PDF To Text

Introduction

System converts scanned documents in PDF format to text. The steps involve converting the PDF pages to images and performing OCR on it.

Development Environment

- Python 3.5.3
- OpenCV 3.4.0
- ImageMagick-6.9.9-43-Q8-x64
- Python wand
- Tesseract 4
- pytesseract 0.2.2
- Ghostscript 9.23

  Above are the list of key python packages used for development.

Reference links

http://docs.wand-py.org/en/latest/guide/install.html#install-imagemagick-on-window https://www.imagemagick.org/download/binaries/ [ImageMagick-6.9.9-43-Q8-x64-dll.exe] https://github.com/tesseract-ocr/tesseract/wiki/4.0-with-LSTM#400-alpha-for-windows http://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-setup-4.00.00dev.exe https://ghostscript.com/download/gsdnld.html

Usage and Deployment

Navigate to the root directory `PdfToText` and use following command to invoke the program. `PTTConfig.properties` property file controls program execution

```
python src\Pipeline.py config\PTTConfig.properties
```

Configuration parameters

Mandatory Parameters

- `logFile` = log/ENLog.log, Path of the log file
- `inputPath` = input/ Path where input pdf are stored
- `tempPath` = temp/ Path to store the temporary files
- `outputPath` = output/ Path where output will be stored
- `magickHome` = C:/Program Files/ImageMagick-6.9.9-Q8/ Absolute path of Image Magic installation directory
- `tesseractHome` = C:/Program Files (x86)/Tesseract-OCR/ Absolute path of Tesseract installation directory
- `tessData` = C:/Program Files (x86)/Tesseract-OCR/tessdata/ Absolute path of directory where Tesseract language models are stored
  Please give the trailing slash / while specifying the path
- `resolution` = 400 , Integer value representing dpi with which images need to be generated, Higher the value better the image quality, But increasing it beyond a point doesn’t help.
- `outputFormat` = tsv , Possible values are pdf, hocr, txt, tsv

Optional parameters

- `logLevel` = INFO, Level of logging required. Possible values are CRITICAL, ERROR, WARNING, INFO, DEBUG, NOTSET. https://docs.python.org/3.6/library/logging.html#levels

Directory Structure

- PdfToText [Root Directory]
    - config [All .property files stays here]
    - input [Input PDF files]
    - log [Default log location]
    - output [Output OCR file stays here]
    - src [Python source folder]
    - temp [Temporary storage location]

This document was edited at https://dillinger.io/