公司所属行业划分的最佳模型（新数据集）

1. 问题界定

在课程案例中，我们已经针对一批数据实现了基于公司概述分类公司所属行业。

现在又得到了一批**新的数据**，但是因为数据涵盖的时间、公司范围等有所不同，需要分析人员重新构造分类模型，评估模型效果。

本次项目中，请参考课程案例的主要思路，基于新的数据集对公司行业进行分类。

2. 数据准备

（1）导入新表 company_industry_overview.csv 的数据后，请检查现有数据集中是否存在数据缺失、数据冗余、字段取值异常的情况，并进行相应的处理。

（2）请参考课程案例中的切词方法和异常记录处理方法，对字段 company_industry 与 company_overview 分别切词，并根据切词后的结果，删除含义上较为异常的记录。

（3）针对现有公司行业字段，请提出相应的数据改造方案，保留若干个最主要行业；针对现有公司概述字段，请通过 TF-IDF 对概述文本进行转换。

3. 数据建模与方法比较

与课程案例保持一致，本次项目中我们依然比较三种模型：决策树、朴素贝叶斯、SVM。

（1）请针对三种模型，在对训练集进行简单重采样的情况下，基于验证集与测试集准确率，确定模型的最优超参数组合。

（2）请在确定三个模型的最优参数组合的情况下，比较五种不同数据重构方式（原训练集、简单重采样、SMOTE 重采样、ADASYN 重采样、简单亚采样），并确定其中效果最好的一个。

（3）在确定三个模型的最优重采样方法后，请确定数据分割中按标签分割与随机分割

哪一种的效果更好。

（4）请比较三个模型的准确率，确定最佳模型进行最终评估。

提高题：

比较本次项目与课程案例中最终模型的准确率可以发现，本次项目中最终模型的训练集准确率较高而测试集准确率较低，存在过拟合的现象，初步判断可能与数据集的质量有关。请尝试更多对数据的清理方法，清除质量不高的数据记录，并查看是否会改善最终模型的效果。

提示：

可以采取的思路有，清除公司概述文本长度过短或者过长的记录；或者人工查看数据集，清除异常记录等。