前期工作

1、将paper2017\_2\_24，paper2017\_2\_25，paper2017\_3\_19（老司机爬的） 三张表插入后去除重复得到临时表paper\_total\_set

2、将paper\_total\_set中 classification\_str字段转为classification，得到和paper表相同的结构

3、将thesis（刘总爬的）转为和paper表一样的临时表paper\_thesis

4、从paper\_total\_set中提取出专家信息得到临时表expert\_from\_paper，从paper\_thesis表中提取出专家信息得到临时表expert\_from\_thesis

5、将paper复制出一张paper\_copy表,将paper\_copy,paper\_thesis,paper\_total\_set整合到一起并去重得到paper\_copy\_new\_set表

6、用unit\_areacode\_map过滤paper表的单位

7、将expert和expert\_from\_thesis表整合去重得到expert\_new

8、根据expert\_new和paper\_copy\_new\_set关联得到expert\_paper\_join\_new

后期paper表的单位处理

1、从新的专利表中抽出新单位去掉人民和多个单位得到unit\_areacode\_map\_new

2、从网上拿到高校信息，研究所信息和各个省份对应的area\_code信息导入数据库

3、将paper表，patent表，project表的单位从unit\_areacode\_map\_new，高校和研究所表中过滤一遍，再按照学校，研究所，企业等关键字匹配得到单位类型，根据省份关键字匹配得到单位所在省份和对应的代码，这些信息不直接写入相应的paper等表中，而是存储在catalog\_unit\_paper\_join\_new，catalog\_unit\_patent\_join\_new，catalog\_unit\_project\_join\_new表中

将来还需做的工作

1、将数据库信息导入HBase或HDFS中

2、对新的paper表（即paper\_copy\_new\_set）单位处理，之前都是对paper表进行单位处理

3、在明确处理怎么处理单位之后，重新抽取专家表

4、对新的paper表和新的专家表在单位明确处理之后进行关联