Colonial Architecture Pipeline

This Python script enables you to input a PDF file from the Colonial Architecture repository and extract meaningful entities from the text.

Operating system

The script was tested on Ubuntu Linux via a Virtual Machine (VirtualBox). It could work using different operating systems, but Frog requires you to install a number of dependencies that are more difficult to install on Windows.

Prerequisites

ImageMagick: http://www.imagemagick.org/script/download.php.
Tesseract: https://github.com/tesseract-ocr/tesseract.
Pyocr: https://github.com/openpaperwork/pyocr.
Frog and its dependencies: https://languagemachines.github.io/frog/.
Python-frog: https://github.com/proycon/python-frog

Installing

Once all the prerequisites are installed, all that is needed is a file to run, located in a folder that can be executed by the script. Change the variables in the script and you are ready to run!