

Contract Parsing & Key Content Extraction

Haoming(Hammond) Liu

Summer 2021

1 Project Abstract

This project builds a set of online APIs for a urban construction corporation, which **parses the contracts (written in Chinese), extract key information and terms, and store in online database for further use.** The core functions are mainly based on the [Tesseract OCR Engine](#), and the structure parsing is achieved by title line recognition and template matching. The choices of backend framework and databases are [Flask](#) and [MySQL](#) respectively. Due to the confidentiality terms, the implementation will neither open source nor elaborate in details.

2 Schematic Diagrams

第一条 项目概况
1.1 项目名称: _____
1.2 项目地址: _____
1.3 承包范围: _____ X号楼X层至X层
1.4 工作内容: 铝合金模板生产加工、运输、卸料、安装、拆除等工作内容。
1.5 承包方式: 包工包料形式承包, 甲方负责提供乙方工人住宿房间。
第二条 合同价款
2.1 本工程采用 <u>固定单价</u> 计价方式(各子项单价及计算规则详见附件1), 合同

Figure 1: Title Line Recognition Results

id	md5_id	keyword	info
49	8a077ef831bc42918610198e0ea3dd44	项目名称	这里是项目名称
50	8a077ef831bc42918610198e0ea3dd44	甲方	样本甲方名称
51	8a077ef831bc42918610198e0ea3dd44	乙方	_____公司
52	8a077ef831bc42918610198e0ea3dd44	签约地点	_____
53	8a077ef831bc42918610198e0ea3dd44	签约时间	2020年123月456日
54	8a077ef831bc42918610198e0ea3dd44	合同编号	null
55	8a077ef831bc42918610198e0ea3dd44	项目地址	这里是项目地址
56	8a077ef831bc42918610198e0ea3dd44	承包方式	包工包料形式承包,甲方负责提供乙方工人住家房...
57	8a077ef831bc42918610198e0ea3dd44	项目工期确定	工期节点,世格按项目部总进度计划要求及任务单...
58	8a077ef831bc42918610198e0ea3dd44	争议	双方在合同执行过程中如发生争议,应本着友好的...
59	8a077ef831bc42918610198e0ea3dd44	承包范围	_____
60	8a077ef831bc42918610198e0ea3dd44	2.1	本工程采用固定单价计价方式(各子项单价及计算...
61	8a077ef831bc42918610198e0ea3dd44	2.2	综合单价包括内容:人工费、材料费、安全文明施...
62	8a077ef831bc42918610198e0ea3dd44	3.1	支付时间:必备条款:合同签订后5个工作日内预付...
63	8a077ef831bc42918610198e0ea3dd44	3.2	乙方所完成的工作量须于每月5日前报送上月已...

Figure 2: Key Content Extraction Results