PDF to Template

Description:

This script is responsible for extracting data from a PDF and then using Flask to display it on a web page.

Project Scope:

The goal is to extract data from an image embedded in a PDF without losing the table format. To achieve this, we will need to recreate the table and structure it in a JSON format to later display it on a web page.

Process:

First step: The PDF is processed with the PDF2Image library to extract images from the PDF in BMP format. It checks if the orientation of the images is correct and corrects it if necessary, using the pytesseract library for this purpose.

Second step: The processed image is sent to the Google Cloud Vision API to obtain the text (OCR), where both the text and the coordinates of the words are retrieved.

Third step: A reference point in the text is identified to locate the words, and then the rows and columns of the table are positioned.

Fourth step: Flask is used to display the data on a web page.

Dependencies:

Technologies: - Gnu/Linux - Python - Google Cloud Vision API (OCR)

Environment: - Debian 12 (Bookworm) - Python 3.11.2 or higher

Debian packages: - tesseract-ocr: 5.3.0 - poppler-utils: 22.12.0-2+b1 [Debian 12/Bookworm]

Python packages: asttokens==2.4.1 blinker==1.7.0 cachelib==0.12.0 cachetools==5.3.2 certifi==2024.2.2 cffi==1.16.0 charset-normalizer==3.3.2 click==8.1.7 cryptography==42.0.2 decorator==5.1.1 executing==2.0.1 Flask==3.0.2 Flask-Session==0.6.0 gevent==24.2.1 google-api-core==2.17.0 google-auth==2.27.0 google-cloud-vision==3.7.0 googleapis-common-protos==1.62.0 greenlet==3.0.3 grpcio==1.60.1 grpcio-status==1.60.1 gunicorn==21.2.0 idna==3.6 ipython==8.21.0 itsdangerous==2.1.2 jedi==0.19.1 Jinja2==3.1.3 MarkupSafe==2.1.5 matplotlib-inline==0.1.6 numpy==1.26.4 opencv-python==4.9.0.80 packaging==23.2 pandas==2.2.1 parso==0.8.3 pdf2image==1.17.0 pdfminer.six==20221105 pexpect==4.9.0 pillow==10.2.0 prompt-toolkit==3.0.43 proto-plus==1.23.0 protobuf==4.25.2 ptyprocess==0.7.0 pure-eval==0.2.2 pyasn1==0.5.1 pyasn1-modules==0.3.0 pycparser==2.21 Pygments==2.17.2 PyPDF2==3.0.1 pypdfium2==4.27.0 pytesseract==0.3.10 python-dateutil==2.9.0.post0 pytz==2024.1 requests==2.31.0 rsa==4.9 six==1.16.0 stack-data==0.6.3 traitlets==5.14.1 tzdata==2024.1 urllib3==2.2.0 wcwidth==0.2.13 Werkzeug==3.0.1 zope.event==5.0 zope.interface==6.2

requirements.txt

How to Install:

Debian packages:

```
sudo apt-get install tesseract-ocr poppler-utils
```

Optional:

```
sudo apt-get install python3-virtualenv
```

Python packages:

```
pip install -r requirements.txt
```

How to Use:

```
gunicorn -w 4 -b 0.0.0.0:8000 app:app --worker-class gevent --timeout 120
```