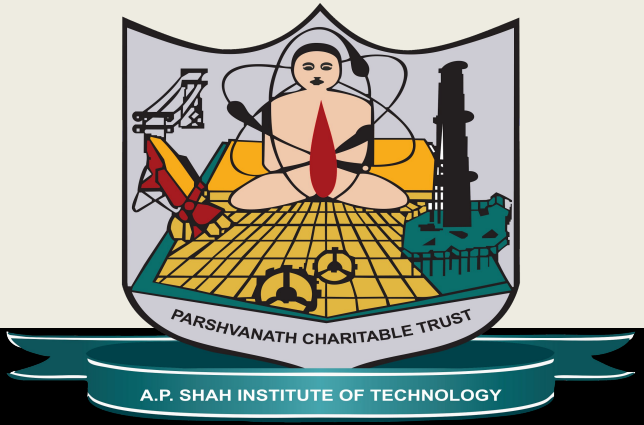


Department of Information Technology
NBA Accredited

A.P. Shah Institute of Technology

G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic Year 2020-2021

A Project Report on

A. I. Based Document Digitization System

Submitted in partial fulfillment of the degree of
Bachelor of Engineering(Sem-8)
in

INFORMATION TECHNOLOGY

By
Sujoy Dev(17104036)
Rashmi Shetty(17104070)
Priya Naik(17104021)

Under the Guidance of
Sameer Nanivadekar & Kiran Deshpansde

1.Project Conception and Initiation

1.1 Abstract

- One of the challenges faced by every corporate industry is maintenance of records, mainly non-digitized type, i.e. hard copies and prints
- While a few of these documents are digitized, the are some which are very crucial in nature and require high level of maintenance, such as, ration card, marks sheet, or etc. Furthermore there are billing invoices.
- Storing these essential documents on a server can be a solution but, it requires manual labour and verification, which might lead to the misplacement of data during the process.
- Our application can act as a maintenance provider by extracting essential data from these documents and storing them in data structures.
- This system will reduce the amount of issues of traditional document maintenance which involves manual verification of data by the human eye, since once the data is extracted it can be verified and tested against a set of rules

1.2 Objectives

- Reduce document maintenance and information retrieval efforts from documents such as invoices, purchases orders, maintenance records, etc. by developing an ecosystem by using **machine learning, Computer Vision** for document data identification, extraction and validation.
- Gain necessary information from the documents and store the data as document format for easy storage, matching and verification of data.
- Update the knowledge-base regularly to gain newer information from the documents.
- Platform to train and test with newer document types.
- Reduce paper dependency and go digital.
- Create a cross platform web application for packaging the technology

1.3 Literature Review

Wei Ruan and Won-sook Lee [5], built a Named Entity Recognition medical imaging procedure recognition system based on conditional random fields (CRF) model with word-based, part-of-speech. The NER model has been trained on a custom annotated dataset of medical notes from I2B2 with the F1 score up to 0.923 for recognizing medical imaging procedure entities. This system can be used to add and recognize new entities by simply

Lu, H. et al.[1], have conceived a method to for better preprocessing of shadowed text images, for which the character recognition performance of Tesseract drops significantly. In this paper, we propose a new method to process the shadowed text images for the Tesseract's optical character recognition engine. First, they performed a local adaptive thresholding to transform the document to gray-scale image into a binary image to capture the contours of texts. Now in order to get rid of the salt-and-pepper noise in the shadow areas they applied a double-filtering algorithm, in which a vertical and horizontal projection method is method is used to remove the noise between texts and after that median filter removes the noise within characters.This type of preprocessed data when provided to Tesseract OCR produces much better result.

1.4 Problem Definition

- To create system for digitization and maintenance of documents. Institutions/users must be able to upload documents(expense bills etc.) and the important data from these documents must be extracted as key pair data (JSON). If the data is not extracted from the documents the users must be able to select the data themselves for training the model.
- Store everything in a centralized repository.

1.5 Scope

- Document digitization make managing, editing, creating and organizing documents easy.
- Digital files can be accessed anywhere, allowing real-time collaboration on a project
- Easy access to online documents

1.6 Technology stack

- Python3 - backend, ML
- NodeJs - file server
- ReactJS - Client/frontend
- MongoDB - Database
- PM2 - Process Manage manager -

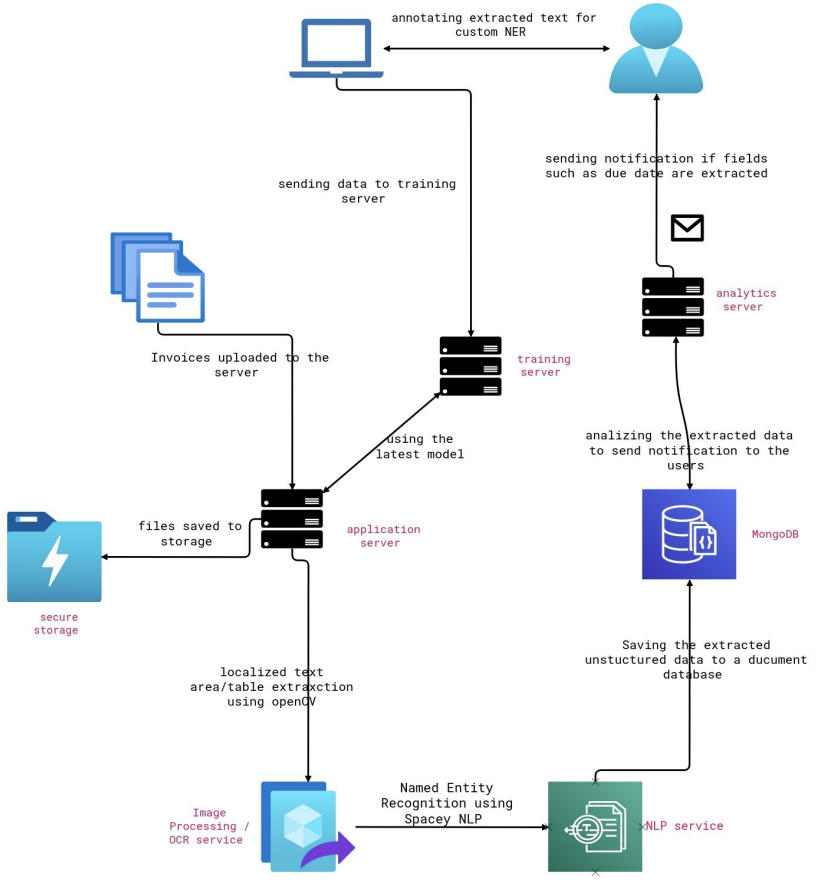
1.7 Benefits for environment & Society

- Document digitization is useful for transactional data management ie banks, manufacturers,
- Automatic Data gathering
- Invoice segmentation etc.
- It reduces human effort

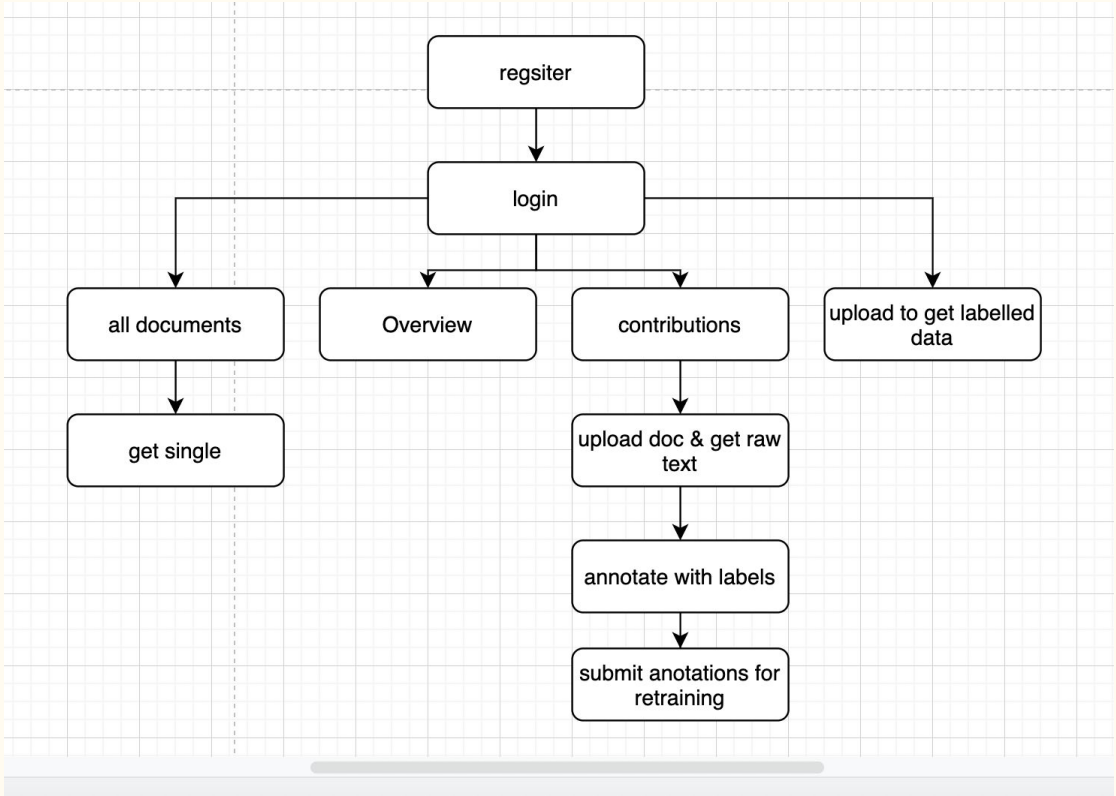
2. Project Design

2.1 Proposed System

A multi server architecture with
Dedicated training and analytics
Server instances



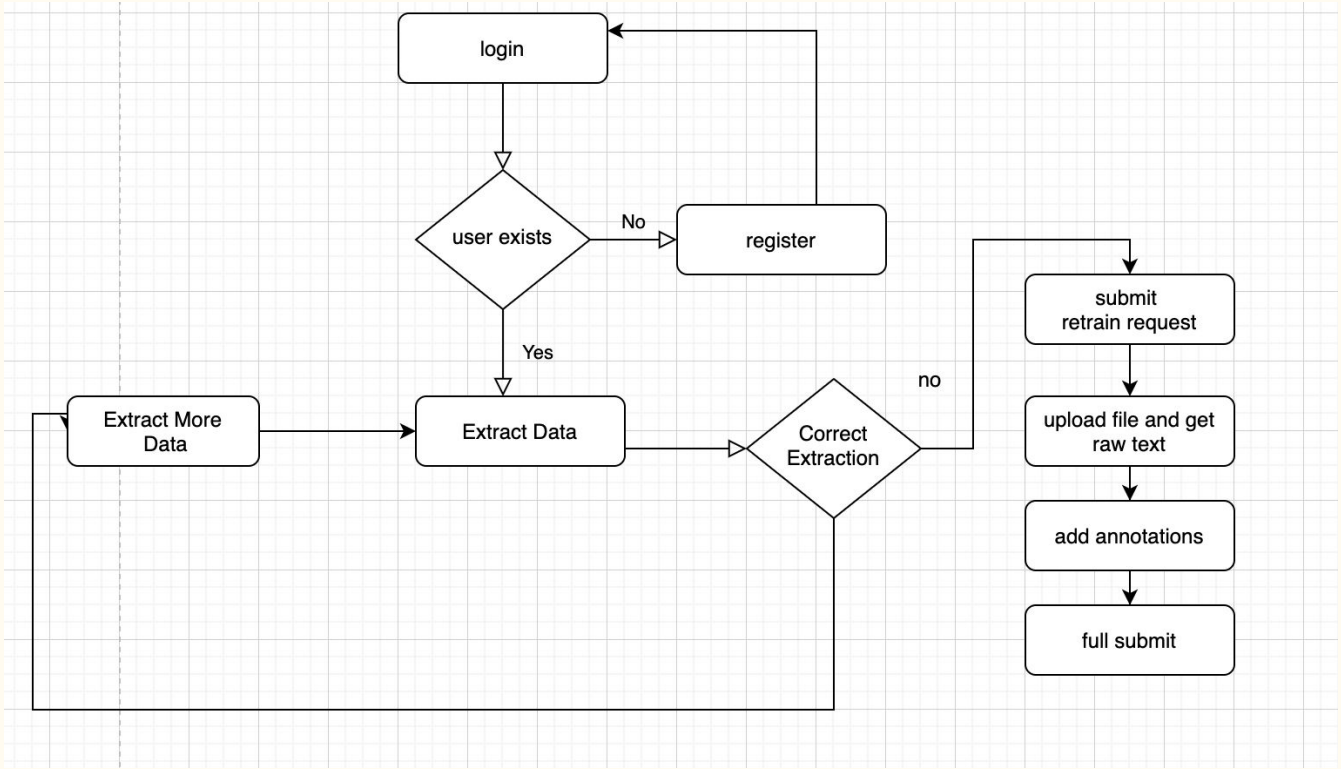
2.2 Design(Flow Of Modules)



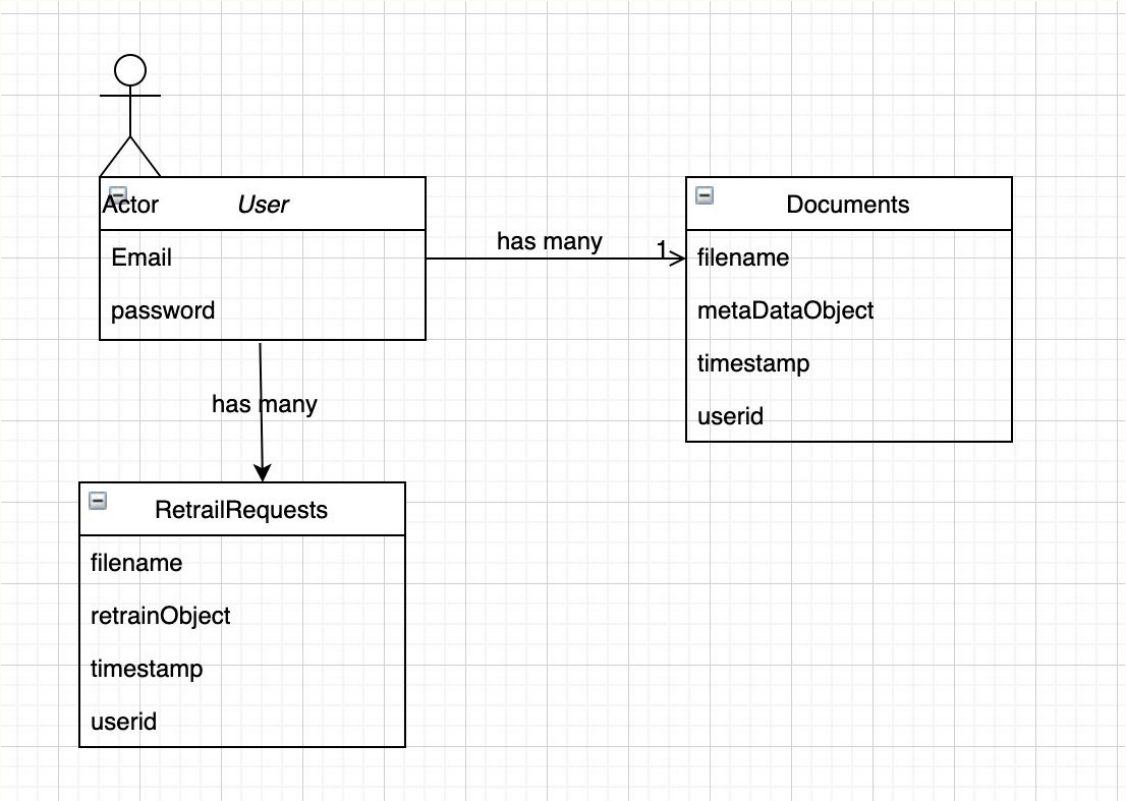
2.3 Description Of Use Case

- As a user i want yo extract key value pairs from documents for personal needs
- I must be able login to a portal
- Upload my document
- Get the key value pair as output containing document metadata
- If the correct data is not extracted Is must be able to teach the system about data I want

2.4 Activity diagram

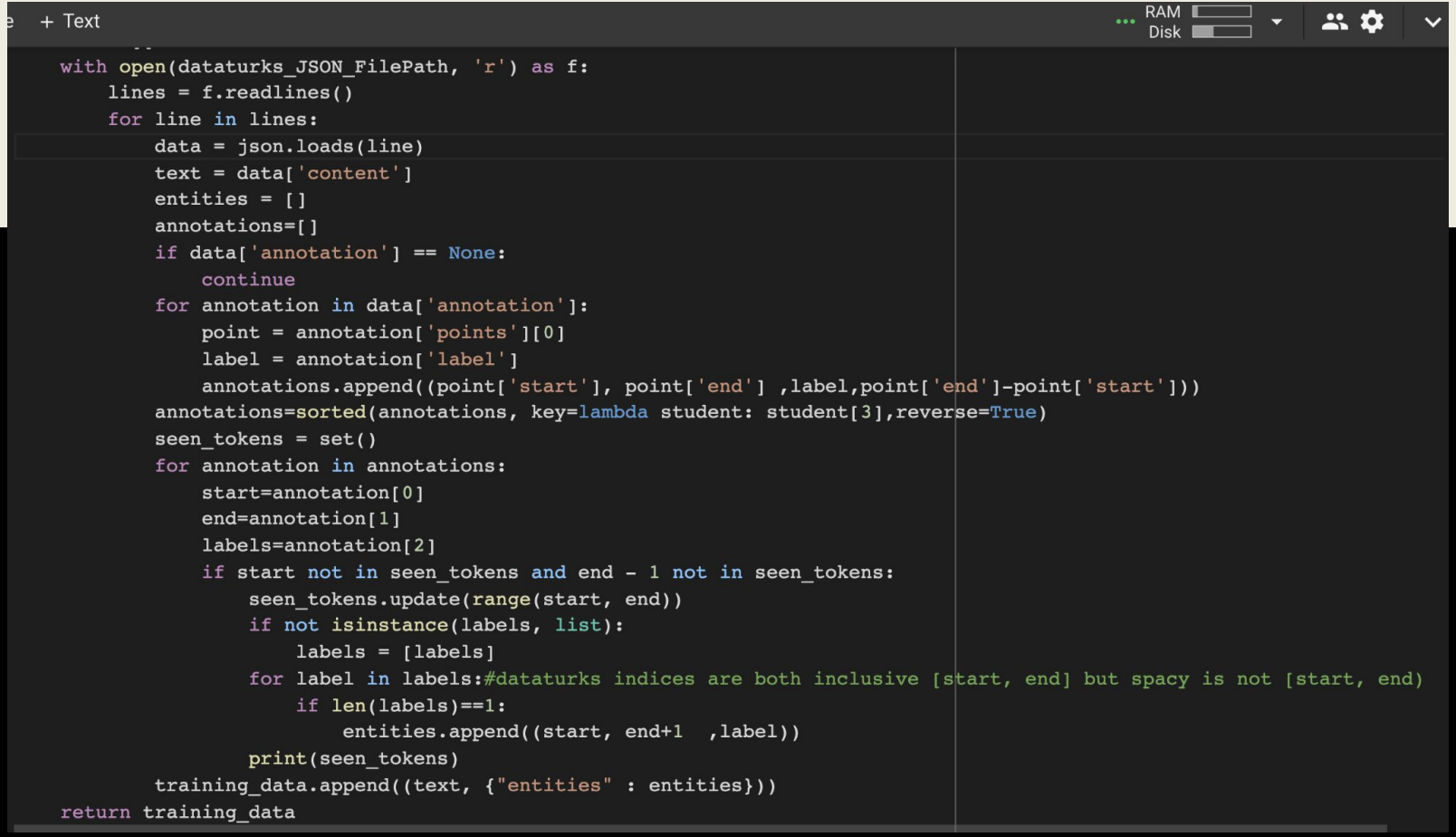


2.5 Class Diagram



3. Implementation

3.1 Data Validation



3.2 NLP update

```
# nlp.create_pipe works for built-ins that are registered with spaCy
if 'ner' not in nlp.pipe_names:
    ner = nlp.create_pipe('ner')
    nlp.add_pipe(ner, last=True)


# add labels
for _, annotations in TRAIN_DATA:
     for ent in annotations.get('entities'):
        ner.add_label(ent[2])

# get names of other pipes to disable them during training
other_pipes = [pipe for pipe in nlp.pipe_names if pipe != 'ner']
with nlp.disable_pipes(*other_pipes):  # only train NER
    optimizer = nlp.begin_training()
    for itn in range(10):
        random.shuffle(TRAIN_DATA)
        losses = {}
        for text, annotations in TRAIN_DATA:
            nlp.update(
                [text],  # batch of texts
                [annotations],  # batch of annotations
                drop=0.2,  # dropout - make it harder to memorise data
                sgd=optimizer,  # callable to update weights
                losses=losses)
```

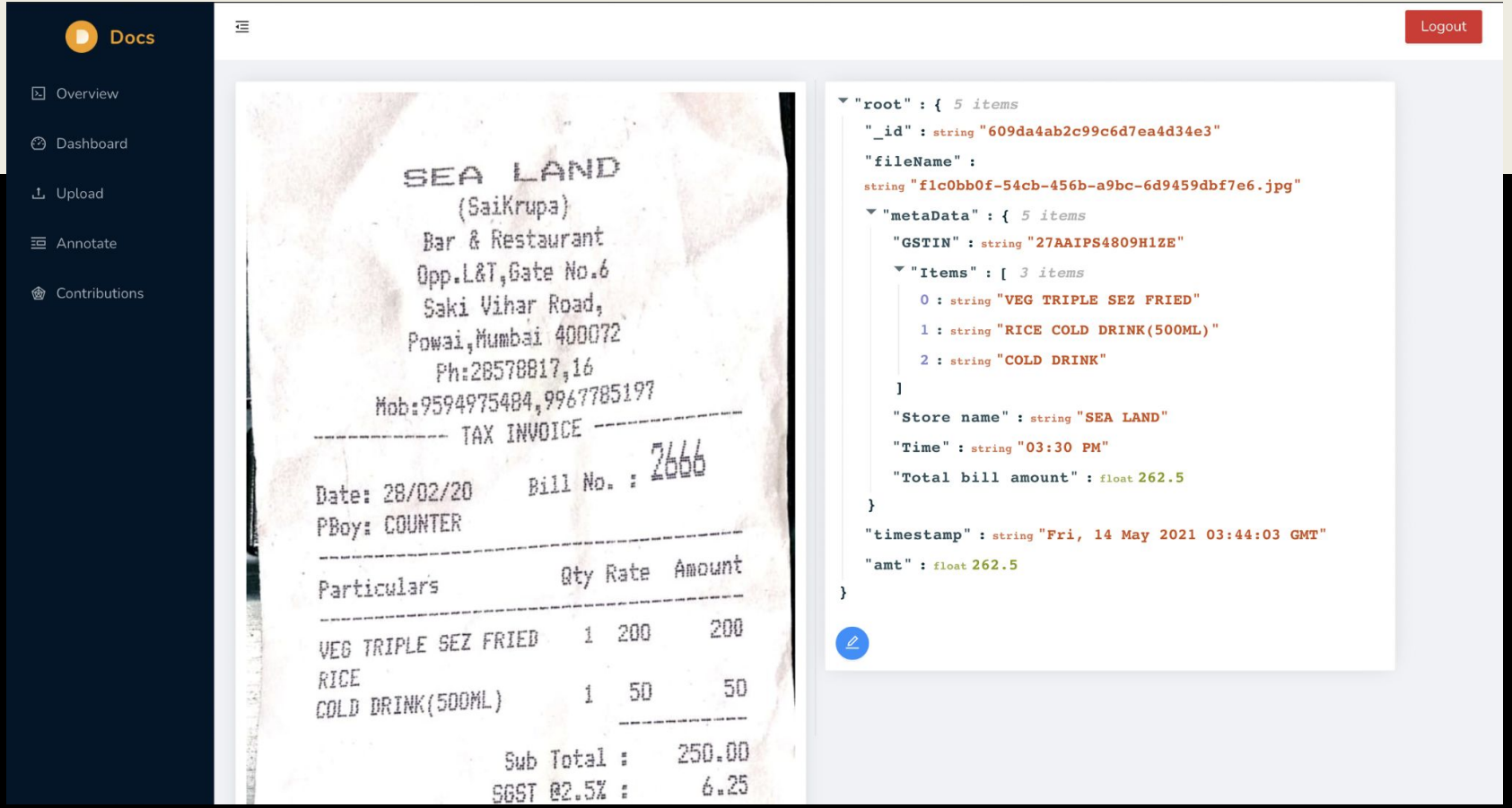
3.2 NLP update

```
Statring iteration 13
{'ner': 2883.609693677259}
Statring iteration 14
{'ner': 2975.905111826582}
Statring iteration 15
{'ner': 3317.8227595397307}
Statring iteration 16
{'ner': 2537.08402598031}
Statring iteration 17
{'ner': 2588.0659047520767}
Statring iteration 18
{'ner': 2496.9433751385095}
Statring iteration 19
{'ner': 2227.142730190494}
Statring iteration 20
{'ner': 2168.7678332011706}
Statring iteration 21
{'ner': 2009.1183573028804}
Statring iteration 22
{'ner': 2242.1861287573984}
Statring iteration 23
{'ner': 2312.8576085695563}
Statring iteration 24
{'ner': 2182.470131357508}
Statring iteration 25
{'ner': 1919.4904161572138}
Statring iteration 26
{'ner': 2110.46642770994}
Statring iteration 27
```

```
[ ] nlp.to_disk("models")
```

```
Statring iteration 89
{'ner': 972.9578853187577}
Statring iteration 90
{'ner': 890.5482174622543}
Statring iteration 91
{'ner': 943.3811822350007}
Statring iteration 92
{'ner': 820.4850192725304}
Statring iteration 93
{'ner': 861.5052266938048}
Statring iteration 94
{'ner': 915.9826556075071}
Statring iteration 95
{'ner': 867.7181525457657}
Statring iteration 96
{'ner': 758.3394671855701}
Statring iteration 97
{'ner': 737.7130508037365}
Statring iteration 98
{'ner': 935.5589474180045}
Statring iteration 99
{'ner': 736.6540348062272}
```

3.4 Digitization



3.5 Custom Training Request

Docs

- Overview
- Dashboard
- Upload
- Annotate
- Contributions

Upload file To get raw text

Choose file 3.jpg

Get Text

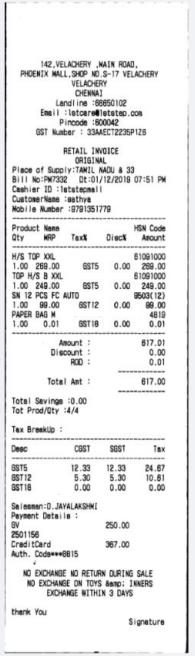
shop address

Add an entity!!

Add

GSTIN

142, VELACHERY ,MAIN ROAD, PHOENIX MALL, SHOP NO.S-17 VELACHERY VELACHERY CHENNAI shop address Landline :66650102 Ema il :1stcare@lststep.com Pincode :600042 GST Number : 33AAECT2235P GSTIN 1Z6 RETAIL INVOICE ORIGINAL Place of Supply:TAMIL NADU & 33 Bill No: PM7332 Invoice number Dt: 01/12/2019 Date 07:51 PM Time Cashier ID :1ststepmall CustomerName :sathya Mobile Number :9791351779 Product Name HSN Code Amount Qty MRP Taxx DiscX H/S TOP Items XXL 1.00 269.00 T OP H/S B XXL Items 1.00 249.00 SN 12 PCS FC AUTO Items 1.00 99.00 PAPER BAG M Items 1.00 0.01 61091000 0.00 269.00 61091000 0.00 GST5 GST5 249.00 9503(12) 0.00 99.00 4819 0.01 GST 12 GST 18 0.00 Amount : Discount : ROD : 617.01 0.00 0.01 Total Amt : 617.00 Total bill amount Total Savings :0.00 Tot Prod/Qty :4/4 Tax Breakup : Desc CGST SGST Tax GST5 GST 12 GST 18 12.33 5.30 0.00 12.33 5.30 0.00 24.67 10.61 0.00 Salesman:D.JAYALAKSHMI Payment Details: GV 2501156 CreditCard Auth. Code***8815 250.00 367.00 NO EXCHANGE NO RETURN DURING SALE NO EXCHANGE ON TOYS & INNERS EXCHANGE WITHIN 3 DAYS thank You Signature



4. Testing

- Client Side UI - Manual Testing to understand user experience
- Backend - automated Testing to simulate API calls

5. Result

- Very Large data set


```
{
  "Date": "28-Nov-19",
  "GSTIN": "33AATCG73851125",
  "Invoice number": "LTN02B1920003774",
  "Items": [
    "VEG RICE BOWL MEA",
    "CLASSIC LEMONADE",
    "MILD BASTING\nGAL",
    "VEG RICE BOWL MEA",
    "CLASSIC LEMONADE",
    "MILD BASTING\n",
    "VEG RICE BOWL MEA\nGAL",
    "CLASSIC LEMONADE",
    "MILD BASTING GAL",
    "QUARTER CHICKEN M",
    "CORN ON THE COB",
    "CLASSIC LEMONADE",
    "MILD BASTING GAL"
  ],
  "Store address": "Unit No: UG-41,PMC,Old Door.No. 66,
  "Store name": "Calito's",
  "Store name-1": "Galito's",
  "Time": "16:47",
  "Total bill amount": 762.0
}
```

Moderate amount of data


```
{
  "GSTIN": "33AAECT2235P 1Z6",
  "Invoice number": "PM7332",
  "Items": [
    "H/S TOP XXL",
    "TOP H/S B XXL",
    "PAPER BAG M"
  ],
  "Store address": "142, VELACHERY ,MAIN ROAD,\r
  "Time": "07:51 PM",
  "Total bill amount": 269.0
}
```

Unknown document schema

```
{
  "Items": []
}
```

6. Conclusion and Future Scope

- With the power of named entity recognition and image processing we were able build a web based application to extract key value pairs from bills and invoices. Extraction is not always effective since there are many limitations such as image quality, small datasets, non expressive data, etc. Retrain queries are mode of gathering additional data and updating the existing model wit that. Such system can take up the time consuming task of data entry and make it automated and faster. By reducing paper dependency we may be able to be truly digitized
- Further with the use of templates and visual markers, template base positional data a can be implemented. A strong intermediary between the user submitted training data and the training server will be effective in filtering out invalid data, hence aiding inadata cleaning and pre-processing. Industrial invoices can be targeted with sufficient training

References

1] Lu, H., Guo, B., Liu, J., Yan, X. (2017).
[2] Sidhwa, H., Kulshrestha, S., Malhotra, S., Virmani, S. (2018).
[3] K.M. Yindumathi, Shilpa Shashikant Chaudhari, R. Aparna. (2020).
[4] Zhang, J., Ren, F., Ni, H., Zhang, Z., Wang, K. (2019).
[5] Wei Ruan; Won-sook Lee (2018).
[6] Internet Archive,
[7] Wikipedia, towardsdatascience,
https://towardsdatascience.com/pre-processing-in-ocr-fc231c6035a7, last accessed 19.10.20
[8} Dataset, https://github.com/zzzDavid/ICDAR-2019-SROIE.git

Paper Publication

- Paper entitled “A. I. Based Document Digitizations” is submitted at “Second International Conference on Secure Syber computing and Communications 2021” And “8th International Conference on Smart Computing Communications”

Thank You