PDF转文档（阿里云市场）

在线试用：

https://try.dhconvert.com/

小程序搜索：

度慧文档转换

概述

使用流程

由于转换需要时间，文件越大页数越多，转换越久，故采用**异步**的方式获得转换结果。即调用转换接口后会获得token，随后有2种方式查询转换结果：

1. 定时轮询结果，调用“查询结果接口”。详细见：**查询结果**
2. 回调，设置callbackurl，当转换结束后，系统会回调该URL直接推送转换结果。详细见：**回调URL**

调用转换接口

> *v1和v2接口功能相同，调用方式不一样，可自选合适的。*
>
> **v1接口包括2种转换方式：**
>
> - 文档是一个下载链接，用HTTP GET方式，见：**文档转换GET_v1**
> - 文档二进制POST到服务器，用HTTP POST方式，见：**文档转换POST_v1**
>
> **v2接口统一为HTTP POST JSON：**
>
> - JSON支持输入：文件url；文件Base64字符串。见：**文档转换_v2**
> - v2版本的API除了异步，还支持同步调用，见：**同步调用**
> - v2版本中的参数如果未出现在v1版本中，v1版本用同样参数也能工作

调用转换API需要签名，详细见文档附录：阿里签名 调用查询结果API无需签名。

阿里云支持从OSS内网直接下载文件，节约流量，见：**阿里云独有部分**

文档转换GET_v1

将单个PDF下载链接转换为目标格式，type是**目标文档的type**，比如要把pdf转为docx，type就是docx

请求参数:

参数	类型	备注	是否必须发送
url	string	文件url，必须http(s)，ftp开头，**需要URL Encoding**	**是**
type	string	小写，转出文件类型，例如docx	**是**
ocr	int	对于扫描的PDF，是否做OCR： 0：不做OCR 1：自动OCR 2：强力OCR：针对有些加密或编码不正确导致的乱码，叠字，未OCR文字等问题 **默认1，建议：不需要OCR选择0，一般情况选择1，出现乱码等问题选择2**	否
language	int	OCR识别语言选项，默认2简体中文： 1：英语 2：简体中文 3：繁体中文 4：法语 5：德语 6：意大利语 7：俄语 8：日文 9：韩文 10：西班牙语 11：葡萄牙语 12：丹麦语 13：荷兰语 14：芬兰语 15：挪威语 16：瑞典语 17：土耳其语	否
excelonesheet	int	如果转为Excel文件，默认0：PDF特定页数以内为一个工作表，否则每页一个工作表；1：一个工作表（如果PDF页数太多，有失败可能）；2：每页一个工作表	否
wordnoimage	int	如果转为Word文件，默认0：需要图片；1：不需要图片	否
wordabsolutelayout	int	如果转为Word文件，默认0：流式布局；1：绝对布局（位置精准，浏览方便，但是编辑方式非流式不利于编辑）	否
textnoformat	int	如果转为Txt文本，默认0： 保持原有布局，1：无需布局	否
textperpage	int	如果转为Txt文本，默认0：所有页转为单个txt文件，1：每一页转为一个txt文件，并且打包为一个zip包	否
imagepdfocroption	int	如果是扫描版PDF，根据ocr参数做ocr。 如果是非扫描版PDF，根据该值做ocr。 默认0：不启用； 1-2：启用，非扫描版对应ocr：1-2 100：启用，非扫描版不做ocr 比如ocr传2，imagepdfocroption传1 那么：扫描版PDF用ocr：2做ocr 非扫描版用ocr：1做ocr	否
password	string	PDF文件的密码，没有密码可以不传或传空	否
pageindexes	string	要转换的PDF页数，默认空全部页，例如：1,3,5-7就是1,3,5,6,7共5页	否
outfilename	string	生成的文件的文件名，默认随机	否

参数	类型	备注	是否必须发送
callbackurl	string	回调URL，转换结束后，会回调该URL，**需要URL Encoding**，详细见 **回调URL**	否

请求示例:

```
https://pdf2doc.market.alicloudapi.com/v1/convert?url=https%3a%2f%2fxxx%2fxxx.pdf&type=docx&ocr=0
```

将所在url地址的pdf文件转为docx，type就是需转换为的文件类型，这个例子里就是docx

输出文件类型**type**可取值： **doc, docx, pptx, xlsx, rtf, txt, ofd**

必须签名才能调用成功，签名见阿里签名规则:

https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/use-cases/call-apis

返回数据结构:

名称	类型	是否必须返回	备注
code	number	是	10000:请求成功
msg	string	是	
result	Dictionary	否	成功后返回

result:

名称	类型	是否必须返回	备注
token	string	是	用于查询结果接口

返回示例(成功状态):

```
{
    "code":10000,
    "msg":"",
    "result":{"token":"xxx"}
}
```

返回示例(失败状态):

```
{
    "code":40001,
    "msg":"ParmNotRight"
}
```

文档转换POST_v1

直接将单个文档POST到服务器，大小限制**8M**

请求参数:

参数	类型	备注	是否必须发送
file	file	要转换的文档，Content-Type使用**multipart/form-data**，最大8M	**是**
type	string	小写，需转换为的文件类型，例如docx	**是**
ocr	int	对于扫描的PDF，是否做OCR： 0：不做OCR 1：自动OCR 2：强力OCR：针对有些加密或编码不正确导致的乱码，叠字，未OCR文字等问题 **默认1，建议：不需要OCR选择0，一般情况选择1，出现乱码等问题选择2**	否
language	int	OCR识别语言选项，默认2简体中文，值同GET方法	否
excelonesheet	int	如果转为Excel文件，默认0：PDF特定页数以内为一个工作表，否则每页一个工作表；1：一个工作表（如果PDF页数太多，有失败可能）； 2：每页一个工作表	否
wordnoimage	int	如果转为Word文件，默认0：需要图片；1：不需要图片	否
wordabsolutelayout	int	如果转为Word文件，默认0：流式布局；1：绝对布局（位置精准，浏览方便，但是编辑方式非流式不利于编辑）	否
textnoformat	int	如果转为Txt文本，默认0： 保持原有布局，1：无需布局	否
textperpage	int	如果转为Txt文本，默认0：所有页转为单个txt文件，1：每一页转为一个txt文件，并且打包为一个zip包	否
imagepdfocroption	int	如果是扫描版PDF，根据ocr参数做ocr。 如果是非扫描版PDF，根据该值做ocr。 默认0：不启用； 1-2：启用，非扫描版对应ocr：1-2 100：启用，非扫描版不做ocr 比如ocr传2，imagepdfocroption传1 那么：扫描版PDF用ocr：2做ocr 非扫描版用ocr：1做ocr	否
password	string	PDF文件的密码，没有密码可以不传或传空	否
pageindexes	string	要转换的PDF页数，默认空全部页，例如：1,3,5-7就是1,3,5,6,7共5页	否
outfilename	string	生成的文件的文件名，默认随机	否
callbackurl	string	回调URL，转换结束后，会回调该URL，详细见 **回调URL**	否

请求示例:

```
https://pdf2doc.market.alicloudapi.com/v1/convert
```

Header中的Content-Type必须是multipart/form-data

输出文件类型**type**可取值：**doc, docx, pptx, xlsx, rtf, txt, ofd**

必须签名才能调用成功，签名见[阿里签名](https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/use-cases/call-apis)规则:

https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/use-cases/call-apis

返回数据结构:

名称	类型	是否必须返回	备注
code	number	是	10000:请求成功
msg	string	是	
result	Dictionary	否	成功后返回

result:

名称	类型	是否必须返回	备注
token	string	是	用于查询结果接口

返回示例(成功状态):

```
{
    "code":10000,
    "msg":"",
    "result":{"token":"xxx"}
}
```

返回示例(失败状态):

```
{
    "code":40001,
    "msg":"ParmNotRight"
}
```

文档转换_v2

异步url:

```
https://pdf2doc.market.alicloudapi.com/v2/convert_async
```

同步url:

```
https://pdf2doc.market.alicloudapi.com/v2/convert_sync
```

HTTP方式： POST

Header中的Content-Type传入application/json

Body是JSON格式，支持以下2种输入源文件的方法：

方法1: 文件url，**最大1500M**:

```
{"input": "http://xxx.pdf", "type": "docx"}
```

方法2: 文件Base64字符串，Base64字符串**最大8M**:

```
{"input": "base64字符串", "type": "docx"}
```

必须签名才能调用成功，签名见[阿里签名]规则:

https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/use-cases/call-apis

支持转换为以下文件格式:

类型	扩展名（type取值）
微软Office文档	doc, docx, pptx, xlsx
开放版式文档	ofd
文本文件	txt, rtf

自定义参数:

参数	类型	备注	默认值
type	string	小写，输出文件类型，例如docx	必须发送
输出文件相关			
ocr	int	对于扫描的PDF，是否做OCR: 0：不做OCR 1：自动OCR 2：强力OCR：针对有些加密或编码不正确导致的乱码，叠字，未OCR文字等问题 **默认1，建议：不需要OCR选择0，一般情况选择1，出现乱码等问题选择2**	1
language	int	OCR识别语言选项，默认2简体中文: 1：英语 2：简体中文 3：繁体中文 4：法语 5：德语 6：意大利语 7：俄语 8：日文	2

		9：韩文 10：西班牙语 11：葡萄牙语 12：丹麦语 13：荷兰语 14：芬兰语 15：挪威语 16：瑞典语 17：土耳其语	
pageindexes	string	要转换的PDF页数，默认空全部页，例如：1,3,5-7就是1,3,5,6,7共5页	空
outfilename	string	生成的文件的文件名，默认随机	空
输出Office文件相关			
wordnoimage	int	如果转为Word文件，默认0：需要图片；1：不需要图片	0
wordabsolutelayout	int	如果转为Word文件，默认0：流式布局；1：绝对布局（位置精准，浏览方便，但是编辑方式非流式不利于编辑）	0
excelonesheet	int	如果转为Excel文件，默认0：PDF特定页数以内为一个工作表，否则每页一个工作表；1：一个工作表（如果PDF页数太多，有失败可能）；2：每页一个工作表	0
输出文本文件相关			
textnoformat	int	如果转为Txt文本，默认0： 保持原有布局，1：无需布局	0
textperpage	int	如果转为Txt文本，默认0：所有页转为单个txt文件，1：每一页转为一个txt文件，并且打包为一个zip包	0
其他			
imagepdfocroption	int	如果是扫描版PDF，根据ocr参数做ocr。 如果是非扫描版PDF，根据该值做ocr。 默认0：不启用； 1-2：启用，非扫描版对应ocr：1-2 100：启用，非扫描版不做ocr 比如ocr传2，imagepdfocroption传1 那么：扫描版PDF用ocr：2做ocr 非扫描版用ocr：1做ocr	0
password	string	PDF文件的密码，没有密码可以不传或传空	空
callbackurl	string	回调URL，转换结束后，会回调该URL，详细见 **回调URL**	空

请求示例:

- 例1: 把PDF文件转为word

```
{"input": "http://xxx.pdf", "type": "docx"}
```

- 例2: 把PDF每一页转为一个txt文件

```
{"input": "http://xxx.pdf", "type": "txt", "textperpage": 1}
```

- 例3: 只把PDF的1，3，5页转为Excel，并放在一个工作表里

```
{"input": "http://xxx.pdf", "type": "xlsx", "excelonesheet": 1, "pageindexes": "1,3,5"}
```

返回数据结构【异步】:

名称	类型	是否必须返回	备注
code	number	是	10000:请求成功
msg	string	是	
result	Dictionary	否	成功后返回

result:

名称	类型	是否必须返回	备注
token	string	是	用于查询结果接口

返回示例(成功状态)【异步】:

```
{
   "code":10000,
   "msg":"",
   "result":{"token":"xxx"}
}
```

返回示例(失败状态)【异步】:

```
{
   "code":40001,
   "msg":"ParmNotRight"
}
```

同步调用

同步调用的最大返回时间是60秒，如果60秒内转换结束则直接返回结果。否则会返回token，之后和异步方式一样可以调用查询结果接口查询该token的转换结果。所以同步调用如果传入的文件过大，无法保证在60秒内结束，则转为异步流程。

返回数据结构【同步】:

名称	类型	是否必须返回	备注
code	number	是	10000:请求成功
msg	string	是	
token	string	是	请求的token
result	Dictionary	否	成功后返回

result:

名称	含义	类型	是否必须返回	备注
fileurl	输出文件地址	string	否（status为Done时返回）	转换出来的文件地址，http和https都支持
count	输入PDF的页数	integer	否（status为Done时返回）	PDF页面总数
filesize	文件大小	integer	否（status为Done时返回）	输出文件大小
status	状态	string	是	Done：转换成功 Failed：转换失败

返回示例(成功状态)【同步】:

```
{
      "code": 10000,
      "msg": "",
```

```
    "result": {
        "count": 1,
        "filesize": 17747,
        "fileurl": "https://file.duhuitech.com/o/xxx/xxx.docx",
        "status": "Done"
    },
    "token": "xxx"
}
```


返回示例(超时状态)【同步】：


```
{
    "code": 40500,
    "msg": "Timeout, query token later",
    "token": "xxx"
}
```

查询结果

请求参数:

参数	类型	备注	是否必须发送
token	string	调用转换接口拿到的token	是

请求示例:

```
https://api.duhuitech.com/q?token=xxx
```

无需签名，无调用次数限制

由于转换需要时间，文件越大页数越多，转换越久，故需要**轮询**查询接口来获得结果。查询频率可以是1s一次，也可以更长一些。**查询后先看status，如果是Done或Failed，则转换结束，停止轮询。如果是Doing或Pending，则继续轮询。**

返回数据结构:

名称	类型	是否必须返回	备注
code	number	是	10000:请求成功
msg	string	是	
token	string	是	请求的token
result	Dictionary	否	成功后返回

result:

名称	含义	类型	是否必须返回	备注
status	状态	string	是	Pending：还未开始 Doing：正在转换 Done：转换成功 Failed：转换失败
progress	进度	number	否（status为Doing时返回）	范围：0.00 - 1.00， 比如0.88表示88%
fileurl	输出文件地址	string	否（status为Done时返回）	转换出来的文件地址，http和https都支持
count	输入PDF的页数	integer	否（status为Done时返回）	PDF页面总数
filesize	文件大小	integer	否（status为Done时返回）	输出文件大小
reason	失败原因	string	否（status为Failed时可能返回）	转换失败的原因

返回示例(成功状态):


```
{
	"code":10000,
	"msg":"",
	"token":"xxx",
	"result":
	{
		"progress":0.02,
		"status":"Doing"
	}
}
```

```
{
	"code":10000,
	"msg":"",
	"token":"xxx",
	"result":
	{
		"status":"Done",
		"fileurl":"https://file.duhuitech.com/o/xxx/xxx.docx",
		"filesize":17747,
		"count":1
	}
}
```


返回示例(失败状态):


```
{
	"code":40000,
	"msg":"No such token"
}
```

备注：

- 文件url方式支持文件大小 *1500M*。
- 转换完成后，在 *1小时* 内下载文件。

回调URL：

用途： 客户可以自行部署服务器，系统转换结束后会调用客户提供的回调URL，直接发送转换结果，从而无需再轮询查询结果。

当设置了回调URL，转换结束后（无论成功失败），系统都会尝试调用该URL，具体如下：

以POST方式调用该URL，Header头中Content-Type: application/json

Body为JSON格式，内容和查询结果的结果相同，例如：

```
{
    "code":10000,
    "msg":"",
    "token":"xxx",
    "result":
    {
        "status":"Done",
        "fileurl":"https://file.duhuitech.com/o/xxx/xxx.docx",
        "filesize":17747,
        "count":1
    }
}
```

服务端收到该POST后需在10秒内返回HTTP STATUS CODE 200，视为调用成功，否则系统认为回调失败，会再次尝试。规则如下：

系统共计最多会调用3次回调URL，如果第一次失败，则等待3秒后尝试第二次，如果第二次失败，则等待5秒后尝试第三次，如果第三次失败，则不再尝试。

回调URL超时时间10秒。

阿里云独有部分：

支持从阿里云OSS内网直接下载文件，目前支持的是上海地区的阿里云OSS内网：

oss-cn-shanghai-internal.aliyuncs.com

文档转换GET或多张图片转换POST里的url地址包含上述域名则自动支持

错误码表：

返回的code如果是10000，代表成功，其余是失败

JSON里返回的code	错误信息
40000	通用错误
40001	参数错误
40002	参数不符合规范
40500	同步调用超时

附录：阿里签名方式

参考链接：

https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/use-cases/call-apis

在调用API商品时，首先您需要了解采用哪种API认证方式，云市场API商品的认证方式主要有以下两种方式。两种方式可同时使用，您可以根据不同情况来选择。

- 简单身份认证（AppCode）
- 签名认证

简单身份认证（AppCode）

简单认证（AppCode）调用API，有两种方式，一种是将AppCode放在Header中进行调用，一种是将AppCode放在Query参数中进行调用。

方式一：将AppCode放在Header中

在请求Header中添加一个Authorization参数。

Authorization字段的值的格式为APPCODE + 半角空格 + APPCODE值。格式如下：

Authorization:APPCODE AppCode值

示例：

Authorization:APPCODE 3F2504E04F8911D39A0C0305E82C3301

方式二：将AppCode放在Query中

在请求Query中添加AppCode参数（同时支持appcode , appCode , APPCODE , APPCode四种写法）。

AppCode参数的值为AppCode的值。

示例：

http://www.aliyum.com?AppCode=3F2504E04F8911D39A0C0305E82C3301

参考链接：https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/user-guide/call-an-api-operation-by-using-an-appcode

签名认证

比较复杂，推荐用阿里自己的SDK来调用，参考链接：https://help.aliyun.com/zh/api-gateway/traditional-api-gateway/user-guide/use-digest-authentication-to-call-an-api