# 分析师报告信息抽取

## 图表数量

原则：图和表都会有名字，计算出现过多少次名字可以等价的认为出现过多少图表

失败经验：

1. 尝试过利用pdfbox提取出一篇pdf中所有的图片，但是这些图片中有很多非图表图的图片，如果尝试区分还需要用到图像识别技术，且图片提取稳定性不高，故放弃。
2. 从计算机程序的角度看pdf，是没有人眼直观感受到的表这一概念的，（there is no table, just characters, fonts and lines）

## 评级信息

原则：