**中山大学数据科学与计算机学院**

**移动信息工程专业-人工智能**

**本科生实验报告**

**（2016学年秋季学期）**

课程名称：Artificial Intelligence

|  |  |  |  |
| --- | --- | --- | --- |
| 教学班级 | **14M2** | 专业（方向） | **移动互联网** |
| 学号 | **14353205** | 姓名 | **刘万里** |

# 实验题目

文本数据集的读写和简单处理

1. 将数据集的数据表示成one-hot矩阵，TF矩阵，TF-IDF矩阵，并分别保存为“onehot”，“TF”，“TFIDF”三个文件 。

2、将数据集的one-hot矩阵表示成三元组矩阵，保存为“smatrix”文件。

# 实验内容

1. 算法原理

因为需要去重且不排序，C++中并没有这样的数据结构来存储，因此按照数据集中的字符串出现的顺序，用一个map<string,int>去存储，int值存储这个字符串第一次出现的下标，再维护一个map<int,int>的RepeatTimes,first为出现的下标，second为在数据集中出现的次数，再维护一个map<int,int>的AppearLines，first存储下标,second存储的是对应下标的字符串出现过的行数，读入文件过程中维护整型变量数据的行数Lines,不同的字符串的个数Index,这样就可以得到一个二维矩阵。并且每次都用一个结构体Three\_Var存储一个字符串对应的行数i，列数j，以及出现过的行数v，并放入Three\_var类型的vector容器中。

OneHot: 遍历的是一个矩阵，如果行数和列数和当前vector中的Three\_Var中的i,j分别相等，则写入1，否则写入0；

TF:对OneHot进行归一化即可。因此需要在读取文件的过程中用一个名为NumsOfLines

的vector记录每一行的单词的个数。

TF-IDF:按照公式计算即可，因为我们在读取文件时维护了每个下标对应的字符串出现

过的行数AppearLines，所以只需计算即可。

Smatrix1:通过Lines和Index遍历矩阵，如果行数和列数和当前vector中的Three\_Var中的i,j分别相等，则得到一个三元组，写入smatrix1.

遍历完每一行进行换行即可得到对应的矩阵。

1. 伪代码

/\*-----------读入数据------------------\*/

While（这一行有数据）

{

如果是一句话的字符串（不是表情，数字之类的）

存储下来，并依次在FirstIndex保存下标

维护好字符串的下标索引Index

维护好RepeatTimes，记录每个字符串重复的次数

记录好这行的字符串数据的个数（即维护好temp）

}

For(遍历RepeatTimes)

{

维护AppearLines;

得到三元组Three\_var，并放入vector中；

}

/\*--------------读取完一行----------------------------\*/

行数Lines++;

维护NumsOfLines来记录每一行出现的字符串个数；

/\*----------------------------读取完成----------------------------\*/

/\*------------------------处理数据并得到结果输出到文档-------------------\*/

pos = 0;//三元组vector的开始下标

For(i = 0；遍历Lines)

{

For(j=0;遍历出现过的字符串Index)

{

If(当前行数i ，下标分别与当前vector三元组的行数，下标相等)//匹配

{

得到结果并输出到对应文档；

}

Else //不匹配

{

三个文档写入0；

}

}

/\*-------------遍历了一行---------------\*/

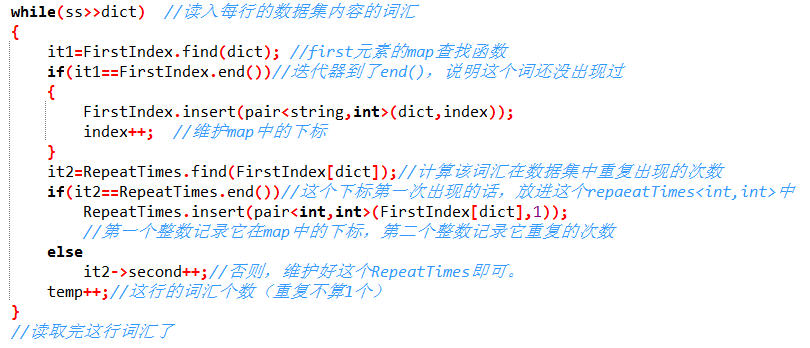
3个文档读入换行符；

}

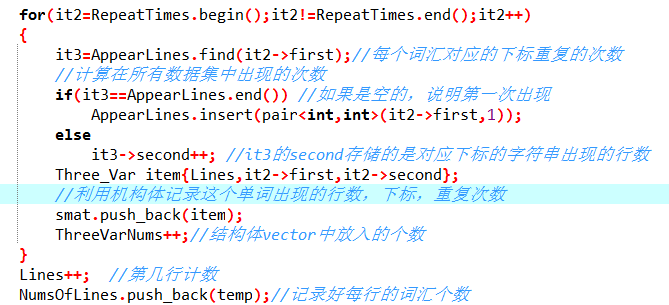
1. 关键代码截图（带注释）

读取文件：

读取一行数据并进行相应的数据结构的维护



读取完一行之后维护好AppearLines这个map，并且维护变量Lines,vector容器 NumsOfLines



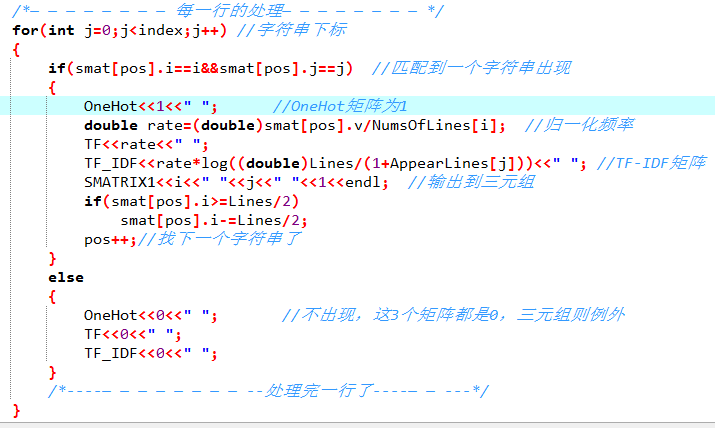
现在进行对数据的处理和计算，输出：

遍历整个矩阵，先遍历每一行：



进行每一行的处理：

进行当前遍历到的这行的这个字符串，查看与三元变量Three\_var的行和列是否对应，对应说明存在，不对应说明不存在这个字符串，OneHot即为0，其余同理。三元组输出时输出的不是出现过的行数，是1即可。



处理完一行就换行：



关闭文件流，完成文件写入操作。

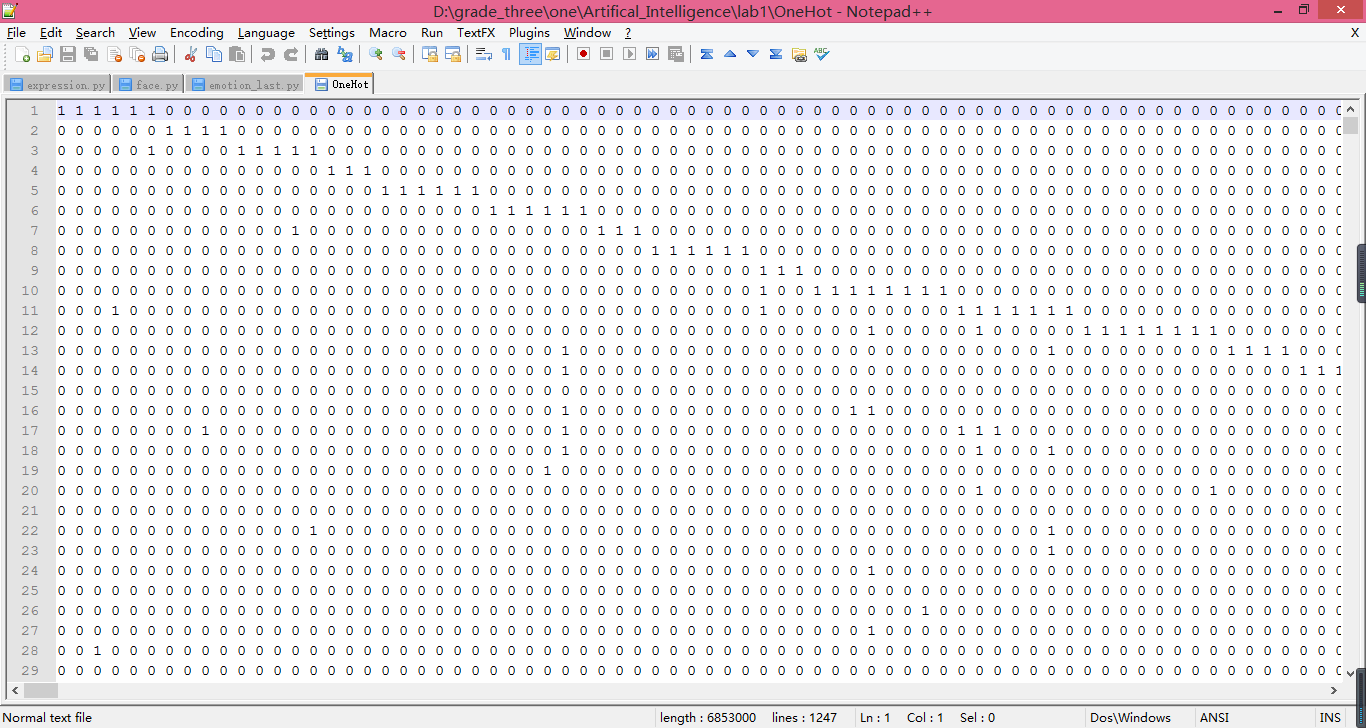
1. 创新点&优化

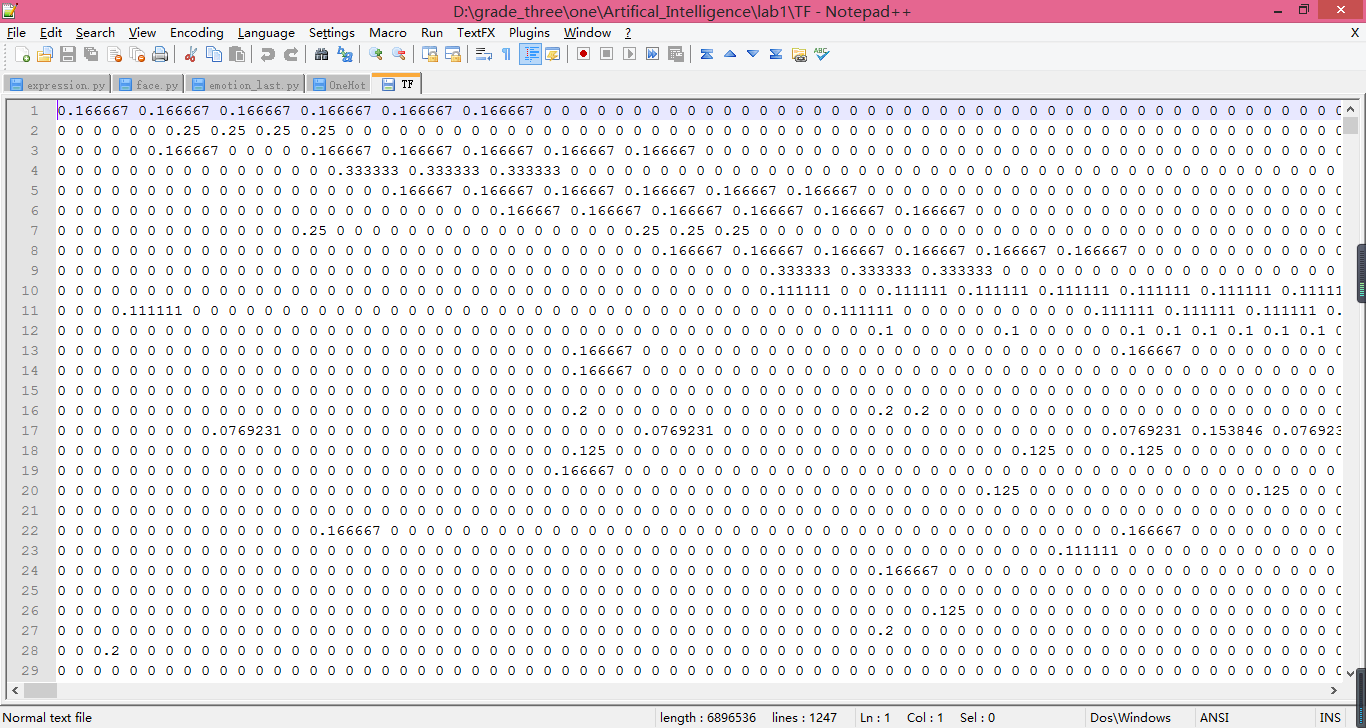
无

# 实验结果及分析

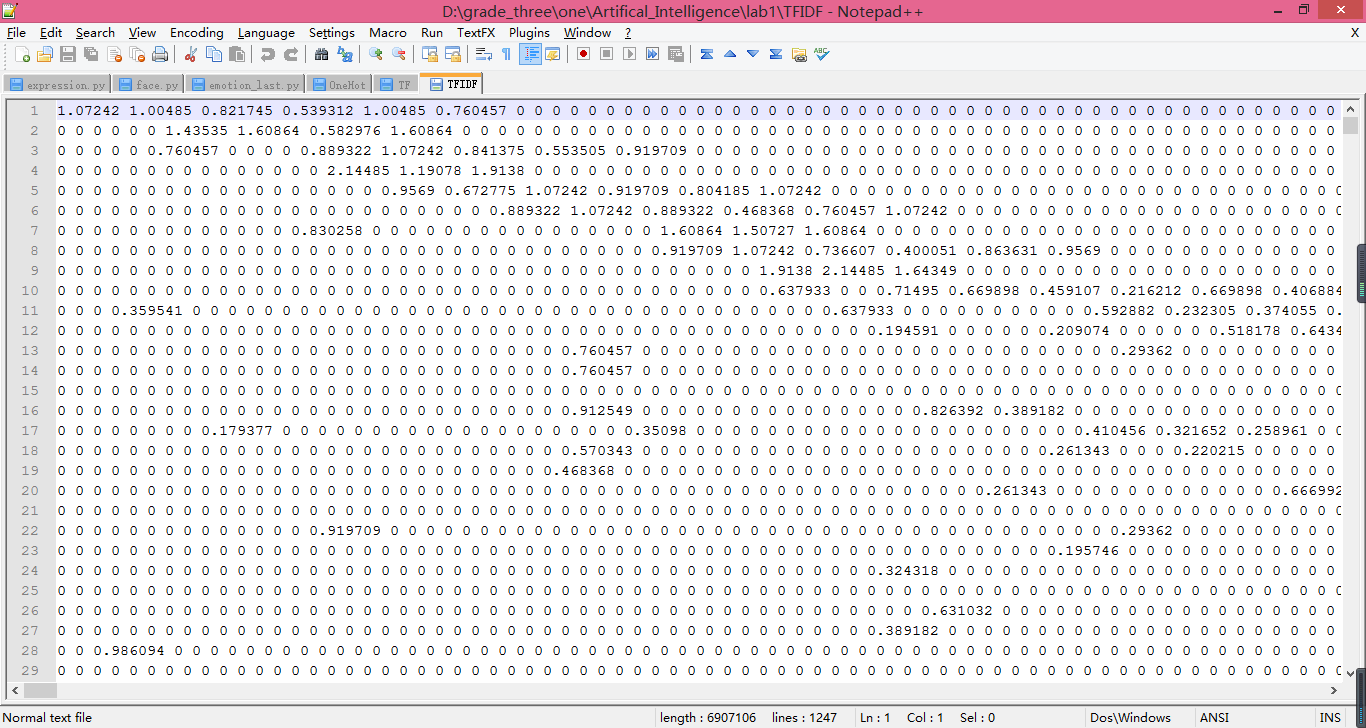
1. 实验结果展示示例（可图可表可文字，尽量可视化）

OneHot:



TF:

TF-IDF:



Smatrix1:

