

岩浆岩表格抽取工具简要说明

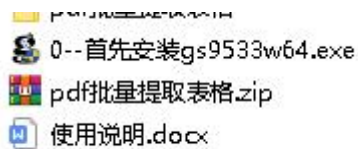
工作中经常碰到需要提取公开发表文献中的表格数据的情况（PDF 格式，caj 格式等），为了提高工作效率和落实 2021 年 4 月份 DDE 昆山工作坊的决议，编写了本程序实现批量处理。

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：

```
PS C:\Users\Ding> pdf批量提取表格.exe -h
本命令需要两个参数，均为字符串。命令的形式如下：
python myname.py 源文件目录 输出目录
----- 源文件目录 ，需要处理的包含pdf、caj、kdh格式文献的目录
----- 输出目录 ，期望输出的目录，最好是已建好的空目录。路径分割符用 /
```

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  pdf批量提取表格.exe ）
- 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

Table 1. Analytical Results of Major, Trace and Rare Earth Elements

转换后的 Excel 截图

线框式表格

Tectonic unit	Locality	Rock	Interpretation	Age (Ma)	Method	Reference
Buqingshan	Taishan	Gabbro	N-MOR	536 ± 4.9	LA-ICPMS zircon U-Pb	Liu et al., 2011a,b
	Delistan	Gabbro	N-MOR	467 ± 6.9	LA-ICPMS zircon U-Pb	Wan et al., 2004
	Huagoule	Gabbro	N-MOR	252 ± 2.1	LA-ICPMS zircon U-Pb	Liu et al., 2011a,b,c
	Kekelake	Gabbro	N-MOR	509 ± 7	LA-ICPMS zircon U-Pb	Feng, 2010
	Kelakeke	Gabbro	N-MOR	542 ± 6	LA-ICPMS zircon U-Pb	Liu et al., 2011a,b
	Buqingshan	Granodiorite	Subduction	449 ± 1.9	LA-ICPMS zircon U-Pb	Liu et al., 2011a,b
	Zhuyi	Monzogranite	Subduction	449 ± 1.6	LA-ICPMS zircon U-Pb	Zhou et al., 2016
	Zhuyi	Biotite-granite	Subduction	449 ± 1.6	LA-ICPMS zircon U-Pb	Zhou et al., 2016
	Yixiehu	Gneiss	Adakite-subduction	436 ± 6.8	LA-ICPMS zircon U-Pb	R. B. Li et al., 2015b
	Yixiehu	Granodiorite	Adakite-subduction	429 ± 5.7	LA-ICPMS zircon U-Pb	R. B. Li et al., 2015b
	Yixiehu	Granodiorite	Adakite-subduction	402 ± 2.8	TRMS zircon U-Pb	Wan et al., 2004
	Wuliandu	Diorite	Subduction	447 ± 2.4	LA-ICPMS zircon U-Pb	Xiong et al., 2015
	Wuliandu	Diorite	Subduction	447 ± 2.4	LA-ICPMS zircon U-Pb	Xiong et al., 2015
	Wuliandu	Granodiorite	Subduction	447 ± 2.5	LA-ICPMS zircon U-Pb	Xiong et al., 2015
	Majoueshan	Gabbro	Subduction	520 ± 1.9	SHRIMP zircon U-Pb	Li et al., 2007
	Deqeni	Diorite	Subduction	493 ± 4.6	SHRIMP zircon U-Pb	Li et al., 2007
	Deqeni	Basalt	N-MOR	345 ± 7.9	walke rock ⁴⁰ Ar/ ³⁹ Ar	L. Chen et al., 2003
	Hegankouli	Granodiorite	Post-collision	222 ± 5	LA-ICPMS zircon U-Pb	G.C. Chen et al., 2013b
	Hegankouli	Migmatite	Post-collision	226 ± 4.1	LA-ICPMS zircon U-Pb	G.C. Chen et al., 2013b
	Gentzhuotuo	Diorite	Subduction	226 ± 1.5	LA-ICPMS zircon U-Pb	Z.C. Li et al., 2013
Aqikekulehu	Granodiorite	Post-collision	219 ± 6.9	biotite ⁴⁰ Ar/ ³⁹ Ar	Hao et al., 2003b	
Aqikekulehu	Diorite	Post-collision	219 ± 1.4	amphibole ⁴⁰ Ar/ ³⁹ Ar	Hao et al., 2003b	
Fore-arc volcanic-clastic rocks	Nachitai	Greywacke	Detrital siltstone	538 ± 13	LA-ICPMS zircon U-Pb	Ji et al., 2015
	Nachitai	Dacite dike	Siltstone	426 ± 2.4	LA-ICPMS zircon U-Pb	Zhou et al., 2010
	Nachitai	Quartz-albite	Subduction	468 ± 9	SHRIMP zircon U-Pb	Feng et al., 2005
	Nachitai	Rhyolite	Subduction	490 ± 4.3	SHRIMP zircon U-Pb	Zhang et al., 2010a,b
Nanwanjiangou Formation	Kelashan	Basalt	Subduction	474 ± 1.9	LA-ICPMS zircon U-Pb	Y. X. Chen et al., 2013
	Nachitai	Rhyolite	Subduction	241 ± 1.7	LA-ICPMS zircon U-Pb	Wu et al., 2010

没有线分割的表格

0	1	2	3	4	5	6
0	Y. Dong et al.					Earth-Science Reviews xxx (xxxx) xxx
1	Table 1 (continued)					
2	Tectonic unit	Locality	Rock	Interpretal Age (Ma)	Method	Reference
3	Buqingshan	Delistan	Gabbro	N-MOR	516 ± 6.3	LA-ICPMS zircon U-Pb
4		Delistan	Gabbro	N-MOR	447 ± 0.9	LA-ICPMS zircon U-Pb
5		Huagoule	Gabbro	N-MOR	333 ± 3.1	LA-ICPMS zircon U-Pb
6		Kekelake	Gabbro	N-MOR	509 ± 7	LA-ICPMS zircon U-Pb
7		Kelakeke	Gabbro	N-MOR	542 ± 6	LA-ICPMS zircon U-Pb
8		Buqingshan	Granodiorite	Subduction	441 ± 6	LA-ICPMS zircon U-Pb
9		Zhuyi	Monzogranite	Subduction	439 ± 1.9	LA-ICPMS zircon U-Pb
10		Zhuyi	Biotite-granite	Subduction	449 ± 1.6	LA-ICPMS zircon U-Pb
11		Yixiehu	Gneiss	Adakite-subduction	436 ± 6.8	LA-ICPMS zircon U-Pb
12		Yixiehu	Granodiorite	Adakite-subduction	429 ± 5.7	LA-ICPMS zircon U-Pb
13		Yixiehu	Granodiorite	Adakite-subduction	402 ± 2.8	TRMS zircon U-Pb
14		Wuliandu	Diorite	Subduction	447 ± 2.4	LA-ICPMS zircon U-Pb
15		Wuliandu	Diorite	Subduction	447 ± 2.4	LA-ICPMS zircon U-Pb
16		Wuliandu	Granodiorite	Subduction	447 ± 2.5	LA-ICPMS zircon U-Pb
17		Majoueshan	Gabbro	Subduction	520 ± 1.9	SHRIMP zircon U-Pb
18		Deqeni	Diorite	Subduction	493 ± 4.6	SHRIMP zircon U-Pb
19		Deqeni	Basalt	N-MOR	345 ± 7.9	walke rock ⁴⁰ Ar/ ³⁹ Ar
20		Hegankouli	Granodiorite	Post-collision	222 ± 5	LA-ICPMS zircon U-Pb
21		Hegankouli	Migmatite	Post-collision	226 ± 4.1	LA-ICPMS zircon U-Pb
22		Gentzhuotuo	Diorite	Subduction	226 ± 1.5	LA-ICPMS zircon U-Pb
23		Aqikekulehu	Granodiorite	Post-collision	219 ± 6.9	biotite ⁴⁰ Ar/ ³⁹ Ar
24		Aqikekulehu	Diorite	Post-collision	219 ± 1.4	amphibole ⁴⁰ Ar/ ³⁹ Ar
25		Fore-arc volcanic-clastic rocks				
26		Nachitai	Greywacke	Detrital siltstone	538 ± 13	LA-ICPMS zircon U-Pb
27		Nachitai	Dacite dike	Siltstone	426 ± 2.4	LA-ICPMS zircon U-Pb
28		Nachitai	Quartz-albite	Subduction	468 ± 9	SHRIMP zircon U-Pb
29		Nachitai	Rhyolite	Subduction	490 ± 4.3	SHRIMP zircon U-Pb
30		Nanwanjiangou Formation				
31		Kelashan	Basalt	Subduction	474 ± 1.9	LA-ICPMS zircon U-Pb
32		Nachitai	Rhyolite	Subduction	241 ± 1.7	LA-ICPMS zircon U-Pb
33						
34						

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认识效果较差。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 pdf 文件，请首先进行 ocr 识别之后另存为 pdf 然后再识别。

4、针对不同语种，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