

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

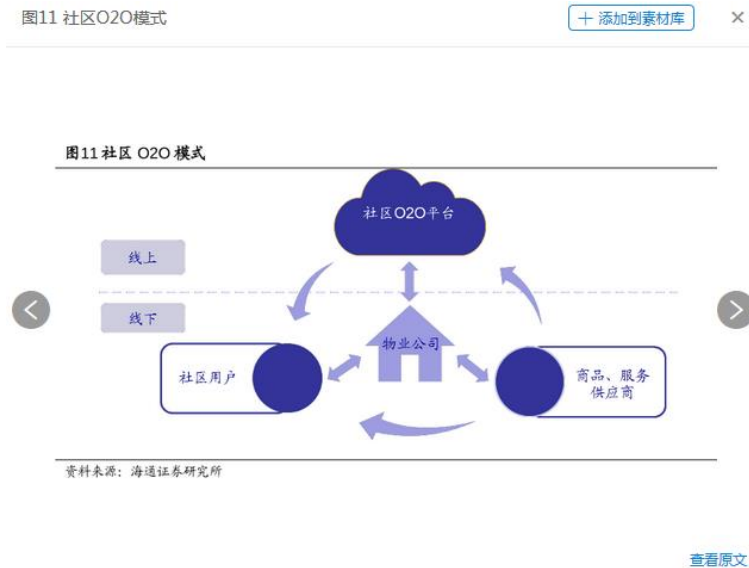
任务 3：PDF 报告深度内容提取（图表目录 1 对 1 图表）

- 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
1 商业本质：味觉记忆筑基业长青，饮食差异造高护城河.....7	图表 1：欧美食品饮料领域过去五年 ROE 均值超过 15%公司一览.....8
1.1 调味品企业基业长青的基石：味觉记忆+生活必需.....7	图表 2：中西饮食差异对比分析.....8
1.2 饮食差异带来行业外护城河高筑.....8	图表 3：调味品、乳制品与白酒行业增长与格局对比分析.....9
1.3 横向对比：调味品强在持续的高盈利能力，板块与龙头公司溢价明显.....9	图表 4：食品饮料细分领域销售净利润率对比（%）.....10
2 行业概况：多品类汇聚大市场，众多细分领域尚待开发.....11	图表 5：食品饮料行业 ROE 前十大企业.....10
2.1 酱油：最大细分市场，结构升级空间大.....13	图表 6：主要调味品企业估值对比（PE TTM）.....10
2.2 食醋：生产与口味呈现明显地区差异，缺乏全国性龙头企业.....15	图表 7：调味品分类.....11
2.3 蚝油：市场渗透率最低，正处快速放量初期.....16	图表 8：调味品按材料和工艺分类.....11
2.4 酱类：品种繁多，老字号等证明全国化发展未来不可.....17	图表 9：1999-2017 年调味品行业规模（亿元）.....12
3 论渠道：餐饮量大而稳，易守难攻；家庭重营销，易攻难守.....18	图表 10：2008-2015 年各调味品收入占比.....12
3.1 餐饮渠道：粘性高，量大且稳定，先发优势明显.....19	图表 11：主要调味品细分领域介绍.....13
3.2 家庭渠道：重品牌与营销，易于产生变革.....20	图表 12：酱油、食醋及类似制品主营收入.....13
4 日本发展启示：餐饮占比持续提升，复合调味料是主流，调味品功能不断细分.....21	图表 13：98-17 年规模企业酱油产量及增速（万吨）.....13
4.1 龟甲万：率先拓展海外市场，突破地域限制，规模效应明显.....25	图表 14：酱油的食用方法分类.....14
5 行业展望：餐饮向上发展与消费升级助推稳步增长，预计年增速 5-10%.....30	图表 15：酱油的品种分类和制作方法.....14
5.1 餐饮行业：外出就餐比例提升，餐饮回暖复苏助推提速.....30	图表 16：中国酿造酱油分级别介绍.....15
5.2 企业有定价权，调味品价格相比成本更快提升.....32	图表 17：中国酿造酱油分级别占比.....15
5.3 产品结构升级，中高端、功能型调味品引领潮流.....33	图表 18：2015 年各美食醋消费分析.....16
6 企业成长路径：集中度提升下，品类扩张+兼并收购+餐饮布局.....35	图表 19：不同地区食醋消费特点比较.....16
6.1 格局演变：从分散走向集中，走品牌化、全国化、平台化道路.....35	图表 20：2006-2015 年中国食醋（50 强/100 强）销售收入（亿元）.....16
6.2 路径分析：品类扩张循序渐进，兼并收购将是常态，餐饮布局至关重要.....37	图表 21：2006-2015 年中国食醋（50 强/100 强）食醋产量（万吨）.....16
	图表 22：09-17 年海天味业蚝油收入及增速.....17
	图表 23：各个地区调味品代表产品介绍.....17

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