NAME-KASA SATYA SAMPATH KUMAR

Project Documentation: PDF to JSON Extractor using Streamlit

1. Project Title

PDF Table Extractor and JSON Converter

---

2. Objective

The objective of this project is to extract tabular data from PDF files, display it on a web-based interface, and convert it into JSON format.
This process simplifies the handling of structured datasets stored in PDFs, making them machine-readable and usable for further applications like data analysis, visualization, and machine learning.

---

3. Technologies Used

- **Python** – Core programming language.
- **Streamlit** – For building the frontend interface.
- **pdfplumber** – To read and extract tabular data from PDF documents.
- **pandas** – To structure and clean tabular data into DataFrames.
- **json** – To convert processed data into JSON format.

---

4. Code Implementation

```
import streamlit as st

import pdfplumber

import pandas as pd

import json

# File uploader

uploaded_file = st.file_uploader("Upload a Dataset PDF", type=["pdf"])

if uploaded_file is not None:

    with pdfplumber.open(uploaded_file) as pdf:

        table = pdf.pages[0].extract_table()

        if table:
```

```
# Convert table into dataframe
df = pd.DataFrame(table[1:], columns=table[0])
# Clean dataframe
df = df.dropna(axis=1, how="all")
df = df.loc[:, df.columns.notna()]
df = df.loc[:, df.columns.str.strip() != ""]
df = df.loc[:, ~df.columns.duplicated()]
# Display dataframe
st.write(df)
# Convert to JSON
json_string = json.dumps(df.to_dict(orient="records"), indent=4)
st.code(json_string, language="json")
 # Save JSON to file
with open("output.json", "w", encoding="utf-8") as f:
   f.write(json_string)
```

---

5. Working Procedure

1. The user uploads a PDF file (dataset.pdf) containing a structured table.
2. The **first page** of the PDF is processed using the pdfplumber library.
3. The extracted table is converted into a **pandas DataFrame**.
4. Data cleaning is performed by:
    - Dropping empty columns.
    - Removing blank headers.
    - Removing duplicate column names.
5. The clean dataset is displayed in a **Streamlit table**.
6. The DataFrame is converted into **JSON format**.
7. The JSON output is displayed on the frontend and saved locally as **output.json**.

NAME-KASA SATYA SAMPATH KUMAR

6. Screenshots

ID	Category	Sub-Category	Entity Name	Address	Postcode	Local Authority	Region	Policy Area	Relevant Policy/Legislation	Policy Summary	Status	Funding (GBP Million)	URL
1	Healthcare	Hospital	St Thomas' Hospital	Westminster Bridge Rd, London	SE1 7EH	Lambeth	London	Health & Social Care	Health and Care Act 2022	Aims to improve the integration of health and social	Operational	1500	https://www.guysandstthomas.nhs.uk/
2	Education	University	University of Manchester	Oxford Rd, Manchester	M13 9PL	Manchester	North West	Higher Education	Higher Education (Freedom of Speech) Act 2023	Strengthens freedom of speech and academic freedom in higher education.	Operational	1200	https://www.manchester.ac.uk/
3	Business	Aerospace	Rolls-Royce Holdings plc	62 Buckingham Gate, London	SW1E 6AT	Westminster	London	Industrial Strategy	UK Research and Development Roadmap	Outlines the government's vision for science and innovation.	Operational	N/A	https://www.rolls-royce.com/
4	Residential	Housing Estate	Barbican Estate	Silk St, London	EC2Y 8DS	City of London	London	Housing & Planning	National Planning Policy Framework	Sets out the government's planning policies for England.	Completed	N/A	https://www.barbican.org.uk/
5	Healthcare	Hospital	Addenbrooke's Hospital	Hills Rd, Cambridge	CB2 0QQ	Cambridge	East of England	Health & Social Care	NHS Long Term Plan	Sets out the priorities for the NHS for the next 10 years.	Operational	950	https://www.cuh.nhs.uk/
6	Education	College	Birmingham Metropolitan College	Jennens Road, Birmingham	B4 7PS	Birmingham	West Midlands	Further Education	Skills and Post-16 Education Act 2022	Aims to improve the quality of post-16 education and training.	Operational	75	https://www.bmet.ac.uk/
7	Business	Defence	BAE Systems	Stirling Square, London	SW1Y 5HD	Westminster	London	Defence & Security	National Security Strategy	Sets out the government's approach to protecting the UK's security.	Operational	N/A	https://www.baesystems.com/
8	Residential	New Town	Poundbury	Poundbury, Dorchester	DT1 3GE	Dorset	South West	Urban Planning	Levelling Up and Regeneration Act 2023	Aims to reduce regional inequalities across the UK.	Under Development	N/A	https://poundbury.co.uk/
9	Healthcare	Hospital	Queen Elizabeth University Hospital	1345 Govan Rd, Glasgow	G51 4TF	Glasgow City	Scotland	Health & Social Care	NHS Recovery Plan (Scotland)	Sets out the Scottish Government's plan for health and social care recovery.	Operational	842	https://www.nhsggc.scot/
10	Education	University	Cardiff University	Park Pl, Cardiff	CF10 3AT	Cardiff	Wales	Higher Education	Tertiary Education and Research (Wales) Act 2022	Establishes a new framework for tertiary education and research in Wales.	Operational	580	https://www.cardiff.ac.uk/
11	Business	Pharmaceuticals	AstraZeneca plc	1 Francis Crick Avenue, Cambridge	CB2 0AA	Cambridge	East of England	Life Sciences	Life Sciences Vision	Sets out a 10-year strategy for the UK life sciences sector.	Operational	N/A	https://www.astrazeneca.com/
12	Residential	Regeneration	Hulme Crescents Regeneration	Hulme, Manchester	M15	Manchester	North West	Urban Regeneration	The Northern Powerhouse Strategy	Aims to boost economic growth in the North of England.	Completed	140	https://www.manchester.gov.uk/
13	Healthcare	Hospital	Royal Victoria Infirmary	Queen Victoria Rd, Newcastle upon Tyne	NE1 4LP	Newcastle upon Tyne	North East	Health & Social Care	People at the Heart of Care: adult social care reform	Sets out a 10-year vision for adult social care in England.	Operational	600	https://www.newcastle-hospitals.nhs.uk/
14	Education	College	Exeter College	Hele Rd, Exeter	EX4 4JS	Devon	South West	Further Education	T Levels	New two-year technical qualifications for 16-19 year olds.	Operational	60	https://exe-coll.ac.uk/
15	Business	Consumer Goods	Unilever	Unilever House, London	EC4Y 0DY	City of London	London	Consumer Goods	The UK's Green Industrial Revolution	A ten-point plan for a green industrial revolution.	Operational	N/A	https://www.unilever.co.uk/
16	Residential	Housing Development	Kidbrooke Village	Kidbrooke, London	SE3	Greenwich	London	Housing Development	The London Plan	The spatial development strategy for the Greater London area.	Under Development	1000	https://www.berkeleygroup.co.uk/developments/london/greenwich/kidbrooke-village
17	Healthcare	Hospital	Belfast City Hospital	Lisburn Rd, Belfast	BT9 7AB	Belfast	Northern Ireland	Health & Social Care	Health and Wellbeing 2026: Delivering Together	Sets out the strategic direction for health and social care in Northern Ireland.	Operational	150	https://belfasttrust.hscni.net/
18	Education	University	University of Edinburgh	Old College, South Bridge, Edinburgh	EH8 9YL	City of Edinburgh	Scotland	Higher Education	The Promise (Scotland)	Aims to improve the lives of children and young people in care.	Operational	1300	https://www.ed.ac.uk/
19	Business	Automotive	Jaguar Land Rover	Abbey Rd, Whitley, Coventry	CV3 4LF	Coventry	West Midlands	Automotive Industry	The Road to Zero	A strategy for zero-emission road transport.	Operational	N/A	https://www.jaguarlandrover.com/
20	Residential	Heritage	Grainger Town, Newcastle	Newcastle upon Tyne	NE1	Newcastle upon Tyne	North East	Heritage & Conservation	National Heritage Act 1980	Makes provision for the protection of ancient monuments and historic buildings.	Completed	40	https://www.newcastle.gov.uk/
21	Healthcare	Cancer Centre	The Christie NHS Foundation Trust	Wilmslow Rd, Manchester	M20 4BX	Manchester	North West	Cancer Care	NHS Cancer Programme	Aims to improve cancer outcomes and services in England.	Operational	300	https://www.christie.nhs.uk/
22	Education	College	Leeds City College	Park Ln, Leeds	LS3 1AA	Leeds	Yorkshire and the H	Further Education	Local Skills Improvement Plans	Aims to put employers at the heart of the skills system.	Operational	70	https://www.leedscitycollege.ac.uk/
23	Business	Pharmaceuticals	GlaxoSmithKline (GSK)	980 Great West Rd, Brentford	TW8 9GS	Hounslow	London	Pharmaceuticals	UK Biopharma Manufacturing Vision	Aims to make the UK the best place in the world to manufacture biopharmaceuticals.	Operational	N/A	https://www.gsk.com/
24	Residential	Social Housing	Park Hill Flats, Sheffield	Sheffield	S2	Sheffield	Yorkshire and the H	Social Housing	Social Housing (Regulation) Act 2023	Aims to improve the regulation of social housing and strengthen tenants' rights.	Completed	100	https://www.urbansplash.co.uk/regeneration/park-hill
25	Healthcare	Children's Hospital	Great Ormond Street Hospital	Great Ormond St, London	WC1N 3JH	Camden	London	Children's Health	Children and Young People's Health Policy	Aims to improve the health and wellbeing of children and young people.	Operational	400	https://www.gosh.nhs.uk/
26	Education	University	Imperial College London	Exhibition Rd, London	SW7 2AZ	Kensington and Chelsea	London	Science & Technology	UK Innovation Strategy	Sets out the government's plan to make the UK a global hub for innovation.	Operational	1100	https://www.imperial.ac.uk/
27	Business	Retail	Tesco PLC	Tesco House, Shire Park, Kestrel Way	AL7 1GA	Welwyn Hatfield	East of England	Retail	The UK's Food Strategy	Sets out the government's vision for a sustainable food system.	Operational	N/A	https://www.tescoplc.com/
28	Residential	Regeneration	New Islington, Manchester	Manchester	M4	Manchester	North West	Urban Regeneration	The Estates Gazette Regeneration Report	Provides insights into the UK's most active regeneration locations.	Completed	300	https://www.urbansplash.co.uk/regeneration/new-islington
29	Healthcare	Cancer Centre	The Royal Marsden NHS Foundation T	203 Fulham Rd, London	SW3 6JJ	Kensington and Chelsea	London	Cancer Research	Dame Tessa Jowell Brain Cancer Mission	Aims to transform brain cancer research and care in the UK.	Operational	250	https://www.royalmarsden.nhs.uk/
30	Education	University	Queen's University Belfast	University Rd, Belfast	BT7 1NN	Belfast	Northern Ireland	Higher Education	The Higher Education Strategy for Northern Ireland	Sets out the strategic direction for higher education in Northern Ireland.	Operational	400	https://www.qub.ac.uk/
31	Business	Technology	DeepMind	6 Pancras Square, London	N1C 4AG	Camden	London	Artificial Intelligence	National AI Strategy	A 10-year plan to make Britain a global AI superpower.	Operational	N/A	https://deepmind.google/
32	Residential	Garden Village	Ebbsfleet Garden City	Ebbsfleet Valley, Kent	DA10	Dartford and Gravesham	South East	Housing & Planning	Garden Communities Programme	A government programme to support the development of new garden towns and villages.	Under Development	250	https://ebbsfleetdc.org.uk/
33	Healthcare	Mental Health	The Priory Hospital Roehampton	Priory Lane, London	SW15 5JJ	Wandsworth	London	Mental Health	NHS Mental Health Implementation Plan	Sets out the delivery plan for the mental health aspects of the NHS Long Term Plan.	Operational	50	https://www.priorygroup.com/locations/priory-hospital-roehampton

streamlit > pdf_text.py > ...

```python
import streamlit as st
import pdfplumber
import pandas as pd
import json
uploaded_file = st.file_uploader("Upload a Dataset PDF", type=["pdf"])
if uploaded_file is not None:
    with pdfplumber.open(uploaded_file) as pdf:
        table = pdf.pages[0].extract_table()
        if table:
            df = pd.DataFrame(table[1:], columns=table[0])
            df = df.dropna(axis=1, how="all")
            df = df.loc[:, df.columns.notna()]
            df = df.loc[:, df.columns.str.strip() != ""]
            df = df.loc[:, ~df.columns.duplicated()]
            st.write(df)
            json_string = json.dumps(df.to_dict(orient="records"), indent=4)
            st.code(json_string, language="json")
            with open("output.json", "w", encoding="utf-8") as f:
                f.write(json_string)
```

---

8. Sample Output


```json
[
    {
        "ID": "1",
        "Category": "Healthcare",
        "Sub-Category": "Hospital",
        "Entity Name": "St Thomas' Hospital",
        "Address": "Westminster Bridge Rd, London",
        "Postcode": "SE1 7EH",
        "Local Authority": "Lambeth",
        "Region": "London",
        "Policy Area": "Health & Social Care",
        "Relevant Policy/Legislation": "Health and Care Act 2022",
        "Policy Summary": "Aims to improve the integration of health and social",
        "Status": "cOapree rsaetirovnicaels.",
        "Funding (GBP Million)": "1500",
        "URL": "https://www.guysandstthomas.nhs.uk/"
    },
    {
```

---

8. Applications

- Extracting structured data from **reports, invoices, research papers, or government datasets**.
- Converting datasets into **machine-readable JSON** for APIs and web apps.
- Data preprocessing for **machine learning and AI projects**.
- Creating digital pipelines for **policy analysis, healthcare, education, and business datasets**.

---

9. Conclusion

This project demonstrates how PDF data can be converted into structured formats for further analysis. Using **Streamlit for UI**, **pdfplumber for extraction**, and **pandas for cleaning**, the system successfully transforms static tables into **dynamic JSON files**, making data more accessible and reusable for modern applications.