pdf22md

A blazingly fast PDF to Markdown converter for macOS.

`pdf22md` is a command-line tool that extracts all text and image content from a PDF file and converts it into a clean Markdown document. It uses Grand Central Dispatch (GCD) to process pages and save images in parallel, making it exceptionally fast for multi-page documents.

Key Features

- **High-Speed Conversion**: Uses all available CPU cores to process PDF pages concurrently
- **Intelligent Heading Detection**: Analyzes font sizes and usage frequency to automatically format titles and headings (`#` , `##` , etc.)
- **Asset Extraction**: Saves raster and vector images into a specified assets folder and links them correctly in the Markdown file
- **Smart Image Formatting**: Automatically chooses between JPEG (for photos) and PNG (for graphics with transparency) to optimize file size and quality
- **Flexible I/O**: Reads from a PDF file or `stdin` and writes to a Markdown file or `stdout`
- **Customizable Rasterization**: Allows setting a custom DPI for converting vector graphics to bitmaps

Installation

Using Homebrew (Coming Soon)

```
brew tap twardoch/pdf22md
brew install pdf22md
```

Building from Source

To build the project manually, you need Xcode Command Line Tools installed.

```
# Clone the repository
git clone https://github.com/twardoch/pdf22md.git
cd pdf22md

# Compile the tool
make

# Install it to /usr/local/bin (optional)
sudo make install
```

Download Pre-built Binary

Pre-built binaries are available from the Releases page.

Usage

```
Usage: pdf22md [-i input.pdf] [-o output.md] [-a assets_folder] [-d dpi]
  Converts PDF documents to Markdown format
  -i <path>: Input PDF file (default: stdin)
  -o <path>: Output Markdown file (default: stdout)
  -a <path>: Assets folder for extracted images
  -d <dpi>: DPI for rasterizing vector graphics (default: 144)
```

Examples

```
# Convert a PDF file to Markdown
pdf22md -i document.pdf -o document.md

# Convert with images saved to an 'assets' folder
pdf22md -i report.pdf -o report.md -a ./assets

# Convert with custom DPI for vector graphics
pdf22md -i presentation.pdf -o presentation.md -a ./images -d 300

# Use with pipes
cat document.pdf | pdf22md > document.md

# Convert and view in less
pdf22md -i manual.pdf | less
```

Requirements

- macOS 10.15 (Catalina) or later
- Xcode Command Line Tools (for building from source)

Project Structure

```
pdf22md/
├── src/                    # Source code
│   ├── main.m              # Entry point
│   ├── PDFMarkdownConverter.*  # Main conversion logic
│   ├── PDFPageProcessor.*      # PDF page processing
│   ├── ContentElement.*        # Content element definitions
│   └── AssetExtractor.*        # Image extraction logic
├── docs/                   # Additional documentation
├── test/                   # Test files
```

```
├── LICENSE                 # MIT License
├── Makefile                # Build configuration
└── README.md               # This file
```

Contributing

Contributions are welcome! Please feel free to submit a Pull Request. For major changes, please open an issue first to discuss what you would like to change.

1. Fork the repository
2. Create your feature branch (`git checkout -b feature/AmazingFeature`)
3. Commit your changes (`git commit -m 'Add some AmazingFeature'`)
4. Push to the branch (`git push origin feature/AmazingFeature`)
5. Open a Pull Request

License

This project is licensed under the MIT License - see the LICENSE file for details.

Acknowledgments

- Built with Apple's PDFKit and Core Graphics frameworks
- Parallel processing powered by Grand Central Dispatch (GCD)
- Inspired by the need for fast, accurate PDF to Markdown conversion

Related Projects

- pdfplumber - Python library for PDF processing
- pdf2md - Another PDF to Markdown converter
- pandoc - Universal document converter

Changelog

See CHANGELOG.md for a list of changes in each version.

Support

If you encounter any issues or have questions, please open an issue on GitHub.