

main

2 Branches

2 Tags

Go to file

Go to file

<> Code

...

 gagb

Merge pull request #91 from PetrAPConsulting/patch-1

18e3f1d · 19 hours ago

171 Commits

	.devcontainer	docs: update README with Devcontai...	3 days ago
	.github/workflows	Initial commit.	last month
	src/markitdown	Merge branch 'main' into update_com...	2 days ago
	tests	Merge branch 'main' into main	3 days ago
	.dockerignore	Added Docker Image for using markitd...	4 days ago
	.gitattributes	Create .gitattributes	5 days ago
	.gitignore	Merge branch 'main' into update_com...	2 days ago
	.pre-commit-config.yaml	Initial commit.	last month
	CODE_OF_CONDUCT.md	CODE_OF_CONDUCT.md committed	last month
	Dockerfile	feat: add devcontainer configuration a...	3 days ago
	LICENSE	LICENSE committed	last month
	README.md	Merge branch 'main' into patch-1	20 hours ago
	SECURITY.md	SECURITY.md committed	last month
	SUPPORT.md	SUPPORT.md committed	last month
	pyproject.toml	feat: add project description in pyproje...	2 days ago

README

Code of conduct

MIT license

Security

Important

(12/19/24) Hello! MarkItDown team members will be resting and recharging with family and friends over the holiday period. Activity/responses on the project may be delayed during the period of Dec 21-Jan 06. We will be excited to engage with you in the new year!

MarkItDown

pypi

v0.0.1a3

downloads

8k/day

MarkItDown is a utility for converting various files to Markdown (e.g., for indexing, text analysis, etc). It supports:

- PDF
- PowerPoint
- Word
- Excel
- Images (EXIF metadata and OCR)
- Audio (EXIF metadata and speech transcription)
- HTML
- Text-based formats (CSV, JSON, XML)
- ZIP files (iterates over contents)

To install MarkItDown, use pip: `pip install markitdown` . Alternatively, you can install it from the source: `pip install -e .`

Usage

Command-Line

```
markitdown path-to-file.pdf > document.md
```

You can also pipe content:

```
cat path-to-file.pdf | markitdown
```

Python API

About

Python tool for converting files and office documents to Markdown.

[#markdown](#) [#pdf](#) [#openai](#) [#microsoft-office](#)
[#autogen](#) [#langchain](#) [#autogen-extension](#)

-  Readme
-  MIT license
-  Code of conduct
-  Security policy
-  Activity
-  Custom properties
-  24.5k stars
-  75 watching
-  914 forks
- Report repository

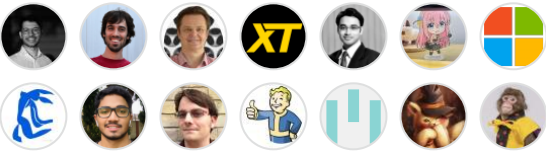
Releases

2 tags

Packages

No packages published

Contributors25



+ 11 contributors

Languages



Basic usage in Python:

```
from markdown import Markdown

md = Markdown()
result = md.convert("test.xlsx")
print(result.text_content)
```

To use Large Language Models for image descriptions, provide `llm_client` and `llm_model` :

```
from markdown import Markdown
from openai import OpenAI

client = OpenAI()
md = Markdown(llm_client=client, llm_model="gpt-4o")
result = md.convert("example.jpg")
print(result.text_content)
```

Docker

```
docker build -t markdown:latest .
docker run --rm -i markdown:latest < ~/your-file.pdf > output.md
```

► Batch Processing Multiple Files

Contributing

This project welcomes contributions and suggestions. Most contributions require you to agree to a Contributor License Agreement (CLA) declaring that you have the right to, and actually do, grant us the rights to use your contribution. For details, visit <https://cla.opensource.microsoft.com>.

When you submit a pull request, a CLA bot will automatically determine whether you need to provide a CLA and decorate the PR appropriately (e.g., status check, comment). Simply follow the instructions provided by the bot. You will only need to do this once across all repos using our CLA.

This project has adopted the [Microsoft Open Source Code of Conduct](#). For more information see the [Code of Conduct FAQ](#) or contact opencode@microsoft.com with any additional questions or comments.

Running Tests and Checks

- Install `hatch` in your environment and run tests:

```
pip install hatch # Other ways of installing hatch: https://hatch.pypa.io/dev/install/
hatch shell
hatch test
```

(Alternative) Use the Devcontainer which has all the dependencies installed:

```
# Reopen the project in Devcontainer and run:
hatch test
```

- Run pre-commit checks before submitting a PR: `pre-commit run --all-files`

Trademarks

This project may contain trademarks or logos for projects, products, or services. Authorized use of Microsoft trademarks or logos is subject to and must follow [Microsoft's Trademark & Brand Guidelines](#). Use of Microsoft trademarks or logos in modified versions of this project must not cause confusion or imply Microsoft sponsorship. Any use of third-party trademarks or logos are subject to those third-party's policies.