## EDI shell脚本职责清单

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | Shell | 职责 | 被调用 | 备注 |
| 1 | /opt/running/edi/edi\_new\_in\_hive.sh | 1. 从HadoopMySQL获取最新的评论数据到HadoopNamenode； 2. 备份HadoopMySQL的该数据； 3. 将获取到的数据load到hive原始数据表   EDI\_N\_PROD\_COMMS；   1. 从原始数据表提取出需要的数据到正式表   EDI\_M\_PROD\_COMMS。 | 5 | 需要注意：确保在备份 |
| 2 | /opt/running/edi/edi\_hive\_do\_op.sh | 1. 根据配置找出最后一次计算的分区号； 2. 对大于上面得到分区号的分区记录做情感分析计算，存储得到的结果； 3. 更新配置最后计算分区。 | 5 |  |
| 3 | /opt/running/edi/edi\_hive\_do\_count\_op.sh | 按品牌型号统计情感分析计算结果写入hive统计表。 | 5 |  |
| 4 | /opt/running/edi/edi\_hive\_to\_mysql.sh | 1. 按需要同步的hive表分别读取该表最后一次同步的分区号； 2. 将分区号大于1取得的分区号的该表记录，按分区循环调用sqoop导出到mysql； 3. 更新该表最后一次同步分区编号 | 5 | 1. 两层循环。 2. 需要mysql有T\_xx的临时表，用以解决sqoop不能同步分区字段的问题。 |
| 5 | /opt/running/edi/edi\_hive\_process.sh | 引用1-4，完成整个过程 |  |  |
| 6 | /opt/running/edi/nkw/  edi\_nkw\_rewrite\_comms.sh | 1. 查询评论信息和评论对象的位置列表，导出到文件。 2. 将1导出的文件分词、切片、混淆成10个基本等大的block，输出到指定路径。 | 8 |  |
| 7 | /opt/running/edi/nkw/  edi\_nkw\_train\_mapper.sh | 1. 是对象提取时map/reduce中的，mapper处理，按测试集名，合并其他9块做训练集； 2. 用训练集做crf训练，输出结果； 3. 用训练结果和测试集做crf测试，得出结果； 4. 上传结果到hdfs. | 8 | 各节点需要安装crf++  crf训练：crf\_learn -f 3 -c 1.5 template\_file train\_file model\_file  crf测试：crf\_test -m model\_file test\_files |
| 8 | /opt/running/edi/nkw/  edi\_nkw\_do\_train.sh | 1. 调用6，得到待训练数据； 2. 发起map/reduce分发任务7给多个节点并行训练； 3. 从hdfs获取训练结果汇总，去重； 4. 更新词典。 |  | 完成整个训练过程  汇总去重:  com.zhongyitech.edi. NLP.test.ToNewWords |