我对数据进行了一次手工筛选，排除掉了针对个人、会计事务所、券商、金融机构等的处罚，因为它们都不涉及年报的财务造假，将排除掉这些情况外的数据条目的F列标记为1。更新后的excel表格保存在data\_new.xlsx中。

一、对excel做初步筛选

1.仅保留F列为1的数据

2.更新B列内容，仅保留原来最外层括号内的内容，有例外情况：编号911、913、982、1018、1022、1024、1026、1028、1029、1030、1035、1036、1044，例外情况的B列要全部保留并不参与后面步骤

注：例外情况的筛选标准：1）无括号 或 2）无“决定书”字样

3.对B列内容做切分，切分符为“、”或者“及”（如果未找到切分符，则直接结束），然后先删掉含有数字的组分，再删掉字数小于等于3的组分（如果删掉含有数字的组分后，所有组分的字数都小于等于3，则保留第一个组分）

二、对document目录里的txt做进一步筛选

1.首先删去含有“IPO”或“重组”字样的（把对应的数据在excel表格的F列改为0、筛选F列为1的项，继续）

2.保留含有“财务”or“虚假”or“虚增”or“虚构”的数据（F列改为2、筛选F列为2的项）

三、再对excel做筛选

在前两步做完之后，如果还出现有多个数据的B列内容相同时，我们仅保留时间靠前的一次数据。

之后手工筛选方式：关注年报是否已经更正，寻找首次年报造假的年份，之后再查找对应公司的股票代码（这块我不知道能不能用程序完成，这一块也可以放到前面来做），下载所需年份的报表。