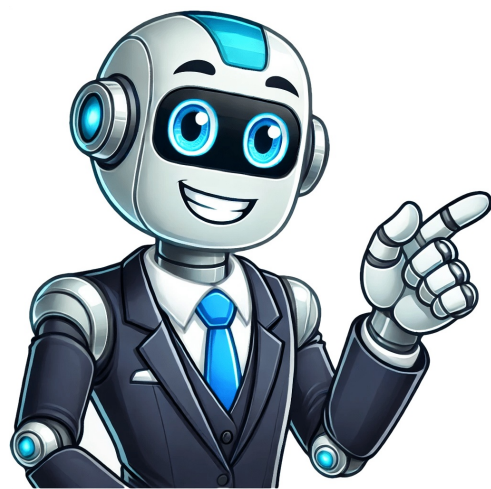


Click to verify



Camelot is a Python library that can help you extract tables from PDFs. Extract tables from PDFs in just a few lines of code: Try it yourself in our interactive quickstart notebook. Or check out a simple example using this pdf. >>> import camelot >>> tables = camelot.read_pdf('foo.pdf') >>> tables >>> tables.export('foo.csv', f='csv', compress=True) # json, excel, html, markdown, sqlite >>> tables[0] >>> tables[0].parsing_report { 'accuracy': 99.02, 'whitespace': 12.24, 'order': 1, 'page': 1 } >>> tables[0].to_csv('foo.csv') # to json, to excel, to html, to markdown, to sqlite >>> tables[0].df # get a pandas DataFrame! Cycle Name KI (1/km) Distance (mi) Percent Fuel Savings Improved Speed Decreased Accel Eliminate Stops Decreased Idle 2012 2 3.30 1.3 5.9% 9.5% 29.2% 17.4% 2145 1 0.68 11.2 2.4% 0.1% 9.5% 2.7% 4234 1 0.59 58.7 8.5% 1.3% 8.5% 3.3% 2032 2 0.17 57.8 21.7% 0.3% 2.7% 1.2% 4171 1 0.07 173.9 58.1% 1.6% 2.1% 0.5% Camelot also comes packaged with a command-line interface! Refer to the QuickStart Guide to quickly get started with Camelot, extract tables from PDFs and explore some basic options. Tip: Visit the parser-comparison-notebook to get an overview of all the packed parsers and their features. Note: Camelot only works with text-based PDFs and not scanned documents. (As Tabula explains, "If you can click and drag to select text in your table in a PDF viewer, then your PDF is text-based".) You can check out some frequently asked questions here. Why Camelot? Configurability: Camelot gives you control over the table extraction process with tweakable settings. Metrics: You can discard bad tables based on metrics like accuracy and whitespace, without having to manually look at each table. Output: Each table is extracted into a pandas DataFrame, which seamlessly integrates into ETL and data analysis workflows. You can also export tables to multiple formats, which include CSV, JSON, Excel, HTML, Markdown, and SqLite. See comparison with similar libraries and tools. Installation Using conda The easiest way to install Camelot is with conda, which is a package manager and environment management system for the Anaconda distribution. conda install -c conda-forge camelot-py Using pip After installing the dependencies (tk and ghostscript), you can also just use pip to install Camelot: pip install "camelot-py[base]" From the source code After installing the dependencies, clone the repo using: git clone and install using pip: cd camelot pip install ". Documentation The documentation is available at . Wrappers camelot-php provides a PHP wrapper on Camelot. Related projects camelot-sharp provides a C sharp implementation of Camelot. Contributing The Contributor's Guide has detailed information about contributing issues, documentation, code, and tests. Versioning Camelot uses Semantic Versioning. For the available versions, see the tags on this repository. For the changelog, you can check out the releases page. License This project is licensed under the MIT License, see the LICENSE file for details. This repository was archived by the owner on Jan 6, 2025. It is now read-only. You can't perform that action at this time. Camelot is a Python library that can help you extract tables from PDFs: Note: You can also check out Excalibur, the web interface to Camelot! Here's how you can extract tables from PDFs. You can check out the PDF used in this example here. >>> import camelot >>> tables = camelot.read_pdf('foo.pdf') >>> tables >>> tables.export('foo.csv', f='csv', compress=True) # json, excel, html, markdown, sqlite >>> tables[0] >>> tables[0].parsing_report { 'accuracy': 99.02, 'whitespace': 12.24, 'order': 1, 'page': 1 } >>> tables[0].to_csv('foo.csv') # to json, to excel, to html, to markdown, to sqlite >>> tables[0].df # get a pandas DataFrame! Cycle Name KI (1/km) Distance (mi) Percent Fuel Savings Improved Speed Decreased Accel Eliminate Stops Decreased Idle 2012 2 3.30 1.3 5.9% 9.5% 29.2% 17.4% 2145 1 0.68 11.2 2.4% 0.1% 9.5% 2.7% 4234 1 0.59 58.7 8.5% 1.3% 8.5% 3.3% 2032 2 0.17 57.8 21.7% 0.3% 2.7% 1.2% 4171 1 0.07 173.9 58.1% 1.6% 2.1% 0.5% Camelot also comes packaged with a command-line interface! Note: Camelot only works with text-based PDFs and not scanned documents. (As Tabula explains, "If you can click and drag to select text in your table in a PDF viewer, then your PDF is text-based".) You can check out some frequently asked questions here. Configurability: Camelot gives you control over the table extraction process with tweakable settings. Metrics: You can discard bad tables based on metrics like accuracy and whitespace, without having to manually look at each table. Output: Each table is extracted into a pandas DataFrame, which seamlessly integrates into ETL and data analysis workflows. You can also export tables to multiple formats, which include CSV, JSON, Excel, HTML, Markdown, and SqLite. See comparison with similar libraries and tools. If Camelot has helped you, please consider supporting its development with a one-time or monthly donation on OpenCollective. The easiest way to install Camelot is with conda, which is a package manager and environment management system for the Anaconda distribution. \$ conda install -c conda-forge camelot-py After installing the dependencies (tk and ghostscript), you can also just use pip to install Camelot: \$ pip install "camelot-py[base]" After installing the dependencies, clone the repo using: \$ git clone and install using pip: \$ cd camelot \$ pip install ". [base]" The documentation is available at . camelot-php provides a PHP wrapper on Camelot. The Contributor's Guide has detailed information about contributing issues, documentation, code, and tests. Camelot uses Semantic Versioning, and the available versions, see the tags on this repository. For the changelog, you can check out HISTORY.md. This project is licensed under the MIT License, see the LICENSE file for details. You will learn how you can extract tables in PDF using camelot library in Python. Camelot is a Python library and a command-line tool that makes it easy for anyone to extract data tables trapped inside PDF files.This part of the documentation covers the steps to install Camelot.Using condaThe easiest way to install Camelot is to install it with conda, which is a package manager and environment management system for the Anaconda distribution.\$ conda install -c conda-forge camelot-pyUsing pipAfter installing the dependencies, which include Tkinter and ghostscript, you can simply use pip to install Camelot:\$ pip install camelot-py[cv]For more information, check the official documentationThe PDF used in this tutorial can be downloaded from hereimport camelotTo extract the PDF# PDF file to extract tables fromfile = "foo.pdf"The PDF file called "foo.pdf" is a normal page that contains one table shown in the image below.TableExtracting all tables in the PDF filetables = camelot.read_pdf(file)"read_pdf()" function extracts all tables in a PDF file.To print the number of tables extracted:# number of tables extractedprint("Total tables extracted:", tables.n)Output:It contains only one table.Printing this table as a Pandas DataFrame:# print the first table as Pandas DataFrameprint(tables[0].df)Output:Exporting the table to a CSV file:# export individuallytables[0].to_csv("foo.csv")Or if you want to export all tables all at once:# or export all in a zipables.export("foo.csv", f="csv", compress=True)The tables can be exported to HTML format:# export to HTMLtables.export("foo.html", f="html")It can also be exported to JSON and Excel formats.Note: Camelot only works with text-based PDFs and not scanned documents When working on Windows, the easiest way to get up and running is through the Conceptive Python SDK. This SDK is a Python distribution targeted at the development and deployment of Qt based applications. This all in one installation of Camelot with all its dependencies is available in the shop. First, make sure you have setup tools installed. Setup tools. If you are using a debian based distribution, you can type: sudo apt-get install python-setuptools Then use easy install to install Camelot, under Linux this would be done by typing: sudo easy_install camelot Linux distributions often offer packages for various applications, including Camelot and its dependencies . When installing Camelot from source, you need to make sure all dependencies are installed and available in your PYTHONPATH. Dependencies In addition to PyQt 4.8 and Qt 4.8, Camelot needs these libraries : SQLAlchemy==0.8.0 Jinja2==2.6 chardet==2.1.1 xlwt==0.7.4 xlrd==0.9.0 Releases The source code of a release can be downloaded from the Python Package Index and then extracted: tar xzvf Camelot-10.07.02.tar.gz Repository The latest and greatest version of the source can be checked out from the Bitbucket repository: hg clone Adapting PYTHONPATH You need to make sure Camelot and all its dependencies are in the PYTHONPATH before you start using it. To verify if you have Camelot installed and available in the PYTHONPATH, fire up a python interpreter: and issue these commands: >>> import camelot >>> print camelot.__version__ >>> import sqlalchemy >>> print sqlalchemy.__version__ >>> import PyQt4 None of them should raise an ImportError. © Copyright 2009 - 2014, Conceptive Engineering. Last updated on Dec 19, 2016. Sphinx theme provided by Read the Docs