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

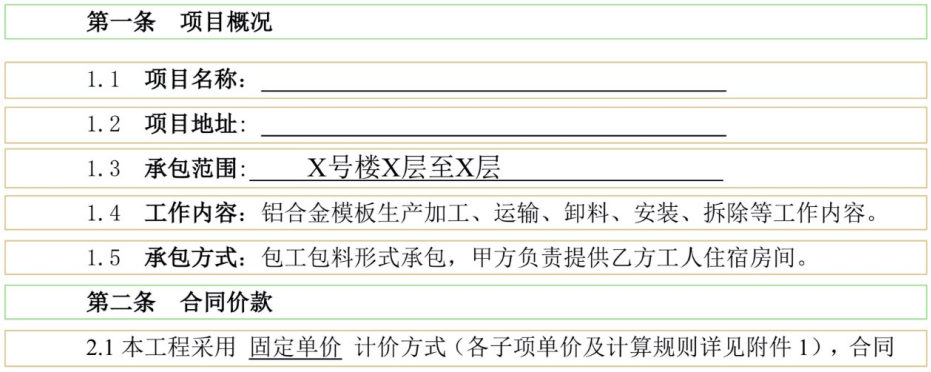

Figure 1: Title Line Recognition Results

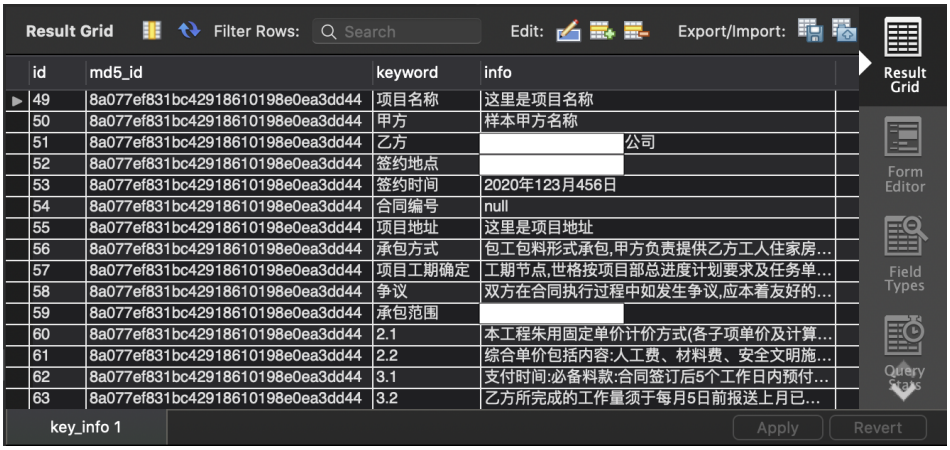

Figure 2: Key Content Extraction Results