

PDF 转文档（阿里）概述：

在线试用：

<https://try.dhconvert.com/>

小程序搜索：

度慧文档转换

2 种转换方式：

单文档转换，文档是一个下载链接，用 HTTP GET 方式，见：[文档转换 GET](#)

单文档转换，文档 POST 到服务器，用 HTTP POST 方式，见：[文档转换 POST](#)

查询转换结果：

由于转换需要时间，文件越大页数越多，转换越久，故系统采用异步的方式获得转换结果。有 2 种方式：

1. 调用转换接口后获得 token，再用 HTTP GET 方式，轮询“查询 query 接口”获得结果。

详细见：[查询 QUERY](#)

2. 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](#)

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

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	否

参数	类型及范围	备注	是否必须发送
		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	
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
imagepdfocroption	int	如果是扫描版 PDF, 根据 ocr 参数做 ocr。	否

参数	类型及范围	备注	是否必须发送
		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	
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
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.ali.duhuitech.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

签名见阿里签名规则。type 可取值：doc，docx，pptx，xlsx，rtf，txt。ocr 可取值 0，1

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	

名称	含义	类型及范围	是否必须返回	备注
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):

<pre>{ "code":10000, "msg": "", "result":{"token":"xxx"} }</pre>
--

返回示例(失败状态):

<pre>{ "code":40001, "msg":"ParmNotRight" }</pre>

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	否

参数	类型及范围	备注	是否必须发送
		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	
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
imagepdfocroption	int	如果是扫描版 PDF, 根据 ocr 参数做 ocr, 否则不做 ocr。默认 0: 不启用; 1: 启用	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
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	类型及范围	备注	是否必须发送
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.ali.duhuitech.com/v1/convert`
Header 中的 Content-Type 必须是 multipart/form-data

签名见阿里签名规则。type 可取值：doc，docx，pptx，xlsx，rtf，txt。ocr 可取值 0，1

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"
}
```

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：

- 上传文件大小不能超过 1000M。
- 转换完成后，在 2 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