人工智能实验（一）

实验报告

姓名：钟怡宁

学号：17341217

班级：教务三班

专业：计算机科学与技术

一：实验目的  
 要求掌握将简单文本数据集处理成TF-IDF矩阵的方法。要求理解掌握KNN最近邻算法，并用KNN最近邻算法解决分类和回归两类问题。

二：算法原理

1. TF-IDF矩阵，用来处理简单的文本数据集，即对文本编码，实现文本的可计算性。该算法分三步，第一，将文本中出现的所有单词进行不重复提取，按照出现次序排列，得到的单词总数为即将建立的TF-IDF矩阵的列数；第二，遍历原文本数据的每一行，将该行中每个单词出现的次数填入单词出现对应的位置，即建立该行文本的向量表达；第三，利用公式：

，或者



来求得TF-IDF矩阵中每行每列的具体数值，其中TF是向量的每一个值标志对应的词语出现的次数的归一化后的频率，公式中分子表示某一行中某个单词出现的次数，分母表示该行中所有单词出现的次数。IDF是逆向文件频率，分子为总的行数或者

文章数，分母为出现了某个单词的行数或者文章数。

1. KNN最近邻算法，KNN邻近算法可以用来解决分类和回归问题。
   1. 分类问题上，第一，处理训练集，测试集和验证集数据部分，进行文本编码，可以直接处理成One-hot矩阵，即某行中的单词如果出现，就标为1，未出现，则标记为0；也可以使用TF-IDF矩阵。第二，选择某种距离衡量标准，计算验证集的特征矩阵的每一行与训练集的特征矩阵的每一行之间的距离，生成一个新的距离矩阵Dis, ,表示验证集的第j行与训练集的第i行之间的距离。第三，选择前k个最短距离所对应的标签，取标签出现次数最多的那一个标签作为最终的分类结果。
   2. 回归问题上，第一，二步与分类问题相同，第三，根据相似度进行加权，选取前k个最短的距离，并把该距离的倒数作为权重，从而计算测试集中的某个案例属于某个标签的概率，例：



然后计算相关系数，在本实验中，计算六个维度上相关系数的平均值后得到最终的相关系数值。

计算相关系数的公式如下：



1. 一些计算距离的方法
   1. 距离计算公式：（表示距离）



当p=1的时候，表示曼哈顿距离，当p=2的时候，表示欧式距离。

* 1. 余弦相似度

，表示文本特征向量。值越大，表示两个文本之间的差异越小。

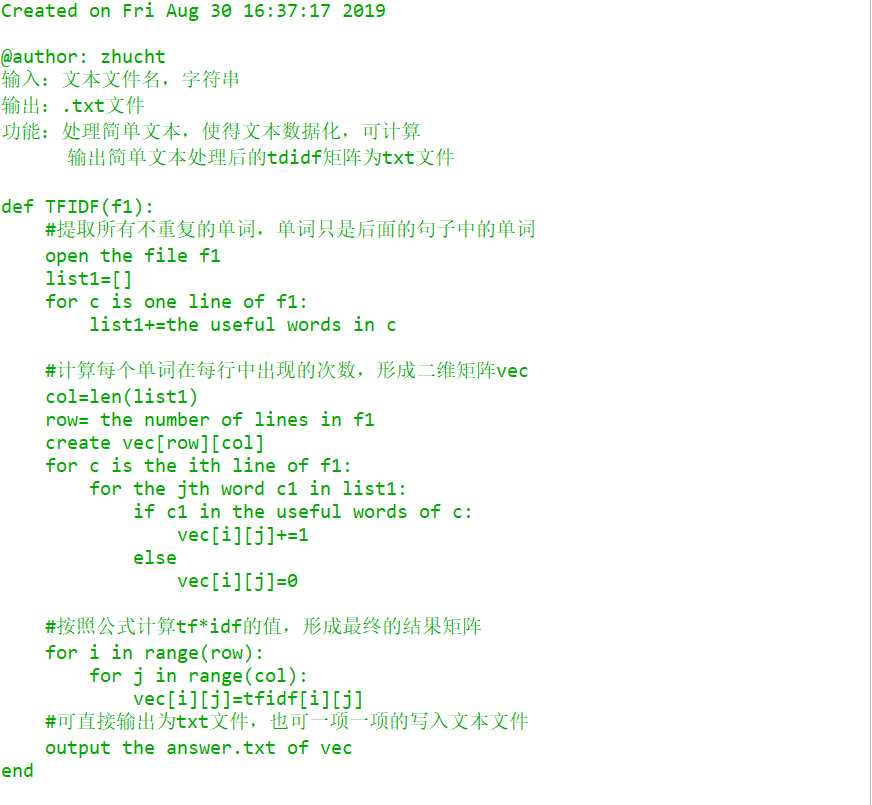
余弦距离：

恒大于等于0。值越小，两个文本之间的差异越小。

三：伪代码或流程图

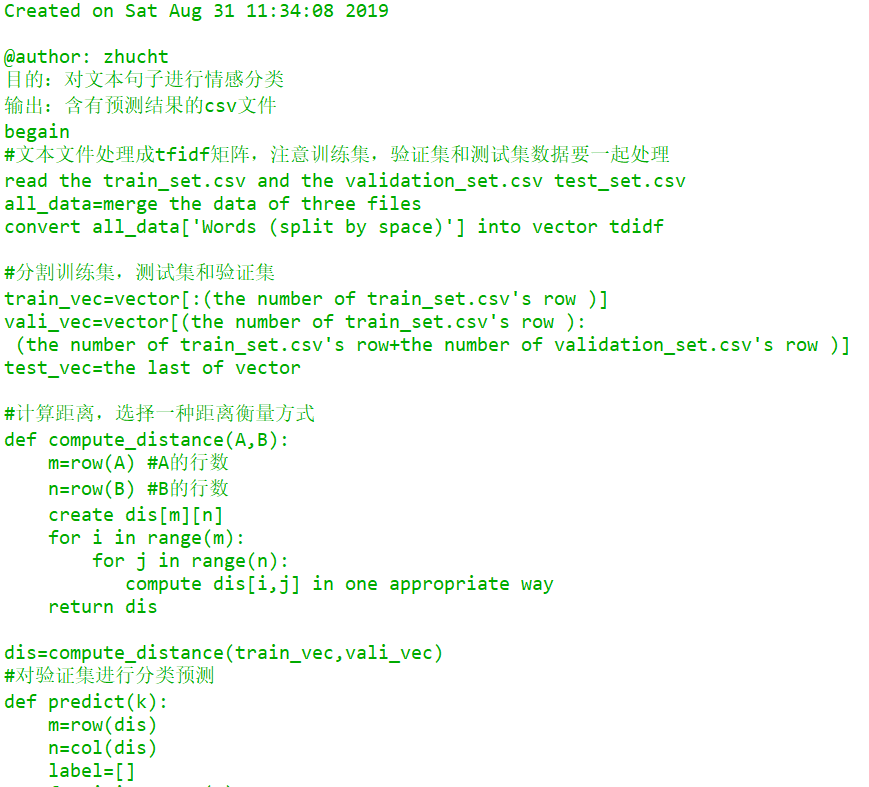
1. 简单文本转TF-IDF矩阵

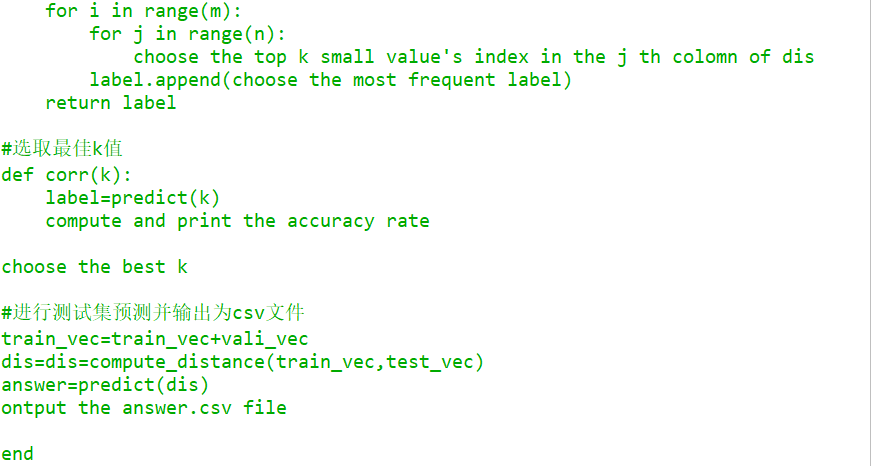
伪代码如下：



1. KNN算法分类

伪代码如下：

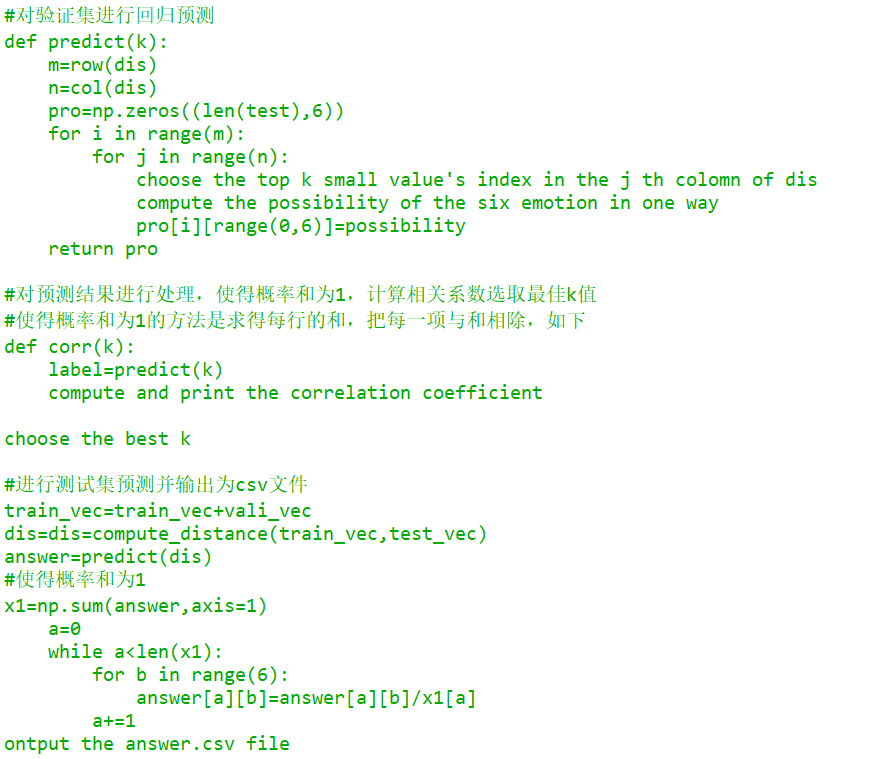




1. KNN算法回归

由于算法对数据得处理部分以及距离衡量计算是一样的，只需考率预测时候得不同之处即可。

伪代码如下：





三：代码展示

1. 普通简单文本转TF-IDF矩阵

公式选择：

，

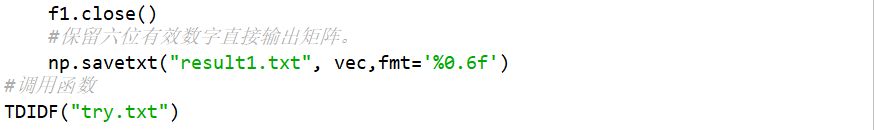




代码截图：







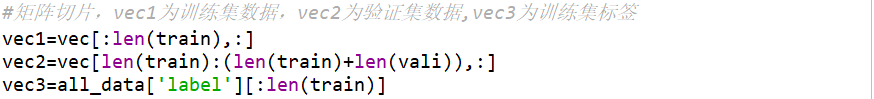
1. KNN分类问题

数据处理采用TFIDF矩阵。

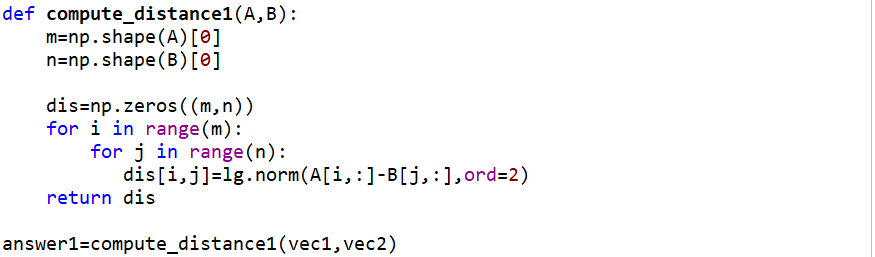
选用距离衡量公式：欧式距离

主要代码截图：

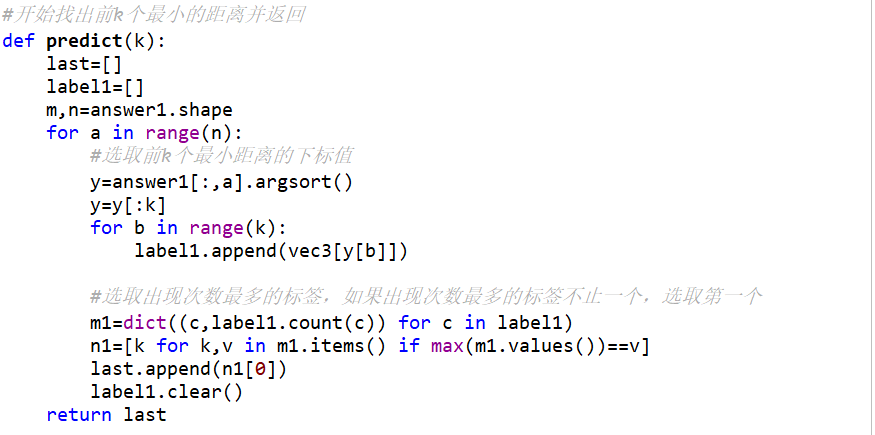
开始的数据预处理以及建立TFIDF矩阵，方法与上文相同，不再赘述。建立好的TFIDF矩阵为vec。然后进行矩阵切片，划分训练集，验证集和测试集如下。



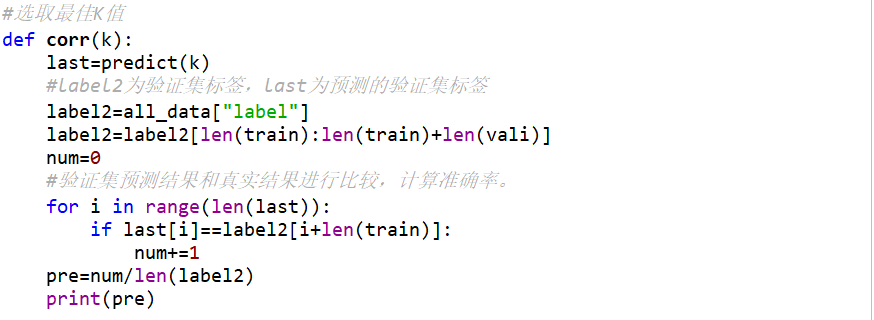
计算距离函数，输入：矩阵A,B要求A,B的列数相同。输出：距离矩阵dis，功能：计算B中矩阵的每一行特征向量与A矩阵每一行特征向量之间的距离。利用欧式距离的计算公式。



分类预测函数，输入：k值，输出：预测标签列表。功能：对验证集或测试集数据进行预测标签分类。取前k个最小距离对应的标签中出现次数最多的标签即可。

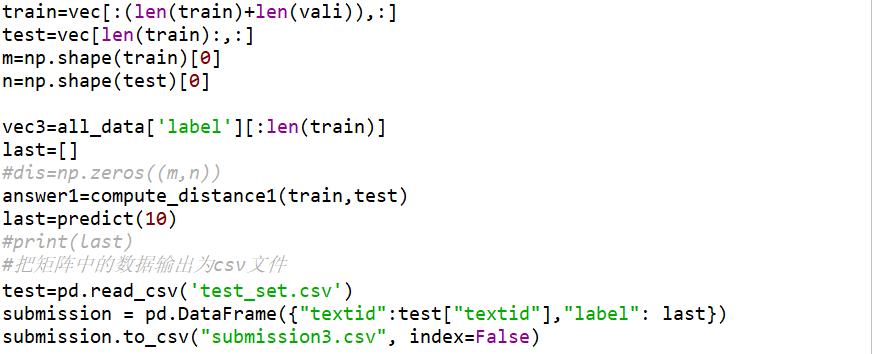


在建好的模型上利用验证集选取最佳k值，主要是利用一个小的循环来实现。



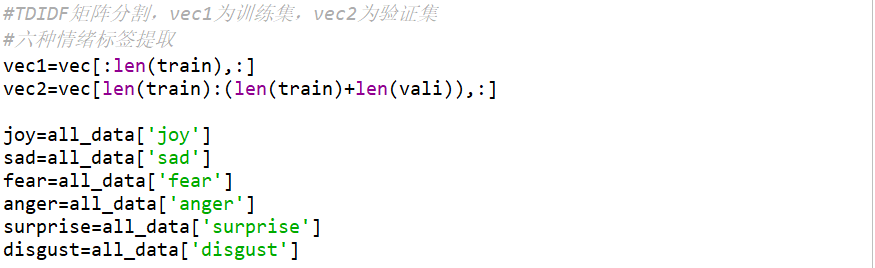


用选取得最佳k值进行测试集的结果预测。



1. KNN回归问题

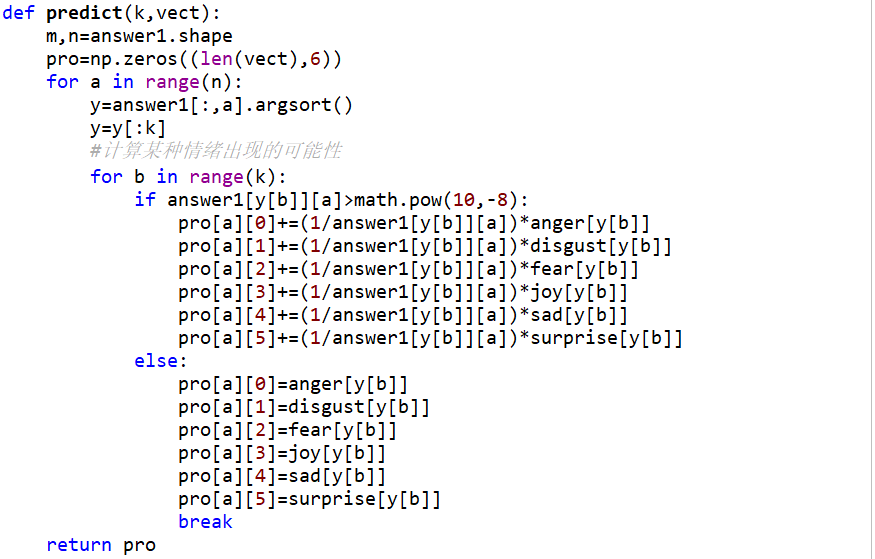
KNN回归问题数据的处理和距离的计算与KNN分类问题相同，这里直接展示不相同的预测部分。



回归预测函数，输入：k值，要预测的集合vect, 输出，预测的6种情绪出现可能性的矩阵。功能：进行每种情绪出现可能性的预测。采用的可能性计算公式原理为加权距离值进行计算，例：



如果，某个距离的值小于一定范围，直接将该测试案例与训练集的对应情绪可能性值相等，不必再进行计算，防止溢出问题的发生。



计算相关系数并选取最佳k值。

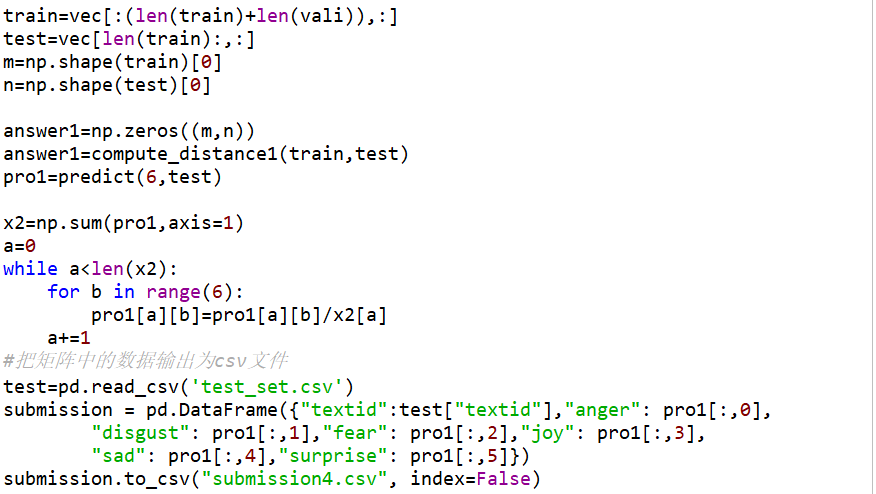
相关系数的计算公式如下：







最后，用得到的最佳k值进行测试集的结果预测，并把结果写入csv文件，代码如下：

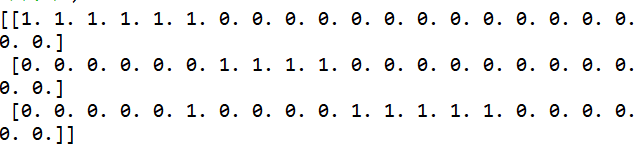


代码展示部分完毕！

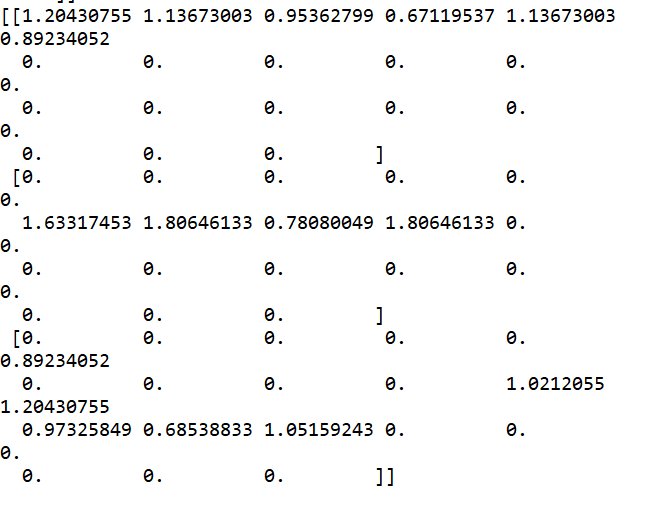
四：实验结果及分析

1. 结果展示和分析
2. TDIDF矩阵转换

部分文本每行部分单词出现次数的展示，可以看出是0，1出现的次数多一些 ，但是估计有>1的数字出现，只是比较少。而且可以看出单词是按照出现顺序排列的。我们以vec[:3,:21]为例：



对应的部分，计算得出得 TDIDF矩阵为：

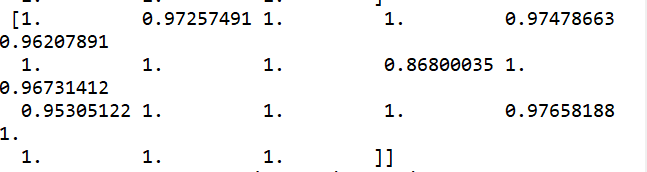


计算结果是有一定的合理性的。

1. KNN分类

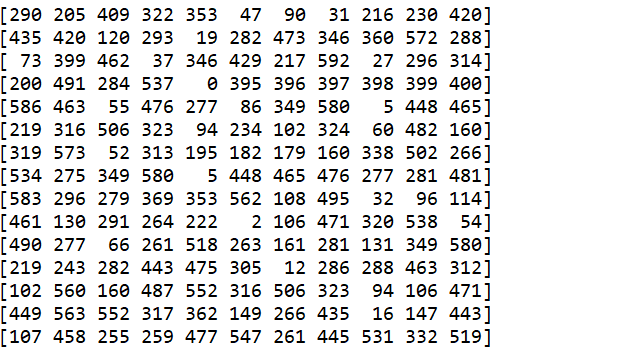
我们以验证集的部分数据为例展示一下数据结果：

下图为计算的余弦距离矩阵的一部分，可以看到大部分数值为1，说明其特征向量的余弦值大都为0，之间的相似度不高，和实际情况基本相同。

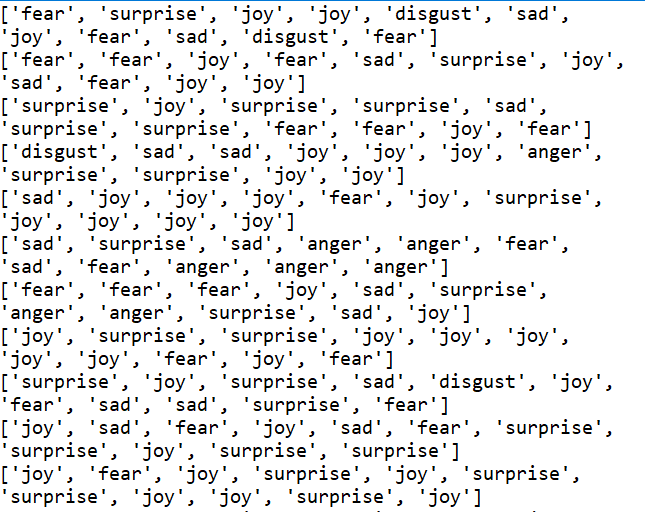


下面就预测过程进行分析，下图是验证集的前11个实例在k=11的时候的预测过程。

1. 找出前k个最小距离值对应的下标



1. 根据下标找出对应的label。



1. 找出每行中出现次数最多的label作为预测结果。



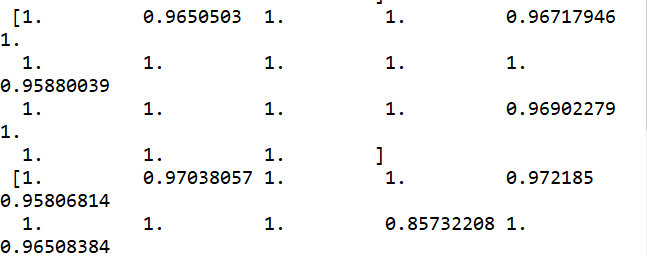
我们可以求得这11个验证集案例的预测准确率为4/11=0.364

实验结果有一定的可取性。

1. KNN回归

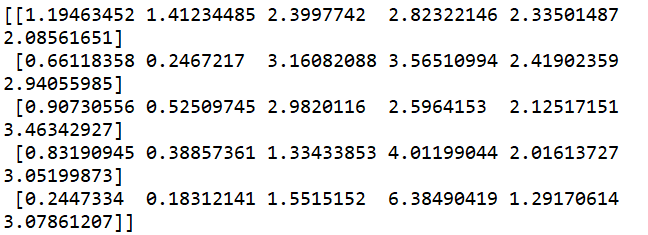
下图为验证集和训练集之间余弦距离矩阵的一部分。期望的是所有距离值都在

0-2之间，下图中的数据符合要求，有一定的合理性。

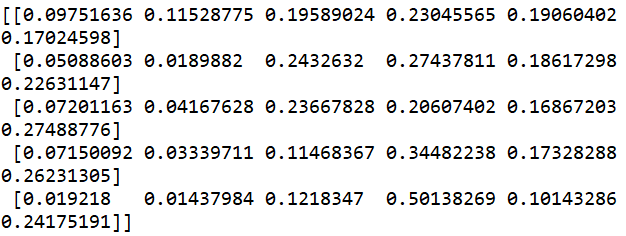


同样以训练集和验证集为例，展示部分预测过程中产生的数据情况。

第一：验证集前5个样例预测的六种情绪出现的相对可能值。期望的情况是六种情绪出现的可能性之和为1，下图的数据和期望偏差很大，需要进行处理，只能说下图中的数据应该是每种情绪出现的相对可能值。

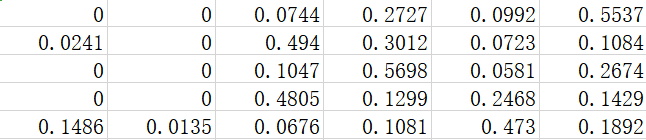


第二：经过处理之后，前五个样例，预测的六种情绪出现的可能值。处理过程是每行中的每个值和每行的和相除，得到新的可能值。



可以检测到，处理之后，每行的数据和为1.

和原数据相比较：



因为样例比较少，相关度很低，相关系数可参考性不大。

1. 模型性能展示和分析

要调整的参数主要有两部分，k值和距离衡量标准。K值主要取6-15中的值进行测试。计算距离的方法有：欧式距离，曼哈顿距离以及余弦距离。

KNN分类，表中数据为准确率

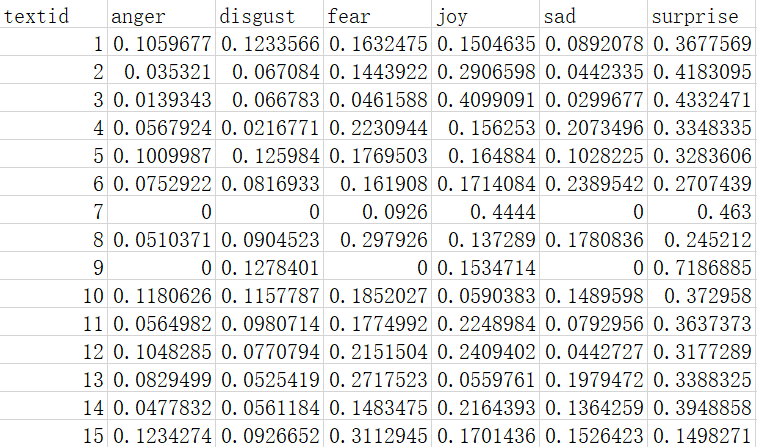
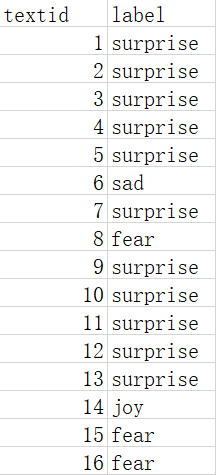
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | K=6 | K=7 | K=10 | K=11 | K=13 | K=14 | K=15 |
| 欧式距离 | 0.225 | 0.232 | 0.199 | 0.203 | 0.206 | 0.209 | 0.209 |
| 曼哈顿距离 | 0.408 | 0.428 | 0.408 | 0.408 | 0.415 | 0.424 | 0.405 |
| 余弦距离 | 0.428 | 0.450 | 0.450 | 0.441 | 0.424 | 0.424 | 0.402 |

KNN回归，表中数据为相关度

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | K=6 | K=7 | K=10 | K=11 | K=13 | K=14 | K=15 |
| 欧式距离 | 0.260 | 0.267 | 0.263 | 0.252 | 0.256 | 0.254 | 0.256 |
| 曼哈顿距离 | 0.316 | 0.320 | 0.337 | 0.333 | 0.334 | 0.334 | 0.329 |
| 余弦距离 | 0.400 | 0.396 | 0.399 | 0.391 | 0.377 | 0.369 | 0.369 |

经对比，可知采用余弦距离进行距离衡量效果是最好的，对于KNN分类，最佳k值为k=10，对于KNN回归，最佳k值为k=6,或者k=10。

1. 输出文件展示



五：思考题

1. 分母加了1，我觉得应该是为了防止分母为0，所给的文本文件可能会有数据缺失，如果不加1，分母可能就为0 了，运算就会出错吧。
2. IDF（逆向文件频率）数值，可以明显的看到，如果某个词在整个文本中出现的次数越少，IDF值就会越大，如果一个词在一个文本中的IDF值很低，而在其他文本中的值很高，就可以用该词来区别这个文本和其他文本。

TF-IDF（词频-逆文件频率）值，是某个单词出现的频率和该词的逆向文件频率的乘积，是衡量一个单词在一个文本中的重要程度的评估方法。

1. 各种情感的概率和为1，则用每种情绪的可能值和所有情绪可能值的和相除，得到新的可能值。这样，得到的结果相加和就为1。