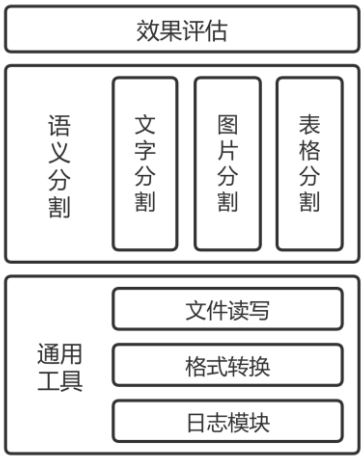
设计文档

2020.02.28

1. 功能结构图

根据工程的需求,将整个工程项目划分为通用工具、语义分割及效果评估三个组成部分。



整个工程主要分为六个模块，分别是核心业务之外的文件读写和格式转换模块；语义分割过程中的文字分割、图片分割和表格分割模块；以及用于测试分割效果的效果评估模块。

2. 通用工具

工程项目当中常用的功能统一放在通用工具之中。包括文件读写模块和格式转换模块。

2.1 文件读写模块

文件读写模块用于整个工程与文件系统之间的交互。

- 文件读入：
  a) 配置文件读入
  b) 标注文件[附1]读入
- 文件写出：
  a) 图片文件写出
  b) JSON 文件[附2]写出
  c) 评估结果写出

2.2 格式转换模块

格式转换模块用于将 PDF 文档转换为 XML 布局对象和图片列表，以及将语义分割结果转换为图片列表和结构化 JSON 文档。

- PDFtoXML：利用 pdfminer 生成 PDF 文件对应的 XML 布局对象
- PDFtoImage：利用 pdf2image 生成 PDF 文件对应的图片
- 坐标域转换：将语义分割结果所在的版面坐标域映射到图片坐标域
- ResultToJSON：将语义分割结果组织成结构化的 JSON 文档
- ResultToImage：将语义分割结果以矩形框的形式绘制在图片上并说明其类型

2.3 日志模块

日志模块用于记录在项目运行过程中 checkpoint 的通过，代码的运行时间，以及可能出现的文件无法解析、数据异常、程序崩溃等异常情况。以便于调试程序、了解程序运行状况并分析定位故障点。

- 日志模块初始化

3 语义分割

利用 XML 版面对象基于规则开展版面分割，分别提取出不同语义成分，并将不同语义成分所在位置正确地组织为同一格式。

3.1 文字分割模块

基于规则对版面对象进行分析，提取出一级和二级语义下的文字部分的内容和坐标。

- 一级语义分割：提取出当前页面内所有文字块的内容及其所在位置的坐标
- 二级语义分割：提取出当前页面内标题、作者、图注、表注、页码、注释和正文的内容及其所在位置的坐标

3.2 图片分割模块

基于规则并对版面对象进行分析，并提取出图片所在位置。

- 图片分割：提取出当前页面内所有图片所在位置的坐标

3.3 表格分割模块

基于规则并对版面对象进行分析，提取出表格的坐标以及内部单元格的内容、坐标及行/列表头信息。

- 表格检测：提取出当前页面内所有表格所在位置的坐标
- 单元格分割：提取出当前页面内每个表格内的每个单元格的内容、所在区域的坐标及其行/列表头信息

4 效果评估

效果评估模块用于量化人工标注与版面分割结果之间的拟合程度。

- 遍历比对：对当前页面下语义分割结果中的每一个区块矩形框，遍历读入的标注文件中相同类型的区块矩形框，计算二者 IoU，如若 IoU 大于该类型的设定阈值，判定该区块框所属类型预测正确。

- IoU 计算：计算两个矩形框面积的交并比

  交并比：“预测的边框”和“真实的边框”的交集和并集的比值。

$$IoU = \frac{InterSection(prediction, groundTruth)}{Union(prediction,\ groundTruth)}$$

- 准确率及召回率计算：对所有语义成分计算语义分割的准确率和召回率

  准确率：对某类语义成分，预测正确的个数与预测出的该类语义成分个数的比值。

$$precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

  召回率：对某类语义成分，预测正确的个数与真实标注的该类语义成分个数的比值。

$$recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

- F1 计算：基于准确率和召回率计算 F1 分数

  F1 分数：准确率和召回率的调和平均，综合地衡量语义分割的效果。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

  其中：$precision$ 为准确率，$recall$ 为召回率。

附录

1. 标注文件格式：

- 文字区域：

  均为二级语义标注，每个文字区域所在矩形框占用标注文件的一行，格式如下：

  Semantic Name: leftUpX, leftUpY, rightDownX, rightDownY

  Semantic Name 为二级语义名称，其后四个标识符从左到右分别表示矩形框的左上角和右下角顶点坐标

- 图片区域：

  每张图片所在矩形框占用标注文件的一行，格式如下：

  Image: leftUpX, leftUpY, rightDownX, rightDownY

- 表格区域：

- 对表格区域标注的第一行是表格的位置坐标，后面紧跟数行其内单元格的位置坐标

  表格：

  每张表格所在矩形框占用标注文件的一行，格式如下：

  Table: leftUpX, leftUpY, rightDownX, rightDownY

  单元格：

  每个单元格所在矩形框占用标注文件的一行，格式如下：

  cell: leftUpX, leftUpY, rightDownX, rightDownY

2. 结构化 Json 文档：

2.1 内部参数说明：

2.1.1 文档参数

父元素名称	元素名称	类型	长度	必填	描述	取值说明
Root	FileName	String	-	是	文件名	被解析的 PDF 文档的文件名称
Root	Pages	Object	-	是	页面列表对象	Pages 为 json 对象，内部存放的数据为该文件下所有页面的版面信息

2.1.2 页面参数

父元素名称	元素名称	类型	长度	必填	描述	取值说明
Pages	PageNo	Int	-	是	页码编号	同级目录下的版面信息所在页码编号
Pages	PageLayout	Object	-	是	版面信息列表对象	PageLayout 为 json 对象，内部存放的数据为该页面内的文字、图片和表格的版面信息

2.1.3 版面参数

父元素名称	元素名称	类型	长度	必填	描述	取值说明
PageLayout	Text	Object	-	否	文字部分列表对象	Text 为 json 对象，内部存放的数据为该页面文字部分的详细信息
PageLayout	Image	Object	-	否	图片部分列表对象	Image 为 json 对象，内部存放的数据为该页面图片部分的位置信息
PageLayout	Table	Object	-	否	表格部分列表对象	Table 为 json 对象，内部存放的数据为该页面表格部分的详细信息

2.1.4 文字区域参数

父元素名称	元素名称	类型	长度	必填	描述	取值说明
Text	Semantic Type	String	-	是	语义类型	该文字块的语义类型
Text	content	String	-	是	文字内容	该文字块的文字内容
Text	Location	List	4	是	位置列表	该列表内的四个元素分别是该文字块所在区域的左上角和右下角顶点坐标
Text	TextLines	Object	-	是	文字行列表对象	TextLines 为 json 对象，内部存放的数据为该文字块内每一行的文字和位置信息

2.1.5 图片区域参数

父元素名称	元素名称	类型	长度	必填	描述	取值说明
Image	location	List	4	是	位置列表	该列表内的四个元素分别是该图片所在区域的左上角和右下角顶点坐标

2.1.6 表格区域参数

父元素名称	元素名称	类型	长度	必填	描述	取值说明
Table	location	List	4	是	位置列表	该列表内的四个元素分别是该表格所在区域的左上角和右下角顶点坐标
Table	cells	Object	-	是	单元格列表对象	cells 为 json 对象，内部存放的数据为该表格内单元格的内容、位置和行/列表头信息

2.2　系统导出的 json 文档示例


```
{
    "FileName": "PubLayNet.pdf",
    "Pages": [
        {
            "PageNo": 1,
            "PageLayout": {
                "Text": [
                    {
                        "SemanticType": "Title",
                        "content": "PubLayNet: largest dataset ever for document layout analysis",
                        "location": [10, 10, 70, 14],
                        "TextLines": [
                            {
                                "content": "PubLayNet: largest dataset ever for document",
                                "location": [10, 10, 70, 12]
                            },
                            {
                                "content": "layout analysis",
                                "location": [30, 12, 50, 14]
                            }
                        ]
                    }
                ],
                "Image": [
                    {
                        "location": [50, 100, 100, 150]
                    }
                ],
                "Table": [
                    {
                        "location": [100, 200, 300, 300],
                        "cells": {
                        }
                    }
                ]
            }
        }
    ]
}
```