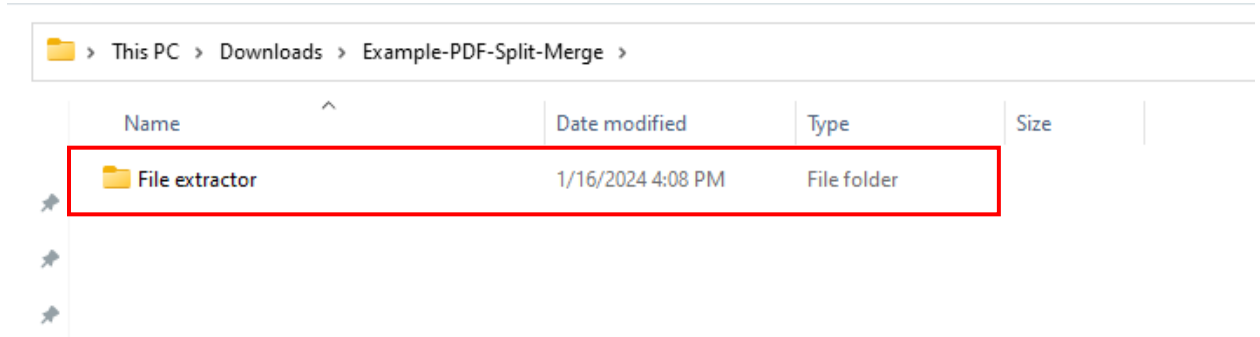


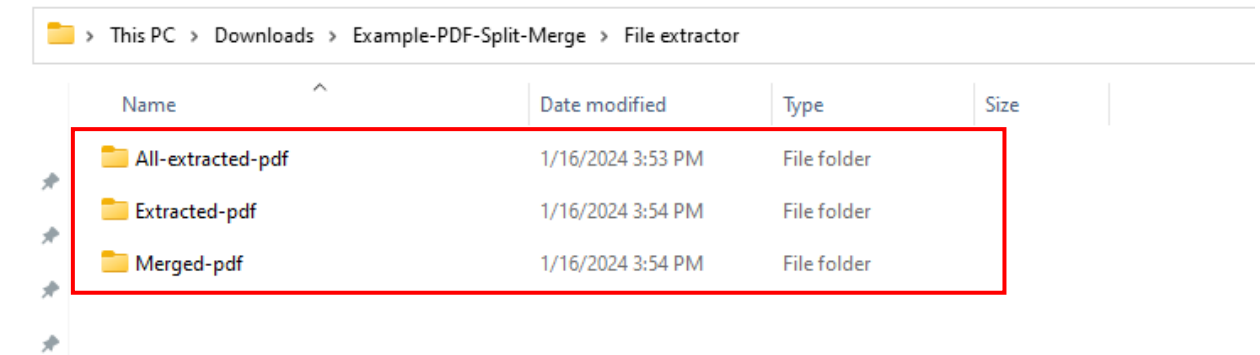
PDF Attachments Split and Merge

Script Execution Instructions:

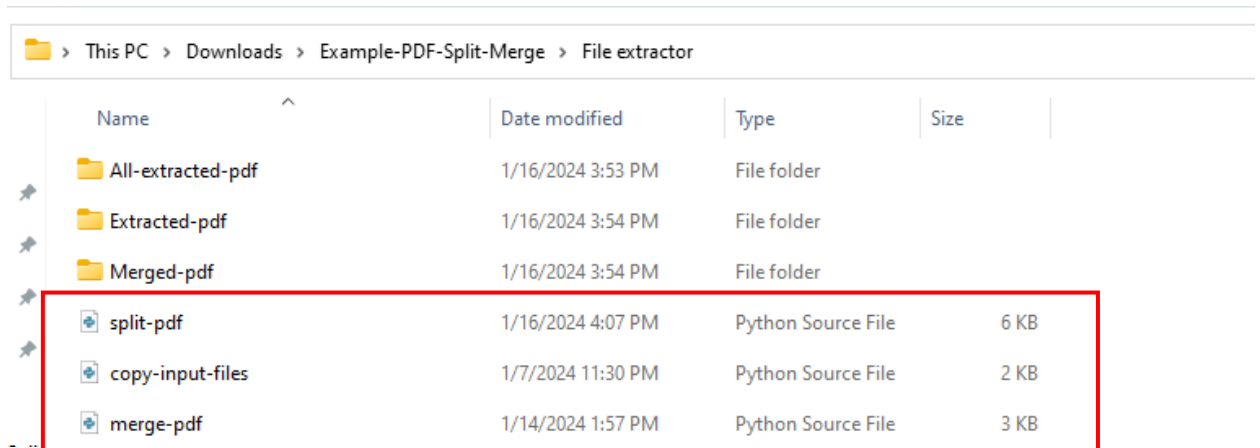
- 1) Create root folder: File extractor



- 2) Create sub folders inside root folder: All-extracted-pdf, Extracted-pdf, Merged-pdf



- 3) Create 3 python scripts with .py extension inside root folder: split-pdf.py, copy-input-files.py, merge-pdf.py



- 4) Place original_file.pdf (original PDF file with attachments) inside root folder.

This PC > Downloads > Example-PDF-Split-Merge > File extractor

| Name | Date modified | Type | Size |
|-------------------|-------------------|--------------------|--------------|
| All-extracted-pdf | 1/16/2024 3:53 PM | File folder | |
| Extracted-pdf | 1/16/2024 3:54 PM | File folder | |
| Merged-pdf | 1/16/2024 3:54 PM | File folder | |
| copy-input-files | 1/7/2024 11:30 PM | Python Source File | 2 KB |
| merge-pdf | 1/14/2024 1:57 PM | Python Source File | 3 KB |
| split-pdf | 1/16/2024 4:07 PM | Python Source File | 6 KB |
| original_file | 1/16/2024 4:13 PM | Adobe Acrobat D... | 1,776,134 KB |

- 5) Set all the path and file name variables in all 3 scripts with correct values.

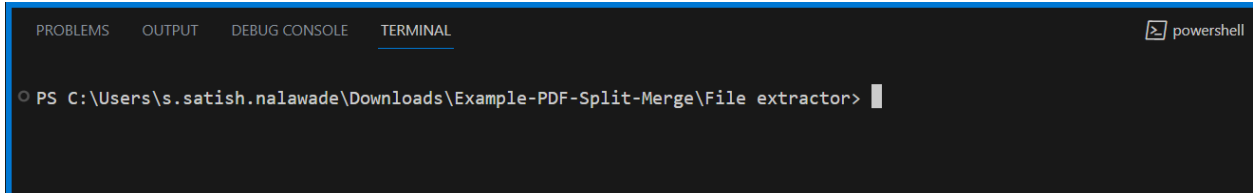
```
split-pdf.py 3
split-pdf.py > ...
114
115
116 #----- Original File (1st Split) -----
117 filename = 'original_file.pdf'
118 path = 'C:\\Users\\s.satish.nalawade\\Downloads\\Example-PDF-Split-Merge\\File extractor\\Extracted-pdf\\'
119 folder_paths_list = []
120 folder_paths_list = extract_pdfs(filename, path)
121 filename_list = []
122
```

```
copy-input-files.py
copy-input-files.py > ...
25 # Example usage
26 root_folder = 'C:\\Users\\s.satish.nalawade\\Downloads\\Example-PDF-Split-Merge\\File extractor\\Extracted-pdf\\'
27 output_folder = 'C:\\Users\\s.satish.nalawade\\Downloads\\Example-PDF-Split-Merge\\File extractor\\All-extracted-pdf\\'
28 copy_files_from_folders(root_folder, output_folder)
```

```
merge-pdf.py 2
merge-pdf.py > ...
28
29 input_path = 'C:\\Users\\s.satish.nalawade\\Downloads\\Example-PDF-Split-Merge\\File extractor\\All-extracted-pdf\\'
30 output_path = 'C:\\Users\\s.satish.nalawade\\Downloads\\Example-PDF-Split-Merge\\File extractor\\Merged-pdf\\'
31 original_pdf_file = 'original_file.pdf'
32 AllNestedFiles_filename = 'AllNestedFiles.xlsx'
33 df = pd.read_excel(AllNestedFiles_filename)
```

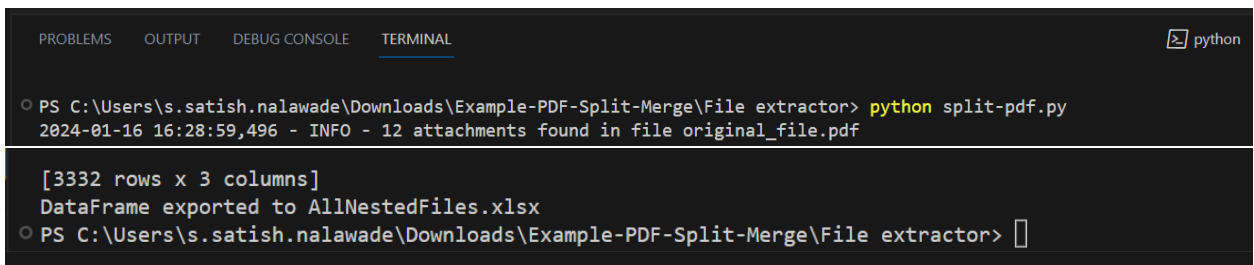
README

- 6) Open cmd or terminal and change current working directory to root folder (cd 'File extractor')



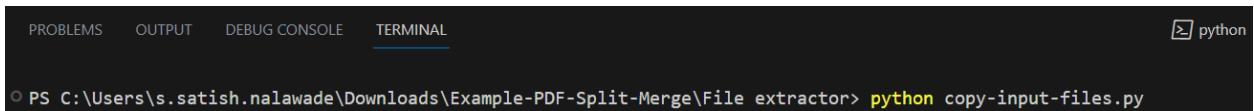
```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL powershell
PS C:\Users\s.satish.nalawade\Downloads\Example-PDF-Split-Merge\File extractor>
```

- 7) Execute python scripts in sequence by executing below commands in cmd or terminal:
python split-pdf.py -> python copy-input-files.py



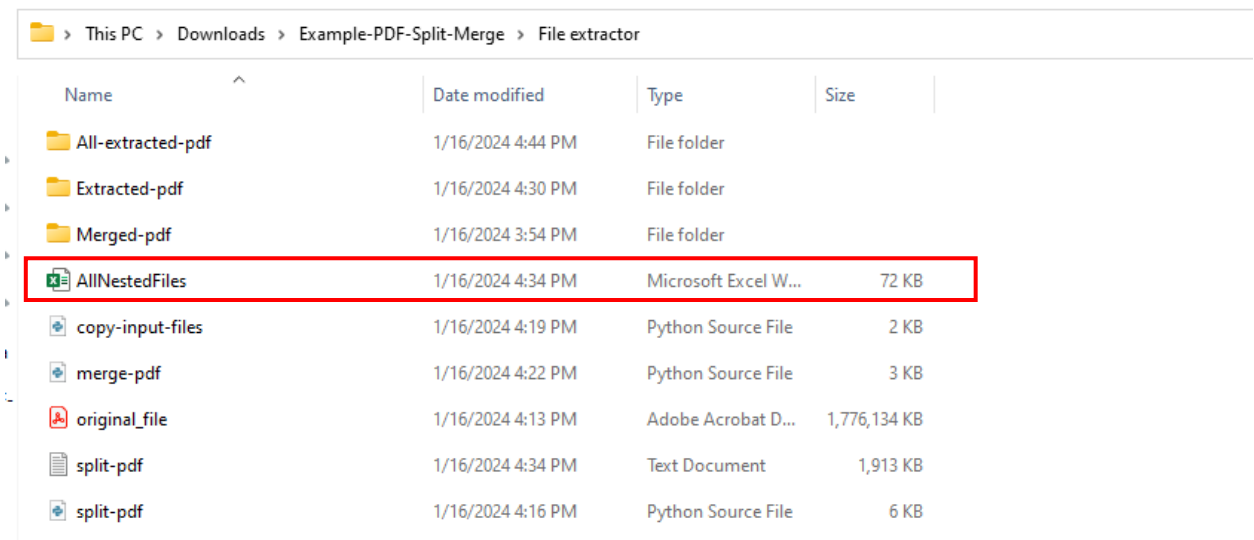
```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL python
PS C:\Users\s.satish.nalawade\Downloads\Example-PDF-Split-Merge\File extractor> python split-pdf.py
2024-01-16 16:28:59,496 - INFO - 12 attachments found in file original_file.pdf

[3332 rows x 3 columns]
DataFrame exported to AllNestedFiles.xlsx
PS C:\Users\s.satish.nalawade\Downloads\Example-PDF-Split-Merge\File extractor>
```



```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL python
PS C:\Users\s.satish.nalawade\Downloads\Example-PDF-Split-Merge\File extractor> python copy-input-files.py
```

- 8) Update 'AllNestedFiles.xlsx' file with grouping logic.



| Name | Date modified | Type | Size |
|-------------------|-------------------|----------------------|--------------|
| All-extracted-pdf | 1/16/2024 4:44 PM | File folder | |
| Extracted-pdf | 1/16/2024 4:30 PM | File folder | |
| Merged-pdf | 1/16/2024 3:54 PM | File folder | |
| AllNestedFiles | 1/16/2024 4:34 PM | Microsoft Excel W... | 72 KB |
| copy-input-files | 1/16/2024 4:19 PM | Python Source File | 2 KB |
| merge-pdf | 1/16/2024 4:22 PM | Python Source File | 3 KB |
| original_file | 1/16/2024 4:13 PM | Adobe Acrobat D... | 1,776,134 KB |
| split-pdf | 1/16/2024 4:34 PM | Text Document | 1,913 KB |
| split-pdf | 1/16/2024 4:16 PM | Python Source File | 6 KB |

Format of 'AllNestedFiles.xlsx' file:

| | A | B | C | D | E |
|---|----------------|-----------------|----------|-----------------------|-------------------|
| 1 | input_filename | output_filename | sequence | final_output_filename | final_page_number |
| 2 | input-1.pdf | output.pdf | 1 | final_output.pdf | 1 |
| 3 | input-2.pdf | output.pdf | 2 | final_output.pdf | 2 |

input_filename: Name of the input files which needs to be merged.

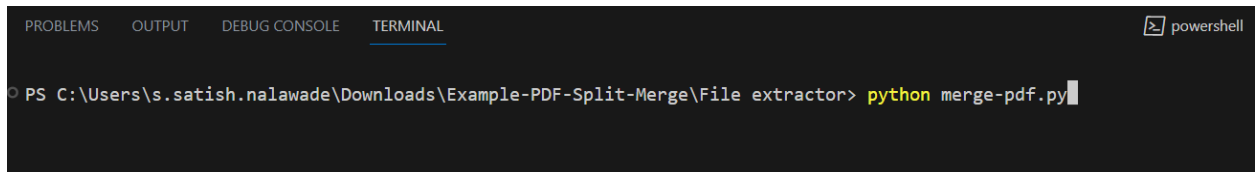
output_filename: Name of the source file from which input files are extracted.

sequence: sequence number of the input files in original file tree hierarchy (sequence of attachments).

final_output_filename: Name of the merged file (output file) to be generated from input files.

final_page_number: Page number of the input files in the final merged file.

- 9) Execute python script by executing below command in cmd or terminal: python merge-pdf.py



```

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL
PS C:\Users\s.satish.nalawade\Downloads\Example-PDF-Split-Merge\File extractor> python merge-pdf.py

```

- 10) Do these validation checks to ensure successful execution:

- Check if all the attachments extracted from original PDF are present in 'Extracted-pdf' folder with the exact hierarchy with sub folders.
- Check if 'AllNestedFiles.xlsx' file is present in 'File extractor' folder and contains all the expected information.
- Check if all the attachments from 'Extracted-pdf' folder and its sub folders are copied in 'All-extracted-pdf' folder.
- Check if all the merged pdf files are present in 'Merged-pdf' folder based on defined grouping logic in 'AllNestedFiles.xlsx'.

Description:

- **split-pdf.py:** extract attachments from original pdf with exact hierarchy in pdf tree represented by sub folders inside Extracted-pdf folder.
- **copy-input-files.py:** copy attachments from Extracted-pdf folder and its sub folders to specific destination folder All-extracted-pdf.
- **merge-pdf.py:** merge all the pdfs from All-extracted-pdf folder based on grouping logic defined in 'AllNestedFiles.xlsx' and place generated merged pdfs in Merged-pdf folder.