# Introduction

The file, *create\_toc.py*, is a Python program designed to be run from the command line. The program creates a table of contents text file called, *refined\_table\_of\_contents.txt*, for an input file called *2025 PFS QPP NPRM -14828.pdf*. Both files must be in the same folder.

This program was designed to help analyze Federal Regulations that are released in PDF form without a table of contents.

The program did not get every section header in the input file correct, but it got the large majority of the headers and pages correct.

The program, create\_toc\_in\_csv.py, outputs the table of contents into a .csv file with the section header in the first column and the page number in the second column. It doesn’t get every header and page number correct, but gets the large majority correct.

# Prerequisites

* You must have a Python environment set up on your machine (recommend Anaconda – get it here: <https://www.anaconda.com/download> )
* Install the following libraries: PyMuPDF, transformers, torch

A black rectangular object with white text

Description automatically generated

# Running the Program

1. Make sure the prerequisites are met
2. Open an anaconda prompt
3. Navigate to the folder where the files, *create\_toc.py* and *2025 PFS QPP NPRM -14828.pdf*, are located
4. If you are using a different PDF file for input, change the pdf\_path variable in *create\_toc.py* appropriately
5. Run the python program

A black rectangle with a white border

Description automatically generated

1. Wait for the program to finish

A screen shot of a computer screen

Description automatically generated

# Code Snippet Where the Input PDF File is Defined

The code snippet below is where the input PDF file is specified. You’ll need to change the definition of pdf\_path if you are using a PDF file with a different name from "2025 PFS QPP NPRM -14828.pdf".

def main():

pdf\_path = "2025 PFS QPP NPRM -14828.pdf"

pdf\_text = extract\_text\_from\_pdf(pdf\_path)

headers = identify\_headers(pdf\_text)

toc = create\_toc(headers)

refined\_toc = refine\_toc\_with\_gpt2(toc)

# Save the refined TOC to a file with UTF-8 encoding

with open("refined\_table\_of\_contents.txt", "w", encoding="utf-8") as file:

file.write(refined\_toc)

print("Table of contents created and saved to 'refined\_table\_of\_contents.txt'.")