20190924：

实验：

YARA规则继续编写(在Linux环境)

IDEA上决策树识别恶意文件实验：找到单个文件最大行数和单个文件最大长度

* test.txt文件中找129长度，一次性最多能训练352行(条)数据，良性数据177行，恶意数据175行，保存在test\_malicious.txt中。test\_malicious.txt这个文件是用来训练模型的，通过该文件的数据进行训练，训练完毕之后，通过代码第673行的YZFile对象调用yanzheng.txt内的每一行数据，得出每一行的预测结果。

计划：用决策树的预测算法(信息增益&信息增益率)，训练5组(test\_malicious1.txt，test\_malicious2.txt，test\_malicious3.txt，test\_malicious4.txt，test\_malicious5.txt)，然后分别去预测同样的测试集(yanzheng.txt)，测试集的数目：良1500条，恶1500条。

文件夹-->“E:\yansan\20190827--20191015 ZhengLiQianSanGeYueDeGongZuo\论文初稿\ML数据”

origin 有这个单词的数据文件是原始数据文件，数量最多

nomarked 未标记数据，处理成格式符合ML计算，可以作为预测数据集，组成yanzheng.txt

marked 标记过的数据，可以用来训练成预测模型，组成test\_malicious1.txt、test\_malicious2.txt、test\_malicious3.txt

20190926：进行实验

(TP真正率🡪良预测为良，FP假正率🡪恶预测为良，

TN真负率🡪良预测为恶，FN假负率🡪恶预测为恶)

calculate.java文件训练

* test\_malicious1.txt(170恶,173良)，测试yanzheng.txt(600恶，482良),得到结果：

FP=21/600 FN=120/600 预测恶意为无效率=459/600

TP=200/482 TN=19/482 预测良性为无效率=263/482

* test\_malicious2.txt(116恶,210良)，测试yanzheng.txt(600恶，483良),得到结果：

FP=36/600 FN=107/600 预测恶意为无效率=457/600

TP=252/482 TN=23/482 预测良性为无效率=207/482

* test\_malicious3.txt(350恶,130良)，测试yanzheng.txt(600恶，483良),得到结果：

FP=12/600 FN=149/600 预测恶意为无效率=439/600

TP=167/482 TN=19/482 预测良性为无效率=296/482