

# 使用说明v.0.2.1

## 安装

```
# python>=3.6
pip install pdfmerger-<version>-py3-none-any.whl
```

## 使用方法

### 自动合并文件

```
pdfmerger.exe merge -h
Usage: pdfmerger merge [OPTIONS]

Sort and merge pdf with page number

Options:
  -f, --files PATH      files to merge. e.g: 1.pdf 2.pdf...
  -d, --directory PATH  input directory
  -o, --output PATH     output path
  -s, --sort            Specify whether to sort files, default is true
  --pattern TEXT        specify search regex pattern for extracting page
                        index,default:(\d+)
  --line-number INTEGER specify line number for extracting page
                        index,default=-1
  --headers PATH        Specify file path to insert header
  -h, --help            Show this message and exit.
```

命令: pdfmerger merge [d/directory] [o/output] [line-number] [pattern]

### 列出目录中的文件清单

- directory 可以通过 **list** 命令查看目录, 如下

```
pdfmerger.exe list <dir>
```

- 例子

```
pdfmerger.exe list .\gitrepo\pdfmerger\pdf_merger\test_data\
PDF list in the .\gitrepo\pdfmerger\pdf_merger\test_data:
[1 ] .\gitrepo\pdfmerger\pdf_merger\test_data\3.pdf
[2 ] .\gitrepo\pdfmerger\pdf_merger\test_data\2.pdf
[3 ] .\gitrepo\pdfmerger\pdf_merger\test_data\4.pdf
```

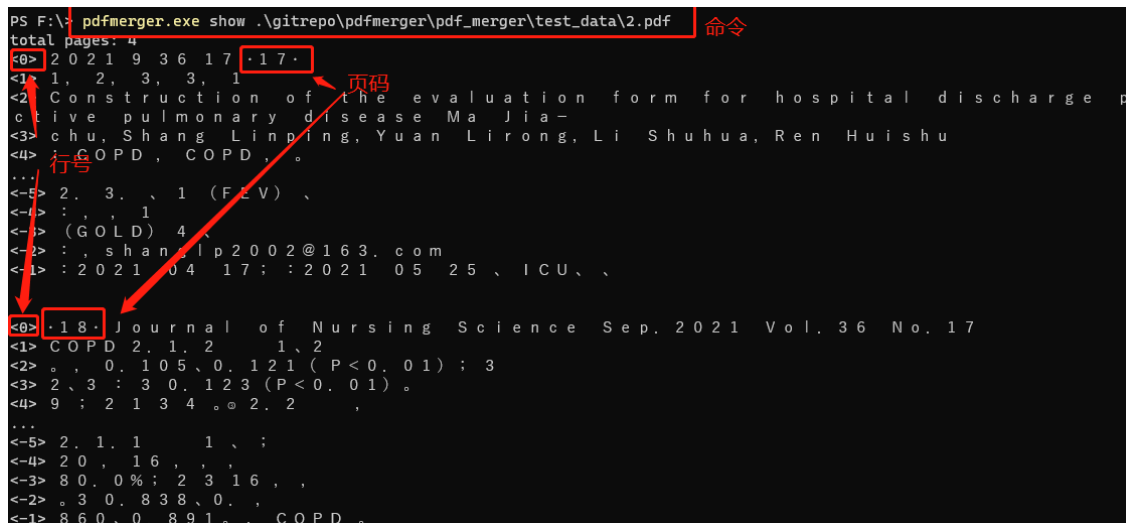
如果上述列出文件是需要合并的文件, 那么就将该目录作为合并的输入目录。

否则, 修改文件匹配模式, 比如加上通配符 (\*.pdf) 等, 或者移除不需要的文件。

## 显示pdf 文件特征信息

- line-number 参数表示页码在文本中的行数，可通过**show** 命令，从上数列表中任意选以文件进行查看特征

```
pdfmerger.exe show .\gitrepo\pdfmerger\pdf_merger\test_data\2.pdf
```



```
PS F:\> pdfmerger.exe show .\gitrepo\pdfmerger\pdf_merger\test_data\2.pdf 命令
total pages: 4
<0> 2 0 2 1 9 3 6 1 7 . 1 7 .
<1> 1, 2, 3, 3, 1
<2> Construction of the evaluation form for hospital discharge p
ctive pulmonary disease Ma Jia-
<3> chu, Shang Linning, Yuan Lirong, Li Shuhua, Ren Huishu
<4> : COPD, COPD, .
...
<-5> 2. 3. , 1 (F E V) ,
<-4> : , , 1
<-3> (GOLD) 4
<-2> : , shan j i p 2 0 0 2 @ 1 6 3 . c o m
<-1> : 2 0 2 1 0 4 1 7 ; : 2 0 2 1 0 5 2 5 , ICU , ,
<0> . 1 8 : Journal of Nursing Science Sep. 2021 Vol. 36 No. 17
<1> COPD 2. 1. 2 1, 2
<2> . , 0. 1 0 5 , 0. 1 2 1 ( P < 0. 0 1 ) ; 3
<3> 2 , 3 : 3 0. 1 2 3 ( P < 0. 0 1 ) .
<4> 9 ; 2 1 3 4 . . . 2. 2
...
<-5> 2. 1. 1 1 , ;
<-4> 2 0 , 1 6 , , ,
<-3> 8 0. 0 % ; 2 3 1 6 , ,
<-2> . 3 0. 8 3 8 , 0. ,
<-1> 8 6 0 , 0. 8 9 1 . . , COPD .
```

如果遇到显示内容乱码，如果页码信息已经正确显示，就可以忽略乱码问题（乱码需要安装对应字体库可解决，安装方法参考github中pdfminer.six项目）。

上述截图的行号就是--line-number需要的参数，上图中为0。

## 测试行号和匹配模式参数

- 页码为17,18，通过观察页码字符旁边的字符，可以修改--pattern 正则表达式来提高匹配精度，比如上例中可以设置匹配模式为--pattern="·(\d+)·"，旁边的特殊字符最好是复制，否则可能不一致，导致匹配失败。为了验证以上参数设置是否可行，可以通过 **test** 命令验证：

```
pdfmerger.exe test .\gitrepo\pdfmerger\pdf_merger\test_data\2.pdf --line-number=0
--pattern="·(\d+)·"
0: PageNumber 17
1: PageNumber 18
```

上述输出“PageNumber 17 PageNumber 18”正是我们需要的页码信息，说明匹配正确(可通过--pages 增加显示页数)，可以将参数用于合并，完整命令如下。

```
pdfmerger.exe merge -d .\gitrepo\pdfmerger\pdf_merger\test_data\ -o
.\out\test.pdf --line-number=0 --pattern="·(\d+)·"
```

**注意：**该命令会将列出的所有文件强行按照页码从小到大进行合并，尽管页码不连续。如果需要添加封面，可以用过--headers=封面路径.pdf, 进行添加到合并文档前部。该版要求封面文件不能和正文文件放在同一目录，后期可改进。