedding 分词构建词向量

通过相应 embedding 方式将之前分词后的词语映射为 n 维词向量，例如“明天”：[0,0,0,0,1]，“天气”：[0,0,0,1,0]等

明天 →	0	0	0	0	0
天气 →	0	0	0	1	0
很好 →	0	0	1	0	0
， →	0	1	0	0	0
出去 →	1	0	0	0	0
玩 →	0	0	0	0	1

第二步：假设将词语都用三维向量来表示，等式左侧是词语，右侧是卷积矩阵，则可利用卷积公式：

$$\begin{pmatrix} \omega_{11} & \omega_{21} & \omega_{31} & \omega_{41} \\ \omega_{12} & \omega_{22} & \omega_{32} & \omega_{42} \\ \omega_{13} & \omega_{23} & \omega_{33} & \omega_{43} \end{pmatrix} \odot \begin{pmatrix} f_{11} & f_{21} \\ f_{12} & f_{22} \\ f_{13} & f_{23} \end{pmatrix}$$

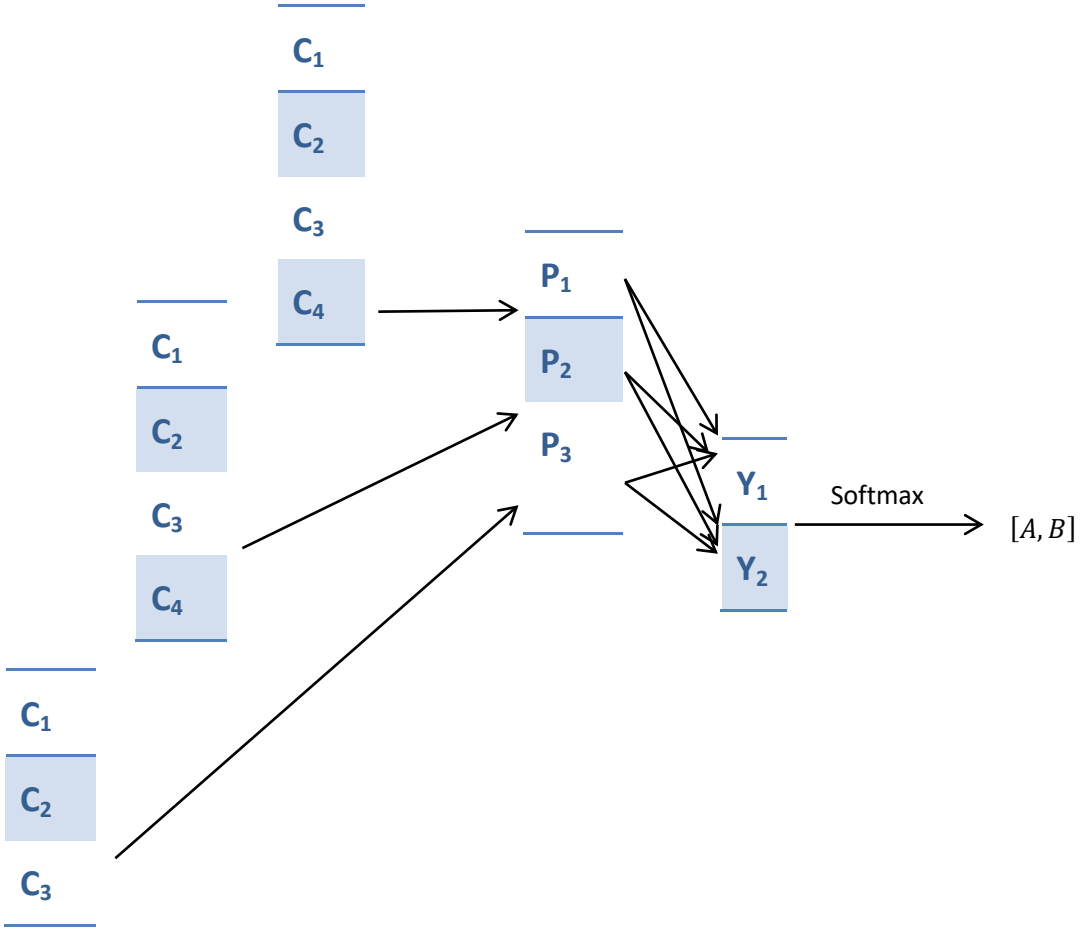
得到输出为：

$$\sigma_1 = \omega_{11}f_{11} + \omega_{12}f_{12} + \omega_{13}f_{13} + \omega_{21}f_{21} + \omega_{22}f_{22} + \omega_{23}f_{23}$$

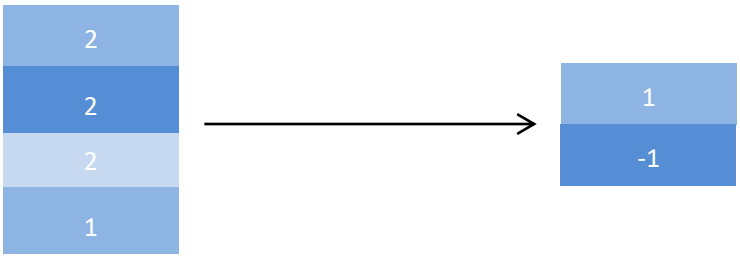
$$\sigma_2 = \omega_{21}f_{11} + \omega_{22}f_{12} + \omega_{23}f_{13} + \omega_{31}f_{21} + \omega_{32}f_{22} + \omega_{33}f_{23}$$

$$\sigma_3 = \omega_{31}f_{11} + \omega_{32}f_{12} + \omega_{33}f_{13} + \omega_{41}f_{21} + \omega_{42}f_{22} + \omega_{43}f_{23}$$

第三步：随后对该结果做最大值池化，则可得σ中的最大值，得出最显著特征。也就是减少模型参数，保证在输出中得到一个定长的输入。



第四步：最后将 max-pooling 的结果连接起来，送到 softmax 中，得到每一个类别。



Soft 示意图

步骤四：对所给出模型进行评价优化，获取有价值的内容。

本文我们采用了 Text-CNN 模型对文本进行分类，分类中运用了几个参数：维度，卷积核的个数等。但在这个过程中，发现数据中每一个类别的数量差异比较大，可能会影响最终效果。随后我们对数据进行了调整，打乱了数据的顺序，并且运用了代价敏感函数使该问题得以解决。

2.3 结果分析

a. 由于传统分类算法对网络留言识别率低，所以需要我们改进，这次最大程度的运用了算法的自动化和智能化，同时也方便工作人员分类，提高了工作人员的工作效率。

b. 本次实验对群众留言样本进行了综合测试，在不同特征数量评估指标下进行判别实验，实验表明，关注热点问题是一个重点，以便及时解决群众的难题，但是同时我们不能只关注热点问题而不理会其他反映少的问题，所以在这一方面还需要改进。

c. 针对群众留言的消息，我们需要及时给予回复，而对于回复的内容，政府也要相应去做到，这就要求相关部门付出实际行动让人民群众感到安心，这样群众留言才发挥了真正的作用。

3.结论：

做好政务公开，确保阳光公正透明已经成为了每一位政府工作者的内心准则。近些年来，随着各类网络问政平台的发展，许多群众也在这些网络平台上做出了关于各个方面的留言。这些留言成为了许多政府工作人员了解社会民意的一个重要渠道。然而，随着留言种类的复杂性的增加，如何高效的对留言采取措施已经成为了一个难题。

通过我们比赛的研究与分析，采用 TD-IDF 算法提取群众留言中的关键词，利用 K-means 算法分类，使用 KNN 算法找出与各个中心相似的元素，最终可以对群众留言分类，以及筛选出热点问题和对答复意见进行评价。

参加这次比赛，不必说我们第一次接触了 python 这一门面向对象、解释型计算机程序设计语言，也不必说第一次学习文本分类的知识内容，单是理解题目要义以及解决题目的各个问题就让我们团队花费了不少时间与精力。

在这个网络时代，随着大数据、云计算、人工智能等技术的发展，我们生活中的许多事都正在由人工转变成计算机工作。作为新时代大学生的我们是国家未来的栋梁之材，正所谓文化强国、科技强国，就是需要我们不断地学习新的理论文化知识和科技手段。因此，努力学习各种知识正是我们现阶段最重要的任务。

4. 参考文献

[1] 李春林, 冯志骥. 基于文本挖掘的新能源汽车用户评论研究[J]. 特区经济, 2020(04):148-151.

[2] 冯梦莹, 李红. 文本卷积神经网络模型在短文本多分类中的应用[J]. 金融科技时代, 2020(01):38-42.

[3] 吴雯雯, 陈振林. 基于蒙特卡洛 k-means 聚类算法的舰船器材分类研究[J]. 计算机测量与控制, 2020, 28(04):222-226.

[4] 刘昀晓, 王东峰, 曹林, 杜康宁, 李萌, 付冲. 基于车辆数据的 k 近邻联合概率数据关联算法[J]. 电讯技术, 2020, 60(04):448-454.

[5] 曾凡锋, 李玉珂, 肖珂. 基于卷积神经网络的语句级新闻分类算法[J]. 计算机工程与设计, 2020, 41(04):978-982.