智能文书档案组件析分引擎

一、概述

1.简介

组件名称：智能文书档案组件析分引擎

建设单位：绍兴市柯桥区档案馆

开发上线时间：2024 年 1 月

2.应用场景

档案利用体系

业务场景：面向档案开放共享需求，通过 AI 引擎，辅助人工分析、切分文书档案中不属于开放共享范围的办理单、过程稿等，提升文书档案管理与利用过程中的共享效果和用户体验。

3.解决问题

难点痛点堵点：文书档案一般由正文、草稿、附件、办理单等组件构成，但由于文书档案的各组件排列顺序并不固定，种类较多、情况复杂，析分、提取、开放其中的正文部分需要耗费大量人力和时间成本，无法高效便捷面向社会安全可控开放档案。

目的和效果：通过研究析分标准，规范确立析分方法，柯桥区档案馆联合行业头部厂商，研发了一个档案智能组件析分引擎。引擎能够通过智能化、自动化和程序化的技术手段提升档案开放共享工作效率，实现对档案正文、草稿、附件、办理单等组件的精准识别、智能析分，确保析分过程既符合档案的原始生成逻辑，又满足开放共享的实际需求。

二、当前运行使用情况

本组件于 2023 年 5 月开始调研，8 月开始研发，2024 年 1 月投入试用，并完成了与柯桥区全域数字化转型试点工作“163”应用场景的对接与集成，作为“开放文书智能化析分应用场景”的核心引擎能力支撑。系统现已应用于柯桥区档案馆、绍兴市档案馆、嘉兴市秀洲区档案馆、衢州市档案馆，累计处理文书档案数据 10 余万件次，不同年代档案的析分平均准确率超过了 80%。具体运行使用情况数据如下：

绍兴市柯桥区档案馆：已累计完成逾 80000 件次档案的智能组件析分；

绍兴市档案馆：已累计完成逾 100000 件次档案的智能组件析分；

秀洲区档案馆：已累计完成约 5500 件次档案的智能组件析分；

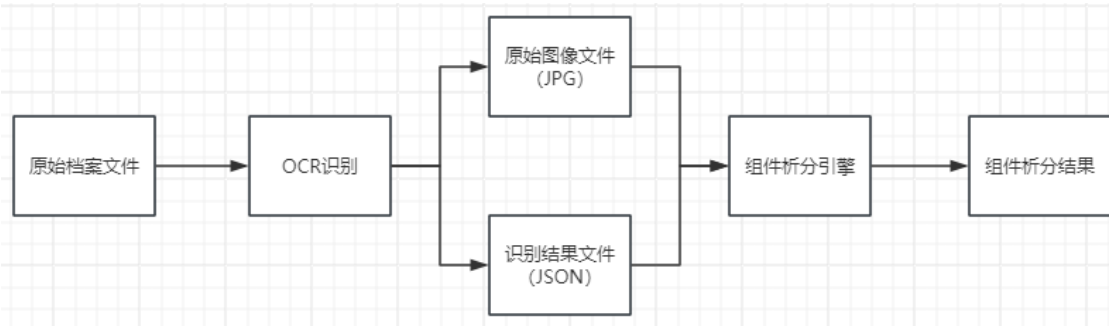
衢州市档案馆：已累计完成逾 20000 件次档案的智能组件析分。

三、核心功能

1.功能清单

属于 AI 引擎类组件，无功能清单。

2.业务流程



四、技术要点

1.部署方法

本组件以微服务形式部署。

非国产服务器建议配置：

cpu:4 核 x86_64；

操作系统：ubuntu 20.04.6；

GPU：NVIDIA GeForce RTX 3060 及以上，可用显存 12GB 以上；

国产信创服务器建议配置：

cpu:鲲鹏 920 及以上

操作系统：银河麒麟高级服务器操作系统 V10；

GPU：Atlas 300I Pro 及以上，可用显存 12GB 以上；

2. 调用规范

请求方式：post

content-type: application/json

请求 json：

```
[    //每页图片转成 base64
    {
        “base64”:””, //第一页图片转成 base64
        “imgPath”:””,//图片的本地路径，base64 可为空
        “ocrPath”:””,//图片的 ocr 识别结果的保存路径
    },
    {
        “base64”:”” ,//第二页的图片转成 base64
        “imgPath”:””,//图片的本地路径，base64 可为空
        “ocrPath”:””,//图片的 ocr 识别结果的保存路径
    },
```

结果 Json：

```
{
  "code": 200,
  //识别状态码，可选值为 200 和 500，分别表示分析成功和分析失败
  "message": "", //分析失败返回的报错信息，分析成功为""
  "results":[
    {
      "start_page_index":1,//起始图像页，从 1 开始计数
      "end_page_index":1,//截止图像页，包含该页
      "class_value":"阅办单"//类型
    },
    {
      "start_page_index":2,//起始图像页，从 1 开始计数
      "end_page_index":4,//截止图像页，包含该页
      "class_value":"正文"//类型
    },
    {
      "start_page_index":5,//起始图像页，从 1 开始计数
      "end_page_index":7,//截止图像页，包含该页
      "class_value":"附件"//类型
    }
  ]
}
```

五、使用指南

1.获取方式（接入、申请、开通账号流程）

接入

2.操作视频或截图

2.1 启动服务

```
(base) shihonghai@hxrj:~/project/dangan/component_segmentation$ bash start.sh
```

2.2 查看服务是否启动

```
(base) shihonghai@hxrj:~/project/dangan/component_segmentation$ tail -f logs/component_segmentation.log
05/13/2025 10:39:34 - INFO - root -   {"code": 200, "message": "success", "results": [{"start_page_index": 1, "end_page_index": 7, "class_value": "正文"}]}
05/13/2025 10:39:34 - INFO - werkzeug -   127.0.0.1 - - [13/May/2025 10:39:34] "POST /process2 HTTP/1.1" 200 -
05/13/2025 10:39:34 - INFO - root -   解析的文件数量为: 9
05/13/2025 10:39:34 - INFO - root -   imgs/28823247-71fa-43e1-990d-8b6b059115c6_298290临时文件移除成功
05/13/2025 10:39:34 - INFO - root -   {"code": 200, "message": "success", "results": [{"start_page_index": 1, "end_page_index": 9, "class_value": "正文"}]}
05/13/2025 10:39:34 - INFO - werkzeug -   127.0.0.1 - - [13/May/2025 10:39:34] "POST /process2 HTTP/1.1" 200 -
05/13/2025 10:39:36 - INFO - root -   解析的文件数量为: 11
05/13/2025 10:39:36 - INFO - root -   imgs/23dc81f4-b496-4436-a30d-5e14aafa74b3_200440临时文件移除成功
05/13/2025 10:39:36 - INFO - root -   {"code": 200, "message": "success", "results": [{"start_page_index": 1, "end_page_index": 11, "class_value": "正文"}]}
05/13/2025 10:39:36 - INFO - werkzeug -   127.0.0.1 - - [13/May/2025 10:39:36] "POST /process2 HTTP/1.1" 200 -
05/13/2025 10:39:36 - INFO - root -   解析的文件数量为: 3
05/13/2025 10:39:36 - INFO - root -   imgs/5ed63482-9dbf-420d-8a63-c5a41e8c5a22_374312临时文件移除成功
05/13/2025 10:39:36 - INFO - root -   {"code": 200, "message": "success", "results": [{"start_page_index": 1, "end_page_index": 3, "class_value": "正文"}]}
05/13/2025 10:39:36 - INFO - werkzeug -   127.0.0.1 - - [13/May/2025 10:39:36] "POST /process2 HTTP/1.1" 200 -
05/13/2025 10:39:36 - INFO - root -   解析的文件数量为: 26
05/13/2025 10:39:38 - INFO - root -   imgs/312d457e-7536-4ff8-a3a6-2e39b26aa0d5_422375临时文件移除成功
05/13/2025 10:39:38 - INFO - root -   {"code": 200, "message": "success", "results": [{"start_page_index": 1, "end_page_index": 2, "class_value": "正文"}, {"start
_page_index": 3, "end_page_index": 26, "class_value": "附件"}]}
05/13/2025 10:39:38 - INFO - werkzeug -   127.0.0.1 - - [13/May/2025 10:39:38] "POST /process2 HTTP/1.1" 200 -
```

日志中出现以下信息表示服务启动成功

```
 * Running on http://127.0.0.1:8872
 * Running on http://192.168.101.181:8872
05/16/2025 08:41:43 - INFO - werkzeug -   Press CTRL+C to quit
```

服务启动成功

2.3 析分结果输出

2.4 关闭服务

```
(base) shihonghai@hxrj:~/project/dangan/component_segmentation$ bash stop.sh
```

六、特点和效益

本引擎充分考虑了在文书档案智能分析和整理两大需要投入大量人力资源的业务环节，能够自动识别和分类文书档案中的各种组件，如正文、附件、发文稿纸等，这不仅提高了档案处理的速度和准确性，而且能够有效地理解和利用档案中的上下文关系，从而提升了整体的文档结构和语义关系捕捉能力。

将人工的档案识别与档案析分工作实现了从“做”到“验”的过渡，把大量机械性、重复性的人工工作，替代为智能化、自动化的计算机计算结果。因此，本项目预计能够在保证工

作成果质量的基础上，降低 70%以上的资源投入与人力成本，显著提高工作效率，为档案工作中的常见工作场景提供直接助力，具有较强的普适性和推广价值。