

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 query 接口”。详细见：**查询 QUERY**
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**

关于同步调用

从 v2 版本的 API 开始，提供同步调用。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 query 接口”查询该 token 的转换结果。

v2 和 v1 版本 API 的区别

- v2 版本的 API 都是 POST JSON，支持输入文件 url；文件 Base64 字符串
- v2 版本的 API 支持同步和异步调用，v1 版本都是异步
- v1 版本的 API 支持 2 种转换方式：输入文件 url；直接 POST 文件二进制
- 其他参数 v1 和 v2 版本一致，之后的新参数只增加到 v2，因此**优先推荐使用 v2 版本 API**
- v1 和 v2 调用的“查询 query 接口”是一样的

调用转换 API 需要签名，详细见文档附录：阿里签名。**调用查询 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**

以下是 v2 接口，v1 接口见附录：**v1 接口**

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

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，需转换为的文件类型，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：做 OCR 2：强制 OCR，如果出现乱码，用 2 以上的选项 3：使用图片转文档引擎做转换，不 清除背景 4：使用图片转文档引擎做转换，清 除背景 5：使用图片转文档引擎做转换，智 能清除背景 默认 1，如果 2 以上无论是否扫描 版都会做 OCR	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 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	2
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； 1-5：启用，非扫描版对应 ocr：1-5 100：启用，非扫描版不做 ocr 比如 ocr 传 5，imagepdfocroption 传 1 那么：扫描版 PDF 用 ocr：5 做 ocr 非扫描版用 ocr：1 做 ocr	0
table	int	如果 ocr 取值 3（包含）以上，也就是使用图片转文档引擎做转换的情况下，转为 Word 或 PPT 是否识别表格。 默认 0：否 1：只识别表格 2：识别表格和下划线	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	含义	类型及范围	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	含义	类型及范围	是否必须返回	备注
token		string	是	用于 query 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"
}
```

返回数据结构【同步】：

名称	含义	类型及范围	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	类型及范围	是否必须返回	备注
fileurl	导出文件地址	string	否（status 为 Done 时返回）	转换出来的文件地址，http 和 https 都支持

名称	含义	类型及范围	是否必须返回	备注
count	总页数	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）	导出文件大小
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.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 QUERY

请求参数:

参数	类型及范围	备注	是否必须发送
token	string		是

请求示例:

https://api.duhuitech.com/q?token=7b799e09e0838919d3ae63d0566683a2

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	含义	类型及范围	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	类型及范围	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 (0.00 - 1.00)	否 (status 为 Doing 时返回)	比如 0.88 表示 88%
fileurl	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
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/7b799e09e0838919d3ae63d0566683a2/cc9c7f1e-03f8-4742-bc37-aab9da191c26.docx"
  }
}
```

返回示例(失败状态):

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

注意：

- 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 Query。

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 Query 的结果相同，例如：

```
{"code":10000,"msg":"","token":"xxx","result":{"status":"Done","fileurl":"https://file.duhuitech.com/o/7b799e09e0838919d3ae63d0566683a2/cc9c7f1e-03f8-4742-bc37-aab9da191c26.docx"}}
```

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：

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

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>

附录：v1 接口（如已使用 v2，无需 v1 接口）

文档转换 GET

将单个 PDF 下载链接转换为其他格式，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	类型及范围	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，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：做 OCR 2：强制 OCR，如果出现乱码，用 2 以上的选项 3：使用图片转文档引擎做转换，不清除背景 4：使用图片转文档引擎做转换，清除背景 5：使用图片转文档引擎做转换，智能清除背景 默认 1，如果 2 以上无论是否扫描版都会做 OCR	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	否

参数	类型及范围	备注	是否必须发送
		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	
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
table	int	如果 ocr 取值 3（包含）以上，也就是使用图片转文档引擎做转换的情况下，转为 Word 或 PPT 是否识别表格。 默认 0: 否 1: 只识别表格 2: 识别表格和下划线	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； 1-5: 启用，非扫描版对应 ocr: 1-5 100: 启用，非扫描版不做 ocr 比如 ocr 传 5，imagepdfocroption 传 1 那么: 扫描版 PDF 用 ocr: 5 做 ocr	否

参数	类型及范围	备注	是否必须发送
		非扫描版用 ocr: 1 做 ocr	
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，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。ocr 可取值 0, 1
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	含义	类型及范围	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	含义	类型及范围	是否必须返回	备注
token		string	是	用于 query 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

直接将单个文档 POST 到服务器，大小限制 8M

请求参数:

参数	类型及范围	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：做 OCR 2：强制 OCR，如果出现乱码，用 2 以上的选项 3：使用图片转文档引擎做转换，不清除背景 4：使用图片转文档引擎做转换，清除背景 5：使用图片转文档引擎做转换，智能清除背景 默认 1，如果 2 以上无论是否扫描版都会做 OCR	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	否

参数	类型及范围	备注	是否必须发送
		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	
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
table	int	如果 ocr 取值 3（包含）以上，也就是使用图片转文档引擎做转换的情况下，转为 Word 或 PPT 是否识别表格。 默认 0：否 1：只识别表格 2：识别表格和下划线	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； 1-5：启用，非扫描版对应 ocr：1-5 100：启用，非扫描版不做 ocr 比如 ocr 传 5，imagepdfocroption 传 1 那么：扫描版 PDF 用 ocr：5 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

参数	类型及范围	备注	是否必须发送
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`。ocr 可取值 0，1
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	含义	类型及范围	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	含义	类型及范围	是否必须返回	备注
token		string	是	用于 query 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"  
  }
```