# 岩浆岩表格抽取工具简要说明

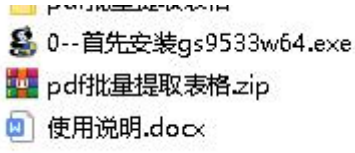
工作中经常碰到需要提取公开发表文献中的表格数据的情况（PDF 格式，caj 格式等），为了提高工作效率和落实 2021 年 4 月份 DDE 昆山工作坊的决议，编写了本程序实现批量处理。

## 一、使用环境

经测试，Win10　64 位,i7 CPU，16 G 内存，500G 硬盘运行正常。

## 二、安装方法：

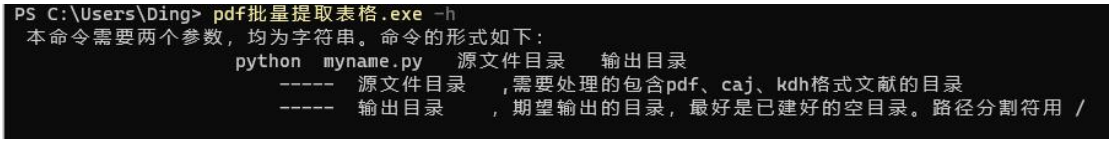
1） 首先下载压缩文件后进行解压，得到三个文件：




2） 安装第三方工具：gs9533w64.exe

3） 进一步解压 “ pdf 批量提取表格.zip” 到硬盘某处，例如 D:\。

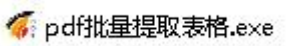
4） （可选）并把解压后的目录配置到 Path 路径中。检验是否配置好的方法是在命令行中输入“pdf 批量提取表格.exe -h ”得到如下界面：




## 三、使用方法：

### 1）目录拖放方式：

① 把需要提取的 pdf 文件和 caj 文件,一起拷贝到程序所在目录的 input 文件夹中，支持多级目录嵌套；

② 执行 pdf 批量提取表格.exe （双击 

）

③ 在 output 目录里面就会出现每个 pdf 对应的文件夹的名字，在各自的文件夹下面有从该 pdf 中提出的表格数据，按照 表格 1.xls，表格 2.xls….方式命名，原始的 pdf 文件以“原始文件.pdf”存在。

④ 所有的 caj 或者 kdh 文件，会首先被转换为 pdf 文件，然后与其它的 pdf 文件一起进行提取。

2）命令行方式：

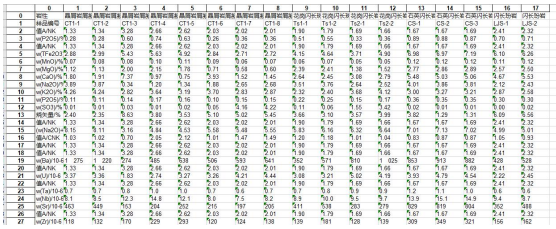
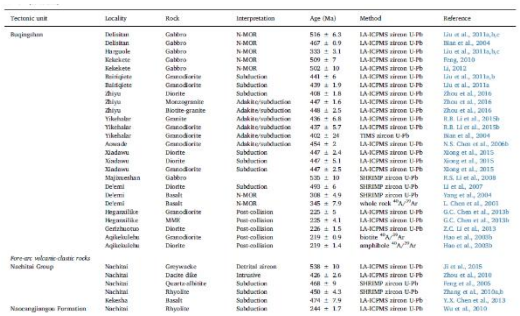
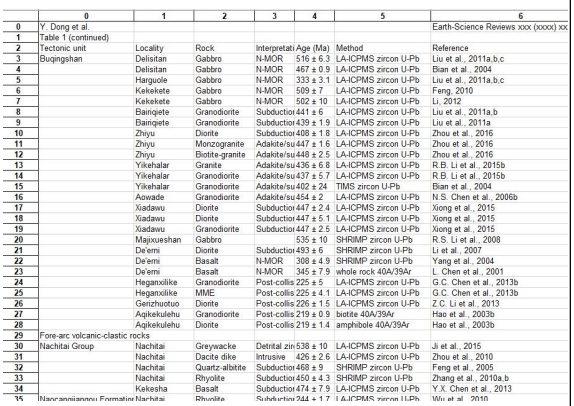
在执行了“二”中的第四步可选步骤后，可以使用命令行方式进行提取，方式类似如下：

```
PS C:\Users\Ding> pdf批量提取表格.exe  D:/表格提取测试  d:/test
```

即在命令行中输入“ pdf 批量提取表格 目录 1  目录 2 ”即可。

## 四、结果解读

实现效果：

| PDF 中截图 | 转换后的 Excel 截图 |
| --- | --- |
|  线框式表格 |  |
|  没有线分割的表格 |  |

## 支持范围：

*只支持文本类型的 pdf 文件表格提取，针对扫描的需要 ocr 的文件，提取不了。针对该类型的文件，请使用 abbyy finereader 首先进行 ocr 识别，然后另存 pdf 之后再进行提取。

*不是所有的 caj 都可以转换为 pdf，转换成功与否参见日志文件中的记录。

*所有执行结果在 log 目录中的日志文件中均有记录。Log 目录为运行时所在

的目录下的，在命令行时尤其注意。

*目前的提取准确率相对较高，但距离人工提取精度尚有较大差距，需人工校对提高器精度。

*针对岩浆岩类型的文档做了局部优化，并且在不断改进中。针对其它学科未作特殊处理。

*表格中的上标以<s></s>圈闭表示。

*如果未能提取到表格，则目录中不存在 excel 文件；如果存在，则请参考 sheet 的名字对应原始 pdf 中的页码。

*针对表格中的表头与内容横竖混排的识别效果较差。有一些小型表格会被忽略掉。

## 五、使用技巧

1、针对 pdf 中竖版排列的表格，在识别之前，请使用 acrobat reader 免费软件旋转为正向（即文字正向排列）。

2、可以通过多级目录嵌套进行文件分类，在提取之前做好备份（正常情况下不会改动 input 文件中的文件，除了会把 caj 转化为 pdf 外）。

3、针对扫描得来的的 pdf 文件，请首先进行 ocr 识别之后另存为 pdf 然后再识别。

4、针对不同余种，例如俄语、日语、阿拉伯语等，建议先通过目录进行分类，然后分别进行处理。中英文的文件可以混同一起处理。

## 六、致谢

本软件引用了 pandas，numpy，pdfminer，mupdf，ghostscript，camelot-py，caj2pdf，openpyxl 等等诸多开源类库，谨对原作者致以诚挚谢意。

## 七、其它

首先感谢各位专家各位老师的关注，并期盼得到您使用中遇到的情况的反馈，您的宝贵意见将有利于改进本软件并提升准确度。

封装为 exe 丧失了一部分参数的手工调整空间，如遇到俄语、阿拉伯语等特殊需求时，请联系我们调整参数。

反馈邮箱：dingyi@cags.ac.cn

DDE 岩浆岩工作组,2021 年 4 月