

Universal Extraction Tool

Turki AlHarbi - Computer Science

Project Advisor/Sponsor: Dr. Donald Roberts

Introduction

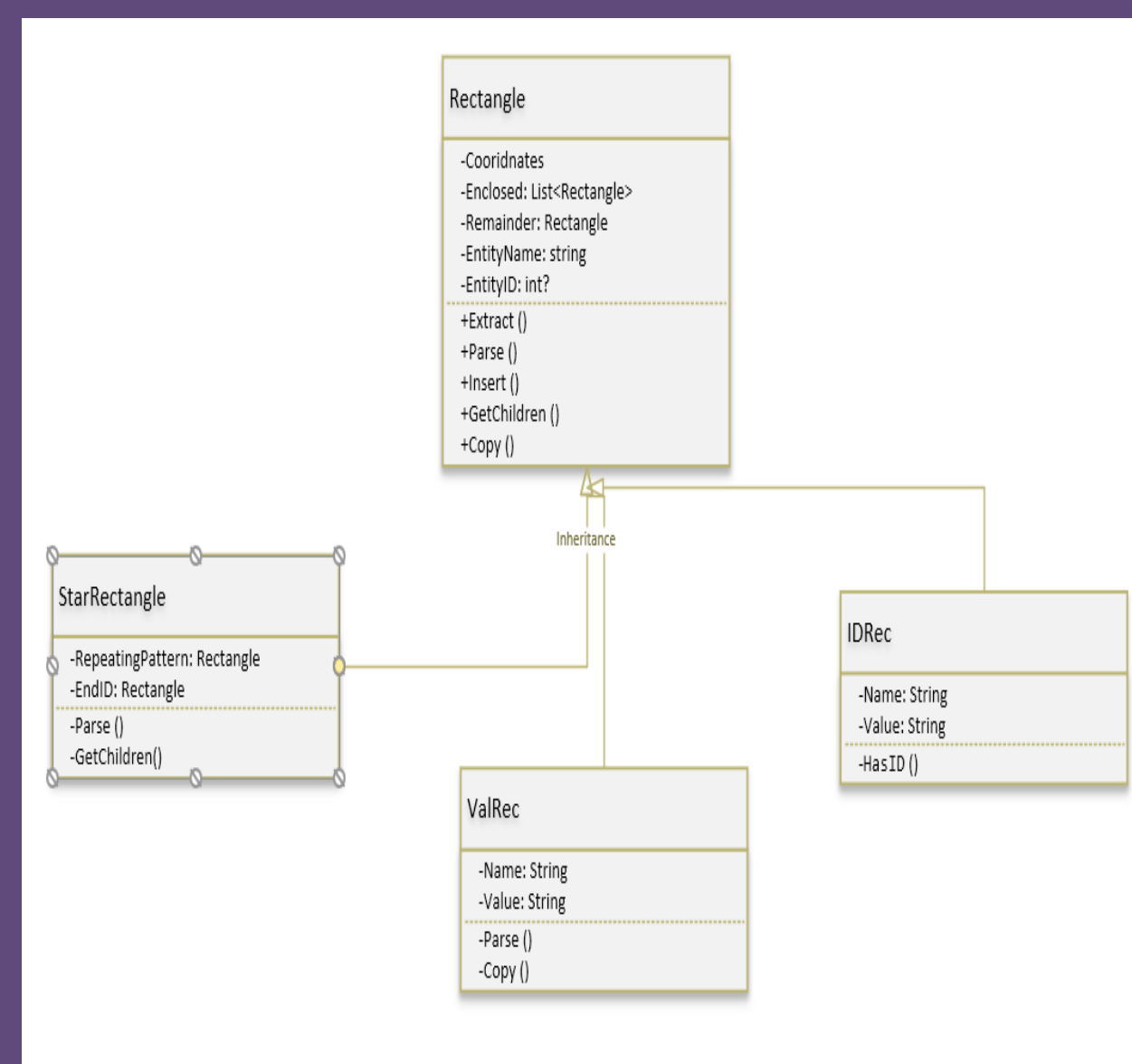
Organizations have many different forms of data such as mail, invoices, and other documents. This project focuses on removing manual data entry which is time consuming and error prone by providing an extraction tool that automates data entry.

Requirements

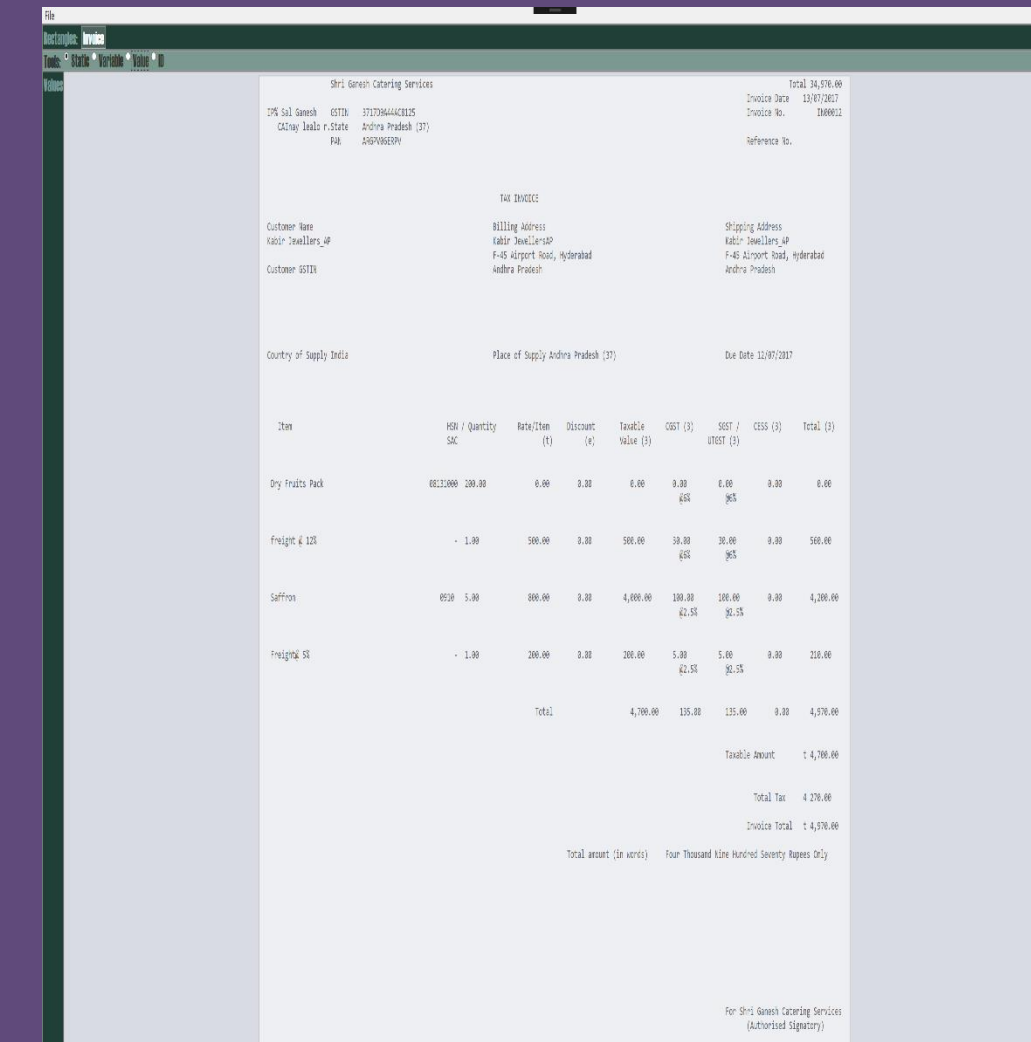
- Must have an easy way of highlighting useful information
- Able to parse documents of variable sizes
- Extracted data can be exported to a database

Design

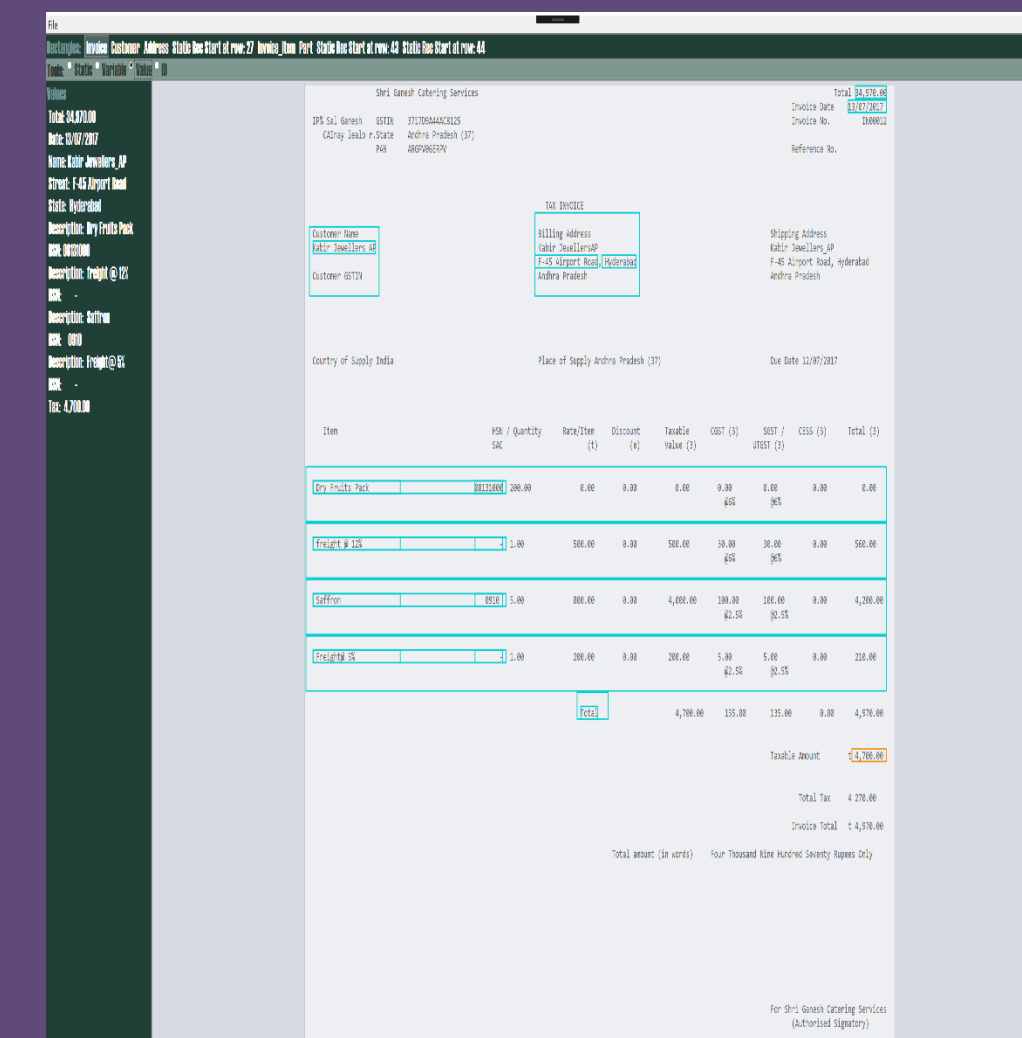
- Information in the text file are rectangles. These rectangles can contain an entity, a value, or a composition of rectangles.
- These rectangles have a position relative to their parent rectangle
- They can be variable in size with a repeating pattern and an end identifier



Results



- Users start by loading a document which they use to highlight data



- They can draw rectangles and specify whether they are an entity, value or a variable rectangle.
- Lastly, they can load other documents of the same form and they get parsed and inserted into a database

Conclusion

- The project is a proof of concept.
- It can effectively parse information from text files and export it.

Future work:

- Add more exporting methods
 - Accounting programs
 - Different types of Databases
- Entity Mapping
 - Allow users to specify the relationships between entities.