

**pdfcrawl**

**version 0.0.1**

**Search & Extract Utility**

July 02, 2021



# Contents

<b>Welcome to pdfcrawl's documentation!</b>	<b>1</b>
pdfcrawl	1
Getting Started	1
Parameter	2
Release Notes	2
<b>0.0.1</b>	<b>2</b>



# Welcome to pdfcrawl's documentation!

## pdfcrawl

**pdfcrawl** utility is meant for searching vast pdf files for certain policy and extract the pages where the match is found.

## Getting Started

- Installation

The utility can be installed and run in multiple ways.

1. Get the latest executable from [Release](#) section of GitHub.
2. Get the **.whl** file from [Release](#) section of GitHub. Then install it using pip command.

### Note

```
pip install pdfcrawl-0.0.1-py3-none-any.whl
```

3. Clone the repo from [here](#)

- Running

1. **Help**

### Note

```
usage: app.py [-h] [-d] -f FILES [FILES ...] [-l LOGFILE] [-n NUM] -g SEARCH_LIST
[SEARCH_LIST ...] -s STATE
```

A commandline utility to crawl on pdf files.

#### optional arguments:

-h, --help

show this help message and exit

-d, --debug

Enable debug output

-f

FILES [FILES ...], --files FILES [FILES ...]  
The absolute path to all files separated by white space.

-l LOGFILE, --log LOGFILE

Log file name.

-n NUM, --num NUM

Number of files to operate on simultaneously. Should not be a number greater than processor in your computer. Default is 4.

-g

SEARCH\_LIST [SEARCH\_LIST ...],  
-grep SEARCH\_LIST [SEARCH\_LIST ...] Strings to search. The order of the string is preserved.

-s STATE, --state STATE

State for which policy needs to be extracted.

1. **Running the executable.**

### Note

```
pdfcrawl -f "pdf_file1","pdf_file2",pdf_file3 -g Headline of Insurance Document -s state
```

#### 1. Running from source code.

### Note

```
C:\Workspace> python -m pdfcrawl.pdfcrawl.app -f "pdf_file1","pdf_file2",pdf_file3 -g Headline of Insurance Document -s state
```

## Parameter

#### 1. Running the executable.

### Note

```
pdfcrawl -f "pdf_file1","pdf_file2",pdf_file3 -g Headline of Insurance Document -s state
```

#### 2. MANDATORY PARAMETERS

- -f / files

### Note

Accept Multiple PDF Files as input. Please note the files in windows have space in it. SO make Sure to quote the file with double("") quotes. For e.g: "D:\Smruti\Python\Practice\pdfscraper\SGIO Amendment Landlord Buildings & Contents\_Test.pdf"

- -g / grep

### Note

Pass the Headlines of the Insurance. This is case sensitive and is searched the order in which it is passed.

- -s / state

### Note

The State of Policy.

## Release Notes

### 0.0.1

Parameter

1. Initial Release