LANGUAGE DETECTION PROCESS FLOW

Pre-Requisite:

1. Config file must be updated with the Input, Output and Log path.
2. Python 3.8.x must be installed.
3. Mandatory Python Libraries to be installed:
   i. Langid
   ii. LangDetect
   iii. Shutil
   iv. Chardet
   v. Math

Key Points:

1. Bot can read Scanned as well as Normal PDFs.
2. Bot can handle the unrecognized characters of the txt file and re-generate the txt file.
3. Bot can identify 97 Languages.
4. Encoding of a file is not a limitation. Code can identify the encoding of the file.

Steps followed in AA for pre-processing of the document:

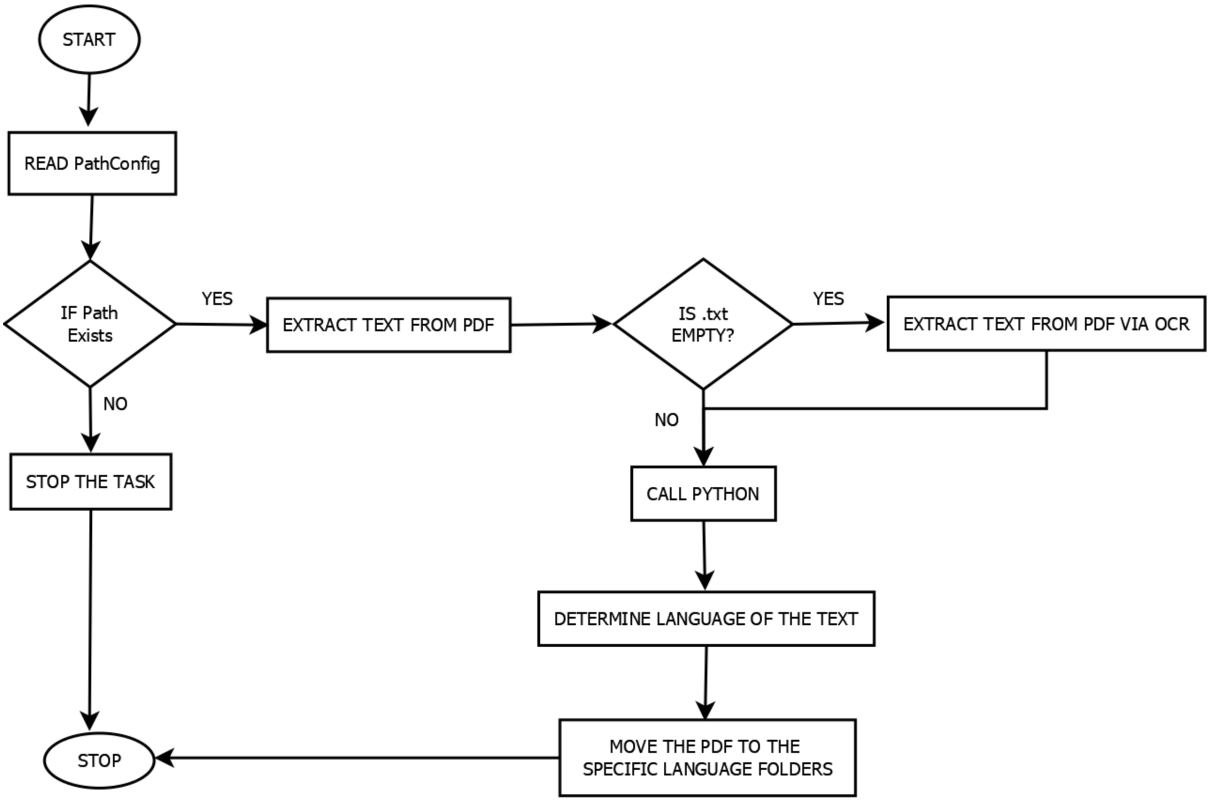
1. Read the PathFile.xml (contains all the required paths)
2. Read the PDFs from the input folder path.
3. Extract the text from the PDF and save the text file
4. In case of scanned PDFs, extract the image of the PDF.
5. Use OCR to extract the text from the image and overwrite the existing .txt files.
6. Call the Python script along with the argument.

Steps followed in Python for getting the language of the document:

1. Loop through each file in the folder of the text files generated.
2. Extract the text from the files and hold it in variable.
3. Check the text string in the 2 Language Detection libraries (Langid, LangDetect [detect(), detect_langs()]) of the Python.
4. Check the for the results and select the high probability of the language detected by the 3 functions.
5. Get the name of the language from the dictionary.
6. Create the folder of the same name and move the file into that folder.

Exceptions Handled:

1. Do not create the folder if it already exists, notified by the PYLogText.txt.
2. Do not move the file if it is already present in the folder, notified by the PYLogText.txt.
3. If paths are not present in config file, user can see the error in AALogFile.txt
4. If no input files are there in the Invoices folder, action will be captured in AALogFile.txt


START
READ PathConfig
IF Path Exists
YES
EXTRACT TEXT FROM PDF
IS .txt EMPTY?
YES
EXTRACT TEXT FROM PDF VIA OCR
NO
STOP THE TASK
NO
CALL PYTHON
DETERMINE LANGUAGE OF THE TEXT
MOVE THE PDF TO THE SPECIFIC LANGUAGE FOLDERS
STOP