# 步骤流程

## 1数据分析

数据采用的是smp2017的中文文本分类任务的数据。一共有31类，文本为句子级别，篇幅较短，数据的分布相对均匀，某些类别的句子特征比较明显。

## 2模型解析

本次采用的是k临近算法，具体实现中调用了scikit-learn中的MultinomialNB，参数调整用的是五折交叉验证。

## 3参数设置

MultinomialNB相对于GaussianNB更适合用于NLP。

分别探索了alpha，fit\_prior 两个参数。

输出报告中看出有较多分类（31类）时，选择较小的平滑度alpha和fit\_prior为True的时候效果最佳。