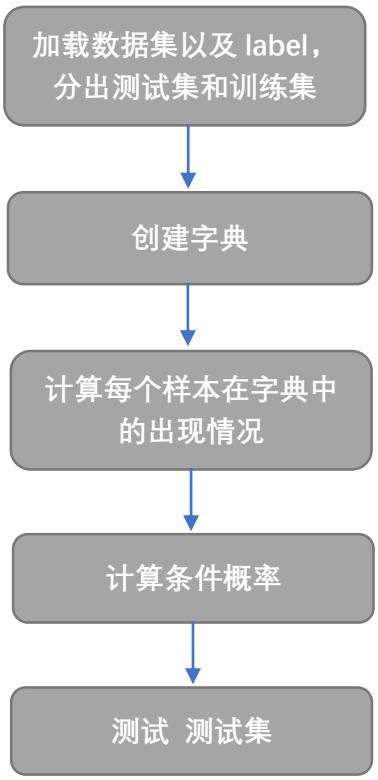
Naïve Bayes

一、Naïve Bayes

朴素贝叶斯是一种构建分类器的简单方法。该分类器模型会给问题实例分配用特征值表示的**类标签**，类标签取自有限集合。**朴素**贝叶斯分类器都假定样本每个特征与其它特征都**不相关**。

二、Program Flow



三、问题

1. 字典巨大

最初跑程序的时候，将字典建的太大，导致后来存放字典 word 频率的矩阵太大，8G 内存根本不够用。后来将字典做了筛选（将在一个文件中出现仅仅一次的 word 丢掉），减小了字典的大小。

2. Label 的存放位置

由于要比较 label 用于比较正确率，所以要将每个文档的 label 存起来，在测试结果出来之后将其 label 取出来，比较实际 label 和预测 label 是否一致。想了一个办法，将字典的第一列存放 label，但是统计词频时又不使用这一列。

四、实验结果

15062 个训练集，3765 个测试集。
3550 个预测**正确**的测试集，**216** 个预测**错误**的测试集。
正确率为 **94.26%**。

```
3550.0 216.0
correct rate:  0.9426
```