百度智能云

TextMind通用文档解析

大模型应用的加速引擎

演讲人：苏崔聪

百度智能云

目录

百度智能云

01 文档解析相关背景

什么是文档解析

百度智能云

文档

文档解析

结构化数据

- 文本内容+位置信息
- 表格内容+结构信息
- 文本属性信息
- 标题目录信息
- 视觉图片信息

为什么要做文档解析

百度智能云

文档解析的重要性不仅在于解决繁杂的文档信息提取问题，**更在于它能为各类业务场景提供核心支持，成为文档信息引擎，推动文档相关的业务信息化和智能化的进程**

通用文档解析的难点

百度智能云

文档格式多

docx、doc、pptx、ppt、txt、pdf、ofd、jpg、xls等数十种常见文档类型

版面内容复杂多样

标题、段落、表格、公式、印章、插图等数十种版面元素，图文混排、多栏等

文档质量不可控

模糊、倾斜、手写体、水印遮挡、版本老旧、内容格式规范化程度低等

信息多模态

文本、表格、插图、统计图、二维码等

大文档挑战

百兆千页级别的大文档对解析系统的稳定性和性能挑战很大

信息非结构化

内容分散、无先验结构、不同模态无关联等

下游任务需求多样

RAG、文档查重、文档比对、合同审核等

实时和大规模处理

场景对实时性、稳定性有高要求，处理大规模的文档解析任务

百度智能云

02 TextMind通用文档解析介绍

TextMind通用文档解析主要核心技术

百度智能云

文档内容结构化引擎

- 阅读顺序重建
- 图文内容关联
- 图 / 表标题关联
- 标题层级划分
- 标题段落树构建
- 文档内容切分

文档内容解析引擎

表格结构识别

- 表格类型识别
- 表格线检测
- 表格角点检测
- 合并单元格检测
- 单元格关系构建

文档版式内容识别

- 排版布局识别
- 印刷文字识别
- 手写文字识别
- 公式识别
- 标题识别
- 表格识别
- 统计图表识别
- 其他元素识别…

文档预处理引擎

- 文档格式转换
- 文档主体矫正
- 文档图像增强
- 印章水印擦除

TextMind通用文档解析概览

百度智能云

文档格式全面

支持docx、pdf、图片、xlsx等17种格式

解析功能丰富

支持文字解析、表格解析、标题层级、阅读顺序等

识别准确率高

版式分析、表格解析等综合识别准确率90%+

解析速度快

支持300M+2K页级长文档，平均解析效率≤200ms/页

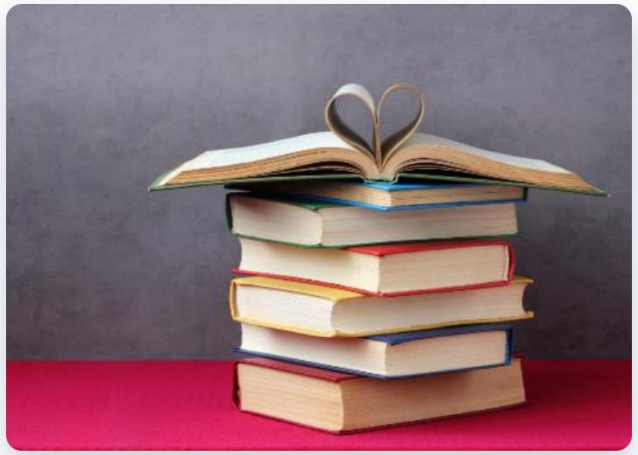
功能演示

https://ai.baidu.com/tech/ocr/doc_parser

TextMind通用文档解析

百度智能云

兼容17种主流文档格式，覆盖企业日常文档



文档格式覆盖全

版式：pdf、jpg、jpeg、png、bmp、tif、tiff、ofd、ppt、pptx

流式：doc、docx、txt、xls、xlsx、 wps、html

支持长文档解析

支持百兆千页级别的长文档文件解析

支持文档大小上限：300M

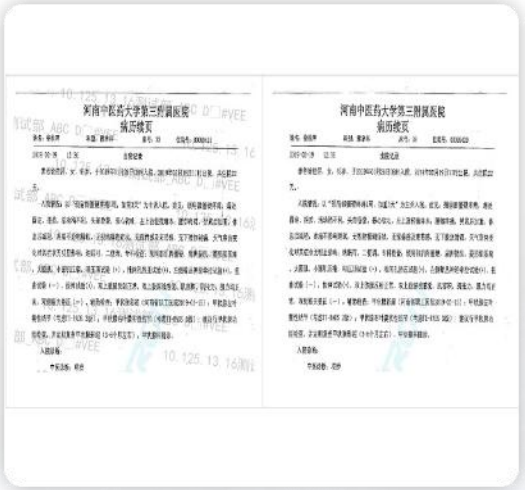
支持文档页数上限：2000页

TextMind通用文档解析

百度智能云

强大的预处理降噪能力，降低低质量图像干扰

原图



.....

印章擦除

减少因印章遮盖导致的比对结果误召，节约人工审核成本

水印擦除

扫描件文档比对、抽取场景，减少水印影响，提升模型准确率

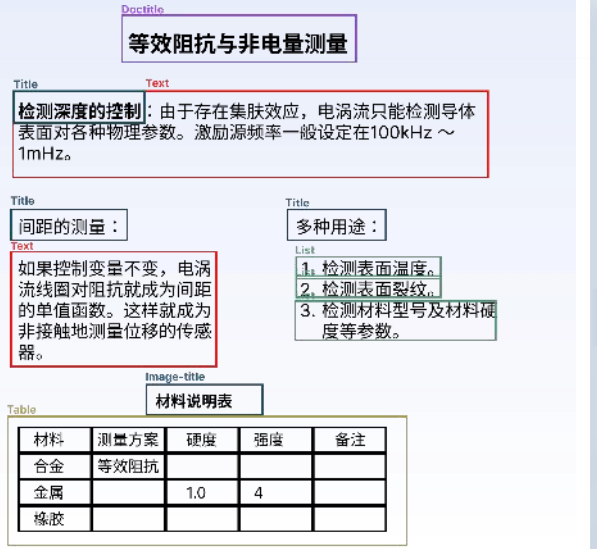
旋转检测与自动修正

避免因旋转角度问题影响OCR识别，用户提交文档时无需对旋转问题进行人工核查，支持0-360度倾斜自动矫正

TextMind通用文档解析

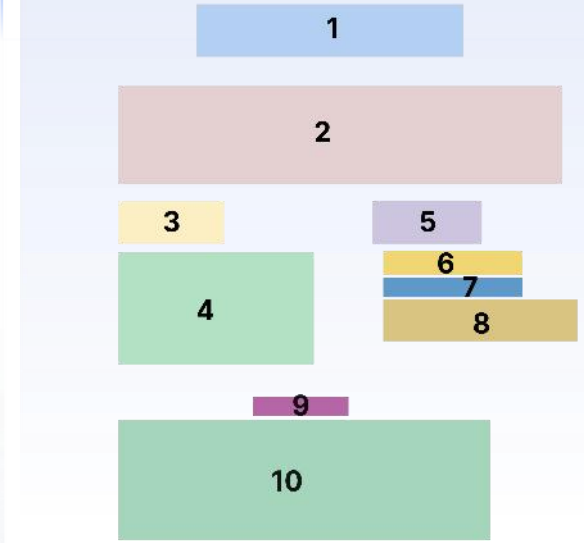
百度智能云

复杂版面的内容解析



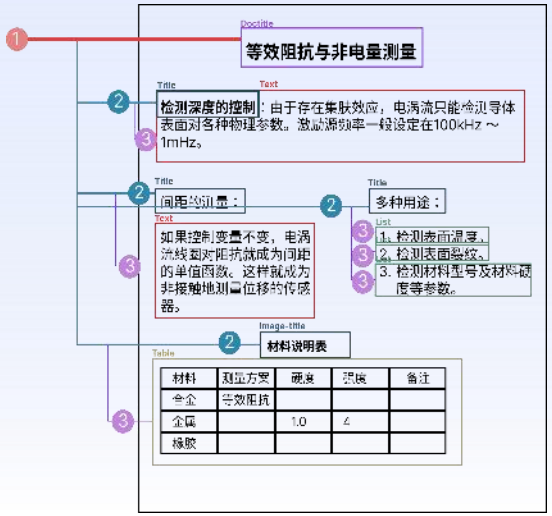
多版面要素识别

识别文档页面**标题、段落、图片、表格、页眉页脚**等元素



阅读顺序

将页面内元素
按阅读顺序重排



布局要素关联

标题层级关系构建、图/表标题关联、跨页/栏要素关联等

TextMind通用文档解析

百度智能云

核心能力梳理与说明

	功能说明	结构化形式	适用场景
版式分析	❖ 支持表格、图表标题、段落、插图、统计图印章、公式、等13种版面类型 ❖ 支持文档内的图片元素通过链接的形式返回	❖ 元素位置坐标+文本内容 ❖ 统计图表通过结构化形式展示 ❖ 图片内容保存为链接，以及图文关联 ❖ 公式内容为latex形式展示	❖ RAG ❖ 多模态 ❖ 文档比对 ❖ 文档审核
表格解析	❖ 支持有线表、无线表的解析结构化 ❖ 支持复杂背景、彩色背景表格解析结构化	❖ 单元格坐标+文本内容 ❖ 表格单元格空间位置关系矩阵构建 ❖ Markdown形式	❖ RAG表格问答 ❖ 信息抽取 ❖ 信息结构化
阅读顺序	❖ 支持单栏、多栏文档、图文混排的阅读顺序构建	❖ 按符合人类阅读习惯的构建文本段落输出	❖ RAG ❖ 文档比对 ❖ 文档翻译
标题层级	❖ 支持扫描件、电子件等文档最多7级标题的解析 ❖ 支持最多7级标题段落层级树的构建	❖ 标题层级通过字段标识，位置信息+文本内容 ❖ 标题段落层级树通过版式元素节点之间的父子关系构建	❖ RAG ❖ 合同审查 ❖ 智慧招采
内容切分	❖ 支持按照标题段落层级树内聚切分 ❖ 按chunk大小、标点切分 ❖ 以上方法混用	❖ 以切分之后的chunk展示 ❖ 每个chunk所属的标题信息	❖ RAG ❖ 智慧招采

TextMind通用文档解析

百度智能云

核心能力展示-复杂表格解析

项　目	年初人口	年内增加	其　中			年内减少	其　中			年末人口	其　中		年末人口		自然增长率‰	年末总户数
			迁入	出生	出生率‰		迁出	出生	出生率‰		男	女	农业人口	流动人口		
团场	16387	304	202	102		538	436	102		16153	8302	7851		260		5849
汉族	16071	302	202	100		538	436	102		15835	8162	7673				
维吾尔族	6									6	3	3				
回族	288	2		2						290	125	165				
蒙古族	2									2	1	1				
其他少数民族	20									20	11	9				
合计	5701	84	53		31	199	161	38		5586	2777	2809				

项目	年初人口	年内增加	其中	其中	其中	年内减少	其中	其中	其中	年末人口	其中	其中	年末人口	年末人口	自然增长率‰	年末总户数
项目	年初人口	年内增加	迁入	出生	出生率‰	年内减少	迁出	出生	出生率‰	年末人口	男	女	农业人口	流动人口	自然增长率‰	年末总户数
团场	16387	304	202	102		538	436	102		16153	8302	7851		260		5849
汉族	16071	302	202	100		538	436	102		15835	8162	7673				
维吾尔族	6									6	3	3				
回族	288	2		2						290	125	165				
蒙古族	2									2	1	1				
其他少数民族	20									20	11	9				
合计	5701	84	53		31	199	161	38		5586	2777	2809				

4.1.3 新兴技术加速上车，产业发展要“广度”也要“深度”

亿欧智库

◆ 座舱搭配基础硬件设施差别越来越小，新技术和产品功能将不断更新并上车。用户的需求和功能随着汽车智能化水平提升和自动驾驶技术的进步而不断变化，不同的产业发展阶段和时间段，用户的刚需都会有所不同。因此一些现在不会马上大规模应用的功能也会“预埋”上车。

◆ 除了新技术的种类和产品将不断丰富，核心技术的深度研发也将是智能座舱极为重要的一部分。核心技术和零部件可以更为有效地提高产品竞争力。例如，大算力芯片、算法、AR技术等。

技术不断创新+预埋，买“期货”将成为常态

- 未来，更多的创新技术将主要出现在安全辅助、人机交互、生活娱乐三个维度，体现为不断上车的“黑科技”，如全息影像下的元宇宙座舱、智能太空舱……
- 智能座舱还将进一步推动技术预埋，从现在大多出现在高端车型上的方案转变为未来的主流方案。例如**HUD**在L0-L1阶段并不是驾驶员的刚需；L2阶段开始，导航、通讯、等功能被纳入；当车辆发展至高等级自动驾驶阶段，HUD将成为标配。**激光雷达**也同样更适用于高等级自动驾驶，很多汽车的“技术发烧友”会预先购买安装了激光雷达的车型，认为购买了一种技术“期货”，将在未来技术发展成熟时进行兑现。

智能座舱将在安全辅助、人机交互、生活娱乐三维度下向更广泛的功能展开

安全辅助	人机交互		生活娱乐	
➢ 基础车况信息显示	➢ 驾驶员行为追踪	➢ 全息影像	➢ 行为追踪	➢ 视频会议
➢ 外界环境融合显示	➢ 眼球追踪	➢ 副驾视频影音	➢ 眼球追踪	➢ 游戏娱乐
➢ 人、车标注	➢ 语音交互	➢ 后排视频影音	➢ 语音交互	➢ AR广告
➢ 空气悬架	➢ 手势交互	➢ 可变结构座椅	➢ 手势交互	➢ 全息影像
➢ ……	➢ 座舱香氛系统	➢ 声音复刻	➢ 电话接听	➢ 智能表面
	➢ ……	➢ ……	➢ 视频影音	➢ ……

核心技术继续深耕，核心技术供应商长板优势凸显

多样化的场景使用和产品要求，对于核心技术的开发和使用都提出了更高的要求。**车企除了让功能变得“更多”之外，还会将目光放在如何让核心功能“更好用”上。**

- **算力合理利用与分配**：随着功能原来越多，各种传感器收集的大量车辆周边信息需要保证能够实时呈现，同时各种休闲娱乐功能也能良好进行。这就需要车机硬件性能更高，同时在开发过程中不占用过多算力，实现数据的最优化渲染呈现。

辅助驾驶功能　娱乐功能　算法

- **语音交互精准度提升**：例如，语音交互技术的精准度已经从2011年的60%提升至2021年的98%，未来还将进一步深耕语音识别的精准度、稳定性以及个性化。

测试场景下，准确率提升　个性化、情感化升级，如声音复刻　全场景下稳定性提升，拾音降噪，音质音效提升

- **高精度、场景化出行体验将推广**：未来座舱将进一步融合驾驶场景，从地图导航到高精度定位，再到场景化出行服务，融合互联网、智能驾驶和智能座舱的产品将受到市场认可，如美行科技车载OS等。

来源：思必驰、美行科技、亿欧智库

45

TextMind通用文档解析

核心能力展示-标题层级树构建

1.1 语言模型的发展历程

图 1.2 基于任务求解能力的四代语言模型的演化过程（图片来源：[10]）

且不容易并行训练，这些缺点限制了早期预训练模型（如 ELMo）的性能。在 2017 年，谷歌提出了基于自注意力机制（Self-Attention）的 Transformer 模型 [12]，通过自注意力机制建模长程序列关系。Transformer 的一个主要优势就是其模型设计对于硬件非常友好，可以通过 GPU 或者 TPU 进行加速训练，这为研发大语言模型提供了可并行优化的神经网络架构。基于 Transformer 架构，谷歌进一步提出了预训练语言模型 BERT [13]，采用了仅有编码器的 Transformer 架构，并通过在大规模无标注数据上使用专门设计的预训练任务来学习双向语言模型。在同期，OpenAI 也迅速采纳了 Transformer 架构，将其用于 GPT-1 [14] 的训练。与 BERT 模型不同的是，GPT-1 采用了仅有解码器的 Transformer 架构，以及基于下一个词元预测的预训练任务进行模型的训练。一般来说，编码器架构被认为更适合去解决自然语言理解任务（如完形填空等），而解码器架构更适合解决自然语言生成任务（如文本摘要等）。以 ELMo、BERT、GPT-1 为代表的预训练语言模型确立了“预训练-微调”这一任务求解范式。其中，预训练阶段旨在通过大规模无标注文本建立模型的基础能力，而微调阶段则使用有标注数据对于模型进行特定任务的适配，从而更好地解决下游的自然语言处理任务。

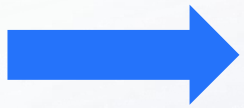
- 大语言模型（Large Language Model, LLM）。研究人员发现，通过规模扩展（如增加模型参数规模或数据规模）通常会带来下游任务的模型性能提升，这种现象通常被称为“扩展法则”（Scaling Law）[15]。一些研究工作尝试训练更大的预训练语言模型（例如 175B 参数的 GPT-3 和 540B 参数的 PaLM）来探索扩展语言模型所带来的性能极限。这些大规模的预训练语言模型在解决复杂任务时表现出了与小型预训练语言模型（例如 330M 参数的 BERT 和 1.5B 参数的 GPT-2）不同的行为。例如，GPT-3 可以通过“上下文学习”（In-Context Learning, ICL）的方式来利用少样本数据解决下游任务，而 GPT-2 则不具备这一能力。这种大模型具有但小模型不具有的能力通常被称为“涌现能力”（Emergent Abilities）。为了区

4

1.2 大语言模型的能力特点

别这一能力上的差异，学术界将这些大型预训练语言模型命名为“大语言模型”[1]（Large Language Model, LLM）[16]。作为大语言模型的一个代表性应用，ChatGPT 将 GPT 系列大语言模型适配到对话任务中，展现出令人震撼的人机对话能力，一经上线就取得了社会的广泛关注。ChatGPT 发布后，与大语言模型相关的 arXiv 论文数量迅速增长（如图 1.1 所示），这一研究方向受到了学术界的高度关注。

通过回顾上述发展历程，可以看到语言模型并不是一个新的技术概念，而是历经了长期的发展历程。早期的语言模型主要面向自然语言的建模和生成任务，而最新的语言模型（如 GPT-4）则侧重于复杂任务的求解。从语言建模到任务求解，这是人工智能科学思维的一次重要跃升，是理解语言模型前沿进展的关键所在。图 1.2 通过任务求解能力的角度对比了四代语言模型所表现出的能力优势与局限性。首先，早期的统计语言模型主要被用于（或辅助用于）解决一些特定任务，主要以信息检索、文本分类、语音识别等传统任务为主。随后，神经语言模型



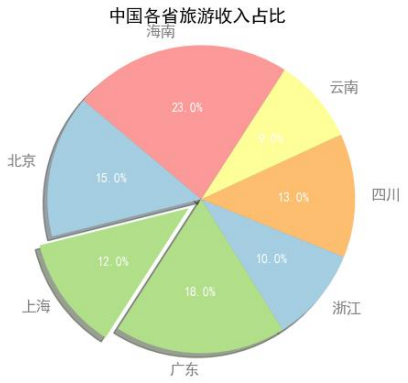
目录

TextMind通用文档解析

百度智能云

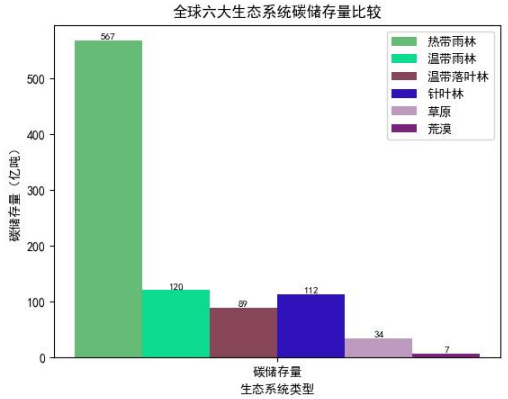
核心能力展示-图表多模态解析，业界首推

饼图



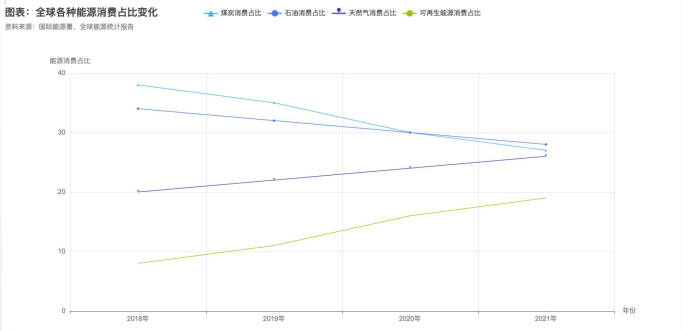

```
{
    "title": "中国各省旅游收入占比",
    "source": "国家旅游局统计数据",
    "x_title": "省份",
    "y_title": "旅游收入占比",
    "values": {
        "北京": "15%",
        "上海": "12%",
        "广东": "18%",
        "浙江": "10%",
        "四川": "13%",
        "云南": "9%",
        "海南": "23%"
    }
}
```


直方图



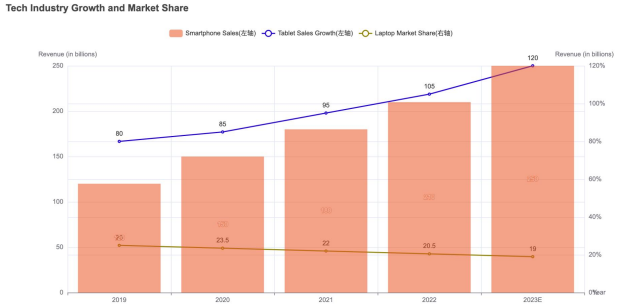

```
{
    "title": "全球六大生态系统碳储存量比较",
    "x_title": "生态系统类型",
    "y_title": "碳储存量（亿吨）",
    "values": {
        "热带雨林": {
            "碳储存量": "567"
        },
        "温带雨林": {
            "碳储存量": "120"
        },
        "温带落叶林": {
            "碳储存量": "89"
        },
        "针叶林": {
            "碳储存量": "112"
        },
        "草原": {
            "碳储存量": "34"
        },
        "荒漠": {
            "碳储存量": "7"
        }
    }
}
```


折线图




```
{
    "title": "图表：全球各种能源消费占比变化",
    "source": "资料来源：国际能源署，全球能源统计报告",
    "x_title": "年份",
    "y_title": "能源消费占比",
    "values": {
        "煤炭消费占比": {
            "2018年": "38%",
            "2019年": "35%",
            "2020年": "30%",
            "2021年": "27%"
        },
        "石油消费占比": {
            "2018年": "34%",
            "2019年": "32%",
            "2020年": "30%",
            "2021年": "28%"
        },
        "天然气消费占比": {
            "2018年": "20%",
            "2019年": "22%",
            "2020年": "24%",
            "2021年": "26%"
        },
        "可再生能源消费占比": {
            "2018年": "8%",
            "2019年": "11%",
            "2020年": "16%",
            "2021年": "19%"
        }
    }
}
```


折线直方图




```
{
    "title": "Tech Industry Growth and Market Share",
    "source": "Tech Market Analytics Report",
    "x_title": "Year",
    "y_title": [
        "Revenue (in billions)",
        "Market Share (%)"
    ],
    "values": {
        "Smartphone Sales(左轴)": {
            "2019": "120",
            "2020": "150",
            "2021": "180",
            "2022": "210",
            "2023E": "250"
        },
        "Laptop Market Share(右轴)": {
            "2019": "25.0%",
            "2020": "23.5%",
            "2021": "22.0%",
            "2022": "20.5%",
            "2023E": "19.0%"
        },
        "Tablet Sales Growth(左轴)": {
            "2019": "80",
            "2020": "85",
            "2021": "95",
            "2022": "105",
            "2023E": "120"
        }
    }
}
```

TextMind通用文档解析

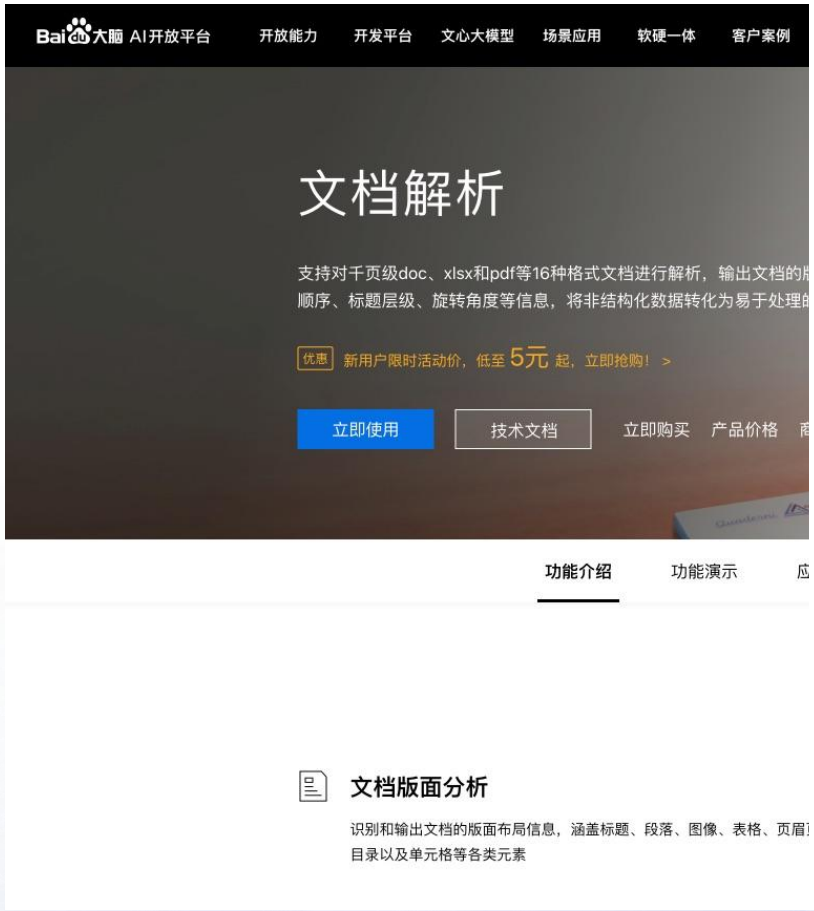
百度智能云

业内同类产品比较

百度优势	百度TextMind文档解析	某云大厂	某OCR垂类厂商	某SaaS厂商
文档格式全面	doc、pdf、图片、xlsx、ppt、ofd、wps、html等17种格式	doc、pdf和图片	pdf和图片	pdf和图片
解析功能丰富	✓ 印章擦除 ✓ 水印擦除 ✓ 角度矫正 — 预处理降噪 ✓ 文字识别 ✓ 表格解析 ✓ 版式分析 — 基础解析 ✓ 标题层级 ✓ 图表解析 ✓ 阅读顺序 ✓ 内容切分 — RAG核心	部分满足	部分满足	部分满足
支持百兆/千页级长文档	单文档页数上限2000页 单文档大小上限300M	单文档页数上限1000页 单文档上限10M	单文档页数上限1页 单文档上限10M	单文档上限1页 单文档上限10M
识别准确率高	版式-92% 表格-90%	版式-83% 表格-62%	版式-85% 表格-72%	版式-70% 表格-68%
解析速度快	TextMind>某OCR垂类厂商>某云大厂>某SaaS厂商			

TextMind通用文档解析使用方式

公有云API



文档解析

更新时间：2024-11-06

接口描述

文档解析支持对doc、pdf、图片、xlsx等16种格式文档进行解析，输出文档的版面、表格、阅读顺序、标题层级、旋转角度等信息，可返回Markdown格式内容，将非结构化数据转化为易于处理的结构化数据，识别准确率可达 90% 以上。

文档解析API服务为**异步接口**，需要先调用**提交请求接口**获取 task_id，然后调用**获取结果接口**进行结果轮询，建议提交请求后5～10 秒轮询。**提交请求接口**QPS为2，**获取结果接口**QPS为10。

提交请求接口

请求说明

请求示例

HTTP 方法：`POST`

请求URL：`https://aip.baidubce.com/rest/2.0/brain/online/v2/parser/task`

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

https://ai.baidu.com/ai-doc/OCR/Klxag8wiy

TextMind通用文档解析使用方式

百度智能云

私有化部署

支持的硬件、操作系统，**完善且灵活的国产化信创适配**

CPU	GPU	操作系统
• X86，Intel、海光等 • ARM，鲲鹏系列、飞腾系列	• NVIDIA T卡、A卡、V100、P卡、L卡系列，或与上述加速卡底层架构一致的其他型号均可支持 • 昆仑芯 R200、R480、RG800、K100、P800 • 华为 910B	• Ubuntu • CentOS • 麒麟

根据实际的硬件资源、操作系统情况，自由组合

百度智能云

03 TextMind通用文档解析应用

TextMind通用文档解析应用

百度智能云

支撑百度内部RAG知识管理问答产品

千帆
AppBuilder

智能客服
客悦

企业知识管理
甄知

……

https://ai.baidu.com/ai-doc/OCR/Klxag8wiy

TextMind通用文档解析应用

百度智能云

支撑行业智能文档分析应用

合同智能审查 | 贸易物流单证校验 | 智慧招采 | 金融业务审查 | 检书业务审查 | 一景一训

TextMind智能文档分析平台

TextMind通用文档解析

https://ai.baidu.com/ai-doc/OCR/Klxag8wiy

百度智能云

Thanks