第一步： 初始化RDD，将有类别信息的数据转化为labeledRDD，无类别信息的数据转化为unlabeledRDD。

第二步：将有类别的数据（labeledRDD）进行广播。

第三步：对无类别信息的数据（unlabeledRDD）执行mapPartation操作，在每个分区中执行如下操作：1.用有类别的数据集训练一个分类器。2.使用训练好分类器去计算无类别数据的信息熵。

第四步：根据熵值对无类别信息的数据排序，选择K个熵值最大的样例作为本次迭代选择的样例，输出到HDFS。

第五步：对第四步中选择的样例进行人工标注。（模拟进行，因为数据本身带有类别信息），并转化为RDD。

第六步：将有类别数据（labeledRDD）与第五步得到的RDD进行合并（union操作），将合并后的RDD作为更新后有类别数据（labeledRDD）。

迭代执行1-6步，输出最后一次迭代得到的有类别数据（labeledRDD）至HDFS。