Turn your documents into data!

Parsr, is a minimal-footprint document (image, pdf) cleaning, parsing and extraction toolchain which generates readily available, organized and usable data for data scientists and developers.

It provides users with clean structured and label-enriched information set for ready-to-use applications ranging from data entry and document analysis automation, archival, and many others.

Currently, Parsr can perform:

1. Document Hierarchy Regeneration - Words, Lines and Paragraphs
2. Headings Detection
3. Key-Value Pair Detection (for the extraction of specific form-based entries)
4. Page Number Detection
5. Header-Footer Detection
6. Link Detection
7. Heading Detection
8. Whitespace Removal

Parsr can generate the following output formats:

1. JSON
2. Markdown
3. Text
4. CSV (for tables), or Pandas Dataframes (see here)
5. PDF

Table of Contents

- Turn your documents into data!
- Table of Contents
- Getting Started
  - Installation
  - Usage
- Documentation
- Contribute
- Third Party Licenses
- License

Getting Started

Installation

– *The advanced installation guide is available here* –

The quickest way to install and run the Parsr API is through the docker image:

```
docker pull axarev/parsr
```

If you also wish to install the GUI for sending documents and visualising results, execute the following:

```
docker pull axarev/parsr-ui-localhost
```

Note: Parsr can also be installed bare-metal (not via Docker containers), the procedure for which is documented in the installation guide.

Usage

– The advanced usage guide is available here –

To run the API, issue:

```
docker run -p 3001:3001 axarev/parsr
```

1. To use the **Jupyter Notebook** and the **python** interface to the Parsr API, follow here.
2. To use the GUI tool (the API needs to already be running), issue:

   ```
   docker run -t -p 8080:80 axarev/parsr-ui-localhost:latest
   ```

   Then, access it through http://localhost:8080.

The API based usage and the command line usage are documented in the advanced usage guide.

Documentation

All documentation files can be found here.

Contribute

Please refer to the contribution guidelines.

Third Party Licenses

Third Party Libraries licenses for its dependencies:

1. **QPDF**: Apache http://qpdf.sourceforge.net
2. **GraphicsMagick**: MIT http://www.graphicsmagick.org/index.html
3. **ImageMagick**: Apache 2.0 https://imagemagick.org/script/license.php
4. **Pdfminer.six**: MIT https://github.com/pdfminer/pdfminer.six/blob/master/LICENSE

5. **Tesseract**: Apache 2.0 https://github.com/tesseract-ocr/tesseract
6. **Camelot**: MIT https://github.com/camelot-dev/camelot
7. **MuPDF** (Optional dependency): AGPL https://mupdf.com/license.html
8. **Pandoc** (Optional dependency): GPL https://github.com/jgm/pandoc

License