Final report

- The code imports necessary libraries and sets up logging.
- It initializes the Adobe PDF Services credentials using a JSON file and creates an execution context.
- A temporary directory is created to store extracted files.
- A loop iterates over a range of 100, representing the number of test data cases to process.
- Inside the loop, an ExtractPDFOperation is created for each iteration and configured with the input source file.
- Extraction options are set, specifying the elements to extract (text and tables), rendition options, and table structure format (CSV).
- The operation is executed, and the result is saved as a temporary ZIP file.
- After the loop, a final ZIP file is created to consolidate all the extracted files.
- Another loop iterates over the range of 100 to unzip the files from the final ZIP file.
- Inside the loop, each ZIP file is opened, and its contents are extracted to a destination path.
- The code then imports additional libraries for further data processing, such as regular expressions (re), CSV, and pandas.
- A list called df_list is created to store the extracted data as dictionaries for DataFrame creation.
- Another loop iterates over a range from 2 to 99 (excluding 100) to process the extracted files.As 0 and 1 have already been extracted.
- Inside the loop, a JSON file is opened and loaded as a dictionary.
- Variables are initialized to store the extracted information.
- A loop iterates through the JSON elements, and based on the index (idx), the relevant information is extracted and assigned to the corresponding variables.
- A text file is written with the extracted values for further processing.
- The text file is opened, and its content is assigned to the content variable.
- Regular expressions are used to perform pattern matching on the content and extract specific information, such as phone numbers, email addresses, due dates, invoice details, and tax information.
- The extracted information is assigned to the respective variables.
- If any information is missing or not found, default values are assigned.
- The extracted information is added as a dictionary entry to the entries list.
- After processing all the test cases, the entries list contains all the extracted data.