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

Image credit: https://commons.wikimedia.org/wiki/File:PDF_file_icon.svg

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

Structured, tabular data

Image credit: https://commons.wikimedia.org/wiki/File:PDF_file_icon.svg

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)
```

https://pymupdf.readthedocs.io/

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()
```

https://pymupdf.readthedocs.io/

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()
```

https://pymupdf.readthedocs.io/

Goal 4:
Extract Tables

Extract Tables: tabula-py

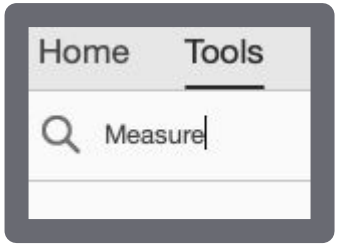
```
pip install tabula-py
```

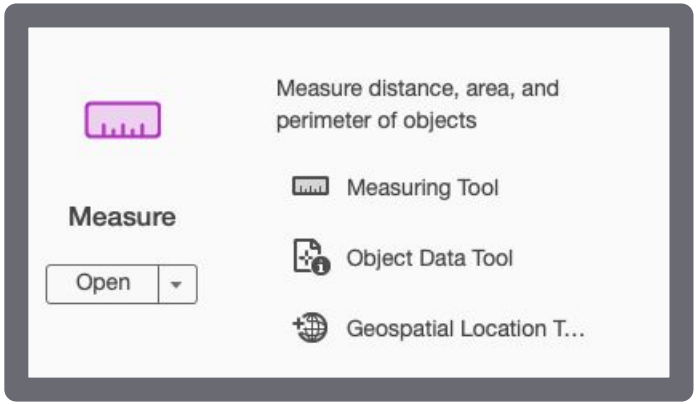
Note: These settings do pretty well out of the box on a full width table in a PDF.

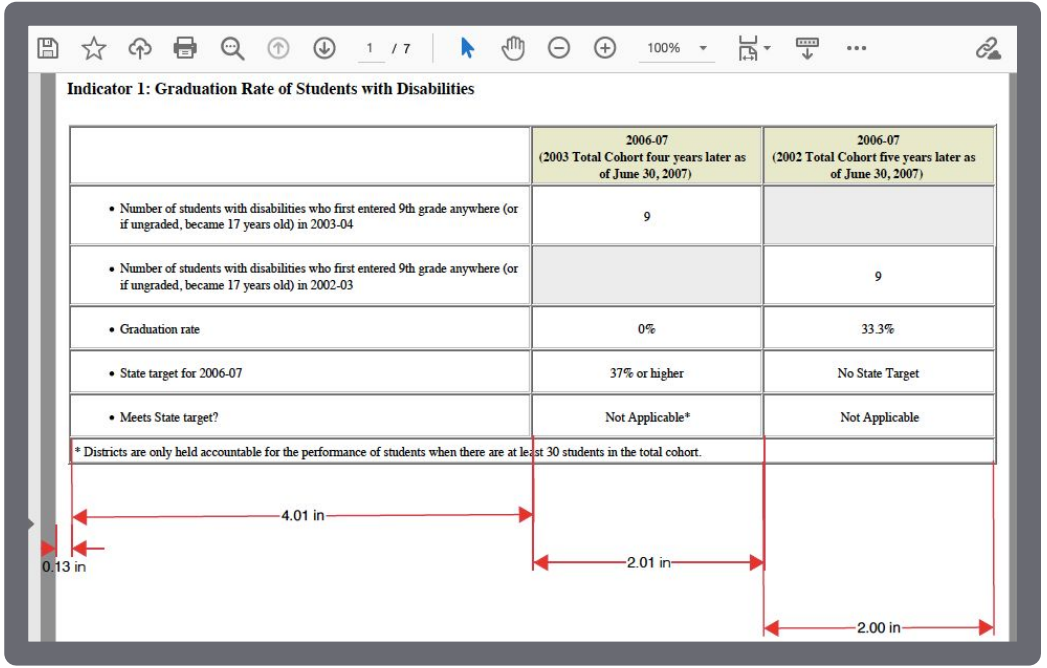
```
import tabula

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

https://tabula-py.readthedocs.io/

Extract Tables: tabula-py





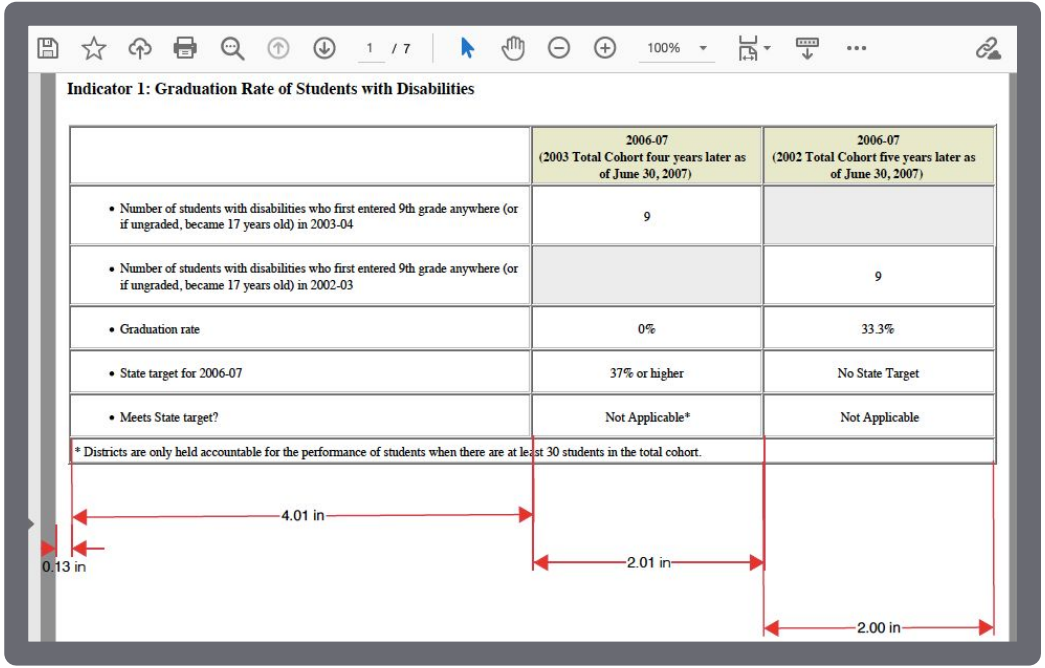


https://tabula-py.readthedocs.io/

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



https://tabula-py.readthedocs.io/

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)
```

https://tabula-py.readthedocs.io/

PDFs are ___a dataset___.

Thank you!!!

Slides & sample code @
github.com/annagodwin/normconf-intro-pdf