首先训练一个模型(Python或Java) （*贝叶斯分类为例）*

Map阶段：

1. 对每一个value进行分词

统计词频 （1\_word， one）

每日词频 (4\_datetime\_word, one)

1. 分词后进行预测 （2\_type\_datetime, likecount）

统计类别数 (3\_type, one)

Reduce阶段：

mos = MultipleOutputs(context).write()

1. prefix == 1

mos.write(word, result,"wordCount");

1. prefix == 2

mos.write(word, result, "predict/"+contents[2]+"\_predict");

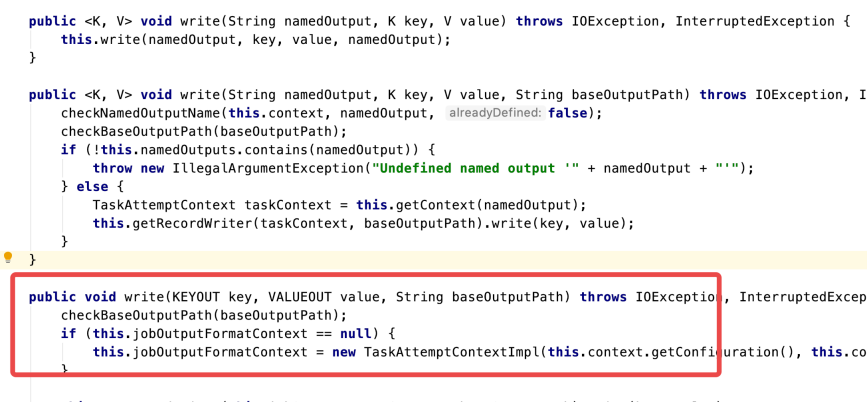
1. prefix == 3

mos.write(word, result, "typeCount");

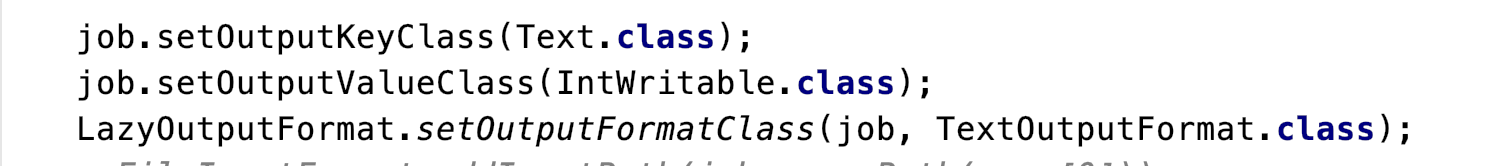
1. prefix == 4

mos.write(word, result, "wordCount/"+contents[2]+"\_wordCount");

Tips（踩坑）:



1、第三种write 可以避免声明MultipleOutputs.addNamedOutput



2、声明LazyOutput使不生成part-XXXXXXX

3、job.getConfiguration().setLong("mapred.max.split.size", 1024\*1024\*10L);

分片太大，无法执行。