Document Classification and Data Extraction

Introduction:

Financial institutions will process and keep a massive volume of consumer documents. These documents are crucial to the company's operations, and there are statutory obligations on how they must be processed and stored. To identify the documents and extract data from them, a significant amount of human processing is required.

Problem Statement:

1) Based on the input file, the documents must be identified, classified, and divided into multiple groups. The user can submit a single file (image/pdf/word document) that contains many documents. Create a library that accepts a user supplied file and recognises and splits numerous documents existing in the file.

Documents to be classified and split are:

- PAN
- Aadhaar (Aadhaar front, Aadhaar back)
- Bank Statement
- ITR/Form 16
- Customer Photograph (Selfie)
- Utility Bill (Power, Water, Gas, Landline etc)
- Cheque Leaf
- Salary Slip/Certificate
- Driving License
- Voter ID
- Passport

2) Once document is classified and split, create a library which accepts split document and extracts the data from it.

The library built should be scalable and secure to process millions of files.

Note:

The first problem statement carries more weightage. If all or most of the documents in problem statement 1 have been addressed, go on to solve problem statement 2

Guidelines for Submission:

- The participants have to prepare a report consisting of 2-3 pages covering which all documents your library is able to identify and extract data ● Approach/Design of your solution
- Attach the code
- The tentative deadline for submission is 25th January 2023
- Link for submission: ecell.in/esummit/i-hack