对SEC的AAER数据我进行了进一步的处理，筛选出了239条有效记录，其中：有效指的是，财务造假and处理对象含有一家公司。现在还需要做的处理是：

1.把数据中含有公司名称的一行做切分处理并保留有效部分

2.删去重复的数据

3.剔除IPO造假的部分

4.寻找首次财务造假的年份

针对excel数据的说明：H列中，1和finance为有效数据，0为无效数据；后面的列是分组筛选列，标明：有无“Inc”字样、有无“CPA”字样、有无“et al”字样，有无“，”字样，是我做人工筛选的辅助指标。

现需用电脑处理的部分：

1.对D列：先删去尾缀（由“Other Release No”或者“Release No”引导），按分隔符“，”和分割分“and”做分割，注意，如果某个分割后的组分为Inc或者Ltd，则将其与前面一个组分相连；删去含有CPA、et al字样的组分，如还有多个组分则保留含有company，Inc，Ltd，Corp字样的组分，\*如果此规则仍旧不能适用则保留第一个组分\*；删去D列相同的数据条目，保留靠后（时间靠前）的那条

2.对document：在已经筛选好的数据条目对应的文档中搜索IPO，reorganization,recapitalization,reconstruction（还可能是其动词形式或分词等），删去对应条目