**Efficient Data Mining and Information Extraction**

Shreya Manepalli

Dept. of Computer Science, PES University, Bangalore

Report submitted to GKN Aerospace India Pvt. Ltd. upon completion of the internship

**TABLE OF CONTENTS**

Abstract

1. Introduction

1.1 Data, structured and unstructured

1.2 Challenges involved in analyzing unstructured data

1.3 Text mining and its importance

2. Project Overview

2.1 Non-Conformance Reports (NCRs) Data Extraction

2.1.1 Non-Conformance Reports

2.1.2 Project Importance

2.2 Color Codes Extraction from Images for the Fokker Plant

2.2.1 Project Importance

2.3 Generalized Data Mining from PDF and Word Documents

2.3.1 Project Importance

3. Objective

4. Methodology

5. Results

6. Testing and Performance

7. Conclusion

8. Future scope

**Abstract**

Every year, a huge volume of unstructured data in form of reports, documents, images etc. are generated at GKN Aerospace. This internship is aimed to leverage data mining techniques to extract valuable information from this vast repository of PDFs, word files, and images enabling the creation of datasets for data analysis using Python.

This work encompasses three distinct projects. In the first project, I extracted specific data from Non-Conformance Reports (NCRs) and organized it into an Excel file. This curated dataset can now be used for data analytics. The second project involved extracting color codes from images of wire harness and systematically storing them in a structured database. The third project revolved around generalized data mining from PDF and Word documents. These extraction algorithms are designed to efficiently capture various data elements, including tables, equations, relevant text, figure/table captions, and data from images. The Python codes developed can be adopted for similar usage with minimum changes.

**1. Introduction**

There is a huge volume of data which if mined and analyzed can provide valuable insights. In GKN Aerospace, huge volumes of reports and documents are generated every year. Thus, a lot of data is locked in those documents. These reports if mined properly can result in the development of a priceless database which in turn can be used for several engineering/aerospace applications.

The primary aim of this internship is to showcase the value and potential of data mining and information extraction from unstructured sources in a real-world industrial context. It will serve as a valuable record of the work accomplished during the two-month internship at GKN Aerospace and contribute to the knowledge base for future data-driven endeavors.

**1.1 Data, structured and unstructured**

Data refers to any collection of facts, statistics, measurements, observations, or information that can be recorded, stored, and analyzed. In the context of computing and technology, data is essential for various purposes, including decision-making, analysis, and generating insights. Data can take many forms, such as text, numbers, images, audio, video, and more. Based on their types, data can be classified into two types:

1. Structured Data:

Structured data refers to information that is organized and formatted in a specific and predefined manner. It follows a fixed schema or data model, where each data item has a well-defined data type and is organized in a tabular format. The schema provides a clear representation of the data, making it easy to understand and work with. Examples of structured data include data in relational databases, spreadsheets, and organized data in CSV/Excel sheets/SQL databases etc.

1. Unstructured Data:

Unstructured data refers to information that does not follow a predefined structure or format. It does not fit into traditional databases with rows and columns, and its organization and content may vary widely. Analyzing and extracting meaningful insights from unstructured data require specialized tools and techniques, such as natural language processing (NLP) for text data and computer vision for image data. Examples of unstructured data include text documents, images, audio files, social media posts, videos, emails, and web pages.

**1.2 Challenges involved in analyzing unstructured data**

Unstructured data generates immense business value, but most organizations have not been able to yield insights because there are so many challenges involved in analyzing unstructured data.

Major challenges in analyzing unstructured data are as follows -

1. This data cannot be analysed with conventional systems:Unstructured data cannot be analysed with current databases because most data analytics databases are designed for structured data, and are not equipped for unstructured data. Therefore, data analytics experts need to find new methods to locate, extract, organise and store data.
2. Unstructured data keeps expanding:Unstructured data continues to grow at an exponential rate. This large volume is going to be a huge challenge in analysing this type of data because the larger the data set, the harder it is to store and analyse data in a way that is timely and efficient.
3. Unstructured data is not always relevant: Ensuring data relevance is a biggest challenge when it comes to analysing unstructured data. This has a huge impact on the reliability and accuracy of findings.
4. Not all unstructured data is high quality:Unstructured data can be very uneven when it comes to quality. The lack of consistency in quality occurs because data is difficult to verify and, therefore, is not always accurate. For example, images and videos all qualify as unstructured data, but that does not make it useful for organisations who are looking for ways to improve sales.

**1.3 Text Mining and its Importance**

Text mining, also known as text data mining or text analytics, is the process of extracting useful information and knowledge from large volumes of unstructured text data. It involves using various techniques from natural language processing (NLP), machine learning, and statistical analysis to analyze, interpret, and derive meaningful insights from textual data. Text mining is crucial in unlocking the hidden value within unstructured text and is important in various fields and industries.

**2. Project Overview**

During the course of this internship, I worked on three different problems as discussed below.

**2.1 Non-Conformance Reports (NCRs) Data Extraction**

In the first project, I focused on extracting specific data from Non-Conformance Reports (NCRs) and organizing it into a structured Excel file. This curated dataset now serves as a valuable resource for data analytics and can be utilized to gain deeper insights into various products and processes.

**2.1.1 Non-Conformance Reports**

Non-Conformance Reports (NCRs) are crucial documents that record instances of non-compliance with quality standards or deviations from specified requirements.

**2.1.2 Project Importance**

Manually reading and entering data from NCRs into Excel can be time-consuming and error-prone. By developing automated data extraction algorithms, I significantly reduced the manual effort involved, leading to increased efficiency and accuracy in data processing.

The curated dataset created from the extracted NCR data now serves as a valuable resource for data analytics and decision-making. Analysts and stakeholders can utilize this structured data to gain deeper insights into various products and processes, identify trends, and take data-driven decisions.

**2.2 Color Codes Extraction from Images for the Fokker Plant**

The second project involved developing algorithms to extract color codes from images and systematically storing them in a structured database.

**2.2.1 Project Importance**

By automating the extraction of color information from images, I streamlined the processes that otherwise would have required manual identification and recording of color codes. This automation not only saves time but also reduces the chances of human errors.

**2.3 Generalized Data Mining from PDF and Word Documents**

The third project centered on the development of versatile extraction algorithms capable of efficiently capturing various data elements from PDF and/or Word documents. These elements include tables, equations, relevant text, figure/table captions, and data from images. The Python codes developed in this project can be adapted for similar data mining tasks with minimal modifications, increasing the overall efficiency of information extraction.

**2.3.1 Project Importance**

Manually extracting data elements such as tables, equations, relevant text, and captions from documents can be time-consuming. These generalized data mining algorithms have significantly reduced the time and effort required for these tasks, leading to improved efficiency in handling large volumes of data.

Analysts and stakeholders can access relevant information quickly, aiding in research, product development, compliance monitoring, and more.

**3. Objective**

The primary objective of this internship was to develop robust data mining and information extraction techniques to handle the vast volume of unstructured data generated at GKN Aerospace. By addressing the specific needs of various stakeholders, these techniques aimed to streamline data processing, enhance data analytics capabilities, and ultimately enable data-driven decision-making. The three distinct projects catered to the requirements of different stakeholder groups within the organization:

**Project 1: Non-Conformance Reports (NCRs) Data Extraction**

Stakeholder: The SPE Team at GKN Aerospace, responsible for analyzing a large number of Non-Conformance Reports (NCRs) that are generated.

Objective: The primary objective of this project was to develop a Python script to extract specific data from Non-Conformance Reports (NCRs) generated and analyzed by the SPE Team at GKN Aerospace. The script aims to automate the data extraction process, as the NCRs contain unstructured data which used to be manually retrieved for analysis. The extracted data will be organized into an excel file, making it suitable for data analytics.

**Project 2:** **Color Codes Extraction from Images for the Fokker Plant**

Stakeholder: The Fokker Plant, focusing on airplane wings and their associated color codes.

Objective: The objective of the second project is extract colors and associated codes from the pdf’s containing images about airplane wings. These color codes and numbers are present in specific formats within the images and can be stored systematically into Excel files which can be used for further data analysis. This project aimed to ease the work of the operator of manually mining the data and assist the Fokker Plant in quickly accessing and utilizing color code information, thereby enhancing their productivity and accuracy in handling data-specific tasks.

**Project 3: Generalized Data Mining from PDF and Word Documents for the GAI Team**

Stakeholder: The entire GAI Team at GKN Aerospace, encompassing various departments and individuals with diverse data mining needs.

Objective: To create generalized data mining algorithms capable of efficiently capturing diverse data elements such as tables, equations, relevant text, figure/table captions, and images from PDF and Word documents. By providing a versatile solution that can be adopted by anyone within the organization, this project aimed to empower the GAI Team with a powerful toolset for data extraction, facilitating a wide range of data mining applications.

Overall, the main objective was to deliver practical and adaptable data mining solutions that catered to the specific requirements of various departments within GKN Aerospace. By automating data extraction processes and structuring the extracted information into databases, the internship sought to decrease the manual effort of operators maximize data utilization, and data analysis, improve productivity and foster data-driven insights across different departments and stakeholders at GKN Aerospace.

**4. Methodology**

**Project1: Data Extraction from Non-Conformance Reports (NCRs)**

Data Collection and Preprocessing

The Python script accepts user input to determine whether the input is a single file or a folder containing multiple NCR PDF files. If the input is a single file, the script performs data extraction directly on that file. If it is a folder, the script iterates through all the PDF files present in the folder.

Modules Used

The script utilizes the tabula [1] library to extract tabular data from the NCR PDF files. For each PDF file, the script reads all tables on every page and stores them in separate sheets within an intermediate Excel file. Additionally, the script extracts specific information from the first page of the PDF files, including NCR numbers, material numbers, and serial numbers affected, and stores them in memory for later use.

Data Transformation and Organization

The extracted data from each PDF file is organized into separate sheets within the Excel file based on the drawing number associated with each report. This segregation facilitates easy access and analysis of relevant information.

Data Validation and Cleaning

To ensure the accuracy and consistency of the extracted data, the script checks for valid sheet names in Excel, and for any missing data in the NCR’s and handles them appropriately.

Working of the code

The script appends the extracted data from each PDF file into the corresponding sheets in the output Excel file. For each NCR, the script appends the item number, description of non-conformance, character number, defect type, measured value, and deviation value.

**Project 2: Color Codes Extraction from Images for the Fokker Plant**

Data Collection and Preprocessing

The Python script first takes the main parent folder containing PDF files as the input. It utilizes the fitz[2] library to convert each PDF file into images with a specific DPI (dots per inch).The converted images are stored in an intermediate output folder, preserving the same folder structure as the original PDF files.

The images are then chunked into smaller regions to enhance readability and to isolate specific information. A function is developed to crop each image into smaller chunks, ensuring each chunk contains meaningful data. The chunking process involves defining the chunk size, adding padding for better separation, and checking for significant data in each chunk using edge detection and contours analysis.

Extracting Colors and their associated codes

The Python script uses the easyocr [3] library to read text from the cropped image chunks. For each chunk, the script processes the text to identify color codes and their associated numbers. The script stores the extracted color codes and numbers in a dictionary, where the color codes act as column headers, and the numbers are the corresponding values.

Working of the code

The Python script generates an Excel file for each PDF file processed, containing extracted color codes and numbers. The Excel file is saved in the final output folder, named after the original PDF file.

Optimization

Threading for Parallel Processing is used to improve efficiency, the script utilizes threading to process multiple image chunks simultaneously. The "concurrent. Futures" module enables parallel execution, reducing the overall processing time.

**Project 3: Generalized Data Mining from PDF and Word Documents**

Data Collection and Preprocessing

The Python script takes pdf or word files as an input and extract various data elements such as tables, equations, relevant text, figure/table captions, and data from images.

Modules used

The tool is built using Python and relies on several libraries for different functionalities, such as PyPDF2[4] and fitz for PDF processing, docx2txt and docx for Word document processing, tabula and pandas for table extraction, and OpenPyXL[5] for Excel manipulation.

Working

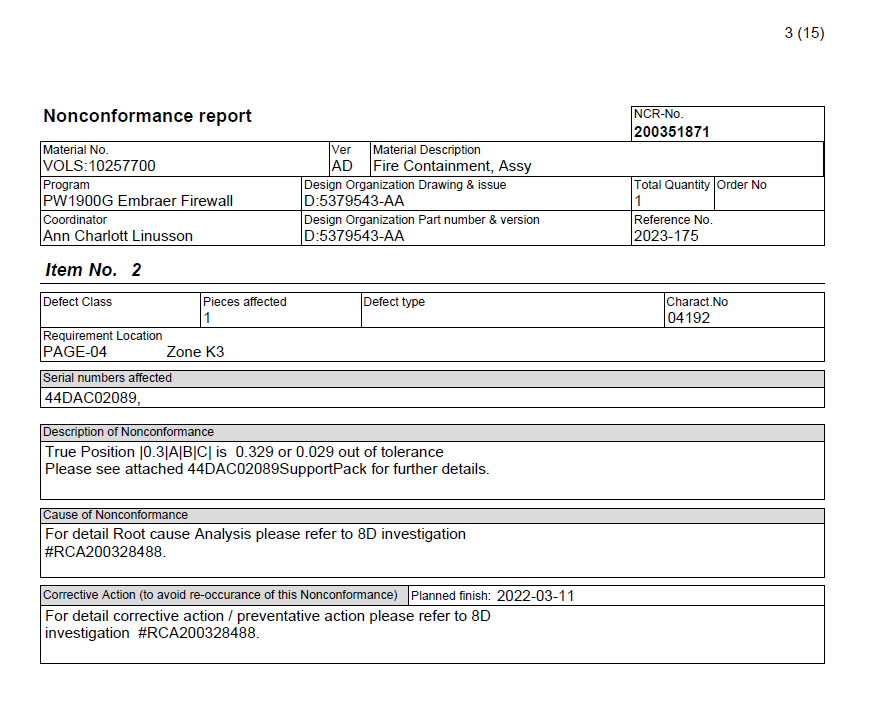
The GUI takes pdf’s and word files as an input. The tables from each file are extracted into excel files which can be used for further analysis. The user can enter multiple words or phrases and text relevant to that keyword is extracted into a document. Similarly there is an option to extract all the equations, figure captions, table captions, and images from the files provided by the user as an input.

**5. Results**

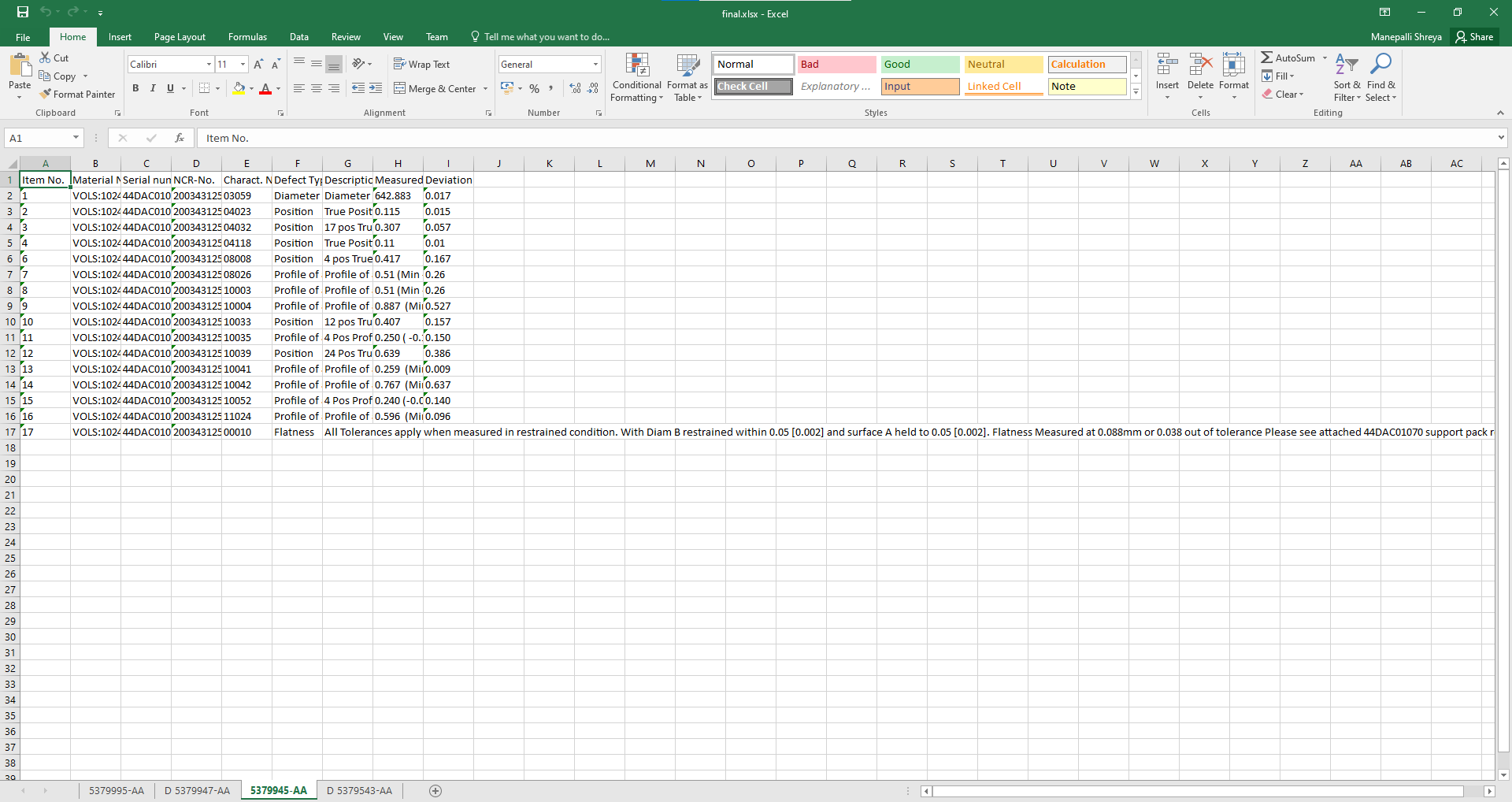
**Project 1: Data Extraction from Non-Conformance Reports (NCRs)**

The final output of the script is a consolidated Excel file containing organized data from all NCRs processed during the operation. The output file can be used for further data analytics and insights, enabling more efficient decision-making and process improvements.

**Input given by the user (Sample NCR Report)**

****

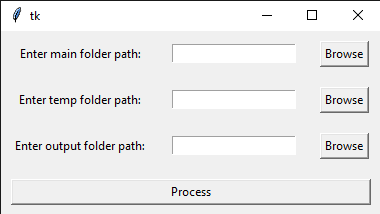
**Final Output Generated**

****

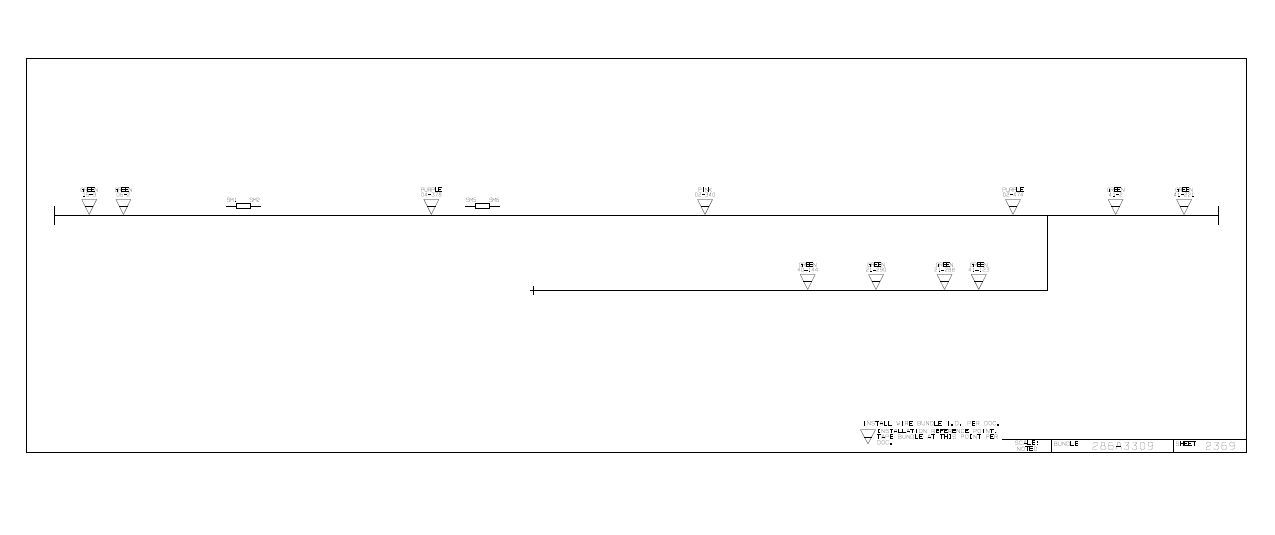
**Project 2: Color Codes Extraction from Images for the Fokker Plant**

The final output of the script is a set of Excel files, each corresponding to a PDF file processed during the operation. Each Excel file contains color codes as column headers and the associated numbers as values. The generated Excel files can be further analyzed and utilized for various purposes, such as data visualization and insights.

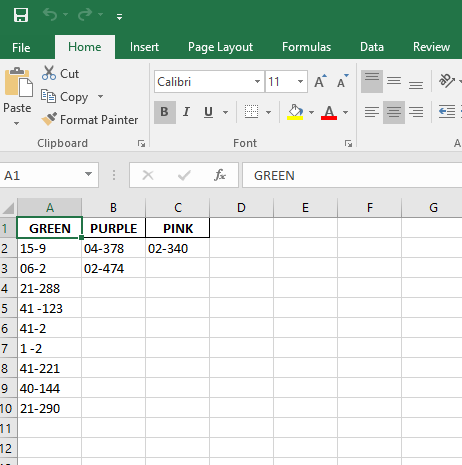
**GUI**

****

**Input given by user (Sample pdf)**

****

**Final Output Generated**

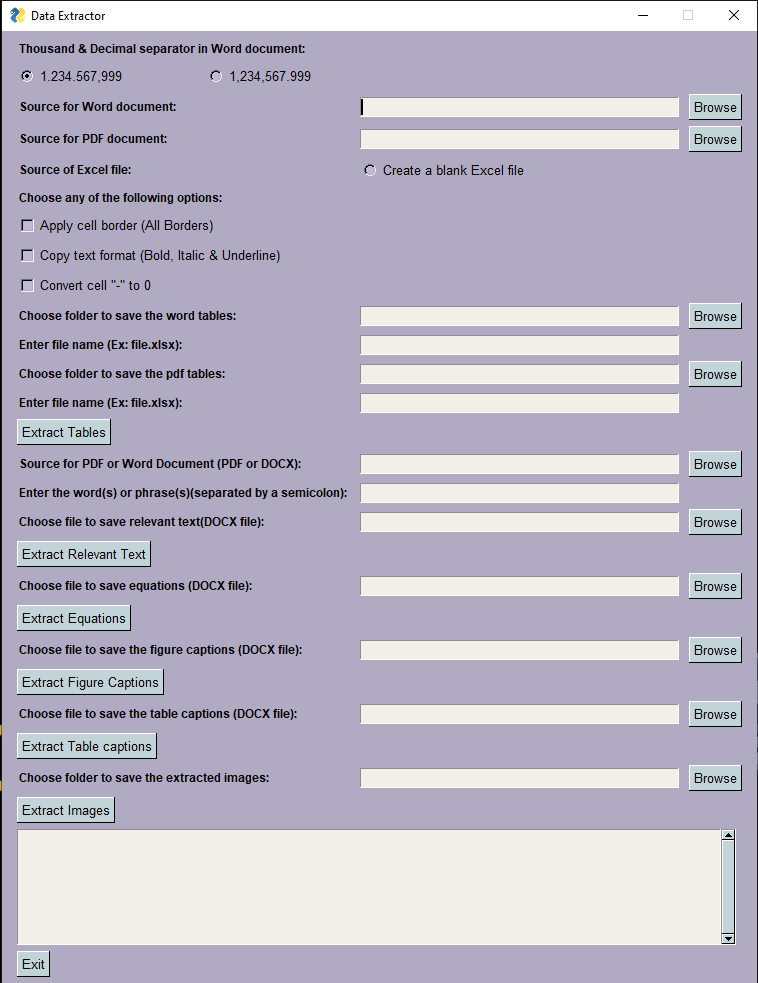
****

**Project 3: Generalized Data Mining from PDF and Word Documents**

The tool is designed to extract various types of data from PDF and Word documents, including tables, equations, relevant text based on keywords, figure captions, table captions, and images. The tool also includes a GUI (Graphical User Interface) using PySimpleGUI, making it easy for users to interact with and use the features.

For clarity, few images related to the GUI and the desired output are provided below.

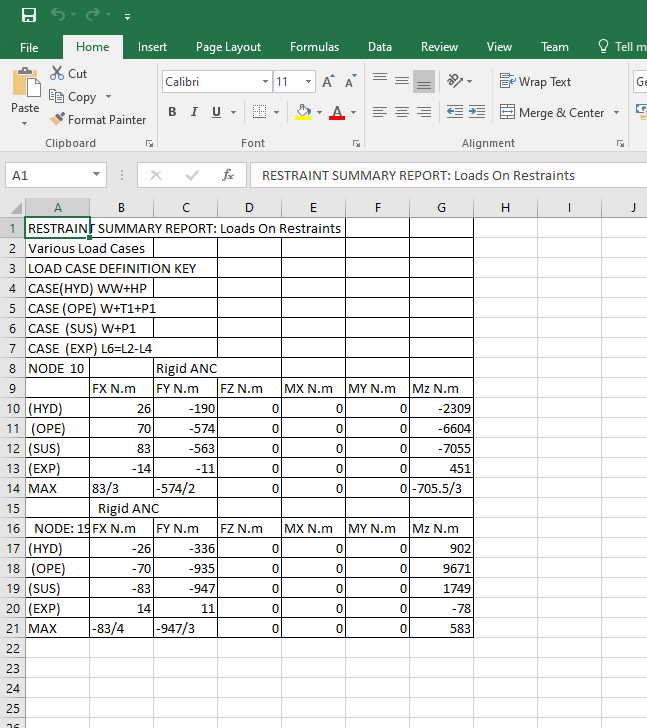
**GUI**

****

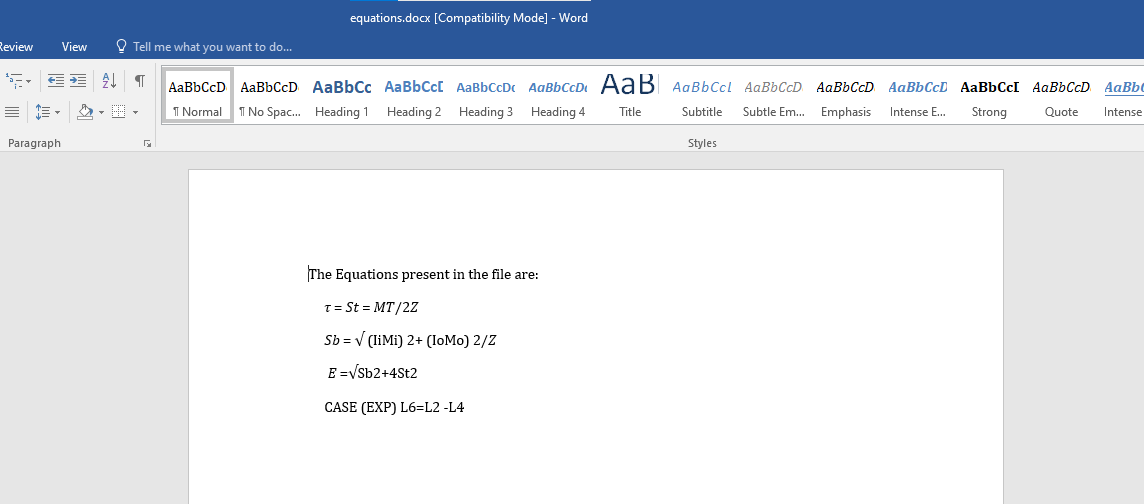
**Input given by the user: (Any pdf or Word Document)**

**Output Generated:**

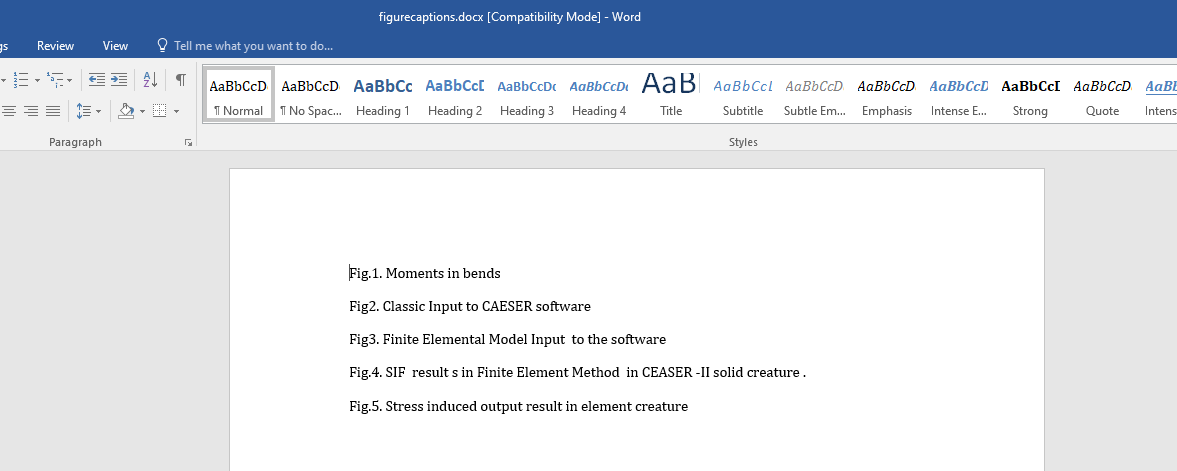
1. **Tables Extracted:**

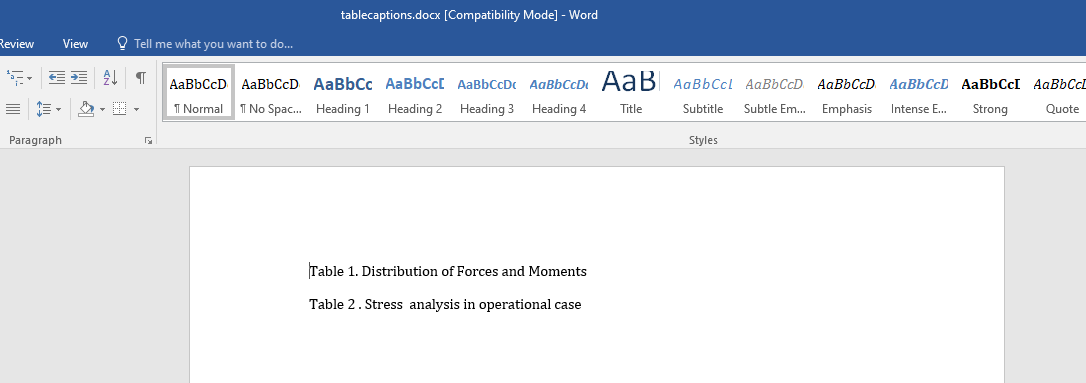


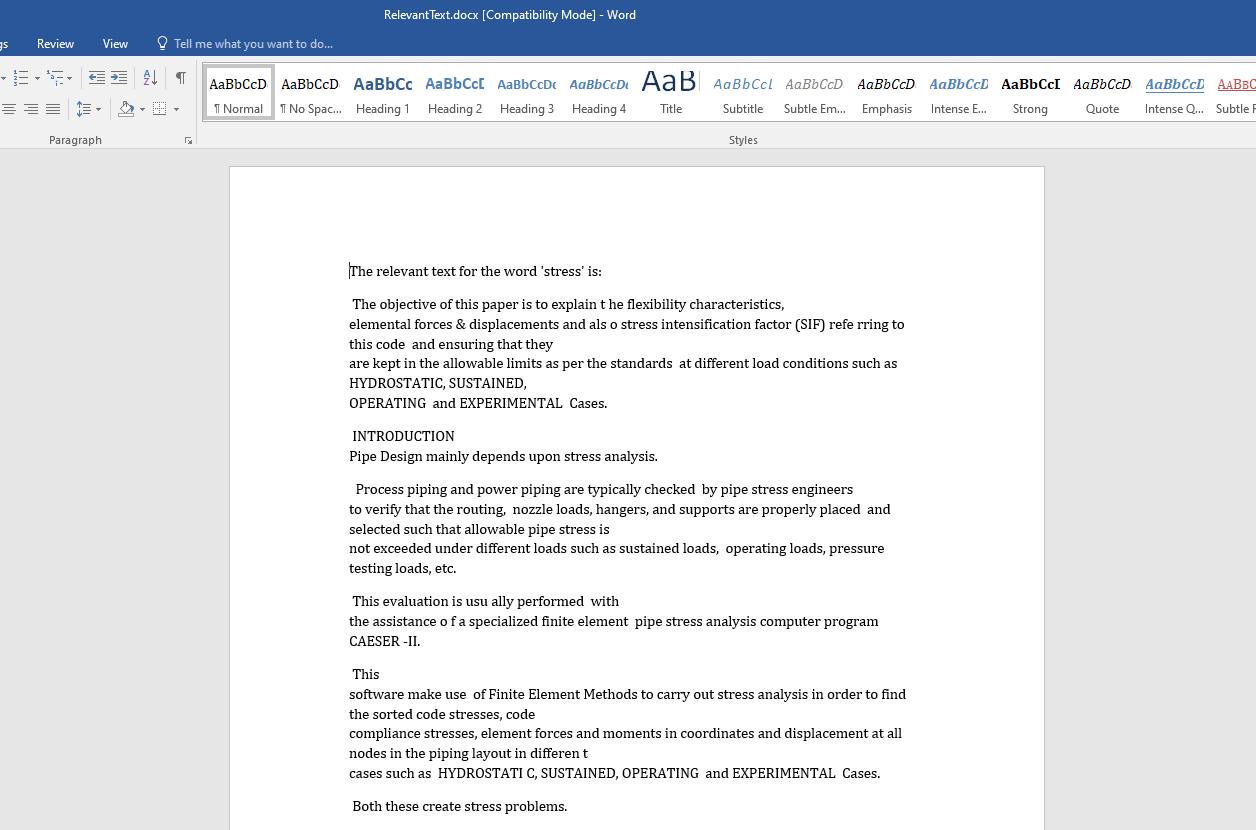
1. **Equations Extracted into a Word Document:**



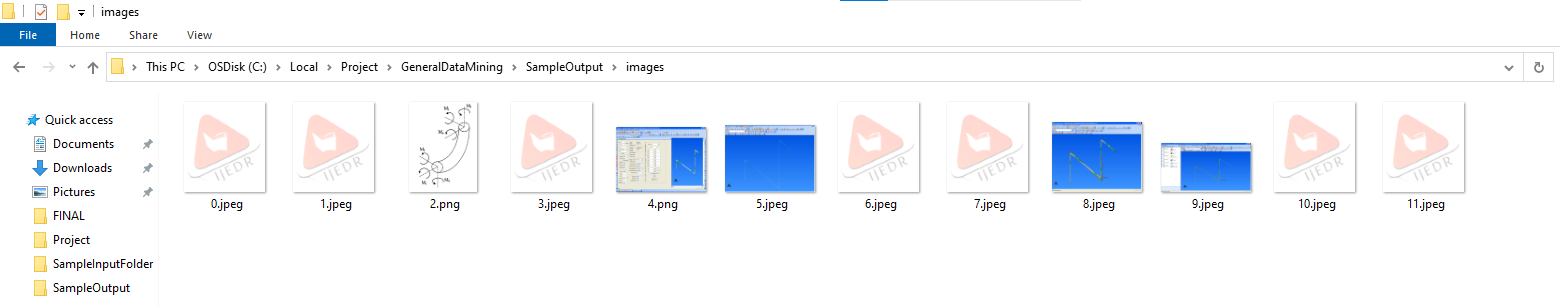
1. **Figure Captions Extracted into a Word Document:**

****

1. **Table Captions Extracted into a Word Document:  
   **
2. **Relevant text based on user input Extracted into a Document:**

****

1. **All Images present in the pdf/word Extracted into a Folder:**

****

**6. Testing and Performance**

All the Python script were tested on various sample files to ensure accurate data extraction and organization. The performance of the script was evaluated on a range of input file sizes to assess its efficiency and handling of larger volumes of unstructured data, image data as well as pdf and word documents.

**7. Conclusion**

The successful implementation of Data Extraction from Non-Conformance Reports (NCRs) project allowed for the efficient extraction of specific data from Non-Conformance Reports (NCRs) and the creation of a structured Excel file for further analysis. The automation of this data extraction process significantly reduces manual efforts and enhances data accessibility and usability, benefiting the SPE Team and the wider organization.

The successful implementation of the color code extraction project allowed for the efficient conversion of PDF files to images and the extraction of color codes and associated numbers from the images. The utilization of parallel processing using threading improved the overall performance and scalability of the solution. The generated Excel files provide valuable datasets on various products, enabling the GKN Aerospace team to gain useful insights for decision-making and process improvements.

The successful implementation of the Generalized Data Mining from PDF and Word Documents project enabled the extraction of various types of data from PDF and Word documents, including tables, equations, relevant text based on keywords, figure captions, table captions, and images which can be used for multiple purposes. The tool also includes a GUI (Graphical User Interface) using PySimpleGUI, making it easy for users to interact with and use the features.

**8. Future Scope**

**Non-Conformance Reports (NCRs) Data Extraction:**

Advanced Analytics: The curated dataset of extracted data from NCRs can be used for advanced analytics and data-driven decision-making. By employing machine learning algorithms and statistical analysis, the company can identify patterns, root causes of non-conformance, and potential areas for process improvement.

Real-time Monitoring: Integrating the automated data extraction process with real-time NCR generation systems can enable the continuous monitoring of non-conformance instances. This proactive approach allows the company to take prompt corrective actions and improve quality.

**Color Code Extraction from Images:**

Expanded Applications: The algorithms developed for color code extraction can be extended to analyze images from different sources, such as visual inspections, quality control checks, and product testing. This opens up opportunities to use color information in various contexts throughout the company's operations.

Integration with Computer Vision Systems: Integrating the color code extraction algorithms with computer vision systems can enhance automation in visual inspection processes, ensuring faster and more reliable detection of color defects.

**Generalized Data Mining from PDF and Word Documents:**

Integration with Data Management Systems: Integrating the data extraction algorithms with the company's data management systems can streamline data entry processes and ensure that valuable information from documents is readily available for analysis.

Support for Additional Document Formats: Expanding the algorithms to support additional document formats beyond PDF and Word, such as PowerPoint presentations and web pages, can broaden the scope of data extraction and analysis.

Automation and Process Optimization: By using these algorithms to automate manual data extraction processes, the company can optimize workflows, reduce manual errors, and free up valuable human resources for higher-value tasks

**Acknowledgement**

I would like to express my heartfelt gratitude to Mukundan Anand, Asha Koshy, Bhattacharya Somesh, Rajguru Chinmay, and the entire Research and Technology team. I would also like to thank Ahmed Haaris, Bapannavar Mabubi and the entire Shared Product Engineering Team. I am grateful to the IT and the HR Team for their constant support. Last but not the least I would like to thank the Site Director Vikram Reddy for his leadership. I am truly grateful for the invaluable contributions of each individual and the teams, without whom this internship would not have been possible.

**References**

[1] Tabula-py, URL: <https://tabula-py.readthedocs.io/en/latest/tabula.html>

[2] PyMuPDF (fitz), URL[: https://pymupdf.readthedocs.io/en/latest/module.html](:%20https:/pymupdf.readthedocs.io/en/latest/module.html)

[3] EasyOCR, 2021. URL: <https://www.jaided.ai/easyocr>

[4] PyPDF2, URL: <https://pypdf2.readthedocs.io/en/3.0.0/>

[5] OpenPyXL, URL[: https://openpyxl.readthedocs.io/en/stable/](:%20https:/openpyxl.readthedocs.io/en/stable/%20)