**ETL4LM接口文档**

1. **获取分段解析结果**

|  |
| --- |
| 容器内默认端口为 8000，宿主机默认端口为8011 |

**请求说明**

**请求参数**

HTTP方法: POST

请求URL: http://host:8011/v1/etl4llm/predict

|  |
| --- |
| host=本地服务部署的IP地址 |

URL参数: 无

Header参数：Content-Type=application/json

Body参数：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **参数** | **默认值** | 是否必填 | **类型** | **说明** |
| filename | 无 | 是 | str | 文件名（必填） |
| b64\_data | 无 | 是 | List[str] | base64编码的文件数据 |
| url | 无 | 否 | str | 通过url指定解析的文件地址，与b64\_data必须且只可以填写一个。读取顺序：b64\_data > url |
| parameters | 无 | 否 | Dict | 请求url文件的请求头  {'headers': {}, 'ssl\_verify': True} |
| end\_pages | 无 | 否 | int | 转换页数限制，为空时默认转换文档全部 |
| force\_ocr | false | 否 | bool | 如果为true,则强制走模型进行 ocr；如果为false，则根据pdf 信息判断是否需要走 ocr，如果不需要，则不走 ocr 识别，改为提取 pdf 文字层。 |
| enable\_formula | true | 否 | bool | 如果为 true，会对一个页面进行 ocr + 公式检测+公式识别。如果为 false，则仅做 ocr。 |
| ocr\_sdk\_url | None | 否 | str | 如果填写则ocr 部分走 dataelem ocr，不填写走 paddleocr |
| for\_gradio | false | 否 | bool | 该接口，bool为false时，返回结果。 |

**返回说明**

|  |  |  |  |
| --- | --- | --- | --- |
| **参数** | **默认值** | **类型** | **说明** |
| status\_code | 无 | int | HTTP 状态码（200 表示成功） |
| partitions | true | List | 结构化内容块 |
| { |  |  |  |
| type | 无 | str | 内容类型（NarrativeText、Title、Image、Equation、Table） |
| text | 无 | str | 提取内容 |
| element\_id | 无 | str | 元素的唯一标识符 |
| metadata | 无 | dict | 解析过程的元信息 |
| { |  |  |  |
| bboxes | 无 | List[List[int]] | 边界框，格式 [x1, y1, x2, y2] |
| pages | 无 | List[int] | 边界框所在的页码 |
| indexes | 无 | List[List[int]] | 边界框提取的文本在合并后的的文本中的序列位置，包含起始和结束[start, end] |
| types | 无 | List[str] | 边界框中元素的类型 |
| } |  |  |  |
| } |  |  |  |

**示例请求代码**

|  |
| --- |
| Python # ocr 使用 paddle ocr url = "http://127.0.0.1:8011/v1/etl4llm/predict" filename = "./demo/demo1.pdf" b64\_data = base64.b64encode(open(filename, "rb").read()).decode() inp = dict(  filename=os.path.basename(filename),   b64\_data=[b64\_data],   force\_ocr=False) resp = requests.post(url, json=inp).json()   # ocr 使用 dataelem ocr url = "http://127.0.0.1:8011/v1/etl4llm/predict" filename = "./demo/demo1.pdf" b64\_data = base64.b64encode(open(filename, "rb").read()).decode() inp = dict(  filename=os.path.basename(filename),   b64\_data=[b64\_data],   force\_ocr=False，  ocr\_sdk\_url="http://192.168.106.20:8502"  ) resp = requests.post(url, json=inp).json() |

**响应格式**

高亮部分为需要关注的参数，其余参数暂可忽略

|  |
| --- |
| Python {  "status\_code": 200,  "status\_message": "success",  "text": null,  "html\_text": null,  "partitions": [], // 文档解析结果  "b64\_pdf": null } |

返回数据示例：

|  |
| --- |
| JSON {  'status\_code': 200,  'status\_message': 'success',  'text': None,  'html\_text': None,  'partitions': [  {  'type': 'NarrativeText',  'text': '统一社会信用代码',  'element\_id': '10d0153f-bc3d-4936-adca-98f0d532175c',  'metadata': {  'extra\_data': {  'bboxes': [  [  104,  138,  221,  151  ]  ],  'pages': [  0  ],  'indexes': [  [  0,  8  ]  ],  'types': [  'paragraph'  ]  }  }  },  {  'type': 'Title',  'text': '营业执照',  'element\_id': '4a4ccb0c-b67e-4c75-a693-404ec9bf42ac',  'metadata': {  'extra\_data': {  'bboxes': [  [  301,  129,  514,  177  ]  ],  'pages': [  0  ],  'indexes': [  [  0,  5  ]  ],  'types': [  'title'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '91310000717854505W证照编号：00000002201906170086',  'element\_id': '444863ec-e857-4b2c-abcc-b67cf7e886f2',  'metadata': {  'extra\_data': {  'bboxes': [  [  103,  159,  197,  169  ],  [  104,  180,  267,  195  ]  ],  'pages': [  0,  0  ],  'indexes': [  [  0,  19  ],  [  19,  44  ]  ],  'types': [  'paragraph',  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '',  'element\_id': '3dec7cbb-d883-4c22-a0c0-904cc130ec69',  'metadata': {  'extra\_data': {  'bboxes': [    ],  'pages': [    ],  'indexes': [    ],  'types': [    ]  }  }  },  {  'type': 'NarrativeText',  'text': '（副本）',  'element\_id': 'e8d0419f-3408-45f3-98b3-2a5c5852084a',  'metadata': {  'extra\_data': {  'bboxes': [  [  379,  179,  437,  202  ]  ],  'pages': [  0  ],  'indexes': [  [  0,  4  ]  ],  'types': [  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '名称尼康映像仪器销售（中国）有限公司注册资本美元1000.0000万类型有限责任公司（台港澳法人独资）',  'element\_id': '0ecc14c3-0d72-494c-9a56-dc67790859ed',  'metadata': {  'extra\_data': {  'bboxes': [  [  102,  228,  339,  245  ],  [  464,  230,  614,  245  ],  [  102,  257,  315,  275  ]  ],  'pages': [  0,  0,  0  ],  'indexes': [  [  0,  18  ],  [  18,  34  ],  [  34,  51  ]  ],  'types': [  'paragraph',  'paragraph',  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '',  'element\_id': '271b11c9-32bc-4667-9dba-164b83281d21',  'metadata': {  'extra\_data': {  'bboxes': [    ],  'pages': [    ],  'indexes': [    ],  'types': [    ]  }  }  },  {  'type': 'NarrativeText',  'text': '',  'element\_id': '5625dfdd-1ddf-4c48-ba8e-e993f83b29b7',  'metadata': {  'extra\_data': {  'bboxes': [    ],  'pages': [    ],  'indexes': [    ],  'types': [    ]  }  }  },  {  'type': 'NarrativeText',  'text': '成立日期2005年04月8日法定代表人松原微',  'element\_id': '2003b53a-2342-46d2-a59a-fd99746ec9f0',  'metadata': {  'extra\_data': {  'bboxes': [  [  463,  259,  604,  273  ],  [  101,  290,  211,  303  ]  ],  'pages': [  0,  0  ],  'indexes': [  [  0,  14  ],  [  14,  22  ]  ],  'types': [  'paragraph',  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '',  'element\_id': '36b9f6e3-3a5d-4ebb-a091-68198f6a24fd',  'metadata': {  'extra\_data': {  'bboxes': [    ],  'pages': [    ],  'indexes': [    ],  'types': [    ]  }  }  },  {  'type': 'NarrativeText',  'text': '营业期限2005年04月8日至2035年04月7日围受日本式会社尼康及其所投资企业的委托，向其提供下列眼务：投资经营决策，资金运作和财务管理，研究开发和技本支持，国内分用及老出口、货物分提等物运作，重接本公司集团内部的共享务及境外会司的服务外包，员工培训与管理及上连相关客询服务，光学收养及其相关产品，日月百货，玩具。文化体育用品。文具、请坦品，家居具品，服装。鞋加和配饰，辅包，电子产品，化妆品的批发，零售分支机构经营佣金代理（拍卖障外），吉出口，晨示【仅限尼康集团产品）并提供相美配套业务及售后服务（不涉及国营贸易管理商品：涉及配销。许可证管理商品的，按国家有关规定办理中请）：上述产品的委托生产：光学仅器及其相关产品的相赁：知识产权咨询（保限尼审集团自有知识产权）。【告须经批准的项目，经相关部门北准后方可开展经营活动】',  'element\_id': 'dfac1f1b-0ca9-499b-9257-ca78a00cd4ff',  'metadata': {  'extra\_data': {  'bboxes': [  [  464,  289,  691,  303  ],  [  158,  316,  450,  331  ],  [  178,  328,  439,  338  ],  [  179,  336,  448,  347  ],  [  179,  345,  439,  355  ],  [  178,  354,  443,  364  ],  [  179,  362,  440,  372  ],  [  179,  371,  448,  382  ],  [  179,  380,  448,  390  ],  [  180,  389,  448,  398  ],  [  179,  397,  428,  408  ],  [  182,  406,  411,  416  ]  ],  'pages': [  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0  ],  'indexes': [  [  0,  25  ],  [  25,  57  ],  [  57,  88  ],  [  88,  118  ],  [  118,  149  ],  [  149,  181  ],  [  181,  210  ],  [  210,  242  ],  [  242,  274  ],  [  274,  306  ],  [  306,  336  ],  [  336,  363  ]  ],  'types': [  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph',  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '',  'element\_id': 'ecdf04ff-06c0-4498-a9a6-4b72893bd853',  'metadata': {  'extra\_data': {  'bboxes': [    ],  'pages': [    ],  'indexes': [    ],  'types': [    ]  }  }  },  {  'type': 'NarrativeText',  'text': '住所上海市黄浦区蒙自路757号1201-1207室',  'element\_id': '76e980b9-3003-4f1a-95a7-629cb7d30611',  'metadata': {  'extra\_data': {  'bboxes': [  [  462,  315,  711,  333  ]  ],  'pages': [  0  ],  'indexes': [  [  0,  25  ]  ],  'types': [  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '登记机关',  'element\_id': '3a9e3ccd-735e-4ca2-af5c-196152e9919d',  'metadata': {  'extra\_data': {  'bboxes': [  [  511,  414,  592,  429  ]  ],  'pages': [  0  ],  'indexes': [  [  0,  4  ]  ],  'types': [  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '2019年06月17日国家市场监督管理总局监制',  'element\_id': 'b23fbaa0-dece-4157-9f09-4c0a387ad945',  'metadata': {  'extra\_data': {  'bboxes': [  [  588,  457,  717,  474  ],  [  648,  514,  759,  528  ]  ],  'pages': [  0,  0  ],  'indexes': [  [  0,  11  ],  [  11,  23  ]  ],  'types': [  'paragraph',  'paragraph'  ]  }  }  },  {  'type': 'NarrativeText',  'text': '',  'element\_id': 'de312cc2-8aef-45f7-a6a2-c05a5871d705',  'metadata': {  'extra\_data': {  'bboxes': [    ],  'pages': [    ],  'indexes': [    ],  'types': [    ]  }  }  }  ],  'b64\_pdf': None } |

2. **获取Markdown解析结果**

|  |
| --- |
| 容器内默认端口为 8000，宿主机默认端口为8011 |

**请求说明**

该接口目前只支持传入本地文件地址。

**请求参数**

HTTP方法: POST

请求URL: http://host:8011/v1/etl4llm/for\_gradio

|  |
| --- |
| host=本地服务部署的IP地址 |

URL参数: 无

Header参数：Content-Type=application/json

Body参数：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **参数** | **默认值** | 是否必填 | **类型** | **说明** |
| file\_path | 无 | 是 | str | 文件路径，目前该接口要求必须是本地路径 |
| filename | 无 | 是 | str | 文件名（必填） |
| b64\_data | 无 | 否 | List[str] | base64编码的文件数据 |
| url | 无 | 否 | str | 通过url指定解析的文件地址，与b64\_data必须填写一个。 |
| parameters | 无 | 否 | Dict | 请求url文件的请求头  {'headers': {}, 'ssl\_verify': True} |
| end\_pages | 无 | 否 | int | 转换页数限制，为空时默认转换文档全部 |
| force\_ocr | false | 否 | bool | 如果为true,则强制走模型进行 ocr；如果为false，则根据pdf 信息判断是否需要走 ocr，如果不需要，则不走 ocr 识别，改为提取 pdf 文字层。 |
| enable\_formula | true | 否 | bool | 如果为 true，会对一个页面进行 ocr + 公式检测+公式识别。如果为 false，则仅做 ocr。 |
| ocr\_sdk\_url | None | 否 | str | 如果填写则ocr 部分走 dataelem ocr，不填写走 paddleocr |
| for\_gradio | true | 否 | bool | 该接口，bool为true时，返回结果。 |

**返回说明**

|  |  |  |  |
| --- | --- | --- | --- |
| **参数** | **默认值** | **类型** | **说明** |
| md | 无 | str | markdown 文本 |
| md\_text | 无 | str | markdown 文本，与md返回结果一样 |
| output\_file | 无 | str | 所有预估文件 zip，包含  md文件  middle.json  origin.pdf  layout.pdf |
| pdf\_show | 无 | str | layout 预估结果可视化pdf |

示例请求代码：

|  |
| --- |
| Python url = "http://127.0.0.1:8011/v1/etl4llm/for\_gradio" file\_path='/ocr\_sdk/demo/demo3.png' inp = dict(  file\_path=file\_path,  filename=os.path.basename(file\_path),  mode="partition") resp = requests.post(url, json=inp).json() print(resp) |

返回数据示例：

|  |
| --- |
| JSON {  "md": "统一社会信用代码\n\n#营业执照\n\n91310000717854505W证照编号：00000002201906170086\n\n\n\n（副本）\n\n名称尼康映像仪器销售（中国）有限公司注册资本美元1000.0000万类型有限责任公司（台港澳法人独资）\n\n\n\n\n\n成立日期2005年04月8日法定代表人松原微\n\n\n\n营业期限2005年04月8日至2035年04月7日围受日本式会社尼康及其所投资企业的委托，向其提供下列眼务：投资经营决策，资金运作和财务管理，研究开发和技本支持，国内分用及老出口、货物分提等物运作，重接本公司集团内部的共享务及境外会司的服务外包，员工培训与管理及上连相关客询服务，光学收养及其相关产品，日月百货，玩具。文化体育用品。文具、请坦品，家居具品，服装。鞋加和配饰，辅包，电子产品，化妆品的批发，零售分支机构经营佣金代理（拍卖障外），吉出口，晨示【仅限尼康集团产品）并提供相美配套业务及售后服务（不涉及国营贸易管理商品：涉及配销。许可证管理商品的，按国家有关规定办理中请）：上述产品的委托生产：光学仅器及其相关产品的相赁：知识产权咨询（保限尼审集团自有知识产权）。【告须经批准的项目，经相关部门北准后方可开展经营活动】\n\n\n\n住所上海市黄浦区蒙自路757号1201-1207室\n\n登记机关\n\n2019年06月17日国家市场监督管理总局监制\n\n",  "md\_text": "统一社会信用代码\n\n#营业执照\n\n91310000717854505W证照编号：00000002201906170086\n\n\n\n（副本）\n\n名称尼康映像仪器销售（中国）有限公司注册资本美元1000.0000万类型有限责任公司（台港澳法人独资）\n\n\n\n\n\n成立日期2005年04月8日法定代表人松原微\n\n\n\n营业期限2005年04月8日至2035年04月7日围受日本式会社尼康及其所投资企业的委托，向其提供下列眼务：投资经营决策，资金运作和财务管理，研究开发和技本支持，国内分用及老出口、货物分提等物运作，重接本公司集团内部的共享务及境外会司的服务外包，员工培训与管理及上连相关客询服务，光学收养及其相关产品，日月百货，玩具。文化体育用品。文具、请坦品，家居具品，服装。鞋加和配饰，辅包，电子产品，化妆品的批发，零售分支机构经营佣金代理（拍卖障外），吉出口，晨示【仅限尼康集团产品）并提供相美配套业务及售后服务（不涉及国营贸易管理商品：涉及配销。许可证管理商品的，按国家有关规定办理中请）：上述产品的委托生产：光学仅器及其相关产品的相赁：知识产权咨询（保限尼审集团自有知识产权）。【告须经批准的项目，经相关部门北准后方可开展经营活动】\n\n\n\n住所上海市黄浦区蒙自路757号1201-1207室\n\n登记机关\n\n2019年06月17日国家市场监督管理总局监制\n\n",  "output\_file": "/ocr\_sdk/service/output/demo3\_56ef4aae-8636-4384-b45a-1442098e2c66.zip",  "pdf\_show": "/ocr\_sdk/service/output/demo3\_570e9ba2-8793-4e64-a3fb-a287a389fbf5/demo3\_layout.pdf" } |