介绍

这是一个半自动抓取PDF表格的demo，虽然它还是有非常多的缺陷，我们只会偶尔拿它批量处理PDF数据，对于一些特殊结构的（多重表头）PDF还需要定制修改某些语句，所以不推荐尝试，仅留作参考

一切输出和交互都在控制台实现：）

使用

使用前：在py文件旁创建一个 Data 的文件夹，并将需要处理的PDF放入，即可运行py：

- 初始化：扫描文件夹中所有PDF做出统计，创建三个文件夹：/FinishPDF、/ProblemPDF、/SaveData
- 接下来逐个PDF读取：
  - 先抓取所有页面的文字呈现，查看PDF是否保留（按下0丢弃/1继续，并回车），如果选择0则PDF将转移至 /ProblemPDF
  - 选择继续则，自动抓取所有表格并自动拼接，生成DataFrame，查看PDF是否保留（按下0丢弃/1继续，并回车）
  - 选择继续则，进入工具箱ToolBox，提供了几个便捷操作：
    - 0： 选择你需要保留的columns， 输入中文，以及逗号隔开（中/英），回车即可
    - 1： 加入一列相同的数据（常用于加入学校名称），用法为：先输入表头名，再输入填入数据
    - 2： 重命名所有columns，输入同等栏数量的标签（逗号隔开，中英文都可），若是数量不匹配不会更新
    - 3： 删除某行（不太好用，因为控制台对DataFrame的输出会压缩行，显示不完全），输入数字，或数字-数字都可以删除，eg：7 或 2-19
    - 4： 可以支持一次撤回，以防错误操作
    - 5： 丢弃，放回 /ProblemPDF
    - 6： 保存
  - 输入院校名称，即可保留为XXXX.csv

如果有任何问题可以随时终止程序

当然这个半自动程序真的有许多bug，如果执意使用本程序但遇到bug请找一下作者..

留个坑

... 当然，想过半自动当然也想过全自动，曾经有方案是：爬取所有表格后，通过词分类模型将不同字但同语义的聚类，实现标签归一后提取数据。但是表格不规则样式与院校对PDF做特殊处理（图片化）还是造成很大影响。方案二是走OCR文本扫描的老路，同时看了一篇图神经解构不规则表格信息提取的文章：GFTE: Graph-based Financial Table Extraction，但还是被不规则处理绊住脚，最终还是决定发起共筑数据集的号召...