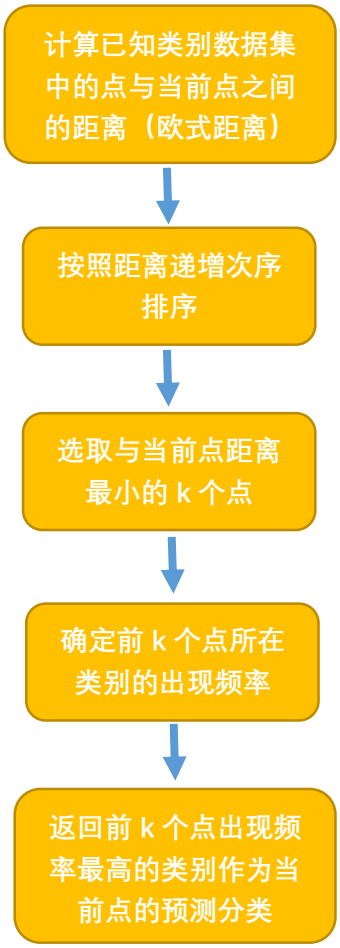
一、KNN

KNN 算法是一种分类器，通过计算当前对象与其余的对象特征向量的距离，并将距离排序，能找到与当前数据距离最近的前 k 个邻居，k 个最近邻居中最常见的分类决定了赋予该对象的类别。

二、Project Flow



三、问题及解决方案

1. 接口问题

在写 VSM 时，没有考虑 KNN 的实现，所以写 KNN 时又返回去重新改了 VSM 的程序，让 VSM 返回计算出来的 TF-IDF，这样在 KNN 里可以直接调用 VSM 函数。

2. 优化问题

其实 KNN 实现起来不难，网上也有很多可以借鉴的，值得注意的是细节问题，比如存 TF-IDF 值的矩阵怎么存放读取、代码速度问题等。还有就是这种大程序跑起来真的很磨人，等好久才能出结果，在从文件数量小到大的转化时，会有一些问题，而且发现问题的过程很漫长。

四、实验结果

实验结果截图：

```
step 4: show the result...
The classify accuracy is: 0.62
```

我自己感觉还可以通过改变字典的构建来继续优化结果。