

一、简介

此工具主要用于 docx 文件的数据提取，并格式化转换为标准 json 结构数据

二、依赖

Python 3.10.14

pip 24.0

三、环境准备

pip install -r requirements.txt

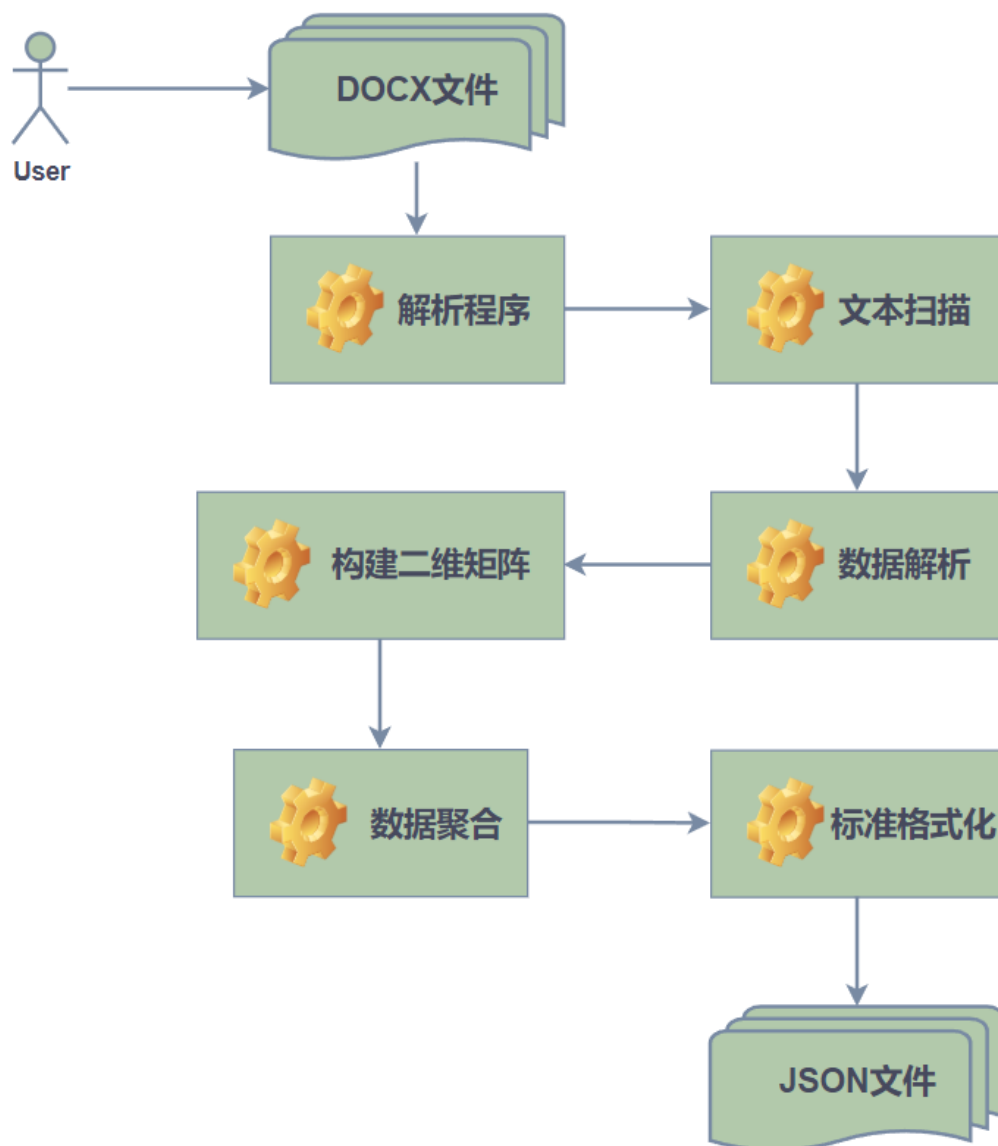
四、使用方法

1 将 docx 文件，置于同级目录下

2 运行转换程序

3 python load_docx2json_bin.py

五、流程图



六、流程描述

- 1 解析程序加载 docx 文件
- 2 对文本进行全量扫描
- 3 提取有效数据
- 4 基于数据样本，构建二维矩阵（5*4 矩阵）
- 5 聚合降噪，去除重复及干扰词条
- 6 对数据进行格式化处理，生成标准化纯净数据

7 转换为 json 格式，输入为 xxx.json 文件