# 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KNeighborsClassifier，参数调整用的是五折交叉验证。

## 3参数设置

分别探索了权重参数weights（uniform, distance），闵可夫斯基度量的次方参数p，邻居数n\_neighbors，kd-tree的叶子数量leaf\_size。

输出报告中看出有较多分类（31类）时，选择较小的n\_neighbors值和较小的leaf\_size的模型效果最佳。