Challenge #2
OCR De-Noising

The BYOB Automation Challenge 2018

Submission By:

Akshay Sharma

Ashish Rana

Technical Components

Platform

- Windows 10

Tools

- Poppler for Windows
- Python

Language Libraries

- elementtree
- python-docx
- pyspellchecker

Problem Overview

- Majority of the business documents today are in Portable Document Format (PDF); a file format for capturing and sending electronic documents in exactly the intended format.
- Unlike Microsoft Word and Excel, key challenge with PDF format is extracting and editing information.
- OCR (Optical Character Recognition) technique is used to identify words in a picture/scanned document and convert it into a machine readable text, that can be processed further with the help of computer.
- Although the technology is matured and uses advanced techniques it quite often produces erroneous output.

Breaking down the problems

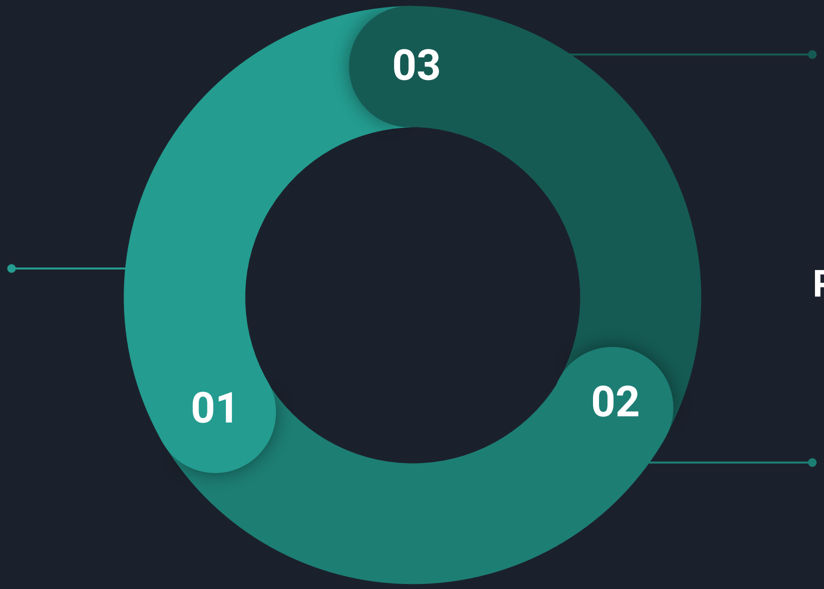
Problem Definition: Identification of specified OCR errors in the document and correction of the same

01 Identification of the OCR errors in a text document

02 De-Noising of OCR-Spelling correction/replacing wrong character with the right character

03 Generate respective DOCX file for every PDF resolved

Solution Design



OCR PDF Files

- Using Poppler run OCR on the provided PDF files
- Generate XML file corresponding to every Optically Read PDF file.
- The XML stores the textual, and syntactical data of the PDF.

Parse XML

- Parse the XML file to resolve OCR errors i.e, De-Noise OCR
- Commonly encountered OCR mistakes and spell checking implemented
- 2D List with multiple nested dictionaries data structure implemented to better store, understand and reuse XML content.

Create DOCX

- OCR results are not assured to provide data in right-left and top-down order.
- Above problem resolved
- Use python-docx library to create DOCX file as closely resembling the original PDF

Additionally:
Multiprocessing with batch processing support is implemented execute the solution for multiple documents parallely

Execution Steps

Step 1
Preprocessing

- Place all PDF documents in PDFs folder.
- Install all pre requisite tools needed
- Execute main.py

Step 2

- Respective DOCX files are generated in DOCX folder with subfolder of filename of PDF

Step 3

- View and analyse the Results
- Achieved good accuracy in OCR reading and De-Noising
- Only minor issues in docx conversion exist

Thank you!