PDFs by the batch

Treating lots of documents like data

Dan Keemahill, he/him
USA TODAY data team
dkeemahill@gannett.com

Tools!

PDF and web scraping from Journalist's Toolbox

Parsing prickly PDFs guide from creators of pdfplumber

Intermediate PDF Manipulation by co-founder of CitizenAudit.org

Google Pinpoint – USA TODAY Network training Sept. 1

Best option: Don't use PDFs for data

Specify a data file type – CSV, Excel spreadsheet, database file, etc. – when filing public records requests.

Take extra care when analyzing data from PDFs

Document your process

Optical Character Recognition (OCR): Beware of errors, double-check numbers before performing calculations

Documents are the data

Documents have metadata

File name, numbers of pages, size, date created, categories

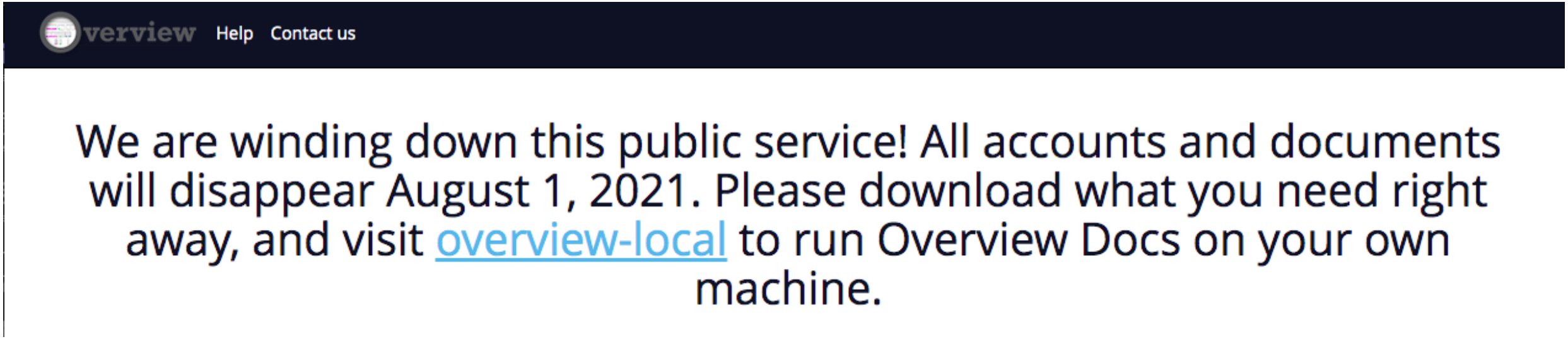
Useful for keeping track of batches of documents

Overview

A free tool that once lived at overviewdocs.com

The software is available as open source, but requires some setup to run

The Overview repository also has utilities for specific tasks



Overview

Performs OCR on image PDFs

Searches across all documents in a collection

Organizes documents with tags

Analyzes "entities" that appear in the documents

Generates a "topic tree" based on keywords

Exports spreadsheets of documents and their text

Demo: Overview

Analyzing bills banning transgender athletes

Created a single database of multiple file types

Assigned tags based on keywords

Identified studies cited in multiple bills

My go-to tools for PDFs with text

Tabula

See USA TODAY Network training

Scans tables

Script-able with libraries in languages including R, Python, JavaScript

pdfplumber

Python library

Extracts tables or text

Pixel-level precision and lots of configuration options

Demo: pdfplumber

This notebook installs pdfplumber and ImageMagick, downloads a PDF report and parses it to a Pandas dataframe

Uses bounding boxes to get data that always appear in the same location

Demo: More pdfplumber

This notebook also installs pdfplumber and ImageMagick, downloads a PDF report and parses it to a Pandas dataframe

Crops a page with multiple tables to get the desired table

Case study: Off Target

The Trace and USA TODAY analyzed 2,000 ATF inspections

Image PDFs that required OCR

Team used 20-step process to enter data in a spreadsheet (Airtable) with data validation

Con: Time consuming

Pro: Confident in accuracy, find stories

Always be learning!

Join USA TODAY Network Data in Microsoft Teams

Visit training.usatodaynetwork.com