附件：

方法很多，比如之前提到的 pypdf。然而用起来其实稍显麻烦，很多操作不够方便。

所以我一般用 pdf2htmlex（github 上有，一个国人项目，非 python）先把 pdf 转 html，接下来再用 bs4 来解析处理。好处是处理 html 的工具非常非常丰富，且 pdf2htmlex 对原页面的效果保持得特别好，特别是对于那些个用 word 和 latex 导出的 pdf 里，大量数据图表里的标签可以很方便地把值抓出来……

方法很多，比如之前提到的 pypdf。然而用起来其实稍显麻烦，很多操作不够方便。

所以我一般用 pdf2htmlex（github 上有，一个国人项目，非 python）先把 pdf 转 html，接下来再用 bs4 来解析处理。好处是处理 html 的工具非常非常丰富，且 pdf2htmlex 对原页面的效果保持得特别好，特别是对于那些个用 word 和 latex 导出的 pdf 里，大量数据图表里的标签可以很方便地把值抓出来……

方法很多，比如之前提到的 pypdf。然而用起来其实稍显麻烦，很多操作不够方便。

所以我一般用 pdf2htmlex（github 上有，一个国人项目，非 python）先把 pdf 转 html，接下来再用 bs4 来解析处理。好处是处理 html 的工具非常非常丰富，

项目	起始日期	终止日期	金额
项目一	2011	2015	1000
项目二	2013	2016	2000
项目三	2014	2018	50000

所以我一般用 pdf2htmlex（github 上有，一个国人项目，非 python）先把 pdf 转 html，接下来再用 bs4 来解析处理。好处是处理 html 的工具非常非常丰富，

项目	终止日期	金额
项目四	2015	1000
项目五	2016	2000
项目六	2018	50000