- PDF与Python
  - create by Dcount
- 一、 简介
  - 1.功能
- 二、提取内容
  - 1.文字内容
  - 2. 提取表格
- 三、分割合并
  - 1. 使用模块：pypdf2
  - 2. 读取写入
  - 3. 拆分
  - 4. 合并
  - 5. 旋转
  - 6. 排序
- 四、加水印和加密解密
  - 1. 批量加水印
  - 2. 加密

PDF与Python

create by Dcount

一、 简介

1.功能

1. 合并PDF
2. 从一堆文件中找到关键数据
3. 批量加密
4. 旋转页面
5. 批量加水印

二、提取内容

1.文字内容

`p.open(path)`:打开PDF `p.pages(number)`:那一页 `pages.extract_text()`：输出文字内容

```
import pdfplumber as p
with p.open("file.pdf") as pdf:
    first_page = pdf.pages[0]
    print(first_page.extract_text())
```

2. 提取表格

`pages.extract_table()` 多个表格`pages.extract_tables()`

- 如果提取有问题就设定参数，查官方文档 提取之后即可存到Excel里
- 去除空行,非空行才加进来,空行元素`''`or`None`;join的作用是元素用逗号拼接起来

```
new = []
for row in talbe:
   if not ''.join([str(item) for item in row ]) == '':
   sheet.append(row)
```

- 合并单词

```
new_row=[]
new_row.append(''.join([]))
sheet.append(new_row)
```

三、分割合并

1. 使用模块：pypdf2

2. 读取写入

`PdfFileReader()`:读取文件 `PdfFileWriter()`:写文件 `pdf.getNumPages()`：读取页数
`pdf.getPage(page)`:读取页面内容 `pdf.addPage(pdf_reader.getPage(page))`:添加页

3. 拆分

一份一份的添加后，保存为不同的文件

```
import pypdf2 as pp
pdf_reader = pp.PdfFileReader('filename.pdf')
for page in range(pdf_reader.getNumPages()):
    pdf_writer = pp.PdfFileWrite()
```

```
    pdf_writer.addPage(pdf_reader.getPage(page))
    with open(f'"路径"{page}.pdf','wb') as out:
        pdf_writer.writer(out)
```

---

4. 合并

一份一份的添加进去

```
import pypdf2 as pp
pdf_writer = pp.PdfFileWrite()

for page in range(16):
    pdf_reader = pp.PdfFileReader('{page}.pdf')
    for i in range(pdf_reader.getPage(page)):
        pdf_writer.addPage(pdf_reader.getPage(i))
with open('merged.pdf','wb') as out:
        pdf.writer(out)
```

---

5. 旋转

选中页面后才可以旋转，只能旋转90的倍数。 `page = pdf_reader.getPage(0).rotateClockwise(度数)`：顺时针旋转 `page = pdf_reader.getPage(1).rotateCounterClockwise(90)`:逆时针旋转

6. 排序

直接按照期望的顺序添加页面即可

四、加水印和加密解密

---

1. 批量加水印

相当于一个水印PDF加一个PDF文档，然后合并。 注意：下面的内容.mergePage(出现在上面的内容) 即：`水印.mergePage(文字)`

```
import pypdf2 as pp
from copy import copy
#读取水印
water = pp.PdfFileReader('水印.pdf')
waterpage = water.getPage(0)
#读取文件
pdf_reader = pp.PdfFileReader('filename.pdf')
pdf_writer = pp.PdfFileWriter()
```

```
for page in range(pdf_reader.getNumPages()):
    origin = pdf_reader.getPage(page)
    new = copy(waterpage)
    new_page.mergePage(origin)
    pdf_writer.addPage(new)
with open('watermarked.pdf','wb') as out:
    pdf_writer.write(out)
```

2. 加密

保存时设密码：`pdf_writer.encrypt(password)` 读取时输入密码：`pdf_reader.decrypt(password)`