需求规格说明

编写：何铁科

审核：

批准：

从倒数第三
水电费
是打发点
阿斯蒂芬

软件工程与软件评测事业部研发部

二〇二〇年 4 月 16 日

~~目录~~

1. 引言

1.1 版本说明

此需求规格说明是根据 RWS-ZNHCS-WJJXFWD-20200416 任务书分解得到。

1.2 项目背景

测试业务流程涉及大量、多种类型的文档，如文档审查、文件索引、关键词检索等任务中，需要提前对 doc、docx、wps、pdf、excel 文件等文件进行内容解析，包括文字、图片、表格等内容。开发出独立的文件解析模块，可以将业务逻辑和文件解析进行逻辑分割，以微服务或者 SDK 的形式提供文件解析能力，可大大降低程序开发模块间的耦合度，提高后续程序管理、修改能力。

1.3 定义

文档解析：将非结构化文档数据转化为结构化条目数据。

1.4 参考资料

《文件解析服务端模块》RWS-ZNHCS-WJJXFWD-20200416 版本。

2. 任务概述

2.1 目标

解析模块需要支持常见 word 文档，如 doc、docx、wps、pdf 等文件解析，解析内容包括标题、段落、表格、并且需要支持获取表格在文档中的位置、图片位置、段落序号等信息，支持字体格式（字体大小、颜色、加粗等）和段落格式信息（打开 word 文档的段落样式，包括段前、段后、行距等）。具体的功能在需求说明中进行分析。第一版内容以文字内容解析为主。

2.2 运行环境

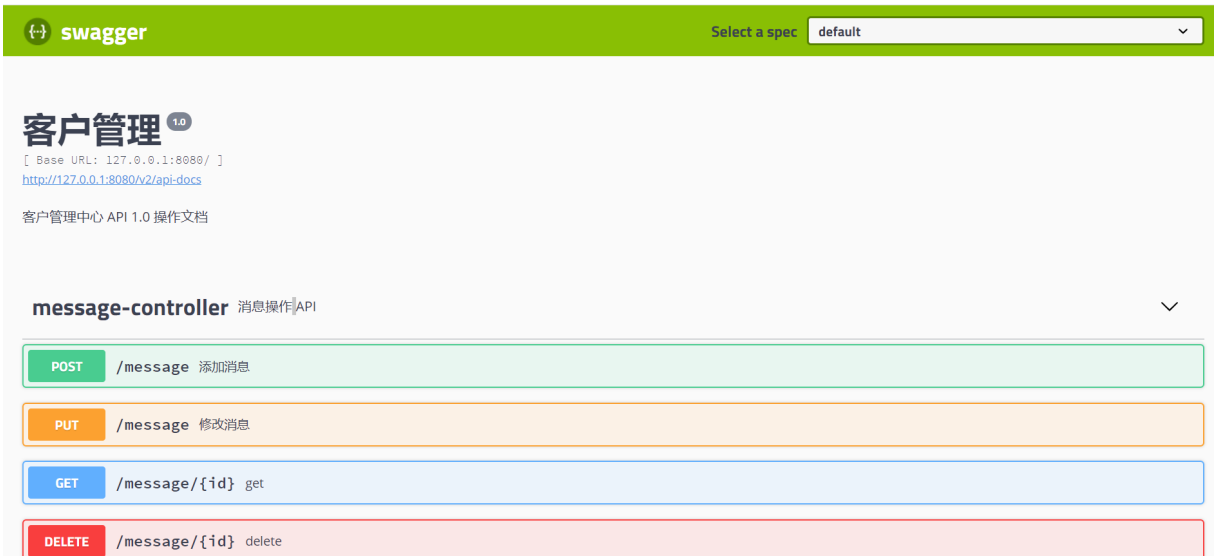
操作系统：Windows10（本地）
支持环境：JDK 8（1.8.0_231）
代码编辑器：IDEA

数据库：无

3. 功能需求

3.1 功能划分

文档载入：提供对支持文本文档的载入功能。
段落内容解析：从文件中提取文档段落信息。
段落格式解析：解析每个段落的基本信息。
标题内容解析：解析文档中的全部标题内容。
图片内容解析：解析文档中的全部图片内容。
段落字单元解析：解析给定段落的字单元集合。
字体格式解析：解析给定段落的字体格式集合。
表格解析：解析文档中的所有表格。
特定标题内容解析：解析给定标题下面所有段落信息。
Swagger 文档：支持采用 Restful 格式进行数据交互，并且采用 Swagger 查看文档接口。



3.2 功能描述

功能概述

文档载入：需要支持对 doc、docx、wps、pdf、excel 文件格式内容解析，即用户传输的文件格式符合这些格式则进行解析，并支持对二进制流文档进行格式判断，依据格式判断结果执行不同格式文件解析，并可对不支持的格式进行提示。

段落内容解析：支持从文件中解析文档段落信息，包括文字内容，自动编号，

以打开 word 看到为依据，解析获取的文字尽可能与其一致。

段落格式解析：解析每个段落的基本信息，包括对齐方式、大纲级别、特殊格式（首行缩进、悬挂缩进）、段前、段后、行距、对齐方式等。

标题内容解析：解析全部标题内容的接口，包括标题文本、标题级别、段落格式等，标题是一种特殊的段落。

图片内容解析：解析全部图片内容的接口，包括图片 base64 编码信息、图片格式、图片大小，获取图片在文档中的位置信息，也就是图片在文档中的段落序号，方便后续解析图片前后文。

段落字单元解析：通常段落由多个具有相同字体格式的单元组成，给定段落 ID，可获取段落的字单元集合。

字体格式解析：给定段落 ID，解析获取段落中所有最小独立单元（字体信息一致的最大连续字符）的字体信息，返回是一个字体列表。

表格解析：可以解析文档中的所有表格，每一个表格包括所有行和列的内容（通常为段落内容），是一个二维矩阵，每一个单元格视为一个段落组合。

特定标题内容获取：给定标题 ID，获取标题下面所有段落信息，标题下的内容以 word 中大纲视图为主。

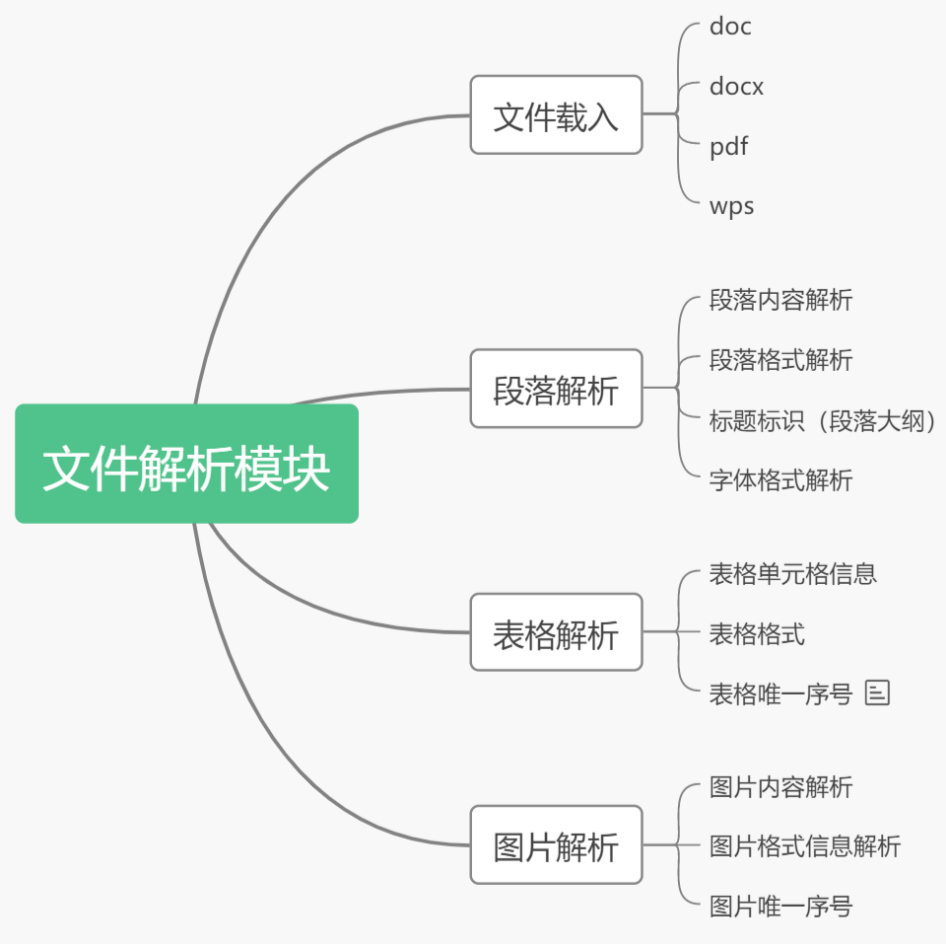


图 3-1 文件解析模块基本功能图

3.2.1 文档载入

（1）功能概述

文档载入：需要支持对 doc、docx、wps、pdf 文件格式内容解析，即用户传输的文件格式符合这些格式则进行解析，并支持对二进制流文档进行格式判断，依据格式判断结果执行不同格式文件解析，并可对不支持的格式进行提示。

（2）验收标准

1. 用户可以对 doc、docx、wps、pdf 等文本文档进行文档解析；
2. 上传大文件，查看解析是否正确；
3. 支持不同大小文档解析。

3.2.2 段落内容解析

（1）功能概述

段落内容解析：支持从文件中解析文档段落信息，包括文字内容，自动编号，以打开 word 看到为依据，解析获取的文字尽可能与其一致。一般 doc、docx、wps 和 excel 文件的基础解析拟基于 wordParserWithPOI 进行再次开发，pdf 类型文件拟采用 pdfbox 作为基本解析服务。

注意：需要对每一个段落按照 word 中显示的顺序进行标号，顺序与 word 中分段一致。

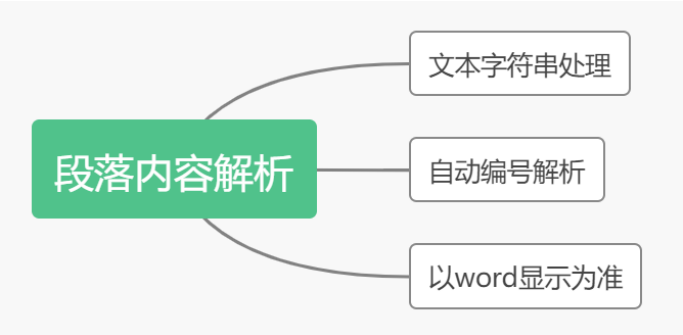


图 3-2 段落内容解析功能图

（2）验收标准

1. 可对 doc、docx、wps、pdf 等文本文档进行段落内容解析；
2. 解析到的段落文字字符串内容与 word 显示一致；
3. 如果有自动编号、则以 word 显示为主，即需要将编号“1.”等解析获取。

3.2.3 段落格式解析

（1）功能概述

段落格式解析：解析每个段落的基本信息，包括对齐方式、大纲级别、特殊格式（首行缩进、悬挂缩进）、段前、段后、行距、对齐方式等。

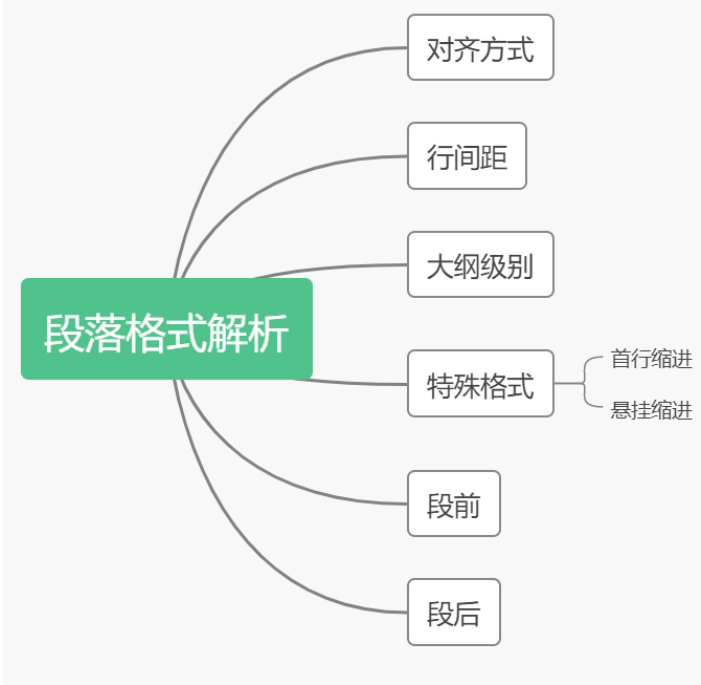


图 3-3 段落格式解析功能图

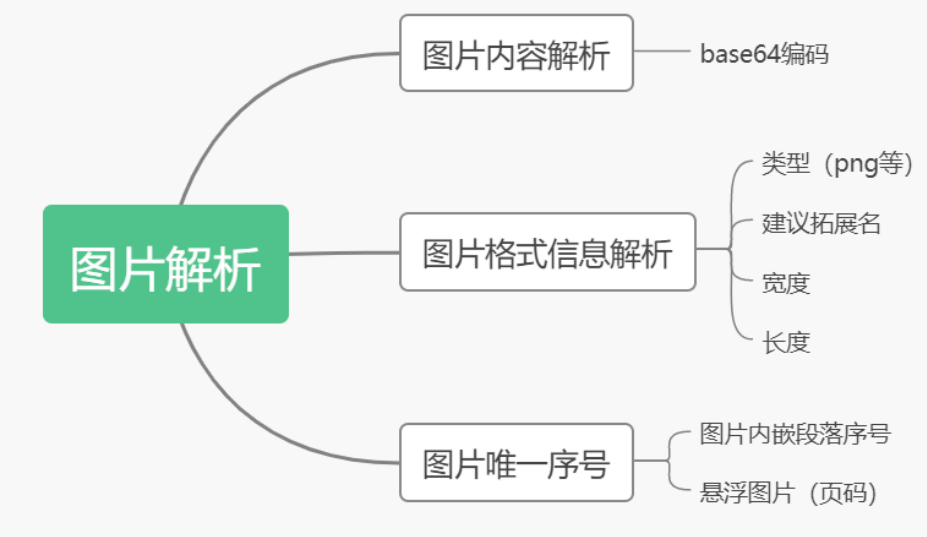


图 3-4 图片解析功能示意图

（2）验收标准

1. 可对 doc、docx、wps、pdf 等文本文档进行段落格式解析；
2. 解析到的段落格式与 word 中的段落格式显示一致。

3.2.4 标题内容解析

（1）功能概述

标题内容解析：解析全部标题内容的接口，包括标题文本、标题级别、段落格式等，标题是一种特殊的段落。

注：标题经常具有自动编号，需要特殊对待处理

（2）验收标准

1. 可对 doc、docx、wps、pdf 等文本文档进行标题内容解析；
2. 对具有自动编号、无自动编号标题可正确解析；
3. 可正确获取标题的大纲级别、标题文本内容、标题格式（段落格式）等。

3.2.5 图片内容解析

（1）功能概述

图片内容解析：解析全部图片内容的接口，包括图片 base64 编码信息、图片格式、图片大小，获取图片在文档中的位置信息，也就是图片在文档中的段落序号，方便后续解析图片前后文。

图片解析需要知道图片所在段落序号、图片 base64 编码、图片长度、图片宽度、图片格式等。

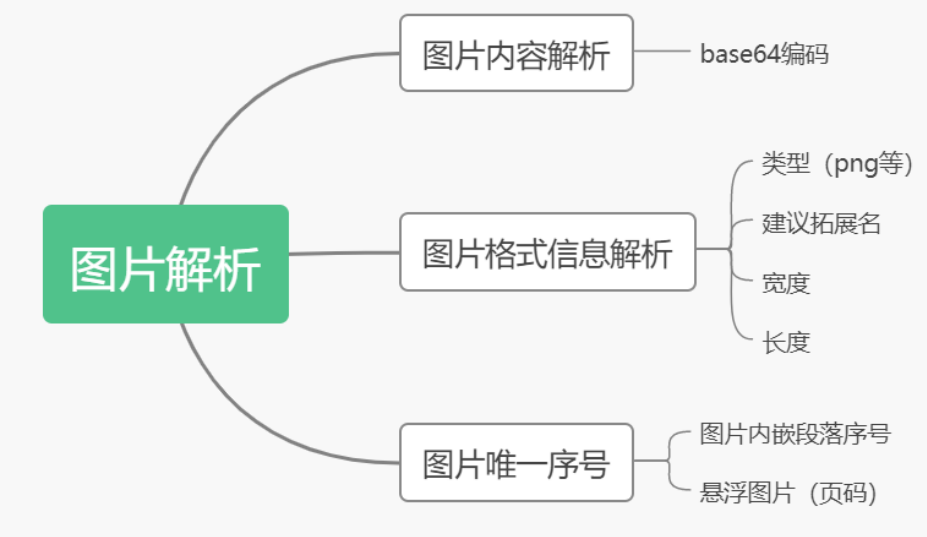


图 3-4 图片解析功能示意图

（2）验收标准

1. 可对 doc、docx、wps、pdf 等文本文档进行图片内容解析；
2. 可正确解析图片内容，可将图片保存在制定位置；

3．可正确解析图片大小等图片格式；
4．可对内嵌的图片解析图片段落位置。

3.2.6 段落字单元解析

（1）功能概述

段落字单元解析：通常段落由多个具有相同字体格式的单元组成，给定段落 ID，可获取段落的字单元集合。

一个段落如“我爱中国”，由于中国为红色字体，则这个段落至少由两个字单元组成，一个是“我爱”，另一个是“中国”，因为为了准确获取每一个段落的字体颜色、加粗等信息，需要获取段落由那些字单元组成。每个字单元由独立的、具有相同字体格式的字符串组成。

（2）验收标准

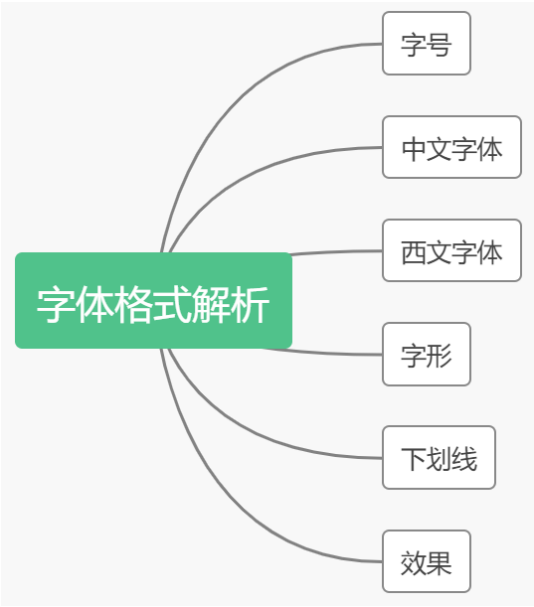
1．可对 doc、docx、wps、pdf 等文本文档进行段落字单元解析；
2．采用部分加粗、斜体、字体大小等方式，将段落人工划分为不同的部分，验证是否可以正确解析字单元，即是否按照正确格式将段落划分为正确的组合；
3．依据“字体格式解析模块”，查验每一部分的字单元的字体是否正确。

3.2.7 字体格式解析

（1）功能概述

字体格式解析：给定段落 ID，解析获取段落中所有最小独立单元（字体信息一致的最大连续字符）的字体信息，返回是一个字体列表。

字体格式需要解析中文字体、西文字体、字号、字形、字体颜色、下划线、效果等。



（2）验收标准

1. 可对 doc、docx、wps、pdf 等文本文档进行字体解析；
2. 根据 word 字体显示的信息对比解析结果。

3.2.8 表格解析

（1）功能概述

表格解析：可以解析文档中的所有表格，每一个表格包括所有行和列的内容（通常为段落内容），是一个二维矩阵，每一个单元格视为一个段落组合。

表格解析需要解析到表格的内容信息，目前支持表格中的段落解析，如果表格中具有多个段落，则分别解析获取。

（2）验收标准

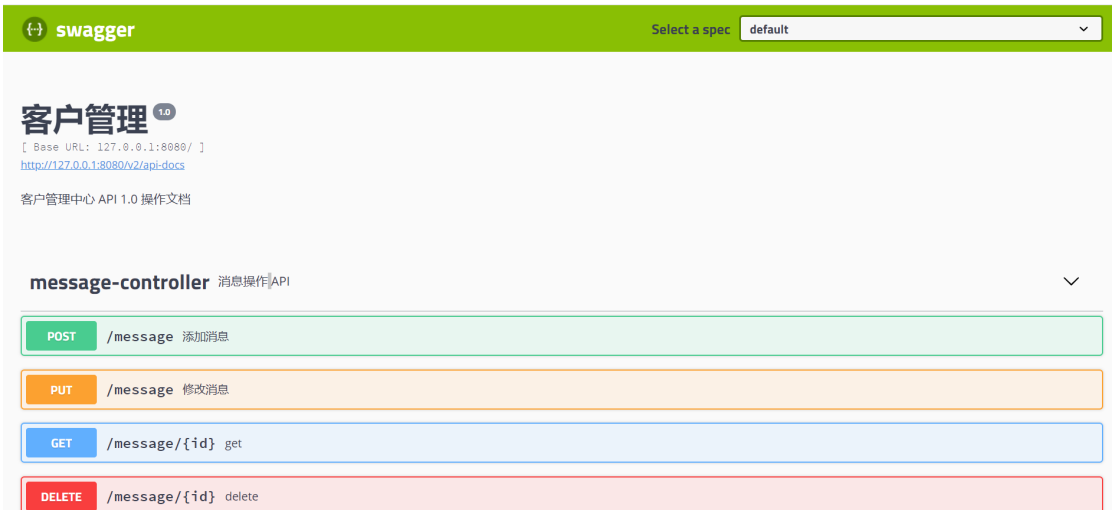
1. 可对 doc、docx、wps、pdf 等文本文档进行表格内容解析，包括段落解析

2. 检查是否可以对表格内容进行充分解析。

3.2.9 Swagger Restful 接口文档

（1）功能概述

需要将现有功能按照 Restful 接口规范进行整理，为了方便测试，需提如下所示的文档接口及说明，方便测试开发。



（2）验收标准

1、API 接口包含上述所有功能；
2、API 参数、返回值说明清楚；

3、API 接口

4. 性能需求

（1）支持文档格式为 doc、docx、pdf、wps 等；
（2）支持在 Linux 下运行；
（3）支持 word 文档大小不小于 200M；
（4）支持文字版 pdf；
（5）支持多份文档同时解析；
（6）支持 Python 直接调用（不开端口）；
（7）提供 python 包，支持用户直接调用。

5. 运行需求

5.1 用户界面

无

5.2 软件接口

后续详细设计给出对应的 API 接口。

5.3 故障处理

无。

6. 其它需求

无

1. 标题
   1.1 标题 1.1
2. 标题 2

表格 2：材料

1	身份证扫描件（正反）	原件彩色扫描件	是的

2	学位证书	地方
3		发的
4	合并单元格 2	地方