**INTERNSHIP AT GKN AEROSPACE**

**Efficient Data Mining and Information Extraction**

**ABSTRACT**

GKN Aerospace generates a substantial volume of unstructured data annually, comprising reports, documents, and images. This internship aimed to leverage data mining techniques to extract valuable information from this vast repository of pdf’s, word files and images enabling the creation of insightful 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 airplane wings and systematically storing them in a structured database. The third project revolved around generalized data mining from PDF and/or 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.

**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 for the Shared Product Engineering (SPE) Team

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 the SPE Team, the Fokker Plant, and the entire GAI Team. 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, data analysis, improve productivity, and foster data-driven insights across different departments and stakeholders at GKN Aerospace.

**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" 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" 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" 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 and fitz for PDF processing, docx2txt and docx for Word document processing, tabula and pandas for table extraction, and OpenPyXL for Excel manipulation.

Working

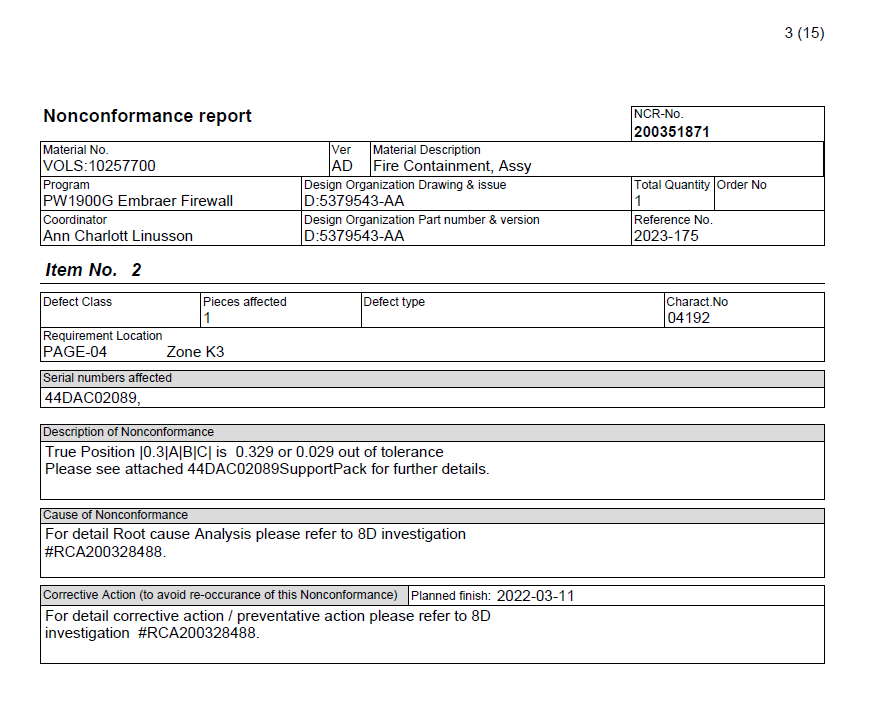
The GUI takes pdf’s and word files as an input. The tables from each file are extracted into excel files which can be used for further analysis. The user can enter multiple words or phrases and text relevant to that keyword is extracted into a document. Similarly there is an option to extract all the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

**RESULTS AND OUTPUT**

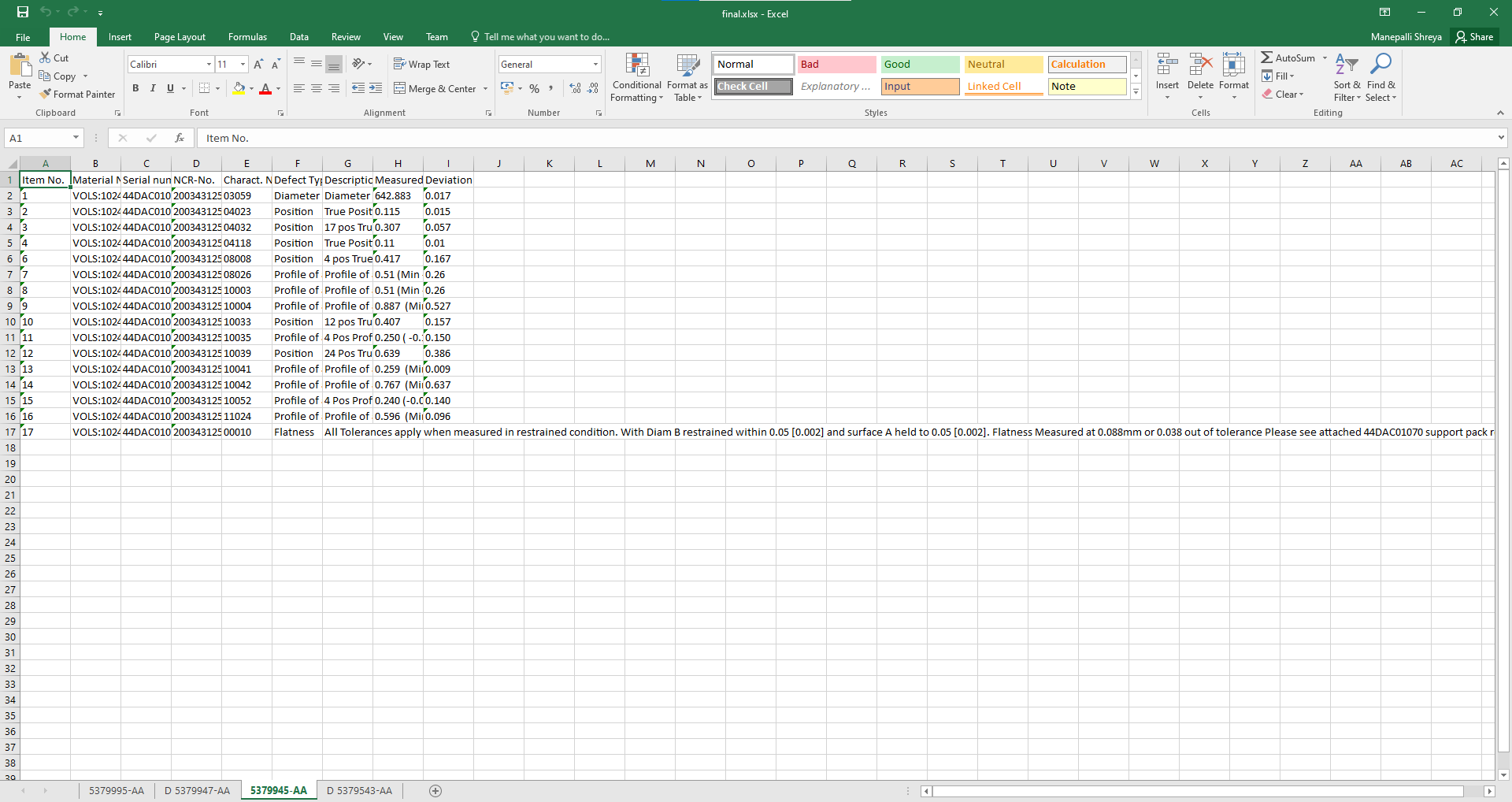
**Project 1: Data Extraction from Non-Conformance Reports (NCRs)**

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)**



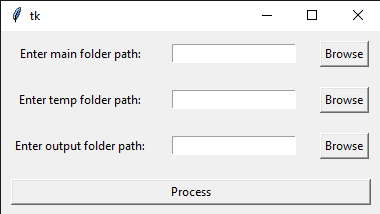
**Final Output Generated**



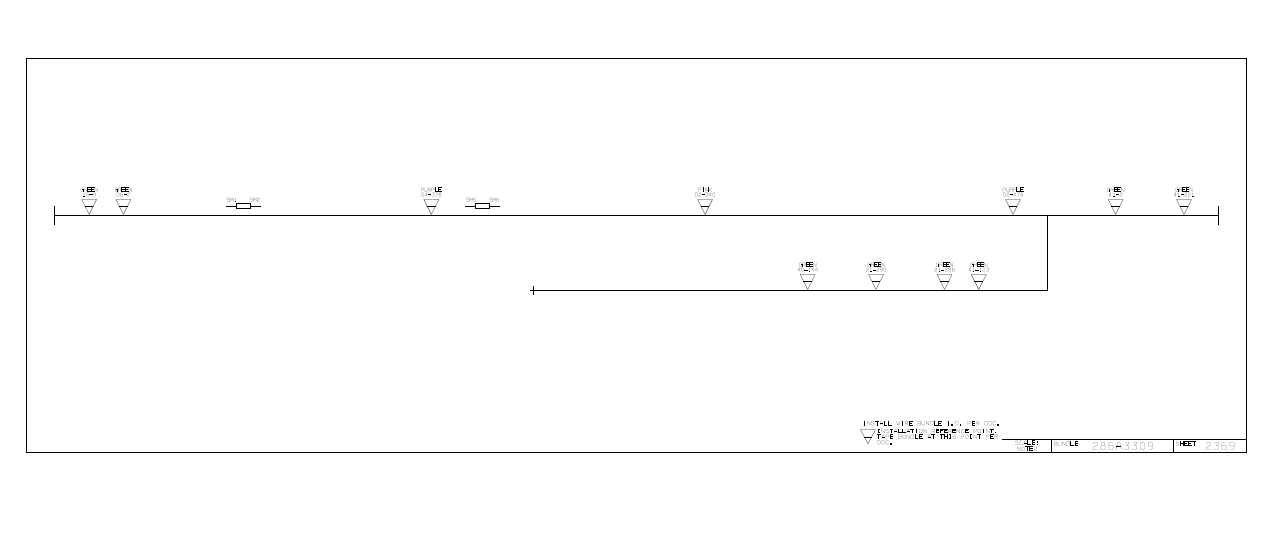
**Project 2: Color Codes Extraction from Images for the Fokker Plant**

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.

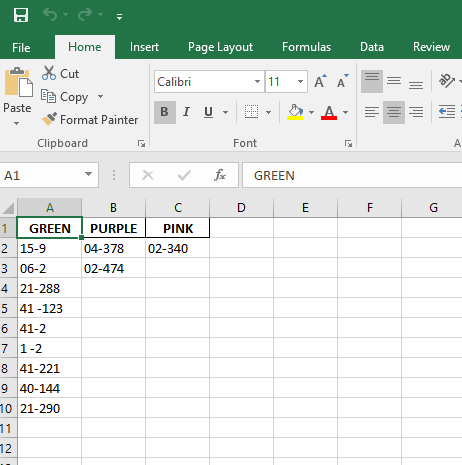
**GUI**



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



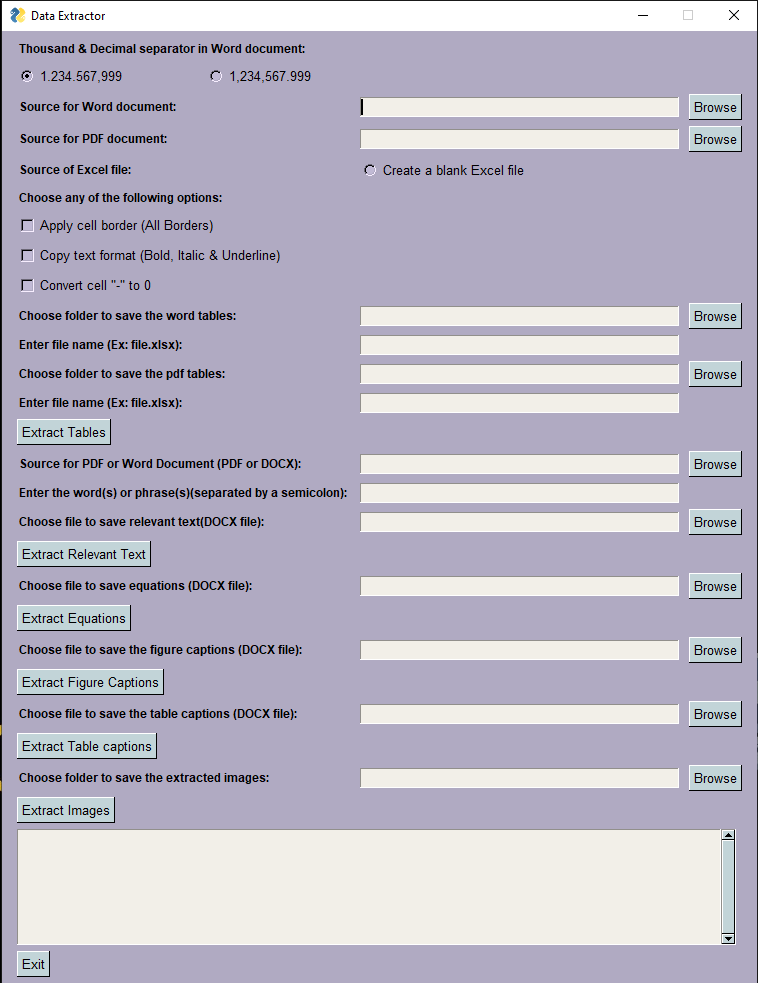
**Final Output Generated**



**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.

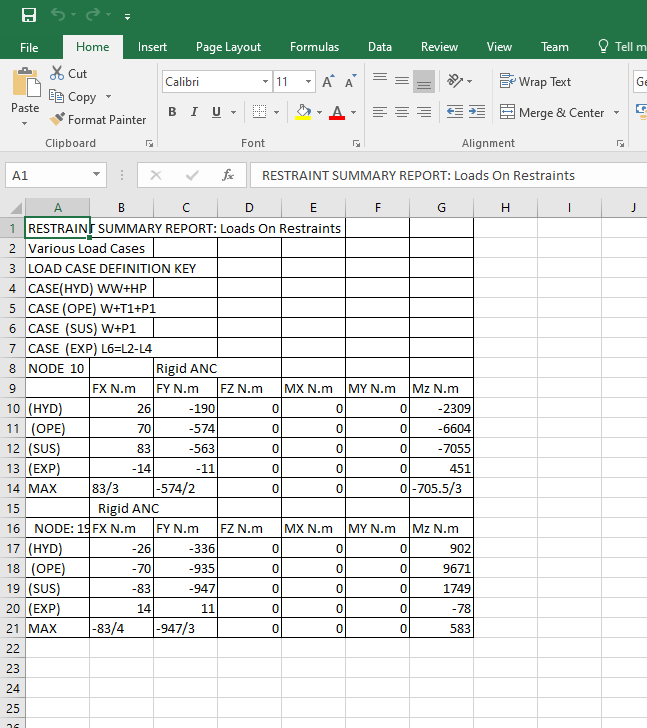
**GUI**



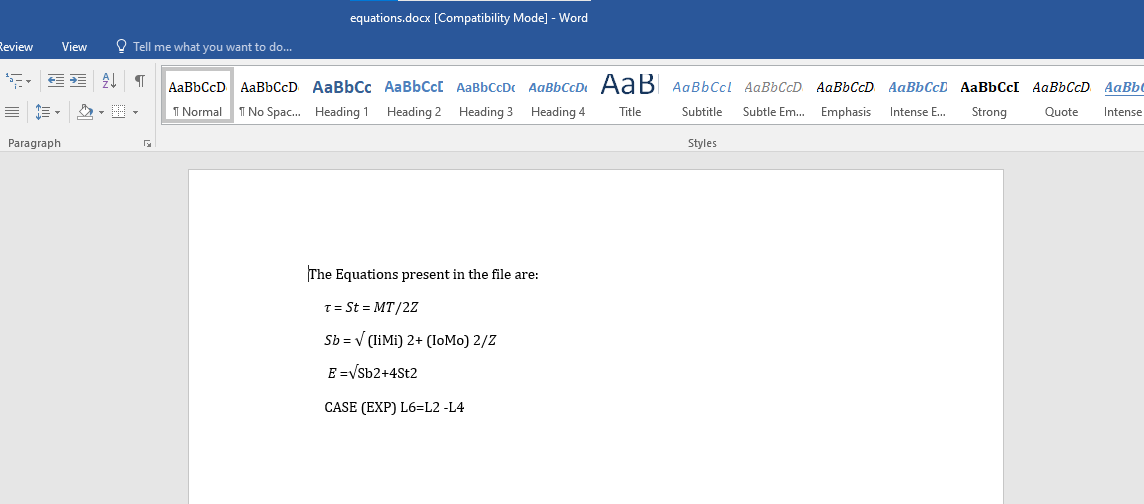
**Input given by the user (Any pdf or word file)**

**Output Generated:**

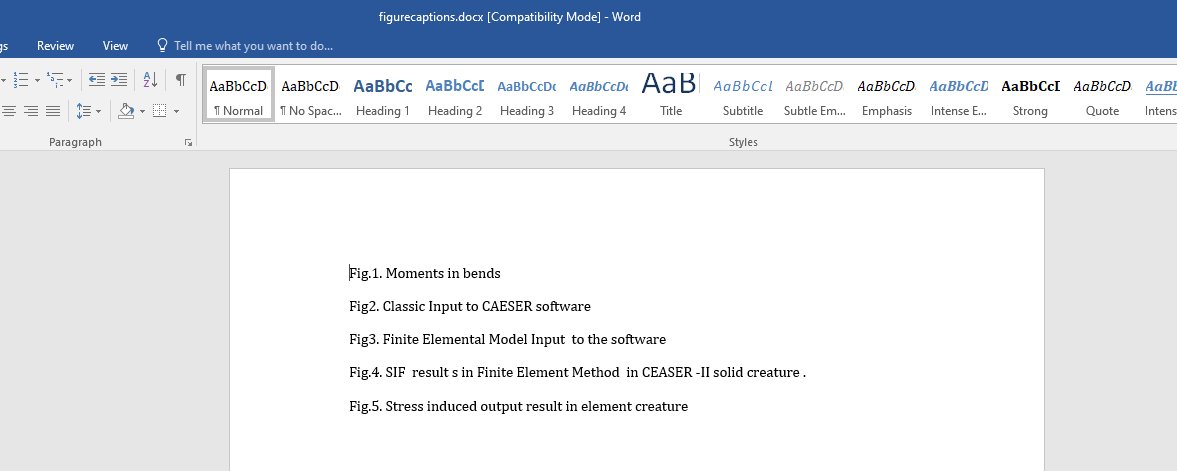
1. **Tables Extracted:**



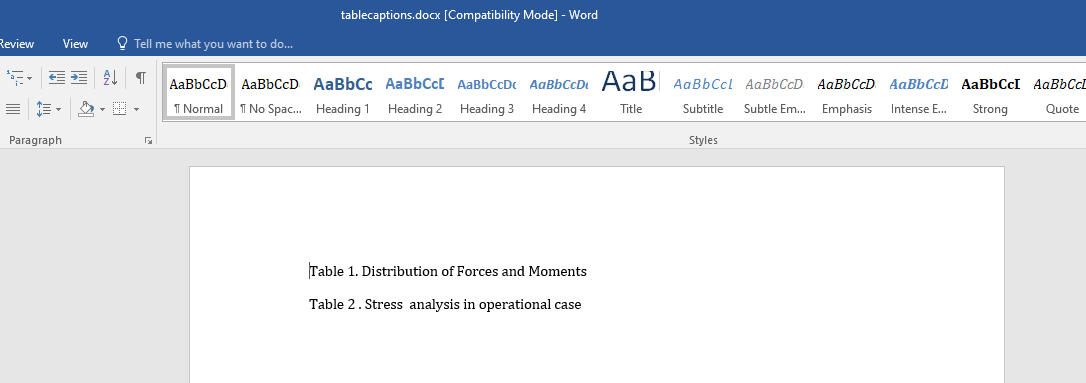
1. **Equations Extracted into a word file:**



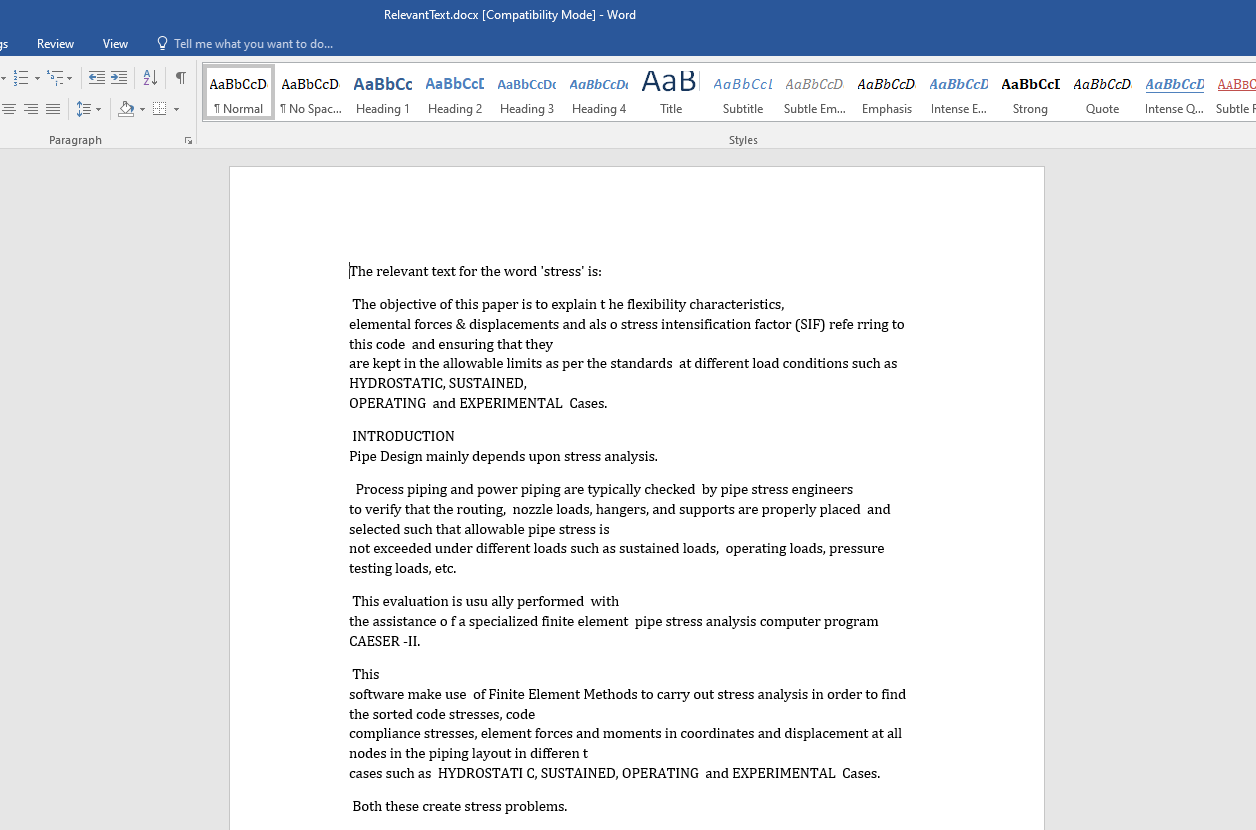
1. **Figure Captions Extracted into a word file:**



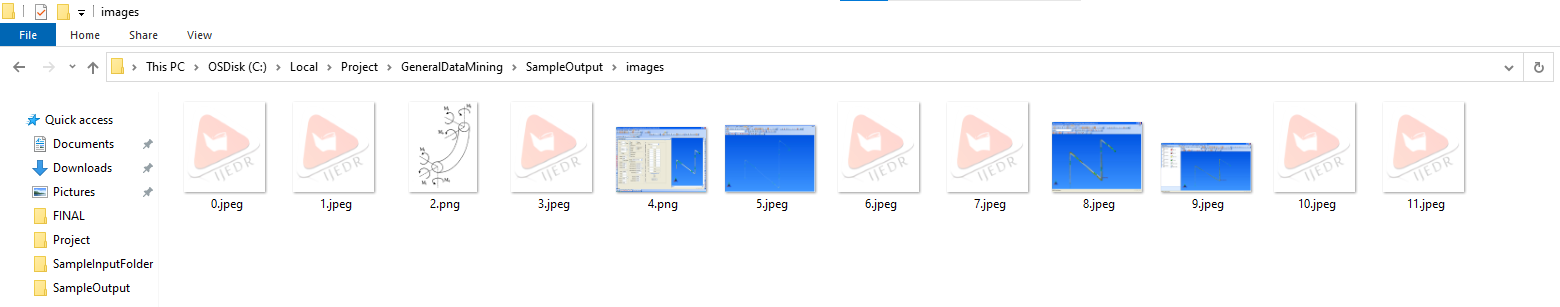
1. **Table Captions Extracted into a word file:**



1. **Relevant text based on user input Extracted into a word file**



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



**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.

**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 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 pruposes. The tool also includes a GUI (Graphical User Interface) using PySimpleGUI, making it easy for users to interact with and use the features.