A Mini-Project Report on

System for secure document digitization and maintenance

Submitted in fulfilment of Mini-Project (ITM605) of

Semester VI

in

Information Technology

By

Sujoy Dev (17104036)

Rashmi Shetty (17104070)

Priya Naik (17104021)

Under the Guidance of

Dr. Sameer Nanivadekar

Prof. Kiran B. Deshpande

Department of Branch Name
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI
2019-2020

CERTIFICATE

This is to certify that

Mr. Sujoy Dev	Student ID 17104036
Ms. Rashmi Shetty	Student ID 17104070
Ms. Priya Naik	Student ID 17104021

has completed all the specified work for submission in Mini-Project (ITM605) of Semester VI, as laid down by the University of Mumbai in a satisfactory manner within the premises of the Institute during the academic year 2019-20.

Prof. Kiran Deshpande
Co-Guide

Dr. Sameer Nanivadekar
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place: A. P. Shah Institute of Technology, Thane
Date:

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academics with honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Sujoy Dev	Rashmi Shetty	Priya Naik
17104036	17104070	17104021

Date:

ABSTRACT

One of the challenges faced by every industry is the maintenance of records, mainly non-digitized type, i.e. hard copies and prints. While a few of these documents are digitized, they are some which are very crucial in nature and require a 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 labor 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.

These data structures can be later used for performing **semantic searches** on the documents, which will have been already tagged by our software application.

Semantic searches will be extremely useful while fetching the required documents using only important tags such as <title>,<document_holders_name>,<date>, etc. This will be useful for quickly tracking the documents and the data.

This system will reduce the number 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 These rules will help us in locating document discrepancies.

CONTENTS

INTRODUCTION

In this project, we present the modeling, implementation, and development of secure document digitization and maintenance platform for documents. The proposed software application will be used to train and test the model by supplying images/pdfs (converted to images), Through this project we will provide semantic search facilities for images by extracting keyphrases/data via **Natural Language Processing**. Furthermore, a cross-platform application can be built on the same technology for scanning documents and extracting data from them and storing them for later use, on the go.

LITERATURE REVIEW

Document Specific Supervised Keyphrase Extraction With Strong Semantic relations

Authors: Huiting Liu, Wang Peng Zhao, Xindong Wu

Findings: Keywords can be extracted via natural language processing after the text has been extracted from the documents. Keyphrases provide semantic metadata that summarizes and characterize documents.

Advantages: Semantic relation can be extracted by NLP application

Disadvantages: No document summarization Publication: IEEE - 22nd October 2019

Keyword Extraction Through Contextual Semantic Analysis of Documents

Authors: Terry Ruas, Willaim Grosky

Findings: Traditional approaches make use of techniques that rely on analyzing just the syntactic aspect of texts, ignoring the meaning they convey and more importantly, the semantic effect of one word over another. Hence we ought to use the following two approaches, the first approach extends the concept of Word Sense Disambiguation (WSD), and the second approach enhances the theory behind traditional lexical chains. These applied techniques also consider distinct levels of abstraction with respect to the meanings of words, in addition to the context in which they appear.

Advantages: Word sense disambiguation and traditional lexical chains and give meaning to the context

Disadvantages: Synset provided by the different algorithms(BSD & FLC) contaminate the score MediaWiki categories Publication: 2017 Conference: the 9th International Conference Research Gate

A System for Handwritten and Printed Text Classification

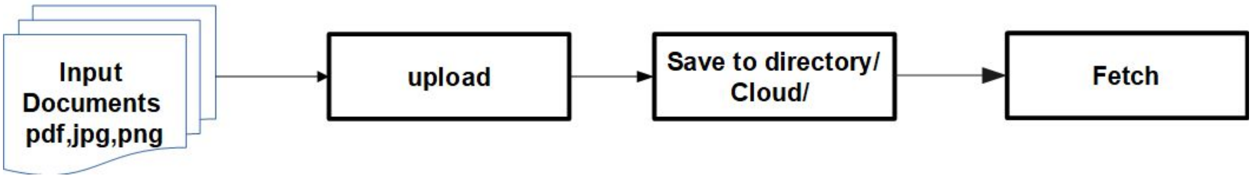
Authors: Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala

Findings: Various features described in her are extracted from these word images and these are used to analyze the differences between machine print and handwritten text. There is no clear separation between machine printed and handwritten text density values, but pixel density can be used to augment the classification efficiency.

Advantages: Can classify between machine printed and handwritten alphabets with 98.6% accuracy

Disadvantages: Only English language support for classification Publication: 2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)

EXISTING SYSTEM ARCHITECTURE / WORKING



- In traditional software digitization techniques, documents, transcripts are only uploaded to a secure storage but the information in those documents remains unread.
- Hence true digitization is not achieved

PROBLEM DEFINITION

To create a **system for secure 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 and stored for **semantic search** later. If the data is not extracted from the documents the users must be able to select the data themselves. Store everything in a centralized repository. Send the users insights about their document,i.e., expiration, validity, etc.

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 **JSON** for easy storage, matching and verification of data.

- Filter through numerous, large document sets within a matter of seconds by entering the keywords to be searched only, such as <date>, <ownerInfo>, through **Natural Language Processing** etc.

- Update the knowledge-base regularly to gain newer information from the documents, using **ELK Stack**.

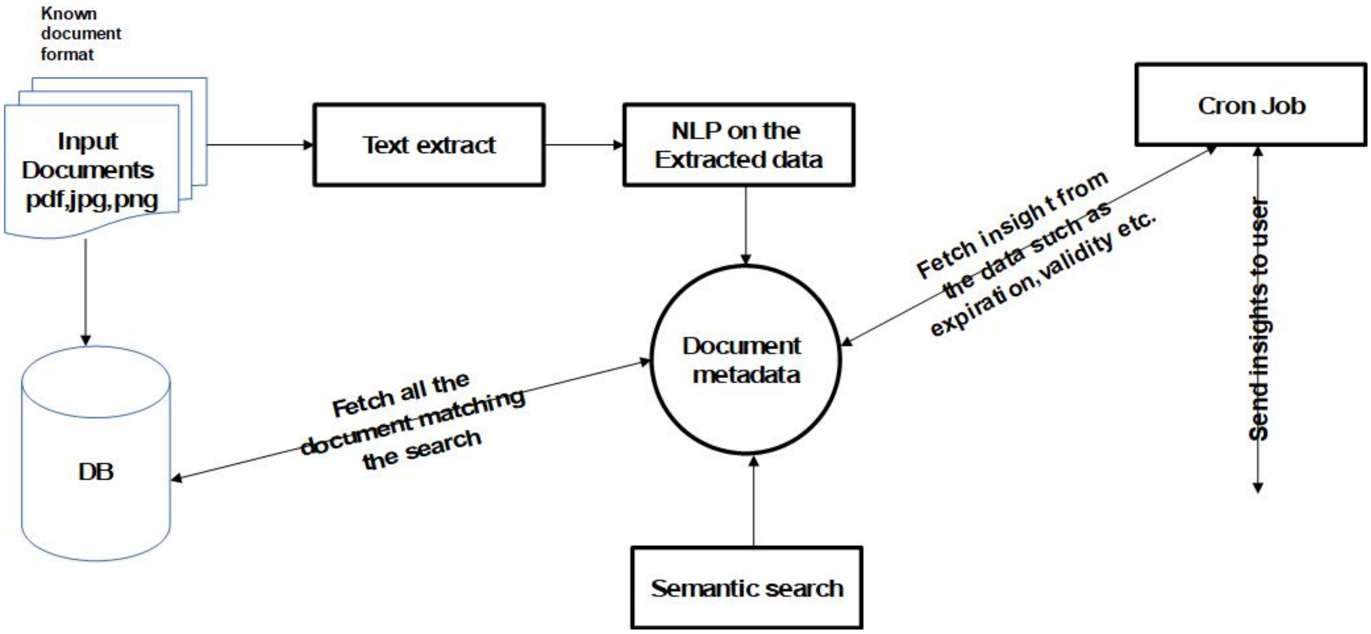
- Platform to train and test with newer document types.

- Reduce paper dependency and go digital.

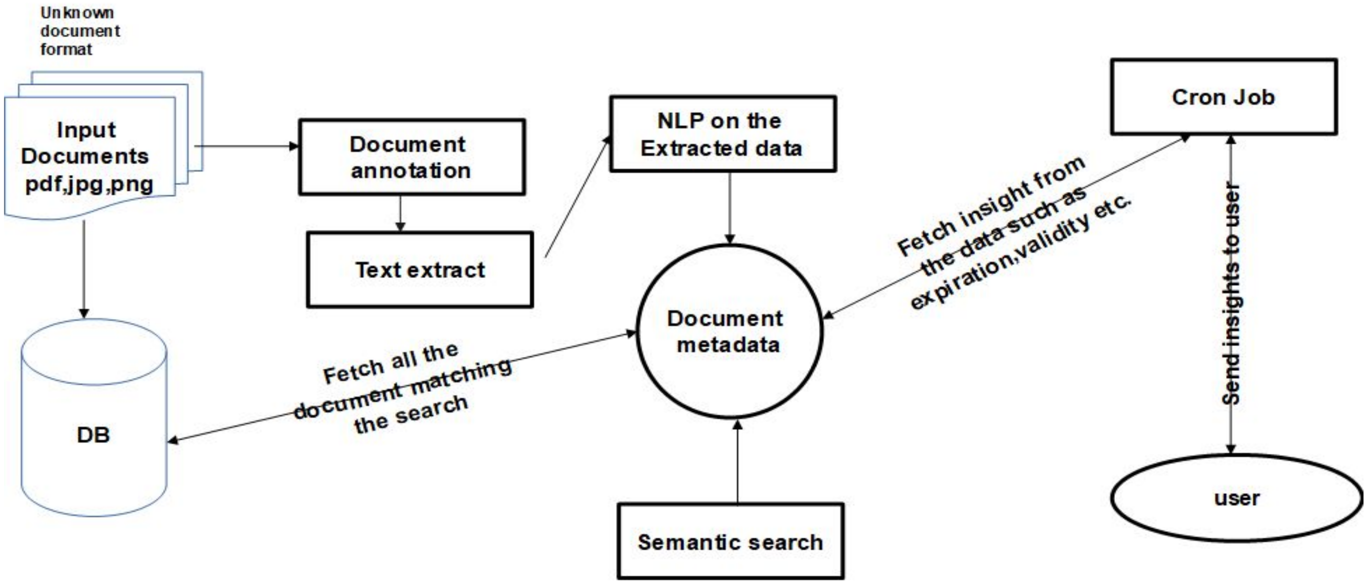
- Create a cross platform desktop application for packaging the technology

PROPOSED SYSTEM ARCHITECTURE / WORKING

case 1: known document format


Known document format
Input Documents pdf,jpg,png
Text extract
NLP on the Extracted data
Cron Job
Fetch insight from the data such as expiration,validity etc.
Send insights to user
Document metadata
Fetch all the document matching the search
DB
Semantic search

case 2: unknown document format


Unknown document format
Input Documents pdf,jpg,png
Document annotation
Text extract
NLP on the Extracted data
Cron Job
Fetch insight from the data such as expiration,validity etc.
Send insights to user
Document metadata
Fetch all the document matching the search
DB
Semantic search
user

PROPOSED TECHNOLOGY STACK

- Extraction and analysis
  - Python
  - OpenCV
  - Keras/Tensorflow
  - Natural Language Toolkit (NLP)

- Application Interface
  - React.Js/Flutter
  - Mysql/Postgres/Mongo
  - Node.Js

REFERENCES

[1] https://ieeexplore.ieee.org/document/8879476

Huiting Liu, Wang Peng Zhao, Xindong Wu. *Document Specific Supervised Keyphrase Extraction With Strong Semantic Relations. IEEE - 22nd October 2019*

[2] https://www.researchgate.net/publication/324074602_Keyword_Extraction_Through_Contextual_Semantic_Analysis_of_Documents

Terry Raus, William Grosky. *Keyword Extraction Through Contextual Semantic Analysis of Documents. 2017 Conference: the 9th International Conference Research Gate*

[3] https://ieeexplore.ieee.org/document/8359046

Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala. *A System for Handwritten and Printed Text Classification. 2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)*