分类算法：贝叶斯。

特征提取流程：

1 根据姓氏辞典把姓氏去掉，留下不带姓氏的名字；

2 特征有三个维度，分别用X1,X2,X3(=X1X2)表示。

如果是单字名，则X1为空格，X2为单字名，X3就是前两者拼接X1X2，

比如郭靖，X1=” ”,X2=”靖”，X3=” 靖”。

如果是双字名，则X1为第一个字，X2为第二个字，X3由前两者拼接，

比如黄药师，则X1=”药”，X2=”师”，X3=”药师”

如果是三个字以上，则只保留最后两个字，当做双字名处理。

一次训练和测试的结果分析：

训练集：300w

男性数量:1421086

女性数量:1578855

单字名数量:53202

男性单字名数量:29204

女性单字名数量:23998

双字名数量:2946739

男性双字名数量:1391882

女性双字名数量:1554857

测试集：100w

|  |  |  |
| --- | --- | --- |
|  | Pre\_male | Pre\_female |
| male | 473695 | 0 |
| female | 5195 | 521092 |

正确分类数量：994787

错误分类数量：5213

准确率：99.5%

PS:可能你会发现，对于错误分类，5213不等于5195+0，因为除了男预测为女，女预测为男，还存在由于数据异常导致无法判断分类是否正确，比如标注数据有误，label是male和female之外的数据，还有比如姓名存在问题，比如都是异常符号。

数据集问题：

1 缺少3个字以上的名字（不带姓氏）的数据。不过实际上3个字以上的名字毕竟是少数，不算大问题。

参考资料：

《基于中文人名用字特征的性别判定方法》