



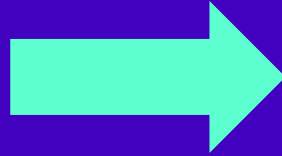
# Intro to PDF Text & Table Extraction

Anna Godwin  
December 2022



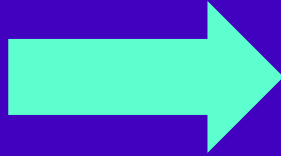
PDFs are \_\_\_\_\_.

PDFs are \_\_\_\_\_ a dataset \_\_\_\_\_.



Lorem ipsum dolor sit amet. Id numquam recusandae et omnis praesentium et laudantium voluptas a numquam impedit. Ut voluptas sint et accusantium soluta eos totam galisum et quasi galisum ad labore ullam in fugiat obcaecati sed eveniet dolorem. Qui soluta voluptas et perferendis omnis non obcaecati labore ut similique beatae? Quo minima galisum aut dolor mollitia et cupiditate nesciunt.

**Unstructured,  
text data**



Pet Name	Pet Type	Trait
Butter	Cat	Pbrrrt
Sugar	Cat	Flops
Sparky	Dog	Wags

**Structured,  
tabular data**

**Unstructured,  
text data**

**+**

**Structured,  
tabular data**



# Common Goals in PDF Extraction

with Python starter code



# Goal 1: Search the Text



# Search the text: pymupdf

```
pip install pymupdf
```

```
import fitz

doc = fitz.open("sample.pdf")

text_list = list()
for page in doc:
    text_list.append(page.get_textpage().extractText())

all_text_str = " ".join(collate_text_list)
```

A series of yellow lightning bolts with green outlines are scattered along the diagonal boundary between the purple and white background.

**Goal 2:**  
**Condense into a**  
**Smaller, Relevant PDF**

# Condense PDF: pymupdf

```
import fitz

doc_orig = fitz.open("sample.pdf")

doc_short = fitz.open() # initialize an empty pdf
doc_short.insert_pdf(doc_orig, from_page=2, to_page=5)
doc_short.insert_pdf(doc_orig, from_page=10, to_page=11)
doc_short.save("short_pdf.pdf")
doc_short.close()
```



# **Goal 3:**

## **Collate Multiple PDFs Into a Single PDF**

# Collate PDFs: pymupdf

```
import fitz

doc_collate = fitz.open() # initialize an empty pdf
for pdf_file in pdf_directory:
    with fitz.open(pdf_file) as f:
        doc_collate.insert_pdf(f)

doc_collate.save("collated_pdf.pdf")
doc_collate.close()
```



# Goal 4: Extract Tables

# Extract Tables: tabula-py

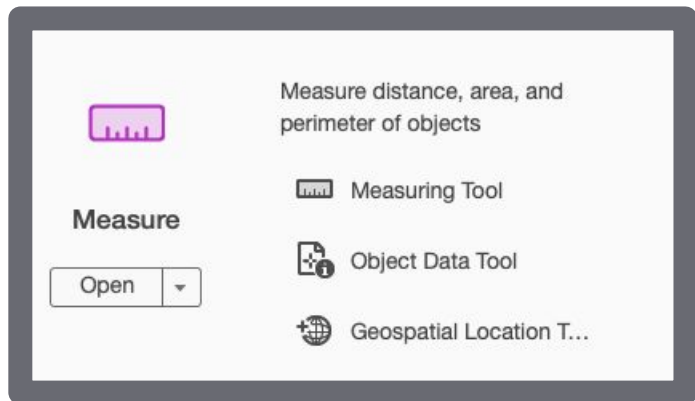
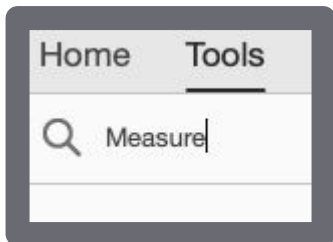
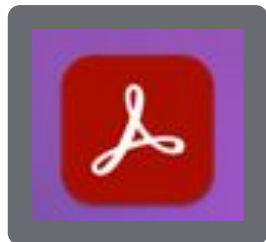
```
pip install tabula-py
```

Note: These settings do pretty well out of the box on a full width table in a PDF.

```
import tabula

output = tabula.read_pdf("test.pdf",
                        guess=False,
                        lattice=False,
                        stream=True,
                        pages=1)
```

# Extract Tables: tabula-py



	2006-07 (2003 Total Cohort four years later as of June 30, 2007)	2006-07 (2002 Total Cohort five years later as of June 30, 2007)
• Number of students with disabilities who first entered 9th grade anywhere (or if ungraded, became 17 years old) in 2003-04	9	
• Number of students with disabilities who first entered 9th grade anywhere (or if ungraded, became 17 years old) in 2002-03		9
• Graduation rate	0%	33.3%
• State target for 2006-07	37% or higher	No State Target
• Meets State target?	Not Applicable*	Not Applicable
* Districts are only held accountable for the performance of students when there are at least 30 students in the total cohort.		



# Extract Tables: tabula-py

Note: Convert the inch measurements to centimeters from the left side of the page

Column Width (in)	Column Width (cm)	Distance from Left Edge (cm)
0.13	0.3	0.3
4.01	10.2	10.5
2.01	5.1	15.6
2.00	5.1	20.7

Indicator 1: Graduation Rate of Students with Disabilities

	2006-07 (2003 Total Cohort four years later as of June 30, 2007)	2006-07 (2002 Total Cohort five years later as of June 30, 2007)
• Number of students with disabilities who first entered 9th grade anywhere (or if ungraded, became 17 years old) in 2003-04	9	
• Number of students with disabilities who first entered 9th grade anywhere (or if ungraded, became 17 years old) in 2002-03		9
• Graduation rate	0%	33.3%
• State target for 2006-07	37% or higher	No State Target
• Meets State target?	Not Applicable*	Not Applicable
* Districts are only held accountable for the performance of students when there are at least 30 students in the total cohort.		

Diagram showing column widths and distances from the left edge:

- 0.13 in (margin)
- 4.01 in (first column width)
- 2.01 in (second column width)
- 2.00 in (third column width)

# Extract Tables: tabula-py

Distance from Left Edge (cm)
0.3
10.5
15.6
20.7

```
import tabula

column_meas = [0.3, 10.5, 15.6, 20.7]

output = tabula.read_pdf("test.pdf",
                          pages=1,
                          columns=column_meas)
```

PDFs are \_\_\_\_\_ a dataset \_\_\_\_\_.

“

**Thank you!!!**

**Slides & sample code @  
[github.com/annagodwin/normconf-intro-pdf](https://github.com/annagodwin/normconf-intro-pdf)**

”