# PDFluxAPI接口说明

## Table of Contents

## 产品介绍

1. PDFluxSaaS提供一套完整易用的REST API，供计算机调用，让您把PDF复杂格式抛诸脑后；
2. 支持将各类PDF格式的金融文档，包括年报、审计报告、IPO 招股说明书、债券募集说明书、评级报告、研究报告等解析为文本段落、表格、图片等内容块的序列，并保留原文档的阅读顺序；
3. 支持中英文多栏复杂排版的PDF文档解析，跨栏和跨页内容块智能自动合并；
4. 智能识别表格的内部结构：单元格合并，单元格文字的对齐方式、缩进、颜色、加粗、斜体等样式信息，跨页和跨栏表格智能合并单元格文字；
5. 智能识别文档的目录结构，支持多达10个层级，长文档信息抽取必不可少；
6. 支撑后续各类文档智能的应用：PDF文档全文检索、文档级别信息抽取等；
7. 使用庖丁科技自研Fin-OCR，高效识别模糊以及含有涂写、水印等干扰因素的文档。

## 使用说明

1. 在 https://saas.pdflux.com/ 中点击申请试用或通过 contact@paodingai.com 邮件联系我们开通试用；
2. 申请成功后我们会提供专属用户名、Secret和Get_Token.py文件；
3. 根据下文中的描述拼接请求URL，再通过运行Get_Token.py来获取带有token的URL；
4. 通过上一步中获取的带有token的URL即可上传文件、下载结果；
5. 除下载接口外，系统接口返回数据格式一般为 json；
6. https 请求提交数据时，如未特殊说明，也使用 json 格式；

**请求成功：**

```
{
  "status": true,
  "errcode": 0,
  "msg": null,
  "data": {}
}
```

**请求异常：**

```
{
  "status": false,
  "errcode": 1,
  "msg": "No File Found",
  "data": null,
}
```

## Token生成方法

**URL均需要添加验证参数：**

- _timestamp-时间戳
- _token-令牌

**验证参数生成方法：**

- 使用python3运行Get_Token.py + 对应URL

**示例：**

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/upload\?
user\=####\&force_update\=true
```

**根据URL生成Token示例（Python3）：**

```
import sys
import hashlib
import urllib.parse
from datetime import datetime

def revise_url(url, extra_params=None, excludes=None):
    extra_params = extra_params or {}
    excludes = excludes or []
    main_url, query = urllib.parse.splitquery(url)
    params = urllib.parse.parse_qs(query) if query else {}
    params.update(extra_params)
    keys = list(params.keys())
    keys.sort()
    params_strings = []
    for key in keys:
        if key in excludes:
            continue
        values = params[key]
        if isinstance(values, list):
            values.sort()
            params_strings.extend(["{}={}".format(key, urllib.parse.quote(str(value)))
for value in values])
        else:
            params_strings.append("{}={}".format(key,
urllib.parse.quote(str(values))))

    return "{}?{}".format(main_url, "&".join(params_strings)) if params_strings else
main_url

def generate_timestamp():
    delta = datetime.utcnow() - datetime.utcfromtimestamp(0)
    return int(delta.total_seconds())

def _generate_token(url, app_id, secret_key, extra_params=None, timestamp=None):
    url = revise_url(url, extra_params=extra_params, excludes=["_token",
"_timestamp"])
    timestamp_now = timestamp or generate_timestamp()
    source = "{}#{}#{}#{}".format(url, app_id, secret_key, timestamp_now)
    token = hashlib.md5(source.encode()).hexdigest()
    return token

def encode_url(url, app_id, secret_key, params=None, timestamp=None):
    timestamp = timestamp or generate_timestamp()
    token = _generate_token(url, app_id, secret_key, params, timestamp)
    extra_params = {
        '_timestamp': timestamp,
        '_token': token
    }
    extra_params.update(params or {})
    url = revise_url(url, extra_params=extra_params)
```

```
    return url

if __name__ == '__main__':
    URL = sys.argv[1]

    url = encode_url(URL, 'pdflux', '请在此处填写您的Secret')  # pdflux是appid，请勿改动
    print(url)
```

## 获取已使用页数、总页数和起止时间

**接口地址：** `/api/v1/saas/usage?user=####`

**请求方式：** `GET`

**请求参数**

| 参数名 | 类型 | 必要性 | 说明 |
| --- | --- | --- | --- |
| user | string | 必填 | 用户名，拼接在URL中，例：user=#### |

**返回字段**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| rest | number | 扫描件剩余页数，整数，可能为负数 |
| total | number | 扫描件总页数，整数 |
| normal_rest | number | 非扫描件剩余页数，正整数 |
| normal_total | number | 非扫描件总页数 |
| trial_start | string | 起始时间 |
| trial_end | string | 截止时间 |

**示例**

- 获取Token：

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/usage?user=pdflux
```

- 获取结果：

```
{
    "data":{
        "rest":99952, // 扫描件剩余页数
        "total":100001, // 扫描件总页数
        "normal_rest":100, // 非扫描件剩余页数
        "normal_total":200, // 非扫描件总页数
        "trial_start":"2020-10-01 00:00:00", // 起始时间
        "trial_end":"2020-10-31 00:00:00", // 截止时间
    },
    "errcode":0,
    "msg":null,
    "status":true
}
```

## 上传文件

**接口地址：** `/api/v1/saas/upload?force_update=true&user=####`

**请求类型：** `POST`

**请求参数**

| 参数名 | 类型 | 必要性 | 说明 |
| --- | --- | --- | --- |
| file | file | 必填 | 待分析文档 |
| user | string | 必填 | 用户名，拼接在URL中，例：user=#### |
| force_update | string | 非必填 | 强制文档重新识别，例如上传重复文件、旧文件时，PDFlux为了经济性默认会优先获取已有的旧结果，增加force_update=true后将会强制重新识别获取最新结果 |

**返回字段**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| uuid | string | 文件 id，后续用于获取结果 |
| checksum | string | 文件hash |
| filename | string | 文件名 |
| filepath | string | 文件路径 |
| parsed | int | -1=解析异常、0=待解析、1=解析中、2=解析完毕 |
| created_utc | int | 创建时间 |
| updated_utc | int | 修改时间 |

**示例**

- 获取Token

```
python3 get_token.py
"https://saas.pdflux.com/api/v1/saas/upload&force_update=true&user=pdflux"
```

- 请求上传：

```
POST  https://saas.pdflux.com/api/v1/saas/upload?
_timestamp=1590560297&_token=bf5bd348e4a414be0aa57899878bd66c&user=pdflux
```

- 返回结果：

```
{
    "status":true,
    "data":{
        "uuid":"b75487ae-09d6-4948-bac5-7924d24bedbb",
        "checksum": "dbc325d09eb6c8af234a57fe62ae6a20",
        "filename":"第二次公开募股.pdf",
        "filepath":"1672345948/f6f8e0b7df6188cd30b37004bbaa2b6d_1816579.pdf",
        "parsed": 1,
        "created_utc":1508467057,
        "updated_utc":1508467057
    }
}
```

## 获取文档处理状态

**接口地址：** `/api/v1/saas/document/<uuid>?user=####`

**请求类型：** `GET`

**请求参数**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| uuid | string | 文件 uuid，必填 |
| user | string | 用户名，必填 |

**返回字段**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| created_utc | int | 创建时间 |
| deleted | int | 删除，0代表未删除、1代表已删除 |
| filename | string | 文件名 |
| parsed | int | -1=解析异常、0=待解析、1=解析中、2=解析完毕 |
| updated_utc | int | 修改时间 |
| uuid | string | 文件 id，后续用于获取结果 |

**示例**

- 获取Token：

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/document/<uuid>?
user=pdflux
```

- 请求获取：

```
GET https://saas.pdflux.com/api/vl/saas/document/<uuid>?
_timestamp=1590560297&_token=bf5bd348e4a414be0aa57899878bd66c&user=pdflux
```

- 返回结果：

```
{
  "data": {
      "created_utc": 1534819038,
      "deleted": 0,
      "filename": "xxx.pdf",
      "id": 8,
      "parsed": 1,
      "updated_utc": 1534820149,
      "uuid": "b5a95e7a-a4ed-11e8-8a4f-8c8590cb4e8f"
  },
  "errcode": 0,
  "msg": null,
  "status": true
}
```


## 获取文档解析结果

**接口地址：** `/api/v1/saas/document/<uuid>/pdftables?user=####`

**请求类型：** `GET`

**请求参数**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| uuid | string | 文件 uuid，必填 |
| user | string | 用户名，必填 |

**解析结果中的部分字段含义**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| paragraphs | string | 段落元素块 |
| tables | string | 表格元素块 |
| images | string | 图片元素块 |
| page_header | string | 页眉元素块 |
| page_footer | string | 页脚元素块 |

*更多字段说明见结果示例及注释

**示例**

- 获取Token

```
python3 Get_Token.py
https://saas.pdflux.com/api/v1/saas/document/<uuid>/pdftables?user=pdflux
```

- 请求获取：

```
GET https://saas.pdflux.com/api/v1/saas/document/<uuid>/pdftables?
_timestamp=1590562575&_token=bc62219c497be230697368b87930eb33&user=pdflux
```

- 返回结果：

```
{
    "document": [
        {
            "id": 2022,
            "uuid": "2afdca4a-5c65-11eb-a1f8-00163e055917", // 文件uuid
            "parsed": 2,
            "filename": "中国广核电力股份有限公司主体与2019年度第一期中期票据信用评级报告（中诚信国际）.pdf", // 文件名
            "created_utc": 1611287515, // 创建时间
            "updated_utc": 1611287688, // 修改事件
            "exceptions": null // 报错信息
        }
    ],
    "pdf_page": [
        {
            "id": 10379,
            "did": 2022,
            "page": 0, // 页面序号、页码
            "meta": {
                "width": 595, // 页面宽度
                "height": 842, // 页面高度
                "page_file_type": null,
                "page_prob": null,
                "is_image": true // 是否是扫描件或图片
            },
            "created_utc": 1611287679,
            "updated_utc": 1611287679
        }, ... // ...表示省略
    ],
    "pdf_elements": [
        {
            "page": 1, // 页面序号、页码
            "elements": [
                {
                    "page": 1, // 页面序号、页码
                    "text": "中诚信国际 CCXI-20182331D-01", // 文字内容
                    "index": 0,
                    "element_file_type": "page_headers" // 元素块类型：页眉
                },
                {
                    "page": 1,
                    "text": "中国广核电力股份有限公司2019年度第一期中期票据信用评级报告",
                    "index": 1,
                    "syllabus": 1, // 和目录的对应关系
                    "element_file_type": "paragraphs" // 元素块类型：段落
                },
                {

                    "unit": "", // 表格单位
```

```
                    "cells": { // 表格单元格
                        "0_0": { //
                            "value": "发行主体" // 单元格内的文字内容
                        },
                        "0_1": { // 单元格位置信息：第一个“0”代表的是行数，第二个”0“代表的是列数，”0_0"代表第一个单元格
                            "value": "中国卜核*力股份有限公司"
                        },
                        ... // 表示省略
                    },
                    "title": "中国广核电力股份有限公司2019年度第一期中期票据信用评级报告",
                    "merged": [ // 单元格合并信息
                        [
                            [0,1],[0,2],[0,3],[0,4] // 表示这4个单元格合并
                        ],
                        [
                            [1,1],[1,2],[1,3],[1,4] // 表示这4个单元格合并
                        ],
                        ... // 表示省略
                    ],
                    "element_file_type": "tables", // 元素块类型：表格
                    "page": 1, // 页码
                    "index": 2 // 在页面中出现的顺序
                },
                {
                    "page": 1,
                    "text": "www.ccxi.com.cn 中国广核电力股份有限公司2019年度第一期中期票据信用评级报告",
                    "index": 3,
                    "element_file_type": "page_footers" // 元素块类型：页脚
                }
            ]
        },
        {
            "page": 2,
            "elements": [
                {
                    "data": "iVBORw0KGgoA......", // 图片内容，Base64格式
                    "page": 2, // 页码
                    "index": 0, // 在页面中出现的顺序
                    "element_file_type": "images"  // 元素块类型：图片
                },
                {
                    "page": 2,
                    "text": "关 注",
                    "index": 1,
                    "element_file_type": "paragraphs",
                },
                ...
            ]
```

```
        },
    ...
    ],
    "syllabus": { // 目录
        "index": -1, // 目录根节点
        "children": [ // 子节点
            {
                "page": 1, // 页码
                "efile_type": "paragraphs",
                "index": 1, // 目录序号
                "level": 1, // 目录层级
                "range": [ // 当前目录包含的元素块
                    1,10
                ],
                "title": "中国广核电力股份有限公司2019年度第一期中期票据信用评级报
告", // 目录标题
                "parent": -1, // 父节点
                "element": 1, // 目录与元素块的对应关系
                "children": [ // 子节点
                    {
                        "page": 2,
                        "efile_type": "paragraphs",
                        "index": 2,
                        "level": 2,
                        "range": [
                            5,10
                        ],
                        "title": "关 注",
                        "parent": 1,
                        "element": 5,
                        "children": []
                    }
                ]
            },
            {
                "page": 3,
                "efile_type": "paragraphs",
                "index": 3,
                "level": 1,
                "range": [
                    10,20
                ],
                "title": "声明",
                "parent": -1,
                "element": 10,
                "children": []
            },
            ...
        ]
    }
}
```

## 下载结果 - Excel

**接口地址：** `/api/v1/saas/document/<uuid>/excel?user=####`

**请求类型：** `GET`

**请求参数**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| uuid | string | 文件 uuid，必填 |
| user | string | 用户名，必填 |

**返回字段**

- 返回内容为 excel 文件

**示例**

- 获取Token

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/document/<uuid>/excel
?user=pdflux
```

- 请求下载：

```
GET https://saas.pdflux.com/api/v1/saas/document/<uuid>/excel?
_timestamp=1590562650&_token=1a014dbd2c719cea3ed04f249b50e39e&user=pdflux
```

## 下载结果 - HTML

**接口地址：** `/api/v1/saas/document/<uuid>/html?user=####`

**请求类型：** `GET`

**请求参数**

| 参数名 | 类型 | 说明 |
| --- | --- | --- |
| uuid | string | 文件 uuid，必填 |
| user | string | 用户名，必填 |

**返回字段**

- 返回内容为 html 文件

**示例**

- 获取Token

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/document/<uuid>/html?user=pdflux
```

- 请求下载：

```
GET https://saas.pdflux.com/api/v1/saas/document/<uuid>/html?_timestamp=1590562650&_token=1a014dbd2c719cea3ed04f249b50e39e&user=pdflux
```

## 发起OCR解析

**接口地址：** `/api/v1/saas/ocr-parse?user=###`

**请求类型：** `POST`

**请求参数**

| 参数名 | 必填 | 类型 | 说明 |
| --- | --- | --- | --- |
| file | 是 | file | 待解析的文档 |
| user | 是 | string | 用户名 |

**上传示例**

- 获取Token

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/ocr-parse?user=pdflux
```

- 返回结果

```
{
    "data": {
        "checksum": "401ca19386a424380d6c67d9879fba03",  // 文件校验和
        "created_utc": 1676449866,  // 上传时间
        "deleted": 0,
        "exceptions": null,
        "filename": "普联软件.PDF",  // 文件名称
        "filepath": "普联软件.PDF",
        "id": 4156,  // 文件ID
        "parsed": 1,  // 解析状态
        "pfb_pushed_status": null,
        "type": 3,
        "uid": -2,  // 上传用户ID
        "updated_utc": 1676449866, // 更新时间
        "uuid": "16fed1e4-ad0b-11ed-917c-0242c0a83002"  // 文件uuid
    },
    "errcode": 0,
    "msg": null,
    "status": true
}
```

## 获取OCR解析结果

**接口地址：** `/api/v1/saas/document/<uuid>/ocr-parse?user=###`

- 请求类型：`GET`
- 请求参数

| 参数名 | 必填 | 类型 | 说明 |
| --- | --- | --- | --- |
| uuid | 是 | string | 文件 uuid |

- 示例-获取Token

```
python3 Get_Token.py https://saas.pdflux.com/api/v1/saas/document/<uuid>/ocr-parse?user=pdflux
```

- 示例-请求下载OCR结果：

```
GET https://saas.pdflux.com/api/v1/saas/document/f97bcec6-5a5b-11ed-ab30-
0242c0a83002/ocr-parse?
_timestamp=1667359271&_token=766d76096d98bdad98015822ea7e7eca&user=pdflux
```