2018-9-20

需求： 制作 PDF 内容提取工具

任务 1： PDF 资料初步筛选（去重、行业类别简单命名、自动归档）

- 通过写脚本自动处理，文件批量更名并通过文件名称进行筛选

任务 2： PDF 报告基础内容提取（标题、日期、页数、目录、图表目录 ）,并制作 TXT 索引

- 通过 PDF 插件对 PDF 内文字进行操作

任务 3： PDF 报告深度内容提取（图表目录 1 对 1 图表）

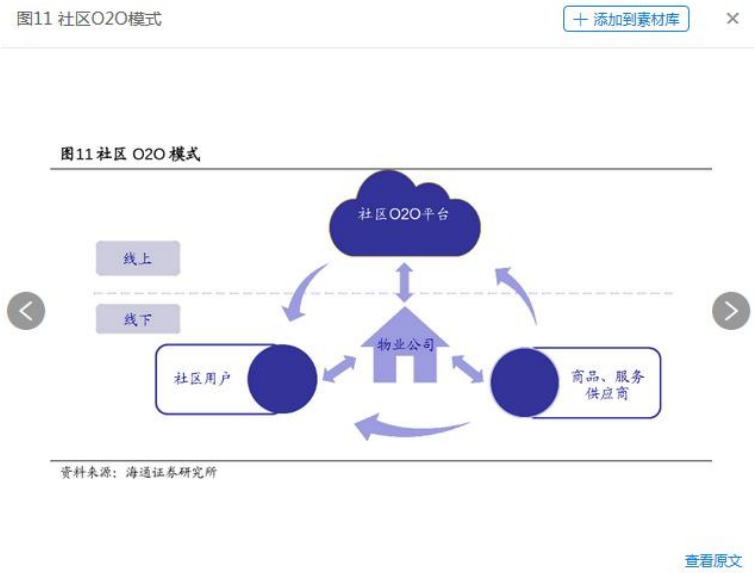
- 通过 PDF 处理插件或 OCR 等技术识别，实现图片自动处理

示例：从 PDF 中提取文章目录和图表目录并写入 TXT

目录

图表目录

示例：从 PDF 自动截图并命名 ，关联研究报告



2018-9-26 信息更新：

解决问题的方式正确，但是作为可交付的产品，需要从用户角度处理细节和代码执行中的问题，无需完美，但需要完整。

任务 1）没有问题

任务 2）作为批量处理成百上千个 PDF 文件的工具，提出一些细节上的建议：

- 待处理的文件放在同一个文件夹下
- 执行代码后，文件夹下新建 3 个文件夹，分别是：研报目录，未处理完成，处理完成
- 执行代码后，将处理好的 TXT 文件放置于 研报目录中，TXT 文件名与 PDF 文件名相同
- 执行代码后，成功处理的 PDF 文件放置入目录 处理完成，，未成功处理的 PDF 放置入 未处理完成

用户处理流程：

用户批量拷贝 PDF 文件到 文件夹下
执行代码（生成文本和图表目录 TXT，重新移动 PDF 文件）
手工检查未处理成功的文件原因，放弃或进行二次清洗

任务 3）目标是完成图片的批量处理，也提出一些细节修改

- 待处理的文件放在同一个文件夹下
- 文件夹下新建 3 个文件夹，分别是：处理完成，未处理完成，文件图表
- 执行代码后，文件图表目录下新建一个目录（名称和文件名相同）存放 PDF 提取的图片，
- 执行代码后，成功处理的 PDF 图片放置入文件图表下的同名目录，并移动 PDF 至处理完成目录
- 执行代码后，未成功处理的 PDF 删除新建的同名目录，避免与成功处理的图片目录混淆，并移动 PDF 至未处理完成目录

用户处理流程：

用户批量拷贝 PDF 文件到 文件夹下
执行代码（自动建目录，读取图片，删除失败的空目录，移动文件）
手工检查未处理成功的文件原因，放弃或进行二次清洗

任务 4）清除安全密码，清除 PDF 已打的水印对象