Project-1 遇到的问题与思考

- 如何选择和构造特征？如果全部词语直接构造tfidf计算效率较低而且数据包含的噪声较多，因此我想到从两个方面着手：
  - 在分词之前将文本中的特殊字符中文标点去除，因为这些符号对提供信息意义不大。在分词之后去掉常用的中文停用词，同样地因为这些停用词在文本出现的频率可能较高但提供的信息很少。
  - 分词完毕后，构建tfidf矩阵时，考虑到稀疏矩阵包含的信息较少而且计算效率低，因此我把每篇文档中出现词频少于此篇文档总词数的1.5%的词汇忽略，最后才构建tfidf矩阵。

- 如何避免过拟合?
  - 将训练样本分为训练集和测试集，训练好模型后，主要观察accuracy，precision，recall和f1score这几个指标在cross-validation的表现和在测试集上的表现。

- 如何应用训练好的模型去判断定位抄袭文章?
  - 首先要挑选出疑似抄袭的文章，用训练好的模型去预测全部样本，得到预测值，用预测值和真实值比较，预测为正类单真实值为负类的样本很可能就是抄袭文章。

  - 对文本进行聚类，根据聚类结果在同一个簇中搜索相似文本。