211275026-陈畅-实验二

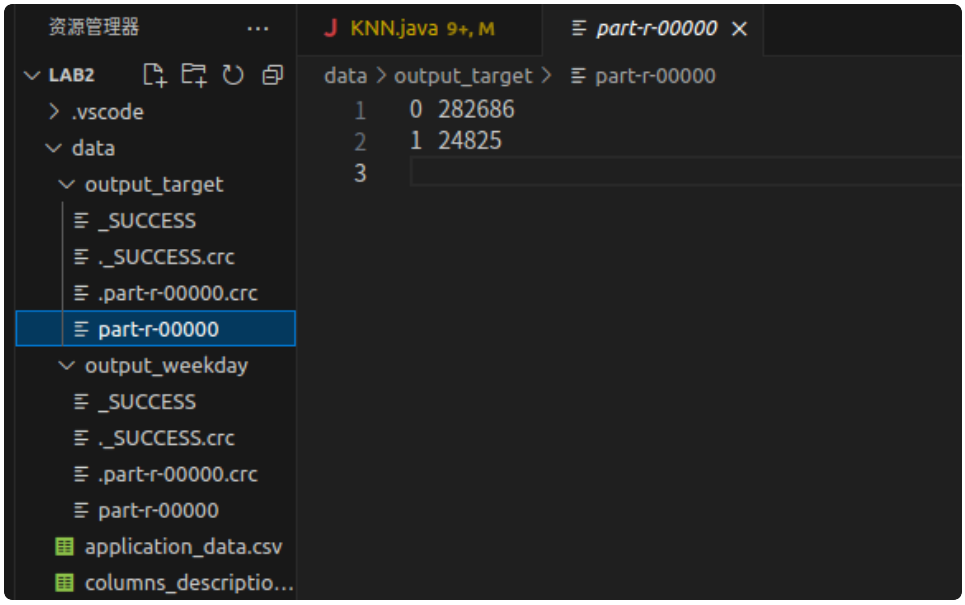
代码仓库链接：https://github.com/Cc17952/FBDP/tree/main/lab2

任务一

设计思路：

- 在map阶段，读取TARGET值，<key,value>=<TARGET,1>;
- 在reduce阶段，对value进行求和sum，返回<key,value>=<TARGET,sum>。

结果展示：

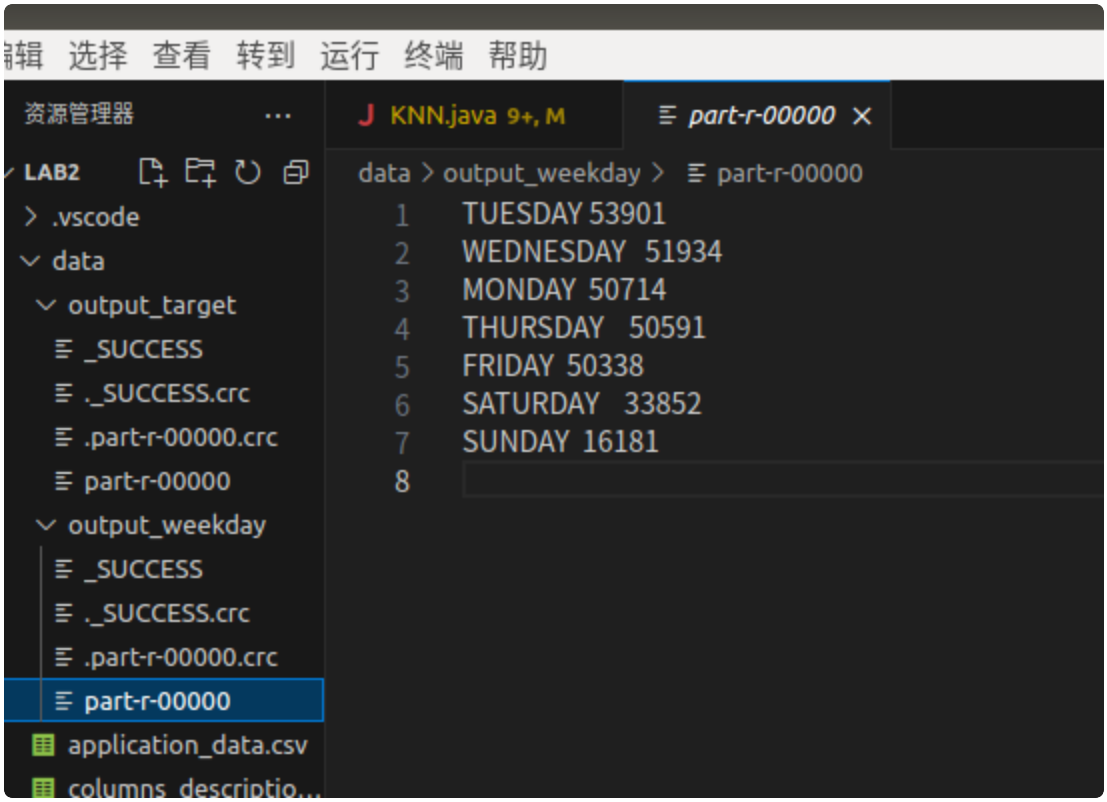


任务二

设计思路：

- map阶段，读取每个星期的值，<key,value>=<Weekday,1>;
- reduce阶段，对value进行求和sum，并且通过map排序，map<string,int>=<Weekday,sum>。

结果展示：



任务三

设计思路：

仅选取部分样本特征，并通过python进行了归一化处理。

```
# ['FLAG_CONT_MOBILE','AMT_INCOME_TOTAL','AMT_CREDIT','FLAG_OWN_CAR','FLAG_OWN_REALTY','REGION_RATING_CLIENT','OBS_30_CNT_SOCIAL_CIRCLE']
```

- map阶段，读取训练集，计算测试样本和训练样本之间的相似度。<key,value>=<index,相似度>
  - 这里的相似度由欧氏距离算得

```
while (itr.hasMoreTokens()) {
    String[] tmp = itr.nextToken().split(",");
    String label = tmp[8];
    List data = new ArrayList();
    for (int i = 1;i<=7;i++){
      data.add(Double.parseDouble(tmp[i]));
    }
    for (int i = 0;i<test.size();i++){
      List tmp2 = (List) test.get(i);
      // 每个测试数据和训练数据的距离(这里使用欧氏距离)
      double dis = 0;
      for (int j=1;j<8;j++){
        dis += Math.pow( (double)tmp2.get(j)-(double)data.get(j),2);
      }
      dis = Math.sqrt(dis);
      // out 为类标签,距离
      String out = label + "," + String.valueOf(dis);
      context.write(new IntWritable(i), new Text(out));
    }
  }
}
```

- reduce阶段，对举例进行排序，选取前K个近邻的作为违规类。
  - $recall_k = \frac{TP}{TP+FN}$
  - $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
  - $f1_k = \frac{2 \cdot precision_k \cdot recall_k}{precision_k + recall_k}$

错误排查

错误1：没有进入reduce阶段

由于代码原因，未进入hadoop的reduce函数，没有完成该任务。

- 猜想原因：map和reduce过程中间context相关内容出现问题
- 尝试设置Job有关map的输出格式，失败。
- 经过多次尝试，发现是对list的循环导致问题，时间不够进行修改，故先将代码提交，后续会再debug。

```
54          data.add(Double.parseDouble(tmp[i]));
55        }
56        // System.out.println(test.size());
57        for (int i = 0; i < test.size(); i++) {
58          List<?> tmp2 = (List<?>) test.get(i);
59
60          // 每个测试数据和训练数据的距离(这里使用欧氏距离)
61          double dis = 0;
62          // String id = "";
63          for (int j = 1; j < 8; j++) {
64            dis += Math.pow((double) tmp2.get(j) - (double) data.get(j), 2);
65          }
66          dis = Math.sqrt(dis);
67
68          // out 为类标签,距离
69          String out = label + "," + String.valueOf(dis);
70          // String id  = tmp2.get(0);
71          System.out.println(i);
72          System.out.println((out));
73          context.write(new IntWritable(i), new Text(out));
74        }
75
```

```
问题 19   输出   调试控制台   终端   端口

hadoop@ubuntu:~/FBDP/lab2$  cd /home/hadoop/FBDP/la
75isigobez87keaejdgxzmmx0.jar com.example.KNN
log4j:WARN No appenders could be found for logger (
log4j:WARN Please initialize the log4j system prope
log4j:WARN See http://logging.apache.org/log4j/1.2/
0
```

仅输出i=0时候的情况，实际test.size()的值为61298

```
log4j:WARN S
61298
0
hadoop@ubunt
```

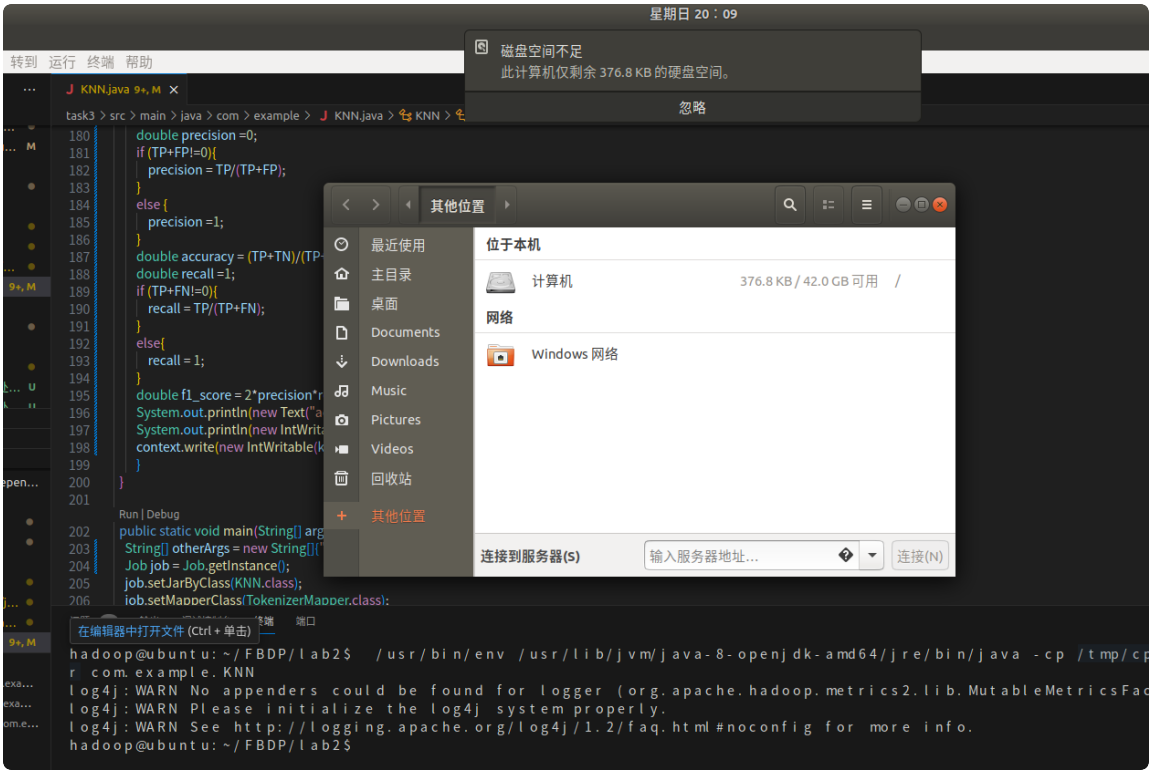
错误2：进入了reduce阶段，但output文件无结果

多写了一行

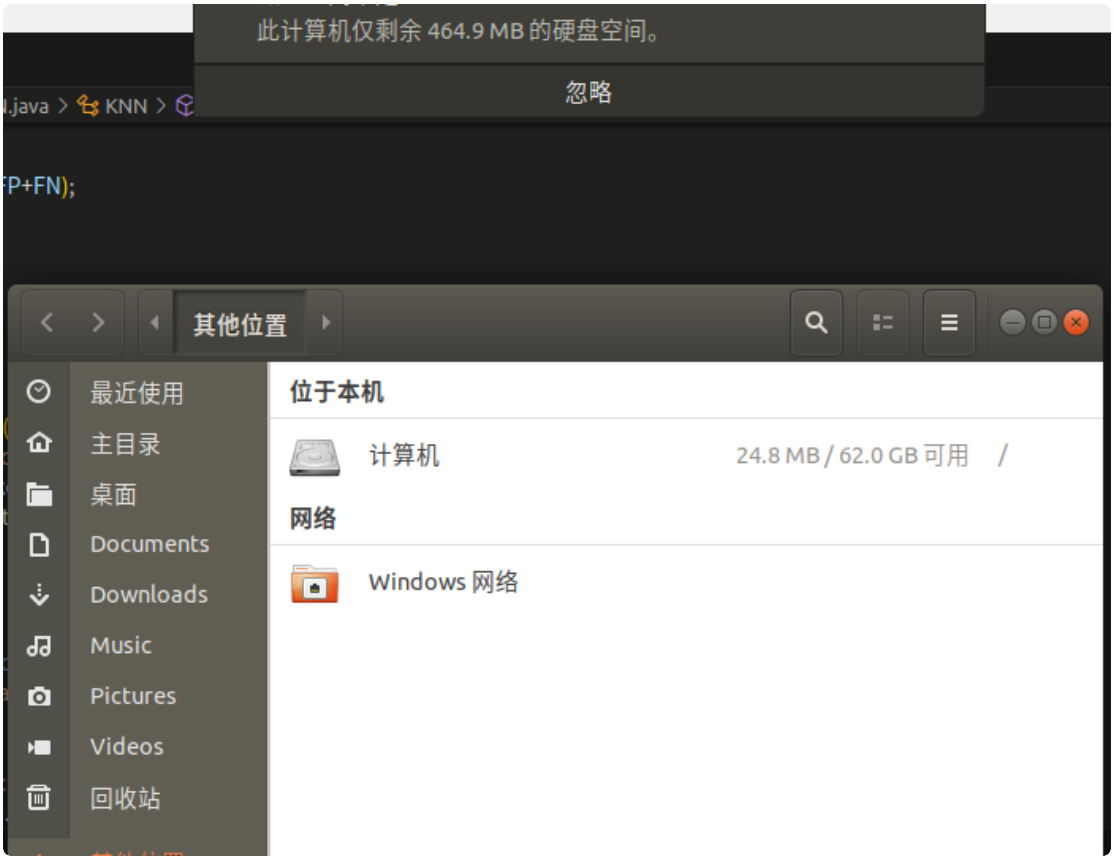
```
job.setMapOutputValueClass(Text.class);
// job.setCombinerClass(TokenizerReducer.class);
job.setReducerClass(TokenizerReducer.class);
```

，导致程序正常运行但没有输出。

错误3：内存不够



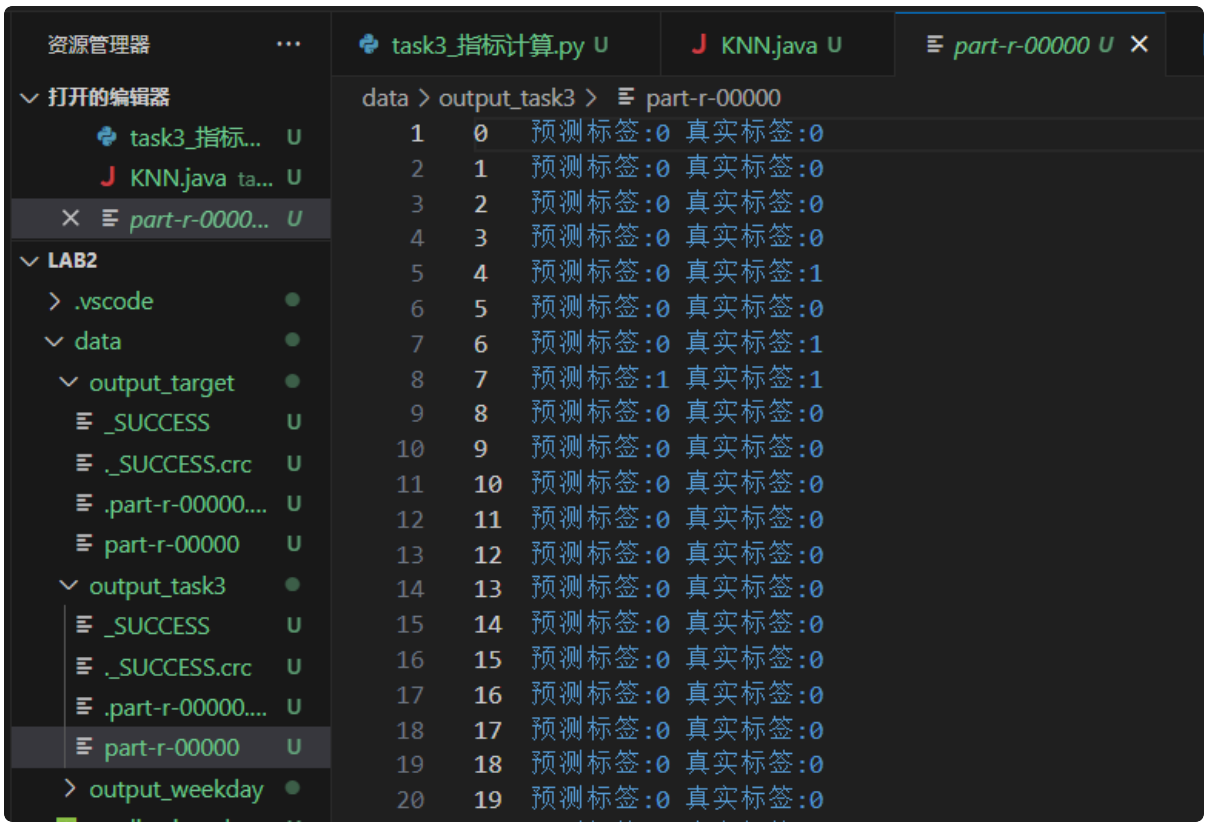
进行单机的hadoop尝试，发现内存不够



扩充之后仍然不足，程序自行停止

最终删减了部分数据集，仅保留了3w条，其中训练集和测试集的比例为8：2。

结果展示：




```
# 以违约作为本类的相关指标如下：
print("precision:"+str(precision))
print("accuracy:"+str(accuracy))
print("recall:"+str(recall))
print("f1_score:"+str(f1_score))
```

问题 输出 调试控制台 终端 端口 Code

```
precision:0.004089979550102225
accuracy:0.9181666666666667
recall:0.3333333333333333
f1_score:0.008080808080808
[Done] exited with code=0 in 0.514 seconds
```