

**Walchand College Of Engineering, Sangli.**

**(An Autonomous Institute)**

**Department of**

**Computer Science and Engineering**

**Mini-Project Synopsis**

on

**Document Key-Value Pairs Extractor**

by

Swapnil Satish Mali (21510054)

Siddhesh Sunil Kitkaru (21510055)

Smit Sandeep Butle (21510056)

Under the Guidance of

**Prof. N. L. Mudegol**

|  |
| --- |
| **Dr. M. A. Shah** |
| HOD |
| Computer Sci. & Engg. Dept, |
| WCE, Sangli |

2023-24

**PROBLEM STATEMENT**

In a data-rich environment, the challenge lies in converting heterogeneous data from sources like GST bills and shop receipts into formats that are both standardized and insightful. The problem is to devise a solution that can adapt to different data layouts and types, automatically extracting key information and presenting it in user-friendly formats.

**ABSTRACT**

This abstract introduces a problem domain focused on the conversion of heterogeneous data, including GST bills and shop receipts, into actionable insights. The domain addresses the intricate challenges of extracting meaningful information, standardizing data formats, and presenting converted outputs like JSON, PDF, Word, or CSV. By employing advanced technologies such as machine learning and natural language processing, the solution aims to enhance data interpretation, user-friendliness, and accuracy. The outcome promises to bridge the gap between raw data and informed choices, shaping the future landscape of data-driven business strategies.

**PROBLEM DOMAIN**

Business documents are a cornerstone of any business transaction. Invoices, receipts, lease agreements, bill of lading are some examples of such documents. The information contained in these documents is critical for maintaining the sanctity of the underlying business transaction and for downstream processes.

To extract information from documents, traditionally, companies resort to manual processes for reviewing and analyzing the documents, identifying relevant fields, and entering data in the computer systems. While there exist solutions to extract information from semi-structured ‘True-PDF’ (ex. e-invoice) documents, these solutions will not cater to fully unstructured or ‘Image-Only’ document types.

It is extremely difficult for a single algorithm to address the problem of data extraction from any arbitrary document. The ubiquity of business documents in industry and the challenge described above make the problem of data extraction extremely attractive for researchers and practitioners.

**OBJECTIVES**

**Data Collection:** To collect user input data tailored to their specific needs and preferences efficiently and accurately, ensuring user-friendly experience.

Ex => SROIE datasetv2 - A grouped and organized dataset of the original ICDAR 2019 SROIE dataset

Source link => https://www.kaggle.com/datasets/urbikn/sroie-datasetv2

**Data Processing:** The solution needs to handle diverse data types and formats, ranging from GST bills to shop receipts. This involves data parsing, normalization, and structuring to ensure consistency and coherence across the converted outputs.

**Information Extraction:** The process of transforming raw, unstructured data into useful information requires advanced techniques for recognizing and extracting key data points. Algorithms that can identify relevant information such as product details, dates, amounts, and vendor information are essential.

**Data Standardization:** Converting heterogeneous data into standardized formats like JSON, PDF, Word, or CSV requires handling various layouts, fonts, and designs. Techniques for standardizing data representation are important to ensure consistency in the converted outputs.

**Usable Outputs:** The solution needs to present the extracted information in formats that are user-friendly and conducive to analysis. This may involve generating visualizations, creating easy-to-read reports, or integrating the converted data into existing software systems.

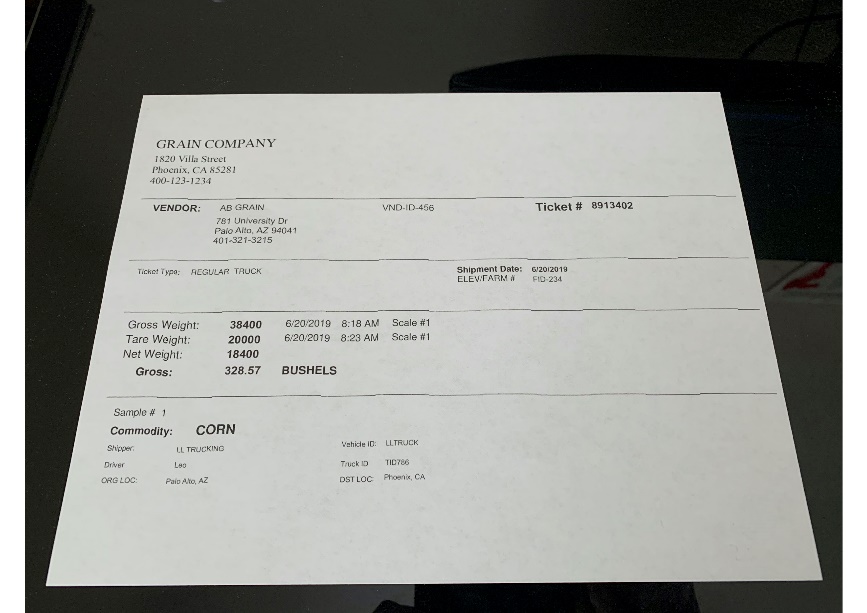
**FUNCTIONAL BLOCK DIAGRAM**

**UML DIAGRAMS**

**METHODOLOGY**

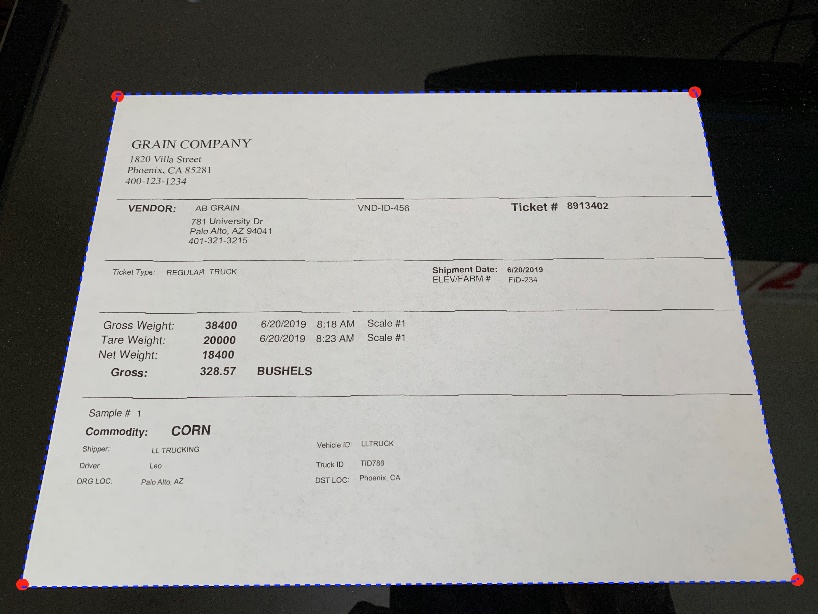
**Document Capture**

Capture a picture of the physical Bill-of-Lading document using a mobile phone or a scanner.



#### **Image pre-processing**

The document after being captured may not always be in a shape or form that we would like to use for extracting information. In this case, you can see that the document is tilted, inclined, and is of poor contrast. This will be transformed into a workable format using corner detection, image registration, and local-histogram equalization techniques.

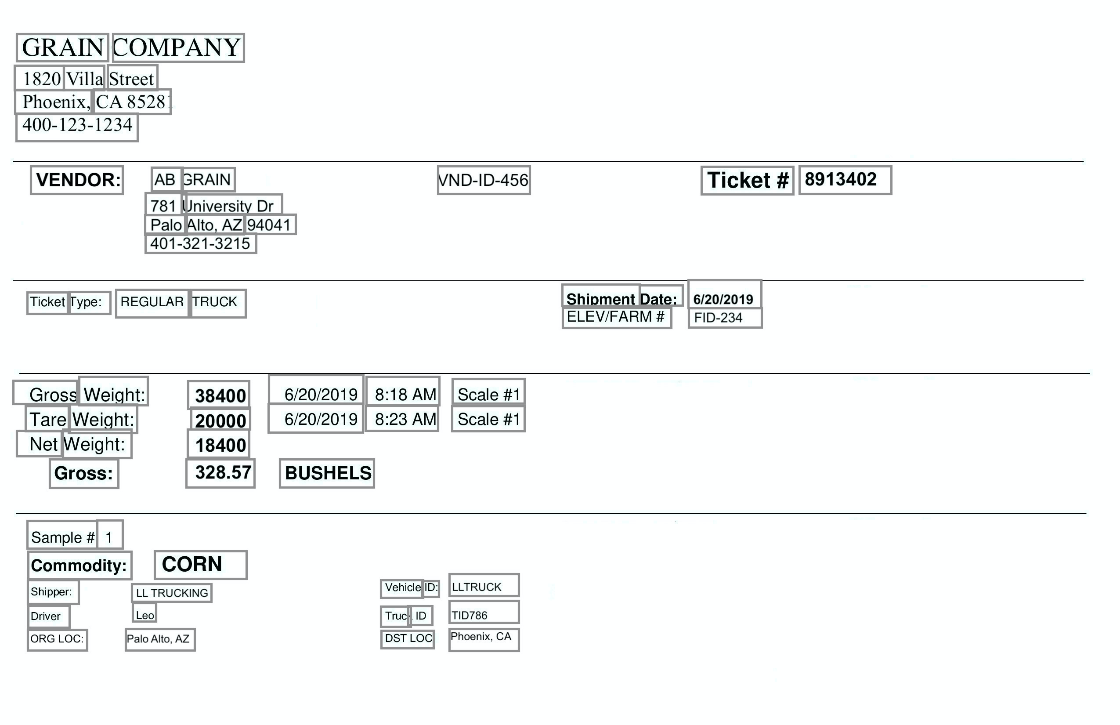
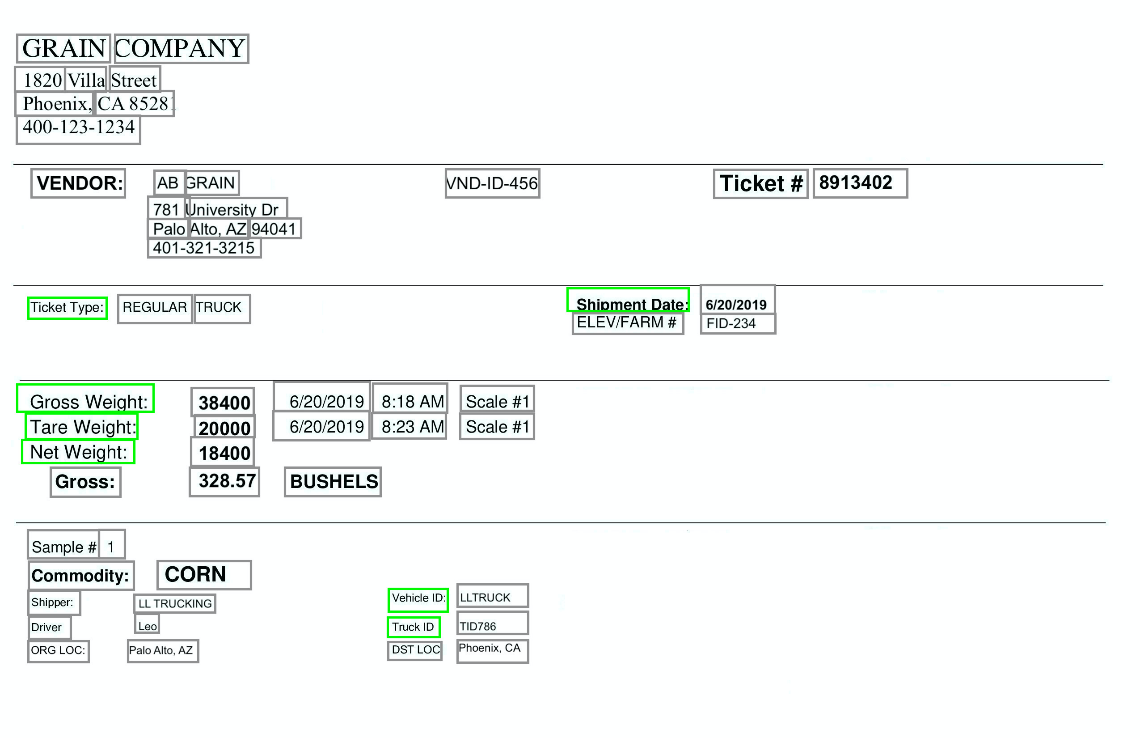


*Corner detection as part of image pre-processing*



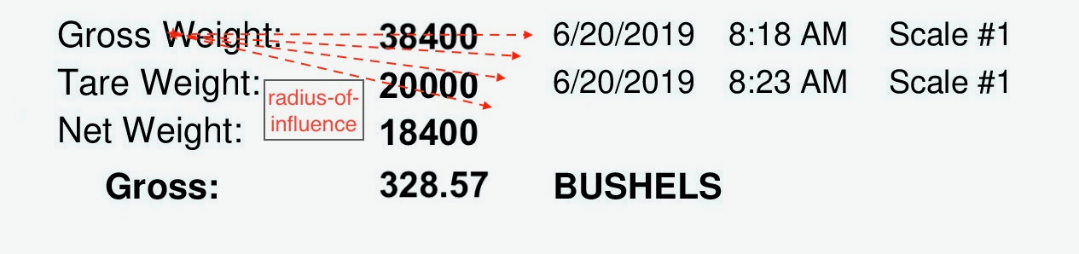
*Transformed bill-of-lading document after pre-processing*

#### **Entity location and centroid determination**

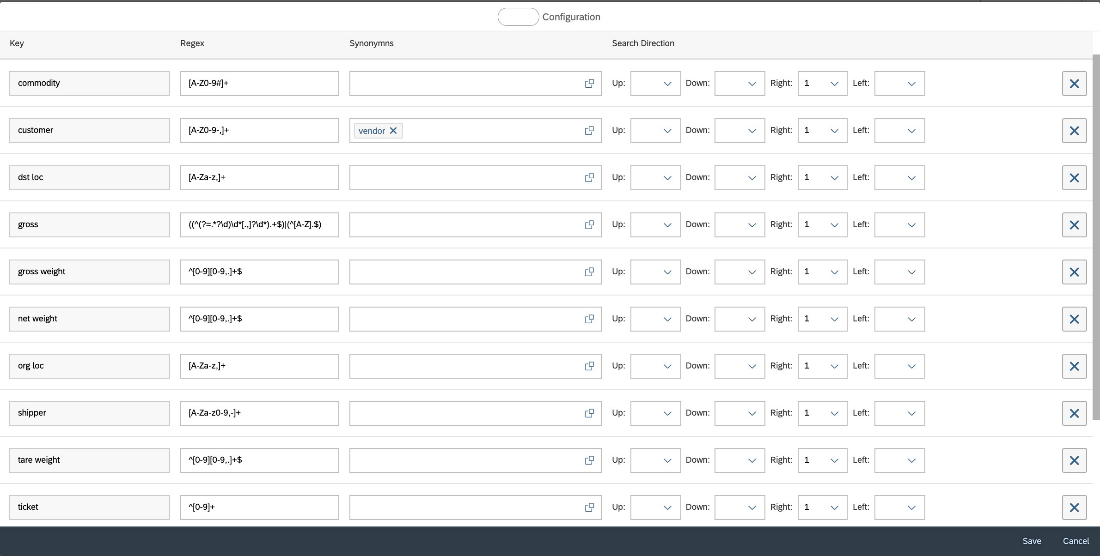
1. Extract all words and word locations from the transformed document image. In this case, an open-source optical character recognition algorithm called ‘Tesseract’ is used to extract the words and word coordinates information.
2. Check if the keys have multiple words. If so, identify and merge the multi-word keys to form a single entity.
3. Calculate the centroid of all the word or entity locations as shown in the example below. A centroid is the arithmetic mean of all the positions in a given shape (rectangle in this case) box. The red star in the center of the box from the following figure represents the centroid of the particular entity.

#### 

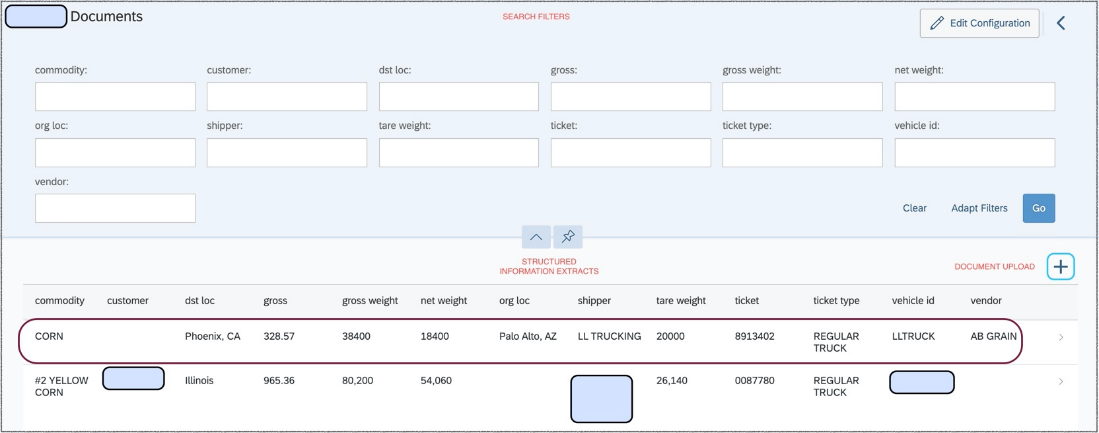
#### **Key-Value extraction**

1. Based on the resolution of the input document, set the radius-of-influence (ex: 1/4th of document width in the x-direction) and draw a hypothetical circle with a center around the key’s centroid to search for the respective value within this space.
2. Within the circle-of-influence around the key, identify the word/s next to the ‘key’ that satisfies the associated regex expression. It is not uncommon to find multiple results within this search.
3. To narrow down the results, set a pre-defined priority based on the angle or direction between the key and the identified value’s centroid. It is common to find the value either to the right side or the bottom of the key location. In this case, the priority angle ranges between 360o – 1o in a clockwise direction.
4. Calculate pair-wise angles between the key and all the identified values that satisfied regular expression.
5. Identify the first value that satisfies the criteria based on the priority angle that is set in step 3.
6. Break the process once the value has been identified and associate the output to the corresponding key. Note that in some cases, it is possible to have multiple words within the value. In this case, if the expected words are few (Ex: < 5-word entities), they can be extracted using regular expression and the distance between the adjacent words. Or if there are more than a handful of words or multiple lines that need to be extracted (Ex: > 5-word entities) then they can be extracted using text-blob analysis which uses morphological image processing techniques to determine the text blob.
7. Continue the above process for the next key for value extraction until you extract all the key-value pairs.

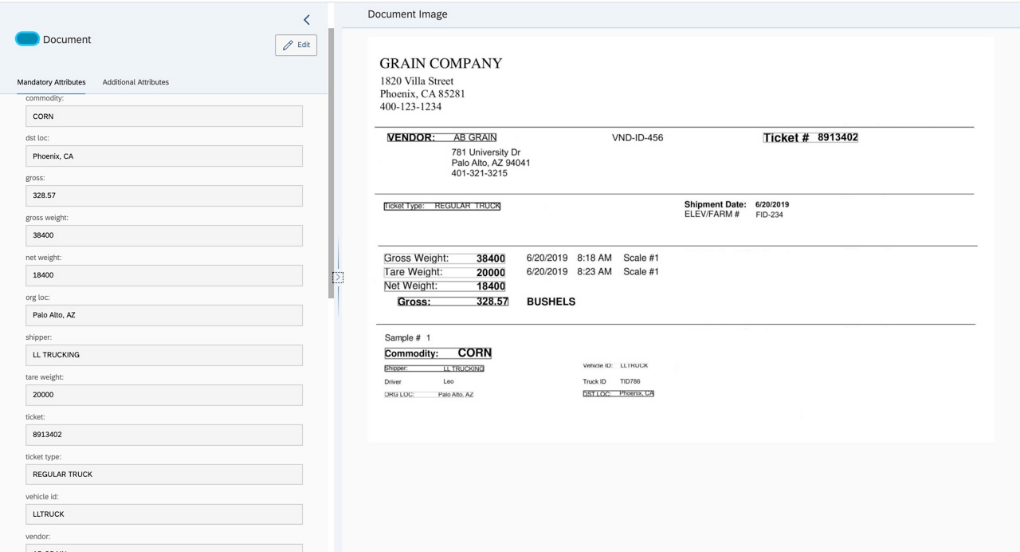
Below are some screenshots of the application that was built to define configurations, upload a document, run and display the extraction content.



Configuration Screen



Application home page that displays filter, document upload options, and the information extracts



Information extracted from the document image along with the location of the extracts

**OUTCOMES/DELIVERABLES**

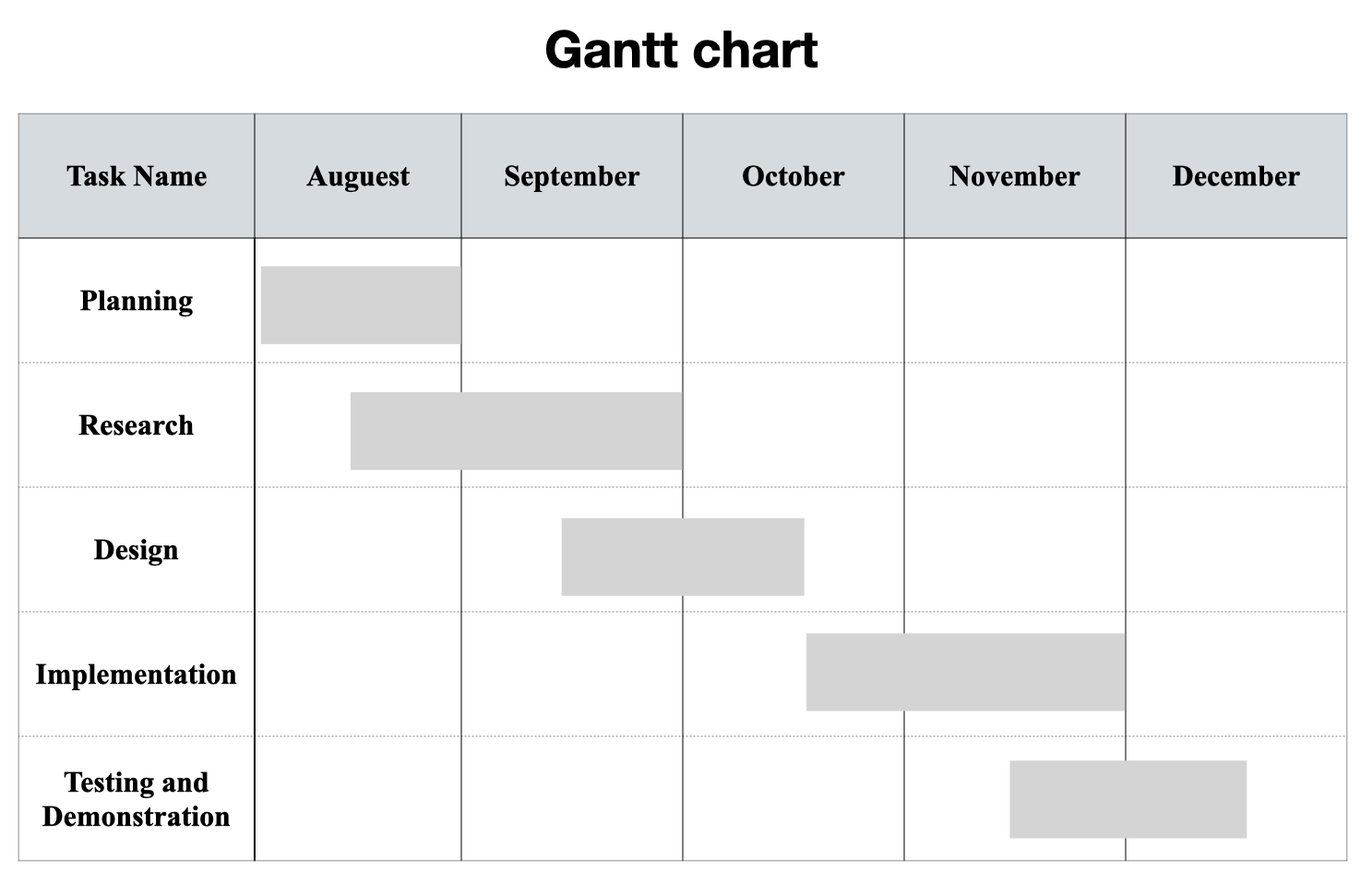
Successfully addressing the "need to extract the key-value fields coming from random layout" problem domain will yield a versatile solution that transforms diverse data types, like GST bills and shop receipts, into actionable insights.

This effort will result in a smart system that transforms messy data into organized forms like PDF, Word documents, spreadsheets (CSV), or structured data (JSON), saving time. These formats will also enable the use of the data as input for various other purposes, enhancing its utility.

**PROJECT POTENTIALS**

We would like to emphasize that our goal is to demonstrate that, good model performance can also be drawn by using alternative and training-free methods based on the problem we are trying to solve. Specifically, the project focuses on extracting key-value pairs out of the unstructured business documents like scanned bill-of-lading.

**PROJECT PLAN**



**REFERENCES**

[1] Raghuveer Kanneganti , [A Training-Free and Layout-Agnostic Approach to Extract Key-Value Pairs from Business Documents | SAP Blogs](https://blogs.sap.com/2020/11/06/a-training-free-and-layout-agnostic-approach-to-extract-key-value-pairs-from-business-documents/).

[2] Glossary of Important GST Terms. https://tallysolutions.com/gst/gst-glossary-terms-definitions/